陈木法文选

（卷四）

Mu-Fa Chen

Selected Papers

(Volume IV)

Beijing Normal University

Contents

Volume IV

Volume I

Volume II

Volume III

Front. Math. China 2014, 9(4): 761–796
DOI 10.1007/s11464-013-0186-5

Unified representation of formulas for single birth processes

Mu-Fa CHEN, Yu-Hui ZHANG

School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems, Ministry of Education, Beijing 100875, China

Abstract Based on a new explicit representation of the solution to the Poisson equation with respect to single birth processes, the unified treatment for various criteria on classical problems (including uniqueness, recurrence, ergodicity, exponential ergodicity, strong ergodicity, as well as extinction probability etc.) for the processes are presented.

Keywords Single birth process, Poisson equation, uniqueness, recurrence, ergodicity, moments of return time
MSC 60J60

1 Introduction

Consider a continuous-time homogeneous Markov chains $\{X(t) : t \geqslant 0\}$, on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with transition probability matrix $P(t) = (p_{ij}(t))$ on a countable state space $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$. We call $\{X(t) : t \geqslant 0\}$ a single birth process if its transition rate (density) matrix $Q = (q_{ij} : i, j \in \mathbb{Z}_+)$ is irreducible and satisfies that $q_{i,i+1} > 0$, $q_{i,i+j} = 0$ for all $i \in \mathbb{Z}_+$ and $j \geqslant 2$. Such a matrix $Q = (q_{ij})$ with $\sum_j q_{ij} = 0$ for every i (conservativity) is called a single birth Q-matrix. Refer to [15]. In the literature, the single birth process is also called upwardly skip-free process, or birth and death process with catastrophes (cf. [1, 2, 3] for instance).

The single birth process, as a natural extension of birth and death process which is a simplest Q-process (Markov chain), has its own origins in practice, refer to the earlier papers [2, 13, 15], for instance. The exit boundary of the process consists at most one single extremal point and so the single birth process is nearly the largest class for which the explicit criteria on classical problems can be expected. Actually, the study on the object is quite fruited and relatively

Received March 1, 2014; accepted April 5, 2014
Corresponding author: Yu-Hui ZHANG, E-mail: zhangyh@bnu.edu.cn

completed (cf. [4, 5, 6, 15, 16, 17]). Based on this advantage, the single birth process becomes a fundamental comparison tool in studying more complex processes, such as infinite-dimensional reaction-diffusion processes. Refer to [4; Chapters 3 and 4, Part Ⅲ] and [15]. Usually, the single birth process is non-symmetric and hence it is regarded as a representative one of the non-symmetric processes. For non-symmetric processes, comparing with the symmetric ones, our knowledge is much limited, except for single birth processes to which much results are known as just mentioned. Up to now, the known results are all presented in some recursive forms. This paper introduces a single unified representation, as well as a unified treatment, of various formulas for single birth processes.

Throughout the paper, we consider only the single birth Q-matrix $Q = (q_{ij})$. Set $q_i = -q_{ii}$ for each $i \in \mathbb{Z}_+$. For a given function c (to be fixed in this and the next sections, and then to be specified case by case), define an operator Ω as follows

$$\Omega g = Qg + cg, \quad \text{where} \quad (Qg)_i = \sum_j q_{ij}(g_j - g_i).$$

Clearly, if $c \leqslant 0$, then Ω is an operator corresponding to a single birth process with killing rates $(-c_i)$.

The following sequences are used throughout this paper.

$$\widetilde{F}_i^{(i)} = 1, \qquad \widetilde{F}_n^{(i)} = \frac{1}{q_{n,n+1}} \sum_{k=i}^{n-1} \tilde{q}_n^{(k)} \widetilde{F}_k^{(i)}, \qquad n > i \geqslant 0, \tag{1.1}$$

$$\tilde{q}_n^{(k)} = q_n^{(k)} - c_n := \sum_{j=0}^{k} q_{nj} - c_n, \qquad 0 \leqslant k < n. \tag{1.2}$$

Note that if $c \leqslant 0$, then $\tilde{q}_n^{(k)} \geqslant 0$ and then $\widetilde{F}_n^{(k)} \geqslant 0$ for every $n > k \geqslant 0$. In what follows, we omit the superscript $\tilde{}$ everywhere in $\widetilde{F}$ and $\tilde{q}$ once $c_i \equiv 0$, and often use the convention that $\sum_{\varnothing} = 0$.

Here is the first of our main results.

Theorem 1.1 *Given a single-birth Q-matrix $Q = (q_{ij})$ and functions c and f, the solution g to the Poisson equation*

$$\Omega g = f \tag{1.3}$$

has the following representation:

$$g_n = g_0 + \sum_{0 \leqslant k \leqslant n-1} \sum_{0 \leqslant j \leqslant k} \frac{\widetilde{F}_k^{(j)}(f_j - c_j g_0)}{q_{j,j+1}}, \qquad n \geqslant 0. \tag{1.4}$$

In particular, the harmonic function g of Ω (i.e., $\Omega g = 0$) can be represented as

$$g_n = g_0 \left(1 - \sum_{0 \leqslant k \leqslant n-1} \sum_{0 \leqslant j \leqslant k} \frac{\widetilde{F}_k^{(j)} c_j}{q_{j,j+1}} \right), \qquad n \geqslant 0.$$

Conversely, for each boundary/initial value $g_0 \in \mathbb{R}$, the function (g_n) defined by (1.4) *is a solution to* (1.3).

For single birth processes, almost all problems we concerned with are related to the solutions to some specific Poisson equation. Here, we unify these equations as (1.3) with different functions c and f which are listed as follow.

Problem	$c_i \in \mathbb{R}$	$f_i \in \mathbb{R}$
Harmonic function	$c_i \in \mathbb{R}$	$f_i \equiv 0$
Uniqueness	$c_i \equiv -\lambda < 0$	$f_i \equiv 0$
Recurrence	$c_i \equiv 0$	$f_i = q_{i0}(1-\delta_{i0})$
Extinction/return probability	$c_i \equiv 0$	$f_i = q_{i0}(1-\delta_{i0})(g_0-1)$
Ergodicity	$c_i \equiv 0$	$f_i = q_{i0}(1-\delta_{i0})g_0 - 1$
Strong ergodicity	$c_i \equiv 0$	$f_i = q_{i0}(1-\delta_{i0})g_0 - 1$
Polynomial moment	$c_i \equiv 0$	$f_i^{(\ell)}$
Exponential moment/ergodicity	$c_i \equiv \lambda > 0$	$f_i = q_{i0}(1-\delta_{i0})(g_0-1)$
Laplace transform of return time	$c_i \equiv -\lambda < 0$	$f_i = q_{i0}(1-\delta_{i0})(g_0-1)$

where $f_i^{(\ell)} = q_{ii_0}(1-\delta_{ii_0})g_{i_0} - \ell\mathbb{E}_i\sigma_{i_0}^{\ell-1}$.

We remark that in the two cases for ergodicity and strong ergodicity, even though the Poisson equation and the functions c and f are the same, but their solutions are required to be finite and bounded, respectively.

The paper is organized as follows. The proof of Theorem 1.1 is given in the next section, using a lemma on the representation of solution to a class of linear equations. Then, Sections 3–7 are devoted, respectively, to the criteria on the problems listed in the table above, and related problems to be specific subsequently. Roughly speaking, the unified treatment presented in the paper consists of the following three steps.

(a) Find out the Poisson equation corresponding to the problem we are interested in.
(b) Apply Theorem 1.1 to get the solution to the Poisson equation.
(c) Work out a criterion for the problem using the solution obtained in (b).

Step (a) is more or less known from the previous study; step (b) is now automatic; hence, our main work is spent on step (c).

For the reader's convenience, several key formulas used often in the proofs are collected into an Appendix in a single page which consists the last page of the paper (so that it can be printed out separately).

2 The Poisson equation

In this section, we consider the solutions of the Poisson equation (1.3) for single birth processes. Let us begin with a simple result for the solution to a class of linear equations.

Lemma 2.1 *For given real numbers $(\alpha_{nk})_{n-1\geqslant k\geqslant 0}$ and $(f_n)_{n\geqslant 0}$, the solution $(g_n)_{n\geqslant 0}$ to the recursive inhomogeneous equations*

$$g_n = \sum_{0\leqslant k\leqslant n-1} \alpha_{nk} g_k + f_n, \qquad n \geqslant 0 \tag{2.1}$$

can be represented as

$$g_n = \sum_{0\leqslant k\leqslant n} \gamma_{nk} f_k, \qquad n \geqslant 0, \tag{2.2}$$

where for fixed $k \geqslant 0$, $(\gamma_{nk})_{n\geqslant k}$ with $\gamma_{kk} = 1$ is the solution to the recursive equations

$$\gamma_{nk} = \sum_{k\leqslant j\leqslant n-1} \alpha_{nj}\gamma_{jk}, \qquad n > k. \tag{2.3}$$

Proof Use induction. For $n = 0$, we have

$$g_0 = f_0 = \gamma_{00} f_0 = \sum_{0\leqslant k\leqslant 0} \gamma_{0k} f_k.$$

Assume that (2.2) holds for all $n \leqslant m$. When $n = m + 1$, from (2.1), we see that

$$\begin{aligned} g_{m+1} &= \sum_{0\leqslant k\leqslant m} \alpha_{m+1,k}\, g_k + f_{m+1} = \sum_{0\leqslant k\leqslant m} \alpha_{m+1,k} \sum_{0\leqslant \ell\leqslant k} \gamma_{k\ell}\, f_\ell + f_{m+1} \\ &= \sum_{0\leqslant \ell\leqslant m} \left(\sum_{\ell\leqslant k\leqslant m} \alpha_{m+1,k}\gamma_{k\ell} \right) f_\ell + f_{m+1} = \sum_{0\leqslant \ell\leqslant m} \gamma_{m+1,\ell}\, f_\ell + f_{m+1} \\ &= \sum_{0\leqslant \ell\leqslant m+1} \gamma_{m+1,\ell}\, f_\ell. \end{aligned}$$

Hence, (2.2) holds for $n = m + 1$. By induction, the representation (2.2) holds for all $n \geqslant 0$. □

Note that the coefficients (α_{nk}) are often fixed and so are (γ_{nk}). Then Lemma 2.1 says that once replacing (α_{nk}) by (γ_{nk}), the solution to (2.1) has a complete representation (2.2), mainly in terms of the inhomogeneous term (f_n) in (2.1).

Without condition $\gamma_{kk} = 1$, equation (2.3) is clearly homogeneous. However, it becomes inhomogeneous under condition $\gamma_{kk} \neq 0$ (then one may assume that $\gamma_{kk} = 1$):

$$\gamma_{nk} = \sum_{k+1 \leqslant j \leqslant n-1} \alpha_{nj}\gamma_{jk} + \alpha_{nk}\gamma_{kk}, \qquad n \geqslant k+1$$

provided $\alpha_{k+1,k} \neq 0$. Otherwise, once $\alpha_{k+1,k} = 0$, by induction, we actually have $\gamma_{nk} = 0$ for all $n \geqslant k+1$. Thus, under condition $\gamma_{kk} = 1$, by Lemma 2.1 (for fixed k), we have the following alternative representation of (γ_{nk}):

$$\gamma_{nk} = \sum_{k+1 \leqslant j \leqslant n} \gamma_{nj}\alpha_{jk}, \qquad n \geqslant k+1.$$

In what follows, we will use the following variant of Lemma 2.1. Replacing the initial 0 by i and the coefficient (α_{nk}) by $(\alpha_{nk}\beta_k)$, respectively, for some non-zero sequence (β_n), and set $h_n = g_n/\beta_n\,(n \geqslant i)$, we obtain the following result.

Corollary 2.2 *The solution $(h_n)_{n\geqslant i}$ to the recursive equations*

$$h_n = \frac{1}{\beta_n}\left(\sum_{i \leqslant k \leqslant n-1} \alpha_{nk}h_k + f_n\right), \qquad n \geqslant i \tag{2.4}$$

can be represented as

$$h_n = \sum_{i \leqslant k \leqslant n} \frac{\gamma_{nk}}{\beta_k} f_k, \qquad n \geqslant i, \tag{2.5}$$

where for each fixed i, $(\gamma_{ni})_{n\geqslant i}$ with $\gamma_{ii} = 1$ is the solution to the equations

$$\gamma_{ni} = \frac{1}{\beta_n} \sum_{i \leqslant k \leqslant n-1} \alpha_{nk}\gamma_{ki}, \qquad n > i.$$

Equivalently,

$$\gamma_{ii} = 1, \qquad \gamma_{ni} = \sum_{i+1 \leqslant k \leqslant n} \frac{\gamma_{nk}}{\beta_k}\alpha_{ki}, \qquad n \geqslant i+1. \tag{2.6}$$

Specifying $\beta_n = q_{n,n+1}$ and $\alpha_{nk} = \tilde{q}_n^{(k)}$ in Corollary 2.2 and using the successive formula of $\widetilde{F}_n^{(k)}$ defined in (1.1), we obtain the following result.

Corollary 2.3 *For given f, the sequence (h_n) defined successively by*

$$h_n = \frac{1}{q_{n,n+1}}\left(f_n + \sum_{i \leqslant k \leqslant n-1} \tilde{q}_n^{(k)}h_k\right), \qquad n \geqslant i$$

has an unified expression as follows

$$h_n = \sum_{k=i}^{n} \frac{\widetilde{F}_n^{(k)}}{q_{k,k+1}} f_k, \qquad n \geqslant i.$$

In particular, the sequence $\left(\widetilde{F}_n^{(k)}\right)$ *defined in* (1.1) *has the following expression*

$$\widetilde{F}_i^{(i)} = 1, \qquad \widetilde{F}_n^{(i)} = \sum_{k=i+1}^{n} \frac{\widetilde{F}_n^{(k)} \tilde{q}_k^{(i)}}{q_{k,k+1}}, \quad n \geqslant i+1. \tag{2.7}$$

Before moving further, let us mention a comparison result for different γ_{nj}, which may be useful elsewhere but not in this paper.

Proposition 2.4 *For each triple* $n \geqslant i > j$*, the following assertion holds:*

$$\gamma_{nj} = \sum_{i \leqslant k \leqslant n} \frac{\gamma_{nk}}{\beta_k} \sum_{j \leqslant \ell \leqslant i-1} \alpha_{k\ell} \gamma_{\ell j}. \tag{2.8}$$

Furthermore, if $\alpha_{nk} \geqslant 0$ *and* $\beta_n > 0$ *for all* $n > k$*, then* $\gamma_{ni}\gamma_{ij} \leqslant \gamma_{nj}$ *for all* $n \geqslant i \geqslant j$*.*

Proof The first assertion is simply a consequence of Corollary 2.2. In fact, for fixed $i > j$, take

$$f_n = \sum_{j \leqslant \ell \leqslant i-1} \alpha_{n\ell}\, \gamma_{\ell j}, \qquad n \geqslant i.$$

Then

$$\gamma_{nj} = \frac{1}{\beta_n} \left[\sum_{i \leqslant \ell \leqslant n-1} \alpha_{n,\ell}\, \gamma_{\ell j} + \sum_{j \leqslant \ell \leqslant i-1} \alpha_{n\ell}\, \gamma_{\ell j} \right] = \frac{1}{\beta_n} \left[\sum_{i \leqslant \ell \leqslant n-1} \alpha_{n\ell}\, \gamma_{\ell j} + f_n \right], \quad n \geqslant i.$$

Hence, by Corollary 2.2, we get

$$\gamma_{nj} = \sum_{i \leqslant k \leqslant n} \frac{\gamma_{nk}}{\beta_k} f_k = \sum_{i \leqslant k \leqslant n} \frac{\gamma_{nk}}{\beta_k} \sum_{j \leqslant \ell \leqslant i-1} \alpha_{k\ell}\, \gamma_{\ell j}, \qquad n \geqslant i.$$

If $\alpha_{nk} \geqslant 0$ and $\beta_n > 0$ for all n and k, then from (2.8), it follows that for all $n > i > j$,

$$\gamma_{nj} = \gamma_{ni}\gamma_{ij} + \sum_{i+1 \leqslant k \leqslant n} \frac{\gamma_{nk}}{\beta_k} \sum_{j \leqslant \ell \leqslant i-1} \alpha_{k\ell} \gamma_{\ell j} \geqslant \gamma_{ni}\gamma_{ij}.$$

In the cases of $n = i$ or $i = j$, the conclusion is trivial. □

Now we turn to prove our first result.

Proof of Theorem 1.1 For each $i \geqslant 0$, we have

$$\begin{aligned}(\Omega g)_i &= q_{i,i+1}(g_{i+1}-g_i) - \sum_{0\leqslant j\leqslant i-1} q_{ij} \sum_{k=j}^{i-1}(g_{k+1}-g_k) + c_i g_i \\ &= q_{i,i+1}(g_{i+1}-g_i) - \sum_{0\leqslant k\leqslant i-1} \sum_{j=0}^{k} q_{ij}(g_{k+1}-g_k) + c_i g_i \\ &= q_{i,i+1}(g_{i+1}-g_i) - \sum_{0\leqslant k\leqslant i-1} \left(\sum_{j=0}^{k} q_{ij} - c_i\right)(g_{k+1}-g_k) + c_i g_0 \\ &= q_{i,i+1}(g_{i+1}-g_i) - \sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)}(g_{k+1}-g_k) + c_i g_0. \end{aligned} \tag{2.9}$$

Denote $g_{k+1}-g_k$ by w_k for $k \geqslant 0$. Then

$$(\Omega g)_i = q_{i,i+1} w_i - \sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)} w_k + c_i g_0, \qquad i \geqslant 0.$$

Now we rewrite the Poisson equation (1.3) as

$$w_i = \frac{1}{q_{i,i+1}} \left(\sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)} w_k + \tilde{f}_i \right), \qquad i \geqslant 0,$$

where $\tilde{f}_i = f_i - c_i g_0$ for $i \geqslant 0$. By Corollary 2.3, we obtain

$$w_i = \sum_{j=0}^{i} \frac{\widetilde{F}_i^{(j)} \tilde{f}_j}{q_{j,j+1}}, \qquad i \geqslant 0.$$

So the solution of the Poisson equation (1.3) satisfies

$$g_i = g_0 + \sum_{k=0}^{i-1} w_k = g_0 + \sum_{k=0}^{i-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)} \tilde{f}_j}{q_{j,j+1}}, \qquad i \geqslant 1.$$

The first assertion is proven. The second assertion is simply a consequence of the first one.

To prove the last assertion of the theorem, noting that by (1.4), we have

$$g_{n+1} - g_n = \sum_{j=0}^{n} \frac{\widetilde{F}_n^{(j)}(f_j - c_j g_0)}{q_{j,j+1}}, \qquad n \geqslant 0.$$

Thus, from (2.9), it follows for each $i \geqslant 0$ that

$$(\Omega g)_i = q_{i,i+1} \sum_{j=0}^{i} \frac{\widetilde{F}_i^{(j)}(f_j - c_j g_0)}{q_{j,j+1}} - \sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)}(f_j - c_j g_0)}{q_{j,j+1}} + c_i g_0.$$

Because (by exchanging the order of sums and using (1.1))

$$\sum_{0\leqslant k\leqslant i-1}\tilde{q}_i^{(k)}\sum_{j=0}^{k}\frac{\widetilde{F}_k^{(j)}(f_j-c_jg_0)}{q_{j,j+1}}=\sum_{0\leqslant j\leqslant i-1}\frac{f_j-c_jg_0}{q_{j,j+1}}\sum_{k=j}^{i-1}\tilde{q}_i^{(k)}\widetilde{F}_k^{(j)}$$
$$=q_{i,i+1}\sum_{0\leqslant j\leqslant i-1}\frac{\widetilde{F}_i^{(j)}(f_j-c_jg_0)}{q_{j,j+1}},$$

we obtain $\Omega g=f$ as required. □

Remark 2.5 (1) One may obtain $\big(\tilde{q}_n^{(k)},\widetilde{F}_n^{(k)}\big)$ from $\big(q_n^{(k)},F_n^{(k)}\big)$ easily replacing the original $Q=(q_{ij})$ by $\widetilde{Q}=(\tilde{q}_{ij})$:

$$\begin{cases}\tilde{q}_{i0}=q_{i0}-c_i,\\ \tilde{q}_{ij}=q_{ij},\quad j\neq 0,\ i\in E.\end{cases}$$

In other words, only the first column of $Q=(q_{ij})$ is modified. Then the original Poisson equation $\Omega g=f$ can be rewritten as $\widetilde{Q}g=\tilde{f}$ with $\tilde{f}_i=f_i-c_ig_0$.

(2) Alternatively, one may enlarge the space E by adding a point, say -1 for instance. Then introduce suitable $\bar{q}_{-1,i}$, $\bar{q}_{i,-1}$, $\bar{g}_{-1}$, and $\bar{f}_{-1}$, so that $\bar{Q}|_E=Q$, $\bar{g}|_E=g$, and $\bar{f}|_E=f$. In this way, one may rewrite $\Omega g=f$ on E as $\bar{Q}\bar{g}=\bar{f}$ on $E\cup\{-1\}$.

(3) To solve the Poisson equation, in view of (2.9), even for the simplest birth–death type, once c appears, it is necessary to go out to the larger class of single birth one, one can not just stay within the class of birth–death processes. Actually, this observation is crucial to solve the Open Problem 9.13 in [7]. Refer to [8; Theorem 2.6].

For the remainder of this section, we consider only the processes on a finite state space $\{0,1,\ldots,N\}$. Note that here the rate $q_{N,N+1}$ is not defined (or setting to be zero), but we allow $c_N\neq 0$. Hence $\widetilde{F}_n^{(k)}$ is defined up to $n=N-1$ only. The next result is a localized version of Theorem 1.1

Proposition 2.6 *Given a single-birth Q-matrix (q_{ij}) and a function c on the finite state space $\{0,1,\ldots,N\}$ $(N\geqslant 1)$, the following assertions hold.*

(i) The solution of the Poisson equation $\Omega g=f$ has the following form:

$$g_n=g_0+\sum_{0\leqslant k\leqslant n-1}\sum_{0\leqslant j\leqslant k}\frac{\widetilde{F}_k^{(j)}(f_j-c_jg_0)}{q_{j,j+1}},\qquad 0\leqslant n\leqslant N,\tag{2.10}$$

with boundary condition

$$c_Ng_0=\sum_{k=0}^{N-1}\tilde{q}_N^{(k)}\sum_{j=0}^{k}\frac{\widetilde{F}_k^{(j)}(f_j-c_jg_0)}{q_{j,j+1}}+f_N\ \text{ or }\ g_0=\frac{f_N+\sum_{k=0}^{N-1}\tilde{q}_N^{(k)}\sum_{j=0}^{k}\widetilde{F}_k^{(j)}f_j/q_{j,j+1}}{c_N+\sum_{k=0}^{N-1}\tilde{q}_N^{(k)}\sum_{j=0}^{k}\widetilde{F}_k^{(j)}c_j/q_{j,j+1}}.$$

(*ii*) *Let* $c \leqslant 0$. *Then the harmonic equation* $\Omega g = 0$ *has only the trivial solution* $g_i \equiv 0$ *iff there exists some* $c_i < 0$.

(*iii*) *The unique solution* g *to the equation* $\Omega g|_{\{0,1,\dots,N-1\}} = 0$ (*locally harmonic*) *with* $g_0 = 1$ *is as follows:*

$$g_n = 1 - \sum_{0\leqslant k\leqslant n-1} \sum_{0\leqslant j\leqslant k} \frac{\widetilde{F}_k^{(j)} c_j}{q_{j,j+1}}, \qquad 0 \leqslant n \leqslant N \tag{2.11}$$

which is increasing once $c \leqslant 0$.

Proof (a) The proof is nearly the same as the one of Theorem 1.1, except we have to take care for the boundary at N. By (2.9), for $0 \leqslant i \leqslant N-1$, we have

$$(\Omega g)_i = q_{i,i+1}(g_{i+1} - g_i) - \sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)}(g_{k+1} - g_k) + c_i g_0.$$

Denote $g_{k+1} - g_k$ by w_k for all $0 \leqslant k < N$. Then

$$\begin{aligned}
(\Omega g)_i &= q_{i,i+1} w_i - \sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)} w_k + c_i g_0, \qquad 0 \leqslant i < N;\\
(\Omega g)_N &= -\sum_{k=0}^{N-1} \tilde{q}_N^{(k)} w_k + c_N g_0.
\end{aligned}$$

Rewrite the Poisson equation as

$$w_i = \frac{1}{q_{i,i+1}} \left(\tilde{f}_i + \sum_{0\leqslant k\leqslant i-1} \tilde{q}_i^{(k)} w_k \right), \qquad 0 \leqslant i < N, \tag{2.12}$$

where $\tilde{f}_i = f_i - c_i g_0$ for all $0 \leqslant i \leqslant N$. By Corollary 2.3, we get

$$w_i = \sum_{j=0}^{i} \frac{\widetilde{F}_i^{(j)} \tilde{f}_j}{q_{j,j+1}}, \qquad 0 \leqslant i < N. \tag{2.13}$$

So the solution of the Poisson equation satisfies

$$g_i = g_0 + \sum_{k=0}^{i-1} w_k = g_0 + \sum_{k=0}^{i-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)} \tilde{f}_j}{q_{j,j+1}}, \qquad 1 \leqslant i \leqslant N.$$

Combining this with the boundary condition $(\Omega g)_N = f_N$ and (2.13), we obtain the first assertion.

(b) We have just seen that the harmonic solution g satisfies

$$g_n = g_0 \left(1 - \sum_{k=0}^{n-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)} c_j}{q_{j,j+1}} \right), \qquad 1 \leqslant n \leqslant N. \tag{2.14}$$

and

$$g_0\left(c_N+\sum_{k=0}^{N-1}\tilde{q}_N^{(k)}\sum_{j=0}^{k}\frac{\widetilde{F}_k^{(j)}c_j}{q_{j,j+1}}\right)=0.$$

When $c\leqslant 0$, by irreducibility, we have not only $\tilde{q}_N^{(N-1)}>0$ but also $\widetilde{F}_{N-1}^{(j)}>0$ for every $j:0\leqslant j\leqslant N-1$. Hence, if there exists some $c_i<0$, then we must have $g_0=0$ by the last equation. Furthermore, by (2.14), we indeed have $g\equiv 0$.

Conversely, if $c_i\equiv 0$, then every constant function $g\neq 0$ is a solution to the equation $\Omega g=0$. Hence the harmonic function g can be non-trivial.

(c) To prove the third assertion, based on the second one, we have to use a smaller space $\{0,1,\ldots,N-1\}$ instead of the original $\{0,1,\ldots,N\}$ to avoid the trivial solution. The assertion now follows from (2.14). □

The next result is exceptional of the paper. Instead of single birth, we consider single death processes on a finite state space. The result may be regarded as a dual of Proposition 2.6. It indicates that a large parts of the study in the paper is meaningful for the single death processes, but we will not go to the details here.

A matrix $Q=(q_{ij})$ is called of single death if $q_{i,i-j}>0$ iff $j=1$ for $i\geqslant 1$.

Proposition 2.7 *Given a single death Q-matrix $Q=(q_{ij})$ and a function (c_i) on the finite state space $\{0,1,\ldots,N\}$, define $\tilde{q}_n^{(k)}=\sum_{j=k}^{N}q_{nj}-c_n$ for $k>n$ and*

$$\widetilde{F}_i^{(i)}=1,\qquad \widetilde{F}_n^{(i)}=\frac{1}{q_{n,n-1}}\sum_{k=n+1}^{i}\tilde{q}_n^{(k)}\widetilde{F}_k^{(i)},\qquad 1\leqslant n<i.$$

Then

(i) the solution g to the Poisson equation $\Omega g=f$ has the following representation:

$$g_n=g_N+\sum_{n+1\leqslant k\leqslant N}\sum_{k\leqslant j\leqslant N}\frac{\widetilde{F}_k^{(j)}(f_j-c_jg_N)}{q_{j,j-1}},\qquad 0\leqslant n\leqslant N$$

with boundary condition

$$c_0g_N=\sum_{k=1}^{N}\tilde{q}_0^{(k)}\sum_{j=k}^{N}\frac{\widetilde{F}_k^{(j)}(f_j-c_jg_N)}{q_{j,j-1}}+f_0.$$

(ii) The unique solution with $g_N=1$ to equation $Qg|_{\{1,2,\ldots,N\}}=0$ is as follows:

$$g_n=1-\sum_{n+1\leqslant k\leqslant N}\sum_{k\leqslant j\leqslant N}\frac{\widetilde{F}_k^{(j)}c_j}{q_{j,j-1}}\qquad (0\leqslant n\leqslant N)$$

which is decreasing in n once $c\leqslant 0$.

Proof For $1 \leqslant i \leqslant N$, we have

$$\begin{aligned}(\Omega g)_i &= q_{i,i-1}(g_{i-1}-g_i)+\sum_{i+1\leqslant j\leqslant N} q_{ij}\sum_{k=i+1}^{j}(g_k-g_{k-1})+c_ig_i\\ &= q_{i,i-1}(g_{i-1}-g_i)+\sum_{i+1\leqslant k\leqslant N}\sum_{j=k}^{N} q_{ij}(g_k-g_{k-1})+c_ig_i\\ &= q_{i,i-1}(g_{i-1}-g_i)-\sum_{i+1\leqslant k\leqslant N}\tilde{q}_i^{(k)}(g_{k-1}-g_k)+c_ig_N.\end{aligned}$$

Denote $g_{k-1}-g_k$ by w_k for all $1 \leqslant k \leqslant N$. Then

$$(\Omega g)_i = q_{i,i-1}w_i-\sum_{i+1\leqslant j\leqslant N}\tilde{q}_i^{(k)}w_k+c_ig_N,\qquad 1\leqslant i\leqslant N;$$

$$(\Omega g)_0 = -\sum_{k=1}^{N}\tilde{q}_0^{(k)}w_k+c_0g_N.$$

Now we rewrite the Poisson equation as

$$w_i=\frac{1}{q_{i,i-1}}\left(\tilde{f}_i+\sum_{i+1\leqslant j\leqslant N}q_i^{(k)}w_k\right),\qquad 1\leqslant i\leqslant N,$$

where $\tilde{f}_i = f_i - c_ig_N$ for all $0 \leqslant i \leqslant N$. As an analogue of Corollary 2.3, by induction, we can verify that

$$w_i=\sum_{j=i}^{N}\frac{\widetilde{F}_i^{(j)}\tilde{f}_j}{q_{j,j-1}},\qquad 1\leqslant i\leqslant N.$$

From the argument above, it follows immediately that

$$g_i=g_N+\sum_{k=i+1}^{N}w_k=g_N+\sum_{i+1\leqslant j\leqslant N}\sum_{k\leqslant j\leqslant N}\frac{\widetilde{F}_k^{(j)}\tilde{f}_j}{q_{j,j-1}},\qquad 0\leqslant i\leqslant N-1.$$

Combining this with the boundary condition $(\Omega g)_0 = f_0$, we finish the proof of the first assertion. The second assertion is derived from the first one immediately. □

3 Uniqueness

Starting from this section, we handle with the problems for single birth processes, listed at the beginning of the paper. First, we study the uniqueness problem. To do so, we need a sequence $(\widetilde{m}_n)$(to be used often subsequently) :

$$\widetilde{m}_0=\frac{1}{q_{01}},\quad \widetilde{m}_n=\frac{1}{q_{n,n+1}}\left(1+\sum_{k=0}^{n-1}\tilde{q}_n^{(k)}\widetilde{m}_k\right),\qquad n\geqslant 1. \tag{3.1}$$

By Corollary 2.3, we have

$$\widetilde{m}_n = \sum_{k=0}^{n} \frac{\widetilde{F}_n^{(k)}}{q_{k,k+1}}, \qquad n \geqslant 0. \tag{3.2}$$

Again, we omit the superscript $\tilde{}$ everywhere in $\widetilde{m}$, $\widetilde{F}$, and $\tilde{q}$ once $c_i \equiv 0$. The following criterion is taken from [4, 15, 16].

Proposition 3.1 *Corresponding to a given single birth Q-matrix $Q = (q_{ij})$ (conservative), the process is unique (non-explosive) iff $\sum_{n=0}^{\infty} m_n = \infty$.*

Proof By [4; Theorems 2.47 and 2.40], the single birth process is unique iff the solution (u_i) to the equation

$$(\lambda + q_i)u_i = \sum_{j \neq i} q_{ij} u_j, \qquad i \geqslant 0; \qquad u_0 = 1 \tag{3.3}$$

is unbounded for some (equivalently for all) $\lambda > 0$. Rewrite (3.3) as

$$\Omega u = Qu - \lambda u = 0; \qquad u_0 = 1.$$

Applying Theorem 1.1 to $c_i \equiv -\lambda$ and $f_i \equiv 0$, we obtain the unique solution:

$$u_n = 1 + \lambda \sum_{0 \leqslant k \leqslant n-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)}}{q_{j,j+1}} = 1 + \lambda \sum_{0 \leqslant k \leqslant n-1} \widetilde{m}_k, \qquad n \geqslant 0.$$

Clearly, u_n is increasing in n and then is unbounded iff $\sum_n \widetilde{m}_n = \infty$. Thus, it remains to show that $\sum_n \widetilde{m}_n = \infty$ iff $\sum_n m_n = \infty$. Combining $\widetilde{m}_n$ with m_n, it is clear that

$$\widetilde{m}_n = \sum_{j=0}^{n} \frac{\widetilde{F}_n^{(j)}}{q_{j,j+1}} \;\downarrow\; \sum_{k=0}^{n} \frac{F_n^{(k)}}{q_{k,k+1}} = m_n \qquad \text{as } \lambda \downarrow 0,$$

since

$$\tilde{q}_n^{(k)} = q_n^{(k)} + \lambda \;\downarrow\; q_n^{(k)} \qquad \text{as } \lambda \downarrow 0.$$

This already shows that the condition $\sum_n m_n = \infty$ is sufficient. It is nearly necessary since the conclusion does not depend on $\lambda > 0$, except there is a jump from $\lambda > 0$ to $\lambda = 0$. Hopefully, we have thus seen some advantage of Theorem 1.1, even though there is still a distance to prove the necessity.

Actually, there are several ways to prove the equivalence

$$\sum_n \widetilde{m}_n = \infty \text{ for a fixed } \lambda > 0 \Longleftrightarrow \sum_n m_n = \infty.$$

From now on, for simplicity, assume that $\lambda = 1$.

(a) Observing that corresponding to the sequence $(\widetilde{m}_n)$, the operator is $\Omega = Q - I$ which may be regarded as a bounded perturbation of the original operator Q. Since these two operators are zero-exit or not simultaneously, the equivalence above holds.

(b) In the original proof (cf. [4; Proof of Theorem 3.16]), it was proved that u_n is unbounded iff $\sum_n m_n = \infty$. Combining this with what proved above, we obtain the required equivalence.

(c) Here is a more direct proof. The idea comes from [20].

Assume that $\sum_{k=0}^{\infty} \widetilde{m}_k = \infty$. If $\sum_{k=0}^{\infty} m_k < \infty$, then there exists N_0 large enough such that for all $n \geqslant N_0$,

$$\widetilde{M}_n := \sum_{k=0}^{n} \widetilde{m}_k > 1 \quad \text{and} \quad K := 2 \sum_{k=N_0+1}^{\infty} m_k < 1.$$

We now prove that for each $n > N_0$,

$$\widetilde{m}_k \leqslant 2m_k \widetilde{M}_{n-1}, \qquad 0 \leqslant k \leqslant n. \tag{3.4}$$

Since $\widetilde{m}_0 = m_0$ and $\widetilde{M}_{n-1} > 1$ (due to the fact that $n - 1 \geqslant N_0$), (3.4) holds in the case of $k = 0$. Assume that (3.4) holds up to $k = \ell - 1 < n$. Then,

$$\begin{aligned}
\widetilde{m}_\ell &= \frac{1}{q_{\ell,\ell+1}} \left(1 + \sum_{k=0}^{\ell-1} q_\ell^{(k)} \widetilde{m}_k + \sum_{k=0}^{\ell-1} \widetilde{m}_k \right) \quad (\text{since } \lambda = 1) \\
&\leqslant \frac{1}{q_{\ell,\ell+1}} \left(1 + \sum_{k=0}^{\ell-1} q_\ell^{(k)} 2m_k \widetilde{M}_{n-1} + \widetilde{M}_{\ell-1} \right) \quad (\text{by assumption}) \\
&\leqslant \frac{1}{q_{\ell,\ell+1}} \left(1 + \sum_{k=0}^{\ell-1} q_\ell^{(k)} m_k \right) 2\widetilde{M}_{n-1} \\
&= 2m_\ell \widetilde{M}_{n-1}.
\end{aligned}$$

So (3.4) holds when $k = \ell$. By induction, we know that (3.4) holds for every $k : 0 \leqslant k \leqslant n$. Now, for each $n > N_0$, we have

$$\widetilde{M}_n = \widetilde{M}_{N_0} + \sum_{k=N_0+1}^{n} \widetilde{m}_k \leqslant \widetilde{M}_{N_0} + \sum_{k=N_0+1}^{n} 2m_k \widetilde{M}_{n-1} \leqslant \widetilde{M}_{N_0} + K\widetilde{M}_{n-1}.$$

Furthermore, we have

$$\begin{aligned}
\widetilde{M}_n &\leqslant \widetilde{M}_{N_0} \big(1 + K + \cdots + K^{n-N_0-1} \big) + K^{n-N_0} \widetilde{M}_{N_0} \\
&= \frac{\widetilde{M}_{N_0}(1 - K^{n-N_0})}{1 - K} + K^{n-N_0} \widetilde{M}_{N_0}.
\end{aligned}$$

Thus, as $n \to \infty$, we would have $\infty \leqslant \widetilde{M}_{N_0}/(1 - K)$ which is a contradiction. Hence, once $\sum_{k=0}^{\infty} \widetilde{m}_k = \infty$, we should also have $\sum_{k=0}^{\infty} m_k = \infty$.

We have therefore completed the proof of the equivalence mentioned above. □

To conclude this section, we mention that the uniqueness problem for the single birth Q-matrix with absorbing set $H = \{0, 1, \ldots, N\}\,(N < \infty)$ can be dealt with by the same approach. Refer to [4; Theorem 3.16] and [14].

4 Recurrence and extinction/return probability

For the recurrence, the following criterion is taken from [4; Theorem 4.52 (1)] and [15].

Proposition 4.1 *Assume the single birth Q-matrix $Q = (q_{ij})$ is non-explosive and irreducible. Then the process is recurrent iff $\sum_{n=0}^{\infty} F_n^{(0)} = \infty$, where $\left(F_n^{(i)}\right)$ was defined in* (1.1) *by setting $c_i \equiv 0$.*

Proof By [4; Lemma 4.51], we know that the single birth process is recurrent iff the equation

$$x_i = \sum_{k\neq 0} \Pi_{ik} x_k, \qquad 0 \leqslant x_i \leqslant 1, \qquad i \geqslant 0 \tag{4.1}$$

has only zero solution, where $\Pi_{ik} = (1-\delta_{ik})q_{ik}/q_i$. It is easily seen that equation (4.1) has a non-trivial solution iff the equation

$$x_i = \sum_{k\neq 0} \Pi_{ik} x_k, \qquad i \geqslant 0; \qquad x_0 = 1$$

has a nonnegative bounded solution. The following fact will be used several times below:

$$x_i = \sum_{k\neq i,i_0} \frac{q_{ik}}{q_i - \lambda} x_k + \frac{\gamma_i}{q_i - \lambda} \Longleftrightarrow (Qx)_i + \lambda x_i = q_{ii_0}(1 - \delta_{ii_0})x_{i_0} - \gamma_i, \tag{4.2}$$

where $\lambda \in \mathbb{R}$ satisfying some suitable condition. Certainly, here we preassume that $x_i \in \mathbb{R}$ for every $i \in E$. By using this fact with $\lambda = 0$ and $i_0 = 0$, we can rewrite the previous equation as

$$(Qx)_0 = 0, \qquad (Qx)_i = q_{i0}, \qquad i \geqslant 1; \qquad x_0 = 1.$$

Applying Theorem 1.1 to $c_i \equiv 0$ and $f_i = q_{i0}(1 - \delta_{i0})$, we obtain the unique solution as follows

$$x_0 = 1, \quad x_n = 1 + \sum_{k=1}^{n-1} \sum_{j=1}^{k} \frac{F_k^{(j)} q_{j0}}{q_{j,j+1}} = 1 + \sum_{k=1}^{n-1} \sum_{j=1}^{k} \frac{F_k^{(j)} q_j^{(0)}}{q_{j,j+1}}, \qquad n \geqslant 1.$$

By (2.7), it follows that

$$x_n = 1 + \sum_{k=1}^{n-1} F_k^{(0)} = \sum_{k=0}^{n-1} F_k^{(0)}, \qquad n \geqslant 1.$$

Clearly, (x_n) is bounded iff $\sum_{k=0}^{\infty} F_k^{(0)} < \infty$. In other words, equation (4.1) has only a trivial solution iff $\sum_{k=0}^{\infty} F_k^{(0)} = \infty$. The assertion is now proven. □

Extinction/return probability

For the remainder of this section, we study the extinction probability. Here the extinction time τ_0 is the first hitting time of the state 0. Thus, this topic is actually a refinement of what studied in the last proposition, in which we pay attention only on the result either $\mathbb{P}_n[\tau_0 < \infty] = 1$ or < 1 rather than its distribution. We will come back this point after the proof of the next proposition. For the extinction problem, the rates $q_{0j}\,(j \neq 0)$ play no rule, so one may assume the state 0 to be an absorbing state. In other words, we may reduce the state space from E to $E_1 := \{1, 2, \ldots\}$, and regard the rate $q_{i0}\,(i \neq 0)$ as a killing from i. Then we need to redefine the sequences $\big(\tilde{q}_n^{(k)}\big)$ and $\big(\widetilde{F}_n^{(k)}\big)$ starting from 1 but not 0. However, for our convenience, we prefer to keep the notation E, $\big(\tilde{q}_n^{(k)}\big)$, $\big(\widetilde{F}_n^{(k)}\big)$ and so on. For this, it is better to use the return time σ_0 instead of the hitting time τ_0. In the case that the state 0 is really an absorbing one, we can add a positive rate q_{01} and assume that the enlarged process becomes irreducible. Then, the solution of $\mathbb{P}_n[\sigma_0 < \infty]$ restricted on E_1 gives us the answer of $\mathbb{P}_n[\tau_0 < \infty]$ on E_1 (as a trivial application of the localization theorem [9; Theorem 3.4.1] or [4; Theorem 2.13]), so we can return to our original problem.

We remark that in the context of denumerable Markov processes, the topic of this section and much more problems were well studied in [9; Chapter IX]. In the present special case, for the single birth processes, the problem was studied in [1; Chapter 9] or [2], using a different technique.

Proposition 4.2 *Let the single birth Q-matrix $Q = (q_{ij})$ be non-explosive and irreducible. Then the return/extinction probability is as follows:*

$$\mathbb{P}_0(\sigma_0 < \infty) = \frac{\sum_{k=1}^{\infty} F_k^{(0)}}{\sum_{k=0}^{\infty} F_k^{(0)}}, \qquad \mathbb{P}_n(\sigma_0 < \infty) = \frac{\sum_{k=n}^{\infty} F_k^{(0)}}{\sum_{k=0}^{\infty} F_k^{(0)}}, \qquad n \geqslant 1.$$

Furthermore, $\mathbb{P}_n(\sigma_0 < \infty) = 1$ for all $n \geqslant 0$ iff $\mathbb{P}_0(\sigma_0 < \infty) = 1$, equivalently iff $\sum_{n=0}^{\infty} F_n^{(0)} = \infty$.

Proof By [4; Lemma 4.46] with $H = \{0\}$, $(\mathbb{P}_i(\sigma_0 < \infty) : i \in E)$ is the minimal nonnegative solution to the equation

$$x_i = \sum_{k \neq 0, i} \frac{q_{ik}}{q_i} x_k + \frac{q_{i0}}{q_i}(1 - \delta_{i0}), \qquad i \in E.$$

The study on recurrence usually starts from here, the lemma [4; Lemma 4.51] used in the last proof simplifies our study on the recurrence problem, as we have just seen above. By (4.2), the last equation is equivalent to

$$(Qx)_i = q_{i0}(1-\delta_{i0})(x_0-1), \qquad i \geqslant 0.$$

Applying Theorem 1.1 to $c_i \equiv 0$ and $f_i = q_{i0}(1-\delta_{i0})(x_0-1)$, we obtain the solution to the last equation:

$$\begin{aligned} x_n &= x_0 + \sum_{0\leqslant k\leqslant n-1}\sum_{0\leqslant j\leqslant k}\frac{F_k^{(j)}}{q_{j,j+1}}q_{j0}(1-\delta_{j0})(x_0-1) \\ &= x_0\left\{1+\sum_{1\leqslant k\leqslant n-1}\sum_{1\leqslant j\leqslant k}\frac{F_k^{(j)}}{q_{j,j+1}}q_j^{(0)}\right\} - \sum_{1\leqslant k\leqslant n-1}\sum_{1\leqslant j\leqslant k}\frac{F_k^{(j)}}{q_{j,j+1}}q_j^{(0)} \\ &= x_0\left(1+\sum_{1\leqslant k\leqslant n-1}F_k^{(0)}\right) - \sum_{1\leqslant k\leqslant n-1}F_k^{(0)}, \qquad n\geqslant 0 \quad \text{(by (2.7))}. \end{aligned}$$

Because $x_n > 0$, it follows that

$$x_0 \geqslant \sup_{n\geqslant 1}\frac{\sum_{k=1}^{n-1}F_k^{(0)}}{\sum_{k=0}^{n-1}F_k^{(0)}} = \sup_{n\geqslant 1}\frac{\sum_{k=0}^{n-1}F_k^{(0)}-1}{\sum_{k=0}^{n-1}F_k^{(0)}} = 1-\frac{1}{\sum_{k=0}^{\infty}F_k^{(0)}}.$$

From here, we obtain the minimal nonnegative solution:

$$x_0^* = 1-\frac{1}{\sum_{k=0}^{\infty}F_k^{(0)}}, \qquad x_n^* = 1-\frac{\sum_{k=0}^{n-1}F_k^{(0)}}{\sum_{k=0}^{\infty}F_k^{(0)}}, \qquad n\geqslant 1.$$

We have thus proved the first assertion. The second one is obvious. □

Rewrite the solution just obtained as follows.

$$1-x_0^* = \frac{1}{\sum_{k=0}^{\infty}F_k^{(0)}}, \qquad 1-x_n^* = \frac{\sum_{k=0}^{n-1}F_k^{(0)}}{\sum_{k=0}^{\infty}F_k^{(0)}}, \qquad n\geqslant 1.$$

Renormalize them so that the initial value becomes 1:

$$x_0 = 1, \qquad x_n = \sum_{k=0}^{n-1}F_k^{(0)}, \qquad n\geqslant 1$$

which is what we obtained in the last proof. We have thus seen the relation between the last two propositions.

The study on the Laplace transform of extinction/return time is delayed to Section 7 (Proposition 7.3 which is based on Lemma 7.1).

5 Ergodicity, strong ergodicity, and the first moment of return time

Let $E = \mathbb{Z}_+$ and $H \subset E$, $H \neq \varnothing, E$. Define $\sigma_H = \inf\{t \geqslant \eta_1 : X(t) \in H\}$, where η_1 is the first jump of the process. When H is a singleton, $H = \{0\}$, for instance, denote $\sigma_{\{0\}}$ by σ_0 for simplicity. We now consider the first moment of the return time σ_0. To do so, we introduce the following lemma (cf. [9; Lemma 9.4.1]).

Lemma 5.1 *Let (q_{ij}) be irreducible and assume that its Q-process is recurrent. Then $(x_i^* := \mathbb{E}_i\sigma_H : i \in E)$ is the minimal nonnegative solution (may be infinite) to the equation*

$$x_i = \frac{1}{q_i} \sum_{k \notin H \cup \{i\}} q_{ik}x_k + \frac{1}{q_i}, \qquad i \in E,$$

where $1 \cdot \infty = \infty$ and $0 \cdot \infty = 0$ by convention.

Proof Let $(y_i^* : i \in E)$ be the minimal nonnegative solution to the equation

$$y_i = \frac{1}{q_i} \sum_{k \notin H \cup \{i\}} q_{ik}y_k + \frac{1}{q_i}, \qquad i \in E.$$

By assumption and [4; Lemma 4.46], the quantity f_{iH} defined there is equal to 1 for every $i \in E$. Then, $(y_i^* : i \in E)$ coincides with $(e_{iH}(0) : i \in E)$ used in [4; Lemma 4.48]. Note that $e_{iH}(0) = \int_0^\infty \mathbb{P}_i(\sigma_H > t)\,\mathrm{d}t = \mathbb{E}_i\sigma_H$. The assertion now follows immediately. □

In what follows, we use often another sequence $(\tilde{d}_n)$ similar to $\big(\widetilde{m}_n\big)$ having different initial value:

$$\tilde{d}_0 = 0, \qquad \tilde{d}_n = \frac{1}{q_{n,n+1}} \left(1 + \sum_{k=0}^{n-1} \tilde{q}_n^{(k)} \tilde{d}_k\right), \qquad n \geqslant 1, \tag{5.1}$$

where $\tilde{q}_n^{(k)}$ is defined in (1.2). By Corollary 2.3, we have

$$\tilde{d}_n = \sum_{1 \leqslant j \leqslant n} \frac{\widetilde{F}_n^{(j)}}{q_{j,j+1}}, \qquad n \geqslant 0 \tag{5.2}$$

which is very much the same as (3.2). Again, we omit the superscript ˜ everywhere in $(\tilde{d}_n)$ once $c_i \equiv 0$. Note that if we rewrite

$$\tilde{d}_n = \frac{1}{q_{n,n+1}} \left(1 + \sum_{1 \leqslant k \leqslant n-1} \tilde{q}_n^{(k)} \tilde{d}_k\right), \qquad n \geqslant 1,$$

$$\widetilde{F}_n^{(0)} = \frac{1}{q_{n,n+1}} \left(\tilde{q}_n^{(0)} + \sum_{1 \leqslant k \leqslant n-1} \tilde{q}_n^{(k)} \widetilde{F}_k^{(0)}\right), \qquad n \geqslant 1,$$

then it is clear that the sequences $\big(\tilde{d}_n\big)_{n\geqslant 1}$ and $\big(\widetilde{F}_n^{(0)}\big)_{n\geqslant 1}$ are also quite close each other.

The main result in this section is as follows. Refer to [4; Theorem 4.52 (2)], [1; Proposition 2.4], and [15, 17, 18].

Proposition 5.2 *Assume that the single birth Q-matrix $Q = (q_{ij})$ is irreducible and corresponding process is recurrent. Then*

$$\mathbb{E}_0\sigma_0 = \frac{1}{q_{01}} + d, \qquad \mathbb{E}_n\sigma_0 = \sum_{k=0}^{n-1}\big(F_k^{(0)}d - d_k\big), \qquad n \geqslant 1,$$

where

$$d = \overline{\lim_{k\to\infty}} \frac{\sum_{n=0}^{k} d_n}{\sum_{n=0}^{k} F_n^{(0)}} = \lim_{n\to\infty} \frac{d_n}{F_n^{(0)}} \text{ if the limit exists.}$$

Furthermore, the process is ergodic (i.e. positive recurrent) iff $d < \infty$; and it is strongly ergodic iff $\sup_{k\in E}\sum_{n=0}^{k}\big(F_n^{(0)}d - d_n\big) < \infty$. Actually, for the last conclusion, the recurrence assumption can be replaced by the uniqueness one.

Proof Let $H = \{0\}$. By Lemma 5.1, $(\mathbb{E}_i\sigma_0 : i \in E)$ is the minimal nonnegative solution (x_i^*) to the equation

$$x_i = \frac{1}{q_i}\sum_{k\notin\{0,i\}} q_{ik}x_k + \frac{1}{q_i}, \qquad i \in E. \tag{5.3}$$

Suppose for a moment that $x_i^* < \infty$ first for some $i \in E$ and then for all i by irreducibility. Next, let (x_i) be a (finite) solution to (5.3). Then, by (4.2), we have

$$(Qx)_i = q_{i0}x_0 - 1, \qquad i \geqslant 1; \qquad (Qx)_0 = -1.$$

Applying Theorem 1.1 to $c = 0$ and $f_i = q_{i0}(1-\delta_{i0})x_0 - 1$ $(i \geqslant 0)$, we obtain the solution to the last equation:

$$x_n = x_0 + \sum_{k=0}^{n-1}\sum_{j=0}^{k}\frac{F_k^{(j)}f_j}{q_{j,j+1}} = x_0\left(1 + \sum_{k=1}^{n-1}\sum_{j=1}^{k}\frac{F_k^{(j)}q_{j0}}{q_{j,j+1}}\right) - \sum_{k=0}^{n-1}\sum_{j=0}^{k}\frac{F_k^{(j)}}{q_{j,j+1}}, \qquad n \geqslant 1.$$

By (2.7) and (5.2), we obtain

$$x_n = x_0\sum_{k=0}^{n-1}F_k^{(0)} - \sum_{k=0}^{n-1}\left(\frac{F_k^{(0)}}{q_{01}} + d_k\right) = \sum_{k=0}^{n-1}\left[F_k^{(0)}\left(x_0 - \frac{1}{q_{01}}\right) - d_k\right], \qquad n \geqslant 1.$$

Since $x_n > 0$, it follows that

$$x_0\sum_{k=0}^{n-1}F_k^{(0)} > \sum_{k=0}^{n-1}\left(\frac{F_k^{(0)}}{q_{01}} + d_k\right), \qquad n \geqslant 1.$$

This gives us

$$x_0 \geqslant \sup_{n\geqslant 1} \frac{\sum_{k=0}^{n-1}(F_k^{(0)}/q_{01} + d_k)}{\sum_{k=0}^{n-1} F_k^{(0)}} = \frac{1}{q_{01}} + \sup_{n\geqslant 1} \frac{\sum_{k=0}^{n-1} d_k}{\sum_{k=0}^{n-1} F_k^{(0)}}.$$

Now, the minimal property implies that

$$x_0^* = \frac{1}{q_{01}} + \sup_{n\geqslant 1} \frac{\sum_{k=0}^{n-1} d_k}{\sum_{k=0}^{n-1} F_k^{(0)}}$$

and then

$$x_n^* = \sum_{k=0}^{n-1} \left(F_k^{(0)} \sup_{n\geqslant 1} \frac{\sum_{j=0}^{n-1} d_j}{\sum_{j=0}^{n-1} F_j^{(0)}} - d_k \right), \qquad n \geqslant 1$$

gives us the solution $(\mathbb{E}_i\sigma_0 : i \in E)$. We claim that the supremum in the last line has to achieved at infinity. Otherwise, if it is achieved at some finite n_0:

$$\frac{\sum_{j=0}^{n_0-1} d_j}{\sum_{j=0}^{n_0-1} F_j^{(0)}} = \sup_{n\geqslant 1} \frac{\sum_{j=0}^{n-1} d_j}{\sum_{j=0}^{n-1} F_j^{(0)}}.$$

Then

$$x_0^* = \frac{1}{q_{01}} + \frac{\sum_{j=0}^{n_0-1} d_j}{\sum_{j=0}^{n_0-1} F_j^{(0)}}$$

and furthermore, $x_{n_0}^* = 0$ which is a contradiction with $x_i^* = \mathbb{E}_i\sigma_0 > 0$. Therefore,

$$\sup_{n\geqslant 1} \frac{\sum_{j=0}^{n-1} d_j}{\sum_{j=0}^{n-1} F_j^{(0)}} = \overline{\lim_{n\to\infty}} \frac{\sum_{j=0}^{n} d_j}{\sum_{j=0}^{n} F_j^{(0)}} =: d$$

as required. The next limit in the expression of d is an application of Stolz's Theorem. Now $d < \infty$ since $x_0^* < \infty$ by assumption. To remove the finiteness assumption of (x_i^*), we claim that the expressions in the first assertion for $\mathbb{E}_n\sigma_0 (= x_n^*)$ still hold even $x_i^* = \infty$, since then we must have $d = \infty$. If otherwise, $d < \infty$, then by the last assertion of Theorem 1.1 and (4.2), we would obtain a finite solution to (5.3), which deduces a contradiction to the assumption $x_i^* = \infty$ by the comparison theorem for the nonnegative solutions (cf. [4; Theorem 2.6]). We have thus proved the first assertion.

Let us remark that the trick used above replacing $\sup_{n\geqslant 1}$ by $\overline{\lim}_{n\to\infty}$ was missed in the previous publications. This trick and the one assuming the finiteness of (x_i^*), will be used several times below but we may not mention it time by time.

Finally, by [4; Theorem 4.44], the single process is ergodic iff $\mathbb{E}_0\sigma_0 < \infty$ which is now equivalent to $d < \infty$. By the same cited theorem, the process is strongly ergodic iff $\sup_{i\in E} \mathbb{E}_i\sigma_0 < \infty$, equivalently, $\sup_{n\in E} \sum_{k=0}^{n} \left(F_k^{(0)} d - d_k\right) < \infty$ which follows from the first assertion. As mentioned in the proof of the

cited book, for ergodicity, the uniqueness assumption is enough instead of the recurrence one. The proof is now finished. □

6 Polynomial moments of hitting time and life time

Polynomial moments of hitting time

We have just studied the first moment of the time of first hitting/return 0 in the last section. Now we study the higher-order moments of the first hitting time.

Fix $i_0 \geqslant 0$. Recall that σ_{i_0} is the time of first return to i_0 after the first jump. For its higher-moments, we have the following result (cf. [19, 21]).

Proposition 6.1 *Assume that the single birth Q-matrix $Q = (q_{ij})$ is irreducible and the corresponding process is $(\ell-1)$-ergodic ($\ell \geqslant 1$), i.e. $\mathbb{E}_i\sigma_{i_0}^{\ell-1} < \infty$ for every $i \geqslant 0$. When $\ell = 1$, assume additionally that the process is unique. Then we have*

$$\mathbb{E}_n\sigma_{i_0}^{\ell} = \begin{cases} \ell\sum_{n\leqslant k\leqslant i_0-1} v_k^{(\ell)} + \big[1 - \sum_{n\leqslant k\leqslant i_0-1} u_k\big]\mathbb{E}_{i_0}\sigma_{i_0}^{\ell}, & 0 \leqslant n \leqslant i_0; \\ -\ell\sum_{i_0\leqslant k\leqslant n-1} v_k^{(\ell)} + \big[1 + \sum_{i_0\leqslant k\leqslant n-1} u_k\big]\mathbb{E}_{i_0}\sigma_{i_0}^{\ell}, & n > i_0; \end{cases}$$

where

$$u_k = \begin{cases} \sum_{j=i_0-1}^{k} {q_{j,j+1}}^{-1} F_k^{(j)} q_{ji_0}(1-\delta_{ji_0}), & k \geqslant i_0, \\ 1, & k = i_0 - 1, \\ 0, & 0 \leqslant k \leqslant i_0 - 2 \end{cases}$$

$$v_k^{(\ell)} = \sum_{j=0}^{k} \frac{F_k^{(j)}}{q_{j,j+1}} \mathbb{E}_j\sigma_{i_0}^{\ell-1}, \qquad k \geqslant 0,$$

$$\begin{aligned} \mathbb{E}_{i_0}\sigma_{i_0}^{\ell} &= \ell \varlimsup_{n\to\infty} \left(\sum_{i_0\leqslant k\leqslant n} v_k^{(\ell)}\right)\left[1 + \sum_{i_0\leqslant k\leqslant n} u_k\right]^{-1} \\ &= \ell \lim_{n\to\infty} \frac{v_n^{(\ell)}}{u_n} \text{ if the limit exists.} \end{aligned}$$

Proof By [9; Theorem 9.3.3] (cf. [4; Proposition 4.56], or [10; Theorem 3.1]), $(y_i^* := \mathbb{E}_i\sigma_{i_0}^{\ell} : i \in E)$ is the the minimal nonnegative solution to the following equation:

$$y_i = \sum_{k\neq i,i_0} \frac{1}{q_i} q_{ik} y_k + \frac{\ell}{q_i}\mathbb{E}_i\sigma_{i_0}^{\ell-1}, \qquad i \in E.$$

As remarked in the last section, we may assume that $y_i^* < \infty$ for every $i \in E$. Then, by (4.2), we obtain the Poisson equation:

$$(Qy)_i = q_{ii_0}(1-\delta_{ii_0})y_{i_0} - \ell\mathbb{E}_i\sigma_{i_0}^{\ell-1}, \qquad i \in E.$$

Applying Theorem 1.1 to $c = 0$ and $f_i = q_{ii_0}(1 - \delta_{ii_0})y_{i_0} - \ell\mathbb{E}_i\sigma_{i_0}^{\ell-1}$, it follows that the solution to the last equation is as follows:

$$y_n = y_0 + \sum_{0\leqslant k\leqslant n-1}\sum_{j=0}^{k}\frac{F_k^{(j)}f_j}{q_{j,j+1}} = y_0 + y_{i_0}\sum_{0\leqslant k\leqslant n-1}u_k - \ell\sum_{0\leqslant k\leqslant n-1}v_k^{(\ell)}, \qquad n\geqslant 0.$$

Here in the summation of u_k, we have used the character of single birth: $q_{ji_0}(1-\delta_{ji_0}) > 0$ only if either $j = i_0 - 1$ or $j \geqslant i_0 + 1$. In particular, by setting $n = i_0$, it follows that

$$y_0 = \ell\sum_{0\leqslant k\leqslant i_0-1}v_k^{(\ell)} + y_{i_0}\biggl(1 - \sum_{0\leqslant k\leqslant i_0-1}u_k\biggr).$$

Return to the original y_n, we get

$$\begin{aligned} y_n &= \ell\biggl[\sum_{0\leqslant k\leqslant i_0-1}v_k^{(\ell)} - \sum_{0\leqslant k\leqslant n-1}v_k^{(\ell)}\biggr] + y_{i_0}\biggl[1 - \sum_{0\leqslant k\leqslant i_0-1}u_k + \sum_{0\leqslant k\leqslant n-1}u_k\biggr] \\ &= \begin{cases} -\ell\sum_{i_0\leqslant k\leqslant n-1}v_k^{(\ell)} + y_{i_0}\bigl[1 + \sum_{i_0\leqslant k\leqslant n-1}u_k\bigr], & n\geqslant i_0+1 \\ \ell\sum_{n\leqslant k\leqslant i_0-1}v_k^{(\ell)} + y_{i_0}\bigl[1 - \sum_{n\leqslant k\leqslant i_0-1}u_k\bigr], & n\leqslant i_0. \end{cases} \end{aligned} \tag{6.1}$$

When $n \leqslant i_0$, since $\sum_{k\leqslant i_0-1}u_k \leqslant 1$ by definition of (u_k), it is clear that $y_n > 0$. When $n \geqslant i_0 + 1$, for $y_n > 0$, one requires the condition

$$y_{i_0} > \frac{\ell\sum_{i_0\leqslant k\leqslant n-1}v_k^{(\ell)}}{1 + \sum_{i_0\leqslant k\leqslant n-1}u_k}$$

and then

$$y_{i_0} \geqslant \sup_{n\geqslant i_0+1}\frac{\ell\sum_{i_0\leqslant k\leqslant n-1}v_k^{(\ell)}}{1 + \sum_{i_0\leqslant k\leqslant n-1}u_k}.$$

By a reason explained in the last section, this leads to

$$y_{i_0}^* = \ell\varlimsup_{n\to\infty}\frac{\sum_{i_0\leqslant k\leqslant n}v_k^{(\ell)}}{1 + \sum_{i_0\leqslant k\leqslant n}u_k}$$

which gives us $\mathbb{E}_{i_0}\sigma_{i_0}^{\ell}$. Combining it with (6.1), we obtain the required assertion. The limit in $\mathbb{E}_{i_0}\sigma_{i_0}^{\ell}$ is again an application of Stolz's Theorem since $\sum_k u_k = \infty$ by the recurrence of the process. To see the last assertion, define a single birth process on $\{i_0, i_0+1, \ldots\}$ (regarding the set $\{0, 1, \ldots, i_0\}$ as a single state) with rates

$$\bar{q}_{ij} = \begin{cases} q_{ij} & \text{if } j\geqslant i_0+1 \\ \sum_{k\leqslant i_0}q_{ik} & \text{if } j = i_0, \qquad i\geqslant i_0. \end{cases}$$

Then $(\bar{q}_{ij})$ is irreducible and recurrent because so is (q_{ij}). Next, as in (1.1), we can define a sequence $\big(\bar{F}_k^{(j)}\big)$ on $\{i_0, i_0+1, \ldots\}$. By induction, it is easy to check that $\bar{F}_k^{(j)} = \widetilde{F}_k^{(j)}$ for every $k \geqslant j \geqslant i_0$. Hence we have

$$\sum_k \bar{F}_k^{(i_0)} = \sum_k \widetilde{F}_k^{(i_0)} = \infty$$

by Proposition 4.1. It should be now easy to see that $\sum_k u_k = \infty$ as claimed. □

Polynomial moments of life time

Recall that τ_n is the time of first hitting the state n. If we start from $i \leqslant n-1$, then τ_n coincides with the time of fist hitting the set $\{n, n+1, \ldots\}$. For the remainder of this section, we are going to study the time $\tau_\infty := \lim_{n\to\infty} \tau_n$. Next, because τ_∞ is actually equal to the life time $\eta := \lim_{n\to\infty} \eta_n$ almost everywhere, where $\{\eta_n\}$ are the successive jumping times:

$$\eta_0 \equiv 0, \qquad \eta_n = \inf\{t \geqslant \eta_{n-1} : X(t) \neq X(\eta_{n-1})\}, \qquad n \geqslant 1,$$

therefore, $\tau_\infty = \infty$ a.e. if the single birth Q-matrix is non-explosive. Thus, the study on the moments of τ_∞ is meaningful only for explosive single birth Q-matrix. The next result is taken from [21].

Proposition 6.2 *Let the single birth Q-matrix $Q = (q_{ij})$ be irreducible and explosive (i.e. $\sum_n m_n < \infty$ by Proposition 3.1). Assume that the minimal process has finite $(\ell-1)$-th moments of τ_∞ for some integer $\ell \geqslant 1$ (i.e. $E_i\tau_\infty^{\ell-1} < \infty$ for all $i \geqslant 0$). Then*

$$\mathbb{E}_n\tau_\infty^\ell = \ell \sum_{k\geqslant n} \bar{m}_k^{(\ell)}, \qquad n \geqslant 0,$$

where

$$\bar{m}_n^{(\ell)} = \frac{1}{q_{n,n+1}}\left[\mathbb{E}_n\tau_\infty^{\ell-1} + \sum_{0\leqslant k\leqslant n-1} q_n^{(k)}\bar{m}_k^{(\ell)}\right] = \sum_{j=0}^{n} \frac{F_n^{(j)}\mathbb{E}_j\tau_\infty^{\ell-1}}{q_{j,j+1}}, \qquad n \geqslant 0.$$

Proof The last equality of $\bar{m}_n^{(\ell)}$ comes from Corollary 2.3. By [4; Proposition 4.56] or [11], we know that $(\mathbb{E}_i\tau_\infty^\ell : i \in E)$ is the the minimal nonnegative solution $(y_i^* : i \in E)$ to the following equation:

$$y_i = \sum_{k\neq i} \frac{1}{q_i} q_{ik} y_k + \frac{\ell}{q_i}\mathbb{E}_i\tau_\infty^{\ell-1}, \qquad i \in E.$$

That is,

$$(Qy)_i = -\ell\mathbb{E}_i\tau_\infty^{\ell-1}, \qquad i \in E.$$

Applying Theorem 1.1 to $c = 0$ and $f_i = -\ell \mathbb{E}_i \tau_\infty^{\ell-1}$ $(i \geqslant 0)$, it follows that the solution to the last equation can be expressed as

$$y_n = y_0 - \ell \sum_{k=0}^{n-1} \sum_{j=0}^{k} \frac{F_k^{(j)} \mathbb{E}_j \tau_\infty^{\ell-1}}{q_{j,j+1}}, \qquad n \geqslant 1.$$

Hence

$$y_n = y_0 - \ell \sum_{k=0}^{n-1} \bar{m}_k^{(\ell)}, \qquad n \geqslant 1.$$

By the nonnegative and minimal properties, it follows that

$$y_0^* = \sup_{n \geqslant 1} \left(\ell \sum_{k=0}^{n-1} \bar{m}_k^{(\ell)} \right) = \ell \sum_{k=0}^{\infty} \bar{m}_k^{(\ell)}, \qquad y_n^* = \ell \sum_{k=n}^{\infty} \bar{m}_k^{(\ell)}, \qquad n \geqslant 1.$$

Hence, we obtain

$$\mathbb{E}_n \tau_\infty^\ell = \ell \sum_{k \geqslant n} \bar{m}_k^{(\ell)}, \qquad n \geqslant 0$$

which is the required assertion. □

7 Exponential ergodicity and Laplace transform of return time

Exponential moments of return time and exponential ergodicity

In this section, we consider the exponential moments of return time. At first, we introduce the following lemma for general Q-matrices.

Lemma 7.1 *Let (q_{ij}) be irreducible and assume that its Q-process is recurrent. Next, let $\lambda \in \mathbb{R}$, $\lambda < q_i$ for every $i \in E$. Then for fixed $H \subset E$, $H \neq \varnothing, E$, $(\mathbb{E}_i \exp(\lambda \sigma_H) : i \in E)$ is the minimal solution to the equation*

$$x_i = \frac{1}{q_i - \lambda} \sum_{k \notin H \cup \{i\}} q_{ik} x_k + \frac{1}{q_i - \lambda} \sum_{k \in H \setminus \{i\}} q_{ik}, \qquad i \in E. \tag{7.1}$$

Proof Let $(y_i^* : i \in E)$ be the minimal nonnegative solution to the equation

$$y_i = \frac{1}{q_i - \lambda} \sum_{k \notin H \cup \{i\}} q_{ik} y_k + \frac{1}{q_i - \lambda}, \qquad i \in E.$$

By the recurrent assumption and [4; Lemma 4.46], the quantity f_{iH} defined there is equal to 1 for every $i \in E$. Then, $(y_i^* : i \in E)$ coincides with $(e_{iH}(\lambda) : i \in E)$ used in [4; Lemma 4.48]. Moreover, by the proof given on [4; page 148], we have $\mathbb{E}_i \exp(\lambda \sigma_H) = 1 + \lambda y_i^*$ for every $i \in E$. Besides, it can be

checked that $(1+\lambda y_i^* : i\in E)$ is a nonnegative solution to equation (7.1). Hence $\mathbb{E}_i \exp(\lambda\sigma_H) = 1+\lambda y_i^* \geqslant x_i^*$ for every $i\in E$, where $(x_i^* : i\in E)$ is the minimal nonnegative solution to equation (7.1). We are now going to prove that $\mathbb{E}_i \exp(\lambda\sigma_H) = x_i^*$ for all $i\in E$. The proof is split into two parts: either $\lambda\geqslant 0$ or $\lambda<0$.

First, let $\lambda\geqslant 0$. It is easily seen that $(x_i^*-1 : i\in E)$ is a nonnegative solution to the equation

$$y_i = \frac{1}{q_i-\lambda}\sum_{k\notin H\cup\{i\}} q_{ik}y_k + \frac{\lambda}{q_i-\lambda}, \qquad i\in E.$$

Hence, $x_i^*-1\geqslant \lambda y_i^*$ since (λy_i^*) is the minimal nonnegative solution to the equation above, by the linear combination theorem [4; Theorem 2.12 (1)]. That is, $x_i^*\geqslant 1+\lambda y_i^*$. Combining what we have proved in the last paragraph, it follows that $x_i^* = \mathbb{E}_i\exp(\lambda\sigma_H)$ for all $i\in E$.

Next, let $\lambda<0$. Denote by $(\bar{y}_i : i\in E)$ the minimal nonnegative solution to the equation

$$y_i = \frac{1}{q_i-\lambda}\sum_{k\notin H\cup\{i\}} q_{ik}y_k + \left[1-\frac{1}{q_i-\lambda}\sum_{k\notin H\cup\{i\}} q_{ik}\right], \qquad i\in E. \tag{7.2}$$

Clearly, we have $\bar{y}_i\leqslant 1$ since $y_i\equiv 1$ is a solution to the equation. We claim that $\bar{y}_i\equiv 1$. To see this, note that $(1-\bar{y}_i : i\in E)$ is the maximal solution to the equation

$$y_i = \frac{1}{q_i-\lambda}\sum_{k\notin H\cup\{i\}} q_{ik}y_k, \qquad 0\leqslant y_i\leqslant 1, \qquad i\in E. \tag{7.3}$$

By a comparison lemma [4; Lemma 3.14], it suffices to show that the equation

$$y_i = \frac{1}{q_i}\sum_{k\notin H\cup\{i\}} q_{ik}y_k, \qquad 0\leqslant y_i\leqslant 1, \qquad i\in E$$

has only trivial (i.e. zero-) solution. Then this follows by the recurrence assumption and [4; Lemma 4.46]. We remark that there is an alternative way to prove that $\bar{y}_i\equiv 1$, using the uniqueness rather than the recurrence assumption. Actually, equation (7.3) is an exit equation for a modified Q-matrix (any local modification of a Q-matrix does not interfere the uniqueness). The exit solution to (7.3) should be zero by uniqueness assumption.

We now return to our main proof. By the linear combination theorem [4; Theorem 2.12 (1)], $(x_i^*-\lambda y_i^* : i\in E)$ is the minimal nonnegative solution to equation (7.2). Hence $x_i^*-\lambda y_i^* = \bar{y}_i\equiv 1$ as we have just proved in the last paragraph. Therefore we conclude that $x_i^* = 1+\lambda y_i^* = \mathbb{E}_i\exp(\lambda\sigma_H)$ for all $i\in E$. We have thus completed the proof of the lemma. □

Now we present our results about the exponential moments of the return time σ_0, which can be referred in [18].

Proposition 7.2 *Let the single birth Q-matrix (q_{ij}) be irreducible. Assume that its process is ergodic. Define $\big(\widetilde{F}_k^{(i)}\big)$ and $\big(\tilde{d}_k\big)$ by setting $c_i \equiv \lambda > 0$. Then for small λ,*

$$\mathbb{E}_0 e^{\lambda\sigma_0} = \frac{q_{01}(1+\lambda\tilde{d})}{q_{01}-\lambda} < \infty \quad \text{and} \quad \mathbb{E}_n e^{\lambda\sigma_0} = 1 + \lambda \sum_{k=0}^{n-1} \Big(\widetilde{F}_k^{(0)}\tilde{d} - \tilde{d}_k\Big) < \infty, \quad n \geqslant 1$$

iff

$$\tilde{d} := \overline{\lim_{n\to\infty}}\, \mathbb{1}_{\{\sum_{k=0}^n \widetilde{F}_k^{(0)} > 0\}} \frac{\sum_{k=0}^n \tilde{d}_k}{\sum_{k=0}^n \widetilde{F}_k^{(0)}} < \infty$$

and

$$\tilde{d} \sum_{k=0}^{n-1} \widetilde{F}_k^{(0)} > \sum_{k=0}^{n-1} \tilde{d}_k \quad \text{whenever} \quad \sum_{k=0}^{n-1} \widetilde{F}_k^{(0)} \leqslant 0 \quad \text{for } n \geqslant 2. \tag{7.4}$$

Furthermore, once $\widetilde{F}_n^{(0)} > 0$ for large enough n and $\sum_n \widetilde{F}_n^{(0)} = \infty$, we have

$$\tilde{d} = \lim_{n\to\infty} \frac{\tilde{d}_n}{\widetilde{F}_n^{(0)}} \text{ if the limit exists.}$$

Finally, the process is exponentially ergodic iff both $\tilde{d} < \infty$ and (7.4) holds.

Proof Let $\lambda \in (0, q_i)$ for every $i \in E$ and set $H = \{0\}$. Then by Lemma 7.1, $(\mathbb{E}_i \mathrm{e}^{\lambda\sigma_0} : i \in E)$ is the minimal solution (x_i^*) of the following equation

$$x_i = \frac{1}{q_i - \lambda} \sum_{k \notin \{0,i\}} q_{ik} x_k + \frac{q_{i0}(1-\delta_{i0})}{q_i - \lambda}, \qquad x_i \geqslant 1, \qquad i \in E.$$

Assume that $x_i^* < \infty$ for every $i \in E$ for a moment, and let (x_i) be a finite nonnegative solution to the last equation. Then, by (4.2), we have

$$(Qx)_i + \lambda x_i = q_{i0}(x_0 - 1), \qquad i \geqslant 1; \qquad (Qx)_0 + \lambda x_0 = 0. \tag{7.5}$$

Applying Theorem 1.1 to $c_i \equiv \lambda$ and $f_i = q_{i0}(1-\delta_{i0})(x_0 - 1)$ for all $i \geqslant 0$, we obtain

$$\begin{aligned} x_n &= x_0 \left(1 - \lambda \sum_{k=0}^{n-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)}}{q_{j,j+1}}\right) + (x_0 - 1) \sum_{k=1}^{n-1} \sum_{j=1}^{k} \frac{\widetilde{F}_k^{(j)} q_{j0}}{q_{j,j+1}} \\ &= x_0 \left(1 - \lambda \sum_{k=0}^{n-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)}}{q_{j,j+1}}\right) + (x_0 - 1) \sum_{k=1}^{n-1} \sum_{j=1}^{k} \frac{\widetilde{F}_k^{(j)} (\tilde{q}_j^{(0)} + \lambda)}{q_{j,j+1}}, \qquad n \geqslant 1. \end{aligned}$$

Due to the explicit representation of $\widetilde{F}_n^{(k)}$, $\widetilde{m}_n$ and $\tilde{d}_n$, given in (2.7), (3.2) and (5.2) respectively, we have not only

$$\widetilde{m}_n = \sum_{0 \leqslant j \leqslant n} \frac{\widetilde{F}_n^{(j)}}{q_{j,j+1}} = \frac{1}{q_{01}} \widetilde{F}_n^{(0)} + \tilde{d}_n, \qquad n \geqslant 0 \tag{7.6}$$

but also that

$$\begin{aligned} x_n &= x_0\bigg(1-\lambda\sum_{k=0}^{n-1}\widetilde{m}_k\bigg)+(x_0-1)\sum_{k=1}^{n-1}\big(\widetilde{F}_k^{(0)}+\lambda\tilde{d}_k\big)\\ &= x_0\bigg(1-\frac{\lambda}{q_{01}}\bigg)\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)}-\sum_{k=0}^{n-1}\big(\widetilde{F}_k^{(0)}+\lambda\tilde{d}_k\big)+1, \qquad n\geqslant 1. \end{aligned} \tag{7.7}$$

Since $x_n>1$, we get

$$x_0\bigg(1-\frac{\lambda}{q_{01}}\bigg)\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)}>\sum_{k=0}^{n-1}\big(\widetilde{F}_k^{(0)}+\lambda\tilde{d}_k\big), \qquad n\geqslant 1.$$

That is

$$\bigg[x_0\bigg(\frac{1}{\lambda}-\frac{1}{q_{01}}\bigg)-\frac{1}{\lambda}\bigg]\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)}>\sum_{k=0}^{n-1}\tilde{d}_k, \qquad n\geqslant 1. \tag{7.8}$$

Note that on the one hand, if $x_0^*=x_0^*(\lambda_0)<\infty$, then $x_0^*=x_0^*(\lambda)<\infty$ for every $\lambda\in(0,\lambda_0)$, by the comparison theorem (cf. [4; Theorem 2.6]). On the other hand, when $\lambda=0$, we have

$$\sum_{k=0}^{n}\widetilde{F}_k^{(0)}=\sum_{k=0}^{n}F_k^{(0)}>0 \quad\text{and}\quad \sum_{k=0}^{n}\tilde{d}_k=\sum_{k=0}^{n}d_k>0, \qquad n\geqslant 1.$$

For each fixed n, $\sum_{k=0}^{n}\widetilde{F}_k^{(0)}$ and $\sum_{k=0}^{n}\tilde{d}_k$ are analytic in λ, and so should be positive for sufficient small λ, say $\lambda\leqslant\lambda_1$ for some $\lambda_1\leqslant\lambda_0$. Then by (7.8), we should have

$$x_0\bigg(\frac{1}{\lambda}-\frac{1}{q_{01}}\bigg)-\frac{1}{\lambda}>0, \qquad \lambda\in(0,\lambda_1)$$

independent of n. Therefore, by the minimal property, we have

$$x_0^*\bigg(\frac{1}{\lambda}-\frac{1}{q_{01}}\bigg)-\frac{1}{\lambda}=\varlimsup_{n\to\infty}\mathbb{1}_{\{\sum_{k=0}^{n}\widetilde{F}_k^{(0)}>0\}}\bigg[\sum_{k=0}^{n}\tilde{d}_k\bigg]\bigg[\sum_{k=0}^{n}\widetilde{F}_k^{(0)}\bigg]^{-1}=\tilde{d},$$

i.e.

$$\mathbb{E}_0\mathrm{e}^{\lambda\sigma_0}=x_0^*=\frac{q_{01}(1+\lambda\tilde{d})}{q_{01}-\lambda}. \tag{7.9}$$

Since x_0^* satisfies (7.8), we obtain condition (7.4). Then

$$\mathbb{E}_n\mathrm{e}^{\lambda\sigma_0}=1+\lambda\sum_{k=0}^{n-1}\Big(\widetilde{F}_k^{(0)}\tilde{d}-\tilde{d}_k\Big), \qquad n\geqslant 1.$$

Conversely, if $\tilde{d}<\infty$ and (7.4) holds. Then starting from $x_0=x_0^*$ given in (7.9) and defining x_n by (7.7), we obtain a solution $(x_i>1: i\in E)$ to (7.5).

By (4.2), we obtain a finite nonnegative solution to the original equation for $\big(\mathbb{E}_i e^{\lambda\sigma_0} : i \in E\big)$, and hence the minimal solution $\big(x_i^* = \mathbb{E}_i e^{\lambda\sigma_0} : i \in E\big)$ should be finite.

Finally, by [4; Theorem 4.44], the process is exponentially ergodic iff $\mathbb{E}_0 \mathrm{e}^{\lambda\sigma_0} < \infty$, equivalently, $\tilde{d} < \infty$ and (7.4) holds. The last assertion of the proposition then follows. □

In contract to the ergodic case, one may study the exponential decay (in the transient case) for which the Poisson equation becomes*

$$Qg + \lambda g = 0, \qquad g > 0.$$

With $c_i \equiv \lambda$, by Theorem 1.1, the solution is

$$g_n = g_0\bigg[1 - \lambda \sum_{0\leqslant k\leqslant n-1} \sum_{0\leqslant j\leqslant k} \frac{\widetilde{F}_k^{(j)}}{q_{j,j+1}}\bigg] = g_0\bigg[1 - \lambda \sum_{0\leqslant k\leqslant n-1} \widetilde{m}_k\bigg], \qquad n \geqslant 0.$$

This is somehow simpler than the previous one. However, these two exponential cases are actually much harder than the others, for instance we do not know at the moment how to remove condition (7.4). That is showing for some $\lambda > 0$, small enough, $\sum_{k=0}^n \widetilde{F}_k^{(0)} > 0$ for all n $\big($or equivalently, $\underline{\lim}_{n\to\infty} \sum_{k=0}^n \widetilde{F}_k^{(0)} > 0\big)$. This seems necessary for the exponential ergodicity since $\sum_{k=0}^\infty \widetilde{F}_k^{(0)} = \infty$ when $\lambda = 0$ by the recurrence (which is much weaker than exponential ergodicity) and λ is allowed to be very small. Actually, to figure out a criterion, one needs much more work using different approaches, refer to [4; Chapter 9] and [7] for some details.

Laplace transform of the return/extinction time

Note that for negative λ, $\mathbb{E}_i \mathrm{e}^{\lambda\sigma_0}$ is the Laplace transform of σ_0. The proof of Proposition 7.2 is still available. So we get the following result.

Proposition 7.3 *Define* $\big(\widetilde{F}_k^{(i)}\big)$ *and* $\big(\tilde{d}_k\big)$ *by* (1.1) *and* (5.1)*, respectively, with* $c_i \equiv -\lambda < 0$*. Let the single birth process be recurrent. Then the Laplace transform of* σ_0 *is given by*

$$\mathbb{E}_0 e^{-\lambda\sigma_0} = \frac{q_{01}(1 - \lambda\tilde{d})}{q_{01} + \lambda}, \qquad \mathbb{E}_n e^{-\lambda\sigma_0} = 1 - \lambda \sum_{k=0}^{n-1} \Big(\widetilde{F}_k^{(0)}\tilde{d} - \tilde{d}_k\Big), \qquad n \geqslant 1,$$

where

$$\tilde{d} = \lim_{n\to\infty} \frac{\sum_{k=0}^{n-1} \tilde{d}_k}{\sum_{k=0}^{n-1} \widetilde{F}_k^{(0)}} = \lim_{n\to\infty} \frac{\tilde{d}_n}{\widetilde{F}_n^{(0)}} \quad \text{if the limit exists.}$$

*See the footnote on page 775

Proof Following the proof of Proposition 7.2, replacing λ by $-\lambda$, we arrive at

$$\begin{aligned} x_n &= x_0\bigg(1+\frac{\lambda}{q_{01}}\bigg)\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)} - \sum_{k=0}^{n-1}\big(\widetilde{F}_k^{(0)} - \lambda\tilde{d}_k\big) + 1, \\ &=: x_0\alpha_{n-1} - \beta_{n-1}, \qquad n \geqslant 1. \end{aligned}$$

By the minimal nonnegative property, $x_0^* = \sup_{n\geqslant 1}\beta_n/\alpha_n$, and then we indeed have

$$x_0^* = \overline{\lim_{n\to\infty}}\frac{\beta_n}{\alpha_n}.$$

We now show that we can replace $\overline{\lim}_{n\to\infty}$ by $\lim_{n\to\infty}$. Noting that on the one hand, since $x_n \in (0,1]$, we have

$$\frac{\beta_n}{\alpha_n} < x_0 \leqslant \frac{\beta_n+1}{\alpha_n}, \qquad n \geqslant 1.$$

On the other hand, following the proof for

$$\sum_k \widetilde{m}_k = \infty \Longleftrightarrow \sum_k m_k = \infty$$

given in Section 3, we can prove that $\sum_k \widetilde{F}_k^{(0)} = \infty$ since $\sum_k F_k^{(0)} = \infty$ by the recurrent assumption (i.e. $\gamma_j \equiv 1$). Hence we can rewrite $\overline{\lim}_{n\to\infty}\beta_n/\alpha_n$ as $\lim_{n\to\infty}\beta_n/\alpha_n$. Therefore, we have

$$\begin{aligned} x_0^* &= \lim_{n\to\infty}\bigg[\sum_{k=0}^{n-1}\big(\widetilde{F}_k^{(0)} - \lambda\tilde{d}_k\big)\bigg]\bigg\{\bigg[1+\frac{\lambda}{q_{01}}\bigg]\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)}\bigg\}^{-1} \\ &= \frac{q_{01}}{q_{01}+\lambda}\lim_{n\to\infty}\bigg[1-\lambda\frac{\sum_{k=0}^{n-1}\tilde{d}_k}{\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)}}\bigg] \\ &= \frac{q_{01}}{q_{01}+\lambda}\big[1-\lambda\tilde{d}\big]. \end{aligned} \quad .$$

Furthermore,

$$x_n^* = (1-\lambda\tilde{d})\sum_{k=0}^{n-1}\widetilde{F}_k^{(0)} - \sum_{k=0}^{n-1}\big(\widetilde{F}_k^{(0)} - \lambda\tilde{d}_k\big) + 1 = 1 - \lambda\sum_{k=0}^{n-1}\Big(\widetilde{F}_k^{(0)}\tilde{d} - \tilde{d}_k\Big), \quad n \geqslant 1.$$

The last limit in $\tilde{d}$ is an application of Stolz's Theorem. □

Exponential moments and Laplace transform of the life time

Now we return to τ_∞.

Proposition 7.4 *Assume that the single birth Q-matrix $Q = (q_{ij})$ is explosive and irreducible. Define $(\widetilde{m}_k)$ by (3.1) with $c_i \equiv \lambda$. For the corresponding minimal process,*

(i) if there exists a $\lambda > 0$ such that $\lambda \sum_{k=0}^{n-1} \widetilde{m}_k < 1$ for every $n > 1$, then

$$\mathbb{E}_n e^{\lambda \tau_\infty} = 1 + \lambda \left[\bar{c} \left(1 - \lambda \sum_{k=0}^{n-1} \widetilde{m}_k \right) - \sum_{k=0}^{n-1} \widetilde{m}_k \right], \qquad n \geqslant 0,$$

where

$$\bar{c} = \overline{\lim_{n\to\infty}} \frac{\sum_{k=0}^{n} \widetilde{m}_k}{1 - \lambda \sum_{k=0}^{n} \widetilde{m}_k}.$$

Furthermore, the process decays exponentially fast provided $\bar{c} < \infty$.

(ii) For $\lambda > 0$, the Laplace transform of τ_∞ is given by

$$\mathbb{E}_n e^{-\lambda \tau_\infty} = \frac{1 + \lambda \sum_{0 \leqslant k \leqslant n-1} \widetilde{m}_k}{1 + \lambda \sum_{k \geqslant 0} \widetilde{m}_k}, \qquad n \geqslant 0.$$

Proof Define

$$e_{i\infty}(\lambda) = \int_0^\infty \mathrm{e}^{\lambda t} \mathbb{P}_i(\tau_\infty > t)\, \mathrm{d}t$$

with $\lambda < q_i$ for all $i \geqslant 0$. Note that the process is explosive and

$$\mathbb{E}_i \mathrm{e}^{\lambda \tau_\infty} = 1 + \lambda e_{i\infty}(\lambda).$$

Because $\mathbb{P}_m(\tau_n < \eta) = 1$ for every pair $m < n$, we have $\mathbb{P}_m(\tau_n < \infty) = 1$ and furthermore $\mathbb{P}_m(\tau_\infty < \infty) = 1$ for every m, as n goes to ∞. Then by [4; Lemma 4.48], $(e_{i\infty}(\lambda))$ is the minimal solution to the equation

$$x_i = \frac{q_i}{q_i - \lambda} \sum_k \Pi_{ik} x_k + \frac{1}{q_i - \lambda}, \qquad i \geqslant 0.$$

By (4.2), we can rewrite the equation as

$$(Qx)_i + \lambda x_i = -1, \qquad i \geqslant 0.$$

Applying Theorem 1.1 to $c_i \equiv \lambda$ and $f_i \equiv -1$, the solution of the equation has the form:

$$\begin{aligned} x_n &= x_0 \left(1 - \lambda \sum_{k=0}^{n-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)}}{q_{j,j+1}} \right) - \sum_{k=0}^{n-1} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)}}{q_{j,j+1}} \\ &= x_0 \left(1 - \lambda \sum_{k=0}^{n-1} \widetilde{m}_k \right) - \sum_{k=0}^{n-1} \widetilde{m}_k, \qquad n \geqslant 1. \end{aligned}$$

Note that $\lambda < q_0 = q_{01}$ and $\lambda \widetilde{m}_0 < 1$. If there exists a positive λ small enough so that $\lambda \sum_{k=0}^{n-1} \widetilde{m}_k < 1$ for every $n > 1$, then by the argument above and the minimal property of the solution, one gets

$$e_{0\infty}(\lambda) = \sup_{n \geqslant 1} \frac{\sum_{k=0}^{n-1} \widetilde{m}_k}{1 - \lambda \sum_{k=0}^{n-1} \widetilde{m}_k} = \overline{\lim_{n\to\infty}} \frac{\sum_{k=0}^{n} \widetilde{m}_k}{1 - \lambda \sum_{k=0}^{n} \widetilde{m}_k} =: \bar{c}$$

and

$$e_{n\infty}(\lambda)=\bar{c}\left(1-\lambda\sum_{k=0}^{n-1}\widetilde{m}_k\right)-\sum_{k=0}^{n-1}\widetilde{m}_k,\qquad n\geqslant 1.$$

Then the first assertion follows.

For the Laplace transform of τ_∞, the argument above still works because now we deal with the case of $-\lambda<0$. By the explosive property, we know that $\sum_{k=0}^{\infty}\widetilde{m}_k<\infty$. Hence we have

$$e_{0\infty}(-\lambda)=\bar{c}=\frac{\sum_{k=0}^{\infty}\widetilde{m}_k}{1+\lambda\sum_{k=0}^{\infty}\widetilde{m}_k}$$

and

$$e_{n\infty}(-\lambda)=\bar{c}\left(1+\lambda\sum_{k=0}^{n-1}\widetilde{m}_k\right)-\sum_{k=0}^{n-1}\widetilde{m}_k=\frac{\sum_{k=n}^{\infty}\widetilde{m}_k}{1+\lambda\sum_{k=0}^{\infty}\widetilde{m}_k},\qquad n\geqslant 1.$$

Finally, we have

$$\mathbb{E}_n\mathrm{e}^{-\lambda\tau_\infty}=1-\frac{\lambda\sum_{k=n}^{\infty}\widetilde{m}_k}{1+\lambda\sum_{k=0}^{\infty}\widetilde{m}_k}=\frac{1+\lambda\sum_{0\leqslant k\leqslant n-1}\widetilde{m}_k}{1+\lambda\sum_{k\geqslant 0}\widetilde{m}_k},\qquad n\geqslant 0.$$

The proof for the second assertion is now finished. □

A more careful study on part (i) of Proposition 7.4, refer to Proposition 7.2.

8 Examples

In the special case of birth–death processes, the problems studied here have rather complete solutions, see for instance [4; Theorem 4.55]. As mentioned in the introduction of the paper, much more models have been studied in the past years. Here we make a little addition. The next example is taken from [3].

Example 8.1 (uniform catastrophes) Let

$$q_{i,i+1}=b\,i,\qquad i\geqslant 0;\qquad q_{ij}=a,\qquad j=0,1,\ldots,i-1;$$

and $q_{ij}=0$ for other $j>i+1$, where a and b are positive constants. Then the extinction of the process has an exponential distribution

$$\mathbb{E}_n\mathrm{e}^{-\lambda\tau_0}=\frac{a}{a+\lambda},\qquad \lambda>0,\ n\geqslant 1.$$

It is surprising that the distribution is independent of b and the starting point n. Redefine $q_{01}=1$. Then the irreducible process is indeed strongly ergodic.

Proof We need to consider the case that $q_{01} > 0$ only. With $c_i \equiv -\lambda \in \mathbb{R}$ and then $\tilde{q}_n^{(k)} = (k+1)a + \lambda$ for $k \leqslant n-1$, by using (1.1), (5.1), and induction, one may check that

$$\widetilde{F}_n^{(0)} = \frac{a+\lambda}{nb} \prod_{1 \leqslant k \leqslant n-1} \left(1 + \frac{(k+1)a+\lambda}{kb}\right), \qquad \prod_{\varnothing} =: 1,$$
$$\tilde{d}_n = \frac{1}{nb} \prod_{1 \leqslant k \leqslant n-1} \left(1 + \frac{(k+1)a+\lambda}{kb}\right), \qquad n \geqslant 1.$$

Since for each fixed $\lambda \in \mathbb{R}$,

$$\log\left(1 + \frac{(n+1)a+\lambda}{nb}\right) \to \log\left(1 + \frac{a}{b}\right) > 0 \qquad \text{as } n \to \infty,$$

we have $\lim_{n\to\infty} \widetilde{F}_n^{(0)} = \infty$ and so $\sum_n \widetilde{F}_n^{(0)} = \infty$. As an application of this fact with $\lambda = 0$, it follows that the process is recurrent (Proposition 4.1) and then should be non-explosive ((7.6) and Proposition 3.1).

Next, because

$$\sum_n \widetilde{F}_n^{(0)} = \infty, \qquad \widetilde{F}_n^{(0)} = (a+\lambda)\tilde{d}_n, \ \ n \geqslant 1,$$

it follows that

$$\tilde{d} = \lim_{n\to\infty} \frac{\tilde{d}_n}{\widetilde{F}_n^{(0)}} = \frac{1}{a+\lambda}.$$

Hence, we have

$$\widetilde{F}_n^{(0)} \tilde{d} = \tilde{d}_n, \qquad n \geqslant 1,$$

From here, when $\lambda = 0$ in particular, we obtain

$$\sup_k \sum_{n=0}^{k} \left(F_n^{(0)} d - d_n\right) = d = a^{-1} < \infty.$$

Hence the process is strongly ergodic by Proposition 5.2.

By using Proposition 7.3, we obtain

$$\mathbb{E}_0 \mathrm{e}^{-\lambda\sigma_0} = \frac{a q_{01}}{(a+\lambda)(q_{01}+\lambda)},$$
$$\mathbb{E}_n \mathrm{e}^{-\lambda\sigma_0} = 1 - \lambda\tilde{d} = \frac{a}{a+\lambda} = \mathbb{E}_n \mathrm{e}^{-\lambda\tau_0}, \qquad n \geqslant 1.$$

Therefore, we have proved the first assertion.

Even though it is now automatic that the process is exponentially ergodic, implied by the strongly ergodicity, we would like to check the effectiveness of Proposition 7.2 for this model. To do so, reset $c_i \equiv \lambda > 0$. Then

$$\widetilde{F}_n^{(0)} = \frac{a-\lambda}{nb} \prod_{1\leqslant k\leqslant n-1} \left(1 + \frac{(k+1)a-\lambda}{kb}\right),$$
$$\tilde{d}_n = \frac{1}{nb} \prod_{1\leqslant k\leqslant n-1} \left(1 + \frac{(k+1)a-\lambda}{kb}\right), \qquad n \geqslant 1.$$

Clearly, $\widetilde{F}_n^{(0)} > 0$ and so does $\tilde{d}_n$ for every $\lambda \in (0, a)$. As we have proved above

$$\sum_n \widetilde{F}_n^{(0)} = \infty, \qquad \tilde{d} = \lim_{n\to\infty} \frac{\tilde{d}_n}{\widetilde{F}_n^{(0)}} = \frac{1}{a-\lambda} < \infty,$$

and hence the process is exponentially ergodic by Proposition 7.2. Actually, we have

$$\mathbb{E}_0 \mathrm{e}^{\lambda\sigma_0} = \frac{aq_{01}}{(a-\lambda)(q_{01}-\lambda)},$$
$$\mathbb{E}_n \mathrm{e}^{\lambda\sigma_0} = \frac{a}{a-\lambda}, \qquad n \geqslant 1, \quad \lambda \in (0, a \wedge q_{01}). \qquad \square$$

Example 8.2 Consider the single birth Q-matrix (q_{ij}) with

$$q_{i0} > 0, \ \ q_{i,i+1} > 0, \ \ q_{ij} = 0 \ \text{ for all other } j \neq i.$$

Let $c_i \in \mathbb{R}$. Then

(1) we have

$$\widetilde{F}_i^{(i)} = 1, \quad \widetilde{F}_n^{(i)} = \frac{q_{n0}-c_n}{q_{n,n+1}} \prod_{i+1\leqslant k\leqslant n-1} \left[1 + \frac{q_{k0}-c_k}{q_{k,k+1}}\right], \tag{8.1}$$
$$\prod_{\varnothing} =: 1, \qquad n > i \geqslant 0,$$

and then $(\tilde{m}_n)$ and $\big(\tilde{d}_n\big)$ are given by (3.2) and (5.2), respectively.

(2) In particular, if $q_{n0} - c_n \equiv q_{10} - c_1$ for every $n \geqslant 1$, then

$$\widetilde{F}_i^{(i)} = 1, \ \ \widetilde{F}_n^{(i)} = \frac{q_{10}-c_1}{q_{n,n+1}} \prod_{k=i+1}^{n-1} \left[1 + \frac{q_{10}-c_1}{q_{k,k+1}}\right], \ \prod_{\varnothing} =: 1, \ \ n > i \geqslant 0,$$
$$\widetilde{m}_0 = \frac{1}{q_{01}}, \ \ \widetilde{m}_n = \frac{1}{q_{n,n+1}} \prod_{k=0}^{n-1} \left[1 + \frac{q_{10}-c_1}{q_{k,k+1}}\right], \qquad n \geqslant 1,$$
$$\tilde{d}_0 = 0, \ \ \tilde{d}_n = \frac{1}{q_{n,n+1}} \prod_{1\leqslant k\leqslant n-1} \left[1 + \frac{q_{10}-c_1}{q_{k,k+1}}\right], \qquad n \geqslant 1.$$

Furthermore, the process is explosive if

$$\kappa' := \lim_{n\to\infty} \frac{n(q_{n+1,n+2} - q_{n,n+1} - q_{10})}{q_{n,n+1} + q_{10}} > 1$$

($q_{n,n+1} = (n+1)^\gamma$ for $\gamma > 1$ for example). Otherwise, if $\kappa' < 1$ ($q_{n,n+1} = (n+1)^\gamma$ for some $\gamma \leqslant 1$ for instance), then the process is unique. If so, the process is indeed strongly ergodic.

Proof (a) By assumption, we have $\tilde{q}_n^{(k)} = q_{n0} - c_n$ for every $k < n$. Hence, by (1.1), we obtain

$$\widetilde{F}_n^{(i)} = \frac{\tilde{q}_n^{(0)}}{q_{n,n+1}} \sum_{k=i}^{n-1} \widetilde{F}_k^{(i)}. \tag{8.2}$$

Thus, to prove (8.1), it suffices to show that

$$\sum_{k=i}^{n-1} \widetilde{F}_k^{(i)} = \prod_{i+1\leqslant k\leqslant n-1} \left[1 + \frac{\tilde{q}_k^{(0)}}{q_{k,k+1}}\right], \qquad n > i \geqslant 0.$$

This clearly holds when $n = i+1$. Suppose that it holds when $n = \ell$, then

$$\begin{aligned}
\sum_{k=i}^{\ell} \widetilde{F}_k^{(i)} &= \sum_{k=i}^{\ell-1} \widetilde{F}_k^{(i)} + \widetilde{F}_\ell^{(i)} \\
&= \sum_{k=i}^{\ell-1} \widetilde{F}_k^{(i)} + \frac{\tilde{q}_\ell^{(0)}}{q_{\ell,\ell+1}} \sum_{k=i}^{\ell-1} \widetilde{F}_k^{(i)} \quad \text{(by (8.2))} \\
&= \left[1 + \frac{\tilde{q}_\ell^{(0)}}{q_{\ell,\ell+1}}\right] \sum_{k=i}^{\ell-1} \widetilde{F}_k^{(i)} \\
&= \prod_{i+1\leqslant k\leqslant \ell} \left[1 + \frac{\tilde{q}_k^{(0)}}{q_{k,k+1}}\right] \quad \text{(by inductive assumption).}
\end{aligned}$$

Therefore, the required assertion holds for $n = \ell$ and it then holds for all $n > i$ by induction. We have thus proved the first assertion.

(b) By assumption, we have $\tilde{q}_n^{(k)} = q_{10} - c_1$ for every $k < n$. Hence, by (3.1) and (5.1), we obtain

$$\begin{aligned}
\widetilde{m}_n &= \frac{1}{q_{n,n+1}} \left(1 + \tilde{q}_1^{(0)} \sum_{k=0}^{n-1} \widetilde{m}_k\right), \qquad n \geqslant 1, \\
\tilde{d}_n &= \frac{1}{q_{n,n+1}} \left(1 + \tilde{q}_1^{(0)} \sum_{k=0}^{n-1} \tilde{d}_k\right), \qquad n \geqslant 1.
\end{aligned}$$

As in the last proof, by using induction, we obtain the explicit expressions of $(\widetilde{m}_n)$ and $\big(\tilde{d}_n\big)$.

To study the divergence of $\sum_n m_n$, we adopt the

Kummer Test Let (u_n) and (v_n) be two sequences of positive numbers. Suppose that $\sum_0^\infty 1/v_n = \infty$ and the limit $\kappa := \lim_{n\to\infty}\kappa_n$ exists, where

$$\kappa_n = v_n \cdot \frac{u_n}{u_{n+1}} - v_{n+1}.$$

Then, the series $\sum u_n$ converges or diverges according to $\kappa > 0$ or $\kappa < 0$ respectively.

Set $v_n \equiv n$ and $u_n = m_n$:

$$m_n = \frac{1}{q_{n,n+1}} \prod_{0\leqslant k\leqslant n-1} \left[1 + \frac{q_{10}}{q_{k,k+1}}\right], \qquad n \geqslant 0.$$

Then

$$v_n \frac{u_n}{u_{n+1}} - v_{n+1} = \frac{n(q_{n+1,n+2} - q_{n,n+1} - q_{10})}{q_{n,n+1} + q_{10}} - 1.$$

Hence $\sum_n u_n < \infty$ if $\kappa' > 1$ $\big($resp. $\sum_n u_n = \infty$ once $\kappa' < 1\big)$. Clearly, $\sum_n m_n = \infty$ implies $\sum_n F_n^{(0)} = \infty$. Hence

$$d = \lim_{n\to\infty} \frac{d_n}{F_n^{(0)}} = \frac{1}{q_{01}}.$$

Furthermore,

$$\sup_{k\in E} \sum_{n=0}^{k} \big(F_n^{(0)} d - d_n\big) = F_0^{(0)} d = d < \infty.$$

This gives us the strong ergodicity by Proposition 5.2.

We mention that Proposition 7.2 (with $0 < c_i \equiv \lambda < q_{10}$) is also available for this model. □

Remark 8.3 For exponential ergodicity, the following sufficient condition

$$M := \sup_{n\geqslant 1} \left[\sum_{k=1}^{n-1} F_k^{(0)}\right] \left[\sum_{j=n}^{\infty} \frac{1}{q_{j,j+1} F_j^{(0)}}\right] < \infty, \tag{8.3}$$

introduced in [12], is sufficient for Example 8.1 but is not for Example 8.2.

Proof It is obvious that $M < \infty$ iff

$$\overline{\lim_{n\to\infty}} \left[\sum_{k=1}^{n-1} F_k^{(0)}\right] \left[\sum_{j=n}^{\infty} \frac{1}{q_{j,j+1} F_j^{(0)}}\right] < \infty. \tag{8.4}$$

For Example 8.1, because $q_{j,j+1}F_j^{(0)}$ is growing exponentially fast and so it is easy to check that $M < \infty$. For Example 8.2, it suffices to consider $q_{n,n+1} = b(n+1)$ for some $b > 0$. By Kummer test, one may show that

$$\sum_{j=n}^{\infty} \frac{1}{q_{j,j+1} F_j^{(0)}} = \infty$$

for suitable $b > 0$ and then $M = \infty$. □

Acknowledgements The authors acknowledge the support by NNSFC (No. 11131003), SRFDP (No. 20100003110005), the "985" project from the Ministry of Education in China, the Fundamental Research Funds for the Central Universities, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Anderson W J. Continuous-Time Markov Chains: An Applications-Oriented Approach. New York: Springer-Verlag, 1991
2. Brockwell P J. The extinction time of a general birth and death processes with catastrophes. J Appl Prob, 1986, 23: 851–858
3. Brockwell P J, Gani J, Resnick S I. Birth, immigration and catastrophe processes. Adv Appl Prob, 1982, 14: 709–731
4. Chen M F. From Markov Chains to Non-Equilibrium Particle Systems (2nd Edition). Singapore: World Scientific, 2004
5. Chen M F. Single birth processes. Chinese Ann Math, 1999, 20B: 77–82
6. Chen M F. Explicit criteria for several types of ergodicity. Chinese J Appl Prob Stat, 2001, 17(2): 1–8
7. Chen M F. Speed of stability for birth-death process. Front Math China, 2010, 5(3): 379–516
8. Chen M F, Zhang X. Isospectral operators. 2014, preprint
9. Hou Z T, Guo Q F. Homogeneous Denumerable Markov Processes (in Chinese), Beijing: Science Press, 1978; English translation, Beijing: Science Press and Springer, 1988
10. Mao Y H. Ergodic degrees for continuous-time Markov chains. Science in China Ser A Mathematics, 2004, 47(2): 161–174
11. Mao Y H. Eigentime identity for transient Markov chains. J Math Anal Appl, 2006, 315(2): 415–424
12. Mao Y H, Zhang Y H. Exponential ergodicity for single-birth processes. J Applied Probab, 2004, 41: 1022–1032
13. Reuter G E H. Competition Processes. In Fourth Berkeley Symposium on Math Stat and Prob, 1961, 2: 421–430
14. Wang L D, Zhang Y H. Criteria for zero-exit (-entrance) of single-birth (-death) Q-matrices. Acta Math. Sinica, 2014, to appear (in Chinese)
15. Yan S J, Chen M F. Multidimensional Q-processes. Chinese Ann Math, 1986, 7B: 90–110
16. Zhang J K. On the generalized birth and death processes (I). Acta Math Sci, 1984, 4: 241–259
17. Zhang Y H. Strong ergodicity for single-birth processes. J Appl Prob, 2001, 38(1): 270–277
18. Zhang Y H. Moments of the first hitting time for single birth processes. J Beijing Normal Univ, 2003, 39(4): 430–434 (in Chinese)
19. Zhang Y H. The hitting time and stationary distribution for single birth processes. J Beijing Normal Univ, 2004, 40(2): 157–161 (in Chinese)
20. Zhang Y H. Birth-death-catastrophe type single birth Q-matrices. J Beijing Normal Univ, 2011, 47(4): 347–350 (in Chinese)
21. Zhang Y H. Expressions on moments of hitting time for single birth process in infinite and finite space. J. Beijing Normal Univ, 2013, 49(5): 445–452 (in Chinese)

Appendix. Key formulas used in the proofs

(A) Solution to the Poisson equation $\Omega g = Qg + cg$:

$$g_n = g_0 + \sum_{0\leqslant k\leqslant n-1}\sum_{0\leqslant j\leqslant k}\frac{\widetilde{F}_k^{(j)}(f_j - c_j g_0)}{q_{j,j+1}}, \qquad n \geqslant 0.$$

(B) Three sequences.

(a) $\widetilde{F}$-sequence:

$$\widetilde{F}_i^{(i)} = 1, \qquad \widetilde{F}_n^{(i)} = \frac{1}{q_{n,n+1}}\sum_{k=i}^{n-1}\tilde{q}_n^{(k)}\widetilde{F}_k^{(i)}, \qquad n > i \geqslant 0, \tag{1.1}$$

where

$$\tilde{q}_n^{(k)} = q_n^{(k)} - c_n := \sum_{j=0}^{k} q_{nj} - c_n, \qquad 0 \leqslant k < n. \tag{1.2}$$

(b) $\widetilde{m}$-sequence:

$$\widetilde{m}_0 = \frac{1}{q_{01}}, \quad \widetilde{m}_n = \frac{1}{q_{n,n+1}}\left(1 + \sum_{k=0}^{n-1}\tilde{q}_n^{(k)}\widetilde{m}_k\right), \qquad n \geqslant 1. \tag{3.1}$$

(c) $\tilde{d}$-sequence:

$$\tilde{d}_0 = 0, \qquad \tilde{d}_n = \frac{1}{q_{n,n+1}}\left(1 + \sum_{k=0}^{n-1}\tilde{q}_n^{(k)}\tilde{d}_k\right), \qquad n \geqslant 1. \tag{5.1}$$

Representation of the three sequences:

$$\widetilde{F}_i^{(i)} = 1, \qquad \widetilde{F}_n^{(i)} = \sum_{k=i+1}^{n}\frac{\widetilde{F}_n^{(k)}\tilde{q}_k^{(i)}}{q_{k,k+1}}, \quad n \geqslant i+1; \tag{2.7}$$

$$\tilde{d}_n = \sum_{1\leqslant k\leqslant n}\frac{\widetilde{F}_n^{(k)}}{q_{k,k+1}}, \tag{5.2}$$

$$\widetilde{m}_n = \sum_{k=0}^{n}\frac{\widetilde{F}_n^{(k)}}{q_{k,k+1}}, \qquad n \geqslant 0. \tag{3.2}$$

Relation of the three sequences:

$$\widetilde{m}_n = \frac{1}{q_{01}}\widetilde{F}_n^{(0)} + \tilde{d}_n, \qquad n \geqslant 0. \tag{7.6}$$

Commun. Math. Stat. (2014) 2(1):17–32

Isospectral operators

Mu-Fa Chen (Beijing Normal University)

Xu Zhang (Beijing University of Technology)

March 6, 2014

Abstract

For a large class of integral operators or second order differential operators, their isospectral (or cospectral) operators are constructed explicitly in terms of h-transform (duality). This provides us a simple way to extend the known knowledge on the spectrum (or the estimation of the principal eigenvalue) from a smaller class of operators to a much larger one. In particular, an open problem about the positivity of the principal eigenvalue for birth–death processes is solved in the paper.

2000 *Mathematics Subject Classification*: 58J53; 37A30.
Key words and phases. Isospectral; harmonic function; integral operator, differential operator.

1 Introduction

Let us consider the elliptic operators

$$L = \sum_{i,j} a_{ij}(x)\partial_{ij}^2 + \sum_i b_i(x)\partial_i + c(x),$$
$$\widetilde{L} = \sum_{i,j} \tilde{a}_{ij}(x)\partial_{ij}^2 + \sum_i \tilde{b}_i(x)\partial_i$$

on $L^2(\mu)$ and $L^2(\tilde{\mu})$ (real) respectively, where $\tilde{\mu} = h^2\mu$ for a given measure μ and some $h \neq 0$. Their main difference is that $c(x) \not\equiv 0$. We are interested in when the operators L and $\widetilde{L}$ are L^2-isospectral in the following sense

$$(Lf, f)_\mu = \left(\widetilde{L}\tilde{f}, \tilde{f}\right)_{\tilde{\mu}}, \qquad \text{for every } \tilde{f} := f/h,\ f \in \mathscr{D}(L).$$

Here is one of our typical results in the note (cf. Theorems 3.1 and 3.6 in Section 3).

Theorem 1.1 (1) Given L on $L^2(\mu)$ having domain $\mathscr{D}(L)$, let $h \neq 0$, μ-a.e. be L-harmonic: $Lh = 0$, μ-a.e., then L is L^2-isospectral to $\widetilde{L}$:

$$\widetilde{L} = L_0 + 2h^{-1}\langle a\nabla h, \nabla\rangle, \qquad \mathscr{D}\big(\widetilde{L}\big) = \{f : fh \in \mathscr{D}(L)\}.$$

where $L_0 = L - c$.

(2) Given $\widetilde{L}$ on $L^2(\tilde{\mu})$ having domain $\mathscr{D}\big(\widetilde{L}\big)$, then for each $h \neq 0$, μ-a.e., $\widetilde{L}$ is L^2-isospectral to L:

$$L = \widetilde{L} - \frac{2}{h}\big\langle \tilde{a}\nabla h, \nabla\big\rangle + \left[\frac{2}{h^2}\big\langle \tilde{a}\nabla h, \nabla h\big\rangle - \frac{1}{h}\widetilde{L}h\right],$$
$$\mathscr{D}(L) = \big\{f : f/h \in \mathscr{D}\big(\widetilde{L}\big)\big\},$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product.

As a typical application of Theorem 1.1, we obtain the next result. To state it, we need to explain the meaning of eigenvalue in different sense. We say that λ is an eigenvalue of L in the ordinary sense if $Lg = \lambda g$ for some $g \neq 0$. It is called a L^2-eigenvalue if additionally, $g \in L^2(\mu)$.

Corollary 1.2 For each $h \in \mathscr{C}^2(\mathbb{R})$, $h \neq 0$, a.e., the operator

$$L^h = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} - \left(x + \frac{h'}{h}\right)\frac{\mathrm{d}}{\mathrm{d}x} + \left[\left(\frac{h'}{h}\right)^2 + x\frac{h'}{h} - \frac{h''}{2h}\right]$$

has L^2-eigenvalues $\lambda_n\big(L^h\big) = -n$ with eigenfunctions

$$g_n(x) = (-1)^n h(x) e^{x^2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\big(e^{-x^2}\big), \qquad n \geqslant 0,$$

respectively. A particular class of L^h is the following

$$L^b = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} - b(x)\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2}\big[b(x)^2 - b'(x) - x^2 + 1\big], \qquad b \in \mathscr{C}^1(\mathbb{R}).$$

Proof. Noting that the Ornstein-Uhlenbeck operator

$$\widetilde{L} = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} - x\frac{\mathrm{d}}{\mathrm{d}x}, \qquad \mathscr{D}\big(\widetilde{L}\big) \supset \mathscr{C}_0^\infty(\mathbb{R})$$

has ordinary eigenvalues $\lambda_n\big(\widetilde{L}\big) = -n$ with eigenfunctions

$$g_n(x) = (-1)^n e^{x^2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\big(e^{-x^2}\big), \qquad n \geqslant 0,$$

respectively (cf. [3; Example 5.1]). Clearly, the polynomial function $g_n \in L^2(\tilde{\mu})$ for every $n \geqslant 0$, where $\tilde{\mu}(\mathrm{d}x) = \exp(-x^2)\mathrm{d}x$. Hence, the eigenvalues

are all L^2-ones. Now, the first assertion follows from part (2) of Theorem 1.1. The last assertion then follows by setting $h = \exp\psi$ with $\psi' = b - x$:

$$\left(\frac{h'}{h}\right)^2 + x\frac{h'}{h} - \frac{h''}{2h} = \psi'^2 + x\psi' - \frac{1}{2}\big(\psi'' + \psi'^2\big) = \psi'\left(x + \frac{1}{2}\psi'\right) - \frac{1}{2}\psi''. \qquad \square$$

Corollary 1.2 says that a large class of operators are all isospectral to the rather simple Ornstein-Uhlenbeck operator. This indicates the value of the study on isospectral operators. It should be pointed out that the technique is still valuable even if you know only some estimates of the principal eigenvalue of $\widetilde{L}$ but have no knowledge on the other part of the spectrum of $\widetilde{L}$, since our knowledge on the principal eigenvalue of L is still rather limited.

Actually, Theorem 1.1 comes from a very simple observation. For completeness, here we write its complex version, even though we will use only its real version later on.

Lemma 1.3 Let $(E, \mathscr{E}, \mu)$ be a measure space and let h be Lebesgue measurable: $E \to \mathbb{C}$, $h \neq 0$, μ-a.s. Then

(1) $\tilde{f} := \mathbb{1}_{[h\neq 0]} f/h$ is an isometry from $L^2(E,\mu)$ to $L^2(E,\tilde{\mu})$ (complex), where $\tilde{\mu} = |h|^2\mu$.

(2) Let L be an operator on $L^2(E,\mu)$ with domain $\mathscr{D}(L)$. Define an operator $\widetilde{L}$ as follows:

$$\widetilde{L}\tilde{f} = \mathbb{1}_{[h\neq 0]}\frac{1}{h}L\big(\tilde{f}h\big), \qquad \mathscr{D}\big(\widetilde{L}\big) = \big\{\tilde{f} \in \mathscr{E} : \tilde{f}h \in \mathscr{D}(L)\big\}. \tag{1}$$

Then the operators $(L, \mathscr{D}(L))$ on $L^2(E,\mu)$ and $\big(\widetilde{L}, \mathscr{D}\big(\widetilde{L}\big)\big)$ on $L^2\big(E,\tilde{\mu}\big)$ are isospectral (say L and $\widetilde{L}$ are L^2-isospectral, for short) (in the following sense):

$$(Lf, f)_\mu = \big(\widetilde{L}\tilde{f}, \tilde{f}\big)_{\tilde{\mu}}, \qquad f \in \mathscr{D}(L).$$

(3) If additionally, $h \in \mathscr{D}(L)$, then $\widetilde{L}\mathbb{1} = 0$, $\tilde{\mu}$-a.e. iff h is L-harmonic: $Lh = 0$, μ-a.s.*

Proof. Recall the inner product in a complex L^2-space:

$$(f, g)_\mu = \int_E f\bar{g}\mathrm{d}\mu.$$

The first assertion is obvious:

$$\int_E |f|^2\mathrm{d}\mu = \int_{E\,[h\neq 0]} \big|\tilde{f}\big|^2|h|^2\mathrm{d}\mu = \int_E \big|\tilde{f}\big|^2\mathrm{d}\tilde{\mu}.$$

*See also Remark 1.4 below.

By definition, for $\tilde{f} \in \mathscr{D}(\widetilde{L})$, we have $\tilde{f}h \in \mathscr{D}(L) \subset L^2(E,\mu)$. Then we have not only $\tilde{f} \in L^2(E,\tilde{\mu})$ but also $L(\tilde{f}h) \in L^2(E,\mu)$. This means that $\widetilde{L}\tilde{f} \in L^2(E,\tilde{\mu})$. Hence, as an operator on $L^2(E,\tilde{\mu})$, $\widetilde{L}$ is well defined. Furthermore, we have

$$(Lf,f)_\mu = \big(L(\tilde{f}h),\tilde{f}h\big)_\mu = \int_E \overline{\tilde{f}h}\, L(\tilde{f}h)\mathrm{d}\mu = \int_E \bar{\tilde{f}}(\bar{h}h)\frac{1}{h}L(\tilde{f}h)\mathrm{d}\mu = \big(\widetilde{L}\tilde{f},\tilde{f}\big)_{\tilde{\mu}}.$$

We have thus proved the second assertion. Clearly, if $h \in \mathscr{D}(L)$, then $\mathbb{1}h = h \in L^2(E,\mu)$ and hence $\mathbb{1} \in L^2(E,\tilde{\mu})$ which implies that $\tilde{\mu}(E) < \infty$. Furthermore, $\mathbb{1} \in \mathscr{D}(\widetilde{L})$ by definition of $\mathscr{D}(\widetilde{L})$. Therefore, the last assertion follows by definition of $\widetilde{L}$. $\square$

For non-symmetric operators, their spectrum can be complex. Hence, it is natural to use the complex L^2-theory. However, in this note, we use the real L^2-spaces only. Thus, the L^2-isospectral (real) here means the spectrum of their symmetrized operators. The last assertion of the lemma suggests us, as we will do often later, to choose h as an L-harmonic function in a weak (pointwise) sense (in other words, h is in a weak domain of L) without assuming $h \in \mathscr{D}(L)$. Then $\widetilde{L}\mathbb{1} = 0$ is meaningful in the weak sense. In this way, we can construct the operator $\widetilde{L}$ explicitly, which is the main goal of this note. Furthermore, part (3) of the lemma has the following extension.

Remark 1.4 For fixed $B \in \mathscr{E}$, $\widetilde{L}\mathbb{1} = 0$, $\tilde{\mu}$-a.e. on B iff $Lh = 0$, μ-a.s. on B.

We will illustrate later an application of this assertion in the context of Markov chains. Clearly, the L-harmonic function is an eigenfunction corresponding to the eigenvalue $\lambda = 0$. However, $\lambda = 0$ is not necessarily an eigenvalue in the L^2-sense unless $h \in L^2(E,\mu)$.

One may write $\widetilde{L} = h^{-1}L(h\,\bullet)$ (μ-a.e.) for short. Because of this, $\widetilde{L}$ is called a h-transform of L. Alternatively, define an operator H:

$$Hf = hf, \qquad \mathscr{D}(H) = \{f \in L^2(E,\mu) : hf \in \mathscr{D}(L)\}.$$

Then, we indeed have $\widetilde{L} = H^{-1}LH$. In view of this, L and $\widetilde{L}$ are similar and so are L^2-isospectral. More generally (without assuming the invertibility of H),

$$H\widetilde{L} = LH.$$

Because of this, L and $\widetilde{L}$ are called dual with respect to H. Therefore, the h-transform is indeed a special duality. For a different dual, refer to [2; §5 and §10]. Note that In the latter case, we were interested in the principal eigenvalue only, but the transform used there is still isospectral. The reason is that the isospectral transform is easier to handle even though it looks rather strong. We remark that when E has boundary ∂E, one may deduce a boundary condition for $\widetilde{L}$ from that of L, based on the transform $\tilde{f} = \mathbb{1}_{[h\neq 0]}f/h$.

Having figured out the dual operators, in the study of their spectrum for Markov processes, it is more convenient in practice to use their extension to the Dirichlet forms, especially for the operator $\big(\widetilde{L}, \mathscr{D}\big(\widetilde{L}\big)\big)$. Generally speaking, Lemma 1.3 says that for a given Dirichlet form $(D, \mathscr{D}(D))$ on $L^2(\mu)$, its dual form $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$ on $L^2(\tilde{\mu})$ is given by

$$\widetilde{D}\big(\tilde{f}\big) = D\big(\tilde{f}h, \tilde{f}h\big), \qquad \mathscr{D}\big(\widetilde{D}\big) = \big\{\tilde{f} \in \mathscr{E} : \tilde{f}h \in \mathscr{D}(D)\big\}.$$

Certainly, one may go to the inverse way, defining $(D, \mathscr{D}(D))$ in terms of $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$. In particular, for the O.-U. operator used in the proof of Corollary 1.2, corresponding to $\big(\widetilde{L}, \mathscr{D}\big(\widetilde{L}\big)\big)$, the Dirichlet form $\big(\widetilde{D}(f), \mathscr{D}\big(\widetilde{D}\big)\big)$ is

$$\begin{aligned}
&\widetilde{D}(f) = \int_{\mathbb{R}} f'^2 e^{-x^2}\,\mathrm{d}x,\\
&\mathscr{D}\big(\widetilde{D}\big) = \big\{f \in L^2(\tilde{\mu}) : \widetilde{D}(f) < \infty\big\} = \bigg\{f : \int_{\mathbb{R}} \big[f^2 + f'^2\big] e^{-x^2}\,\mathrm{d}x < \infty\bigg\}.
\end{aligned}$$

In the case that the potential term c^h (the last term) in L^h is non-positive, then L^h corresponds to the operator of a diffusion having killing rate $-c^h$, to which we certainly have a Dirichlet form $\big(D^h, \mathscr{D}\big(D^h\big)\big)$ on $L^2(\mu^h)$:

$$\begin{aligned}
&D^h(f) = \int_{\mathbb{R}} \big[f'^2(x) - c^h(x) f^2(x)\big] e^{-x^2}\,\frac{\mathrm{d}x}{h(x)^2},\\
&\mathscr{D}\big(D^h\big) = \bigg\{f : \int_{\mathbb{R}} \big[f^2 + (f'h - fh')^2\big] e^{-x^2}\,\mathrm{d}x < \infty\bigg\},\\
&c^h(x) = \bigg[\bigg(\frac{h'}{h}\bigg)^2 + x\frac{h'}{h} - \frac{h''}{2h}\bigg](x), \qquad \mu^h(\mathrm{d}x) = e^{-x^2}\,\frac{\mathrm{d}x}{h(x)^2}.
\end{aligned}$$

Here $\mathscr{D}\big(D^h\big)$ is deduced from $\mathscr{D}\big(\widetilde{D}\big)$, based on Lemma 1.3. For general $c^h(x) \in \mathbb{R}$, this symmetric form may not be a Dirichlet one even though it does have nonnegative spectrum in view of our isospectral property. Actually, Lemma 1.3 is meaningful in a very general setup rather than Markov processes.

The h-transform, or the Doob's h-transform is a well-known topic in probability/potential theory. Here we mention only two related papers [9, 10] where the tool is used to study the principal eigenvalue. In [9], the following model

$$\begin{aligned}
&L = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x} a \frac{\mathrm{d}}{\mathrm{d}x} - \frac{1}{2}\bigg(\frac{b^2}{a} + b'\bigg),\\
&\widetilde{L} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x} a \frac{\mathrm{d}}{\mathrm{d}x} + b\frac{\mathrm{d}}{\mathrm{d}x},\\
&h(x) = \exp\bigg[\int_0^x \frac{b}{a}(y)\mathrm{d}y\bigg]
\end{aligned}$$

is carefully handled and applied to multi-dimensional diffusion operators. In [10], a class of symmetric Markov processes having killings are studied and some upper and lower estimates for the first eigenvalue are presented.

The remainder of this note is organized as follows. In the next two sections, we apply Lemma 1.3, respectively, to two special classes of operators: either integral operators for Markov pure jump processes or the operators for diffusions.

2 Integral operators

Theorem 2.1 Let $(q(x), q(x,\mathrm{d}y))$ be a totally stable and conservative q-pair on $(E,\mathscr{E},\mu)$ (cf. [1; Definition 1.9]). For a given function $c\in\mathscr{E}$ with $c\leqslant q$, define an operator Ω

$$\Omega f(x)=\int_E q(x,\mathrm{d}y)\big[f(y)-f(x)\big]+c(x)f(x),\qquad x\in E$$

with domain $\mathscr{D}(\Omega)\subset L^2(E,\mu)$. Next, let $h\,(>0,\mu\text{-a.e.})$ be Ω-harmonic (if exists): $\Omega\, h=0$, μ-a.e. on E. Define a new totally stable and conservative q-pair $\big(\tilde{q}(x),\tilde{q}(x,\mathrm{d}y)\big)$ as follows.

$$\begin{aligned}\tilde{q}(x,A)&=\mathbb{1}_{[h(x)\neq 0]}\frac{1}{h(x)}\int_A q(x,\mathrm{d}y)h(y),\qquad A\in\mathscr{E},\\ \tilde{q}(x)&=\tilde{q}(x,E),\quad \mu\text{-a.e. } x\in E.\end{aligned}$$

Set

$$\begin{aligned}\widetilde{\Omega}f(x)&=\int_E\tilde{q}(x,\mathrm{d}y)\big[f(y)-f(x)\big],\qquad \mu\text{-a.e. } x\in E,\\ \mathscr{D}\big(\widetilde{\Omega}\big)&=\big\{\tilde{f}\in\mathscr{E}:\tilde{f}h\in\mathscr{D}(\Omega)\big\}.\end{aligned}$$

Then Ω and $\widetilde{\Omega}$ are L^2-isospectral.

Proof. Noting that $h\,(>0,\mu\text{-a.e.})$ is Ω-harmonic by assumption, we have

$$\big[q(x)-c(x)\big]h(x)=\int_E q(x,\mathrm{d}y)h(y)\geqslant 0.$$

Hence h is $q(x,\cdot)$-integrable for a.e.-$x\in E$ and moreover $q\geqslant c$. Therefore, the new q-pair $\big(\tilde{q}(x),\tilde{q}(x,\mathrm{d}y)\big)$ is totally stable. It is clearly conservative. By definition of $\widetilde{\Omega}$, we have on the set $[h>0]$,

$$\begin{aligned}\widetilde{\Omega}(f)(x)&=\int_E\tilde{q}(x,\mathrm{d}y)\big[f(y)-f(x)\big]\\ &=\frac{1}{h(x)}\int_E q(x,\mathrm{d}y)\big\{\big[(fh)(y)-(fh)(x)\big]+f(x)\big[h(x)-h(y)\big]\big\}\Big]\\ &=\frac{1}{h(x)}\bigg[\int_E q(x,\mathrm{d}y)\big[(fh)(y)-(fh)(x)\big]-f(x)\int_E q(x,\mathrm{d}y)\big[h(y)-h(x)\big]\\ &=\frac{1}{h(x)}\big[\Omega(fh)(x)-c(fh)(x)-f(x)[\Omega h(x)-(ch)(x)]\big]\\ &=\frac{1}{h(x)}\big[\Omega(fh)(x)-f(x)\Omega h(x)\big].\end{aligned}$$

Now, by harmonic property of h, the right-hand side is equal to

$$\frac{1}{h(x)}\Omega(fh)(x) \qquad \text{on } [h>0].$$

The assertion then follows from Lemma 1.3. □

We mention that the positive condition of h used in the theorem is to keep $\big(\tilde q(x),\tilde q(x,\mathrm{d}y)\big)$ to be a q-pair. This is certainly not necessary in a general context: considering general integral kernel instead of the nonnegative one.

The inverse of the last theorem goes as follows.

Theorem 2.2 Given a totally stable and conservative q-pair $\big(\tilde q(x),\tilde q(x,\mathrm{d}y)\big)$ and a positive $\mathscr{E}$-measurable function h such that h^{-1} is $\tilde q(x,\cdot)$-integrable for each $x\in E$, the operator $\big(\widetilde\Omega,\mathscr{D}(\widetilde\Omega)\big)$ on $L^2(E,\tilde\mu)$ corresponding to the q-pair $\big(\tilde q(x),\tilde q(x,\mathrm{d}y)\big)$ is L^2-isospectral to the following operator Ω on $L^2(E,\mu)$ $\big(\mu:=h^{-2}\tilde\mu\big)$:

$$\Omega f(x)=\int_E q(x,\mathrm{d}y)[f(y)-f(x)]+c(x)f(x),$$
$$\mathscr{D}(\Omega)=\big\{f\in\mathscr{E}: f/h\in\mathscr{D}(\widetilde\Omega)\big\}\subset L^2(E,\mu),$$

where

$$q(x,\mathrm{d}y)=h(x)\frac{\tilde q(x,\mathrm{d}y)}{h(y)},$$
$$c(x)=\int_E\tilde q(x,\mathrm{d}y)\left[\frac{h(x)}{h(y)}-1\right],\qquad x\in E.$$

Proof. It is simply a use of the duality $\Omega=H\widetilde\Omega H^{-1}$, noting the property that $\Omega h=0$ is now automatic since $\widetilde\Omega 1=0$. The remainder of the proof is mainly a careful computation. □

It is the place to discuss the existence of a positive Ω-harmonic function. Let $c(x)<q(x)$, $x\in E$. Choose and fix a reference point $\theta\in E$. By [1; Theorem 2.2], there exists uniquely the minimal solution $(h^*(x):x\in E)$ with $h^*(\theta)=1$ to the following nonnegative equation

$$h(x)=\int_{E\setminus\{\theta\}}\frac{q(x,\mathrm{d}y)}{q(x)-c(x)}h(y)+\frac{q(x,\{\theta\})}{q(x)-c(x)},\qquad x\neq\theta. \tag{2}$$

Moreover, the solution can be obtained in the following way: let

$$h^{(1)}(x)=\frac{q(x,\{\theta\})}{q(x)-c(x)},\qquad x\neq\theta,$$
$$h^{(n+1)}(x)=\int_{E\setminus\{\theta\}}\frac{q(x,\mathrm{d}y)}{q(x)-c(x)}h^{(n)}(y)+\frac{q(x,\{\theta\})}{q(x)-c(x)},\qquad x\neq\theta,\ n\geqslant 1.$$

Then for each $x\neq\theta$, $h^{(n)}(x)\uparrow h^*(x)\in[0,\infty]$ as $n\to\infty$.

Proposition 2.3 Let $c(x) < q(x)$ for every $x \in E$ and assume that $q(x, \{\theta\}) > 0$ for some $x \neq \theta$. Then the equation $\Omega h = 0$ has a non-trivial (finite) solution iff the minimal solution $(h^*(x) : x \in E)$ to (2) is finite. Equivalently, there is a finite f satisfying the inequality

$$f(x) \geqslant \int_{E\setminus\{\theta\}} \frac{q(x, \mathrm{d}y)}{q(x) - c(x)} f(y) + \frac{q(x, \{\theta\})}{q(x) - c(x)}, \qquad x \neq \theta.$$

Then we actually have $f(x) \geqslant h^*(x)$ for every $x \in E$.†

Proof. For a given finite non-trivial Ω-harmonic function h, choosing $h(\theta) = 1$, one may write down immediately equation (2).

Conversely, a finite solution h^* to (2) is clearly a Ω-harmonic function. From the construction given above, it is also clear that $h^*(x) > 0$ once $q(x, \{\theta\}) > 0$. The last assertion of the proposition is essentially a comparison theorem [1; Theorem 2.6]. □

It is clear from the proof above, to obtain a positive harmonic h, some irreducible condition is necessary. Noting that it is often practical to find an explicit comparison function f, and $h^{(n)}$ for each n is already explicit, we have explicit estimates of h^* which may not be easy to obtain explicitly.

Before moving further, we discuss an alternative way to describe the Ω-harmonic function. Suppose that $\sup_x c(x) < \infty$. Then by a shift if necessary, we may and will assume for a moment that $\sup_x c(x) \leqslant 0$. Define

$$\begin{aligned} z^{(0)}(x) &= 1, \qquad x \in E, \\ z^{(n+1)}(x) &= \int_E \frac{q(x, \mathrm{d}y)}{q(x) - c(x)} z^{(n)}(y), \qquad x \in E,\ n \geqslant 1. \end{aligned}$$

Then $z^{(n)}(x) \downarrow \bar{z}(x)$ as $n \to \infty$ for each $x \in E$. This is an analog of the maximal exit solution in the study of q-processes, cf. [1; Lemma 2.39]. The proof for the conclusion is easy, simply use the property

$$\frac{q(x, E)}{q(x) - c(x)} \leqslant 1, \qquad x \in E.$$

Remark 2.4 Let $\sup_x c(x) \leqslant 0$. Then a bounded Ω-harmonic function is non-zero iff so is the maximal solution $\bar{z}$ constructed above.

To apply the previous results, Theorem 2.1 for instance, to finite state spaces, say $E = \{0, 1, \ldots, N\}$ for some $N \geqslant 3$, one meets a problem about the existence of positive Ω-harmonic h. For which, there $N + 1$ homogeneous equations with $N + 1$ variables $h_0, h_1, \ldots, h_N$. Because of the homogeneous

†Correction. Here the uniqueness of the solution h to the equation $\Omega h = 0$ with $h(\theta) = 1$ up to a positive constant is needed. Otherwise, $(h^*(x) : x \in E)$ is only a lower bound of h.

property in h, one may assume that $h_0 = 1$ once a non-trivial solution h exists with $h_0 \neq 0$ for instance. Thus, we have only N free variables in $N+1$ equations. Then a finite non-trivial solution often does not exist (or equivalently, the minimal solution given in Proposition 2.3 may be infinite). To overcome this difficulty, one has to decrease the number of equations. This is the reason we will adopt a local harmonic condition below. Then, one needs non-trivial $\tilde{c}_i$ in the corresponding operator $\widetilde{\Omega}$.

Theorem 2.5 Let $E = \{0, 1, \dots, N\}$ for some $N \geqslant 3$ and let $Q = (q_{ij})$ be a conservative Q-matrix on E. For given $(c_i : i = 0, 1, \dots, N)$ with $c_i \leqslant q_i := -q_{ii}$ for $i = 0, 1, \dots, N-1$, set $\Omega = Q + \operatorname{diag}(c_i)$. Next, let $h > 0$ be Ω-harmonic on $\{0, 1, \dots, N-1\}$, i.e.,

$$\Omega h = 0 \qquad \text{on } \{0, 1, \dots, N-1\}.$$

Define $\tilde{q}_{ij}$ $(i, j \in E)$ as in Theorem 2.1:

$$\tilde{q}_{ij} = h_i^{-1} q_{ij} h_j, \qquad i, j \in E,\ i \neq j; \qquad \tilde{q}_{ii} = -\sum_{k \neq i} \tilde{q}_{ik},\ i \in E.$$

Next, define $\tilde{c}_i = 0$ on $\{0, 1, \dots, N-1\}$ and

$$\tilde{c}_N = c_N + \sum_{j \leqslant N-1} q_{Nj} \left(\frac{h_j}{h_N} - 1 \right).$$

Denote by $\widetilde{\Omega}$ the operator corresponding to the matrix $\big(\tilde{q}_{ij}\big) + \operatorname{diag}(\tilde{c}_i)$. Then Ω and $\widetilde{\Omega}$ are L^2-isospectral. Besides, we have $\tilde{q}_{ii} + \tilde{c}_i = q_{ii} + c_i$ for each $i \in E$.

Proof. Following the proof of Theorem 2.1, restricted to $\{0, 1, \dots, N-1\}$, we see that

$$\widetilde{\Omega} \tilde{f}(i) = \frac{1}{h_i} \Omega\big(\tilde{f} h\big)(i) \qquad \text{on } \{0, 1, \dots, N-1\}.$$

We now show that this equality also holds for $i = N$.

$$\begin{aligned}
\widetilde{\Omega} f(N) &= \sum_{j \leqslant N} \tilde{q}_{Nj} (f_j - f_N) + \tilde{c}_N f_N \\
&= \frac{1}{h_N} \sum_{j \leqslant N} q_{Nj} \big[(fh)_j - (fh)_N \big] - \frac{f_N}{h_N} \sum_{j \leqslant N} q_{Nj} (h_j - h_N) + \tilde{c}_N f_N \\
&= \frac{1}{h_N} \Omega(fh)(N) - \frac{1}{h_N} c_N h_N f_N - \frac{f_N}{h_N} \sum_{j \leqslant N} q_{Nj} (h_j - h_N) + \tilde{c}_N f_N \\
&= \frac{1}{h_N} \Omega(fh)(N).
\end{aligned}$$

From Remark 1.4, it follows that $\tilde{c}_i = 0$ on $\{0, 1, \dots, N-1\}$. The required main assertion now follows from Lemma 1.3. The last assertion is then easy

to check using the harmonic property on $\{0,1,\dots,N-1\}$ and the expression of $\tilde{c}_N$:

$$\tilde{q}_{ii}=-\frac{1}{h_i}\sum_{j\neq i}q_{ij}h_j=-\frac{1}{h_i}\big[\Omega h(i)-(q_{ii}+c_i)h_i\big]=q_{ii}+c_i,\quad i\leqslant N-1;$$

$$\tilde{c}_N=c_N+\sum_{j\leqslant N-1}q_{Nj}\left[\frac{h_j}{h_N}-1\right]=c_N+\sum_{j\leqslant N-1}\tilde{q}_{Nj}-\sum_{j\leqslant N-1}q_{Nj}=c_N-\tilde{q}_{NN}+q_{NN}.$$

□

A typical application of Theorem 2.1 to the single birth processes is presented in [12]. In this case, the Ω-harmonic function has a very simple expression (cf. [5; Theorem 1.1]). In particular, for the killing case, the function is not only positive but also non-decreasing. It is interesting to note that for single birth processes, the function h-dual is again the same type, but the measure μ-dual

$$\bar{q}_{ij}=\frac{\mu_j q_{ji}}{\mu_i},\qquad i,j\in E$$

maps the single birth type to the single death type. Next, for birth–death processes with birth and death rates b_i and a_i, respectively, and with killing rates $-c_i\geqslant 0$, we have

$$\tilde{a}_i=a_i\frac{h_{i-1}}{h_i}\,(\leqslant a_i),\ i\geqslant 1, h_0=1,\qquad \tilde{b}_i=b_i\frac{h_{i+1}}{h_i}\,(\geqslant b_i),\ i\geqslant 0.$$

Then

$$\tilde{\mu}_i=\frac{\tilde{b}_0\dots\tilde{b}_{i-1}}{\tilde{a}_1\dots\tilde{a}_i}=\frac{b_0\dots b_{i-1}}{a_1\dots a_i}h_i^2=h_i^2\mu_i,\qquad \hat{\tilde{\nu}}_i=\frac{1}{\tilde{\mu}_i\tilde{b}_i}=\frac{1}{h_ih_{i+1}}\hat{\nu}_i,\qquad i\geqslant 0.$$

For finite state space, we have

$$\tilde{c}_N=c_N+a_N\left(\frac{h_{N-1}}{h_N}-1\right).$$

Clearly, $\tilde{c}_N\leqslant 0$ since so does c_N. However, the story is still meaningful for general $c_i\in\mathbb{R}$ satisfying $c_i\leqslant a_i+b_i$ for all $i\geqslant 0$.

To conclude this section, we answer an open question for birth–death processes with state space $\{0,1,2,\dots\}$. For this, we need some notation. Given birth rates $b_i>0\,(i\geqslant 0)$, death rates $a_i>0\,(i\geqslant 1)$ and killing rates $-c_i\geqslant 0\,(i\geqslant 0)$, define

$$\tilde{q}_n^{(k)}=\begin{cases}-c_n, & 0\leqslant k\leqslant n-2\\ a_n-c_n, & k=n-1,\end{cases}$$

$$\widetilde{F}_i^{(i)}=1,\ \ \widetilde{F}_n^{(i)}=\frac{1}{b_n}\sum_{k=i}^{n-1}\tilde{q}_n^{(k)}\widetilde{F}_k^{(i)},\qquad n>i\geqslant 0,$$

$$h_n = 1 - \sum_{0\leqslant k\leqslant n-1} \sum_{0\leqslant j\leqslant k} \widetilde{F}_k^{(j)} \frac{c_j}{b_j}, \qquad n \geqslant 0.$$

Next, define the principal eigenvalue λ_0 as follows.

$$\lambda_0 = \inf \left\{ \sum_{k\geqslant 0} \mu_k \big[b_k (f_{k+1} - f_k)^2 - c_k f_k^2 \big] : \sum_{k\geqslant 0} \mu_k f_k^2 = 1,\ f \text{ has finite support} \right\}.$$

Here is a solution to the Open Problem 9.13 in [2].

Theorem 2.6 For birth–death processes as above, we have $\tilde{\delta} \leqslant \lambda_0^{-1} \leqslant 4\tilde{\delta}$, where

$$\tilde{\delta} = \sup_{n\geqslant 0} \sum_{j=0}^{n} \tilde{\mu}_j \sum_{k\geqslant n} \hat{\tilde{\nu}}_k = \sup_{n\geqslant 0} \sum_{j=0}^{n} \mu_j h_j^2 \sum_{k\geqslant n} \frac{1}{h_k h_{k+1} \mu_k b_k}.$$

In particular, $\lambda_0 > 0$ iff $\tilde{\delta} < \infty$.

Proof. The harmonic function h we need for applying Theorem 2.1 is given by [5; Theorem 1.1]. Then the result follows by applying [2; Theorem 3.1] to the process with rates $(\tilde{b}_i, \tilde{a}_i)$ and using $\tilde{\mu}_i$ and $\hat{\tilde{\nu}}_k$ just computed above. □

3 Differential operators

We now turn to study the second-order differential operators.

Theorem 3.1 Consider the elliptic operator

$$L = \sum_{i,j} a_{ij}(x) \partial_{ij}^2 + \sum_i b_i(x) \partial_i + c(x)$$

with a domain $\mathscr{D}(L)$, and let $h \neq 0$ a.e. (with respect to Lebesgue measure) be L-harmonic. Here

$$\partial_i = \mathrm{d}/\mathrm{d}x_i, \qquad \partial_{ij}^2 = \partial_i \partial_j.$$

Define

$$\widetilde{L} = \sum_{i,j} \tilde{a}_{ij}(x) \partial_{ij}^2 + \sum_i \tilde{b}_i(x) \partial_i,$$

with domain $\mathscr{D}\big(\widetilde{L}\big)$ defined in Lemma 1.3, where

$$\tilde{a}_{ij}(x) = a_{ij}(x), \qquad \tilde{b}_i(x) = b_i(x) + \frac{2}{h(x)} \sum_j a_{ij}(x) \partial_j h(x)$$

for all i, j and a.e.-x. Then L and $\widetilde{L}$ are L^2-isospectral.

Proof. Noting that by the symmetry of the matrix (a_{ij}), we have

$$\begin{aligned} L(fh) &= \sum_{i,j} a_{ij}\partial_{ij}^2(fh) + \sum_i b_i\partial_i(fh) + cfh \\ &= \sum_{i,j} a_{ij}\big[\big(\partial_{ij}^2 f\big)h + 2\partial_i f\partial_j h \\ &\quad + f\big(\partial_{ij}^2 h\big)\big] + \sum_i b_i\big[\big(\partial_i f\big)h + f\partial_i h\big] + f(ch) \\ &= hLf + fLh - cfh + 2\sum_{i,j} a_{ij}\partial_j h\partial_i f \quad \text{a.e.} \end{aligned}$$

Because h is L-harmonic, we obtain

$$\frac{1}{h}L(fh) = (Lf - cf) + \frac{2}{h}\sum_i \bigg(\sum_j a_{ij}\partial_j h\bigg)\partial_i f, \quad \text{a.e.}$$

From which, one reads out the coefficients $\tilde{a}_{ij}(x)$ and $\tilde{b}_i(x)$ of $\widetilde{L}$. $\square$

For short, if we set $L_0 = L - c$, then we have

$$\begin{aligned} \widetilde{L} &= L_0 + \frac{2}{h}\langle a\nabla h, \nabla\rangle \\ &= L_0 + 2\langle a\nabla \log h, \nabla\rangle \quad \text{if } h > 0. \end{aligned}$$

Remark 3.2 In one-dimensional case, denoting by $(a(x), b(x), c(x))$ the coefficients of L, we can represent L as

$$L = \frac{\mathrm{d}}{\mathrm{d}\mu}\frac{\mathrm{d}}{\mathrm{d}\hat{\nu}} + c(x),$$

where

$$\mathrm{d}\mu(x) = \frac{e^{C(x)}}{a(x)}\mathrm{d}x, \quad \mathrm{d}\hat{\nu}(x) = e^{-C(x)}\mathrm{d}x, \quad C(x) = \int_\theta^x \frac{b}{a}(z)\mathrm{d}z,$$

and θ is a reference point. Then the (dual) operator $\widetilde{L}$ can be written as

$$\widetilde{L} = \frac{\mathrm{d}}{\mathrm{d}\tilde{\mu}}\frac{\mathrm{d}}{\mathrm{d}\hat{\tilde{\nu}}} = \frac{\mathrm{d}}{\mathrm{d}(h^2\mu)}\frac{\mathrm{d}}{\mathrm{d}\big(h^{-2}\hat{\nu}\big)}.$$

Here are simple examples of L-harmonic functions.

Example 3.3 Let $E = \mathbb{R}$ or $(0, \infty)$.

(1) The function $h(x) = x$ is L-harmonic (a.e.) on E for

$$L = \gamma(x)\big(\partial_{xx}^2 + V(x)\partial_x - V(x)/x\big),$$

where the functions V and γ are arbitrary.

(2) The function $h(x) = x^2$ is L-harmonic (a.e.) on E for

$$L = \gamma(x)\big(x\partial_{xx}^2 + \partial_x - 4/x\big),$$

where the function γ is again arbitrary.

In dimension one, the existence and uniqueness of L-harmonic function, as well as an approximating (constructing) procedure, can be found from [11; Theorems 1.2.1 and 2.2.1]. To see the positivity of h in general dimensions, suppose that L is self-adjoint and $\sup_x c(x) \leqslant 0$. Then the spectrum of $-L$ should be nonnegative. If the principal eigenvalue λ_0 of L (i.e. the minimal eigenvalue of $-L$) is zero, then, the L-harmonic function is just a non-trivial eigenfunction corresponding to the eigenvalue $\lambda_0 = 0$ and hence should be nonnegative. The function h should be positive inside the domain based on the maximum principal. Next, if $\lambda_0 > 0$, then replacing L by a shift $L + \lambda_0$, its principal eigenvalue becomes zero, we can continue the study as above, and finally shifting back to the original operator.

In higher dimensional case, the harmonic function may not be unique. We remark that the positive solution of L-harmonic functions for Schrödinger operator $L = \Delta + c(x)$ was examined in [7] in detail, and for elliptic operators in [8] with probabilistic representation.

Example 3.4 ([7; (1.2)]) The L-harmonic function h for $L = \Delta - 1$ can be represented as

$$h(x) = \int_{S^{n-1}} e^{x\cdot\omega} \mathrm{d}\mu(\omega),$$

where μ is a nonnegative measure on the unique sphere S^{n-1}.

The next example is a particular case of Corollary 1.2. Its duality relation was mentioned in [6; §6. Example of O.U.-process and harmonic oscillator], without mention the L-harmonic property of h.

Example 3.5 On $\mathbb{R}$, the function $h(x) = \exp[-x^2/2]$ is L-harmonic:

$$L = \frac{1}{2}\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} + 1 - x^2\right).$$

Its dual is the O.U.-operator:

$$\widetilde{L} = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} - x\frac{\mathrm{d}}{\mathrm{d}x}.$$

Furthermore, L has L^2-eigenvalues $\lambda_n = n\,(n \geqslant 0)$ with eigenfunctions

$$g_n(x) = (-1)^n e^{x^2/2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\big(e^{-x^2}\big), \qquad n \geqslant 0,$$

respectively.

We have just seen an example of the application of known results having $\tilde{c}(x) = 0$ to the one having $c(x) \neq 0$. This indicates a general result as follows.

Theorem 3.6 Given an elliptic operator

$$\widetilde{L} = \sum_{i,j} \tilde{a}_{ij}(x)\partial_{ij}^2 + \sum_i \tilde{b}_i(x)\partial_i, \qquad \mathscr{D}\big(\tilde{L}\big) \subset L^2\big(\tilde{\mu}\big),$$

for each $h \in \mathscr{C}^2$, $h \neq 0$ a.e., $\widetilde{L}$ is L^2-isospectral to L:

$$L = \sum_{i,j} a_{ij}(x)\partial_{ij}^2 + \sum_i b_i(x)\partial_i + c(x), \qquad \mathscr{D}(L) = \big\{f \in \mathscr{E} : f/h \in \mathscr{D}\big(\widetilde{L}\big)\big\},$$

where

$$\begin{aligned} a_{ij}(x) &= \tilde{a}_{ij}(x), \\ b_i(x) &= \tilde{b}_i(x) - \frac{2}{h(x)} \sum_j \tilde{a}_{ij}(x)\partial_j h(x) \ \text{ on } [h \neq 0], \\ c(x) &= \frac{2}{h(x)^2} \sum_{i,j} \tilde{a}_{ij}(x)\partial_i h(x)\partial_j h(x) - \frac{1}{h(x)}\widetilde{L}h(x) \qquad \text{on } [h \neq 0]. \end{aligned}$$

Briefly,

$$\begin{aligned} L &= \widetilde{L} - \frac{2}{h}\big\langle \tilde{a}\nabla h, \nabla\big\rangle + \left[\frac{2}{h^2}\big\langle \tilde{a}\nabla h, \nabla h\big\rangle - \frac{1}{h}\widetilde{L}h\right] \\ &= \widetilde{L} - 2\big\langle \tilde{a}\nabla \log h, \nabla\big\rangle + \Big\{2\big\langle \tilde{a}\nabla \log h, \nabla \log h\big\rangle - h^{-1}\big\langle \tilde{a}\nabla, \nabla h\big\rangle \\ &\qquad + \big\langle \tilde{b}, \nabla \log h\big\rangle\Big\} \quad \text{if } h > 0. \end{aligned}$$

Proof. In parallel to the pure jump case, this is simply a use of the duality $L = H\widetilde{L}H^{-1}$, noting the property that $Lh = 0$ is now automatic since $\widetilde{L}1 = 0$. The remainder of the proof is mainly a careful computation. Actually,

$$\widetilde{L}\left(\frac{f}{h}\right) = \frac{1}{h}\widetilde{L}f + f\widetilde{L}\left(\frac{1}{h}\right) + 2\left\langle \tilde{a}\nabla\left(\frac{1}{h}\right), \nabla f\right\rangle.$$

Hence

$$h\widetilde{L}\left(\frac{f}{h}\right) = \widetilde{L}f + 2h\left\langle \tilde{a}\nabla\left(\frac{1}{h}\right), \nabla f\right\rangle + fh\widetilde{L}\left(\frac{1}{h}\right).$$

From this, it is ready to write down the coefficients of L. □

Corollary 3.7 For given $\widetilde{L}$ and $h = \exp\psi$, the dual operator L takes the following form

$$L = \widetilde{L} - 2\big\langle \tilde{a}\nabla\psi, \nabla\big\rangle + \Big\{\big\langle \tilde{a}\nabla\psi, \nabla\psi\big\rangle - \widetilde{L}\psi\Big\}.$$

We remark that Corollary 3.7 provides us an alternative way to construct the isospectral operator in dimension one. Suppose that we are given an operator

$$\overline{L} = \bar{a}(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \bar{b}(x)\frac{\mathrm{d}}{\mathrm{d}x} + \bar{c}(x).$$

We want to construct $\widetilde{L}$ in terms of the operator L given in Corollary 3.7. First, instead of solving the second order harmonic equation $\overline{L}h = 0$, we need to solve the first order Riccati equation for ϕ:

$$\bar{a}\phi' + \bar{a}\phi^2 + \bar{b}\phi + \bar{c} = 0$$

to which there is a standard iterative procedure in ODE. Next, let ψ satisfy $\psi' = \phi$ and define $\tilde{b} = 2\bar{a}\phi + \bar{b}$. Then we have $L = \overline{L}$. With this $\tilde{b}$ and $\tilde{a} := \bar{a}$, we obtain the operator $\widetilde{L}$ as required.

As an application of the last theorem, one can obtain a lot of examples from [3, 4]. We remark that each $\widetilde{L}$ corresponds to a large class of L since h is quite arbitrary.

The natural higher-dimensional extension of Example 3.5 is as follows.

Example 3.8 The dual of $L = \frac{1}{2}\sum_i \left(\partial_{ii}^2 + 1 - x_i^2\right)$ is $\widetilde{L} = \frac{1}{2}\sum_i \left(\partial_{ii}^2 - 2x_i\partial_i\right)$. The function h takes the form $h(x) = \exp[-|x|^2/2]$ rather than $\sum_i \exp\left[-x_i^2/2\right]$. The operator L has eigenvalue $n\,(n \geqslant 0)$ with multiplicity $\#\{(k_1, k_2, \ldots, k_d) : k_1 + k_2 + \ldots + k_d = n\}$, here $\#$ means the cardinality of the set following.

Proof. For the higher-dimensional O.U.-operator $\widetilde{L}$, we have eigenvalues $\{\sum_{i=1}^d k_i : k_i = 0, 1, \ldots\}$. Corresponding to each $\sum_{i=1}^d k_i$, the eigenfunction is $g(x) := \prod_{i=1}^d g_{k_i}^{(i)}(x_i)$ (where each $g_n^{(i)}$ is the function g_n given in the proof of Corollary 1.2):

$$\widetilde{L}g(x) = -\sum_{i=1}^d k_i g_{k_i}^{(i)}(x_i)\prod_{j\neq i} g_{k_j}^{(j)}(x_j) = -\left(\sum_{i=1}^d k_i\right)g(x).$$

Therefore, $\widetilde{L}$ has eigenvalue $n\,(n \geqslant 0)$ with multiplicity $\#\{(k_1, k_2, \ldots, k_d) : k_1 + k_2 + \ldots + k_d = n\}$. From here, it is easy to write down the eigenvalues of L and their corresponding eigenfunctions. □

Acknowledgments. The results of the paper were presented several times in our seminar, from which the authors are benefited a lot from the discussions and suggestions. Research supported in part by the National Natural Science Foundation of China (No. 11131003), the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Chen, M.F. (2004). *From Markov Chains to Non-equilibrium Particle Systems*. World Scientific. 2nd ed. (1st ed., 1992).

[2] Chen, M.F. (2010). Speed of stability for birth–death processes, Front. Math. China 5(3), 379–515.

[3] Chen, M.F. (2012a). *Basic estimates of stability rate for one-dimensional diffusions*. Chapter 6 in "Probability Approximations and Beyond", 75–99, Lecture Notes in Statistics 205, eds. A.D. Barbour, H.P. Chan and D. Siegmund.

[4] Chen, M.F. (2012b). *Lower bounds of the principal eigenvalue in dimension one*. Front. Math. China 2012, 7(4): 645–668.

[5] Chen, M.F. and Zhang, Y.H. (2014). *Unified representation of formulas for single birth processes*. Preprint.

[6] Jansen, S. and Kurt, N. (2012). *On the notion(s) of duality for Markov processes*. arXiv:1210.7193.

[7] Murata, M. (1986). *Structure of positive solutions to* $(-\Delta + V)u = 0$ *in* $\mathbb{R}^n$. Duke Math. J. 53(4): 869-943.

[8] Pinsky, R.G. (1995). *Positive Harmonic Functions and Diffusion*. Cambridge University Press.

[9] Pinsky, R.G. (2009). *Explicit and almost explicit spectral calculations for diffusion operators*. J. Funct. Anal. 256(10): 3279C3312.

[10] Wang, J. (2012). *Sharp bounds for the first eigenvalue of symmetric Markov processes and their applications*. Acta Math. Sin. Eng. Ser. 28(10): 1995C2010.

[11] Zettl, A. (2005). *Sturm–Liouville Theory*. AMS, Providence, Rhode Island.

[12] Zhang, X. (2013).*On the eigenvalues of birth-death processes with killing*. Preprint.

Mu-Fa Chen

School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.

E-mail: mfchen@bnu.edu.cn

Home page: http://math.bnu.edu.cn/~chenmf/main_eng.htm

Xu Zhang

College of Applied Sciences, Beijing University of Technology, Beijing 100022, The People's Republic of China.

E-mail: zhangxu@bjut.edu.cn

Commun. Math. Stat. 2014, 2(3): 279–309

Criteria for Discrete Spectrum of 1D Operators

Mu-Fa Chen

(Beijing Normal University)

October 21, 2014

Abstract

For discrete spectrum of 1D second-order differential/difference operators (with or without potential (killing), with the maximal/minimal domain), a pair of unified dual criteria are presented in terms of two explicit measures and the harmonic function of the operators. Interestingly, these criteria can be read out from the ones for the exponential convergence of four types of stability studied earlier, simply replacing the 'finite supremum' by 'vanishing at infinity'. Except a dual technique, the main tool used here is a transform in terms of the harmonic function, to which two new practical algorithms are introduced in the discrete context and two successive approximation schemes are reviewed in the continuous context. All of them are illustrated by examples. The main body of the paper is devoted to the hard part of the story, the easier part but powerful one is delayed to the end of the paper.

2000 *Mathematics Subject Classification*: 34L05, 60J27, 60J60

Key words and phrases. Discrete spectrum; essential spectrum; tridiagonal matrix (birth–death process); second-order differential operator (diffusion); killing.

Received: 17 December 2014 / Accepted: 23 December 2014

1 Introduction

The spectral theory is an active research subject, not only in mathematics but also in physics. The discrete spectrum has an especial meaning in quantum physics, it represents the discrete levels of energy. From the Internet, one may find a large number of publications in the field (more than $50,000$ webpages in the scholar search for "discrete spectrum"). From the search, we learnt that the theory was begun in early 1900s, mainly from the interaction of mathematics and physics, by F. Riesz, D. Hilbert, H. Weyl, J. von Neumann, and many others. In particular, the concept of "essential spectrum" used below was first introduced by H. Weyl in 1910. Surprisingly, in such a long

time-developed field, the known complete results are still rather limited, even in dimension one. We will review some of the related results case by case subsequently.

This paper deals with one-dimensional case only. Mainly, the results come from three resources: (i) Mao's criteria [10] in the ergodic case; (ii) the Karlin–McGregor's dual technique (cf. [3]); and (iii) an isospectral transform introduced recently by the author and X. Zhang (2014). The last point is essential different from the known approach (comparing with [12, 7]). If the harmonic function is replaced by the ground state (i.e., the eigenfunction corresponding to the principal eigenvalue), then the transform in (iii) is just the H-transform often used in the study of spectral gap for Schrödinger operators. Certainly, in practice, it is important to estimate the harmonic function. For this, we introduce some easier algorithms in the discrete context and review two successive approximation schemes in the continuous context.

A large part of the paper (5 sections: §2–§6) deals with the discrete space. A typical result of the paper is presented in the next section (Theorem 2.1), its proof is given in §3. Some illustrating examples are also presented in the next section, their proofs are delayed to §6. The new algorithms are presented in §4 and §5. The continuous analog of the results in the discrete case is presented in the last section (§7) of the paper. Additionally, a powerful application of our approach is illustrated by Corollary 7.9 and Examples 7.10 and 7.11.

2 Main results in discrete case

Given a tridiagonal matrix $Q^c = \{q_{ij}\}$ on $E := \{0, 1, 2, \ldots\}$: $q_{i,i+1} = b_i > 0\,(i \geqslant 0)$, $q_{i,i-1} = a_i > 0\,(i \geqslant 1)$, $q_{i,i} = -(a_i + b_i + c_i)$, where $c_i \geqslant 0\,(i \geqslant 0)$, and $q_{i,j} = 0$ for other $j \neq i$. From probabilistic language, this matrix corresponds to a birth–death process with birth rates b_i, death rates a_i and killing rates c_i. Corresponding to the matrix Q^c, we have an operator

$$\Omega^c f(k) = b_k(f_{k+1} - f_k) + a_k(f_{k-1} - f_k) - c_k f_k, \qquad k \in E, \;\; a_0 := 0.$$

In what follows, we need two measures μ and $\hat{\nu}$ on E:

$$\mu_0 = 1, \;\; \mu_n = \frac{b_0 \cdots b_{n-1}}{a_1 \cdots a_n}, \quad n \geqslant 1; \qquad \hat{\nu}_n = \frac{1}{\mu_n b_n}, \quad n \geqslant 0.$$

Corresponding to the operator Ω^c, on $L^2(\mu)$, there are two quadratic (Dirichlet) forms

$$D^c(f) = \sum_{k \geqslant 0} \mu_k \big[b_k(f_{k+1} - f_k)^2 + c_k f_k^2\big]$$

either with the maximal domain

$$\mathscr{D}_{\max}(D^c) = \{f \in L^2(\mu) : D^c(f) < \infty\}$$

or with the minimal one $\mathscr{D}_{\min}(D^c)$ which is the smallest closure of

$$\{f \in L^2(\mu) : f \text{ has a finite support}\}$$

with respect to the norm $\|\cdot\|_D$: $\|f\|_D^2 = \|f\|_{L^2(\mu)}^2 + D^c(f)$. The spectrum we are going to study is with respect to these Dirichlet forms. We say that $\big(D^c, \mathscr{D}_{\min}(D^c)\big)$ has discrete spectrum $\big($equivalently, the essential spectrum of $\big(D^c, \mathscr{D}_{\min}(D^c)\big)$, denoted by $\sigma_{\text{ess}}\big(\Omega^c_{\min}\big)$, is empty$\big)$ if its spectrum consists only isolated eigenvalues of finite multiplicity. For an operator L, we have

$$\text{spectrum of } L = \text{discrete part} + \text{essential part}.$$

Hence the statement "L has discrete spectrum" is exactly the same as "$\sigma_{\text{ess}}(L) = \varnothing$". To state our first main result, we need some notation. Define

$$u_i = \frac{a_i}{b_i}, \qquad v_i = \frac{c_i}{b_i}, \qquad \xi_i = 1 + u_i + v_i, \qquad i \geqslant 0;$$

$$r_0 = \frac{1}{1+v_0}, \quad r_n = \cfrac{1}{\xi_n - \cfrac{u_n}{\xi_{n-1} - \cfrac{u_{n-1}}{\ddots\, \xi_2 - \cfrac{u_2}{\xi_1 - \cfrac{u_1}{1+v_0}}}}} = \frac{1}{\xi_n - u_n r_{n-1}}, \qquad n \geqslant 1;$$

$$h_0 = 1, \qquad h_n = \left(\prod_{k=0}^{n-1} r_k\right)^{-1}, \qquad n \geqslant 1.$$

For simplicity, we write

$$\text{Spec}\big(\Omega^c_{\min}\big) = \text{The } L^2(\mu)\text{-spectrum of } \big(D^c, \mathscr{D}_{\min}(D^c)\big).$$

Similarly, we have $\text{Spec}\big(\Omega^c_{\max}\big)$.

Theorem 2.1 (1) Let $\sum_{k=0}^{\infty}(h_k h_{k+1}\mu_k b_k)^{-1} < \infty$. Then $\text{Spec}\big(\Omega^c_{\min}\big)$ is discrete iff

$$\lim_{n\to\infty} \sum_{j=0}^{n} \mu_j h_j^2 \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1}\mu_k b_k} = 0.$$

(2) Let $\sum_{j=0}^{\infty} \mu_j h_j^2 < \infty$. Then $\text{Spec}\big(\Omega^c_{\max}\big)$ is discrete iff

$$\lim_{n\to\infty} \sum_{j=n+1}^{\infty} \mu_j h_j^2 \sum_{k=0}^{n} \frac{1}{h_k h_{k+1}\mu_k b_k} = 0.$$

(3) Let $\sum_{k=0}^{\infty}(h_k h_{k+1}\mu_k b_k)^{-1} = \infty = \sum_{j=0}^{\infty} \mu_j h_j^2$. Then $\text{Spec}\big(\Omega^c_{\min}\big) = \text{Spec}\big(\Omega^c_{\max}\big)$ is not discrete.

Corollary 2.2 If $\sigma_{\rm ess}\big(\Omega^c_{\min}\big) = \varnothing$, then $\lambda_0\big(\Omega^c_{\min}\big) > 0$, where

$$\lambda_0\big(\Omega^c_{\min}\big) = \inf\big\{D^c(f) : f \in \mathscr{D}_{\min}(D^c), \|f\|_{L^2(\mu)} = 1\big\}.$$

Proof. Once $\sigma_{\rm ess}\big(\Omega^c_{\min}\big) = \varnothing$, by Theorem 2.1 (1), it is obvious that

$$\sup_n \sum_{j=0}^{n} \mu_j h_j^2 \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1} \mu_k b_k} < \infty.$$

Then the conclusion follows from [5; Theorem 2.6]. □

Remark 2.3 If $c_i \equiv 0$. Then $v_i \equiv 0$ and $\xi_n \equiv 1 + u_n$. Since $r_0 = 1$ and $r_n = (\xi_n - u_n r_{n-1})^{-1}$, by induction, it is obvious to see that $r_n \equiv 1$ and then $h_n \equiv 1$.

When $c_i \equiv 0$, we drop the superscript c from Ω^c and D^c for simplicity. In this case, part (2) of the theorem is due to [10; Theorem 1.2]. Under the same condition, a parallel spectral property of the birth–death processes has recently obtained by [13]. The criteria in the present general setup seem to be new. Let us mention that different sums $\sum_n^\infty$ and $\sum_{n+1}^\infty$ are used respectively in the first two parts of Theorem 2.1.

Before moving further, let us explain the reasons for the partition of three parts given in the theorem.

Remark 2.4 Consider $c_i \equiv 0$ only for simplicity.

(a) First, let $\sum_n \mu_n < \infty$. If furthermore $\sum_n (\mu_n b_n)^{-1} = \infty$, then the corresponding unique birth–death process is ergodic. It becomes exponentially ergodic iff the first non-trivial "eigenvalue" λ_1 $\big($or the spectral gap $\inf\{{\rm Spec}(\Omega)\backslash\{0\}\}\big)$ is positive. Equivalently,

$$\sup_{n\geqslant 1} \sum_{k=0}^{n-1} \frac{1}{\mu_k b_k} \sum_{j=n}^{\infty} \mu_j < \infty$$

(cf. [2; Theorem 9.25]). One may compare this condition with part (2) of Theorem 2.1 having $h_n \equiv 1$. Clearly, this is a necessary condition for ${\rm Spec}\big(\Omega\big)$ to be discrete. The exponential ergodicity means that the process will return to the origin exponentially fast. Hence with probability one, it will never go to infinity.

(b) Conversely, if $\sum_n (\mu_n b_n)^{-1} < \infty$. Then the process is transient. It decays (or "goes to infinity") exponentially fast iff

$$\sup_{n\geqslant 1} \sum_{j=0}^{n} \mu_j \sum_{k=n}^{\infty} \frac{1}{\mu_k b_k} < \infty.$$

Refer to [3; Theorem 3.1] for more details. One may compare this condition with part (1) of Theorem 2.1 having $h_n \equiv 1$. This conclusion holds even without the uniqueness assumption:

$$\sum_{k=0}^{\infty} \frac{1}{\mu_k b_k} \sum_{j=0}^{k} \mu_j = \infty.$$

(cf. [2; Corollary 3.18] or [3; (1.2)]).

(c) Let $\sum_n \mu_n < \infty$ and $\mathscr{D}_{\min}(D) \neq \mathscr{D}_{\max}(D)$. From [3; Proposition 1.3]), it is known that $\mathscr{D}_{\min}(D) = \mathscr{D}_{\max}(D)$ iff

$$\sum_{k=0}^{\infty} \left(\frac{1}{\mu_k b_k} + \mu_k \right) = \infty.$$

Hence we have also $\sum_{k=0}^{\infty} (\mu_k b_k)^{-1} < \infty$. In this case, we should study their spectrum separately. For the maximal one $\big(D, \mathscr{D}_{\max}(D)\big)$, the solution is given by part (2) of the theorem. For the minimal one, the solution is given in part (1). In this case, both $\mathrm{Spec}(\Omega_{\min})$ and $\mathrm{Spec}(\Omega_{\max})$ are discrete. In [5; Theorem 2.6], the principal eigenvalue is studied only in a case for $\Omega^c_{\min}$. The other three cases (cf. [3]) should be in parallel. For instance, $\Omega^c_{\max}$ corresponds to an extended Hardy inequality:

$$\|f\|^2_{L^2(\mu)} \leqslant AD^c(f), \qquad f \in L^2(\mu),$$

where A is a constant. However, for $\Omega^c_{\min}$, the condition "$f \in L^2(\mu)$" in the last line should be replaced by "f has finite support".

(d) As for part (3) of the theorem, since part (1) remains true even if $\sum_n (\mu_n b_n)^{-1} = \infty$. Dually, part (2) remains true even if $\sum_n \mu_n = \infty$. Alternatively, in case (3), the birth-death is zero recurrent and so the spectrum can not be discrete. Actually, it can not have exponential decay. Otherwise,

$$\infty = \int_0^{\infty} p_{ii}(t)\mathrm{d}t \leqslant C \int_0^{\infty} e^{-\lambda_0 t} < \infty.$$

Besides, $\mathscr{D}_{\min}(D) = \mathscr{D}_{\max}(D)$. The assertion is now clear.

The next four simple examples show that the three parts in Theorem 2.1 are independent. Note that in what follows, we do not care about b_0 and a_0 since a change of finite number of the coefficients does not effect our conclusion (in general, the essential spectrum is invariant under compact perturbations).

Example 2.5 Let $b_n = n^4$ and $\mu_n = n^{-2}$. Then both $\mathrm{Spec}(\Omega_{\min})$ and $\mathrm{Spec}(\Omega_{\max})$ are discrete.

Proof. Since $\hat{\nu}_n = n^{-2}$, we have $\sum_n \mu_n < \infty$ and $\sum_n \hat{\nu}_n < \infty$. The assertion follows from the first two parts of Theorem 2.1. □

Example 2.6 Let $c_n \equiv 0$, $b_n = a_{n+1} = n^\gamma\,(\gamma \geqslant 0)$. Then $\mathrm{Spec}(\Omega_{\min})$ is discrete iff $\gamma > 2$. In particular, if $\gamma \in [0,1]$, then $\mathrm{Spec}(\Omega_{\min}) = \mathrm{Spec}(\Omega_{\max})$ is not discrete.

Proof. Because $\mu_n \sim 1$, $\hat{\nu}_n \sim n^{-\gamma}$. Hence $\sum_n \hat{\nu}_n < \infty$ iff $\gamma > 1$,

$$\sum_0^n \mu_k \sum_n^\infty \hat{\nu}_j \sim n^{2-\gamma}.$$

The main assertion follows from the last two parts of Theorem 2.1. In the particular case that $\gamma \in [0,1]$, we have $\sum_n \mu_n = \infty$ and $\sum_n \hat{\nu}_n = \infty$. The assertion follows from part (3) of Theorem 2.1. □

Dually, we have the following example.

Example 2.7 Let $c_n \equiv 0$, $a_n = b_n = n^\gamma\,(\gamma \geqslant 0)$. Then $\mathrm{Spec}\big(\Omega^c_{\max}\big)$ is discrete iff $\gamma > 2$. In particular, when $\gamma \in [0,1]$, then $\mathrm{Spec}(\Omega_{\min}) = \mathrm{Spec}(\Omega_{\max})$ is not discrete.

Since a local modification of the rates does not make influence to our conclusion, we obtain the next result.

Example 2.8 If $c_i \neq 0$ only on a finite set, then the conclusions of the last three examples remain the same.

The next three examples are much more technical since their (c_n) are not local. This is what we have to pay by our approach. The proofs are delayed to Section 6.

Example 2.9 Let $a_n = b_n = 1$ and $c_n \downarrow 0$. Then $\mathrm{Spec}\big(\Omega^c_{\min}\big)$ is not discrete or equivalently $\sigma_{\mathrm{ess}}\big(\Omega^c_{\min}\big) \neq \varnothing$.

Example 2.10 Let $a_n = b_n = (n+1)/4$, $c_n = 9(n+1)/16$. Then $\sigma_{\mathrm{ess}}\big(\Omega^c_{\min}\big) = \varnothing$.

Example 2.11 Let $a_n = b_n = (n+1)^2$, $c_n = 5+10/(5n-12)$. Then $\sigma_{\mathrm{ess}}\big(\Omega^c_{\min}\big) \neq \varnothing$.

3 Proof of Theorem 2.1

(a) The computation of the Ω^c-harmonic function h (i.e., $\Omega^c h = 0$) used in the theorem is delayed to Section 5.

(b) By using h, one can reduce the case of $c_i \not\equiv 0$ to the one that $c_i \equiv 0$. Roughly speaking, the idea goes as follows. Let $h \neq 0$, μ-a.e. Then the mapping $f \to \tilde{f}$: $\tilde{f} = \mathbb{1}_{[h\neq 0]} f/h$ is an isometry from $L^2(\mu)$ to $L^2(\tilde{\mu})$, where $\tilde{\mu} = h^2\mu$. Next, for given operator $\big(\Omega^c, \mathscr{D}\big(\Omega^c\big)\big)$, one may introduce an operator

$\widetilde{\Omega}$ on $L^2(\tilde{\mu})$ (without killing) with deduced domain $\mathscr{D}\big(\widetilde{\Omega}\big)$ from $\mathscr{D}\big(\Omega^c\big)$ under the mapping $f \to \tilde{f}$ such that

$$\big(\Omega^c f,\, f\big)_\mu = \big(\widetilde{\Omega}\tilde{f},\, \tilde{f}\big)_{\tilde{\mu}}, \qquad f \in \mathscr{D}\big(\Omega^c\big).$$

This implies that the corresponding quadratic form $\big(D^c, \mathscr{D}\big(D^c\big)\big)$ on $L^2(\mu)$ coincides with $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$ on $L^2(\tilde{\mu})$ under the same mapping, and hence

$$\mathrm{Spec}_\mu\big(\Omega^c\big) = \mathrm{Spec}_{\tilde{\mu}}\big(\widetilde{\Omega}\big).$$

Refer to [5; Lemma 1.3 and §2]. Actually, as studied in the cited paper, this idea works in a rather general setup.

From now on in this section, we assume that $c_i \equiv 0$.

(c) Consider first $\mathrm{Spec}\big(\Omega_{\max}\big)$. Without loss of generality, assume that $\mu(E) < \infty$. Otherwise, $\sigma_{\mathrm{ess}}(\Omega_{\max}) \neq \varnothing$. (Actually, in this case, the spectral gap vanishes and so the spectrum can not be discrete.) By [10; Theorem 1.2], $\sigma_{\mathrm{ess}}(\Omega) = \varnothing$ iff

$$\lim_{n\to\infty} \mu[n,\infty) \sum_{j=0}^{n-1} \frac{1}{\mu_j b_j} = \lim_{n\to\infty} \hat{\nu}[0,n]\, \mu[n+1,\infty) = 0,$$

where (μ_n) and $(\hat{\nu}_n)$ are defined at the beginning of the paper. This is the condition given in Theorem 2.1 (2) with $h_k \equiv 1$. Here we remark that in the original [10; Theorem 1.2], the non-explosive (uniqueness) assumption was made. However, as mentioned in [3; §6], one can use the maximal process instead of the uniqueness condition. This remains true in the present setup, since the basic estimates for the principal eigenvalue used in [10; Theorem 2.4] do not change if the uniqueness condition is replaced by the use of the maximal process, as proved in [3; §4].

(d) Define a dual birth–death process on $\{0, 1, 2, \ldots\}$ by

$$b_i^* = a_{i+1}, \qquad a_i^* = b_i, \qquad i \geqslant 0.$$

Similar to (μ_n) and $(\hat{\nu})$, we have

$$\mu_0^* = 1, \;\; \mu_n^* = \frac{b_0^* \cdots b_{n-1}^*}{a_1^* \cdots a_n^*}, \quad n \geqslant 1; \qquad \hat{\nu}_n^* = \frac{1}{\mu_n^* b_n^*}, \quad n \geqslant 0.$$

Then

$$\mu_n = \frac{a_0^* \cdots a_{n-1}^*}{b_0^* \cdots b_{n-1}^*} = \frac{a_0^*}{\mu_{n-1}^* b_{n-1}^*} = a_0^* \hat{\nu}_{n-1}^*, \qquad n \geqslant 1.$$

$$\hat{\nu}_n = \frac{1}{\mu_n b_n} = \frac{1}{a_0^* \hat{\nu}_{n-1}^* a_n^*} = \frac{\mu_{n-1}^* b_{n-1}^*}{a_0^* a_n^*} = \frac{\mu_n^*}{a_0^*}, \qquad n \geqslant 1.$$

The last equality holds also at $n = 0$, and then

$$\hat{\nu}_n = \frac{1}{\mu_n b_n} = \frac{\mu_n^*}{a_0^*}, \qquad n \geqslant 0.$$

Therefore,

$$\hat{\nu}[0,n]\,\mu[n+1,\infty) = \frac{1}{a_0^*}\mu^*[0,n]\,a_0^*\,\hat{\nu}^*[n,\infty) = \mu^*[0,n]\,\hat{\nu}^*[n,\infty).$$

Clearly, we have $\hat{\nu}^*(E) < \infty$ iff $\mu(E) < \infty$.

Next, define

$$M = \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 & \dots \\ 0 & \mu_1 & \mu_2 & \mu_3 & \dots \\ 0 & 0 & \mu_2 & \mu_3 & \dots \\ 0 & 0 & 0 & \mu_3 & \dots \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}, \quad M^{-1} = \begin{bmatrix} \frac{1}{\mu_0} & -\frac{1}{\mu_0} & 0 & 0 & \dots \\ 0 & \frac{1}{\mu_1} & -\frac{1}{\mu_1} & 0 & \dots \\ 0 & 0 & \frac{1}{\mu_2} & -\frac{1}{\mu_2} & \dots \\ 0 & 0 & 0 & \frac{1}{\mu_3} & \dots \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}.$$

Then we have $\Omega^* = M\Omega M^{-1}$ or equivalently, $Q^* = MQM^{-1}$. In other words, Ω and Ω^* are similar and so have the same spectrum (one may worry the domain problem of the operators, but they can be approximated by finite ones, as used often in the literature, see for instance [3]). Now, we can read from proof (c) above for a criterion for $\mathrm{Spec}(\Omega^*_{\min})$ to have discrete spectrum: $\sigma_{\mathrm{ess}}(\Omega^*_{\min}) = \varnothing$ iff

$$\lim_{n\to\infty} \mu^*[0,n]\,\hat{\nu}^*[n,\infty) = 0.$$

Ignoring the superscript $*$, this is the condition given in Theorem 2.1 (1) with $h_k \equiv 1$.

4 An algorithm for (h_i) in the "lower-triangle" case.

To get a representation of the harmonic function h, as mentioned in [6; Remark 2.5 (3)], even in the special case of birth–death processes, we originally still had to go to a more general setup: the "lower-triangle" matrix (or single birth process). For those reader who is interested in the tridiagonal case only, one may jump from here to the next section. The matrix we are working in this section is as follows: $q_{i,i+1} > 0$ for each $i \geqslant 0$ but $q_{ij} \geqslant 0$ can be arbitrary for every $j < i$. For each $(c_i \in \mathbb{R})$, the operator Ω^c becomes

$$\Omega^c f(i) = \sum_{j<i} q_{ij}(f_j - f_i) + q_{i,i+1}(f_{i+1} - f_i) - c_i f_i, \qquad i \geqslant 0.$$

To be consistence to what used in the last section, we replace c_i used in [6] by $-c_i$ here. Following [6; Theorem 1.1], we adopt the notation:

$$\tilde{q}_n^{(k)} = \sum_{j=0}^{k} q_{nj} + c_n \ \ (\text{here } c_n \in \mathbb{R}!), \qquad 0 \leqslant k < n,$$

$$\widetilde{F}_i^{(i)} = 1, \quad \widetilde{F}_n^{(i)} = \frac{1}{q_{n,n+1}} \sum_{k=i}^{n-1} \tilde{q}_n^{(k)} \widetilde{F}_k^{(i)}, \qquad n > i \geqslant 0,$$

$$g_n = g_0 + \sum_{0 \leqslant k \leqslant n-1} \sum_{0 \leqslant j \leqslant k} \widetilde{F}_k^{(j)} \frac{f_j + c_j g_0}{q_{j,j+1}} \quad \left[\sum_{\varnothing} := 0 \right], \qquad n \geqslant 0.$$

The theorem just cited says that (g_n) is the solution to the Poisson equation

$$\Omega^c g = f \quad \text{on } E = \{0, 1, \cdots\}.$$

In particular, when $f = 0$, this g gives us the unified formula of Ω^c-harmonic function h.

We now introduce an alternative algorithm for $\{\widetilde{F}_n^{(i)}\}_{n \geqslant i \geqslant 0}$ (and then for $\{g_n\}_{n \geqslant 0}$). This is meaningful since it is the most important sequence used in [6]. The advantage of the new algorithm given in (1) below is that at the kth step in computing $G_{\cdot,k}^{(i)}$, we use $G_{\cdot,k-1}^{(i)}$ only but not $G_{\cdot,s}^{(i)}$ all s: $i \leqslant s \leqslant k-2$, as in the original computation for $\widetilde{F}_n^{(i)}$ where the whole family $\{\widetilde{F}_s^{(i)}\}_{s=i}^{n-1}$ is required.

Proposition 4.1 Let

$$u_\ell^{(i)} = \frac{\tilde{q}_{i+\ell}^{(i)}}{q_{i+\ell, i+\ell+1}}, \qquad i \geqslant 0, \ \ell \geqslant 1.$$

Fix $i \geqslant 0$, define $\{G_{\ell,k}^{(i)} : \ell \geqslant k\}_{k \geqslant 1}$, recursively in k, by

$$G_{\ell,k}^{(i)} = G_{\ell,k-1}^{(i)} + u_{\ell-k+1}^{(i+k-1)} G_{k-1,k-1}^{(i)}, \qquad (\ell \geqslant) k \geqslant 2 \tag{1}$$

with initial condition

$$G_{\ell,1}^{(i)} = u_\ell^{(i)}, \qquad \ell \geqslant 1.$$

Then, with $G_{0,0}^{(i)} \equiv 1$, we have the following alternative representation.

(1) For each $m \geqslant 0$ and $i \geqslant 0$,

$$\widetilde{F}_{i+m}^{(i)} = G_{m,m}^{(i)}.$$

(2) For each $n \geqslant 0$,

$$g_n = g_0 + \sum_{0 \leqslant j \leqslant n-1} v_j \sum_{k=0}^{n-j-1} G_{k,k}^{(j)},$$

where

$$v_j = \frac{f_j + c_j g_0}{q_{j,j+1}}, \qquad j \geqslant 0.$$

Proof. (a) To prove part (1) of the proposition, by [6; (2.7)], we have

$$\widetilde{F}_i^{(i)} = 1, \quad \widetilde{F}_n^{(i)} = \sum_{k=i+1}^{n} \widetilde{F}_n^{(k)} \frac{\tilde{q}_k^{(i)}}{q_{k,\,k+1}}, \qquad n \geqslant i+1.$$

Rewrite

$$\widetilde{F}_n^{(i)} = \sum_{\ell=1}^{n-i} \widetilde{F}_n^{(i+\ell)} \frac{\tilde{q}_{i+\ell}^{(i)}}{q_{i+\ell,\,i+\ell+1}}, \qquad n \geqslant i+1.$$

For simplicity, let

$$m = n - i, \quad f_m^{(i)} = \widetilde{F}_{m+i}^{(i)}, \quad u_\ell^{(i)} = \frac{\tilde{q}_{i+\ell}^{(i)}}{q_{i+\ell,\,i+\ell+1}}.$$

Then we have

$$f_0^{(i)} = 1, \quad f_m^{(i)} = \sum_{\ell=1}^{m} f_{m-\ell}^{(i+\ell)} u_\ell^{(i)}, \qquad m \geqslant 1,\ i \geqslant 0. \tag{2}$$

The goal of the construction of $\{G_{\cdot,k}^{(i)}\}$ is for each k: $1 \leqslant k \leqslant m$, express $f_m^{(i)}$ as

$$f_m^{(i)} = \sum_{\ell=k}^{m} f_{m-\ell}^{(i+\ell)} G_{\ell,k}^{(i)}.$$

Clearly, $f_1^{(i)} = u_1^{(i)}$. Next, by (2), we have

$$f_{m-1}^{(i+1)} = \sum_{s=1}^{m-1} f_{m-1-s}^{(i+1+s)} u_s^{(i+1)} = \sum_{s=2}^{m} f_{m-s}^{(i+s)} u_{s-1}^{(i+1)}, \qquad m \geqslant 2.$$

Hence by (2) again, it follows that

$$f_m^{(i)} = \sum_{\ell=2}^{m} f_{m-\ell}^{(i+\ell)} u_\ell^{(i)} + f_{m-1}^{(i+1)} u_1^{(i)} = \sum_{\ell=2}^{m} f_{m-\ell}^{(i+\ell)} \big[u_\ell^{(i)} + u_{\ell-1}^{(i+1)} u_1^{(i)} \big].$$

Comparing this and (2), it is clear that for replacing the set $\{1, 2, \ldots, m\}$ by $\{2, 3, \ldots, m\}$ in the summation, we should replace the term

$$u_\ell^{(i)} =: G_{\ell,1}^{(i)}, \qquad (m \geqslant)\ \ell \geqslant 1 \quad \text{(at the first step)}$$

by

$$u_\ell^{(i)} + u_{\ell-1}^{(i+1)} u_1^{(i)} = G_{\ell,1}^{(i)} + u_{\ell-1}^{(i+1)} G_{1,1}^{(i)} =: G_{\ell,2}^{(i)}, \ \ (m \geqslant)\ \ell \geqslant 2.$$

Then, we have

$$f_m^{(i)} = \sum_{\ell=2}^{m} f_{m-\ell}^{(i+\ell)} G_{\ell,2}^{(i)}, \qquad m \geqslant 2 \quad \text{(at the second step)} \tag{3}$$

Similarly, by (2), we have

$$f_{m-2}^{(i+2)} = \sum_{s=1}^{m-2} f_{m-2-s}^{(i+2+s)} u_s^{(i+2)} = \sum_{s=3}^{m} f_{m-s}^{(i+s)} u_{s-2}^{(i+2)}, \qquad m \geqslant 3.$$

Inserting this into (3), it follows that

$$f_m^{(i)} = \sum_{\ell=3}^{m} f_{m-\ell}^{(i+\ell)} G_{\ell,3}^{(i)}, \qquad m \geqslant 3 \quad \text{(at the third step)}$$

with

$$G_{\ell,3}^{(i)} = G_{\ell,2}^{(i)} + u_{\ell-2}^{(i+2)} G_{2,2}^{(i)}, \qquad (m \geqslant)\, \ell \geqslant 3.$$

One may continue the construction of $G_{\cdot,k}^{(i)}$ recursively in k. In particular, with

$$\begin{aligned} f_m^{(i)} &= \sum_{\ell=m-1}^{m} f_{m-\ell}^{(i+\ell)} G_{\ell,\,m-1}^{(i)} \\ &= f_0^{(i+m)} G_{m,\,m-1}^{(i)} + f_1^{(i+m-1)} G_{m-1,\,m-1}^{(i)} \\ &= G_{m,\,m-1}^{(i)} + u_1^{(i+m-1)} G_{m-1,\,m-1}^{(i)} \quad \text{(at } (m-1) \text{ th step)} \end{aligned}$$

and

$$f_m^{(i)} = f_0^{(i+m)} G_{m,m}^{(i)} = G_{m,m}^{(i)} \quad \text{(by (2))},$$

at last, we obtain

$$f_m^{(i)} = G_{m,m}^{(i)} = G_{m,\,m-1}^{(i)} + u_1^{(i+m-1)} G_{m-1,\,m-1}^{(i)} \quad \text{(at the } m \text{ th step)}$$

for $m \geqslant 2$ and $i \geqslant 0$. We have thus proved not only (1) but also the first assertion of the proposition.

(b) To prove part (2) of the proposition, we rewrite g_n as

$$g_n = g_0 + \sum_{0 \leqslant j \leqslant n-1} v_j \sum_{k=j}^{n-1} \widetilde{F}_k^{(j)}, \qquad n \geqslant 0.$$

Then the second assertion follows from the first one of the proposition. □

Remark 4.2 From (1), it follows that

$$G_{k,k}^{(i)} \geqslant u_1^{(i+k-1)} G_{k-1,\,k-1}^{(i)}.$$

Successively, we get

$$\begin{aligned} G_{k,k}^{(i)} &\geqslant u_1^{(i+k-1)} u_1^{(i+k-2)} G_{k-2,\,k-2}^{(i)} \\ &\cdots\cdots \\ &\geqslant u_1^{(i+k-1)} u_1^{(i+k-2)} \cdots u_1^{(i+1)} G_{1,1}^{(i)} \\ &= \prod_{s=0}^{k-1} u_1^{(i+s)}. \end{aligned}$$

We have thus obtained a lower bound of $G^{(i)}_{m,m}$ (and then lower bound of g_n):

$$G^{(i)}_{m,m} \geqslant \prod_{s=0}^{m-1} \frac{\tilde{q}^{(i+s)}_{i+s+1}}{q_{i+s+1,\,i+s+2}}.$$

When $c_i \equiv 0$, we return to the original $F^{(0)}_m$:

$$F^{(0)}_m = G^{(0)}_{m,m} = \prod_{s=0}^{m-1} \frac{a_{s+1}}{b_{s+1}}.$$

5 An algorithm for (h_i) in the tridiagonal case.

We now come back to the birth–death processes and look for a simpler algorithm for the Ω^c-harmonic function h.

Lemma 5.1 For a birth–death process with killing, the Ω^c-harmonic function h:

$$b_i(h_{i+1} - h_i) + a_i(h_{i-1} - h_i) - c_i h_i = 0, \qquad i \geqslant 0$$

can be expressed by the following recursive formula

$$\begin{cases} h_0 = 1, \\ h_1 = 1 + v_0, \\ h_i = (1 + u_{i-1} + v_{i-1})h_{i-1} - u_{i-1}h_{i-2}, \qquad i \geqslant 2, \end{cases}$$

where

$$u_i = \frac{a_i}{b_i}, \qquad v_i = \frac{c_i}{b_i}, \qquad i \geqslant 0.$$

From Lemma 5.1, it is clear that the sequence (h_n) is completely determined by the sequences (u_n) and (v_n).

Next, we introduce a first-order difference equation instead the second-order one used in the last lemma. To do so, set

$$r_i = \frac{h_i}{h_{i+1}}, \qquad i \geqslant 0, \qquad r_0 = \frac{1}{1 + v_0}.$$

By induction, we have $h_i \geqslant (1 + v_{i-1})h_{i-1}$ and hence $r_i \leqslant (1 + v_i)^{-1}$. From

$$h_{n+1} = (1 + u_n + v_n)h_n - u_n h_{n-1}, \qquad n \geqslant 1,$$

we get

$$1 = (1 + u_n + v_n)r_n - u_n r_{n-1} r_n = (1 + u_n + v_n - u_n r_{n-1})r_n, \qquad n \geqslant 1.$$

Clearly, we have

$$1 + u_n + v_n - u_n r_{n-1} = 1 + v_n + u_n(1 - r_{n-1}) \geqslant 1 + v_n \geqslant 1.$$

The next result says that we can describe (h_n) by (r_n) which has a simpler expression.

Proposition 5.2 Let (u_n) and (v_n) be given in the last lemma, set $\xi_n = 1 + u_n + v_n$. Then

$$r_0 = \frac{1}{1+v_0}, \quad r_n = \frac{1}{\xi_n - u_n r_{n-1}} \leqslant \left[1 + v_n + \frac{u_n v_{n-1}}{1+v_{n-1}}\right]^{-1}, \quad n \geqslant 1.$$

Furthermore, the sequences $\{r_n\}$ and $\{h_n\}$ are presented in Theorem 2.1.

In what follows, we are going to work out some more explicit bounds of (r_n) and a more practical corollary of our main criterion (Theorem 2.1). We will pay a particular attention to the case that $u_n \equiv 1$ which is more attractive since then the principal eigenvalue $\lambda_0\big(\Omega^c_{\min}\big) = 0$ ($\Rightarrow \sigma_{\rm ess}\big(\Omega^c_{\min}\big) \neq \varnothing$) once $v_n \equiv 0$. Thus, one may get some impression about the role played by (c_n).

Lemma 5.3 If

$$u_n\Big(\xi_{n-1} - \sqrt{\xi_{n-1}^2 - 4u_{n-1}}\Big) \leqslant u_{n-1}\big(\xi_n - \sqrt{\xi_n^2 - 4u_n}\big)$$

for large n, then by a local modification of the rates (a_i, c_i) if necessary, we have

$$r_n \leqslant \frac{\xi_n - \sqrt{\xi_n^2 - 4u_n}}{2u_n}, \qquad n \geqslant 1 \tag{4}$$

and then

$$h_n \geqslant (1+v_0)\prod_{k=1}^{n-1} \frac{\xi_k + \sqrt{\xi_k^2 - 4u_k}}{2}, \qquad n \geqslant 1.$$

Besides, for $r_{n-1} \leqslant r_n$, condition (4) is necessary.

Proof. Let us start the proof of an informal description of the idea of the lemma. Suppose that $r_n \sim x$ as $n \to \infty$. From the second equation given in Proposition 5.2, we obtain an approximating equation

$$x = \frac{1}{\xi_n - u_n x}.$$

Since $x \leqslant 1$, we have only one solution

$$x = \frac{\xi_n - \sqrt{\xi_n^2 - 4u_n}}{2u_n}.$$

This suggests us the upper bound

$$r_n \leqslant \frac{\xi_n - \sqrt{\xi_n^2 - 4u_n}}{2u_n}$$

for large n. This leads to the conclusion of the lemma.

(a) Assume that condition in the lemma holds starting from n_0, and suppose that (4) holds for $n-1\,(n\geqslant n_0)$. Then we have

$$\begin{aligned} r_n &= \frac{1}{\xi_n - u_n r_{n-1}} \\ &\leqslant \frac{1}{\xi_n - u_n(2u_{n-1})^{-1}\Big(\xi_{n-1} - \sqrt{\xi_{n-1}^2 - 4u_{n-1}}\Big)} \\ &= \frac{2u_{n-1}}{2u_{n-1}\xi_n - u_n\Big(\xi_{n-1} - \sqrt{\xi_{n-1}^2 - 4u_{n-1}}\Big)}. \end{aligned}$$

We now show that the right-hand side is upper bounded by

$$\frac{\xi_n - \sqrt{\xi_n^2 - 4u_n}}{2u_n} = \frac{2}{\xi_n + \sqrt{\xi_n^2 - 4u_n}}.$$

Or equivalently,

$$\frac{u_{n-1}}{2u_{n-1}\xi_n - u_n\Big(\xi_{n-1} - \sqrt{\xi_{n-1}^2 - 4u_{n-1}}\Big)} \leqslant \frac{1}{\xi_n + \sqrt{\xi_n^2 - 4u_n}}.$$

This clearly holds by the condition of the lemma. We have thus obtained (4) for n and then completed the second step of the induction argument.

(b) The proof for the last assertion of the lemma is similar: from $r_{n-1}\leqslant r_n$, one obtains

$$r_n = \frac{1}{\xi_n - u_n r_{n-1}} \leqslant \frac{1}{\xi_n - u_n r_{n-1}}.$$

Solving this inequality and noting that $r_n\leqslant 1$, we obtain again condition (4).

(c) It remains to show that (4) holds for every $n\leqslant n_0-1$ by a suitable modification of the rates, and then complete the induction argument. To see this, we may modify the rates (a_i, c_i) step by step. Let us start at $n=1$. First, let $c_1=0$. Then $v_1=0$. Moreover,

$$\frac{\xi_1 - \sqrt{\xi_1^2 - 4u_1}}{2u_1} = \frac{1+u_1-|1-u_1|}{2u_1} = \begin{cases} 1 & \text{if } u_1\leqslant 1 \\ u_1^{-1} & \text{if } u_1 > 1. \end{cases}$$

Hence we can simply choose $a_1\leqslant b_1$ which implies that $u_1\leqslant 1$. At the same time,

$$r_1 = \frac{1}{\xi_1 - u_1 r_0} = \frac{1}{1+u_1(1-r_0)} \leqslant 1.$$

Therefore, for the modified rates, the assertion holds at $n=1$. Note that this modification does not change anything of r_n for $n\geqslant 3$ and (a_n, b_n, c_n) for $n\geqslant 2$. Besides, for smaller r_1, we have smaller r_2. Continuing the modification step by step, we can arrived at the required conclusion. □

In particular, if $u_n \equiv 1$, the condition of the lemma becomes

$$\sqrt{v_{n-1}(4+v_{n-1})} - v_{n-1} \geqslant \sqrt{v_n(4+v_n)} - v_n$$

which holds once v_n is decreasing in n since the function $\sqrt{x(x+4)} - x$ is increasing in x.

Lemma 5.4 Given two sequences $\{p_n\}$ and $\{q_n\}$, suppose that $q_n \uparrow\uparrow \infty$ as $(n_0 \leqslant) n \uparrow \infty$.

(1) If

$$\frac{p_{n+1} - p_n}{q_{n+1} - q_n} \geqslant \eta, \qquad n \geqslant n_0,$$

then

$$\varliminf_n \frac{p_n}{q_n} \geqslant \eta.$$

(2) Dually, if

$$\frac{p_{n+1} - p_n}{q_{n+1} - q_n} \leqslant \varepsilon, \qquad n \geqslant n_0,$$

then

$$\varlimsup_n \frac{p_n}{q_n} \leqslant \varepsilon.$$

Proof. Here we prove part (1) of the lemma only. Since $q_n \uparrow\uparrow$, by assumption and the proportional property, we have

$$\frac{p_{n+1} - p_{n_0}}{q_{n+1} - q_{n_0}} = \frac{(p_{n+1} - p_n) + \cdots + (p_{n_0+1} - p_{n_0})}{(q_{n+1} - q_n) + \cdots + (q_{n_0+1} - q_{n_0})} \geqslant \eta, \qquad n \geqslant n_0.$$

Because $q_n \uparrow \infty$, we have

$$\varliminf_n \frac{p_n}{q_n} = \varliminf_n \frac{p_{n+1}/q_{n+1} - p_{n_0}/q_{n+1}}{1 - q_{n_0}/q_{n+1}} = \sup_{m > n_0} \inf_{n > m} \frac{p_{n+1} - p_{n_0}}{q_{n+1} - q_{n_0}} \geqslant \eta$$

as required. □

With some obvious change, one may prove the following result.

Lemma 5.5 Suppose that $q_n \downarrow\downarrow 0$ as $(n_0 \leqslant) n \uparrow \infty$.

(1) If

$$\frac{p_n - p_{n+1}}{q_n - q_{n+1}} \geqslant \eta, \qquad n \geqslant n_0,$$

then

$$\varliminf_n \frac{p_n}{q_n} \geqslant \eta.$$

(2) If

$$\frac{p_n - p_{n+1}}{q_n - q_{n+1}} \leqslant \varepsilon, \qquad n \geqslant n_0,$$

then

$$\overline{\lim_n} \frac{p_n}{q_n} \leqslant \varepsilon.$$

Corollary 5.6 Let

$$A_n = \sum_0^n \mu_i h_i^2, \qquad B_n = \sum_{k \geqslant n} \frac{1}{h_k h_{k+1} \mu_k b_k}.$$

If $B_n = \infty$ and $\lim_n A_n = \infty$, then $\sigma_{\rm ess}(\Omega_{\rm min}^c) \neq \varnothing$. Next, assume that $B_n < \infty$.

(1) If $\inf_{n \gg 1} a_n > 0$ and $\lim_n h_n^2 \mu_n \sqrt{a_n} B_n = 0$, then $\lim_n A_n B_n = 0$ and so $\sigma_{\rm ess}\big(\Omega_{\rm min}^c\big) = \varnothing$.

(2) If either $\underline{\lim}_n h_n^2 \mu_n B_n > 0$ or $\underline{\lim}_n h_n^2 \mu_n \sqrt{a_n} B_n > 0$ plus $\inf_{n \gg 1} r_n > 0$, then $\underline{\lim}_n A_n B_n > 0$ and so $\sigma_{\rm ess}\big(\Omega_{\rm min}^c\big) \neq \varnothing$.

Proof. The trivial case that $B_n = \infty$ is easy by our criterion. Now, assume that $B_n < \infty$. Note that $B_n^{-1} \uparrow\uparrow \infty$ as $n \uparrow \infty$. We have

$$\begin{aligned} \frac{A_{n+1} - A_n}{B_{n+1}^{-1} - B_n^{-1}} &= \frac{\mu_{n+1} h_{n+1}^2 B_n B_{n+1}}{1/(h_n h_{n+1} \mu_n b_n)} \\ &= h_n h_{n+1}^3 \mu_n \mu_{n+1} b_n B_{n+1} \left(B_{n+1} + \frac{1}{h_n h_{n+1} \mu_n b_n} \right) \\ &= \big(h_{n+1}^2 \mu_{n+1} \sqrt{a_{n+1}} B_{n+1} \big)^2 r_n + h_{n+1}^2 \mu_{n+1} B_{n+1}. \end{aligned}$$

By part (2) of Lemma 5.4, it follows that

$$\lim_n A_n B_n = 0 \quad \text{once} \quad \lim_n h_n^2 \mu_n \sqrt{a_n}\, B_n = 0.$$

We have proved part (1) of the corollary. The proof of part (2) is similar. □

Lemma 5.7 Assume that

$$r_n < \left(\frac{b_n}{u_n a_{n+1}} \right)^{1/4}, \qquad n \gg 1$$

and $\lim_n h_n^2 \mu_n \sqrt{a_n} = \infty$.

(1) If $\lim_n \big[b_n / \sqrt{a_n} - r_n^2 \sqrt{a_{n+1}} \big] = \infty$, then $\lim_n h_n^2 \mu_n \sqrt{a_n}\, B_n = 0$.

(2) If $\inf_{n \gg 1} r_n > 0$ and $\overline{\lim}_n \big[b_n / \sqrt{a_n} - r_n^2 \sqrt{a_{n+1}} \big] < \infty$, then $\underline{\lim}_n h_n^2 \mu_n \sqrt{a_n}\, B_n > 0$.

Proof. Note that $h_n^2\mu_n\sqrt{a_n}$ is strictly increasing iff

$$r_n < \sqrt{\frac{b_n}{\sqrt{a_n a_{n+1}}}} = \left(\frac{b_n}{u_n a_{n+1}}\right)^{1/4}.$$

If $\lim_n h_n^2\mu_n\sqrt{a_n} = \infty$, then by Lemma 5.5, the study of the limit

$$h_n^2\mu_n\sqrt{a_n}\,B_n = \frac{B_n}{1/(h_n^2\mu_n\sqrt{a_n}\,)}$$

can be reduced to examine the limit of

$$\frac{1/(h_n h_{n+1}\mu_n b_n)}{1/(h_n^2\mu_n\sqrt{a_n}) - 1/(h_{n+1}^2\mu_{n+1}\sqrt{a_{n+1}})} = \frac{r_n}{b_n/\sqrt{a_n} - r_n^2\sqrt{a_{n+1}}}. \quad \square$$

The next result shows that once we know the precise leading order of the summands, the computation used in Theorem 2.1 becomes much easier.

Lemma 5.8 (1) If both $\mu_n h_n^2$ and $\mu_n b_n h_n h_{n+1}$ have algebraic tail (i.e., $\sim n^\alpha$ for some $\alpha > 0$), then

$$\sum_{j=0}^{n} \mu_j h_j^2 \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1}\mu_k b_k} \sim \frac{n^2}{b_n} r_n \qquad \text{as } n \to \infty.$$

$$\sum_{j=n+1}^{\infty} \mu_j h_j^2 \sum_{k=0}^{n} \frac{1}{h_k h_{k+1}\mu_k b_k} \sim \frac{n^2}{a_{n+1} r_n} \qquad \text{as } n \to \infty.$$

(2) If both $\mu_n h_n^2$ and $\mu_n b_n h_n h_{n+1}$ have exponential tail (i.e., $\sim e^{\alpha n}$ for some $\alpha > 0$), then

$$\sum_{j=0}^{n} \mu_j h_j^2 \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1}\mu_k b_k} \sim \frac{r_n}{b_n} \qquad \text{as } n \to \infty.$$

$$\sum_{j=n+1}^{\infty} \mu_j h_j^2 \sum_{k=0}^{n} \frac{1}{h_k h_{k+1}\mu_k b_k} \sim \frac{1}{a_{n+1} r_n} \qquad \text{as } n \to \infty.$$

Proof. Let $\mu_n h_n^2 \sim n^\alpha$ and $\mu_n b_n h_n h_{n+1} \sim n^\beta$. Then

$$\sum_{j=0}^{n} \mu_j h_j^2 \sim n\mu_n h_n^2, \qquad \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1}\mu_k b_k} \sim \frac{n}{h_n h_{n+1}\mu_n b_n}.$$

Hence we obtain the first assertion in part (1). The other assertion can be proved similarly. $\square$

6 Proofs of Examples 2.9–2.11

Proof of Example 2.9 The conclusion that $\sigma_{\rm ess}\big(\Omega^c_{\rm min}\big) \neq \varnothing$ is actually known since the principal eigenvalue $\lambda_0\big(\Omega^c_{\rm min}\big) = 0$ by [3; Example 9.16] which implies the required assertion.

We now prove the assertion by our new criterion. First, noting that $h_n \uparrow$, if $h_\infty := \lim_n h_n < \infty$, then $B_n = \infty$ and so the conclusion follows by Corollary 5.6. Next, let $h_\infty = \infty$. Then $\lim_n h_n^2 \mu_n \sqrt{a_n} = \infty$. Because $c_n \downarrow 0$,

$$r_n < 1 = \left(\frac{b_n}{u_n a_{n+1}}\right)^{1/4}, \quad n \geqslant 1, \qquad \lim_n r_n = 1,$$

we have $\overline{\lim}_n \big[b_n/\sqrt{a_n} - r_n^2\sqrt{a_{n+1}}\big] = 0$ and so the assertion that $\sigma_{\rm ess}\big(\Omega^c_{\rm min}\big) \neq \varnothing$ follows by using part (2) of Lemma 5.7 and part (2) of Corollary 5.6. □

Proof of Example 2.10 The model is modified from [3; Example 9.19] where it was proved that $\lambda_0(\Omega^c_{\rm min}) > 0$. The key for this example is that $u_n \equiv 1$ and $v_n \equiv 9/4$. Hence $r_n \sim 1/4$ and then $h_n \sim 4^n$. Because $\mu_n \sim n^{-1}$, we have $\sum_n \mu_n h_n^2 = \infty$ and $\lim_n h_n^2 \mu_n \sqrt{a_n} = \infty$. It is obvious that

$$r_n < \left(\frac{b_n}{u_n a_{n+1}}\right)^{1/4} = \left(\frac{n+1}{n+2}\right)^{1/4}, \qquad n \geqslant 1.$$

Besides, we have

$$b_n/\sqrt{a_n} - r_n^2\sqrt{a_{n+1}} \sim \frac{15}{16}\sqrt{n} \qquad \text{as } \ n \to \infty.$$

The assertion that $\sigma_{\rm ess}\big(\Omega^c_{\rm min}\big) = \varnothing$ now follows from part (1) of Lemma 5.7 and part (1) of Corollary 5.6.

Lemma 5.8 is applicable to this example. Because $r_n \sim 1/4$ and then $h_n \sim 4^n$, we are in the case of exponential tail. By the first assertion in part (2) of Lemma 5.8, we have

$$\sum_{j=0}^{n} \mu_j h_j^2 \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1} \mu_k b_k} \sim \frac{r_n}{b_n} \sim \frac{1}{n} \sim 0 \qquad \text{as } \ n \to \infty.$$

The required assertion then follows from part (1) of Theorem 2.1. □

Proof of Example 2.11 The model is modified from [3; Example 9.20] where it was proved that $\lambda_0\big(\Omega^c_{\rm min}\big) > 0$. We have $u_n \equiv 1$ and

$$v_n = \frac{1}{(n+1)^2}\left[5 + \frac{10}{5n-12}\right]$$

which is decreasing in $n \geqslant 3$. Because we are studying a property at infinity, without loss of generality, we may apply Lemma 5.3 to derive

$$r_n \leqslant \frac{2 + v_n - \sqrt{(4+v_n)v_n}}{2} < \left(\frac{n+1}{n+2}\right)^{1/2}. \qquad n \geqslant 3. \tag{5}$$

From (5), we get

$$h_n = \left(\prod_{k=0}^{n-1} r_k\right)^{-1} > \left(\prod_{k=0}^{n-1}\left(\frac{k+1}{k+2}\right)^{1/2}\right)^{-1} = (n+1)^{1/2}.$$

Hence $\mu_n h_n^2 \geqslant n^{-1}$ and so $\sum \mu_n h_n^2 = \infty$.

To estimate $\overline{\lim}_n \left[b_n/\sqrt{a_n} - r_n^2\sqrt{a_{n+1}}\right]$, we need a lower bound of r_n. An easier way to do so is modifying the rather precise upper bound of r_n:

$$r_n \leqslant \frac{\xi_n - \sqrt{\xi_n^2 - 4u_n}}{2u_n} = \frac{\xi_n}{2u_n}\left[1 - \sqrt{1 - \frac{4u_n}{\xi_n^2}}\right].$$

Clearly, we need only to look for a lower bound of $-\sqrt{1 - 4u_n/\xi_n^2}$ as $n \to \infty$. For this example, $u_n \equiv 1$, $\xi_n = 2 + v_n$. Since $v_n \to 0$, it is clear that $-\sqrt{1 - 4u_n/\xi_n^2} \sim 0$ as $n \to \infty$. Thus, it is natural to approximate this by second-order polynomials of $1/n$:

$$-\sqrt{1 - \frac{4u_n}{\xi_n^2}} \sim -\frac{2.23615}{n} + \frac{1.81327}{n^2}.$$

This leads us to choose the following lower bound:

$$-\sqrt{1 - \frac{4u_n}{\xi_n^2}} \geqslant -\frac{3}{n} + \frac{2}{n^2}.$$

Then

$$r_n > \left(1 + \frac{v_n}{2}\right)\left(1 - \frac{3}{n} + \frac{2}{n^2}\right) \text{ (by a numerical check) } =: \eta_n \text{ (Fig. 1).}$$

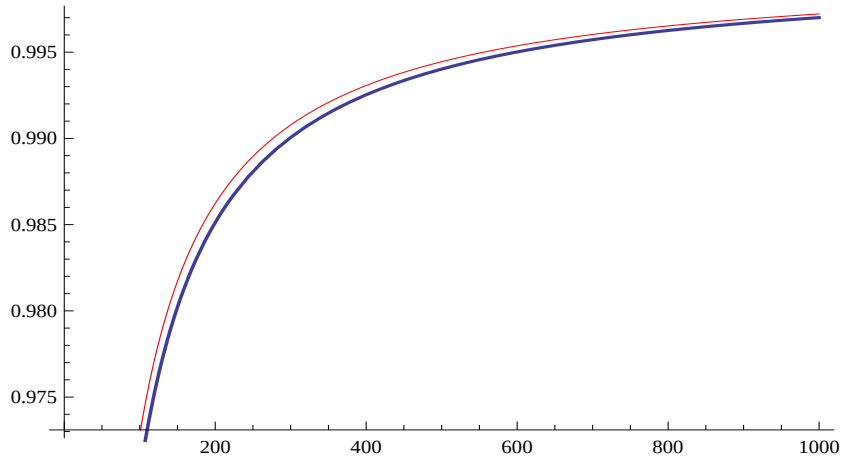


Figure 1 The top curve is r_n. The bottom curve is $\left(1 + \frac{v_n}{2}\right)\left(1 - \frac{3}{n} + \frac{2}{n^2}\right)$.

Furthermore,

$$\overline{\lim_n}\big[b_n/\sqrt{a_n} - r_n^2\sqrt{a_{n+1}}\big] \leqslant \overline{\lim_n}\big[n+1-\eta_n^2(n+2)\big] = 5 < \infty.$$

We mention that there is enough freedom in choosing the lower bound. For instance, replacing $2/n^2$ by $5/n^2$ in the lower bound above, the result is the same. By using part (2) of Lemma 5.7 and part (2) of Corollary 5.6, we obtain $\sigma_{\rm ess}\big(\Omega^c_{\min}\big) \neq \varnothing$.

Alternatively, we can also use Lemma 5.8 to prove this example. We have seen that $r_n \sim 1-\alpha/n$ for some $\alpha > 0$. This means that $h_n \sim n^\alpha$ and hence we are in the case of algebraic tail. By the first assertion in part (1) of Lemma 5.8, we have

$$\sum_{j=0}^n \mu_j h_j^2 \sum_{k=n}^\infty \frac{1}{h_k h_{k+1}\mu_k b_k} \sim \frac{n^2}{b_n} r_n \sim r_n \sim 1 \qquad \text{as } n \to \infty.$$

Then the required assertion follows from part (1) of Theorem 2.1. □

7 Elliptic differential operators (Diffusions)

Consider the elliptic differential (diffusion) operator

$$L^c = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} - c(x), \qquad a(x) > 0,\ c(x) \geqslant 0$$

on $E := (0,\infty)$ or $\mathbb{R}$. Define two measures

$$\mu(\mathrm{d}x) = \frac{e^{C(x)}}{a(x)}\mathrm{d}x, \qquad \nu(\mathrm{d}x) = e^{C(x)}\mathrm{d}x,$$

where $C(x) = \int_\theta^x (b/a)(y)\mathrm{d}y$ and θ is a reference point. Define also a measure deduced from ν

$$\hat\nu(\mathrm{d}x) = e^{-C(x)}\mathrm{d}x.$$

Corresponding to the operator, we have the following Dirichlet form

$$D^c(f) = \int_E f'(x)^2\nu(\mathrm{d}x) + \int_E c(x)f(x)^2\mu(\mathrm{d}x)$$

with domains: either the maximal one

$$\mathscr{D}_{\max}(D^c) = \big\{f \in L^2(\mu) : f \text{ is absolutely continuous and } D^c(f) < \infty\big\},$$

or the minimal one $\mathscr{D}_{\min}(D^c)$ which is the smallest closure of the set

$$\big\{f \in \mathscr{C}^2(E) : f \text{ has a compact support}\big\}$$

with respect to the norm $\|\cdot\|_D$, as in the discrete case (§2).

In parallel to Theorem 2.1, we have the following result.

Theorem 7.1 Let $E = (0, \infty)$ and $h \neq 0$-a.e. be an L^c-harmonic function (to be constructed in Theorem 7.4 below): $L^c h = 0$, a.e.

(1) If $\hat{\nu}\big(h^{-2}\big) < \infty$, then $\sigma_{\rm ess}(L^c_{\rm min}) = \varnothing$ iff

$$\lim_{x\to\infty} \mu\big(h^2 \mathbb{1}_{(0,x)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(x,\infty)}\big) = \lim_{x\to\infty} \int_0^x h^2 \mathrm{d}\mu \int_x^\infty \frac{1}{h^2}\mathrm{d}\hat{\nu} = 0.$$

(2) If $\mu\big(h^2\big) < \infty$, then $\sigma_{\rm ess}(L^c_{\rm max}) = \varnothing$ iff

$$\lim_{x\to\infty} \mu\big(h^2 \mathbb{1}_{(x,\infty)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(0,x)}\big) = \lim_{x\to\infty} \int_x^\infty h^2 \mathrm{d}\mu \int_0^x \frac{1}{h^2}\mathrm{d}\hat{\nu} = 0.$$

(3) If $\hat{\nu}\big(h^{-2}\big) = \infty = \mu\big(h^2\big)$, then $\sigma_{\rm ess}(L^c_{\rm min}) = \sigma_{\rm ess}(L^c_{\rm max}) \neq \varnothing$.

When $c(x) \equiv 0$ and $b(x) \equiv 0$, the first two parts of the theorem may go back to [9]. When $c(x) \equiv 0$, part (1) was presented in [1; Theorem 4.1] and [7; Example 6.1]; under the same condition $c(x) \equiv 0$, part (2) of the theorem is due to [10; Theorem 1.1], again replacing the uniqueness condition by a use of the maximal process. In the general setup, a different criterion for part (1) was presented in [12] and [7; Corollary 5.3], assuming some weak smooth conditions on the coefficients of the operator. Unfortunately, we are unable to state here their results in a short way. In particular, in [7; Corollary 5.3], the family of intervals $\{I(x) : x \in E\}$ is assumed to be existence but is not explicitly constructed. When $b(x) \equiv 0$ in L^c, a compact criterion for part (1) was presented in [12]. For general b in part(1), it was also handled in [9, 12] by a standard change of variables (called time-change in probabilistic language). Unfortunately, as far as we know, the conditions of these general results are usually not easy to verify in practice and so a different approach should be meaningful. Here is a key difference between the approaches, the time-change technique eliminates the first-order differential term b and our H-transform eliminates the killing (or potential) term c.

Corollary 7.2 If $\sigma_{\rm ess}(L^c_{\rm min}) = \varnothing$, then $\lambda_0(L^c_{\rm min}) > 0$.

To study a construction (existence and uniqueness) of an a.e. L^c-harmonic function, we need the following hypothesis.

Hypotheses 7.3 Let $J \subset \mathbb{R}$. Suppose that

(1) $a > 0$ on J;

(2) b/a and c/a are locally integrable with respect to the Lebesgue measure.

In an earlier version, we assumed that e^C/a is locally integrable. Actually, this is equivalent to the local integrability of b/a since C and then e^C are locally bounded due to the assumption that b/a is locally integrable.

Theorem 7.4 Under Hypothesis 7.3, for every $\gamma^{(0)}$, $\gamma^{(1)} \in \mathbb{R}$, an L^c-a.e. harmonic function f always exists. More precisely, a function f can be chosen from the first component of F^* obtained uniquely by the following successive approximation scheme.

(1) *The first successive approximation scheme.* Define

$$F^{(1)}(x) = F(\theta) = \begin{pmatrix} \gamma^{(0)} \\ \gamma^{(1)} \end{pmatrix}, \quad F^{(n+1)}(x) = F(\theta) + \int_\theta^x GF^{(n)}, \quad x \in J,\ n \geqslant 1, \tag{6}$$

where $G(x) = \begin{pmatrix} 0 & e^{-C} \\ ce^C/a & 0 \end{pmatrix}$. Then

$$F^{(n)} \to \begin{pmatrix} f \\ e^C f' \end{pmatrix} =: F^* \qquad as\ n \to \infty \tag{7}$$

uniformly on each compact subinterval of J. In other words, F^* is the unique solution to the equation

$$F(x) = F(\theta) + \int_\theta^x GF, \qquad x \in J \tag{8}$$

and so it is absolutely continuous on each compact subinterval of J.

(2) *The second successive approximation scheme.* Define

$$\widetilde{F}^{(1)}(x) = F(\theta), \qquad \widetilde{F}^{(n+1)}(x) = \int_\theta^x G\widetilde{F}^{(n)}, \quad x \in J,\ n \geqslant 1, \tag{9}$$

then $F^* = \sum_{n=1}^\infty \widetilde{F}^{(n)}$ (which is the so-called Peano-Baker series).

Proof. (a) Part (1) is taken from [14; Theorem 1.2.1 and its proof plus Theorem 2.2.1].

(b) By induction, it is easy to check that $F^{(n)} = \sum_{k=1}^n \widetilde{F}^{(k)}$. Then part (2) follows from part (1). □

A simple way to understand Theorem 7.4 is to look at its differential form of (8):

$$F' = GF, \qquad \text{a.e.} \tag{10}$$

From (9), one sees that the sequence $\{\widetilde{F}^{(n)}\}_{n\geqslant 1}$ is given by a one-step algorithm, as the one for $\{r_n\}_{n\geqslant 1}$ used in the discrete case. Then F^* is given by the summation of $\{\widetilde{F}^{(n)}\}$, which is different from the discrete situation where h is defined by a product of $\{r_n^{-1}\}$.

Theorem 7.5 Let $c, \gamma^{(0)}$, $\gamma^{(1)} \geqslant 0$. Then under Hypothesis 7.3,

(1) the solution F^* constructed in Theorem 7.4 is actually the (finite) minimal nonnegative solution to (8). Furthermore, $F^{(n)} \uparrow F^*$ (pointwise) as $n \to \infty$.

(2) Let $\bar{F}$ be a solution to the inequality

$$F(x) \geqslant F(\theta) + \int_\theta^x GF, \qquad x \in J \tag{11}$$

or more simplicity, to the inequality

$$F' \geqslant GF, \qquad \text{with } \bar{F}(\theta) \geqslant F(\theta). \tag{12}$$

Then $\bar{F} \geqslant F^*$.

Proof. Apply [2; Theorems 2.2, 2.9, and 2.6]). □

Example 7.6 Let

$$L^c = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - c(x), \quad c(x) := \frac{1}{4}x^{2\alpha-2} + \frac{\alpha-1}{2}x^{\alpha-2}, \qquad \alpha \geqslant 1.$$

Then $\sigma_{\mathrm{ess}}(L^c_{\min}) = \varnothing$ if $\alpha > 1$ and $\sigma_{\mathrm{ess}}(L^c_{\min}) \neq \varnothing$ if $\alpha = 1$.

Proof. By Theorem 7.4, we have $F^* = \begin{pmatrix} h \\ e^C h' \end{pmatrix}$. Hence, it is natural to choose $F(\theta) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Because of this, we may denote the first component of $F^{(n)}$ by $h^{(n)}$. Similarly, we have $\tilde{h}^{(n)}$ from the second successive approximation scheme.

First, by using Mathematica, we have $\tilde{h}^{(2n)} = 0$,

$$\begin{aligned}
\tilde{h}^{(1)}(x) &= 1, \\
\tilde{h}^{(3)}(x) &= \frac{x^\alpha}{2\alpha} + \frac{x^{2\alpha}}{8\alpha(2\alpha-1)}, \\
\tilde{h}^{(5)}(x) &= \frac{(\alpha-1)x^{2\alpha}}{8\alpha^2(2\alpha-1)} + \frac{(5\alpha-3)x^{3\alpha}}{48\alpha^2(2\alpha-1)(3\alpha-1)} + \frac{x^{4\alpha}}{128\alpha^2(2\alpha-1)(4\alpha-1)}, \\
\tilde{h}^{(7)}(x) &= \frac{(\alpha-1)^2x^{3\alpha}}{48\alpha^3\,(6\alpha^2-5\alpha+1)} + \frac{(\alpha-1)(7\alpha-3)x^{4\alpha}}{192\alpha^3(2\alpha-1)(3\alpha-1)(4\alpha-1)} \\
&\quad + \frac{\left(89\alpha^2-80\alpha+15\right)x^{5\alpha}}{3840\alpha^3\,(120\alpha^4-154\alpha^3+71\alpha^2-14\alpha+1)} \\
&\quad + \frac{x^{6\alpha}}{3072\alpha^3\,(48\alpha^3-44\alpha^2+12\alpha-1)}.
\end{aligned}$$

Their leading orders are as follows:

$$\tilde{h}^{(1)}(x) = 1,\ \tilde{h}^{(3)}(x) \sim \frac{1}{4}\left(\frac{x^\alpha}{2\alpha}\right)^2,\ \tilde{h}^{(5)}(x) \sim \frac{1}{64}\left(\frac{x^\alpha}{2\alpha}\right)^4,\ \tilde{h}^{(7)}(x) \sim \frac{1}{2304}\left(\frac{x^\alpha}{2\alpha}\right)^6.$$

More simply, one may use $x^{2\alpha-2}/4$ instead of the original $c(x)$, one gets the same leading order of $\tilde{h}^{(2n+1)}$. From this, we guess that $h = \sum_{n=1}^\infty \tilde{h}^{(n)}$ looks

like $\exp\frac{x^\alpha}{2\alpha}$. This becomes more clear when we use the first successive approximation scheme.

$$h^{(1)}(x)=h^{(2)}(x)=1,$$

$$h^{(3)}(x)=h^{(4)}(x)=\underbrace{1+\frac{x^\alpha}{2\alpha}}+\frac{x^{2\alpha}}{8\alpha(2\alpha-1)},$$

$$h^{(5)}(x)=h^{(6)}(x)=\underbrace{1+\frac{x^\alpha}{2\alpha}+\frac{x^{2\alpha}}{8\alpha^2}}+\frac{(5\alpha-3)x^{3\alpha}}{48\alpha^2\,(6\alpha^2-5\alpha+1)}+\frac{x^{4\alpha}}{128\alpha^2\,(8\alpha^2-6\alpha+1)},$$

$$\begin{aligned}h^{(7)}(x)=h^{(8)}(x)=&\underbrace{1+\frac{x^\alpha}{2\alpha}+\frac{x^{2\alpha}}{8\alpha^2}+\frac{x^{3\alpha}}{48\alpha^3}}+\frac{\left(23\alpha^2-23\alpha+6\right)x^{4\alpha}}{384\alpha^3(3\alpha-1)\,(8\alpha^2-6\alpha+1)}\\&+\frac{\left(89\alpha^2-80\alpha+15\right)x^{5\alpha}}{3840\alpha^3(3\alpha-1)(5\alpha-1)\,(8\alpha^2-6\alpha+1)}\\&+\frac{x^{6\alpha}}{3072\alpha^3(6\alpha-1)\,(8\alpha^2-6\alpha+1)}.\end{aligned}$$

Obviously, $h^{(n)}$ is approximating to

$$\exp\left[\frac{x^\alpha}{2\alpha}\right]=\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{x^\alpha}{2\alpha}\right)^n$$

step by step. Since $\alpha\geqslant 1$, we have seen that

$$h^{(2n+2)}\geqslant\sum_{k=0}^{n}\frac{1}{k!}\left(\frac{x^\alpha}{2\alpha}\right)^k,\qquad n=0,1,2,3.$$

This leads to the lower estimate of h: $h(x)\geqslant\exp\frac{x^\alpha}{2\alpha}$. We are now going to show that the equality sign here holds.

In general, in order to check that $F^*=\begin{pmatrix}h\\ e^{C}h'\end{pmatrix}$, it is easier to check (10). With $h=\exp\psi$, from equation (10), it follows that

$$a(\psi''+\psi'^2)+b\psi'=c,$$

or equivalently,

$$\psi''+\psi'^2+\frac{b}{a}\psi'=\frac{c}{a}. \tag{13}$$

In the present case, it is simply

$$\psi''(x)+\psi'(x)^2=\frac{1}{4}x^{2\alpha-2}+\frac{\alpha-1}{2}x^{\alpha-2}=\left(\frac{x^{\alpha-1}}{2}\right)^2+\frac{\alpha-1}{2}x^{\alpha-2}.$$

From this, we obtain $\psi'(x)=x^{\alpha-1}/2$ and then $\psi(x)=x^\alpha/(2\alpha)$. Having h at hand, the assertion of the lemma follows from Theorem 7.1. Since h is

increasing, $\mu(h^2) = \infty$, we need only parts (1) and (3) of Theorem 7.1. The details are delayed since this one is actually a particular case of Example 7.10 (2) with $b = 0$.

We remark that the precise leading order of h at infinity is required for our purpose, the natural lower estimate $h \geqslant h^{(n)}$ for fixed n is usually not enough. Nevertheless, the successive approximation schemes are still effective to provide practical lower bound of h. An upper bound of h is often easier to obtain by using (12). We also remark that the simplest way to prove Example 7.6 is using the following Molchanov's criterion ([11], see also [8; page 90, Theorem 6]): if $b = 0$, $a = 1$, and c is lower bounded, then $\sigma_{\rm ess}(L^c_{\rm min}) = \varnothing$ iff

$$\text{for each } \theta > 0, \qquad \int_x^{x+\theta} c \to \infty \quad \text{as } x \to \infty.$$

From this remark, it should be clear that there is quite a distance from the last special case to our general setup.

With a little modification of the proof for the last example in the case of $\alpha = 2$, it follows that $h(x) := \exp[x^2/2]$ is harmonic of the following operator

$$L^c = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - c(x), \quad c(x) := x^2 + 1.$$

Then, by using a shift, we obtain the following result.

Example 7.7 The one-dimensional harmonic oscillator

$$L^c = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - c(x), \quad c(x) := x^2$$

has discrete spectrum.

Actually, it is known that the eigenvalues of the last operator $-L^c$ are simple: $\lambda_n = 2n + 1$, $n = 0, 1, \ldots$ with eigenfunction

$$g_n(x) = (-1)^n e^{x^2/2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-x^2}, \qquad n = 0, 1, \ldots,$$

respectively. By symmetry, the conclusion holds not only on the half-line but also on the whole line.

Example 7.8 Let $E = (0, \infty)$, $\gamma \geqslant 10/9$, and

$$L^c = (1+x)^\gamma \frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{4\gamma}{5}(1+x)^{\gamma-1} \frac{\mathrm{d}}{\mathrm{d}x} - c(x), \quad c(x) := \frac{\gamma(9\gamma - 10)}{100}(1+x)^{\gamma-2}.$$

Then $\sigma_{\rm ess}(L^c_{\rm min}) = \varnothing$ if $\gamma > 2$ and $\sigma_{\rm ess}(L^c_{\rm min}) \neq \varnothing$ if $\gamma \in [10/9, 2]$.

Proof. We remark that condition $\gamma \geqslant 10/9$ is for $c(x) \geqslant 0$.

First, we look for the L^c-harmonic function h having the form $h = \exp\psi$ for some ψ. Then, by (13), we have

$$(1+x)^2(\psi'' + \psi'^2) + \frac{4\gamma}{5}(1+x)\psi' = \frac{\gamma(9\gamma - 10)}{100}.$$

This equation suggests us first that $\psi' = \beta(1+x)^{-1}$ for some constant β, and then $\beta = \gamma/10$. Hence, we obtain $h(x) = (1+x)^{\beta}$.

Next, we have

$$\begin{aligned} &C(x) = \int_0^x \frac{b}{a} = \frac{4\gamma}{5}\log(1+x), \qquad e^{C(x)} = (1+x)^{4\gamma/5}, \\ &\mu(\mathrm{d}x) = (1+x)^{-\gamma/5}\mathrm{d}x, \qquad \hat{\nu}(\mathrm{d}x) = (1+x)^{-4\gamma/5}\mathrm{d}x, \\ &\mu\big(h^2\mathbb{1}_{(0,x)}\big) = x, \qquad \hat{\nu}\big(h^{-2}\mathbb{1}_{(x,\infty)}\big) = \frac{x}{(1+x)^{\gamma}}. \end{aligned}$$

Therefore, $\mu\big(h^2\mathbb{1}_{(0,x)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(x,\infty)}\big) \sim x^{2-\gamma}$ as $x \to \infty$. The result now follows from Theorem 7.1 (1). □

Actually, this example is a special case of Example 7.11 (2).

Up to now, we have studied in one direction: reducing the case that $c(x) \not\equiv 0$ to the one $c(x) \equiv 0$. Certainly, we can go to the opposite direction: extending the result from $c(x) \equiv 0$ to $c(x) \not\equiv 0$. This is actually much easier but is very powerful. For simplicity, we restrict ourselves to the special case that $h > 0$. Then one may write $h = \exp\psi$ for some ψ. This leads to the next result which is a special case of [5; Corollary 3.7].

Corollary 7.9 Given

$$\widetilde{L} = \tilde{a}(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \tilde{b}(x)\frac{\mathrm{d}}{\mathrm{d}x} \quad \text{with domain } \mathscr{D}\big(\widetilde{L}\big),\ \tilde{a}(x) > 0$$

and $\psi \in \mathscr{C}^2(E)\,(E \subset \mathbb{R})$, define

$$\begin{aligned} L &= \widetilde{L} - 2\tilde{a}\psi'\frac{\mathrm{d}}{\mathrm{d}x} + \Big[\tilde{a}\psi'^2 - \widetilde{L}\psi\Big] \\ &= \tilde{a}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \big[\tilde{b} - 2\tilde{a}\psi'\big]\frac{\mathrm{d}}{\mathrm{d}x} + \Big[\tilde{a}\psi'^2 - \tilde{a}\psi'' - \tilde{b}\psi'\Big], \\ \mathscr{D}(L) &= \big\{f\exp[-\psi] \in L^2(\tilde{\mu}) : f\exp[-\psi] \in \mathscr{D}\big(\widetilde{L}\big)\big\}. \end{aligned} \tag{14}$$

Then $(L, \mathscr{D}(L))$ and $\big(\widetilde{L}, \mathscr{D}\big(\widetilde{L}\big)\big)$ are isospectral $\big($in particular, $\sigma_{\mathrm{ess}}(L) = \sigma_{\mathrm{ess}}\big(\widetilde{L}\big)\big)$. Furthermore, if we replace ψ' by

$$\psi' = \frac{\tilde{b} - b}{2\tilde{a}} \ \text{ for varying } b \tag{15}$$

$\big($assuming $\tilde{a}, \tilde{b}, b \in \mathscr{C}^1(E)\big)$, then the operator L becomes

$$L^b = \tilde{a}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2}\bigg[\frac{b^2 - \tilde{b}^2}{2\tilde{a}} - \tilde{a}\frac{\mathrm{d}}{\mathrm{d}x}\bigg(\frac{\tilde{b} - b}{\tilde{a}}\bigg)\bigg].$$

Corresponding to $\mathscr{D}_{\max}\big(\widetilde{L}\big)$, we have $\mathscr{D}_{\max}(L)$ defined by (14) in terms of ψ. Then, we have $L_{\max}$. Furthermore, we have $L^b_{\max}$ in terms of (15). Similarly, corresponding to $\mathscr{D}_{\min}\big(\widetilde{L}\big)$, we have $\mathscr{D}_{\min}(L)$, $L_{\min}$, and $L^b_{\min}$, respectively.

Example 7.10 Let $E = (0, \infty)$, $\alpha > 0$ and $b \in \mathscr{C}(E)$.

(1) Define

$$\widetilde{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - x^{\alpha-1}\frac{\mathrm{d}}{\mathrm{d}x},$$
$$L^b = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2}\bigg[\frac{1}{2}b(x)^2 + b'(x) - \bigg(\frac{1}{2}x^\alpha - \alpha + 1\bigg)x^{\alpha-2}\bigg].$$

Then for each b, $L^b_{\max}$ and $\widetilde{L}_{\max}$ are isospectral, $\sigma_{\mathrm{ess}}\big(L^b_{\max}\big) = \varnothing$ if $\alpha > 1$ and $\sigma_{\mathrm{ess}}\big(L^b_{\max}\big) \neq \varnothing$ if $\alpha \in (0, 1]$.

(2) Define

$$\widetilde{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^{\alpha-1}\frac{\mathrm{d}}{\mathrm{d}x},$$
$$L^b = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2}\bigg[\frac{1}{2}b(x)^2 + b'(x) - \bigg(\frac{1}{2}x^\alpha + \alpha - 1\bigg)x^{\alpha-2}\bigg].$$

Then for each b, $L^b_{\min}$ and $\widetilde{L}_{\min}$ are isospectral, $\sigma_{\mathrm{ess}}\big(L^b_{\min}\big) = \varnothing$ if $\alpha > 1$ and $\sigma_{\mathrm{ess}}\big(L^b_{\min}\big) \neq \varnothing$ if $\alpha \in (0, 1]$.

Proof. Note that $\hat{\nu}(E) = \infty$. By [10; Example 4.1], for the operator

$$L_0 = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - x^{\alpha-1}\frac{\mathrm{d}}{\mathrm{d}x} \quad \text{on} \quad (0, \infty)$$

with the maximal domain, we have $\sigma_{\mathrm{ess}}(L_0) = \varnothing$ if $\alpha > 1$ and $\sigma_{\mathrm{ess}}(L_0) \neq \varnothing$ if $\alpha \in (0, 1]$. Actually,

$$C(x) = -\int_0^x x^{\alpha-1} \sim -x^\alpha,\ \mu(x, \infty) = \int_x^\infty e^{C(x)} \sim x^{1-\alpha}e^{-x^\alpha},\ \hat{\nu}(0, x) = \int_0^x e^{-C(x)} \sim x^{1-\alpha}e^{x^\alpha}$$

as $x \to \infty$. Hence $\hat{\nu}(0, x)\mu(x, \infty) \sim x^{2(1-\alpha)}$ as $x \to \infty$. The required conclusion now follows from Theorem 7.1 (2) with $h = 1$. Then, by Corollary 7.9, we obtain part (1).

To prove part (2), recall that for the differential operator

$$L = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x},$$

as an analog of the duality for birth–death processes used in Section 3 (part (d)), its dual operator $\widehat{L}$ takes the following form:

$$\widehat{L} = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \left(\frac{\mathrm{d}}{\mathrm{d}x}a(x) - b(x)\right)\frac{\mathrm{d}}{\mathrm{d}x}$$

(cf. [3; (10.6)] or [4; §3.2]). Hence, the operator

$$\widetilde{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^{\alpha-1}\frac{\mathrm{d}}{\mathrm{d}x}$$

with the minimal domain is a dual of L_0 with the maximal domain (cf. [3; (10.6)]) and so they have the same spectrum. Thus, part (2) follows again from Corollary 7.9. □

Example 7.11 Let $E = (0, \infty)$, $\gamma > 1$ and $b \in \mathscr{C}(E)$.

(1) Define

$$\widetilde{L} = (1+x)^\gamma \frac{\mathrm{d}^2}{\mathrm{d}x^2},$$
$$L^b = (1+x)^\gamma \frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2}\left[\frac{b(x)^2}{2(1+x)^\gamma} + b'(x) - \frac{\gamma b(x)}{1+x}\right].$$

Then for each b, $L^b_{\max}$ and $\widetilde{L}_{\max}$ are isospectral, $\sigma_{\mathrm{ess}}\big(L^b_{\max}\big) = \varnothing$ if $\gamma > 2$ and $\sigma_{\mathrm{ess}}\big(L^b_{\max}\big) \neq \varnothing$ if $\gamma \in (1, 2]$.

(2) Define

$$\widetilde{L} = (1+x)^\gamma \frac{\mathrm{d}^2}{\mathrm{d}x^2} + \gamma(1+x)^{\gamma-1}\frac{\mathrm{d}}{\mathrm{d}x},$$
$$L^b = (1+x)^\gamma \frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{2}\left[\frac{b(x)^2}{2(1+x)^\gamma} - \frac{\gamma b(x)}{1+x} + b'(x) - \gamma\left(\frac{\gamma}{2} - 1\right)(1+x)^{\gamma-2}\right].$$

Then for each b, $L^b_{\min}$ and $\widetilde{L}_{\min}$ are isospectral, $\sigma_{\mathrm{ess}}\big(L^b_{\min}\big) = \varnothing$ if $\gamma > 2$ and $\sigma_{\mathrm{ess}}\big(L^b_{\min}\big) \neq \varnothing$ if $\gamma \in (1, 2]$.

Proof. As in the proof of Example 7.10, it suffices to study the spectrum of the operator

$$L_0 = (1+x)^\gamma \frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

with the maximal domain. Clearly,

$$\mu(\mathrm{d}x) = (1+x)^{-\gamma}\mathrm{d}x, \quad \hat{\nu}(\mathrm{d}x) = \mathrm{d}x, \quad \hat{\nu}(0,\infty) = \infty, \quad \mu(E) < \infty \text{ if } \gamma > 1.$$

We are in the case of Lemma 5.8 (1): $\hat{\nu}(0,x)\mu(x,\infty) \sim x^{2-\gamma}$ as $x \to \infty$. Hence, by Theorem 7.1 (2) with $h = 1$, L_0 has discrete spectrum if $\gamma > 2$ and otherwise, if $\gamma \in (1, 2]$. □

Remark 7.12 The condition $c(x) \geqslant 0$ used in the paper has some probabilistic meaning (killing rate), but it is not necessary, as we have seen from Example 7.10 (2) with $b(x) \equiv 0$ and $\alpha \in (0,1)$. Everything should be the same if c is lower bounded which can be reduced to the nonnegative case by using a shift. The last 'bounded below' condition is still not necessary, refer to [5].

Up to now, we have studied the half-space only. The case of whole line is in parallel. To see this, fix the reference point $\theta = 0$ and use the measures μ and $\hat{\nu}$ defined at the beginning of this section. For simplicity, here we write down only the symmetric case (which means that μ are finite or not simultaneously on $(-\infty, 0)$ and $(0, \infty)$, and similarly for $\hat{\nu}$). The other cases may be handled in parallel.

Theorem 7.13 Let $E = \mathbb{R}$ and $h \neq 0$-a.e. be an L^c-harmonic function constructed in Theorem 7.4.

(1) If $\hat{\nu}\big(h^{-2}\big) < \infty$, then $\sigma_{\rm ess}(L^c_{\rm min}) = \varnothing$ iff

$$\lim_{x\to\infty}\Big[\mu\big(h^2\mathbb{1}_{(0,x)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(x,\infty)}\big) + \mu\big(h^2\mathbb{1}_{(-x,0)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(-\infty,-x)}\big)\Big] = 0.$$

(2) If $\mu\big(h^2\big) < \infty$, then $\sigma_{\rm ess}(L^c_{\rm max}) = \varnothing$ iff

$$\lim_{x\to\infty}\Big[\mu\big(h^2\mathbb{1}_{(x,\infty)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(0,x)}\big) + \mu\big(h^2\mathbb{1}_{(-\infty,-x)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(-x,0)}\big)\Big] = 0.$$

(3) If $\hat{\nu}\big(h^{-2}\mathbb{1}_{(-\infty,0)}\big) = \hat{\nu}\big(h^{-2}\mathbb{1}_{(0,\infty)}\big) = \infty = \mu\big(h^2\mathbb{1}_{(-\infty,0)}\big) = \mu\big(h^2\mathbb{1}_{(0,\infty)}\big)$, then $\sigma_{\rm ess}(L^c_{\rm min}) = \sigma_{\rm ess}(L^c_{\rm max}) \neq \varnothing$.

Proof. (a) As in the proof of Theorem 2.1, by [5; Theorem 3.1], it suffices to consider only the case that $c(x) \equiv 0$.

(b) Part (2) of the theorem follows from [10; Theorem 2.5 (2)].

(c) Part (1) of the theorem is a dual of part (2). Refer to [3; (10.6)] or [4; §3.2].

(d) In the present symmetric case, part (3) is obvious in view of Theorem 7.1 (3). □

The approach used in this paper is meaningful in a quite general setup. For instance, one may refer to [5] for some isospectral operators in higher dimensions.

Acknowledgments. The author thanks S. Kotani for introducing [7] and [9] to him and R. Oĭnarov for sending him the original version of [12]. Thanks are also given to H.J. Zhang and Z.W. Liao for their corrections of an earlier version of the paper. Research supported in part by the National Natural Science Foundation of China (No. 11131003), the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Ahlbrandt, C.D., Hinton, D.B. and Lewis, R.T. (1981). *Necessary and sufficient conditions for the discreteness of the spectrum of certain singular differential operators*. Canad. J. Math. 33, 229–246.

[2] Chen, M.F. (2004). *From Markov Chains to Non-equilibrium Particle Systems* (1st edn., 1992), 2nd edn. World Scientificm Singapore.

[3] Chen, M.F. (2010). *Speed of stability for birth–death processes*. Front. Math. China 5(3), 379–515.

[4] Chen, M.F. (2011). *Basic estimates of stability rate for one-dimensional diffusions*. Chapter 6 in "Probability Approximations and Beyond", pp. 75–99. Lecture Notes in Statistics 205.

[5] Chen, M.F. and Zhang, X. (2014). *Isospectral operators*. Commun. Math. Stat. 2, 17–32.

[6] Chen, M.F. and Zhang, Y.H. (2014). *Unified representation of formulas for single birth processes*. Front. Math. China 9(4), 761–796.

[7] Ćurgus, B. and Read, T.T. (2002). *Discreteness of the spectrum of second-Order differential operators and associated embedding theorems*. J. Differential Equations 184, 526–548.

[8] Glazman, I.M. (1965). *Direct Methods of Qualitative Spectral Analysis of Singular Differential Operators*. Israel Program for Scientific Translations, Jerusalem.

[9] Kac, I.S. and Kreň, M.G. (1958). *Criteria for the discreteness of the spectrum of a singular string* (in Russian). Izv. Vysš. Učebn. Zaved. Matematika 2(3), 136–153.

[10] Mao, Y.H. (2006). *On empty essential spectrum for Markov processes in dimension one*. Acta Math. Sin. Eng. Ser. 22(3), 807–812.

[11] Molchanov, A.M. (1953). *The conditions for the discreteness of the spectrum of self-adjoint second order differential equations*. Trudy Moskov. Mat. Obshch. 2, 169–200 (In Russian).

[12] Oĭnarov, R. and Otelbaev, M. (1988). *A criterion for the discreteness of the spectrum of the general Sturm-Liouville operator, and embedding theorems connected with it* (in Russian), Differentsial'nye Uravneniya 24(4), 584–591; translation in Differential Equations 24 (4), 402–408.

[13] van Doorn, E.A. (2014) *Spectral properties of birth-death polynomials*. Preprint.

[14] Zettl, A. (2005). *Sturm–Liouville Theory*. AMS, Providence, Rhode Island.

Mu-Fa Chen

School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.

E-mail: mfchen@bnu.edu.cn

Home page: http://math.bnu.edu.cn/~chenmf/main_eng.htm

应用概率统计 第三十一卷
第二期 2015年4月

Chinese Journal of Applied Probability
and Statistics Vol.31 No.2 Apr. 2015

p.213--224

综述报告

Practical Criterion for Uniqueness of Q-Processes *

CHEN MUFA

(*School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing, 100875*)

Abstract

The note begins with a short story on seeking for a practical sufficiency theorem for the uniqueness of time-continuous Markov jump processes, starting around 1977. The general result was obtained in 1985 for the processes with general state spaces. To see the sufficient conditions are sharp, a dual criterion for non-uniqueness was obtained in 1991. This note is restricted however to the discrete state space (then the processes are called Q-processes or Markov chains), for which the sufficient conditions just mentioned are showing at the end of the note to be necessary. Some examples are included to illustrate that the sufficient conditions either for uniqueness or for non-uniqueness are not only powerful but also sharp.

Keywords: Criterion, uniqueness, Markov chain, Markov jump process.

AMS Subject Classification: 60J27.

Let E be a countable set with elements $i, j, k, \ldots$. A matrix $Q = (q_{ij} : i, j \in E)$ is called a Q-matrix if its non-diagonals are nonnegative and $\sum\limits_{j \in E} q_{ij} \leq 0$ for every $i \in E$. Throughout this note, we restrict ourselves to the special case that the Q-matrix is totally stable $q_i := -q_{ii} < \infty$ and conservative $q_i = \sum\limits_{j \neq i} q_{ij}$ for every $i \in E$. It is called bounded if $\sup\limits_{i \in E} q_i < \infty$. For a given Q-matrix $Q = (q_{ij})$ on E, a sub-Markovian semigroup $\{P(t) = (p_{ij}(t) : i, j \in E)\}_{t \geq 0}$ is called a Q-process if

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t)\Big|_{t=0} = Q \qquad \text{(pointwise)}.$$

*The research was supported in part by the National Natural Science Foundation of China (11131003), the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Received January 15, 2015.

doi: 10.3969/j.issn.1001-4268.2015.02.010

The Q-processes may not be unique in general, but there always exists the minimal one, due to Feller (1940, Theorem 1), denoted by $P^{\min}(t) = (p_{ij}^{\min}(t) : i, j \in E)$. For more than half-century ago, some criteria for the uniqueness were known.

Theorem 1 The Q-process is unique (equivalently, the minimal process $P^{\min}(t)$ is not explosive) iff one of the following equivalent conditions holds:

(C1) $\sum\limits_{j\in E} p_{ij}^{\min}(t) = 1$ for every $i \in E$ and $t \geq 0$.

(C2) $\sum\limits_{n=1}^{\infty} q_{X^{\min}(\tau_n)}^{-1} = \infty$, P_i-a.s., where τ_n is the nth jump time of the minimal process $\{X^{\min}(t) : t \geq 0\}$ corresponding to $P^{\min}(t)$.

(C3) The equation

$$(\lambda I - Q)u = 0, \qquad 0 \leq u \leq 1, \tag{1}$$

has only zero solution for some (equivalently, for all) $\lambda > 0$.

Criterion (C1) goes back to Feller (1940). Criterion (C2) is due to Dobrushin (1952). Criterion (C3) is due to Feller (1957) and Reuter (1957). Refer also to Chung (1967; Part II, § 19, Theorem 1), or Gikhman and Skorokhod (1975; Chap. 3, § 2, Theorems 3 and 4).

The earlier Criterion (C1) often requires a further effort in practice, rather than a direct application. In particular, the proof of the powerful sufficiency theorem (Theorem 2 below) is based on it.

Criterion (C2) is effective in some cases. For instance in the simplest case that $M := \sup\limits_{i\in E} q_i < \infty$, since

$$\sum_{n=1}^{\infty} q_{X^{\min}(\tau_n)}^{-1} \geq \sum_{n=1}^{\infty} M^{-1} = \infty,$$

we obtain the uniqueness of the processes. For pure birth process (i.e., $q_{i,i+1} > 0$ and $q_{ij} = 0$ for all $j \neq i$, $i, j \geq 0$), Criterion (C2) says that the process is unique iff

$$\sum_{n=1}^{\infty} \frac{1}{q_{n,n+1}} = \infty. \tag{2}$$

Besides, if the minimal process is recurrent, then the term q_k^{-1} will appears infinitely often in the summation, hence the process should be unique according to the criterion.

Criterion (C3) is more effective once equation (1) is solvable. More precisely, it is the case if the exit boundary consists at most a single point, for instance the pure birth processes, the birth-death processes or more general the single birth processes (i.e., for $j > i \geq 0$, $q_{ij} > 0$ iff $j = i + 1$; for $0 \leq j < i$, q_{ij} is nonnegative but free). We will come back this story soon.

However, the next model stopped our study for several years at the beginning of the study (1977–1978) on non-equilibrium particle systems. To state our model, we use operator Ω instead of the matrix Q:

$$\Omega f(i) = \sum_{j\in E} q_{ij}(f_j - f_i), \qquad i \in E.$$

Of course, in this case, $\Omega f = Qf$. For a Markov chain on a countable set E, by a transform, one often assumes that E is simply the set $\mathbb{Z}_+ = \{0, 1, \ldots\}$. However, such a transform ignores the original geometry of E and may not be convenient in multidimensional case. To state our model, we need some notation. Let $i = (i_u : u \in S)$ and define its updates $i^{u\pm}$ and $i^{u,v}$ as follows:

$$i_w^{u\pm} = \begin{cases} i_u \pm 1 & w = u; \\ i_w & w \neq u, \end{cases} \qquad i_w^{u,v} = \begin{cases} i_u - 1 & w = u; \\ i_v + 1 & w = v; \\ i_w & w \neq u, v, \end{cases} \qquad w \in S.$$

Example 1 (Schlögl's second model) Let S be a finite set and $E = \mathbb{Z}_+^S$. Define a Markov chain on E with operator

$$\begin{aligned} \Omega f(i) = &\sum_{u\in S} \{b(i_u)[f(i^{u+}) - f(i)] + a(i_u)[f(i^{u-}) - f(i)]\} \\ &+ \sum_{u,v} i_u p(u,v)[f(i^{u,v}) - f(i)], \qquad i = (i_u : u \in S) \in E, \end{aligned}$$

where $(p(u,v) : u, v \in S)$ is a "simple" random walk on S, and

$$\begin{aligned} b(k) &= \beta_0 + \beta_2 k(k-1), \qquad \beta_0,\, \beta_2 > 0, \\ a(k) &= \delta_1 k + \delta_3 k(k-1)(k-2), \qquad \delta_1,\, \delta_3 > 0. \end{aligned}$$

Here in the first sum of Ω, in each vessel u, there is a birth-death process with birth rate $b(k)$ and death rate $a(k)$, respectively. This is called the reaction part of the model. The reactions in different vessels are independent. In the second sum of Ω, a particle from vessel u moves to vessel v. This is called the diffusion part of the model. Thus, it is actually a finite-dimensional reaction-diffusion processes. Replacing the finite S with $S = \mathbb{Z}^d$, we obtain formally an operator of infinite-dimensional reaction-diffusion process which is a typical model from the non-equilibrium statistical physics. Even though the large systems are quite popular today, in that period, it was rather unusual to study such a non-equilibrium system. Our original program is to rebuild the mathematical ground of non-equilibrium statistical physics (cf. Chen (2004; Part IV). An earlier paper on this topic appeared in 1985 (see Chen, 1985)). For this, the model is meaningful only if it

is ergodic in every finite dimension. Thus, the finite dimensional model consists the first doorsill of our program.

In 1983, the author and Yan (see Yan and Chen, 1986), using a comparison technique, overcame this doorsill, based on a systemic study on the single birth processes. To which, we obtained explicit criteria not only for uniqueness but also for ergodicity and so on. This goes back to Yan and Chen (1986), Chen (1986a). Refer to Chen (2004) for updates and to Chen and Zhang (2014) for a unified treatment. After two more years, using an approximating approach, we obtained a powerful sufficiency theorem as stated below.

Theorem 2 (Uniqueness criterion) Let $Q = (q_{ij})$ be a Q-matrix on a countable set E. Then the corresponding Q-process is unique iff the following two conditions hold simultaneously.

(U1) There exist $E_n \uparrow E$ as $n \uparrow \infty$ and a nonnegative function φ such that $\sup\limits_{i\in E_n} q_i < \infty$ and $\lim\limits_{n\to\infty} \inf\limits_{i\notin E_n} \varphi_i = \infty$.

(U2) There exists a constant $c \in \mathbb{R}$ such that $Q\varphi \le c\varphi$.

Certainly, for Schlögl's model for instance, in condition (U2), it is more convenient to use $\Omega\varphi$ instead of $Q\varphi$. Besides, an important fact should be very helpful in practice: if φ satisfies the conditions with $c \ge 0$, then so does $M + \varphi$ for every constant $M \ge 0$. In particular, a local modification of Q does not interfere the conclusion.

From Chen (2004; Parts I and II), it is now clear that a large part of the theory of Q-processes can be generalized to the so-called Markov jump processes on general state space. To save the space, we will not really go to the last subject but it is worth to mention the extension. We now use the codes "GS" and "DS" to distinguish the "general state space" and the "discrete state space", respectively. The sufficient part of the last theorem first appeared in Chen (1986a; Theorem 2.37 (GS)) and Chen (1986b; Theorem (16) (GS)). Because it is regarded as one of the author's favourite contributions to the theory of Markov jump processes, this result was then introduced several times in the author's publications: Chen (1991; Theorem 1.11 (DS)), Chen and Yan (1991; Theorem 3.9 (GS)), Chen (2004; Theorem 2.25 (GS)), Chen (1997; Theorem 2.1 (DS)), Chen (2005; Theorem 9.4 (DS)), and Chen and Mao (2007; Theorem 2.9 (DS)).

Theorem 2 is often accompanied in the publications just listed by the next simpler result.

Corollary 1 Suppose that there exist a function $\varphi \ge q$ and a constant $c \in \mathbb{R}$ such that $Q\varphi \le c\varphi$ on E. Then the Q-process is unique.

Proof Set $E_n = \{i \in E : q_i \le n\}$. If $M := \sup\limits_{i\in E} q_i < \infty$, then for large enough n,

we have $E_n = E$ and so $\inf_{k\notin E_n} q_k = \infty$ by standard convention $\inf_{\emptyset} \varphi = \infty$. In this case, condition (U2) is trivial with $\varphi = 1 + M$. If $M = \infty$, then $\inf_{k\notin E_n} q_k \geq n \to \infty$ as $n \to \infty$. Combining this with (U2), the conclusion follows from Theorem 2. □

Corollary 1 is almost explicit since one can simply specify $\varphi = 1 + q$. This enables us to use it easier in practice. However, such a specification makes the assumption becomes a little stronger. We will come back this point later.

Let us make some remarks about the conditions in Theorem 2. Condition (U2) is a relax of the equation in (1): finding a solution to an inequality is easier than finding a solution to the corresponding equality. Criterion (C3) says that there is only trivial bounded solution to the equation (1). Conversely, if a solution of the equation is fixed at some point, say θ, such that $\varphi_\theta = 1$, then the solution φ should be unbounded. This leads to the condition $\lim_{n\to\infty} \inf_{k\notin E_n} \varphi_k = \infty$ in (U1). Using this idea, we prove that the assumptions in Theorem 2 are necessary for single birth processes (see Chen, 2004; Remark 3.20). The reason we allow some subset of E_n to be infinite is to rule out some region of E, on which $\sup_{i\in E_n} q_i < \infty$. The key in the proof of this result is an economic approximation by bounded Q-processes. Certainly, the necessity shows that the assumptions of the theorem are sharp, and is valuable as illustrated by Chen (1986b; Theorem (25)). However, it does not mean that the inverse of the conditions can be used in practice to show the non-uniqueness of the processes. Hence, we went to an opposite way proving the following criterion (see Chen, 2004; Theorem 2.27 (GS), its proof in 2nd edition uses Lemma ~~5.18~~ 5.17 rather than Lemma 5.15).

Theorem 3 (Non-uniqueness criterion) For a given Q-matrix Q on a countable set E, the Q-processes are not unique if for some (equivalently, for all) $c > 0$, there is a bounded function φ with $\sup_{k\in E} \varphi_k > 0$ such that $Q\varphi \geq c\varphi$. Conversely, these conditions plus $\varphi \geq 0$ are also necessary.

We remark that three results (Theorems 2, 3 and Corollary 1), we have talked so far are specialized from their original case in GS to the one in DS. Theorems 2 and 3 are somehow the extensions of Criterion (C3) in two opposite directions. As we will see soon that the extended theorems are much effective than the original Criterion (C3). Using two opposite sufficiency results instead of a single criterion is often meaningful. For instance, for recurrence, we have a criterion (see Chen, 2004; Proposition 4.21) which is accompanied with more practical criteria (see Chen, 2004; Theorems 4.24 and 4.25) for the recurrence and transiency, respectively. As a companion to Chen (2004; Theorem 4.25), refer to Meyn and Tweedie (2009; Theorem 8.0.2) and Hairer (2010; Proposition 1.3) or more recent criteria. Next, for ergodicity and nonergodicity, refer to Chen (2004;

Theorem 4.45 (1)) and Kim and Lee (2008; Theorem 1), respectively. For various stability speeds/principal eigenvalues, in Chen (2005), we have not only the classical variational formula, but also dual variational formulas to describe their lower and upper bounds, respectively.

It is interesting that there is now a direct way to prove the necessity of Theorem 2 in the context of DS based on a recent result by Spieksma (2014).

Theorem 4 Everything is the same as in Theorem 2 except (U1) is replaced by

(U1)′ In the original (U1), assume in addition that each E_n is finite and ignore "$\sup_{i\in E_n} q_i < \infty$".

It is now the position to illustrate by examples the power of our results and compare conditions (U1) and (U1)′.

The next two examples show that in Theorem 2, the condition "$\lim_{n\to\infty} \varphi_n = \infty$" is not necessary, which is however necessary in a criterion for recurrence used in the proof of Theorem 4 (see its proof below).

Example 2 Let E be a countable set and $Q = (q_{ij})$ be a bounded conservative Q-matrix on E. Then assumptions of Theorem 2 hold but its test function φ can be bounded.

Proof (a) Simply set $E_n \equiv E$ (may be infinite) for every $n \geq 1$ and $\varphi_i \equiv 1$. Then it is obvious that $0 = Q\varphi \leq \varphi$ and $\lim_n \inf_{i\notin E_n} \varphi_i = \infty$ since $\inf_{\emptyset} \varphi = \infty$ by the standard convention. Hence by Theorem 2, the process is unique. As we have seen before, Corollary 1 is also applicable in such a trivial case.

(b) Knowing that the process is unique, then by Theorem 4, there should exist a φ satisfying (U1)′, as well as (U2). The problem is that the resulting φ is not explicitly known when E is infinite. In this sense, Theorem 4 is theoretic correct but not practical in such simplest case. □

Example 3 Let $E = \mathbb{Z}_+$ and $Q^{(1)}$ be a bounded conservative Q-matrix on E. Denote its test function by $\varphi^{(1)} \equiv 1$ as in the last example. Next, let $Q^{(2)}$ be a conservative Q-matrix on E satisfying the assumptions of Theorem 2 with a sequence of finite subsets $\{E_n\}_{n\geq 1}$ and a test function $\varphi^{(2)}$. Finally, we construct a new Q as follows: on the odd numbers in E, we use the transition mechanism of $Q^{(1)}$, and on the even numbers in E, we adopt the one of $Q^{(2)}$. Define $\varphi = \varphi^{(1)}$ on the odd numbers and $\varphi = \varphi^{(2)}$ on the even numbers. Then the assumptions of Theorem 2 hold but its test function φ_n has no limit as $n \to \infty$: $\overline{\lim}_{n\to\infty} \varphi_n = \infty$ and $\underline{\lim}_{n\to\infty} \varphi_n = 1$.

Proof First, note that for the original $Q^{(2)}$ on E, because each E_n is a finite subset

of E, the condition $\lim\limits_{n\to\infty}\inf\limits_{k\notin E_n}\varphi_k^{(2)}=\infty$ is equivalent to $\lim\limits_{n\to\infty}\varphi_n^{(2)}=\infty$. Therefore, we have

$$\overline{\lim_{n\to\infty}}\,\varphi_n=\lim_{n\to\infty}\varphi_n^{(2)}=\infty,\qquad \underline{\lim_{n\to\infty}}\,\varphi_n=\lim_{n\to\infty}\varphi_n^{(1)}=1.$$

To show the assumptions in Theorem 2 hold, simply let $E_0=\{\text{odd integers}\}$, and let $E_n\,(n\geq 1)$ be the union of E_0 and the natural modification of the original E_n used for $Q^{(2)}$. Then the resulting $E_n\uparrow E$ as $n\to\infty$, $\sup\limits_{k\in E_n}q_k<\infty$ for each $n\geq 0$, and

$$\lim_{n\to\infty}\inf_{k\notin E_n}\varphi_k=\lim_{n\to\infty}\inf_{k\notin E_n}\varphi_k^{(2)}=\lim_{n\to\infty}\varphi_n^{(2)}=\infty.$$

Finally, because of the independence of $Q^{(1)}$ and $Q^{(2)}$, $\varphi^{(1)}$ and $\varphi^{(2)}$, the condition $Q\varphi\leq\max\{c_2,1\}\varphi$ on the set of odd numbers follows from

$$Q^{(1)}\varphi^{(1)}\leq\varphi^{(1)}\qquad\text{on } E;$$

and the same condition on the set of even numbers follows from

$$Q^{(2)}\varphi^{(2)}\leq c_2\varphi^{(2)}\qquad\text{on } E.$$

We have thus obtained the required conclusion.

As mentioned in the last proof, in the present situation, we do not know how to use Theorem 4. □

Note that the last matrix Q is reducible. However, we can add a connection between 0 and 1 to produce an irreducible version of the example. This is not essential since a local modification does not interfere the uniqueness problem. Furthermore, one may replace the set {odd integers} or {even integers} by any infinite subset of E, but not E itself, the set of primer numbers for instance. The conclusion of Example 3 remains the same by an obvious modification.

The point is that some E_n is allowed to be infinite in (U1) but not in (U1)′.

Example 4 The pure birth process is unique iff (2) holds. In particular, set $q_{n,n+1}=$ the nth primer, then Theorem 2 is suitable but Corollary 1 fails.

Proof Note that $q_k=q_{k,k+1}$ for $k\geq 0$.

(a) If $\sum\limits_k q_k^{-1}=\infty$, set $E_n=\{0,1,\ldots,n\}$ and

$$\varphi_k=1+\sum_{1\leq j\leq k-1}\frac{1}{q_j}\to\infty\qquad\text{as } k\to\infty.$$

Then $Q\varphi\leq\varphi$ and so Theorem 2 gives us the uniqueness of the processes. In the particular case that $q_{n,n+1}=n+1$, the above φ has order $\log n$. However, we can also choose $\varphi_n=1+n$ and apply Theorem 2. This shows that there are some freedom in choosing φ.

(b) If $M := \sum_k q_k^{-1} < \infty$, set E_n as above and

$$\varphi_k = \frac{1}{2} + \sum_{1 \le j \le k-1} \frac{1}{q_j} - M \in \Big[\frac{1}{2} - M, \frac{1}{2}\Big].$$

Then $\sup_k \varphi_k = 1/2 > 0$, $Q\varphi \ge \varphi$, and so by Theorem 3, the processes are not unique. We remark that it would be awful to use the necessity in Theorems 2 or 4 to prove this non-uniqueness property.

(c) The last assertion is due to J.L. Zheng (cf. Chen, 1986a; Example 2.3.12; or Chen, 2004; Example 2.26). □

Proof of the uniqueness for Example 1 For $i \in E = \mathbb{Z}_+^S$, define its level by $|i| = \sum_{u \in S} i_u$ and set $E_n = \{i \in E : |i| \le n\}$ for $n \ge 1$.

(a) Next, define $\varphi(i) = 1 + |i|$. Then it is clear that $\lim_{n \to \infty} \inf_{k \notin E_n} \varphi(k) = \infty$. Because the diffusions do not change the levels, we have

$$\Omega\varphi(|i|) = \sum_{u \in S} [b(i_u) - a(i_u)] = \sum_{u \in S} \big[\alpha_0 - \alpha_1 i_u + \alpha_2 i_u^2 - \alpha_3 i_u^3\big]$$

for some positive $\{\alpha_k\}_{k=0}^3$. Next, since

$$\sum_{u \in S} i_u^2 \le |i|^2, \qquad \frac{1}{|S|} \sum_{u \in S} i_u^3 \ge \Big(\frac{|i|}{|S|}\Big)^3 \text{ (Jensen's inequality)},$$

where $|S|$ is the cardinality of S (finite but arbitrary), we have

$$\Omega\varphi(|i|) \le \alpha_0' - \alpha_1'|i| + \alpha_2'|i|^2 - \alpha_3'|i|^3$$

for some positive $\{\alpha_k'\}_{k=0}^3$. Now, because the right-hand side becomes negative for large enough $|i|$, it is clear that $\Omega\varphi(|i|) \le c\varphi(|i|)$ for every $i \in E$ and large enough c. The assertion now follows from Theorem 2. Hopefully, we have seen the role played by the geometry of E. The proof shows the power of our result. A good sufficiency result may be more effective than a criterion.

(b) It is also possible to use Corollary 1 to prove the required assertion, simply choose $\varphi(i) = \gamma\Big(1 + \sum_{u \in S} i_u^3\Big)$. First, choose γ large enough so that $\varphi \ge q$. Next, choose c large enough so that $\Omega\varphi \le c\varphi$. □

It is worthy to mention that in accompany to Theorem 2, we also have a similar, practical sufficiency result for (exponential) ergodicity. Refer to Chen (1989; Theorem 3 (GS)), Chen (1991; Theorem 1.18 (DS)), Chen (2004; Corollary 4.49 (DS) and Theorem 14.1 (GS)).

In the past nearly 30 years, Theorem 2 and Corollary 1 have very successful applications. A list of the literature was collected in Chen (2005; § 9.2). Certainly, the results used a lot by the author (in Chen (2004) for instance). In particular, it was used at the first step to construct a large class of infinite-dimensional processes (see Chen, 2004; § 13.2), 15 models are included in Chen (2004; § 13.4). Corollary 1 with some extension was used by Song (1988) in a quite earlier stage for Markov decision processes moving from bounded to unbounded situation. It is now quite often to see the influence of the study on Markov jump processes to the theory of Markov decision processes. Based on Chen (1986b), Theorem 2 was collected into Anderson (1991; Corollary 2.2.16), its originality was unfortunately ignored, even though the original paper (see Chen; 1986b) is included in the references of the book. For some corrections and comments on the last book, refer to Chen (1996). Very recently, Theorem 2 (GS) is applied by Chen and Ma (2014) to genetic study having continuous state space. Finally, we mention that the results have already extended to the time-inhomogeneous case by Zheng and Zheng (1987) and Zheng (1993) using the martingale approach.

Before going to the proofs, note that equation (1) is equivalent to

$$\Pi(\lambda)u = u, \qquad 0 \le u \le 1 \ \text{ on } \ E, \qquad \lambda > 0, \tag{3}$$

where

$$\Pi(\lambda) = \Big(\frac{(1-\delta_{ij})q_{ij}}{\lambda + q_i} : i, j \in E\Big).$$

Here the matrix $\Pi(\lambda)$ is sub-stochastic. We introduce a fictitious state Δ and define on the enlarged state space $E_\Delta = E \cup \{\Delta\}$ a new transition probability matrix

$$\Pi_{ij}^{\Delta}(\lambda) = \begin{cases} \Pi_{ij}(\lambda) & \text{if } i, j \in E; \\ \dfrac{\lambda}{\lambda + q_i} & \text{if } i \in E,\ j = \Delta; \\ p_j & \text{if } i = \Delta,\ j \in E, \end{cases}$$

where $(p_j : j \in E)$ is a positive probability measure on E. The enlarged transition probability matrix is irreducible even the original one may be not.

Lemma 1 The equation (1) has zero solution only iff so does the equation

$$\Pi^{\Delta}(\lambda)(u\mathbb{1}_E) = u, \qquad 0 \le u \le 1 \ \text{ on } \ E_\Delta, \qquad \lambda > 0. \tag{4}$$

Thus, the original Q-process is unique iff the $\Pi^{\Delta}(\lambda)$-chain is recurrent.

Proof Noting that $u_\Delta = \sum\limits_{k \in E} p_k u_k$, it is clear that $u_\Delta = 0$ iff $u_k = 0$ for all $k \in E$ since $p_k > 0$ for all $k \in E$. Equation (4) restricted to E coincides with (3) and then (1). This proves the first assertion.

To prove the second assertion, it suffices to note that $\Pi^{\Delta}(\lambda)$-chain is recurrent iff equation (4) has only trivial solution. The last result comes from Yan and Chen (1986), Chen (1986a; Lemma 12.1.27), or Chen (2004; Lemma 4.51). We remark here that the regularity assumption used in the cited references can be replaced by the minimal process, due to the equivalence of recurrence of the minimal process and its embedded chain. Refer to Chen (1986a; Lemma 12.3.1), or Chen (2004; Theorem 4.34). □

Proof of Theorem 4 When $|E| < \infty$, the conclusion is trivial and the assumptions hold for the specific $E_n \equiv E$ and $\varphi_i \equiv 1$ as seen from proof (a) of Example 2. Hence we may assume that $E = \mathbb{Z}_+$. Since each E_n is finite, the condition $\lim\limits_{n} \inf\limits_{k \notin E_n} \varphi_k = \infty$ becomes $\lim\limits_{n \to \infty} \varphi_n = \infty$. In this case, conditions (U1)′ and (U2) consist a criterion for the recurrence of the Markov chain $\Pi^{\Delta}(\lambda)$, refer to Chen (2004; Theorem 4.24) and its references within.

We remark that it is at this point, the finiteness of E_n is required and so the present sufficiency proof is not suitable for Theorem 2. At the moment, we do not know how to extend the necessity result of Theorem 4 from DS to GS.

Here is a part of an alternative proof given in Spieksma (2014). Let $P^{\min}(\lambda)$ be the Laplace transform of $P^{\min}(t)$. Using the second successive approximation scheme for the backward Kolmogorov equation (goes back to Feller (1940; Theorem 1)), we obtain

$$P^{\min}(\lambda) = \sum_{n=0}^{\infty} \Pi(\lambda)^n \mathrm{diag}\Big(\frac{1}{\lambda + q}\Big)$$

(cf. Chen, 2004, page 75, line -6). Hence

$$\lambda P^{\min}(\lambda) \text{ column } (1) = \sum_{n=0}^{\infty} \Pi(\lambda)^n \text{ column } \Big(\frac{\lambda}{\lambda + q}\Big).$$

The process is unique iff the left-hand side equals 1 at some/every $i \in E$, the right-hand side is the probabilistic decomposition of the time that the Markov chain $\Pi^{\Delta}(\lambda)$ starts from some $i \in E$, first visits Δ at some step $n \geq 1$, which equals 1 iff the irreducible Markov chain $\Pi^{\Delta}(\lambda)$ is recurrent. We have thus come back to the last lemma. □

Proof of Theorem 2 Here we adopt a circle argument.

(U1)′ + (U2) $\Longrightarrow$ (U1) + (U2). This is easy since (U1) is weaker that (U1)′.

(U1) + (U2) $\Longrightarrow$ uniqueness. This is the sufficiency part of Theorem 2 and was proved long time ago, even for GS.

Uniqueness $\Longrightarrow$ (U1)′ + (U2). This is the necessity part of Theorem 4. □

We remark that a similar phenomena is appeared in Theorem 3, the conditions for sufficiency are weaker than the ones for necessity. As we have seen from Example 4, this is very helpful in practice. However, these conditions are actually equivalent: conditions for necessity $\Longrightarrow$ conditions for sufficiency $\Longrightarrow$ non-uniqueness $\Longrightarrow$ conditions for necessity.

In view of these discussions, one may combine Theorems 2 and 4 into one having the style of Theorem 3.

In conclusion, this note as well as the practice during the past 30 years confirm that the sufficient part of Theorem 2 and Theorem (Criterion) 3 are not only powerful but also sharp, even though at the moment we are still unable to prove the necessity part of Theorem 2 for general state spaces.

Acknowledgments The author thanks Yong-Hua Mao for bringing Spieksma (2014) to the attention.

References

[1] Anderson, W.J., *Continuous-time Markov Chains: An Applications-oriented Approach* (*Springer Series in Statistics*), Springer-Verlag, New York, 1991.

[2] Chen, M.F., Infinite Dimensional Reaction-diffusion Processes, *Acta Mathematica Sinica, New Series*, **1(3)**(1985), 261–273.

[3] Chen, M.F., *Jump Processes and Interacting Particle Systems* (in Chinese), Beijing Normal University Press, 1986a.

[4] Chen, M.F., Coupling for jump processes, *Acta Mathematica Sinica, New Series*, **2(2)**(1986b), 123–136.

[5] Chen, M.F., Stationary distributions of infinite particle systems with non-compact state spaces, *Acta Mathematica Scientia*, **9(1)**(1989), 7–19.

[6] Chen, M.F., On three classical problems for Markov chains with continuous time parameters, *Journal of Applied Probability*, **28(2)**(1991), 305–320.

[7] Chen, M.F., A comment on the book "Continuous-Time Markov Chains" by W.J. Anderson, *Chinese Journal of Applied Probability and Statistics*, **12(1)**(1996), 55–59.

[8] Chen, M.F., Reaction-diffusion processes, *Chinese Science Bulletin*, **42(23)**(1997), 2465–2474 (Chinese Series); *Science Bulletin*, **43(17)**(1998), 1409–1420 (English Series).

[9] Chen, M.F., *From Markov Chains to Non-equilibrium Particle Systems*, Second Edition (First Edition, 1992), World Scientific, 2004.

[10] Chen, M.F., *Eigenvalues, Inequalities, and Ergodic Theory* (*Probability and Its Applications*), Springer, London, 2005.

[11] Chen, M.F. and Mao, Y.H., *Introduction to Stochastic Processes* (in Chinese), Higher Education Press, 2007.

[12] Chen, M.F. and Yan, S.J., Jump processes and particle systems, In *Probability Theory and Its Applications in China* (Edited by: Yan, S.J., Yang C.C. and Wang, J.G.), Providence, American Mathematical Society, **118**(1991), 23–57.

[13] Chen, M.F. and Zhang, Y.H., Unified representation of formulas for single birth processes, *Frontiers of Mathematics in China*, **9(4)**(2014), 761–796.

[14] Chen, X. and Ma, Z.M., A transformation of Markov jump processes and applications in genetic study, *Discrete and Continuous Dynamical Systems*, **34(12)**(2014), 5061–5084.

[15] Chung, K.L., *Markov Chains: With Stationary Transition Probabilities*, Second Edition (First Edition, 1960), Springer-Verlag, Berlin, 1967.

[16] Dobrushin, R.L., Regularity conditions for Markov processes with countably many states (in Russian), *Uspekhi Matematicheskih Nauk* (New Series), **7(6)**(1952), 185–191.

[17] Feller, W., On the integro-differential equations of purely discontinuous Markoff processes, *Transactions of the American Mathematical Society*, **48(3)**(1940), 488–515.

[18] Feller, W., On boundaries and lateral conditions for the Kolmogorov differential equations, *Annals of Mathematics*, **65(3)**(1957), 527–570.

[19] Gikhman, I.I. and Skorokhod, A.V., *The Theory Of Stochastic Processes II*, Springer, New York, 1975.

[20] Hairer, M., *Convergence of Markov Processes*, Lecture Notes, http://www.hairer.org/notes/Convergence.pdf, 2010.

[21] Kim, B. and Lee, I., Tests for nonergodicity of denumerable continuous time Markov processes, *Computers and Mathematics with Applications*, **55(6)**(2008), 1310–1321.

[22] Meyn, S.P. and Tweedie, R.L., *Markov Chains and Stochastic Stability*, Second Edition, Cambridge University Press, 2009.

[23] Reuter, G.E.H., Denumerable Markov processes and the associated contraction semigroups on l, *Acta Mathematica*, **97(1-4)**(1957), 1–46.

[24] Song, J.S., Continuous time Markov decision processes (CTMDP) with non-uniformly bounded transition rates, *Science in China, Series A*, **12(11)**(1988), 1281–1291.

[25] Spieksma, F.M., Countable state Markov processes: non-explosiveness and moment function, To appear in *Probability in the Engineering and Informational Sciences*, 2014.

[26] Yan, S.J. and Chen, M.F., Multi-dimensional Q-processes, *Chinese Annals of Mathematics, Series B*, **7(1)**(1986), 90–110.

[27] Zheng, J.L., Phase transitions of ising model on lattice fractals, martingale approach for Q-processes (in Chinese), Ph.D. thesis, Beijing Normal University, 1993.

[28] Zheng, J.L. and Zheng, X.G., A martingale approach to Q-processes (abstract), *Science Bulletin*, **32(21)**(1987), 1457–1459.

Q过程唯一性的实用判别准则

陈 木 法

(北京师范大学数学科学学院; 北京师范大学数学与复杂系统教育部重点实验室, 北京, 100875)

首先简要介绍自1977年左右开始的寻找连续时间马尔可夫跳过程实用的唯一性充分条件的故事. 对于一般状态空间的一般性结果是1985年得到的. 为展示这些充分条件的精确性, 于1991年找到非唯一性的对偶判别准则. 本文主要限于离散空间(此时的跳过程也称为Q过程或马尔可夫链). 在这种情况下, 我们将在文末证明上述充分条件也是必要的. 我们还将举例说明不论对于唯一性或非唯一性, 我们的充分条件不仅强有力, 而且精确.

关键词: 判别准则, 唯一性, 马尔可夫链, 马尔可夫跳过程.

学科分类号: O211.62.

Extended abstract. In: Souvenir Booklet of the 24th International Workshop on Matrices and Statistics (25-28 May 2015), Haikou City, Hainan, China. Pages 68–72. Ed. Jeffrey J. Hunter.
Spec. Matrices 2016; 4: 9–12

Unified Speed Estimation of Various Stabilities

Mu-Fa Chen

(Beijing Normal University)

March 14, 2015

The main topic of this talk is the speed estimation of stability/instability. The word "various" comes with no surprising since there are a lot of different types of stability/instability and each of them has its own natural distance to measure. However, the adjective "unified" is very much unexpected. The talk surveys our recent progress on the topic, made in the past five years or so.

In the next section, we introduce our first unified result: Theorem 1. Then, several extensions or generalizations of Theorem 1 are collected briefly in Section 2.

1 Basic estimates of the first non-trivial eigenvalue

Here is our first stability, the exponential stability in the ergodic case. Given a Markov chain on a countable E with transition probability $P(t) = (p_{ij}(t) : i, j \in E)$ $(t \geqslant 0)$, in the irreducible ergodic case, we have a stationary distribution π: $\pi P(t) = \pi$ for all $t \geqslant 0$. Then, we have

$$p_{ij}(t) \to \pi_j \quad \text{as} \quad t \to \infty \quad \text{for all } i, j.$$

We are now looking for the exponential convergence speed (rate) ε:

$$|p_{ij}(t) - \pi_j| \leqslant C_i e^{-\varepsilon t}, \qquad t \geqslant 0,\ i, j \in E.$$

Define the Q-matrix by

$$Q = (q_{ij} : i, j \in E) = \frac{\mathrm{d}}{\mathrm{d}\, t} P(t)\Big|_{t=0} \text{ (pointwise)}.$$

In the reversible case, we have $\varepsilon_{\max} = \lambda_1$, where λ_1 is the smallest (the first nontrivial) eigenvalue of $-Q$: $Qg = -\lambda g$ for some $g \neq$ constant.

Let us now consider a simpler birth–death Q-matrix on $E = \{0, 1, 2, \ldots\}$:

$$Q = \begin{pmatrix} -b_0 & b_0 & 0 & 0 & \ldots \\ a_1 & -(a_1 + b_1) & b_1 & 0 & \ldots \\ 0 & a_2 & -(a_2 + b_2) & b_2 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \end{pmatrix},$$

where $a_k,\ b_k > 0$. Since the sum of each row equals 0, we have $Q\mathbb{1} = \mathbb{0} = 0 \cdot \mathbb{1}$, where $\mathbb{1}$ is the vector having elements 1 everywhere and $\mathbb{0}$ is the zero vector. This means that the Q-matrix has a trivial eigenvalue $\lambda_0 = 0$ with eigenvector $\mathbb{1}$. Our question is what is the next eigenvalue λ_1 of $-Q$?

Actually, the story is much harder than it looks like, as shown in [3; pages 1–3], even for $E = \{0, 1, 2, 3\}$. The reader is urged strongly to have some personal computation or have a look at the pages just mentioned.

We now show that the story is even much more complicated. Let $E = \{0, 1, \ldots, N\}$ with $N < \infty$ for a moment. Consider the eigenvalue problem:

$$Qg = -\lambda g, \qquad g \neq 0$$

with Dirichlet boundary at 0: $g_0 = 0$ and Neumann boundary at N: $g_N = g_{N+1}$. Using codes 'D' and 'N', we may denote this minimal eigenvalue λ by λ^{DN}. Actually, the DN case is well studied in the history. Obviously, except the DN case, we should have three more cases: ND, DD, and NN. The last one, λ^{NN}, denotes the ergodic rate λ_1 just mentioned above, for which the constraint is not at the endpoints but is having mean zero.

In the non-ergodic case, the symmetric measure μ can not be finite. Hence, the exponential convergence rate is changed to be the exponential decay rate:

$$\begin{aligned}
&\text{ergodic case}: \ |p_{ij}(t) - \pi_j| \leqslant C_i e^{-\varepsilon t},\ t \geqslant 0,\ \varepsilon_{\max} = \lambda^{\mathrm{NN}};\\
&\text{non-ergodic case}: \ p_{ij}(t) \leqslant C_i e^{-\varepsilon t},\ t \geqslant 0,\ \varepsilon_{\max} = \lambda^{\#}, \qquad i, j \in E,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{where } \# = \text{DN, ND, or DD.}
\end{aligned}$$

Altogether, there are four cases: NN, DD, DN, and ND.

To state our main result, we need a standard notion. Return to our general state space $E = \{0, 1, \cdots, N\}$, $N \leqslant \infty$. Define

$$\mu_0 = 1, \quad \mu_n = \frac{b_0 b_1 \cdots b_{n-1}}{a_1 a_2 \cdots a_n}, \qquad 1 \leqslant n \leqslant N.$$

For general $N \leqslant \infty$, the principal eigenvalue $\lambda^{\#}$ defined above has to be extended to the largest λ satisfying

$$\sqrt{\lambda}\, \|f\|_{\mu,2} \leqslant \|\partial f\|_{\nu,2} \tag{1}$$

with one of the four boundary conditions, where $\|\cdot\|_{\mu,q}$ denotes the $L^q(\mu)$-norm and

$$\nu_i = \begin{cases} \nu_i^- = \mu_i a_i & i \leqslant \theta \\ \nu_i^+ = \mu_i b_i & \theta \leqslant i < N+1; \end{cases}$$

$$\partial_i f = \begin{cases} (\partial_i f)^- = f_{i-1} - f_i & i \leqslant \theta \\ (\partial_i f)^+ = f_{i+1} - f_i & \theta \leqslant i < N+1 \end{cases}$$

and $\theta \in E$ is a reference point.

The author started to study $\lambda^{\rm NN}$ in 1988 (cf. [2, 3]), but the following result (the first unified exponential rate estimation) was obtained in 2010 [5] only.

Theorem 1 For the first non-trivial eigenvalue $\lambda^{\#}$ defined above, we have the following unified basic estimates:

$$\big(4\kappa^{\#}\big)^{-1} \leqslant \lambda^{\#} \leqslant \big(\kappa^{\#}\big)^{-1},$$

where

$$\big(\kappa^{\rm NN}\big)^{-1} = \inf_{n,\, m\in E,\, m<n} \left[\left(\sum_{i=0}^{m}\mu_i\right)^{-1} + \left(\sum_{i=n}^{N}\mu_i\right)^{-1}\right]\left(\sum_{j=m}^{n-1}\frac{1}{\mu_j b_j}\right)^{-1}$$

$$\big(\kappa^{\rm DD}\big)^{-1} = \inf_{n,\, m\in E,\, m\leqslant n} \left[\left(\sum_{i=0}^{m}\frac{1}{\mu_i a_i}\right)^{-1} + \left(\sum_{i=n}^{N}\frac{1}{\mu_i b_i}\right)^{-1}\right]\left(\sum_{j=m}^{n}\mu_j\right)^{-1}$$

$$\kappa^{\rm DN} = \sup_{n\in E}\left(\sum_{i=0}^{n}\frac{1}{\mu_i a_i}\right)^{-1}\left(\sum_{j=n}^{N}\mu_j\right)^{-1}$$

$$\kappa^{\rm ND} = \sup_{n\in E}\left[\left(\sum_{i=0}^{n}\mu_i\right)^{-1}\left(\sum_{j=n}^{N}\frac{1}{\mu_j b_j}\right)^{-1}.\right.$$

In particular, $\lambda^{\#} > 0$ iff $\kappa^{\#} < \infty$.

Note that if we define $\hat{\nu}_k = (\mu_k b_k)^{-1}$, and in the DD and DN cases, under the sum $\sum_{k=0}^{m}$, we modify $\hat{\nu}_k$ to be $(\mu_k a_k)^{-1}$ (noting that when $k \in E$, $\mu_k b_k = \mu_{k+1} a_{k+1}$), then the basic estimates given in the theorem can be described completely by two measures μ and $\hat{\nu}$. The upper and lower bounds are the same up to a universal constant 4 only. It is easy to see that the two endpoints 0 and N are symmetric in the constants $\kappa^{\rm NN}$ and $\kappa^{\rm DD}$.

Finally, we mention that the DN and ND cases are known around 1970 in harmonic analysis, our main contribution is for the cases of DD and N-N, especially the two isoperimetric constants $\kappa^{\rm NN}$ and $\kappa^{\rm DD}$ (come from [5; Corollaries 7.8 and 7.9]). In the proof of the DD and NN cases, three advanced mathematical tools are used and its proof given in [5] consists of five steps. Later, a direct elementary proof was found in [6]. It then leads to the study in the next section.

2 Generalizations

2.1 Bilateral case

Clearly, the birth-death process studied in the last section can be extended to the bilateral one with state space $E = \{i : -M-1 < i < N+1\}$, where $M, N \leqslant \infty$, and with evolution rates: $q_{i,i+1} = b_i$, $q_{i,i-1} = a_i$, and $q_{ij} = 0$ for

other $j \neq i$, $i, j \in E$. In this case, the symmetric measure μ is defined as follows.

$$\mu_{\theta+n} = \frac{a_{\theta-1}a_{\theta-2}\cdots a_{\theta+n+1}}{b_\theta b_{\theta-1}\cdots b_{\theta+n}}, \qquad -M-1-\theta < n \leqslant -2,$$
$$\mu_{\theta-1} = \frac{1}{b_\theta b_{\theta-1}}, \qquad \mu_\theta = \frac{1}{a_\theta b_\theta}, \qquad \mu_{\theta+1} = \frac{1}{a_\theta a_{\theta+1}},$$
$$\mu_{\theta+n} = \frac{b_{\theta+1}b_{\theta+2}\cdots b_{\theta+n-1}}{a_\theta a_{\theta+1}\cdots a_{\theta+n}}, \qquad 2 \leqslant n < N+1-\theta.$$

where $\theta \in E$ is a reference point. In this bilateral case, Theorem 1 remains the same. Refer to [5].

2.2 Bilateral Hardy-type inequalities

Obviously, the Poincaré inequalities (1) can be generalized to the following

$$\|f\|_{\mu,\,q} \leqslant A^{\#}\|\partial f\|_{\nu,\,p}, \qquad f \in L^q(\mu) \tag{2}$$

for $p, q \in [1, \infty]$. This and the parallel inequalities with different boundary condition consist of the bilateral Hardy-type inequalities. When $q \geqslant p$, a generalization of Theorem 1 is given in [7] in the continuous context and in [14] in the discrete one.

2.3 Normed linear space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$

In many applications (Sobolev inequalities, logarithmic Sobolev inequalities, Nash inequalities, and so on), the L^q-norm in (2) is not enough. This leads to the extension to a normed linear space $\mathbb{B}$ which is a linear subset of Borel measurable functions on (E, μ) with a specific norm $\|\cdot\|_{\mathbb{B}}$. In other words, instead of (2), we study the following Hardy-type inequalities

$$\||f|^q\|_{\mathbb{B}}^{1/q} \leqslant A_{\mathbb{B}}^{\#}\|\partial f\|_{\nu,\,p}, \qquad f \in \mathbb{B}$$

with different boundary conditions as before. Our result is presented in [1, 7]. For the last two topics, some popular reports are presented in [7–12].

2.4 Birth–death processes with killing

For the remainder of this section, we consider the birth–death processes with killing on $E = \{0, 1, 2, \ldots, N\}$, $N \leqslant \infty$. Its Q-matrix becomes

$$Q^c = \begin{pmatrix} -(b_0+c_0) & b_0 & 0 & 0 & \cdots \\ a_1 & -(a_1+b_1+c_1) & b_1 & 0 & \cdots \\ 0 & a_2 & -(a_2+b_2+c_2) & b_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

with $a_i > 0,\ b_i > 0$, and $c_i \geqslant 0$ for every $i \in E$. Clearly, this is a special type of tridiagonal or Jacobi's matrix. Assume that $c_i \not\equiv 0$ on $(0, N)$, otherwise, we would return to Section 1. Even though the spectral problem becomes much harder than before, since a new sequence of parameter (c_i) is added, we are lucky to obtain a result in parallel to Theorem 1. Refer to [13, 9].

2.5 Discrete spectrum

We say that the matrix Q^c (or its quadratic form) on $L^2(\mu)$ has discrete spectrum if its spectrum consists of only eigenvalues with finite multiplicity. Since an operator on a finite space is compact and hence must have discrete spectrum, we need only consider an infinite state space. Next, since the whole line can be split into two half lines, without loss of generality, we assume that $E = \{0, 1, \ldots\}$. In this subsection, we allow $c_i|_{(0,N-1)} \equiv 0$. This problem is solved completely by [9; Theorem 2.1], based on [13]. From the last cited paper, one finds an interesting story on isospectral operators.

Acknowledgments. This paper is an extended abstract of a plenary lecture presented at "24th International Workshop on Matrices and Statistics," the author acknowledges a kind invitation by the Scientific Program Committee, especially Professor Jeffrey J. Hunter.

The references given below are only those the talk is based on. It is regretted that a large number of publications in the active research area is omitted here, otherwise, the list would be too long. For more references on the related sub-topics, the reader is urged to look at the related papers below.

References

[1] Chen, M.F. (2003). *Variational formulas of Poincaré-type inequalities for birth-death processes.* Acta Math. Sin. Eng. Ser. 19(4): 625-644.

[2] Chen, M.F. (2004). *From Markov Chains to Non-equilibrium Particle Systems.* World Scientific. 2nd ed. (1st ed., 1992).

[3] Chen, M.F. (2005). *Eigenvalues, Inequalities, and Ergodic Theory.* Springer, London.

[4] Chen, M.F. (2007). *Exponential convergence rate in entropy.* Front. Math. China, 2(3): 329–358.

[5] Chen M.F. (2010). *Speed of stability for birth–death processes.* Front Math China 5(3): 379–515.

[6] Chen, M.F. (2012). *Lower bounds of principal eigenvalue in dimension one.* Front. Math. China 7(4): 645–668.

[7] Chen, M.F. (2013a). *Bilateral Hardy-type inequalities.* Acta Math Sin Eng Ser. 29(1): 1–32.

[8] Chen, M.F. (2013b). *Bilateral Hardy-type inequalities and application to geometry.* Mathmedia 37(2): 12–32; Math. Bulletin 52(8/9) (in Chinese).

[9] Chen, M.F. (2014). *Criteria for discrete spectrum of 1D operators.* Commu. Math. Stat. 2: 279–309

[10] Chen, M.F. (2015a). *Criteria for two spectral problems of 1D operators* (in Chinese). Sci Sin Math, 44(1):
[11] Chen, M.F. (2015b). *The optimal constant in Hardy-type inequalities.* Acta Math. Sinica, Eng. Ser.
[12] Chen, M.F. (2015c). *Progress on Hardy-type inequalities.* Chapter 6 in the book "Festschrift Masatoshi Fukushima", eds: Z.Q. Chen, N. Jacob, M. Takeda, and T. Uemura, World Sci.
[13] Chen, M.F. and Zhang, X. (2014) *Isospectral operators.* Commu Math Stat 2: 17–32.
[14] Liao, Z.W. (2015). *Discrete weighted Hardy inequalities with different boundary conditions.* arXiv:1508.04601.

Mu-Fa Chen
School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.
E-mail: mfchen@bnu.edu.cn
Home page: http://math.bnu.edu.cn/~chenmf/main_eng.htm

应用概率统计　第 32 卷
第 1 期　2016 年 2 月

Chinese Journal of Applied Probability and Statistics
Feb., 2016, Vol. 32, No. 1, pp. 1-22
doi: 10.3969/j.issn.1001-4268.2016.01.001

Survey

Unified Speed Estimation of Various Stabilities *

CHEN MuFa

(*School of Mathematical Sciences, Beijing Normal University; Laboratory of Mathematics and Complex Systems* (*Beijing Normal University*), *Ministry of Education, Beijing, 100875, China*)

Abstract: To study some infinite-dimensional subject (the phase transitions in statistical physics, for instance), several mathematical tools are developed. One of them is the speed estimation of various stabilities/instabilities. This paper collects some unexpected, unified, nearly sharp basic estimates of various types of stability/instability for the simplest class of Markov processes, the birth-death processes. Some motivations and a part of extensions are also discussed. The paper is based on a talk presented recently in several international conferences.

Keywords: stability; the first (non-trivial) eigenvalue; Hardy-type inequality; killing; speed estimation; criterion; birth-death process

2010 Mathematics Subject Classification: 60J27

The main topic of this talk is the speed estimation of stability/instability. The word "various" comes with no surprising since there are a lot of different types of stability/instability and each of them has its own natural distance to measure. However, the adjective "unified" is very much unexpected. The talk surveys our recent progress on the topic, made in the past five years or so. The restriction on our recent work is due to the fact it seems now not practical to have a comprehensive review in a paper for the rapidly developing subject.

In the next section, we begin with two models to show the importance of the topic. Then the difficulty of the problem is illustrated. The first unified result is Theorem 3 presented in Section 2. If one is in hurry to know a representative result, who may jump from here to Section 2. The subsequent sections of the paper is devoted to various extension or generalization of Theorem 3.

*The research was supported in part by the National Natural Science Foundation of China (No. 11131003), the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Received October 28, 2015.

§1. Motivation

We show the importance of speed estimation by two concrete models.

Economic Model

For a given Markov chain with countable state space E and transition matrix $P = (p_{ij} : i, j \in E)$, very often, we have a stationary distribution $\pi = (\pi_i : i \in E)$ such that $\pi = \pi P$ and furthermore

$$\pi = \pi P^n, \qquad n \geqslant 1.$$

In the probabilistic language, the last formula says that if the chain starts at π, then its distribution at every time n is the same π. This is the meaning of the invariance measure and is a very useful stability in practice. We are now going to show a different meaning of this property.

Recall that for a finite E, the last property is a special case of the Perron–Frobenius theorem. Let $A = (a_{ij} : i, j \in E)$ be a nonnegative irreducible (prime) matrix and denote by $\rho(A)$ and u (must be positive, up to a constant), respectively, the maximal eigenvalue of A and its left-eigenvector (row). Then $uA = \rho(A)u$ and furthermore

$$uA^n = \rho(A)^n u, \qquad n \geqslant 1.$$

When A is invertible, this is equivalent to

$$uA^{-n} = \rho(A)^{-n} u, \qquad n \geqslant 1.$$

Regarding u as the initial input x_0, then

$$x_n := x_0 A^{-n}, \qquad n \geqslant 1$$

is the output at the nth year of the economic model with structure matrix A. This is the well-known input–output method in economy. Clearly, we have the equilibrium:

$$x_0 = u \Longrightarrow x_n = \rho(A)^{-n} u, \qquad n \geqslant 1.$$

This means that starting from $x_0 = u$, the economy has an optimal growing speed $\rho(A)^{-n}$ in some sense. This stability result comes with no surprising. The surprising point, due to Loo-Keng Hua (1984), is the following (cf. [25] or [3; Chapter 10]): if we start from $x_0 \neq u$ up to a constant, then the economy will be collapsed, that is, the collapse time

$$T^{x_0} = \inf\{n : x_n \not> 0 \text{ (some of the components of } x_n \text{ is } \leqslant 0)\}$$

is finite. After Hua proved his result, he wrote a letter to Zhen-Ting Hou and Kai-Lai Chung, separately, saying that "every Markov chain should have a natural starting point". The starting point (distribution) is (assuming to be irreducible for simplicity) clearly the specific entrance law, i.e., the stationary distribution $\pi : \pi P^{-n} \equiv \pi$. Otherwise, if $\mu \neq \pi$, then μP^{-n} has only a finite life-time.

Let us now return to Hua's conclusion. In the case that T^{x_0} is very large, ten thousand years for instance, then we do not need to take care of it. However, it is not the case in practice. Here is a simple example.

Example 1 Let

$$A = \frac{1}{100}\begin{pmatrix} 25 & 14 \\ 40 & 12 \end{pmatrix}.$$

Then the left-eigenvector corresponding to the maximal eigenvalue is

$$u = \big(5\big(\sqrt{2409}+13\big)/7, 20\big) \approx (44.34397483, 20).$$

With respect to different x_0, the collapse time is given in Table 1.

Table 1 The collapse time with respect to x_0

x_0	T^{x_0}
$(44, 20)$	3
$(44.344, 20)$	8
$(44.34397483, 20)$	13

This shows that the economy is very sensitive!

For input $x_0 = (44.344, 20)$, the collapse time $T^{x_0} = 8$. This is okay since the economic plan in our country lasts for 5 years. Now, if we make ± 0.01 perturbation of A with probability $1/6$, respectively, then

$$\mathbb{P}^{x_0}\big[T^{x_0} \leqslant 3\big] = 0.74.$$

Hence, a modulation of the input x_0 is needed at the end of the second year. How can we imagine such a fast collapse in advance? The textbook [18] begins with this attractive model.

The importance of the speed (the collapse rate) estimation should be clear now. Generally speaking, a subject is still incomplete if knowing qualitative theory only without quantitative estimation. Unfortunately, the analytic estimation for the collapse rate of this model is still largely open (cf. [3; Chapter 10] for instance). However, there is a recent interesting progress on computing the maximal eigenpair $(\rho(A), u)$, refer to [17]. The

efficient initials used in [17] are based on Theorems 3, 7 and Corollary 8 in this paper. Thus, one may regard [17] as a remarkable application of the theory introduced here.

The next model is taken from statistical physics, which is an infinite-dimensional model.

Phase Transition: φ^4 Euclidean Quantum Field on the Lattice

Consider the following operator L for the φ^4-model with state space $\mathbb{R}^{\mathbb{Z}^d}$:

$$L = \sum_{i\in\mathbb{Z}^d} \big[\partial_{ii} - (u'(x_i) + \partial_i H)\partial_i\big],$$

where

$$u(x_i) = x_i^4 - \beta x_i^2, \qquad H(x) = -2J \sum_{\langle ij\rangle} x_i x_j, \qquad \beta \geqslant 0,\ J \geqslant 0,$$

and the pair $\langle ij\rangle$ is the nearest neighbors in $\mathbb{Z}^d$. For $\Lambda \Subset \mathbb{Z}^d$ (finite subset), let

$$U_\Lambda^\omega(x_\Lambda) = \sum_{i\in\Lambda} u(x_i) - J \sum_{\langle ij\rangle:\, i,j\in\Lambda} x_i x_j - J \sum_{\langle ij\rangle:\, i\in\Lambda,\, j\notin\Lambda} x_i\omega_j, \qquad \omega \in \mathbb{R}^{\mathbb{Z}^d}.$$

Define the conditional Gibbs distribution $\pi_U^{\Lambda,\omega}(\mathrm{d}x_\Lambda)$ with boundary ω and the Dirichlet form $D_U^{\Lambda,\omega}(f)$ (more precisely, the diagonal elements of a Dirichlet form) on the L^2-space $L^2\big(\pi_U^{\Lambda,\omega}\big)$ as follows.

$$\pi_U^{\Lambda,\omega}(\mathrm{d}x_\Lambda) = \mathrm{e}^{-U_\Lambda^\omega(x_\Lambda)}/Z_\Lambda^\omega, \qquad D_U^{\Lambda,\omega}(f) = \int_{\mathbb{R}^\Lambda} |\nabla f|^2\, \mathrm{d}\pi_U^{\Lambda,\omega}, \qquad \Lambda \Subset \mathbb{Z}^d,\ \omega \in \mathbb{R}^{\mathbb{Z}^d},$$

where Z_Λ^ω is a normalizing constant. In general, knowing $D(f)$ on $L^2(\mu)$ with a general measure μ, it is standard to write down the quadratic form $D(f,g)$ on $L^2(\mu)$ by the quadrilateral rule:

$$D(f,g) = \frac{1}{4}\big[D(f+g) - D(f-g)\big].$$

In particular, here we have

$$D_U^{\Lambda,\omega}(f,g) = \int_{\mathbb{R}^\Lambda} \langle \nabla f,\, \nabla g\rangle_{L^2\left(\mathbb{R}^\Lambda,\, \pi_U^{\Lambda,\omega}\right)} \mathrm{d}\pi_U^{\Lambda,\omega}, \qquad \Lambda \Subset \mathbb{Z}^d,\ \omega \in \mathbb{R}^{\mathbb{Z}^d}.$$

Having these quantities at hand, we can define the local first non-trivial eigenvalue $\lambda_1^{\beta,J}(\Lambda,\omega)$ which describes the exponential stability rate and the local logarithmic Sobolev constant $\sigma^{\beta,J}(\Lambda,\omega)$ which describes the exponential stability rate in entropy, as will be defined below.

Again, in general, for a given probability measure π and a Dirichlet form $D(f)$ on the L^2-space $L^2(\pi)$ with norm $\|\cdot\|$, we can define the largest λ_1 and σ satisfying the following inequalities, respectively:

- Poincaré inequality [H. Poincaré, 1890]:

$$\lambda_1\|f-\pi(f)\|^2 \leqslant D(f), \qquad f\in L^2(\pi),\ \ \|\cdot\|:=\|\cdot\|_{L^2(\pi)}.$$

- Logarithmic Sobolev inequality [L. Gross, 1976]:

$$2^{-1}\sigma\,\mathrm{Ent}(f^2)\leqslant D(f), \qquad f\in L^2(\pi),$$

where $\mathrm{Ent}(f)=\int\big(f\log(f/\|f\|_1)\big)\mathrm{d}\pi$ and $\log=\log_{\mathrm{e}}$, $\|\cdot\|_r$ is the $L^r(\pi)$-norm.

The last inequality [23, 24] is now quite popular since it was used by G. Perelman (2002) in the study of Poincaré conjecture.

Next, let $(P_t)_{t\geqslant 0}$ be the symmetric Markov semigroup determined by the Dirichlet form. Then

- the Poincaré inequality is equivalent to the L^2-exponential stability

$$\|P_tf-\pi(f)\|\leqslant\|f\|\mathrm{e}^{-\varepsilon t}, \qquad f\in L^2(\pi),\ t\geqslant 0,$$

with the maximal rate $\varepsilon_{\max}=\lambda_1$.

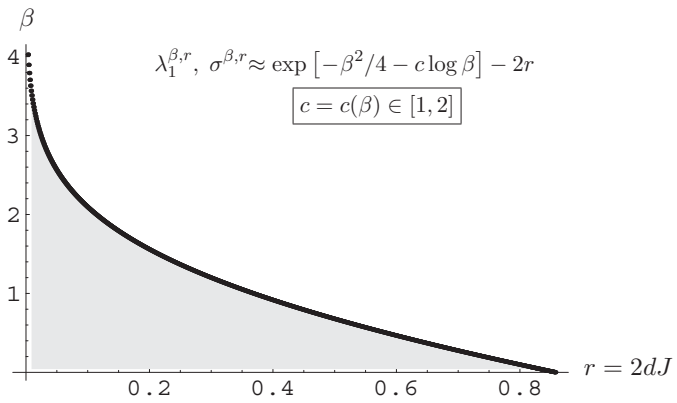
- The logarithmic Sobolev inequality implies the exponential stability in entropy

$$\mathrm{Ent}(P_tf)\leqslant\mathrm{Ent}(f)\mathrm{e}^{-2\sigma t}, \qquad f\in L^2_+(\pi),\ t\geqslant 0.$$

Refer to [3] for more details.

Theorem 2 ([5]) For the φ^4-model, we have

$$\inf_{\Lambda\Subset\mathbb{Z}^d}\inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}}\lambda_1^{\beta,J}(\Lambda,\omega)\approx\inf_{\Lambda\Subset\mathbb{Z}^d}\inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}}\sigma^{\beta,J}(\Lambda,\omega)\approx\exp\big[-\beta^2/4-c\log\beta\big]-4dJ.$$



In the shadow region, $\inf_{\Lambda\in\mathbb{Z}^d}\inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}}\lambda_1^{\beta,r}(\Lambda,\omega)>0$. This means that the system is exponentially ergodic for smaller β (higher temperature) and smaller r (weaker interaction), uniformly in finite subset Λ and in boundary ω. We mention that here the leading term $\beta^2/4$ is sharp. In the region which is a little away from the shadow one, $\lambda_1^{\beta,r}(\Lambda,\omega)$ becomes zero and so the system cannot be (exponentially) ergodic. For more details, refer to [5]. Clearly, the quantitative estimation of $\lambda_1^{\beta,r}(\Lambda,\omega)$ plays a key role in Theorem 2.

The author learnt this approach in 1988. For this, we were very glad since we can use the well-developed spectral theory to study the new topic — phase transitions to which the known techniques are rather limited, due to the fact it is an infinite-dimensional object. Now, the question is: how to find the leading eigenvalue for some typical Markov processes? Certainly, the first class of processes we should choose is the one-dimensional ones.

Birth-Death Processes

Let us now borrow some materials from [3; pages 1–3]. Consider a birth-death Q-matrix on $E=\{0,1,2,\ldots\}$:

$$Q=(q_{ij}:i,j\in E)=\begin{pmatrix}-b_0 & b_0 & 0 & 0 & \ldots\\ a_1 & -(a_1+b_1) & b_1 & 0 & \ldots\\ 0 & a_2 & -(a_2+b_2) & b_2 & \ldots\\ \vdots & \ddots & \ddots & \ddots & \end{pmatrix},$$

where a_k, $b_k>0$. Since the sum of each row equals 0, we have $Q\mathbb{1}=\mathbb{0}=0\cdot\mathbb{1}$, where $\mathbb{1}$ is the vector having elements 1 everywhere and $\mathbb{0}$ is the zero vector. This means that the Q-matrix has a trivial eigenvalue $\lambda_0=0$ with eigenvector $\mathbb{1}$. Our question is what the next eigenvalue λ_1 of $-Q$ is?

To get a concrete feeling about the difficulties of the topic, let us look at the following simple examples.

When $E=\{0,1\}$,

$$Q=\begin{pmatrix}-b_0 & b_0\\ a_1 & -a_1\end{pmatrix},$$

it is trivial that $\lambda_1=a_1+b_0$. Everyone is happy to see this result, since if either a_1 or b_0 increases, so does λ_1. If we go one more step, $E=\{0,1,2\}$,

$$Q=\begin{pmatrix}-b_0 & b_0 & 0\\ a_1 & -(a_1+b_1) & b_1\\ 0 & a_2 & -a_2\end{pmatrix},$$

then we have four parameters, b_0, b_1 and a_1, a_2. In this case,

$$\lambda_1 = 2^{-1}\big[a_1 + a_2 + b_0 + b_1 - \sqrt{(a_1 - a_2 + b_0 - b_1)^2 + 4a_1b_1}\,\big].$$

It is disappointing to see this result, since how the parameters effect on λ_1 is not clear at all. When $E = \{0, 1, 2, 3\}$,

$$Q = \begin{pmatrix} -b_0 & b_0 & 0 & 0 \\ a_1 & -(a_1 + b_1) & b_1 & 0 \\ 0 & a_2 & -(a_2 + b_2) & b_2 \\ 0 & 0 & a_3 & -a_3 \end{pmatrix},$$

we have six parameters: $b_0, b_1, b_2, a_1, a_2, a_3$. The solution is expressed by the three quantities B, C, and D:

$$\lambda_1 = \frac{D}{3} - \frac{C}{3 \cdot 2^{1/3}} + \frac{2^{1/3}\,(3\,B - D^2)}{3\,C},$$

where the quantities D, B, and C are not too complicated:

$$\begin{aligned} D &= a_1 + a_2 + a_3 + b_0 + b_1 + b_2, \\ B &= a_3b_0 + a_2(a_3 + b_0) + a_3b_1 + b_0b_1 + b_0b_2 + b_1b_2 + a_1(a_2 + a_3 + b_2), \\ C &= \big(A + \sqrt{4(3\,B - D^2)^3 + A^2}\,\big)^{1/3}. \end{aligned}$$

However, in the last expression, another quantity, A, is involved. What, then, is A?

$$\begin{aligned} A = {} & -2a_1^3 - 2a_2^3 - 2a_3^3 + 3a_3^2b_0 + 3a_3b_0^2 - 2b_0^3 + 3a_3^2b_1 - 12a_3b_0b_1 + 3b_0^2b_1 \\ & + 3a_3b_1^2 + 3b_0b_1^2 - 2b_1^3 - 6a_3^2b_2 + 6a_3b_0b_2 + 3b_0^2b_2 + 6a_3b_1b_2 - 12b_0b_1b_2 \\ & + 3b_1^2b_2 - 6a_3b_2^2 + 3b_0b_2^2 + 3b_1b_2^2 - 2b_2^3 + 3a_1^2(a_2 + a_3 - 2b_0 - 2b_1 + b_2) \\ & + 3a_2^2\big[a_3 + b_0 - 2(b_1 + b_2)\big] \\ & + 3a_2\big[a_3^2 + b_0^2 - 2b_1^2 - b_1b_2 - 2b_2^2 - a_3(4b_0 - 2b_1 + b_2) + 2b_0(b_1 + b_2)\big] \\ & + 3a_1\big[a_2^2 + a_3^2 - 2b_0^2 - b_0b_1 - 2b_1^2 - a_2(4a_3 - 2b_0 + b_1 - 2b_2) \\ & \qquad + 2b_0b_2 + 2b_1b_2 + b_2^2 + 2a_3(b_0 + b_1 + b_2)\big], \end{aligned}$$

computed using Mathematica. One should be shocked, at least I was, to see this result, since the roles of the parameters are completely hidden! Of course, everyone understands that it is impossible to compute λ_1 explicitly when the size of the matrix is greater than five!

Now, how about the estimation of λ_1? To see this, let us consider the perturbation of the eigenvalues and eigenfunctions. We consider the infinite state space $E = \{0, 1, 2, \ldots\}$.

Denote by g and Degree(g), respectively, the eigenfunction of λ_1 and the degree of g when g is polynomial. Three examples of the perturbation of λ_1 and Degree(g) are listed in Table 2.

Table 2 Perturbation of λ_1 and Degree(g)

$b_i\,(i \geqslant 0)$	$a_i\,(i \geqslant 1)$	λ_1	Degree(g)
$i + c\,(c > 0)$	$2\,i$	1	1
$i + 1$	$2\,i + 3$	2	2
$i + 1$	$2\,i + (4 + \sqrt{2})$	3	3

The first line is the well-known linear model, for which $\lambda_1 = 1$, independent of the constant $c > 0$, and g is linear. Next, keeping the same birth rate, $b_i = i + 1$, the perturbation of the death rate a_i from $2\,i$ to $2\,i + 3$ and then to $2\,i + 4 + \sqrt{2}$ leads to the change of λ_1 from one to two and then to three. More surprisingly, the eigenfunction g is changed from linear to quadratic and then to cubic. As seen from these examples, the first non-trivial eigenvalue is very sensitive. Hence, in general, it is very hard to estimate λ_1.

§2. Basic Estimates of the First Non-Trivial Eigenvalue

Actually, the story is much more complicated. Let $E = \{0, 1, \ldots, N\}$ with $N < \infty$ for a moment. Consider the eigenvalue problem:

$$-Qg = \lambda g, \qquad g \neq 0$$

with Dirichlet boundary at 0: $g_0 = 0$ and Neumann boundary at N: $g_N = g_{N+1}$. Using codes 'D' and 'N', we may denote the minimal eigenvalue λ by λ^{DN}. Actually, the DN case is well studied in the context of Hardy-type inequalities (we will come back to this topic later). Under its inspiration, we obtained a criterion for $\lambda_1 > 0$ in 2000 (cf. [3; Theorem 5.7]). Next, except the DN case, we should have three more cases: ND, DD, and NN. The last one, λ^{NN}, denotes the ergodic rate λ_1 discussed in the last section, for which the constraint is not at the endpoints but is having mean zero.

In the non-ergodic case, the symmetric measure μ cannot be finite. The Poincaré inequality becomes

$$\lambda^{\#}\|f\|^2 \leqslant D(f), \qquad f \in L^2(\pi)$$

which is equivalent to the L^2-exponential stability

$$\|P_t f\| \leqslant \|f\|\mathrm{e}^{-\varepsilon t}, \qquad f \in L^2(\mu),\ t \geqslant 0,$$

the maximal rate $\varepsilon_{\max} = \lambda^{\#}$, where $\# =$ DN, ND, or DD. Furthermore, if we replace the L^2-norm by pointwise convergence, for Markov chains at least, we also have the same exponential stability rate, even in the ergodic case (cf. [3,6]):

$$\begin{aligned}&\text{ergodic case}: \ |p_{ij}(t)-\pi_j| \leqslant C_i \mathrm{e}^{-\varepsilon t}, \ t \geqslant 0, \ \varepsilon_{\max} = \lambda^{\mathrm{NN}};\\ &\text{non-ergodic case}: \ p_{ij}(t) \leqslant C_i \mathrm{e}^{-\varepsilon t}, \ t \geqslant 0, \ \varepsilon_{\max} = \lambda^{\#}, \qquad i,j \in E,\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{where } \# = \text{DN, ND, or DD.}\end{aligned}$$

To make a more precise definition of $\lambda^{\#}$, we follow [13]. Let $E = \{i : -M-1 < i < N+1\}$, $M, N \leqslant \infty$. That is, we are now considering the bilateral case. The birth-death Q-matrix becomes: for $i, j \in E$, $q_{i,i+1} = b_i > 0$, $q_{i,i-1} = a_i > 0$, $q_{ii} = -(a_i + b_i)$, and $q_{ij} = 0$ provided $|i-j| > 1$. Choose a reference point $\theta \in E$ and define

$$\begin{aligned}&\mu_{\theta+n} = \frac{a_{\theta-1}a_{\theta-2}\cdots a_{\theta+n+1}}{b_\theta b_{\theta-1}\cdots b_{\theta+n}}, \qquad -M-1-\theta < n \leqslant -2,\\ &\mu_{\theta-1} = \frac{1}{b_\theta b_{\theta-1}}, \qquad \mu_\theta = \frac{1}{a_\theta b_\theta}, \qquad \mu_{\theta+1} = \frac{1}{a_\theta a_{\theta+1}},\\ &\mu_{\theta+n} = \frac{b_{\theta+1}b_{\theta+2}\cdots b_{\theta+n-1}}{a_\theta a_{\theta+1}\cdots a_{\theta+n}}, \qquad 2 \leqslant n < N+1-\theta.\end{aligned}$$

Note that for $k > \theta$, (a_k, b_k) plays the role as so does (b_k, a_k) for $k < \theta$. When θ varies, the measure (μ_i^θ) keeps the same up to a constant depending on θ only. Thus, when $M < \infty$ for instance, one can simply set $\theta = -M$ and define

$$\mu_\theta = 1, \qquad \mu_{\theta+n} = \frac{b_\theta b_{\theta+1}\cdots b_{\theta+n-1}}{a_{\theta+1}a_{\theta+2}\cdots a_{\theta+n}}, \qquad 1 \leqslant n \leqslant N+M.$$

Similarly, when $N < \infty$, we can define a sequence $(\mu_{\theta-n} : 0 \leqslant n \leqslant N+M)$ with $\theta = N$. However, when $M = \infty = N$, a reference point θ in the open interval $(-M, N)$ is necessary. Corresponding to the matrix Q, the Dirichlet form (diagonal elements) on $L^2(\mu)$ is defined by

$$D(f) = \sum_{-M-1<i\leqslant\theta} \mu_i a_i (f_i - f_{i-1})^2 + \sum_{\theta\leqslant i<N+1} \mu_i b_i (f_{i+1} - f_i)^2.$$

Since $\mu_{i+1}a_{i+1} = \mu_i b_i$ once $i \in E$, this expression does not depend on the choice of $\theta \in E$. In the ND case, simply set $\theta = -M-1$; conversely, in the DN case, set $\theta = N+1$. Next, we have the following boundary condition:

$$\begin{aligned}&\text{if } M < \infty, \text{ then } \begin{cases} f_{-M-1} = 0, & \text{Dirichlet boundary}\\ f_{-M-1} = f_{-M}, & \text{Neumann boundary,}\end{cases}\\ &\text{if } N < \infty, \text{ then } \begin{cases} f_{N+1} = 0, & \text{Dirichlet boundary}\\ f_{N+1} = f_N, & \text{Neumann boundary.}\end{cases}\end{aligned}$$

Noting that here when $N<\infty$, we assume $b_N>0$. Similarly, when $M<\infty$, we assume $a_{-M}>0$. This is helpful to describe the boundary conditions. However, for Neumann boundary at N for instance, usually we assume $b_N=0$ and omit "$f_{N+1}=f_N$". Since we now allow infinite intervals, the quantities $\lambda^{\#}$ defined above need to be extended. To do so, write $\mu(f)=\sum\limits_{k\in E}\mu_k f_k$. Then we have $\mu[\alpha,\beta]=\mu\big(\mathbb{1}_{[\alpha,\beta]}\big)$. Clearly, $\mu[\alpha,N]=\mu[\alpha,N)$ once $N=\infty$. First, define

$$\lambda^{\rm DD}=\inf\big\{D(f):f \text{ has finite support and } \mu\big(f^2\big)=1\big\}.$$

Next, when $\mu[\theta,N]<\infty$, define

$$\lambda^{\rm DN}=\inf\big\{D(f):\exists\, m,n\in E,\ m<n \text{ such that } f=\mathbb{1}_{[m,N]}f_{\bullet\wedge n} \text{ and } \mu\big(f^2\big)=1\big\},$$

where $\alpha\wedge\beta=\min\{\alpha,\beta\}$ and similarly, $\alpha\vee\beta=\max\{\alpha,\beta\}$. In words, here f vanishes on $[-M,m-1]$ and is a constant on $[n,N]$. Dually, when $\mu[-M,\theta]<\infty$, define

$$\lambda^{\rm ND}=\inf\big\{D(f):\exists\, m,n\in E,\ m<n \text{ such that } f=\mathbb{1}_{[-M,n]}f_{\bullet\vee m} \text{ and } \mu\big(f^2\big)=1\big\}.$$

Finally, when $\mu[-M,N]<\infty$, define

$$\begin{aligned}\lambda^{\rm NN}&=\inf\big\{D(f):\mu(f)=0,\ \mu\big(f^2\big)=1\big\}\\&=\inf\big\{D(f):\exists\, m,n\in E,\ m<n \text{ such that } f=f_{m\vee\bullet\wedge n},\ \mu(f)=0\ \&\ \mu\big(f^2\big)=1\big\}.\end{aligned}$$

The author started to study $\lambda^{\rm NN}$ in 1988, but the following result (the first unified exponential rate estimation) was obtained in 2010 only.

Theorem 3 For the first non-trivial eigenvalue $\lambda^{\#}$ defined above, we have the following unified basic estimates:

$$\big(4\kappa^{\#}\big)^{-1}\leqslant\lambda^{\#}\leqslant\big(\kappa^{\#}\big)^{-1},\tag{1}$$

where

$$\big(\kappa^{\rm NN}\big)^{-1}=\inf_{n,m\in E,\,m<n}\Big[\Big(\sum_{i=-M}^{m}\mu_i\Big)^{-1}+\Big(\sum_{i=n}^{N}\mu_i\Big)^{-1}\Big]\Big(\sum_{j=m}^{n-1}\frac{1}{\mu_j b_j}\Big)^{-1},\tag{2}$$

$$\big(\kappa^{\rm DD}\big)^{-1}=\inf_{n,m\in E,\,m\leqslant n}\Big[\Big(\sum_{i=-M}^{m}\frac{1}{\mu_i a_i}\Big)^{-1}+\Big(\sum_{i=n}^{N}\frac{1}{\mu_i b_i}\Big)^{-1}\Big]\Big(\sum_{j=m}^{n}\mu_j\Big)^{-1},\tag{3}$$

$$\kappa^{\rm DN}=\sup_{n\in E}\Big(\sum_{i=-M}^{n}\frac{1}{\mu_i a_i}\Big)^{-1}\Big(\sum_{j=n}^{N}\mu_j\Big)^{-1},\tag{4}$$

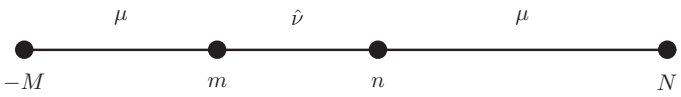
$$\kappa^{\rm ND}=\sup_{n\in E}\Big(\sum_{i=-M}^{n}\mu_i\Big)^{-1}\Big(\sum_{j=n}^{N}\frac{1}{\mu_j b_j}\Big)^{-1}.\tag{5}$$

In particular, $\lambda^{\#}>0$ iff $\kappa^{\#}<\infty$.

We remark that if we define $\widehat{\nu}_k = (\mu_k b_k)^{-1}$, and in the DD and DN cases, under the sum $\sum_{k=-M}^{m}$ we modify $\widehat{\nu}_k$ to be $(\mu_k a_k)^{-1}$ (noting that when $k \in E$, $\mu_k b_k = \mu_{k+1} a_{k+1}$), then the basic estimates given in the theorem can be described completely by two measures μ and $\widehat{\nu}$. The upper and lower bounds are the same up to a universal constant 4 only. It is easy to see that the two endpoints $-M$ and N are symmetric in these two constants: κ^{NN} and κ^{DD}.

Since this is the central result of the paper, it maybe helpful for our reader to write down $\kappa^{\#}$ step by step. Let us begin with κ^{NN}.

- Since we are in the bilateral case (having the same boundaries at the two endpoints), in the ergodic case especially, the process starting from different endpoints may have different rates to go to the middle part of the interval. Based on this, we need two parameters, say m and n with $m < n$. The state space $[-M, N]$ is then divided by m and n into three parts: the left-hand part $[-M, m]$, the right-hand part $[n, N]$, and the middle one $[m, n-1]$.



- Measure the left-hand and the right-hand subintervals by μ and the middle one by $\widehat{\nu}$, respectively:

$$\kappa = \kappa^{\mathrm{NN}}: \qquad \mu[-M, m] \qquad \mu[n, N] \qquad \widehat{\nu}[m, n-1].$$

- Make inverse everywhere:

$$\kappa^{-1}: \qquad \mu[-M, m]^{-1} \qquad \mu[n, N]^{-1} \qquad \widehat{\nu}[m, n-1]^{-1}.$$

- Finally, multiplying the last term with the sum of the first two terms and making infimum with respect to $m < n$, we get the expression of $\left(\kappa^{\mathrm{NN}}\right)^{-1}$ given in (2).

Every step is quite natural except the second one: why we use μ but not $\widehat{\nu}$ in the first two terms? This is because we are in the ergodic case, μ is a finite measure. If μ is replaced by $\widehat{\nu}$, since $\widehat{\nu}(-\infty, m]$ and $\widehat{\nu}[n, \infty)$ are infinite when $M, N = \infty$, one would get zero for these terms and so the quantity is trivial. A sensitive point here is that we use plus, rather than maximum in the last step. Otherwise, even though the resulting bounds are equivalent to ours but it then would produce a factor 8 rather than 4 as we expected. We have thus completed the description of the first, the most important quantity κ^{NN}.

- To get κ^{DD}, simply apply the rule: the exchange of the codes D and N simultaneously in $\kappa^{\#}$ leads to the exchange of the measures μ and $\widehat{\nu}$ in their representation.
- For κ^{DN}, let us examine (3) more carefully. When $N=\infty$ and $\widehat{\nu}[n,\infty)=\infty$, the second term in the sum of (3) disappeared. In other words, the boundary condition D on the right endpoint is replaced by N. Then the variable n is free and so can be removed. Therefore we obtain formula (4). We remark however that the relation between λ^{DN} and κ^{DN} remains the same even if $\widehat{\nu}[n,\infty)<\infty$.
- For κ^{ND}, (5) follows from (4) simply using again our rule. Here we mention that (5) can be formally obtained from (2) by removing the second term in the sum. Actually, (5) is formally a reverse of (4), and so is somehow an easy consequence of (4).

For the last two constants in (4) and (5) respectively, we write down only the case that $M=0$ in [6]. If $M=\infty$, then we can first use finite M instead of 0, and then go to the limit $M\to\infty$ to arrive the required result. When $M<\infty$ and the left endpoint is the Dirichlet boundary, the constant here is different from that in [6]: here we adopt $-M$, but in [6] we use $-M+1$. The reason for this point is to guarantee the first non-trivial eigenvalue in the NN case coincides with the principal eigenvalue in the DD case. Thus, when $M,N<\infty$, the cardinality of the state space in DD case is one less than that in the NN case.

Finally, we mention that the DN and ND cases are known around 1970 in harmonic analysis, our main contribution is for the cases of DD and NN, especially the two isoperimetric constants κ^{NN} and κ^{DD} (come from [6; Corollaries 7.8 and 7.9]). In computing κ^{NN} and κ^{DD}, for fixed $m\leqslant n$, we simply choose the reference point in defining (μ_k) and $D(f)$ to be the same: $\theta=m$. In the proof of the DD and NN cases, three advanced mathematical tools (the coupling method, the dual technique, and the capacitary method) are used and its proof given in [6] consists of five steps. Later, a direct elementary proof was found in [9]. It then leads to the study in the next section.

§3. Bilateral Hardy-Type Inequalities and Extension

Bilateral Hardy-Type Inequalities

Recall that in the definition of $D(f)$, the reference point $\theta\in E$ is free, hence it can be rewritten as

$$D(f)=\sum_{i\in E}\nu_i(\partial_i f)^2=:\|\partial f\|^2_{\nu,2},\qquad f\in L^2(\mu),$$

where $\|\cdot\|_{\mu,q}$ denotes the $L^q(\mu)$-norm and

$$\nu_i = \begin{cases} \nu_i^- = \mu_i a_i & -M-1 < i \leqslant \theta \\ \nu_i^+ = \mu_i b_i & \theta \leqslant i < N+1; \end{cases}$$

$$\partial_i f = \begin{cases} (\partial_i f)^- = f_{i-1} - f_i & -M-1 < i \leqslant \theta \\ (\partial_i f)^+ = f_{i+1} - f_i & \theta \leqslant i < N+1. \end{cases}$$

Thus, the Poincaré inequalities, in the DD case for instance, can be rewritten as

$$\sqrt{\lambda^{\mathrm{DD}}}\,\|f\|_{\mu,2} \leqslant \|\partial f\|_{\nu,2}, \qquad f \in L^2(\mu),\ \ f_{-M-1} = 0 = f_{N+1}.$$

This leads to the following generalization

$$\|f\|_{\mu,q} \leqslant A^{\mathrm{DD}} \|\partial f\|_{\nu,p}, \qquad f \in L^q(\mu),\ \ f_{-M-1} = 0 = f_{N+1}$$

for $p, q \in [1, \infty]$. This and the parallel inequalities with different boundary condition consist of the bilateral Hardy-type inequalities. For a large number of references and results on the Hardy inequalities, refer to [26, 29].

Except the typical L^2-case, a motivation of the last class of inequalities is studying the algebraic stability (ergodic case):

$$\|P_t f - \pi(f)\| \leqslant C\|f\|_1/t^\alpha, \qquad t > 0$$

for some $\alpha > 0$, here $\|\cdot\|_r$ is the $L^r(\mu)$-norm. This is equivalent to the Nash inequality

$$\|f - \pi(f)\|^{2+4/\gamma} \leqslant A_N D(f)\|f\|_1^{4/\gamma}, \qquad f \in L^2(\pi),\ \gamma > 2.$$

Which then is equivalent to the Sobolev-type inequality

$$\|f - \pi(f)\|^2_{2\gamma/(\gamma-2)} \leqslant A_S D(f), \qquad f \in L^2(\pi),\ \gamma > 2.$$

Clearly, the last one is a Hardy-type inequality with $q = 2\gamma/(\gamma-2)$ and $p = 2$.

In the continuous context, the next result is due to [10]. In the present discrete context, it is due to [27].

Theorem 4 Let $q \geqslant p > 1$ and $\widehat{\nu}_k = \nu_k^{1-p^*}$, $p^* = p/(p-1)$. Then we have $B_*^{\#} \leqslant A^{\#} \leqslant k_{q,p} B^{\#*}$, where

$$k_{q,p} = \Big[\frac{q-p}{p\,\mathsf{B}\big(p/(q-p),\, p(q-1)/(q-p)\big)}\Big]^{1/p-1/q} \qquad \text{if } q > p$$
$$= p^{1/p} p^{*1/p^*} \qquad \text{if } q = p,$$

and $\mathsf{B}(\alpha,\beta)$ is the Beta function, and in the DD case especially,

$$B^{\mathrm{DD}*}=\sup_{m\leqslant n}\frac{\mu[m,n]^{1/q}}{\left\{\widehat{\nu}[-M,m]^{-q/p^*}+\widehat{\nu}[n,N]^{-q/p^*}\right\}^{1/q}},$$
$$B_*^{\mathrm{DD}}=\sup_{m\leqslant n}\frac{\mu[m,n]^{1/q}}{\left\{\widehat{\nu}[-M,m]^{1-p}+\widehat{\nu}[n,N]^{1-p}\right\}^{1/p}}.$$

When $\#=\mathrm{DD}$ or NN, we also have $B_*^{\mathrm{DD}}\leqslant B^{\mathrm{DD}*}\leqslant 2^{1/p-1/q}B_*^{\mathrm{DD}}$. Besides, $2^{1/p-1/q}k_{q,p}\leqslant 2$ for $q\geqslant p$.

In short, we have the unified estimates: $B_*^{\#}\leqslant A^{\#}\leqslant 2B_*^{\#}$.

To save space, here and in what follows, we omit the quantities $B^{\mathrm{NN}*}$, B_*^{NN}, $B^{\mathrm{DN}*}=B_*^{\mathrm{DN}}$, and $B^{\mathrm{ND}*}=B_*^{\mathrm{ND}}$.

Extension to Normed Linear Space $(\mathbb{B},\|\cdot\|_{\mathbb{B}},\mu)$

Noting that $x\leqslant x\log x\leqslant x^{1+\varepsilon}$ for large x and then

$$\|f\|_1\leqslant \mathrm{Ent}(f)\leqslant \|f\|_{1+\varepsilon}^{1+\varepsilon}$$

for every $\varepsilon>0$, it is clear that an interpolation of L^p-spaces is needed in order to study the logarithmic Sobolev inequality. This leads to the setup of the normed linear space $\mathbb{B}$ introduced below.

Let $\mathbb{B}$ be a linear subset of Borel measurable functions on $(X,\mathscr{X},\mu)$ with norm $\|\cdot\|_{\mathbb{B}}$ having the following properties.

Hypotheses 5 (Hypotheses (H))

(H1) If $\mu(X)=\infty$, then $\mathbb{1}_K\in\mathbb{B}$ for each compact K.

(H2) If $h\in\mathbb{B}$ and $|f|\leqslant h$, then $f\in\mathbb{B}$.

(H3) $\|f\|_{\mathbb{B}}=\sup_{g\in\mathscr{G}}\int_X|f|g\mathrm{d}\mu$, where $\mathscr{G}\subset\mathscr{X}/\mathbb{R}_+$.

A typical example is $\mathscr{G}=L^p$ for some $p>1$, then $\mathbb{B}=L^{p^*}$. In the study of logarithmic Sobolev inequality, we use $\mathscr{G}=\left\{g\geqslant 0:\int_X \mathrm{e}^g\mathrm{d}\mu\leqslant \mathrm{e}^2+1\right\}$, where μ is a probability measure.

We can now study the following Hardy-type inequality

$$\||f|^q\|_{\mathbb{B}}^{1/q}\leqslant A_{\mathbb{B}}^{\#}\|\partial f\|_{\nu,p},\qquad f\in\mathbb{B}$$

with different boundary conditions as before. For instance, in the DD case: $f_{-M-1}=0=f_{N+1}$.

Theorem 6 ([1, 10]) Let $q \geqslant p > 1$. Then under (H), the optimal $A_{\mathbb{B}}^{\#}$ in the Hardy-type inequalities satisfies $B_{\mathbb{B}*}^{\#} \leqslant A_{\mathbb{B}}^{\#} \leqslant k_{q,p} B_{\mathbb{B}}^{\#*}$, where in the DD case especially,

$$B_{\mathbb{B}}^{\mathrm{DD}*} = \sup_{m \leqslant n} \frac{\|\mathbb{1}_{[m,n]}\|_{\mathbb{B}}^{1/q}}{\left\{\widehat{\nu}[-M,m]^{-q/p^*} + \widehat{\nu}[n,N]^{-q/p^*}\right\}^{1/q}}$$

$$B_{\mathbb{B}*}^{\mathrm{DD}} = \sup_{m \leqslant n} \frac{\|\mathbb{1}_{[m,n]}\|_{\mathbb{B}}^{1/q}}{\left\{\widehat{\nu}[-M,m]^{1-p} + \widehat{\nu}[n,N]^{1-p}\right\}^{1/p}}.$$

Shortly, we have the unified estimates: $B_{\mathbb{B}*}^{\#} \leqslant A_{\mathbb{B}}^{\#} \leqslant 2B_{\mathbb{B}*}^{\#}$.

We mention that in the NN case, in the last two results, we adopt either a stronger constraint that $q = 2 \geqslant p > 1$ in [10] or a different constraint replacing $\pi(f) = 0$ by another condition in [27].

The Hardy-type inequalities are an active research topic in harmonic analysis. For more information on it and for a popular report on the results in this section, refer to [10, 11, 15].

§4. The Case with Killing

We now turn to consider the birth-death processes with killing ($c_i \geqslant 0 : i \in E$). Assume that $c_i \not\equiv 0$ on $(-M, N)$, otherwise, we would return to Section 2. Its Q-matrix is the same as above except the diagonal elements $-(a_i + b_i)$ are replaced by $-(a_i + b_i + c_i)$, $i \in E$. When $N < \infty$, since we can replace c_N by $c_N + b_N$ if necessary, we may assume that $b_N = 0$ once $N < \infty$. Throughout this section, we will do so, as well as for a_{-M} for simplicity. Denote by Q^c this extended Q-matrix. The Dirichlet form now becomes

$$D^c(f) = \sum_{i \in E} \nu_i (\partial_i f)^2 + \sum_{i \in E} \mu_i c_i f_i^2, \qquad f \in L^2(\mu).$$

Q^c-Harmonic Function

The harmonic function h we are going to construct is non-zero everywhere on $[-M, N]$, and is local in the sense that $Q^c h = 0$ on $(-M, N)$ rather than on $[-M, N]$, which are different at the endpoints if M or N is finite. To do so, again, let $\theta \in E$ to be a reference point. Set

$$u_{\theta+i} = \frac{a_{\theta+i}}{b_{\theta+i}}, \quad v_{\theta+i} = \frac{c_{\theta+i}}{b_{\theta+i}}, \quad \xi_{\theta+i} = 1 + u_{\theta+i} + v_{\theta+i}, \quad r_{\theta+i} = \frac{h_{\theta+i}}{h_{\theta+i+1}},$$
$$0 \leqslant i < N + 1 - \theta.$$

When $1 \leqslant n < N+1-\theta$, from the harmonic equation, it follows that

$$1 = \xi_{\theta+n} r_{\theta+n} - u_{\theta+n} r_{\theta+n-1} r_{\theta+n} = r_{\theta+n}(\xi_{\theta+n} - u_{\theta+n} r_{\theta+n-1}).$$

From this, we obtain a recursive equation on the right-hand side (i.e. $n \geqslant 1$):

$$r_{\theta+n} = \frac{1}{\xi_{\theta+n} - u_{\theta+n} r_{\theta+n-1}} = \cfrac{1}{\xi_{\theta+n} - \cfrac{u_{\theta+n}}{\xi_{\theta+n-1} - \cfrac{u_{\theta+n-1}}{\ddots\ \xi_{\theta+1} - u_{\theta+1} r_\theta}}}, \quad 1 \leqslant n < N+1-\theta.$$

In parallel, set

$$u_{\theta+i} = \frac{b_{\theta+i}}{a_{\theta+i}}, \quad v_{\theta+i} = \frac{c_{\theta+i}}{a_{\theta+i}}, \quad \xi_{\theta+i} = 1 + u_{\theta+i} + v_{\theta+i}, \quad r_{\theta+i} = \frac{h_{\theta+i+1}}{h_{\theta+i}}, \quad -M-1-\theta < i \leqslant -1.$$

When $-M-\theta < n \leqslant -1$, applying once again the harmonic equation, we obtain

$$1 = \xi_{\theta+n} r_{\theta+n-1} - u_{\theta+n} r_{\theta+n-1} r_{\theta+n} = r_{\theta+n-1}(\xi_{\theta+n} - u_{\theta+n} r_{\theta+n}).$$

Hence we obtain a recursive equation on the left-hand side (i.e. $n \leqslant -1$):

$$r_{\theta+n} = \frac{1}{\xi_{\theta+n+1} - u_{\theta+n+1} r_{\theta+n+1}} = \cfrac{1}{\xi_{\theta+n+1} - \cfrac{u_{\theta+n+1}}{\xi_{\theta+n+2} - \cfrac{u_{\theta+n+2}}{\ddots\ \xi_{\theta-1} - u_{\theta-1} r_{\theta-1}}}}, \quad -M-1-\theta < n \leqslant -1.$$

Now, define the initial values

$$r_{\theta-1} = r_\theta = 1 - \frac{c_\theta}{a_\theta + b_\theta + c_\theta}.$$

This comes from the harmonic equation at θ (by convention $h_\theta = 1$). Actually, these two numbers $r_{\theta-1}$ and r_θ satisfying the equation are not necessarily equal, there is a freedom here. From this, it follows that the solution r_n, even allowing up to a constant, may still not be unique. Hence the function h defined below may also not be unique. Finally, a

required positive harmonic function $(h_n : n \in E)$ is given by

$$h_{\theta+n} = \begin{cases} 1, & n = 0 \\ \prod_{k=0}^{n-1} r_{\theta+k}^{-1}, & 1 \leqslant n < N+1-\theta \\ \prod_{k=1}^{-n} r_{\theta-k}, & -M-1-\theta < n \leqslant -1. \end{cases}$$

Here we fix such a specific h, the subsequent discussions do not depend on the choice of h. Even though the construction looks lengthy, it is still quite easy to show by induction that

$$c_i \equiv 0 \Longrightarrow h_i \equiv 1.$$

Thus, every result with $c_i \equiv 0$ can be obtained from the one containing h by setting $h_i \equiv 1$. Remark that when $M < \infty$, we may set $\theta = -M$, reducing the problem to the one with half space on the right. Similarly, there may be only the half space on the left.

Isospectral Operators

Having h at hand, we can now define a dual birth-death Q-matrix $\widetilde{Q}$ as follows:

$$\widetilde{a}_i = a_i \frac{h_{i-1}}{h_i}, \qquad \widetilde{b}_i = b_i \frac{h_{i+1}}{h_i}, \qquad i \in E,$$

On $\{-M+1, \ldots, N-1\}$, we have $\widetilde{c}_i = 0$; but

$$\widetilde{c}_N = c_N + a_N\Big(1 - \frac{h_{N-1}}{h_N}\Big), \qquad \text{if } N < \infty,$$
$$\widetilde{c}_{-M} = c_{-M} + b_{-M}\Big(1 - \frac{h_{-M+1}}{h_{-M}}\Big), \qquad \text{if } M < \infty.$$

In other words, the dual birth-death matrix is conservative everywhere, except at the finite boundaries, at which it is non-conservative. Next, let

$$\widetilde{\mu} = h^2\mu : \ \widetilde{\mu}_i = h_i^2\mu_i, \qquad i \in E,$$

where the measure μ is the same as what we have used before. When $M = \infty = N$, there are four boundary conditions as in §2 according to $\sum_{k \geqslant \theta} \widetilde{\mu}_k$ and $\sum_{k \leqslant \theta} \widetilde{\mu}_k$ being finite or infinite, respectively. Certainly, it can be simpler in some special case, $M < \infty$ for instance. Then one may fix $\theta = -M$ and ignore the DD case. When $N < \infty$, it is the ND case by the definition of h. When $N = \infty$, there are either the ND or NN case according to $\sum_{k \geqslant \theta} \widetilde{\mu}_k = \infty$ or $< \infty$, respectively.

Under the mapping $f \to \widetilde{f} := f/h$, using the quadratic form D^c with a given domain $\mathscr{D}(D^c)$, one deduces on $L^2(\widetilde{\mu})$ a dual quadratic form $\widetilde{D}$ with domain $\mathscr{D}\big(\widetilde{D}\big)$ as follows:

$$\mathscr{D}\big(\widetilde{D}\big) = \big\{\widetilde{f} \in L^2(\widetilde{\mu}) : \widetilde{f}h \in \mathscr{D}(D^c)\big\}.$$

The next result is taken from [22; Section 2].

Theorem 7 The quadratic form $(D^c, \mathscr{D}(D^c))$ on $L^2(\mu)$ and the one $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$ on $L^2(\widetilde{\mu})$ have the same spectrum. In details, the mapping $f \to \widetilde{f} := f/h$ from $L^2(\mu)$ to $L^2(\widetilde{\mu})$ possesses the following properties:

(i) one-to-one and isometric.

(ii) $f \in \mathscr{D}(D^c)$ iff $\widetilde{f} \in \mathscr{D}\big(\widetilde{D}\big)$ and $\widetilde{D}\big(\widetilde{f}\big) = D^c(f)$.

As a corollary of the theorem, we see that $(D^c, \mathscr{D}(D^c))$ and $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$ have the same principal (or the first non-trivial) eigenvalues $\lambda_c^{\#} = \widetilde{\lambda}^{\#}$. Applying Theorem 3 to $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$, we have the basic estimates for $\widetilde{\lambda}^{\#}$ in terms of $\widetilde{\kappa}^{\#}$ which is then the same as $\widetilde{\kappa}_c^{\#}$ used for the basic estimates of $\lambda_c^{\#}$. We have thus obtained the following result.

Corollary 8 ([12, 22]) The basic estimates hold:

$$\big(4\widetilde{\kappa}_c^{\#}\big)^{-1} \leqslant \lambda_c^{\#} \leqslant \big(\widetilde{\kappa}_c^{\#}\big)^{-1},$$

where $\widetilde{\kappa}_c^{\#}$ are obtained from $\kappa^{\#}$ given in Theorem 3 plus the remark after the theorem, replacing μ and $\widehat{\nu}$ by

$$\widetilde{\mu}_j = \mu_j h_j^2, \qquad \widehat{\widetilde{\nu}}_j = \frac{\widehat{\nu}_j}{h_j h_{j+1}}, \qquad j \in E,$$

respectively.

We remark that in different from the other three cases, in the NN case, the minimal eigenvalue $\widetilde{\lambda}^{\min}$ of $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$ is zero, our $\widetilde{\lambda}^{\mathrm{NN}}$ is the first non-trivial eigenvalue (also called spectral gap) of $\big(\widetilde{D}, \mathscr{D}\big(\widetilde{D}\big)\big)$. This leads to the same conclusion about $\lambda_c^{\min}\,(=0)$ and λ_c^{NN} of $(D^c, \mathscr{D}(D^c))$.

Discrete Spectrum

The last result of this section below exhibits the power of Theorem 7. We say that the quadratic form $(D^c, \mathscr{D}(D^c))$ on $L^2(\mu)$ has discrete spectrum if its spectrum consists of only eigenvalues with finite multiplicity. Since an operator on a finite space is compact and hence must have discrete spectrum, we need only consider an infinite state space. Next, since the whole line can be split into two half lines, without loss of generality, we assume that $E = \{0, 1, \ldots\}$.

To state our result, we specify two domains for D^c. The first one is simply

$$\mathscr{D}_{\max}(D^c) = \{f \in L^2(\mu) : D^c(f) < \infty\}.$$

The second one $\mathscr{D}_{\min}(D^c)$ is the minimal closure of the set

$$\{f \in L^2(\mu) : f \text{ has finite support}\}$$

with respect to the norm $\|\cdot\|_D$: $\|f\|_D^2 = \|f\|_{L^2(\mu)}^2 + D^c(f)$.

The result below is taken from [12; Theorem 2.1].

Theorem 9 Set $E = \{0, 1, \ldots\}$.

(i) Let $\sum_{k=0}^{\infty} (h_k h_{k+1} \mu_k b_k)^{-1} < \infty$. Then $(D^c, \mathscr{D}_{\min}(D^c))$ has discrete spectrum iff

$$\lim_{n\to\infty} \sum_{j=0}^{n} \mu_j h_j^2 \sum_{k=n}^{\infty} \frac{1}{h_k h_{k+1} \mu_k b_k} = 0.$$

(ii) Let $\sum_{j=0}^{\infty} \mu_j h_j^2 < \infty$. Then $(D^c, \mathscr{D}_{\max}(D^c))$ has discrete spectrum iff

$$\lim_{n\to\infty} \sum_{j=n+1}^{\infty} \mu_j h_j^2 \sum_{k=0}^{n} \frac{1}{h_k h_{k+1} \mu_k b_k} = 0.$$

(iii) Let $\sum_{k=0}^{\infty} (h_k h_{k+1} \mu_k b_k)^{-1} = \infty = \sum_{j=0}^{\infty} \mu_j h_j^2$. Then the spectrum of each of $(D^c, \mathscr{D}_{\min}(D^c))$ and $(D^c, \mathscr{D}_{\max}(D^c))$ is not discrete.

When $c_i \equiv 0$, Theorem 9 (ii) comes from [28; Theorem 1.2]. Starting from which and using Theorem 7, one can prove Theorem 9.

To conclude this section, we mentioned that even though up to now, we have studied in one direction only: reducing the case that $c_i|_{(0,N)} \not\equiv 0$ to the one $c_i|_{(0,N)} \equiv 0$. Certainly, we can go to the opposite direction: extending the result from $c_i|_{(0,N)} \equiv 0$ to $c_i|_{(0,N)} \not\equiv 0$. This is actually much easier but is very powerful (cf. [12, 22]).

§5. Further Reading

The Optimal Factor

The optimal universal factor 4 in Theorem 3 or more general one $k_{q,\,p}$ ($\leqslant 2$) in Theorem 4 can be improved further. For instance, the latter one can be improved up to $\sqrt{2}$ as in many practical models. Quite often, we can even produce new upper and lower bounds which are nearly the same. Refer to [7, 11] for the application to the Riemannian geometry and to [14] for more recent progress.

Dual Variational Formulas

Our main contribution to the topic is a class of dual variational formulas, from which we deduce the corresponding approximating procedures for the upper/lower bounds of the optimal constants $\lambda^{\#}$ or A in the inequalities. What we mentioned in the last paragraph about the $\sqrt{2}$-estimates come from the first step (explicit) of our approximating procedures, which then deduces the basic estimates given in Theorem 3. This shows that our approach is essentially different from those used in harmonic analysis. For this approach, one of the key observations is a mimic of the eigenfunction corresponding to the principal eigenvalue as explained in [3; the paragraph including (3.13)]. This story has lasted for a long period, for the results up to 2004, refer to [3]. For the later developments, refer to [6, 8, 9, 14, 19–21]. Surprisingly, much precise results can be obtained by a numerical algorithm, refer to [17].

Even though we have almost not touched the diffusions in this paper, the story is parallel for one-dimensional diffusions (included in the references just listed), sometimes, may be easier. For instance, in the diffusion context, the logarithmic Sobolev inequality is equivalent to the exponential stability of the semigroup in the entropy. However, it is not so in the discrete context. For birth-death processes, a criterion for the exponential stability in entropy has been opened for quite a long time, refer to [4] for more information.

Higher Dimension

The one-dimensional results are often essential in the study of higher dimensional one, as we used many times before, refer to [2; Theorem 14.10] or [3; Chapter 9], as well as [5, 7, 17]. However, in the higher dimensions, the topic is still largely open.

As mentioned before, because of a challenge of phase transitions which belong to the infinite-dimensional mathematics, we re-examined the finite-dimensional mathematical tools, and came back to dimension one. In contract to the finite-dimensional situation, the infinite-dimensional mathematics (needed not only in physics, but also in biology, big data, networks, et al.) is still an undeveloped area: there is a sea of open problems, the known results consist of a few of islands only.

Acknowledgments This paper is an extension of a plenary lecture presented at "24th International Workshop on Matrices and Statistics", the author acknowledges a kind invitation by the Scientific Program Committee, especially Professor Jeffrey J. Hunter. An extended abstract of the talk has appeared in [16]. The talk was also presented at the conference on "Markov Processes and Stochastic Models: In honor of the 80$^{\text{th}}$ Birthday of

Professor Zhen-Ting Hou" (June 23–25, 2015, Central South University). The same topic in the context of diffusion was talked at "International Conference on Stochastic Analysis and Related Topics" (August 3–8, 2015, Wuhan University), and also talked briefly in a workshop (April 23, 2015) with some probabilistic guests from Taiwan at Jiangsu Normal University. The author acknowledges the invitations from the organizers and the financial support from their institutes.

The references given below are only that the talk is based on. It is regretted that a large number of publications in the active research area is omitted here, otherwise, the list would be too long. For more references on the related sub-topics, the reader is urged to look at the related papers below.

References

[1] Chen M F. Variational formulas of Poincaré-type inequalities for birth-death processes [J]. *Acta Math. Sin. Eng. Ser.*, 2003, **19(4)**: 625–644.

[2] Chen M F. *From Markov Chains to Non-equilibrium Particle Systems* [M]. 2nd ed. (1st ed., 1992) Singapore: World Scientific, 2004.

[3] Chen M F. *Eigenvalues, Inequalities, and Ergodic Theory* [M]. London: Springer, 2005.

[4] Chen M F. Exponential convergence rate in entropy [J]. *Front. Math. China*, 2007, **2(3)**: 329–358.

[5] Chen M F. Spectral gap and logarithmic Sobolev constant for continuous spin systems [J]. *Acta Math. Sin. Eng. Ser.*, 2008, **24(5)**: 705–736.

[6] Chen M F. Speed of stability for birth-death processes [J]. *Front. Math. China*, 2010, **5(3)**: 379–515.

[7] Chen M F. General estimate of the first eigenvalue on manifolds [J]. *Front. Math. China*, 2011, **6(6)**, 1025–1043.

[8] Chen M F. Basic estimates of stability rate for one-dimensional diffusions [M] // Barbour A, Chan H P, Siegmund D. *Lecture Notes in Statistics – Proceedings 205: Probability Approximations and Beyond*. New York: Springer, 2012: 75–99.

[9] Chen M F. Lower bounds of principal eigenvalue in dimension one [J]. *Front. Math. China*, 2012, **7(4)**: 645–668.

[10] Chen M F. Bilateral Hardy-type inequalities [J]. *Acta Math. Sin. Eng. Ser.*, 2013, **29(1)**: 1–32.

[11] Chen M F. Bilateral Hardy-type inequalities and application to geometry [J]. *Mathmedia*, 2013, **37(2)**: 12–32; *Shuxue Tongbao*, 2013, **52(8/9)**: 1–6. (in Chinese)

[12] Chen M F. Criteria for discrete spectrum of 1D operators [J]. *Commu. Math. Stat.*, 2014, **2(3)**: 279–309.

[13] Chen M F. Criteria for two spectral problems of 1D operators [J]. *Sci. Sin. Math.*, 2015, **45(5)**: 429–438. (in Chinese)

[14] Chen M F. The optimal constant in Hardy-type inequalities [J]. *Acta Math. Sin. Eng. Ser.*, 2015, **31(5)**: 731–754.

[15] Chen M F. Progress on Hardy-type inequalities [M] // Chen Z Q, Jacob N, Takeda M, et al. *Festschrift Masatoshi Fukushima*. Singapore: World Scientific, 2015: 131–142.

[16] Chen M F. Unified speed estimation of various stabilities (extended abstract) [C] // Hunter J, Puntanen S, Von Rosen D. *Proceedings of the 24th International Workshop on Matrices and Statistics, held in May 2015 at Haikou, Hainan, China*. Special Issue of Special Matrices, 2015.

[17] Chen M F. Efficient initials for computing the maximal eigenpair [J]. Preprint, 2016.

[18] Chen M F, Mao Y H. *Introduction to Stochastic Processes* [M]. Beijing: Higher Edu. Press, 2007. (in Chinese)

[19] Chen M F, Wang L D, Zhang Y H. Mixed principal eigenvalues in dimension one [J]. *Front. Math. China*, 2013, **8(2)**: 317–343.

[20] Chen M F, Wang L D, Zhang Y H. Mixed eigenvalues of discrete p-Laplacian [J]. *Front. Math. China*, 2014, **9(6)**: 1261–1292.

[21] Chen M F, Wang L D, Zhang Y H. Mixed eigenvalues of p-Laplacian [J]. *Front. Math. China*, 2015, **10(2)**: 249–274.

[22] Chen M F, Zhang X. Isospectral operators [J]. *Commu. Math. Stat.*, 2014, **2(1)**: 17–32.

[23] Gross L. Logarithmic Sobolev inequalities [J]. *Amer. J. Math.*, 1976, **97(4)**: 1061–1083.

[24] Gross L. Logarithmic Sobolev inequalities and contractivity properties of semigroups [M] // Fabes E, Fukushima M, Gross L, et al. *Lecture Notes in Mathematics 1563: Dirichlet Forms*. Berlin: Springer, 1993: 54–88.

[25] Hua L K. Mathematical theory of global optimization on planned economy, (II) and (III) [J]. *Chinese Sci. Bull.*, 1984, **29(13)**: 769–772. (in Chinese)

[26] Kufner A, Maligranda L, Persson L E. *The Hardy Inequality: About its History and Some Related Results* [M]. Pilsen: Vydavatelsky Servis, 2007.

[27] Liao Z W. Discrete weighted Hardy inequalities with different kinds of boundary conditions [OL]. 2015 [2015-08-19]. http://arxiv.org/abs/1508.04601.

[28] Mao Y H. On the empty essential spectrum for Markov processes in dimension one [J]. *Acta Math. Sin. Eng. Ser.*, 2006, **22(3)**: 807–812.

[29] Opic B, Kufner A. *Hardy-type Inequalities* [M]. New York: Longman, 1990.

各种稳定性统一的速度估计

陈 木 法

(北京师范大学数学科学学院; 北京师范大学数学与复杂系统教育部重点实验室, 北京, 100875)

摘 要: 为研究一些无穷维对象(例如统计物理中的相变), 人们发展了若干数学工具, 其一为各种稳定性/不稳定性的速度估计. 本文对于最简单的一类马尔可夫过程—生灭过程的各种稳定性/不稳定性, 汇集了一些预料不到的、统一的、近乎精确的基本估计. 还讨论了若干背景和扩充. 本文源于在几个国际会议上的报告.

关键词: 稳定性; 第一非平凡特征值; Hardy型不等式; 杀死; 速度估计; 判准; 生灭过程

中图分类号: O211.62

Front. Math. China 2016, 11(6): 1379–1418
DOI 10.1007/s11464-016-0573-4

RESEARCH ARTICLE

Efficient initials for computing maximal eigenpair

Mu-Fa CHEN

School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, China

Abstract This paper introduces some efficient initials for a well-known algorithm (an inverse iteration) for computing the maximal eigenpair of a class of real matrices. The initials not only avoid the collapse of the algorithm but are also unexpectedly efficient. The initials presented here are based on our analytic estimates of the maximal eigenvalue and a mimic of its eigenvector for many years of accumulation in the study of stochastic stability speed. In parallel, the same problem for computing the next to the maximal eigenpair is also studied.

Keywords Perron-Frobenius theorem, power iteration, Rayleigh quotient iteration, efficient initial, tridiagonal matrix, Q-matrix
MSC 15A18, 65F15, 93E15, 60J27

1 Introduction. Two algorithms and a typical example

Consider a nonnegative irreducible matrix $A = (a_{ij})$ on $E := \{0, 1, \ldots, N\}$, $N < \infty$. By the well-known Perron-Frobenius theorem, the matrix has uniquely a positive eigenvalue $\rho(A)$ having positive left-eigenvector and positive right-eigenvector. Moreover, both the left-eigenspace and the right-eigenspace of $\rho(A)$ have dimension one. This eigenvalue is maximal in the sense that for every other eigenvalue λ_k, we have $\rho(A) \geqslant |\lambda_k|$. The last equality sign appears only if A has a period $p > 1$. For instance, for

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

we have $p = 2$ and the eigenvalues of A are $\pm\sqrt{2}$ and 0. However, we may

Received October 22, 2015; accepted June 23, 2016
E-mail: mfchen@bnu.edu.cn

assume that $\rho(A) > |\lambda_k|$ for every other eigenvalue λ_k. Actually, if $\lambda = \rho e^{i\theta}$ with $\theta \neq k\pi/2$ for every odd $k \in \mathbb{Z}$, then for every $\varepsilon > 0$, we have $\rho + \varepsilon > |\rho e^{i\theta} + \varepsilon|$. This means that the required assertion holds for the shifted pair $\rho + \varepsilon$ and $\lambda + \varepsilon$. In other words, an analog of the Perron-Frobenius theorem is meaningful for the matrices having nonnegative off-diagonal elements only, their diagonal elements can be arbitrary but real. By a shift if necessary, such a matrix can be transformed into a nonnegative one: the maximal eigenvector is kept but their maximal eigenvalues are shifted from one to the other. In this paper, we are interested in computing $\rho(A)$ and its corresponding eigenvector. This is a very important problem, we will come back to its motivation in the next section.

There are mainly two popular algorithms for this problem. Unless otherwise stated, the eigenvector below means the right-eigenvector. Then, the maximal eigenpair (the maximal eigenvalue and its eigenvector) is denoted by $(\rho(A), g)$.

Power iteration *Given an initial vector $v_0 \in \mathbb{R}^{N+1}$ having a nonzero component in the direction of g with $\|v_0\| = 1$, define*

$$v_k = \frac{Av_{k-1}}{\|Av_{k-1}\|}, \quad z_k = \|Av_k\|, \quad k \geqslant 1, \tag{1}$$

where $\|\cdot\|$ is an arbitrary but fixed vector norm. Then v_k converges to the eigenvector g of $\rho(A)$ and z_k converges to $\rho(A)$ as $k \to \infty$.

Even it is not necessary, in the next algorithm, we fix the vector norm to be the Euclidean one (or equivalently, the ℓ^2-norm). Actually, a refined choice is using the inner product and the norm in the space $L^2(\mu)$ for a suitable measure μ to be specified case by case, as illustrated by the improved algorithm given at the end of Sections 3 and 4. See also Section 6.

Rayleigh quotient iteration (a variation of inverse iteration) *Choose a pair (z_0, v_0) as an approximation of $(\rho(A), g)$ with $v_0^* v_0 = 1$, where v^* is the transpose of v. In particular, one may set $z_0 = v_0^* A v_0$ for a given v_0. At the kth ($k \geqslant 1$) step, solve the linear equation in w_k:*

$$(A - z_{k-1}I)w_k = v_{k-1}, \tag{2}$$

where I is the identity matrix on E, and define

$$v_k = \frac{w_k}{\sqrt{w_k^* w_k}}, \quad z_k = v_k^* A v_k.$$

If the pair (z_0, v_0) is close enough to $(\rho(A), g)$, then (z_k, v_k) converges to $(\rho(A), g)$ as $k \to \infty$.

In what follows, unless otherwise stated, we fix z_0 to be the particular choice just defined. We now use a typical example (which will be studied time by time in the paper) to illustrate the effectiveness and their difference of the above two algorithms.

Example 1 Let $E = \{0, 1, \ldots, 7\}$ and

$$Q = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -5 & 2^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2^2 & -13 & 3^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3^2 & -25 & 4^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4^2 & -41 & 5^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5^2 & -61 & 6^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6^2 & -85 & 7^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7^2 & -113 \end{pmatrix}.$$

Then we have $\rho(Q) \approx -0.525268$ with eigenvector

$$\approx (55.878,\ 26.5271,\ 15.7059,\ 9.97983,\ 6.43129,\ 4.0251,\ 2.2954,\ 1.0)^*.$$

Starting from v_0 which is the normalized vector of

$$(1,\ 0.587624,\ 0.426178,\ 0.329975,\ 0.260701,\ 0.204394,\ 0.153593,\ 0.101142)^*,$$

the power iteration (applied to the nonnegative $A := 113\,I + Q$) arrives at $-0.525268 \approx \rho(Q)$ after 990 iterations. Here, we adopt the ℓ^1-norm:

$$\|v\| = \sum_{k \in E} |v_k|.$$

We now give a little more details about the computations for this example.

Table 1 gives us partial outputs of $(k, -z_k)$. The corresponding figure below shows that $-z_k$ decreases quickly for small k, but the convergence goes very slow for large k.

Table 1 Partial outputs of $(k, -z_k)$

k	$-z_k$	k	$-z_k$	k	$-z_k$
0	2.11289	14	0.877012	100	0.589332
1	1.42407	15	0.86311	120	0.574136
2	1.37537	16	0.850338	140	0.56279
3	1.22712	17	0.838548	160	0.554157
4	1.1711	18	0.827619	180	0.547529
5	1.10933	19	0.817449	200	0.542423
6	1.06711	20	0.807953	300	0.529909
7	1.02949	30	0.738257	400	0.526517
8	0.998685	40	0.694746	500	0.525603
9	0.971749	50	0.664453	600	0.525358
10	0.948331	60	0.641946	700	0.525292
11	0.927544	70	0.624473	800	0.525274
12	0.908975	80	0.610468	900	0.52527
13	0.892223	90	0.598963	$\geqslant 990$	0.525268

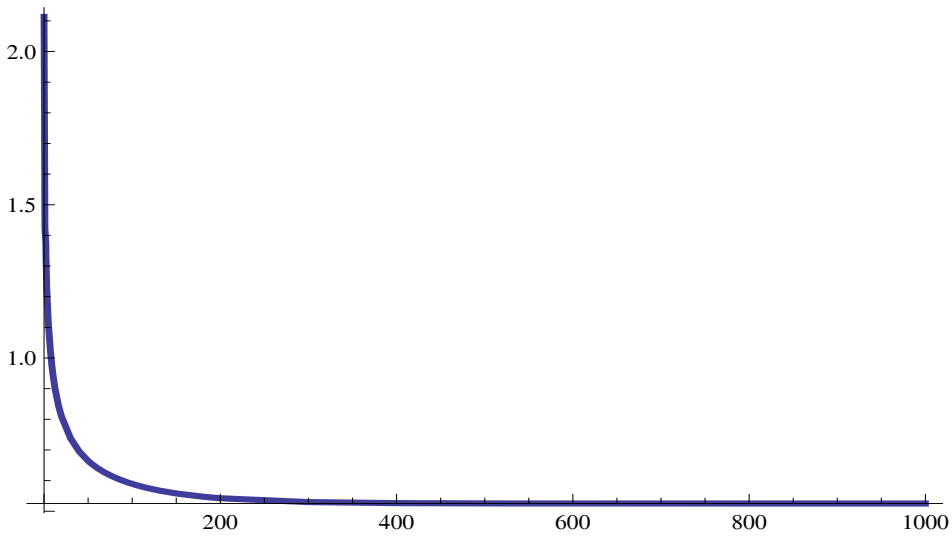


Fig. 1 Figure of $-z_k$ for $k = 0, 1, \ldots, 1000$

The advantage of the first algorithm is that it allows us to use a quite arbitrary positive initial vector v_0. The reason why the convergence of the example at the beginning steps goes quite fast is because we have used a very good initial v_0, as will be studied in Section 3. However, for larger k, the convergence becomes very slow, that is the limitation of this algorithm. From Fig. 1, it is clear that one may stop the computation at 300 iterations since then the results are almost the same. However, we keep going on until the six precisely significant digits as limited by a computer using Mathematica 9. The reason for doing so is for the comparison with other algorithms to be studied later.

Certainly, we expect the second algorithm to be more efficient. Now, what can we expect? Since this problem is often used in practice, we would be very happy if a new algorithm can reduce the number of iterations seriously, say, 500 for instance. Certainly, we would be very surprising if it can be reduced to 250. Let us think this question more carefully. Suppose that we are now interested in the maximal eigenvalue only, and suppose that we know it is located on $(0, 1)$ (actually, as we will see by Proposition 11 below, the maximal eigenvalue of $A := 113\,I + Q$ is located in $(0, 113)$). We may use the Golden Section Search (a famous method in optimization theory), its speed is about 0.618^{-1}. Then, to obtain the six precisely significant digits as in the last example, one needs at least 24 iterations since $10^{-6} \approx 0.618^{24}$. Suppose that we can adopt a faster algorithm, the Bisection Method. Then it requires about 20 iterations since $10^{-6} \approx 2^{-20}$. Hence, it is reasonable if an algorithm uses more than 20 iterations to arrive at the same precise level. Having this analysis in mind, we were shocked when the next result came to us.

Example 2 The matrix Q and the initial vector v_0 are the same as in the last example but we now adopt the ℓ^2-norm. The Rayleigh quotient iteration

(applied to Q) starts at

$$z_0 = v_0^* Q v_0 \approx -0.78458$$

and then arrives at the same result as in the last example at the second step:

$$z_1 \approx -0.528215, \quad z_2 \approx -0.525268.$$

Example 2 is the main illustrating example (which will be further improved by Example 7 below) of this paper. It shows that the second algorithm can be extremely powerful. The key to this result is that we have chosen an efficient initial vector v_0 and then the resulting z_0 is also close to $\rho(Q)$. It may be the position to compare the use of ℓ^1-norm and ℓ^2-one. Let everything be the same as in the last example but replacing the ℓ^2-norm by the ℓ^1-one. Then the iteration starts at $z_0 \approx -0.367937$ and arrives at the same result at the third step:

$$z_1 \approx -0.509272, \quad z_2 \approx -0.52509, \quad z_3 \approx -0.525268.$$

The result comes with no surprising: it is easier to use the ℓ^1-norm in the computation but it is a little less efficient than using the ℓ^2-norm.

We have seen the power of the second algorithm. However, "too good" is dangerous. Each eigenvalue $\lambda_k \neq \rho(A)$ can be a pitfall of the algorithm provided either z_0 is close enough to λ_k or v_0 is close enough to the eigenvector g_k of λ_k. The next example illustrates the latter situation. For which a simpler v_0 deduces its corresponding z_0 to be more close to λ_2 rather than λ_3.

Here and in what follows, we often use the so-called Q-matrix

$$Q = (q_{ij} \colon i, j \in E),$$

which means that $q_{ij} \geqslant 0$ for every pair $i \neq j$ and $\sum_{j \in E} q_{ij} \leqslant 0$ for every $i \in E$. This implies the intrinsic use of probabilistic idea. For convenience, we often write by

$$0 < \lambda_0 < |\lambda_1| \leqslant |\lambda_2| \leqslant \cdots,$$

where $\{\lambda_j\}$ is the sequence of the eigenvalues of $-Q$. Then $\lambda_0 = -\rho(Q)$.

Example 3 The matrix Q is the same as the last example and we use again the ℓ^2-norm. Replace Q by $-Q$ (then the corresponding $z_k > 0$). Choose the initial vector v_0 to be the normalized uniform vector:

$$v_0 = \frac{1}{\sqrt{8}} \{1, 1, 1, 1, 1, 1, 1, 1\}.$$

Then with the particular choice given in the algorithm

$$z_0 = v_0^*(-Q)v_0 = 8,$$

we obtain the following output at the first 4 steps of the iterations:

$$z_1 \approx 4.78557, \quad z_2 \approx 5.67061, \quad z_3 \approx 5.91766, \quad z_4 \approx 5.91867 \approx \lambda_2.$$

The first two eigenvalues of $-Q$ are

$$\lambda_0 \approx 0.525268, \quad \lambda_1 \approx 2.00758, \quad \lambda_3 \approx 13.709,$$

respectively. Hence, the limit λ_2 is quite away from what we are interested in.

By the way, let us mention that in practice, we can stop our computation once the components of the first output v_1 have different signs, and try to choose a new initial pair (v_0, z_0). This is due to the fact that the maximal vector should be positive/negative up to a constant. Here, in the last example,

$$\begin{aligned} v_1 = (&-0.26762,\ 0.242432,\ -0.522646,\ -0.579319,\\ &-0.423469,\ -0.253452, -0.124365,\ -0.0425044)^*. \end{aligned}$$

Each of the components is negative except the second one.

The next example shows that we can still arrive at the expected result for a good initial z_0 even if v_0 is quite rough.

Example 4 Everything is the same as in the last example except

$$z_0 = 2.05768^{-1} \approx 0.485985.$$

Then $\{z_k\}$ approaches to λ_0 at the second step:

$$z_1 \approx 0.525998, \quad z_2 \approx 0.525268.$$

This paper is organized as follows. In the next section, we first review the five sources of the motivation for our problem. Then we recall the known convergence of these algorithms. From the above examples, we have seen that the second algorithm is much more attractive. To which, we need a careful design in choosing the initial pair (v_0, z_0). Clearly, an efficient initial pair is just a good estimate of the pair in advance. This itself is a hard topic in the study of eigenvalue problem and so it is understandable that the initial problem is still largely open in the eigenvalue computation theory. A complete, analytic (explicit) solution to this problem is presented in Section 3 first for tridiagonal matrices (after a suitable relabeling if necessary), and then for a class of more general matrices in terms of the so-called Lanczos tridiagonalization procedure. The main extension to the general situation is presented in Section 4 which consists of two subsections. In the first one, we concentrated on the construction of z_0 for a fixed simplest v_0. The second one in even more technical, in which we are mainly working on the construction of v_0. A number of examples are illustrated, case by case, for the results in the paper. It is remarkable that only the one-step iteration scheme, as illustrated by the two algorithms used above, is used in the paper. In Section 5, we make either additional proofs of some results in the main context, or additional remarks on related problems. In particular, we prove a convergence result of our approximating procedure for the principal eigenvalue of birth–death processes which have been studied for a long period up to now. A summary for the use

of the algorithms up to Section 4 is given at the end of Section 5. The study on the next eigenpair is delayed to Section 6.

2 Motivation of problem and convergence of algorithms

In this section, for the reader's convenience, we recall briefly the motivation of our problem and the well-known convergence of the two algorithms introduced in the first section.

2.1 Motivation

It seems not necessary to mention the value of the study on the problem since the matrix eigenvalue computation is used almost everywhere. The next five sources reflect more or less the road where we started and finally arrive here.

Google's PageRank

When we search an expression from the network, a large number of webpages are collected. The question is how to output them on the screen of our computer. For this, we need to rank the pages. The procedure goes as follows. According to the connections of the websites, we get a nonnegative matrix A. To which we have the largest eigenvalue $\rho(A)$ and its corresponding positive left-eigenvalue. The normalized left-eigenvector gives us an order of the webpages, that is the PageRank as we required. Nowadays, there are a large number of publications on Google's PageRank, see for instance [12], in which the power iteration is studied but not the inverse iteration.

Global optimization of planned economy

Regarding the matrix A as a structure matrix in economy, Hua [11] proved that the optimal input of the planned economy is the left-eigenvector u of $\rho(A)$. Surprisingly, Hua [11] also proved that if one uses a different input rather than u, then the economy will go to collapse (i.e., some components of the product in the economic system will become less or equal to zero). Mathematically, this situation is very much the same as the last one, but in a completely different context. As far as I know, the practical algorithms for $(u, \rho(A))$ were not studied carefully during that period, except a formula was mentioned in [11]:

$$\rho(A) = \lim_{\ell\to\infty} \Big(\frac{\text{Trace } (A^\ell)}{N+1}\Big)^{1/\ell}.$$

Stationary distribution of time-discrete Markov chain

If A itself is a transition probability matrix, then the left-eigenvector corresponding to the largest eigenvalue one is nothing but the stationary distribution of the corresponding Markov chain. This explains the stability meaning in the two situations just discussed above. Based on this idea, we obtained a probabilistic proof of Hua's collapse theorem. Refer also to [4, Chapter 10] for additional story and related references.

Computing the stationary distribution of a given Markov chain is very

important in practice and so has been studied quite a lot in the past years, including the so-called Markov Chain Monte Carlo (MCMC), perfect/backward coupling, and so on.

Exponential decay of time-continuous Markov chain

The maximal eigenvalue $\rho(Q)$ in Example 1 describes the exponential decay rate of the Markov chain with semigroup $(P_t = \mathrm{e}^{tQ} : t \geqslant 0)$. The present paper is based on our study on this topic, as will be seen from the subsequent sections.

Phase transitions

The last topic and the investigation on related stability speed are actually motivated from the study on phase transitions in statistical mechanics (cf. [3,4] for more references within). This is a challenge topic in mathematics since it is mainly in infinite-dimensional setting. To which, the mathematical tools are rather limited. Therefore, we have to look for new tools or develop some known traditional tools. To this end, we have already visited several branches of mathematics, including the computation theory. We are now glad to be able to say something on the last field after a long trip of the study.

In the second part of this section, we review some well-known facts on the convergence of the algorithms.

2.2 Convergence of algorithms

Here is the convergence of the power iteration. In this subsection, we suppose that the given matrix A (not necessarily nonnegative) has the dominant eigenvalue λ_0 (i.e., $|\lambda_0| > |\lambda_j|$ for all other eigenvalues λ_j) which is simple. The extension to the periodic situation is also possible, but is omitted here, one simply replaces the convergence of the original sequence by a subsequence.

Lemma 5 *Suppose that the initial vector v_0 has a nonzero component in the direction of the dominant eigenvector g. Then*

$$v_k = \frac{A^k v_0}{\|A^k v_0\|} \to g, \quad v_k^* A v_k \to \lambda_0, \quad k \to \infty.$$

Moreover,

$$\lim_{n\to\infty} \frac{A^n v_0}{A^{n-1} v_0} = \lambda_0,$$

where for given vectors u and v, the ratio u/v is understood as the quotient function of the functions u and v.

Proof Suppose that the eigenvalues are all different for simplicity. Otherwise, one simply uses the Jordan representation of matrices. Write

$$v_0 = \sum_{j=0}^{N} c_j g_j$$

for some constants (c_j) with $g_0 = g$. Then $c_0 \neq 0$ by assumption and

$$A^k v_0 = \sum_{j=0}^{N} c_j \lambda_j^k g_j = c_0 \lambda_0^k \Big[g + \sum_{j=1}^{N} \frac{c_j}{c_0} \Big(\frac{\lambda_j}{\lambda_0} \Big)^k g_j \Big].$$

Since $|\lambda_j/\lambda_0| < 1$ for each $j \geqslant 1$ and $\|g\| = 1$, we have

$$\frac{A^k v_0}{\|A^k v_0\|} \to \frac{c_0}{|c_0|} g, \quad k \to \infty,$$

and then

$$v_k^* A v_k \to g^* A g = g^* \lambda_0 g = \lambda_0, \quad k \to \infty.$$

We have thus proved the main assertion of the lemma. The proof of the last assertion is similar. □

Clearly, the convergence speed in the lemma is

$$\Big| \frac{\lambda_1}{\lambda_0} \Big|^k, \quad |\lambda_1| := \max\{|\lambda_j| \colon j > 0\}.$$

The next result is the convergence for the inverse iteration.

Lemma 6 *Under the assumption of the last lemma, for each z close to λ_0, we have*

$$v_k = \frac{(A - zI)^{-k} v_0}{\|(A - zI)^{-k} v_0\|} \to g, \quad v_k^* A v_k \to \lambda_0, \quad k \to \infty.$$

Moreover,

$$\lim_{n \to \infty} \frac{(A - zI)^{-n} v_0}{(A - zI)^{-n+1} v_0} = \frac{1}{\lambda_0 - z}.$$

Proof Note that for z close to λ_0, the dominant eigenvalue of the matrix $(A - zI)^{-1}$ is $(\lambda_0 - z)^{-1}$ with the same dominant eigenvector g as the one for A. The proof is very much the same as the previous one. □

The iteration given in Lemma 6 is called the inverse iteration. It is remarkable that the convergence speed in this lemma is

$$\Big| \frac{\lambda_0 - z}{\lambda_1 - z} \Big|^k \sim \Big| \frac{\lambda_0 - z}{\lambda_1 - \lambda_0} \Big|^k$$

when z is sufficiently close to λ_0. At the kth step, replacing z by the Rayleigh quotient approximation $z_k = v_k^* A v_k$, we obtain the Rayleigh quotient iteration as described in the first section. Clearly, the last algorithm is an acceleration of the inverse iteration. The price is that the initial z_0 has to be chosen close to λ_0 which is usually not explicitly known. Otherwise, if z_0 is chosen close to some $\lambda_j \neq \lambda_0$, then a similar proof of Lemma 6 shows that $v_k^* A v_k$ converges to the pitfall $\lambda_j \neq \lambda_0$. In practice, once $z = z_0$ is chosen in a suitable neighborhood of λ_0, the sequence $z = z_k$ converges to λ_0 rapidly, as illustrated by Examples 2

and 4. More precisely, Example 1 applies the power iteration to $A := 113I + Q$, its convergence speed is

$$\sim \Big(\frac{113-\lambda_1}{113-\lambda_0}\Big)^k \approx \Big(\frac{113-2.00758}{113-0.525268}\Big)^k, \quad k \to \infty.$$

Examples 2 and 4 use the Rayleight quotient iteration which has the convergence speed

$$\sim \prod_{j=0}^{k} \frac{\lambda_0 - z_j}{\lambda_1 - z_j}, \quad k \to \infty.$$

Since $z_k \to \lambda_0$, the last convergence is much fast than the previous one. Honestly, this still does not answer the reason why the inverse algorithm in Example 2 can achieve the six significant digits at the second iteration.

3 Efficient initials. Tridiagonal case

Again, assume that $A = (a_{ij})$ on $E = \{0, 1, \ldots, N\}$, $N < \infty$, is irreducible and having non-negative off-diagonal elements. Assume also that the matrix is tridiagonal (after a suitable relabeling if necessary): $a_{ij} = 0$ unless $|i-j| \leqslant 1$. By a shift $Q := A - mI$ if necessary, where I is the identity matrix on E and

$$m = \max_{i\in E} \sum_{j\in E} a_{ij},$$

one may assume that

$$Q = \begin{pmatrix} -(b_0+c_0) & b_0 & 0 & 0 & \cdots \\ a_1 & -(a_1+b_1+c_1) & b_1 & 0 & \cdots \\ 0 & a_2 & -(a_2+b_2+c_2) & b_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & a_N & -(a_N+c_N) \end{pmatrix},$$

where $a_i, b_i > 0$, $c_i \geqslant 0$ but $c_i \not\equiv 0$. Define

$$\mu_0 = 1, \quad \mu_n = \mu_{n-1}\frac{b_{n-1}}{a_n} = \frac{b_0 b_1 \cdots b_{n-1}}{a_1 a_2 \cdots a_n}, \quad 1 \leqslant n \leqslant N.$$

We now split our discussion into two cases.

Case 1 Let

$$c_0 = c_1 = \cdots = c_{N-1} = 0.$$

Then we may assume that $c_N > 0$. Otherwise, Q has the trivial maximal eigenvalue 0 with eigenvector with components being one everywhere. In this case, we rewrite c_N as b_N, ignoring the sequence (c_i), and define

$$\varphi_n = \sum_{k=n}^{N} \frac{1}{\mu_k b_k}, \quad 0 \leqslant n \leqslant N. \tag{3}$$

Case 2 Let some of c_i $(i = 0, 1, \ldots, N-1)$ be positive. Then, we need more work. Define

$$r_0 = 1 + \frac{c_0}{b_0}, \quad r_n = 1 + \frac{a_n + c_n}{b_n} - \frac{a_n}{b_n r_{n-1}}, \quad 1 \leqslant n < N,$$
$$h_0 = 1, \quad h_n = h_{n-1} r_{n-1} = \prod_{k=0}^{n-1} r_k, \quad 1 \leqslant n \leqslant N,$$

and additionally,

$$h_{N+1} = c_N h_N + a_N (h_{N-1} - h_N).$$

Finally, define

$$\varphi_n = \sum_{k=n}^{N} \frac{1}{h_k h_{k+1} \mu_k b_k}, \quad 0 \leqslant n \leqslant N, \tag{4}$$

with a convention that $b_N = 1$ to save our notation.

We remark that in the special case that $c_0 = c_1 = \cdots = c_{N-1} = 0$, by induction, it is easy to check that

$$r_0 = r_1 = \cdots = r_{N-1} = 1,$$

and hence,

$$h_0 = h_1 = \cdots = h_N = 1.$$

Furthermore, $h_{N+1} = c_N$. Thus, once replacing c_N by b_N, we return to (3) from (4).

To state our algorithm, we need one more quantity:

$$\delta_1 = \max_{0 \leqslant n \leqslant N} \left[\sqrt{\varphi_n} \sum_{k=0}^{n} \mu_k h_k^2 \sqrt{\varphi_k} + \frac{1}{\sqrt{\varphi_n}} \sum_{j=n+1}^{N} \mu_j h_j^2 \varphi_j^{3/2} \right].$$

Rayleigh quotient iteration in tridiagonal case *For given tridiagonal matrix A, define m, (a_i, b_i, c_i), (h_i), (φ_i), and δ_1 as above. Set*

$$\widetilde{v}_0(i) = h_i \sqrt{\varphi_i}, \quad 0 \leqslant i \leqslant N, \quad v_0 = \frac{\widetilde{v}_0}{\sqrt{\widetilde{v}_0^* \widetilde{v}_0}}, \quad z_0 = \frac{1}{\delta_1}.$$

At the kth step $(k \geqslant 1)$, solve the linear equation in w_k:

$$(-Q - z_{k-1} I) w_k = v_{k-1}, \tag{5}$$

and define

$$v_k = \frac{w_k}{\sqrt{w_k^* w_k}}, \quad z_k = v_k^* (-Q) v_k.$$

Then v_k converges to g and $m - z_k$ converges to $\rho(A)$ as $k \to \infty$.

It is an essential point that the choice of z_0 avoids the collapse since we have known that $\lambda_0(Q) = \lambda_{\min}(-Q)$ (the minimal eigenvalue of $-Q$) $\geqslant \delta_1^{-1}$ by [5,

Corollary 3.3]. As an application of this result to Example 1, we have $c_i \equiv 0$ but $b_7 = 64$ and then $h_i \equiv 1$. We can define φ by (3) and then $\widetilde{v}_0 = \sqrt{\varphi}$ which is the one, up to a free factor $\sqrt{\varphi_0}$, used in Example 1. This is the meaning of "very good" claimed in the first section. We now compute the minimal eigenvalue of $-Q$ using not only $\widetilde{v}_0$ but also δ_1.

Example 7 The matrix Q and the vector $\widetilde{v}_0$ are the same as in Example 1:

$$(1,\ 0.587624,\ 0.426178,\ 0.329975,\ 0.260701,\ 0.204394,\ 0.153593,\ 0.101142)^*.$$

We have $\delta_1 = 2.05768$. Then, with the new $z_0 := \delta_1^{-1} \approx 0.485985$, the Rayleigh quotient iteration arrives at the expected estimate at the second step:

$$z_1 \approx 0.525313, \quad z_2 \approx 0.525268.$$

Comparing the approximation value of z_1 here and that in Example 2, it is clear that this result is sharper than Example 2 (see also the comment below Corollary 12).

Now, let us discuss the effectiveness of our algorithm with respect to the size N of the matrix. In computational mathematics, one often expects the number of iterations M grows up no more than N^α for some $\alpha > 0$. It is unusual if $M \approx \log N$ for large N. To this question, considering the basic Example 1 with varying N, the answer given below is worked out by Yue-Shuang Li using the algorithm introduced in this section and the software MatLab on a notebook. In the first line of Table 2, the reason we use $N+1$ rather than N is that the space is labeled starting at 0 but not 1.

Table 2 For different N, eigenvalue λ_0, its lower bound δ_1^{-1}, and z_1, z_2

$N+1$	$z_0 = \delta_1^{-1}$	z_1	$z_2 = \lambda_0$
100	0.348549	0.376437	0.376383
500	0.310195	0.338402	0.338329
1000	0.299089	0.32732	0.32724
5000	0.281156	0.308623	0.308529
7500	0.277865	0.305016	0.304918
10000	0.275762	0.30266	0.302561

Is it believable? Yes, we have justified the outputs in two different ways: in each case, first, the outputs starting from z_2 become the same (which actually coincides with the output of λ_0). Second, by using v_2, we can find upper/lower estimates $\overline{\xi}/\underline{\xi}$ of λ_0 such that $z_2 \in (\underline{\xi}, \overline{\xi})$, and moreover,

$$\frac{\overline{\xi}}{\underline{\xi}} \approx 1 + 10^{-5}.$$

The next example is due to Hua [11] in the study of economic optimization (cf. [4, Chapter 10]). Note that here we are studying the right-eigenvector, the matrix A used below is the transpose of the original one.

Example 8 *Let*

$$A = \frac{1}{100}\begin{pmatrix} 25 & 40 \\ 14 & 12 \end{pmatrix}.$$

Then

$$\rho(A) = \frac{37 + \sqrt{2409}}{200} \approx 0.430408.$$

With the initials:

$$v_0 \approx (0.429166,\ 0.220573)^*, \quad z_0 := \delta_1^{-1} \approx 0.212077,$$

the iteration arrives at the expected result at the second step $(n = 2)$:

$$0.65 - z_0 \approx 0.437923; \quad 0.65 - z_1 \approx 0.430603; \quad 0.65 - z_2 \approx 0.430408.$$

Proof First, we have $m = 65/100$ and then

$$Q = \begin{pmatrix} -\dfrac{2}{5} & \dfrac{2}{5} \\ \dfrac{7}{50} & -\dfrac{53}{100} \end{pmatrix}.$$

In this case, we ignore (c_i) but let $b_1 > 0$. Actually, we have

$$b_0 = \frac{2}{5}, \quad b_1 = \frac{39}{100}; \quad a_1 = \frac{7}{50}; \quad \mu_0 = 1,\ \mu_1 = \frac{20}{7}; \quad \varphi_0 = \frac{265}{78},\ \varphi_1 = \frac{35}{39}.$$

Therefore,

$$v_0 = \left(\sqrt{\frac{53}{67}}, \sqrt{\frac{14}{67}}\right), \quad z_0^{-1} = \frac{5(2809 + 40\sqrt{742}\,)}{4134}.$$

The conclusion now follows by our algorithm. □

An additional example for the algorithm presented in this section is delayed to Example 22.

Before moving further, let us introduce an algorithm for (and then a representation of) the solution to equation (5). This is mainly used in theoretic analysis rather than numerical computation. The idea is meaningful in a more general setup and comes from [9, Theorem 1.1, Proposition 2.6] plus a modification [6, Proposition 4.1]. Given a number $z \in \mathbb{R}$ and a vector v, consider the equation for the vector w:

$$Qw + zw = -v. \tag{6}$$

To do so, we need some notation. Fix i: $0 \leqslant i \leqslant N - 1$, and set

$$\alpha_\ell^{(i)} = \frac{1}{b_{i+\ell}} \begin{cases} c_{i+\ell} - z + a_{i+\ell}, & 1 = \ell \leqslant N - i, \\ c_{i+\ell} - z, & 2 \leqslant \ell \leqslant N - i. \end{cases}$$

Next, define the vector $G^{(i)}_{\cdot,1}$ by $G^{(i)}_{\ell,1} = \alpha^{(i)}_{\ell}$ for $\ell = 1, 2, \ldots, N-i$ and define recursively in $k = 2, 3, \ldots, N-i$, the vector $G^{(i)}_{\cdot,k}$ by

$$G^{(i)}_{\ell,k} = G^{(i)}_{\ell,k-1} + \alpha^{(i+k-1)}_{\ell-k+1} G^{(i)}_{k-1,k-1}, \quad \ell = k, k+1, \ldots, N-i. \tag{7}$$

Note that here for computing $G^{(i)}_{\cdot,k}$, we use only $G^{(i)}_{\cdot,k-1}$ but not the others $G^{(i)}_{\cdot,j}$ with $j \leqslant k-2$.

Proposition 9 *Let $N \geqslant 1$ and $G^{(\cdot)}_{0,0} \equiv 1$. Then the solution $w = (w_k \colon k \in E)$ to equation (6) has the following representation*:

$$w_n = \frac{v_N + M_{N-1}(v)}{c_N - z + M_{N-1}(c_\cdot - z)} \left[1 + N_{n-1}(c_\cdot - z)\right] - N_{n-1}(v), \quad 0 \leqslant n \leqslant N,$$

where for each vector h, $N_{-1}(h) = 0$ and

$$M_{N-1}(h) = c_N \sum_{j=0}^{N-1} \frac{h_j}{b_j} G^{(j)}_{N-j,N-j},$$
$$N_n(h) = \sum_{j=0}^{n} \frac{h_j}{b_j} \sum_{k=0}^{n-j} G^{(j)}_{k,k}, \quad 0 \leqslant n < N.$$

The proof of this result is delayed to Section 5.

From now on, we are going to treat general real matrices. This is a hard task and will be the main goal of the next section. Here, we study a special case only. In computational mathematics, there is a well-known Lanczos tridiagonalization procedure making a matrix to be tridiagonal one. That is, for a given A, constructing a nonsingular B such that $B^{-1}AB =: T$ becomes a tridiagonal matrix. We will come back to the procedure soon. Here is an example (the details are omitted).

Example 10 Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 1 \\ 3 & 2 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{10} & 3/\sqrt{13} \\ 0 & 3/\sqrt{10} & -2/\sqrt{13} \end{pmatrix}.$$

Then

$$T = B^{-1}AB = \begin{pmatrix} 1 & 11/\sqrt{10} & 0 \\ \sqrt{10} & 25/11 & 20\sqrt{130}/143 \\ 0 & \sqrt{130}/11 & 8/11 \end{pmatrix}.$$

We have

$$\rho(A) = \rho(T) = 3 + \sqrt{5} \approx 5.23607.$$

Our algorithm arrives at the same result at the second step of the iterations $(n = 2)$:

$$(n = 0)\ 5.43937; \quad (n = 1)\ 5.23996; \quad (n = 2)\ 5.23607.$$

It is the position to recommend an improved algorithm as follows. The point is to use the inner product $(\cdot,\cdot)_\mu$ and norm $\|\cdot\|_\mu$ in the space $L^2(\mu)$ since (μ_k) may not be a constant as in Example 1.

Improved algorithm Given $\widetilde{v}_0$ and δ_1 as above, redefine $v_0 = \widetilde{v}_0/\|\widetilde{v}_0\|_\mu$ and

$$z_0 = \xi\delta_1^{-1} + (1-\xi)(v_0, -Qv_0)_\mu, \quad \xi \in [0,1].$$

For $k \geqslant 1$, define w_k as before but redefine

$$v_k = \frac{w_k}{\|w_k\|_\mu}, \quad z_k = (v_k, -Qv_k)_\mu.$$

With $\xi = 7/8$, Example 7 and Table 2 are improved as Table 2′.

Table 2′ Example 7 and Table 2 are improved using new z_0 with $\xi = 7/8$

$N+1$	z_0	z_1	$z_2 = \lambda_0$	upper/lower
8	0.523309	0.525268	0.525268	$1+10^{-11}$
100	0.387333	0.376393	0.376383	$1+10^{-8}$
500	0.349147	0.338342	0.338329	$1+10^{-7}$
1000	0.338027	0.327254	0.32724	$1+10^{-7}$
5000	0.319895	0.30855	0.308529	$1+10^{-7}$
7500	0.316529	0.304942	0.304918	$1+10^{-7}$
10000	0.31437	0.302586	0.302561	$1+10^{-7}$

The last column is the order of the ratio of the upper and lower bounds of λ_0 in terms of v_2, as will be explained below, above Example 13.

Table 3 gives two more examples.

Table 3 Outputs using improved z_0 with $\xi = 7/8$

Example	z_0	z_1	$z_2 = \lambda_0$
8	0.436733	0.430407	0.430408
10	5.36161	5.23578	5.23607

Appendix of Section 3 Algorithm for Lanczos tridiagonalization

For a given A, the aim is choosing a nonsingular Q such that

$$Q^{-1}AQ = T = \begin{pmatrix} c_1 & b_1 & \cdots & \cdots & 0 \\ a_1 & c_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & \cdots & a_{n-1} & c_n \end{pmatrix}.$$

Note that the notation here is somehow different from the other part of the paper. To do so, we use the following column partitionings:

$$Q = [q_1 \mid q_2 \mid \cdots \mid q_n], \quad (Q^{-1})^* = \widetilde{Q} = [\widetilde{q}_1 \mid \widetilde{q}_2 \mid \cdots \mid \widetilde{q}_n].$$

Let

$$q_0 = 0, \quad \widetilde{q}_0 = 0, \quad b_0 = 0, \quad a_0 = 0.$$

Choose unit vectors q_1 and $\widetilde{q}_1$ such that $\widetilde{q}_1^* q_1 = 1$. Define

$$\begin{aligned} c_k &= \widetilde{q}_k^* A q_k, \quad k \geqslant 1, \\ r_k &= (A - c_k I) q_k - a_{k-1} q_{k-1}, \quad k \geqslant 1, \\ \widetilde{r}_k &= (A - c_k I)^* \widetilde{q}_k - b_{k-1} \widetilde{q}_{k-1}, \quad k \geqslant 1, \\ b_k &= \|r_k\|_2, \quad a_k = \frac{\widetilde{r}_k^* r_k}{b_k}, \quad k \geqslant 1, \\ q_k &= \frac{r_{k-1}}{b_{k-1}}, \quad \widetilde{q}_k = \frac{\widetilde{r}_{k-1}}{a_{k-1}}, \quad k \geqslant 2. \end{aligned}$$

For Example 10, we simply choose

$$q = (1, 0, 0)^*, \quad \widetilde{q} = (1, 0, 0)^*.$$

Generally speaking, there is a question in choosing initial q_0 and $\widetilde{q}_0$. More generally, it should be meaningful to know for what A, the resulting matrix have positive a_k and b_k for every k.

4 Efficient initials. General case

A general algorithm for the efficient initials will be introduced later in the second subsection. The algorithm introduced in the next subsection is easier and quite general, but may be less efficient.

4.1 Fix uniformly distributed initial vector v_0

In this subsection, we fix the uniformly distributed initial vector

$$v_0 = \frac{(1, 1, \ldots, 1)}{\sqrt{N+1}}.$$

This is the easiest choice of v_0 since it does not use any information from the eigenvector g of $\rho(A)$ except its positivity property. On the other hand, this means that the choice is less efficient and it can be even broken as shown by Example 3. The effectiveness of this v_0 depends heavily on the choice of z_0. For which, here we introduce three effective choices.

Choice I Let $A = (a_{ij}: i, j \in E)$ be nonnegative and set $z_0 = \sup_{i \in E} A_i$, where $A_i = \sum_{j \in E} a_{ij}$. This universal choice comes from the fact that $\sup_{i \in E} A_i$ is an upper bound of $\rho(A)$, which can be seen by setting $x_i \equiv 1$ in the next result.

Proposition 11 *For a nonnegative irreducible matrix A with maximal eigenvalue $\rho(A)$, the Collatz–Wielandt formula holds*:

$$\sup_{x>0} \min_{i \in E} \frac{(Ax)_i}{x_i} = \rho(A) = \inf_{x>0} \max_{i \in E} \frac{(Ax)_i}{x_i}.$$

For the present (v_0, z_0), even though it is not necessary, one may replace (2) by

$$(z_{k-1}I - A)w_k = v_{k-1}. \tag{8}$$

This choice of z_0 avoids the collapse of the algorithm since

$$0 < z_0 - \rho(A) < |z_0 - \lambda|$$

for every eigenvalue $\lambda \neq \rho(A)$ of A.

Let us now introduce an important application of Proposition 11. First, if we replace A and $\rho(A)$ with $-Q$ and λ_0, respectively, the same conclusion holds, as shown in the next corollary (the proof is delayed to Section 5). Actually, the corollary holds in a much more general setup. Refer to [3, Theorem 9.5].

Corollary 12 *For Q-matrix, the Collatz–Wielandt formula becomes*

$$\sup_{x>0} \min_{i\in E} \frac{(-Qx)_i}{x_i} = \lambda_0(Q) = \inf_{x>0} \max_{i\in E} \frac{(-Qx)_i}{x_i}.$$

Thus, instead of the mean estimate given in these algorithm, we can produce pointwise estimates. To do so, we need only to compute the ratio $(-Q)v_k/v_k$. For instance, in Example 2, the ratio $(-Q)v_2/v_2$ is as follows:

$$0.525197,\ 0.5254,\ 0.52553,\ 0.525623,\ 0.525693,\ 0.525747,\ 0.525787,\ 0.525816.$$

Therefore, we obtain

$$0.525197 \leqslant \lambda_0 \leqslant 0.525816$$

and the ratio of the upper/lower bounds is ≈ 1.00118. Next, for Example 7, the ratio $(-Q)v_2/v_2$ is as follows:

$$0.525268,\ 0.525268,\ 0.525267, 0.525267, 0.525267, 0.525267, 0.525267, 0.525267.$$

Hence, we have

$$0.525267 \leqslant \lambda_0 \leqslant 0.525268$$

and the ratio of the upper/lower bounds is $\approx 1 + 10^{-6}$. Actually, if we apply the estimates given in [5, Theorem 2.4 (3)] (with $\operatorname{supp}(f) = E$)

$$z_2 \wedge \sup_{i\in E} \frac{f_i}{g_i} \geqslant \lambda_0 \geqslant \inf_{i\in E} \frac{f_i}{g_i},$$

$$g_i := \sum_{k\in E} \mu_k f_k \varphi_{i\vee k} = \varphi_i \sum_{k=0}^{i} \mu_k f_k + \sum_{k=i+1}^{N} \mu_k \varphi_k f_k,$$

$$\varphi_i := \sum_{k=i}^{N} \frac{1}{\mu_k b_k} \quad \text{(for this example, } \mu_k \equiv 1,\ b_i = (i+1)^2\text{)},$$

to the test function $f = v_2$ with a more precise output, the upper/lower bounds can be improved as $\approx 1 + 10^{-7}$. Hence, the estimate $\lambda_0 \approx 0.525268$ is indeed

sharp up to the six precisely significant digits. This shows that the estimates in the latter example are much better than the former one.

Example 13 Let A be the same as in Example 10. Then $\rho(A) \approx 5.23607$ and $z_0 = 6$. The Rayleigh quotient iteration gives us

$$z_1 \approx 5.27273, \quad z_2 \approx 5.23639, \quad z_3 \approx 5.23607.$$

Example 14 Let

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix}.$$

Then $\rho(A) \approx 36.2094$ and $z_0 = 58$. The Rayleigh quotient iteration gives us

$$z_1 \approx 37.3442, \quad z_2 \approx 36.2674, \quad z_3 \approx 36.2095, \quad z_4 \approx 36.2094.$$

Example 15 Let

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 14 & 11 & 0 \\ 9 & 10 & 11 & 1 \\ 5 & 6 & 7 & 8 \end{pmatrix}.$$

This matrix has complex eigenvalues:

$$24.0293,\ 7.72254,\ 1.1241 + 2.40522\,\mathrm{i},\ 1.1241 - 2.40522\,\mathrm{i}.$$

Hence, $\rho(A) \approx 24.0293$ and $z_0 = 31$. The Rayleigh quotient iteration gives us

$$z_1 \approx 24.4393, \quad z_2 \approx 24.0385, \quad z_3 \approx 24.0293.$$

Example 16 Let Q be the same as in Example 1 and let

$$A = 113\, I + Q.$$

Then $z_0 = 113$. Recall that $\lambda_{\min}(-Q) \approx 0.525268$. For $k = 1, 2, 3$, the Rayleigh quotient iteration gives us $113 - z_k$ as follows:

$$113 - z_1 \approx 0.602312, \quad 113 - z_2 \approx 0.525463, \quad 113 - z_3 \approx 0.525268.$$

Alternatively, one may apply the algorithm directly to $-Q$ with $z_0 = 0$.

We remark that the algorithm is meaningful for any

$$z_0 \geqslant \sup_{i \in E} \sum_{j \in E} a_{ij}.$$

For instance, if we choose $z_0 = 200$ rather than $z_0 = 6$ used in Example 13, then the successive results of the iterations are as follows:

$$z_1 \approx 5.33546, \quad z_2 \approx 5.24182, \quad z_3 \approx 5.23608, \quad z_4 \approx 5.23607.$$

The convergence becomes slower as we can imagine. In other words, a larger initial z_0 is less efficient. In view of Proposition 11, we have

$$0 < \rho(A) \leqslant 113.$$

It seems that there is a large room for us to choose z_0. Yes or no? It is yes, since the last estimates are rather rough, each choice $z_0 \in [111.7, 113]$ is also available. The answer is also no, since if we choose $z_0 = 111.6$, then we will go to the pitfall λ_1 $(> \lambda_0)$. Hence, it is rather sensitive to find a useful z_0 except Choice I. Noting that

$$\rho(A) \approx 113 - 0.525268,$$

the reason why the rough Choice I is still efficient for this model should be clear.

We have thus studied the model introduced in Example 1 six times with different initials. The results are collected in Table 4. Among them, the worst one is Example 3 and the best one is Example 7 which uses the whole power of the algorithm introduced in Section 3. The "Uniform" is the present Choice I and the "Auto" means automatic one given by the algorithm, as we will come back in Choice II below.

Table 4 Comparison of examples with different initials

same Q	v_0	z_0	# of iterations
Example 1	$\widetilde{v}_0$	power	10^3
Example 2	$\widetilde{v}_0$	auto	2
Example 3	uniform	auto	collapse
Example 4	uniform	δ_1^{-1}	2
Example 7	$\widetilde{v}_0$	δ_1^{-1}	2
Example 16	uniform	113	3

In conclusion, even though the present choice (v_0, z_0) may not be very efficient, but it works in a very general setup. This algorithm works even for a more general class of matrices, without assuming the nonnegative property, once you have an upper estimate of the largest eigenvalue of A. Clearly, for large-scale matrix, Choice I is meaningful only for the sparse ones.

Choice II Simply use the particular choice given in the Rayleigh quotient iteration: $z_0 = v_0^* A v_0$. This simple choice is quite natural and so is often used in practice. However, there is a dangerous here since v_0 is chosen roughly, the algorithm may lead to an incorrect limit, as illustrated by Example 3.

With the present z_0, the computation results for Examples 13–15 are listed in Table 5.

Table 5 Output (z_1, z_2, z_3) of Examples 13–15

Example	z_1	z_2	$z_3 = \lambda_0$
13	5.24183	5.23608	5.23607
14	35.8428	36.2127	36.2094
15	23.7316	24.0317	24.0293

Combining (z_1, z_2) here with those given in the last part, it is clear that the present choice of z_0, once works, is better than Choice I.

Choice III This is based on a comparison technique. For given $A = (a_{ij})$ having the property $a_{i,i+1} + a_{i+1,i} > 0$ for every i, we introduce the symmetrized matrix $(A + A^*)/2$. (This symmetrizing procedure may be omitted if both $a_{i,i+1} > 0$ and $a_{i+1,i} > 0$ for every i.) Denote by $(\alpha_i, \beta_i, \gamma_i)$ the tridiagonal part (where γ_i are the diagonal elements) taken from the symmetrized matrix. By assumption, we have $\alpha_i > 0$ and $\beta_i > 0$. We can then follow the last section to choose a z_0 first for the tridiagonal matrix and then regarding it as an approximation of z_0 for the original A. One may worry that we have lost too much in the last step. Yes, it may be so. However, the key is to avoid the collapse. The smaller estimate z_0 of $\lambda_{\min}(-Q)$ is not really serious since the algorithm can repair it rapidly, as shown by the next example.

Example 17 Let A be the same as in Example 15. Then

$$\frac{1}{2}(A + A^*) = \begin{pmatrix} 1 & 5/2 & 9/2 & 5/2 \\ 5/2 & 14 & 21/2 & 3 \\ 9/2 & 21/2 & 11 & 4 \\ 5/2 & 3 & 4 & 8 \end{pmatrix}.$$

From this, we obtain a tridiagonal matrix

$$T = \begin{pmatrix} 1 & 5/2 & 0 & 0 \\ 5/2 & 14 & 21/2 & 0 \\ 0 & 21/2 & 11 & 4 \\ 0 & 0 & 4 & 8 \end{pmatrix},$$

and then

$$Q = T - 27I = \begin{pmatrix} -26 & 5/2 & 0 & 0 \\ 5/2 & -13 & 21/2 & 0 \\ 0 & 21/2 & -16 & 4 \\ 0 & 0 & 4 & -19 \end{pmatrix}.$$

According to what we did in Section 3, we have $z_0 \approx 1/0.321526$ for $-Q$. Then, we have

$$z_0 \approx 27 - \frac{1}{0.321526}$$

for T. This is regarded as an approximation of z_0 for A. Starting from here and using the Rayleigh quotient iteration, we obtain the successive approximation of $\rho(A)$ as follows:

$$z_1 \approx 24.0125, \quad z_2 \approx 24.0293,$$

as we expected. Picking up the tridiagonal part directly from A (without using the symmetrizing procedure), the same approach leads to the following output:

$$z_0 \approx 28 - \frac{1}{0.23307} \approx 23.7094, \quad z_1 \approx 23.9901, \quad z_2 \approx 24.0293.$$

Let us remark that the three choices of z_0 in this subsection are independent of the initial v_0 used here and so can be also used in the next subsection. Certainly, there are other approaches can be used to deduce an approximation of the required z_0. For instance, Cheeger's approach [3, §9.5], which is meaningful in a very general setup. Since it takes account of all subset of E (except the emptyset), the number of computations is of order 2^N. This approach as well as the capacitary one (cf. [4, Chapter 7]) needs to be simplified to fit the present setup. In practice, one often uses Proposition 11 or Corollary 12 to get an upper/lower bound in terms of a suitable test sequence (x_i). Refer also to [4, Theorem 3.6] which uses test weights. These approaches depend heavily on the working models.

4.2 Efficient initial vector v_0

In general, it is much more difficult to choose an efficient initial v_0 than z_0. Here is our algorithm.

A general algorithm

Let $A = (a_{ij}\colon i, j \in E)$ be a given irreducible matrix having nonnegative off-diagonal elements. Once again, denote by $\rho(A)$ the maximal eigenvalue of A. If $A_i := \sum_{j\in E} a_{ij}$ is a constant (independent of $i \in E$), then we have $\rho(A) \equiv A_i$ with right-eigenvector $\mathbb{1}$ (its components are all equal to 1). From now on, we assume that A_i are not a constant.

We introduce our algorithm in four steps.

Step 1 When $A_i \leqslant 0$ for every $i \in E$, one can jump from here to Step 2 below by setting $Q = A$. Otherwise, let $\max_{i\in E} A_i > 0$. Define

$$Q = A - \big(\max_{i\in E} A_i\big) I.$$

Then the sum of each row of Q is less or equal to zero and at least one of the rows is less than zero since A_i is not a constant. Now, if

$$Q_0 = Q_1 = \cdots = Q_{N-1} = 0$$

but $Q_N < 0$ ($Q_k := \sum_j q_{kj}$), then one can jump from here to Step 3 with $h_i \equiv 1$.

Step 2 Assume that $Q_k < 0$ for some $k \leqslant N-1$. Denote by

$$h = (h_0, h_1, \ldots, h_N)^*$$

with $h_0 = 1$ the solution to the equation

$$Q^{\setminus N\text{'s row}} h = 0,$$

where $Q^{\setminus k\text{'s row}}$ is obtained from Q removing its k's row $(q_{k0}, q_{k1}, \ldots, q_{kN})$. In the case that

$$c_N + \sum_{j\leqslant N-1} q_{Nj}\Big(1 - \frac{h_j}{h_N}\Big)$$

is much smaller than

$$\sum_{j\leqslant N-1} q_{Nj}\frac{h_j}{h_N}$$

(say, 1 : 100 for instance), one can jump from here to (10) with $x_i \equiv 1$ (cf. Example 21 in the case of $b_4 = 0.01$).

Step 3 Let $(h_i\colon i \in E)$ be constructed in the last step. Define $q_i = -q_{ii}$, $i \in E$. Let $x = (x_0, x_1, \ldots, x_N)^*$ (with $x_0 = 1$) be the solution to the equation

$$x^{\backslash 0\text{'s row}} = P^{\backslash 0\text{'s row}}\, x, \tag{9}$$

where

$$P = (p_{ij}\colon i, j \in E)\colon \quad p_{ii} = 0, \quad p_{ij} = \frac{q_{ij}h_j}{q_i h_i}, \quad j \neq i;$$

or in the matrix form,

$$P = \mathrm{Diag}((q_i h_i)^{-1})Q\mathrm{Diag}(h_i) + I.$$

Refer to the comments below Examples 21 and 22 for the constraint $x_0 = 1$. Here, the sequence (x_i) is an extension of (φ_i) used in Section 3 (cf. Lemma 24 below).

Step 4 We are now ready to state our algorithm as follows. Define a (column) vector $\widetilde{v}_0$ with components

$$\widetilde{v}_0(i) = h_i\sqrt{x_i}, \quad i = 0, 1, \ldots, N. \tag{10}$$

Let

$$v_0 = \frac{\widetilde{v}_0}{\sqrt{\widetilde{v}_0^*\widetilde{v}_0}}, \quad z_0 = v_0^*(-Q)v_0.$$

In general, for $k \geqslant 1$, let w_k be the solution to the equation

$$(-Q - z_{k-1}I)w_k = v_{k-1},$$

and define

$$v_k = \frac{w_k}{\sqrt{w_k^* w_k}}, \quad z_k = v_k^*(-Q)v_k.$$

Then z_k and v_k are approximations of the minimal eigenvalue $\lambda_0 = \lambda_{\min}(-Q)$ of $-Q$ and its eigenvector, respectively. If we replace $-Q$ by A everywhere in this step, then the resulting z_k and v_k are approximations of $\rho(A)$ and its eigenvector g, respectively. Obviously, from Step 1, it follows that

$$\lambda_{\min}(-Q) + \rho(A) = \max_{i\in E} A_i.$$

Hence,

$$\lambda_0 = \lambda_{\min}(-Q) > \alpha \Longleftrightarrow \rho(A) \leqslant \max_{i\in E} A_i - \alpha.$$

This gives the relationship of a lower estimate of λ_0 and an upper estimate of $\rho(A)$.

Example 18 *Let A be given in Example* 10. *Then*

$$\rho(A) = 3 + \sqrt{5} \approx 5.23607.$$

Our algorithm here gives us

$$z_1 \approx 5.23883, \quad z_2 \approx 5.23607.$$

Proof Since $\max_i A_i = 6$, we have

$$Q = A - 6\,I = \begin{pmatrix} -5 & 2 & 3 \\ 1 & -4 & 1 \\ 3 & 2 & -5 \end{pmatrix}.$$

Next, we have

$$h_0 = 1, \quad h_1 = \frac{4}{7}, \quad h_2 = \frac{9}{7},$$

and

$$x_0 = 1, \quad x_1 = \frac{7}{9}, \quad x_2 = \frac{49}{81}.$$

From these, we obtain

$$\widetilde{v}_0 = (1,\, h_1\sqrt{x_1},\, h_2\sqrt{x_2}\,)^* = \Big(1,\, \frac{4}{3\sqrt{7}},\, 1\Big)^*.$$

Now, with

$$v_0 = \frac{\widetilde{v}_0}{\sqrt{\widetilde{v}_0^*\widetilde{v}_0}}, \quad z_0 = v_0^* A v_0 \approx 5.11616,$$

we can apply the Rayleigh quotient iteration in two steps to obtain the conclusion. □

Example 19 *Let A be the same as in Example* 14. *Then $\rho(A) \approx 36.2094$. By using* (10),

$$v_0 = (0.348213,\, 0.244601,\, 0.389728,\, 0.816719)^*,$$

the Rayleigh quotient iteration starts at $z_0 \approx 34.4924$ and gives us

$$z_1 \approx 36.1469, \quad z_2 \approx 36.2095, \quad z_3 \approx 36.2094.$$

Proof We have

$$Q = A - 58I = \begin{pmatrix} -57 & 2 & 3 & 4 \\ 5 & -52 & 7 & 8 \\ 9 & 10 & -47 & 12 \\ 13 & 14 & 15 & -42 \end{pmatrix}.$$

Next, we have

$$h_0 = 1, \quad h_1 = \frac{59}{27}, \quad h_2 = \frac{91}{27}, \quad h_3 = \frac{287}{27}.$$

Furthermore, we have

$$x_0 = 1, \quad x_1 = \frac{189}{1829}, \quad x_2 = \frac{7155}{64883}, \quad x_3 = \frac{243}{4991}.$$

Then the conclusion follows from the iteration. □

Example 20 *Let A be the same as in Example* 15. *Then $\rho(A) \approx 24.0293$. By using the algorithm in Section* 4.2, *the Rayleigh quotient iteration starts at $31 - z_0 \approx 22.6424$ and gives us for $k = 1, 2, 3$,*

$$31 - z_k \approx 24.1046, \quad 24.0298, \quad 24.0293,$$

respectively.

Proof We have

$$Q = A - 31I = \begin{pmatrix} -30 & 2 & 0 & 0 \\ 3 & -17 & 11 & 0 \\ 9 & 10 & -20 & 1 \\ 5 & 6 & 7 & -23 \end{pmatrix}.$$

Then, we have

$$h_0 = 1, \quad h_1 = 15, \quad h_2 = \frac{252}{11}, \quad h_3 = \frac{3291}{11};$$
$$x_0 = 1, \quad x_1 = \frac{3691}{76575}, \quad x_2 = \frac{1694}{45945}, \quad x_3 = \frac{7447}{3360111};$$
$$v_0 = (0.140655,\ 0.463208,\ 0.61873,\ 0.61873).$$

The conclusion follows by the algorithm. □

It is interesting to compare this example with Examples 15 and 17.

Actually, to show that our algorithm is reasonable, one may ignore the part using the H-transform and jump to the last step on Q-matrix since the transform does not change the spectrum. Thus, one needs to compare the maximal eigenvector g and its approximation (x_i). As mentioned before, this depends heavily on the rate $b_N = c_N$. Here is an example of sparse matrix.

Example 21 Let

$$Q = \begin{pmatrix} -3 & 2 & 0 & 1 & 0 \\ 4 & -7 & 3 & 0 & 0 \\ 0 & 5 & -5 & 0 & 0 \\ 10 & 0 & 0 & -16 & 6 \\ 0 & 0 & 0 & 11 & -11 - b_4 \end{pmatrix}.$$

Corresponding to different b_4, the maximal eigenvector g (normalized so that the first component to be one) and its approximation $(\sqrt{x_i})$ (up to a positive constant) are given in Table 6.

Table 6 For different b_4, vectors g and $(\sqrt{x_i})$ (Example 21)

b_4	g	$\sqrt{x}$ up to a constant
0.01	$(1, 1.00011, 1.00017, 0.999498, 0.998616)^*$	$(1, 1, 1, 0.999728, 0.999274)^*$
1	$(1, 1.00992, 1.0149, 0.955637, 0.877794)^*$	$(1, 1, 1, 0.9759, 0.934353)^*$
100	$(1, 1.08011, 1.1211, 0.656961, 0.0652116)^*$	$(1, 1, 1, 0.805682, 0.253629)^*$

The corresponding output of our algorithm is given in Table 7.

Table 7 For different b_4, eigenvalue λ_0 and z_1, z_2, z_3 (Example 21)

b_4	z_1	z_2	$z_3 = \lambda_0$
0.01	0.000278573	0.000278686	
1	0.0236258	0.0245174	0.0245175
100	0.200058	0.182609	0.182819

Our original purpose to design the Q-matrix in the last example is for a test of sparse matrix. The solution $x_0 = x_1 = x_2 = 1$ leads us to think about the transition machinery of the Q-matrix. Here is the graphic structure of the Q-matrix:

$$\textcircled{2} \leftrightarrows \textcircled{1} \leftrightarrows \textcircled{0} \leftrightarrows \textcircled{3} \leftrightarrows \textcircled{4}.$$

As we will see at the end of Section 5, x_i is the probability of the process first hitting 0 starting from i (which is exactly the probabilistic meaning of the construction of (x_i) given in our general algorithm). Now, starting from 2, there is only one way to go to 0, and hence x_2 should be equal to 1. So does x_1. From this graph, it follows that the matrix is indeed tridiagonal after a relabeling (simply exchange the labels ② and ⓪):

$$\textcircled{0} \leftrightarrows \textcircled{1} \leftrightarrows \textcircled{2} \leftrightarrows \textcircled{3} \leftrightarrows \textcircled{4}.$$

As a comparison, we present the next result using the algorithms given in Sections 4 and 3, respectively.

Example 22 Let

$$Q = \begin{pmatrix} -5 & 5 & 0 & 0 & 0 \\ 3 & -7 & 4 & 0 & 0 \\ 0 & 2 & -3 & 1 & 0 \\ 0 & 0 & 10 & -16 & 6 \\ 0 & 0 & 0 & 11 & -11 - b_4 \end{pmatrix}.$$

Corresponding to different b_4, the maximal eigenvector g and its approximation $(\sqrt{x_i})$ are given in Table 8.

Table 8 For different b_4, vectors g and $(\sqrt{x_i}\,)$ (Example 22)

b_4	g	$\sqrt{x}$ up to a constant
0.01	$(1, 0.999944, 0.999833, 0.999331, 0.998449)^*$	$(1, 0.999819, 0.999682, 0.99941, 0.998956)^*$
1	$(1, 0.995096, 0.98532, 0.941608, 0.864908)^*$	$(1, 0.984848, 0.973329, 0.949871, 0.909433)^*$
100	$(1, 0.963436, 0.89198, 0.585996, 0.0581675)^*$	$(1, 0.91325, 0.842344, 0.678661, 0.213643)^*$

The corresponding output (z_k) of the algorithm in Section 4 is given in Table 9.

Table 9 For different b_4, eigenvalue λ_0 and z_1, z_2, z_3 (Example 22)

b_4	z_1	z_2	$z_3 = \lambda_0$
0.01	0.000278548	0.000278686	
1	0.0234222	0.0245174	0.0245175
100	0.13342	0.182541	0.182819

The output (z_k) of the algorithm in Section 3 is given in Table 10.

Table 10 For different b_4, eigenvalue λ_0, its lower bound δ_1^{-1} and z_1, z_2 (Example 22)

b_4	$z_0 = \delta_1^{-1}$	z_1	$z_2 = \lambda_0$
0.01	0.00027867	0.000278686	
1	0.0244003	0.024519	0.0245175
100	0.179806	0.182912	0.182819
10^6	0.191917	0.195239	0.195145

Once again, one sees the efficiency of our algorithm.

Comparing the last two examples, especially their g and $\sqrt{x_i}$, it is obvious that the latter is better than the former one. This suggests us to choose the starting point 0 carefully. Here is an easier way to do so. First, define a sequence $\{E_\ell\}$ of level sets as follows. Let $E_0 = \{N\}$ and $E_1 = \{i \in E\colon a_{iN} > 0\}$. At the kth step, set

$$E_k = \{i \in E \setminus (E_0 + E_1 + \cdots + E_{k-1})\colon \exists\, j \in E_{k-1} \text{ such that } a_{ij} > 0\}.$$

The procedure should be stopped at m if $E_{m+1} = \emptyset$. Because of the irreducibility, each $i \in E$ should belong to one of the level sets. Finally, regard one of $i_m \in E_m$ satisfying

$$a_{i_m j_{m-1}} = \min\{a_{ij}\colon i \in E_m,\, j \in E_{m-1}\}$$

as our initial 0. However, for initial $\widetilde{v}_0$, in practice, it is not necessary to relabeling the states as we did in Example 22. What we need is only replace the constraint $x_0 = 1$ by $x_{i_m} = 1$ (at the same time, "removing the first line" is replaced by "removing the i_m's line" in constructing the required matrix) in solving (x_i) without change the original matrix A or Q. One may need the relabeling in computing δ_1 defined in Section 3.

To conclude this subsection, we introduce a new construction of z_0 based on v_0 defined by our general algorithm. It is an extension of $z_0 = \delta_1^{-1}$ given in Section 3. To do so, we use Q, (h_i), and (x_i) defined at the beginning of this subsection. Let $\widetilde{Q}_0$ be the matrix obtained from

$$\widetilde{Q} := \mathrm{Diag}(h_i)^{-1} Q \mathrm{Diag}(h_i)$$

by modifying the last diagonal element $\widetilde{q}_{N,N}$ so that the sum of its last row becomes zero (i.e., removing the killing c_N). Next, let $\mu := (\mu_0, \mu_1, \ldots, \mu_N)$ with $\mu_0 = 1$ be the solution to the equation

$$\mu \widetilde{Q}_0 = 0.$$

Since there are only N variables $\mu_1, \mu_2, \ldots, \mu_N$, one may get the solution μ from the equation

$$\widetilde{Q}^{*\,\setminus \text{the last row}} \mu^* = 0.$$

Here, we remark that for a large class of Q-matrix Q, there is an explicit representation of μ in terms of the non-diagonal elements of Q, refer to [3, Chapter 7]. Now, our new initial z_0 is defined to be δ_1^{-1}:

$$\delta_1 = \frac{1}{1 - x_1} \max_{0 \leqslant n \leqslant N} \left[\sqrt{x_n} \sum_{k=0}^{n} \mu_k \sqrt{x_k} + \frac{1}{\sqrt{x_n}} \sum_{n+1 \leqslant j \leqslant N} \mu_j x_j^{3/2} \right]. \tag{11}$$

In contrast to the above examples which use only the automatic $z_0 = v_0^* A v_0$ (or $z_0 = v_0^*(-Q)v_0$), here we use (11). Remember that this initial z_0 is for $-Q$, when we go back to the original A, its initial becomes $\max_{i \in E} \sum_{j \in E} a_{ij} - z_0$. The outputs of Examples 18–20 using δ_1^{-1} are listed in Table 11.

Table 11 Outputs of Examples 18–20 using δ_1^{-1}

Example	z_0	z_1	z_2	$z_3 = \lambda_0$
18	5.90016	5.22268	5.23611	5.23607
19	57.2719	36.236	36.2097	36.2094
20	30.3886	23.7436	24.0347	24.0293

Finally, we have an improved algorithm (for Q) as stated in Section 3 (below Example 10) based on the use of $L^2(\mu)$ and the convex combination:

$$z_0 = \xi \delta_1^{-1} + (1 - \xi)(v_0, -Q v_0)_\mu, \quad \xi \in [0, 1].$$

The outputs of Examples 18–20 using the new z_0 with $\xi = 1/3$ are listed in Table 12.

Table 12 Outputs of Examples 18–20 using new z_0 with $\xi = 1/3$

Example	z_0	z_1	z_2	$z_3 = \lambda_0$
18	5.04169	5.24358	5.23608	5.23607
19	35.4952	36.2657	36.2095	36.2094
20	24.0583	24.0213	24.0293	

This combination becomes more serious when N is large since in that case $(v_0, -Qv_0)_\mu$ is often an upper bound of λ_0, which may be much closer to other $\lambda_j \neq \lambda_0$ and so the algorithm would converge to λ_j but not λ_0. Certainly, the convex combination idea is also meaningful for the first two choices of z_0 introduced in the first subsection.

5 Additional remarks and proofs

In this section, we first prove a new result related to our earlier study. Then we present some proofs of the results given in the last two sections. Finally, we will make some remarks on the results studied so far in the previous sections.

The next result solves an open question kept in our mind for many years. For a given birth–death matrix Q on E with $c_0 = c_1 = \cdots = c_{N-1} = 0$ and $b_N := c_N > 0$, and a positive function f on E, define

$$\Pi(f)(i) = \frac{1}{f_i} \sum_{j=i}^{N} \frac{1}{\mu_j b_j} \sum_{k=0}^{j} \mu_k f_k, \quad i \in E.$$

Proposition 23 *For Q and Π given above, let f_1 (> 0 on E) be arbitrarily given function and define successively $f_{n+1} = f_n \Pi(f_n)$. Then this algorithm coincides with the inverse iteration given in Lemma 6 with $z = 0$, even for infinite N. Furthermore, we have*

$$\lambda_0 = \lambda_{\min}(-Q) = \lim_{n\to\infty} \Pi(f_n)(i)^{-1}$$

for each $i \in E$. In particular, we have

$$\lim_{n\to\infty} \min_{i\in E} \Pi(f_n)(i) = \frac{1}{\lambda_0} = \lim_{n\to\infty} \max_{i\in E} \Pi(f_n)(i).$$

Proof Consider the Poisson equation: $-Qf = g$ for a given g. The solution is given by $f = g\Pi(g)$ ([5, (2.7)–(2.9)]). It can be also written as $f = (-Q)^{-1}g$. By setting $g = f_1$ and $f = f_2$, it follows that

$$f_2 = (-Q)^{-1} f_1 = f_1 \Pi(f_1).$$

Now, by iteration, we get

$$f_{n+1} = (-Q)^{-n} f_1 = f_n \Pi(f_n), \quad n \geqslant 1.$$

We have thus proved the first assertion. Therefore,

$$\Pi(f_n) = \frac{f_{n+1}}{f_n} = \frac{(-Q)^{-n}(f_1)}{(-Q)^{-n+1}(f_1)} \to \frac{1}{\lambda_0}, \quad n \to \infty,$$

by the last assertion of Lemma 6 with $z = 0$. The last assertion of the proposition then follows since on a finite set, the pointwise convergence implies the uniform one. □

We remark that the last proposition is meaningful once the Poisson equation $-Qf = g$ is solvable. In parallel, Lemma 6 improves the approximating procedures studied in [5] and related publications.

Now, we turn to prove Proposition 9 and Corollary 12.

Proof of Proposition 9 (a) First, we follow the setup and notation in [9] (where a more general situation is studied) for a moment. Define

$$M_{N-1}(h) = \sum_{k=0}^{N-1} \widetilde{q}_N^{(k)} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)} h_j}{q_{j,j+1}},$$

$$N_n(h) = \sum_{k=0}^{n} \sum_{j=0}^{k} \frac{\widetilde{F}_k^{(j)} h_j}{q_{j,j+1}}, \quad 0 \leqslant n < N.$$

Then the solution given in [9, Proposition 2.6] can be rewritten as

$$g_n = \frac{f_N + M_{N-1}(f)}{c_N + M_{N-1}(c_\cdot)} [1 - N_{n-1}(c_\cdot)] + N_{n-1}(f), \quad N_{-1} := 0, \quad 0 \leqslant n \leqslant N.$$

By an exchange of the order of the summations, we can rewrite M_n and N_n as follows:

$$M_{N-1}(h) = \sum_{j=0}^{N-1} \frac{h_j}{q_{j,j+1}} \sum_{k=j}^{N-1} \widetilde{q}_N^{(k)} \widetilde{F}_k^{(j)},$$

$$N_n(h) = \sum_{j=0}^{n} \frac{h_j}{q_{j,j+1}} \sum_{k=j}^{n} \widetilde{F}_k^{(j)}, \quad 0 \leqslant n < N.$$

Here, for finite N, the element $q_{N,N+1}$ is replaced by c_N by our convention. Thus, by [9, (1.1)], we get

$$M_{N-1}(h) = c_N \sum_{j=0}^{N-1} \frac{h_j}{q_{j,j+1}} \widetilde{F}_N^{(j)}.$$

By [6, Proposition 4.1], we have $\widetilde{F}_{i+m}^{(i)} = G_{m,m}^{(i)}$. It follows that

$$M_{N-1}(h) = c_N \sum_{j=0}^{N-1} \frac{h_j}{q_{j,\,j+1}} G_{N-j,N-j}^{(j)},$$

$$N_n(h) = \sum_{j=0}^{n} \frac{h_j}{q_{j,j+1}} \sum_{k=0}^{n-j} G_{k,k}^{(j)}, \quad 0 \leqslant n < N.$$

Applying this solution to the birth–death context and setting $f=-v$, $g=w$, replacing the original $c_\cdot$ used in [9] by $z-c_\cdot$, we obtain

$$g_n=\frac{-v_N-M_{N-1}(v)}{z-c_N+M_{N-1}(z-c_\cdot)}\,[1-N_{n-1}(z-c_\cdot)]-N_{n-1}(v),\quad 0\leqslant n\leqslant N.$$

Equivalently,

$$g_n=\frac{v_N+M_{N-1}(v)}{c_N-z+M_{N-1}(c_\cdot-z)}\,[1+N_{n-1}(c_\cdot-z)]-N_{n-1}(v),\quad 0\leqslant n\leqslant N.$$

This gives us the required conclusion. □

Proof of Corollary 12 The proof is quite straightforward. Choose m large enough such that

$$A:=mI+Q$$

is a nonnegative matrix. Then $-Q=mI-A$. Hence,

$$\lambda_0(Q)=m-\rho(A).$$

The proof now is a direct application of the Collatz–Wielandt formula:

$$m-\rho(A)=m-\inf_{x>0}\max_i\frac{(Ax)_i}{x_i}=\sup_{x>0}\min_i\frac{(-Qx)_i}{x_i},$$
$$m-\rho(A)=m-\sup_{x>0}\min_i\frac{(Ax)_i}{x_i}=\inf_{x>0}\max_i\frac{(-Qx)_i}{x_i}.$$ □

It is now ready to make some additional remarks on the results in the previous sections. The two algorithms as well as their convergence and the Collatz–Wielandt formula can be found easily from Wikipedia. From which, one knows that the Power Iteration was first appeared in 1929 [14] and the Inverse Iteration appeared in 1944 [15]. These algorithms are taught for undergraduate students on the course of computations and are included in many books, see for instance [10,13,16]. In particular, Appendix of Section 3 is modified from [10, pp. 584, 585].

We now say a few words about the unusual word "complete" used at the end of the first section for the results obtained in Section 3. Actually, this is one of the 16 situations with $N\leqslant\infty$ we have worked out so far to have a unified estimation of the principal eigenvalue:

$$(4\delta)^{-1}\leqslant\delta_1^{-1}\leqslant\lambda_0\leqslant{\delta_1'}^{-1}\leqslant\delta^{-1}\tag{12}$$

for some constants δ,δ_1, and δ_1', where δ_1 is the one we have used in Section 3 for the initial z_0. Besides, we often have in practice that $1\leqslant\delta_1/\delta_1'\leqslant 2$. Thus, the efficiency of the initial (v_0,z_0) introduced in Section 3 comes with no surprising. More precisely, the initial (v_0,z_0) is taken from the first step of our approximating procedure: [5, Theorem 3.3 (1), (3.4)]. Example 1 here is a

truncated one from [5, Example 3.6] where $N = \infty$, $\lambda_0 = 1/4$, and $\delta_1 = 4$ which is sharp. Certainly, this is still not enough to claim that we can arrive at such a precise approximation in the second iteration. The story on the estimation of the principal eigenvalue, or more general on the estimation of the stability speed is too long to talk here and so the author is planning to publish a survey article [7]. For earlier progress, refer to [4] which includes a lot of information up to 2004, or a more recent paper [5].

Next, we discuss the sequence $(h_0, h_1, \dots, h_N)$ used in Sections 3 and 4. The role of the sequence is to keep the same spectrum of the original Q and its H-transform $\widetilde{Q}$:

$$\widetilde{Q} = \mathrm{Diag}(h_i)^{-1} Q \mathrm{Diag}(h_i). \tag{13}$$

Certainly, Q and $\widetilde{Q}$ have the same diagonals. Next, define

$$P = (p_{ij} \colon i, j \in E) := \mathrm{Diag}(q_i^{-1})\, \widetilde{Q} + I, \tag{14}$$

which is the matrix used in Section 4. Note that even though the sequence (c_i) in the original Q can be non-zero, the resulting $\widetilde{c}_k = 0$ for every $k < N$ but $\widetilde{c}_N > 0$ for the matrix $\widetilde{Q}$. For a given measure μ, set $\widetilde{\mu} = h^2\mu$ (i.e., $\widetilde{\mu}_i = h_i^2 \mu_i$ for each $i \in E$), the transform $\widetilde{f} = f/h$ gives us an isometry between $L^2(\mu)$ and $L^2(\widetilde{\mu})$ and then an isospectrum of Q on $L^2(\mu)$ and $\widetilde{Q}$ on $L^2(\widetilde{\mu})$. This technique is due to [8]. See also [6]. Now, if $\widetilde{g}$ is an approximating eigenvector corresponding to $\widetilde{\lambda}_0$ of $\widetilde{Q}$, then, $g := h\widetilde{g}$ is an approximating eigenvector corresponding to λ_0 of Q, due to the isospectral property of Q and $\widetilde{Q}$. Because

$$\|\widetilde{g}\|_{L^2(\widetilde{\mu})} = \|g\|_{L^2(\mu)}, \quad \big(\widetilde{g}, \widetilde{Q}\widetilde{g}\big)_{\widetilde{\mu}} = (g, Qg)_\mu,$$

by [8], we have

$$\frac{\big(\widetilde{g}, -\widetilde{Q}\widetilde{g}\big)_{\widetilde{\mu}}}{\|\widetilde{g}\|_{L^2(\widetilde{\mu})}} = \frac{(g, -Qg)_\mu}{\|g\|_{L^2(\mu)}} = \frac{g^*(-Q)g}{\sqrt{g^*g}}. \tag{15}$$

Here, we assume that $\mu_k \equiv 1$ for simplicity. This means that we can estimate the maximal eigenpair (λ_0, g) of Q in terms of the one $\big(\widetilde{\lambda}_0, \widetilde{g}\big)$ of $\widetilde{Q}$. More precisely, the maximal eigenvalue $\widetilde{g}$ of $\widetilde{Q}$ is approximated by φ in the context of Section 3 (or by $x = (x_i)$ in Section 4). Now, in Section 3 for instance, $\widetilde{v}_0 = h\sqrt{\varphi}$ is an approximation of the maximal eigenvector g of Q. With $v_0 = \widetilde{v}_0/\sqrt{\widetilde{v}_0^* v_0}$, equation (15) leads to our first approximation of λ_0:

$$v_0^*(-Q)v_0 = z_0.$$

Now, our task is to show that the sequence (x_i) defined in Section 4 is an extension of (φ_i) given in Section 3. To this end, recall that the matrix $\widetilde{Q}$ defined by (13) is again a Q-matrix. Hence, the matrix $P = (p_{ij} \colon i, j \in E)$ defined by (14) is just the embedding chain of $\widetilde{Q}$. Note that here $p_{ii} = 0$ for

each $i \in E$. By the construction of (h_i), we have $\sum_{j\in E} p_{ij} = 1$ for each $i \leqslant N-1$ but $\sum_{j\in E} p_{Nj} < 1$, refer to [8]. The equation for (x_i) in (9) can be rewritten as

$$x_n = \sum_{j\in E} p_{ij}x_j, \quad 1 \leqslant n \leqslant N, \quad x_0 = 1. \tag{16}$$

In probabilistic language, the solution (x_i) (or the minimal solution (x_i^*) when $N = \infty$) to equation (16) is the probability of first hitting 0 of the Q-process with Q-matrix $\widetilde{Q}$ or its embedding sub-Markov chain with transition matrix $P = (p_{ij})$, starting from i. Refer to [3, Lemma 4.46].

We are now going to prove the following result.

Lemma 24 *For birth–death matrix, the solution* (x_i) *to equation* (16) *coincides with* (φ_i) (*up to a constant*) *used in Section* 3.

Before prove Lemma 24, let us discuss the relation of these sequence with the recurrence of the Markov chain in the case of $N = \infty$. First, it is known by [3, Theorem 4.55 (1) and the second line of p. 161] that a birth–death process is recurrent if and only if

$$b_0 \sum_{n=1}^{\infty} \frac{a_1a_2\cdots a_n}{b_1b_2\cdots b_n} = b_0 \sum_{n=1}^{\infty} \frac{1}{\mu_n b_n} = \infty.$$

For simplicity, set

$$F_n^{(0)} = \frac{a_1a_2\cdots a_n}{b_1b_2\cdots b_n}, \quad n \geqslant 1.$$

The sequence $\{F_n^{(0)}\}_{n\geqslant 1}$ is a very special case of $\{\widetilde{F}_n^{(j)}\}_{n\geqslant 1}$ used in the proof of Proposition 9. Refer to [9] and [3, §4.5] for more details. Note that (φ_n) is just the tail series of $\sum_{n=1}^{\infty} F_n^{(0)}$ provided $N = \infty$. On the other hand, by [3, Lemma 4.46], the process is recurrent if and only if the minimal solution (x_i^*) to the equation (16),

$$x_n = \frac{b_n}{a_n + b_n}\, x_{n+1} + \frac{a_n}{a_n + b_n}\, x_{n-1}, \quad n \geqslant 1, \quad x_0 := 1,$$

is equal to one identically. Rewrite the equation as

$$x_n - x_{n+1} = \frac{a_n}{b_n}\,(x_{n-1} - x_n), \quad n \geqslant 1.$$

By induction, it follows that

$$x_n - x_{n+1} = F_n^{(0)}(x_0 - x_1), \quad n \geqslant 1.$$

Hence,

$$x_n - x_{N+1} = (x_0 - x_1)\sum_{k=n}^{N} F_k^{(0)}, \quad x_1 - x_n = (x_0 - x_1)\sum_{k=1}^{n-1} F_k^{(0)}, \quad n \geqslant 1.$$

Equivalently,

$$x_n - x_{N+1} = (x_0 - x_1)\sum_{k=n}^{N} F_k^{(0)}, \quad x_0 - x_n = (x_0 - x_1)\sum_{k=0}^{n-1} F_k^{(0)}, \quad n \geqslant 0,$$

since $F_0^{(0)} = 1$ by convention. If $\sum_{k=0}^{\infty} F_k^{(0)} = \infty$, then from the second equation, we must have $x_1 = 1$ (since $x_0 = 1$) and then have the unique solution $x_i \equiv 1$. Therefore, the minimal solution $x_i^* \equiv 1$ and so the process is recurrent. Conversely, if $\sum_{k=0}^{\infty} F_k^{(0)} < \infty$, then from the first equation above, we obtain

$$x_0 - x_1 = \frac{x_0 - x_\infty}{\sum_{j=0}^{\infty} F_j^{(0)}},$$

and then

$$x_n - x_\infty = \frac{x_0 - x_\infty}{\sum_{j=0}^{\infty} F_j^{(0)}} \sum_{k=n}^{\infty} F_k^{(0)}, \quad n \geqslant 0.$$

Equivalently,

$$x_n = \frac{\sum_{k=n}^{\infty} F_k^{(0)}}{\sum_{j=0}^{\infty} F_j^{(0)}} + x_\infty \frac{\sum_{k=0}^{n-1} F_k^{(0)}}{\sum_{j=0}^{\infty} F_j^{(0)}}, \quad n \geqslant 0.$$

Clearly, for each given $x_\infty \in [0, 1]$, using this formula, we obtain a solution (x_n) to the equation. Thus, the minimal solution should be as follows:

$$x_n^* = \frac{\sum_{k=n}^{\infty} F_k^{(0)}}{\sum_{j=0}^{\infty} F_j^{(0)}}, \quad n \geqslant 0,$$

which is clearly less than one for $n \geqslant 1$ since $\sum_{j=0}^{\infty} F_j^{(0)} < \infty$.

Proof of Lemma 24 For finite state $\{0, 1, \ldots, N\}$, since there is a killing $b_N > 0$, the minimal solution is as follows:

$$x_n^* = \frac{\sum_{k=n}^{N} F_k^{(0)}}{\sum_{j=0}^{N} F_j^{(0)}}, \quad n = 0, 1, \ldots, N.$$

In other words, up to a constant, we have

$$\varphi_n = \sum_{k=n}^{N} F_k^{(0)} = \frac{1}{1 - x_1^*} x_n^*, \quad n = 0, 1, \ldots, N.$$

That is what we required. □

Finally, we remark that the story for one-dimensional diffusions should be in parallel to Section 3. The algorithm presented in Section 4 may not be complete since the lack of an analog of (12).

Summary

This paper deals with the efficient initials for the Rayleigh quotient iteration. Here are suggestions for the use of the results in the previous sections of the paper on computing the maximal eigenpair.

(i) If the iterations are easy (small size of A, for instance), one simply adopts the simplest algorithm: Section 4.1 with Choice I, or more effectively, with the convex combination of Choice I and Choice II:

$$z_0 = \xi \max_{i\in E} A_i + (1-\xi) v_0^* A v_0, \quad \xi \in [0,1].$$

More especially, $\xi = 7/8$ for instance. Certainly, one may use Choice III for z_0.

(ii) If the given matrix is nearly tridiagonal (after a suitable relabeling if necessary) or the Lanczos tridiagonalization procedure is suitable, one use the method introduced in Section 3. The computation there is rather explicit and it works even for $N = \infty$.

(iii) In general, one uses the algorithm given in Section 4.2. Note that at each step of the Rayleigh quotient iteration, one has to solve a linear equation. Here, for the initials, we have to solve two more linear equations.

6 Next to maximal eigenpair

After an earlier version of the paper containing the first five sections was submitted, the author found a natural way to study the next to the maximal eigenpair. In this section, we restrict ourselves to the easier case that $A_i := \sum_{j\in E} a_{ij}$ is a constant. Then the maximal eigenpair is simply $(A_0, \mathbb{1})$ (where $\mathbb{1}$ is the constant function having value 1 everywhere), as mentioned before. By a shift if necessary, we return to the problem for a Q-matrix which is especially valuable since its next eigenvalue describes the ergodic rate of the corresponding Markov chain. In this setup, the minimal eigenpair ($\lambda_0 = 0$, $g_0 = \mathbb{1}$) of $-Q$ is known and we are looking for the next eigenpair (λ_1, g_1). Clearly, g_1 should be orthogonal to g_0 in $L^2(\pi)$-sense for the stationary distribution π of the process corresponding to the given matrix Q. This is the reason why we often use $v - \pi v$ in what follows for constructing a mimic of the eigenvector g_1. Besides, we need the assumption that $\lambda_1 > |\lambda_j|$ for every $j > 1$ to guarantee the convergence of our algorithms.

Once again, let us begin our study with a tridiagonal conservative Q-matrix

$$Q = \begin{pmatrix} -b_0 & b_0 & 0 & 0 & \cdots \\ a_1 & -(a_1+b_1) & b_1 & 0 & \cdots \\ 0 & a_2 & -(a_2+b_2) & b_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & a_N & -a_N \end{pmatrix},$$

where $a_i, b_i > 0$. Define $(\mu_k \colon k \in E)$ as in Section 3. Then we have the

probability distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_N)$: $\pi_k = \mu_k / \sum_{j\in E} \mu_j$. Again, denote by $(\cdot,\cdot)_\mu$ and $\|\cdot\|_\mu$ the inner product and norm in $L^2(\mu)$, respectively. Next, set

$$\varphi_n = \sum_{j\leqslant n-1} \frac{1}{\mu_j b_j}, \quad n \in E.$$

To define our initial v_0, let

$$\widetilde{v}_0 = (\sqrt{\varphi_0}, \sqrt{\varphi_1}, \ldots, \sqrt{\varphi_N})^*, \quad \overline{v}_0 = \widetilde{v}_0 - \pi\widetilde{v}_0.$$

We can now introduce our algorithm in the present situation as follows. Choose initials

$$v_0 = \frac{\overline{v}_0}{\|\overline{v}_0\|_\mu}, \quad z_0 = \frac{(\overline{v}_0, -Q\widetilde{v}_0)_\mu}{\|\overline{v}_0\|_\mu^2}. \tag{17}$$

At the kth step ($k \geqslant 1$), let w_k be the solution to the equation

$$(-Q - z_{k-1})w_k = v_{k-1}$$

and set

$$v_k = \frac{w_k}{\|w_k\|_\mu}, \quad z_k = (v_k, -Qv_k)_\mu.$$

We remark that here in defining v_k ($k \geqslant 1$), we do not need to use $w_k - \pi w_k$. The reason is as follows. If $\pi v = 0$ and w solves the equation

$$(-Q - z)w = v$$

for some constant $z \neq 0$, then

$$0 = \pi v = \pi(-Q - z)w = -z\pi w,$$

and so $\pi w = 0$. Therefore, we have $\pi w_k = 0$ for each $k \geqslant 1$ since so does the initial v_0: $\pi v_0 = 0$.

Instead of z_0 given in (17), there is another choice. Define

$$\eta_1 = \max_{0\leqslant i\leqslant N-1} \frac{1}{\mu_i b_i[\widetilde{v}_0(i+1) - \widetilde{v}_0(i)]} \sum_{j=i+1}^{N} \mu_j \overline{v}_0(j).$$

Then one may choose

$$z_0 = \eta_1^{-1} \tag{18}$$

as an initial.

Here, the initials $\widetilde{v}_0$ and z_0 are taken from [2, Theorem 2.2 (1)] or [4, Theorem 1.5 (2)]. Certainly, we can adopt the convex combination of those given in (17) and (18):

$$z_0 = \xi\eta_1^{-1} + (1-\xi)(\overline{v}_0, -Q\widetilde{v}_0)_\mu \|\overline{v}_0\|_\mu^{-2}, \quad \xi \in [0,1]. \tag{19}$$

We now consider an example modified from Example 1.

Example 25 Let $E=\{0,1,\ldots,7\}$ and

$$Q=\begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -5 & 2^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2^2 & -13 & 3^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3^2 & -25 & 4^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4^2 & -41 & 5^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5^2 & -61 & 6^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6^2 & -85 & 7^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7^2 & -7^2 \end{pmatrix}.$$

Then we have $\mu_k\equiv 1$, $\lambda_1(Q)\approx 0.820539$ with eigenvector

$$\approx(-3.95053,-0.708966,0.246859,0.649164,0.842169,0.93805,0.983254,1)^*.$$

Starting from $\overline{v}_0$:

$$(-4.79299,\ -0.0815238,\ 0.474589,\ 0.70372,\ 0.828504,\ 0.906932,\ 0.960767,\ 1)^*,$$

for different initial z_0, the outputs are given in Table 13.

Table 13 Outputs for different initial z_0 (Example 25)

choice	z_0	z_1	$z_2=\lambda_1$
(17)	0.902633	0.820614	0.820539
(18)	0.456343	0.8216	0.820539
(19)	0.724117	0.820629	0.820539

We remark that for this and the next example, the parameter ξ in (19) is specified to be $2/5$.

The next example has non-trivial (μ_k).

Example 26 Let

$$Q=\begin{pmatrix} -5 & 5 & 0 & 0 & 0 \\ 3 & -7 & 4 & 0 & 0 \\ 0 & 2 & -3 & 1 & 0 \\ 0 & 0 & 10 & -16 & 6 \\ 0 & 0 & 0 & 11 & -11 \end{pmatrix}.$$

Then

$$\mu_0=1,\quad \mu_1=\frac{5}{3},\quad \mu_2=\frac{10}{3},\quad \mu_3=\frac{1}{3},\quad \mu_4=\frac{2}{11}.$$

The eigenvalues of $-Q$ are as follows:

$$22.348,\quad 10.6857,\quad 5.92951,\quad 3.03673,\quad 0.$$

With

$$\widetilde{v}_0=\frac{1}{2\sqrt{5}}\,(0,2,\sqrt{7},\sqrt{13},\sqrt{23}\,)$$

for different initial z_0, the outputs are given in Table 14.

Table 14 Outputs for different initial z_0 (Example 26)

choice	z_0	z_1	$z_2 = \lambda_1$
(17)	3.84977	3.05196	3.03673
(18)	1.72924	3.05715	3.03673
(19)	3.00156	3.03675	3.03673

Next, consider the general conservative Q-matrices $Q = (q_{ij}\colon i, j \in E)$. Here, the conservativity means that $\sum_{j\in E} q_{ij} = 0$ for every $i \in E$. Next, define an auxiliary Q-matrix Q_1 which coincides with Q except replacing the element q_{NN} by cq_{NN}, where $c > 1$ is an arbitrary constant and is fixed to be 1000 in what follows for simplicity.

Following Section 4 (replacing Q by Q_1), let $(x_0, x_1, \ldots, x_N)$ (with $x_0 = 1$) be the solution to the equation

$$x^{\backslash 0\text{'s row}} = P^{\backslash 0\text{'s row}}\, x, \tag{20}$$

where

$$P = \mathrm{Diag}(q_0^{-1}, q_1^{-1}, \ldots, q_{N-1,N-1}^{-1}, (cq_{NN})^{-1})Q_1 + I.$$

To go further, we need $\mu = (\mu_0, \mu_1, \ldots, \mu_N)$ with $\mu_0 = 1$, which is the same as defined in Section 4: the solution to the equation

$$Q^{*\,\backslash \text{the last row}}\mu^* = 0.$$

Having x and μ at hand, we are ready to define our initials. For each $r \in [0, 1]$, to be specified later, define

$$\begin{gathered}\widetilde{v}_0 = (r, \sqrt{1-x_1}, \sqrt{1-x_2}, \ldots, \sqrt{1-x_N}\,)^*, \quad \overline{v}_0 = \widetilde{v}_0 - \frac{\mu\widetilde{v}_0}{\sum_{k=0}^N \mu_k},\\ v_0 = \frac{\overline{v}_0}{\|\overline{v}_0\|_\mu}, \quad z_0 = \frac{(\overline{v}_0, -Q\widetilde{v}_0)_\mu}{\|\overline{v}_0\|_\mu^2}.\end{gathered} \tag{21}$$

Because $\widetilde{v}_0$ depends on r, so do $\overline{v}_0$, v_0, and $z_0 =: z_0(r)$. Choose $r_0 \in [0, 1]$ so that

$$z_0(r_0) \approx \inf_{r\in[0,1]} z_0(r).$$

Corresponding to this specified r_0, we obtain our initials v_0 and z_0. This minimizing procedure in r is necessary for avoiding collapse since we are in a more sensitive situation than before. Then the iteration procedure is exactly the same as we used several times before.

The reason we adopt a large $c = 1000$ here is that for a larger c, its minimal eigenvalue $\lambda_0(Q_1)$ is closer to, but less than, the eigenvalue $\lambda_1(Q)$ we are interested. Refer to [1, Proposition 3.2] for more details. Thus, one may regard the former as an approximation of the latter. In other words, we can use an alternative initial

$$z_0 = \lambda_0(Q_1) \text{ or its estimates studied in previous sections.} \tag{22}$$

Certainly, one can define a convex combination of those given in (21) and (22) in an obvious way, but it is omitted here. The use of $\lambda_0(Q_1)$ seems necessary (especially for large N) to avoid some pitfall, as mentioned before.

The next example is interesting for which some of its eigenvalues are complex but the one we are interested is real.

Example 27 Let

$$Q = \begin{pmatrix} -30 & 30 & 0 & 0 \\ 1/5 & -17 & 84/5 & 0 \\ 11/28 & 275/42 & -20 & 1097/84 \\ 55/3291 & 330/1097 & 588/1097 & -2809/3291 \end{pmatrix}.$$

Then

$$Q_1 = \begin{pmatrix} -30 & 30 & 0 & 0 \\ 1/5 & -17 & 84/5 & 0 \\ 11/28 & 275/42 & -20 & 1097/84 \\ 55/3291 & 330/1097 & 588/1097 & -2809000/3291 \end{pmatrix}.$$

The eigenvalues of $-Q$ and $-Q_1$ are

$$29.8411 + 2.45214\,\mathrm{i}, \quad 29.8411 - 2.45214\,\mathrm{i}, \quad 8.17131, \quad 0,$$

and

$$853.548, \quad 29.8249 + 2.46241\,\mathrm{i}, \quad 29.8249 - 2.46241\,\mathrm{i}, \quad 7.34195,$$

respectively. Using (21) with $r_0 \approx 0.951$, the output is

$$z_0 \approx 7.73667, \quad z_1 \approx 8.15021, \quad z_2 \approx 8.17129, \quad z_3 \approx 8.17131.$$

While using (22), the output is

$$z_0 \approx 7.34195, \quad z_1 \approx 8.13216, \quad z_2 \approx 8.17124, \quad z_3 \approx 8.17131.$$

Here is one more example.

Example 28 Let

$$Q = \begin{pmatrix} -57 & 118/27 & 91/9 & 1148/27 \\ 135/59 & -52 & 637/59 & 2296/59 \\ 243/91 & 590/91 & -47 & 492/13 \\ 351/287 & 118/41 & 195/41 & -62/7 \end{pmatrix}.$$

Then

$$Q_1 = \begin{pmatrix} -57 & 118/27 & 91/9 & 1148/27 \\ 135/59 & -52 & 637/59 & 2296/59 \\ 243/91 & 590/91 & -47 & 492/13 \\ 351/287 & 118/41 & 195/41 & -62000/7 \end{pmatrix}.$$

The eigenvalues of $-Q$ and $-Q_1$ are

$$59.3118, \quad 58, \quad 47.5454, \quad 0,$$

and

$$8857.18, \quad 59.2467, \quad 58, \quad 38.7143,$$

respectively. Using (21) with $r_0 \approx 0.953$, the output is

$$z_0 \approx 47.5318, \quad z_1 \approx 47.5453, \quad z_2 \approx 47.5454.$$

While using (22), the output is

$$z_0 \approx 38.7143, \quad z_1 \approx 47.5343, \quad z_2 \approx 47.5453, \quad z_3 \approx 47.5454.$$

Acknowledgements The main results of the paper have been reported at Anhui Normal University, Jiangsu Normal University, the International Workshop on SDEs and Numerical Methods at Shanghai Normal University, Workshop on Markov Processes and Their Applications at Hunan University of Arts and Science, and Workshop of Probability Theory with Applications at University of Macau. The author acknowledges Professors Dong-Jin Zhu, Wan-Ding Ding, Ying-Chao Xie, Xue-Rong Mao, Xiang-Qun Yang, Xu-Yan Xiang, Jie Xiong, Li-Hu Xu, and their teams for very warm hospitality and financial support. The author also thanks Ms. Yue-Shuang Li for her assistance in computing large matrices. This work was supported in part by the National Natural Science Foundation of China (Grant No. 11131003), the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Chen M F. Explicit bounds of the first eigenvalue. Sci China Ser A, 2000, 43(10): 1051–1059
2. Chen M F. Variational formulas and approximation theorems for the first eigenvalue. Sci China Ser A, 2001, 44(4): 409–418
3. Chen M F. From Markov Chains to Non-equilibrium Particle Systems. 2nd ed. Singapore: World Scientific, 2004
4. Chen M F. Eigenvalues, Inequalities, and Ergodic Theory. London: Springer, 2005
5. Chen M F. Speed of stability for birth–death processes. Front Math China, 2010, 5(3): 379–515
6. Chen M F. Criteria for discrete spectrum of 1D operators. Commun Math Stat, 2014, 2: 279–309
7. Chen M F. Unified speed estimation of various stabilities. Chinese J Appl Probab Statist, 2016, 32(1): 1–22
8. Chen M F, Zhang X. Isospectral operators. Commun Math Stat, 2014, 2: 17–32
9. Chen M F, Zhang Y H. Unified representation of formulas for single birth processes. Front Math China, 2014, 9(4): 761–796
10. Golub G H, van Loan C F. Matrix Computations. 4th ed. Baltimore: Johns Hopkins Univ Press, 2013
11. Hua L K. Mathematical theory of global optimization on planned economy, (II) and (III). Kexue Tongbao, 1984, 13: 769–772 (in Chinese)

12. Langville A N, Meyer C D. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton: Princeton Univ Press, 2006
13. Meyer C. Matrix Analysis and Applied Linear Algebra. Philadelphia: SIAM, 2000
14. von Mises R, Pollaczek-Geiringer H. Praktische Verfahren der Gleichungsaufösung. ZAMM Z Angew Math Mech, 1929, 9: 152–164
15. Wielandt H. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme. Teil V: Bestimmung höherer Eigenwerte durch gebrochene Iteration. Bericht B 44/J/37, Aerodynamische Versuchsanstalt Göttingen, Germany, 1944
16. Wilkinson J H. The Algebraic Eigenvalue Problem. Oxford: Oxford Univ Press, 1965

Advances in Mathematics (China) 2017
Vol. 46, No. 4, 281–297.

The Charming Leading Eigenpair

Mu-Fa Chen

(Beijing Normal University)

June 8, 2016

Abstract The leading eigenpair (the couple of eigenvalue and its eigenvector) or the first nontrivial one has different names in different contexts. It is the maximal one in the matrix theory. The talk starts from our new results on computing the maximal eigenpair of matrices. For the unexpected results, our contribution is the efficient initial value for a known algorithm. The initial value comes from our recent theoretic study on the estimation of the leading eigenvalues. To which we have luckily obtained unified estimates which consist of the second part of the talk. In the third part of the talk, the original motivation of the study along this direction is explained in terms of a specific model. The paper is concluded by a brief overview of our study on the leading eigenvalue, or more generally on the speed of various stabilities.

2000 *Mathematics Subject Classification*: 15A18, 65F15, 93E15
Key words and phrases. Leading eigenpair, efficient initial, tridiagonal matrix, speed estimation, Hardy (Poincaré)-type inequality, φ^4-model.

§1 Computing the maximal eigenpair

We begin with the following Perron-Frobenius theorem. For positive A (pointwise), the result is due to Perron, and in the nonnegative irreducible case, it is due to Frobenius. The theorem says there exists uniquely a maximal eigenvalue $\rho(A) > 0$ with positive left-eigenvector u and right-eigenvector g:

$$uA = \lambda u, \qquad Ag = \lambda g, \qquad \lambda = \rho(A).$$

These eigenvectors are also unique up to a constant.

Here is a simplest example due to Luo-Geng Hua (Loo-Keng Hua) (1984) (refer to [3; Chapter 10] for references within):

Example 1 (Hua, 1984) Let

$$A = \frac{1}{100}\begin{pmatrix} 25 & 14 \\ 40 & 12 \end{pmatrix}.$$

Then its maximal eigenvalue $\rho(A)$, the left-eigenvector u, and right-eigenvalue g are, respectively, as follows

$$\begin{aligned} \rho(A) &= \big(37 + \sqrt{2409}\,\big)/200, \\ u &= \big(5\big(13+\sqrt{2409}\,\big)/7,\ 20\big) \approx (44.34397483, 20), \\ g &= \big(\big(13+\sqrt{2409}\,\big)/4,\ 20\big)^*. \end{aligned}$$

Such a simple matrix is already enough to show the great importance of computing the maximal eigenpair. Recall a simple description of an economic system is using its structure matrix (the matrix of expanding coefficients) A, which is nonnegative, irreducible and invertible. Then the well-known input-output method can be expressed as

$$x_n = x_0 A^{-n}, \qquad n \geqslant 1.$$

where x_0 is the input (row vector) and $x_n = \left(x_n^{(0)}, \cdots, x_n^{(d)}\right)$ is the output of the products we are interested at the nth year. In 1984, Hua proved the following fundamental theorem:

Theorem 2 (Hua's Fundamental Theorem, 1984)

- The optimal choice of x_0 is u, it has the fastest grow: $x_n = x_0\,\rho(A)^{-n}$.
- Except some very special A, if $x_0 \neq u$, then the economic system will be collapsed. That is, some component of the products at some year becomes nonpositive.

Certainly, we do not care if the collapse time is very large, say 10^4 years for instance. However, it is not the case in practice. Table 1 shows the collapse time of Example 1 for the initials different from u.

Table 1 Input and collapse time

x_0	Collapse time n
(44, 20)	3
(44.344, 20)	8
(44.34397483, 20)	13

If we take only the integer part of u as x_0, then the system collapses at the third year; if we take 3 decimals, then the system collapses at the eighth year; finally, if we take all 8 decimals, then the system collapses at the thirteenth year. This result clearly shows the importance of the study on the maximal eigenpair. We need not only high precision but also need to face large systems.

We now study how to compute the maximal eigenpair. Before doing so, let us make two remarks.

1) We need to study the right-eigenvector g only. Otherwise, use the transpose A^* instead of A.

2) The matrix A is required to be irreducible with nonnegative off-diagonal elements, its diagonal elements can be arbitrary. Otherwise, use a shift $A+mI$ for large m:

$$(A + mI)g = \lambda g \Longleftrightarrow Ag = (\lambda - m)g,$$

their eigenvector remains to be the same but the maximal eigenvalues are shifted.

Consider the following example.

Example 3 Consider the matrix

$$Q = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -5 & 2^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2^2 & -13 & 3^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3^2 & -25 & 4^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4^2 & -41 & 5^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5^2 & -61 & 6^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6^2 & -85 & 7^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7^2 & -113 \end{pmatrix}.$$

The main character of the matrix is the sequence $\{k^2\}$. For this Q, the maximal eigenvalue is -0.525268 with eigenvector:

$$g \approx (55.878,\ 26.5271,\ 15.7059,\ 9.97983,\ 6.43129,\ 4.0251,\ 2.2954,\ 1)^*,$$

where the vector v^* = the transpose of v.

Actually, this matrix is truncated from the corresponding infinite one, in which case we have known that the maximal eigenvalue is $-1/4$ (refer to [5; Example 3.6]).

We now want to practice the standard algorithms in matrix eigenvalue computation. The first method in computing the maximal eigenpair is the *Power Iteration*, introduced in 1929. Starting from a vector v_0 having a nonzero component in the direction of g, normalized with respect to a norm $\|\cdot\|$. At the kth step, iterate v_k by the formula

$$v_k = \frac{Av_{k-1}}{\|Av_{k-1}\|}, \quad z_k = \|Av_k\|, \qquad k \geqslant 1.$$

Then we have the convergence: $v_k \to g$ and $z_k \to \rho(Q)$ as $k \to \infty$. If we rewrite v_k as

$$v_k = \frac{A^k v_0}{\|A^k v_0\|},$$

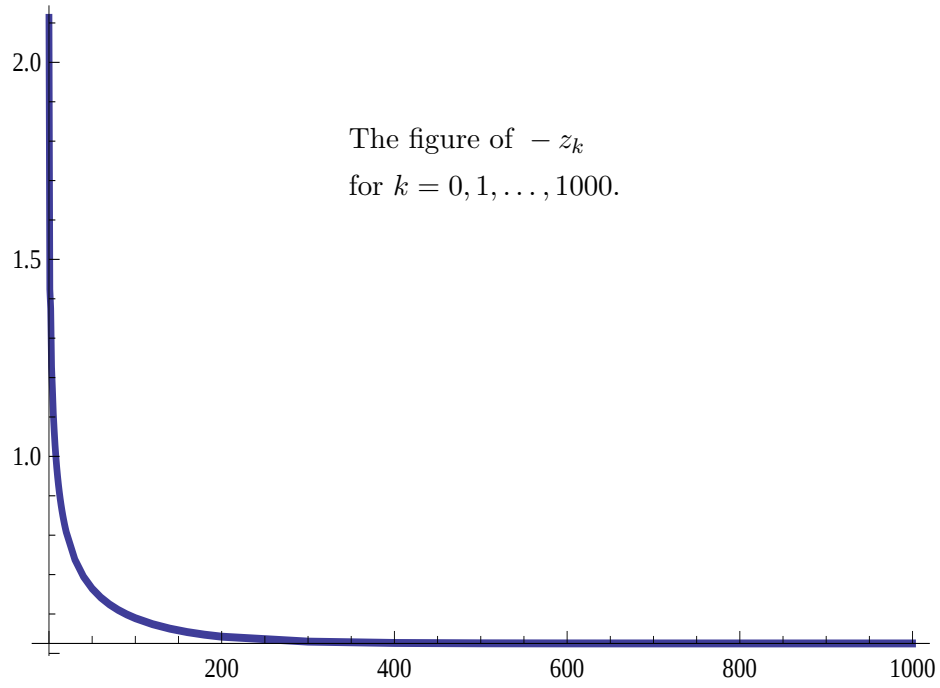
one sees where the name "power" comes from. For our example, to use the Power Iteration, we adopt the ℓ^1-norm and choose $v_0 = \tilde{v}_0/\|\tilde{v}_0\|$, where

$$\tilde{v}_0 = (1,\ 0.587624,\ 0.426178,\ 0.329975,\ 0.260701,\ 0.204394, 0.153593, 0.101142)^*.$$

This initial comes from a formula to be given in the last part of this section. Comparing it with g, noting that the eigenvector g decays from 56 to 1, here $\tilde{v}_0$ decays from 10 to 1, one may worry about the effectiveness of the choice of v_0. Anyhow, having the experience of computing its eigensystem, I expect to finish the computation in a few of seconds. Unfortunately, I got a difficult time to compute the maximal eigenpair for this simple example. Altogether, I computed it for 180 times, not in one day, using 1000 iterations. The printed pdf-file of the outputs has 64 pages. Here are some data.

Table 2 Outputs $(k, -z_k)$

0	2.11289		50	0.664453
1	1.42407		100	0.589332
2	1.37537	Computing	200	0.542423
3	1.22712	180 times,	300	0.529909
4	1.1711	10^3 iterations,	400	0.526517
5	1.10933	64 pages.	500	0.525603
6	1.06711		600	0.525358
7	1.02949		700	0.525292
8	0.998685	$(k, -z_k)$	800	0.525274
9	0.971749		900	0.52527
10	0.948331		$\geqslant 990$	0.525268



The figure of $-z_k$
for $k = 0, 1, \ldots, 1000$.

The first ten iterations reduce the estimate of the maximal eigenvalue from 2 to 1. It is quite good. Then, we receive the wished output only at the 990th iteration. The corresponding figure shows that the convergence of z_k goes quickly at the beginning of the iterations. This means that our initial v_0 is good enough. Then the convergence goes very slow which means that the Power Iteration Algorithm converges very slowly.

Let us now move to the second algorithm in computing the maximal eigenpair, the *Rayleigh Quotient Iteration* (RQI), a variant of the *Inverse It-*

eration presented in 1944. Here we use the ℓ^2-norm. Starting from an approximating pair (z_0, v_0) of the maximal pair $(\rho(A), g)$ with $v_0^* v_0 = 1$, use the following iteration.

$$v_k = \frac{(A - z_{k-1}I)^{-1} v_{k-1}}{\|(A - z_{k-1}I)^{-1} v_{k-1}\|}, \qquad z_k = v_k^* A v_k, \qquad k \geqslant 1.$$

If (z_0, v_0) is close enough to $(\rho(A), g)$, then

$$v_k \to g \quad \text{and} \quad z_k \to \rho(A) \quad \text{as } k \to \infty.$$

Before moving further, let us make a remark about this algorithm. Without using the shift $z_{k-1}I$, it is the original Inverse Iteration:

$$v_k = \frac{A^{-1} v_{k-1}}{\|A^{-1} v_{k-1}\|} \Longleftrightarrow v_k = \frac{A^{-k} v_0}{\|A^{-k} v_0\|} \quad \text{i.e. the input-output method.}$$

From this, one may obtain a short proof of Hua's magical assertion in his fundamental theorem. The use of a constant shift zI for z closed enough to $\rho(A)$ enables us to compute the eigenvector corresponding to $\rho(A)$ rather than $\lambda_{\min}(A)$. The use of a variant shift $z_{k-1}I$ is for accelerating the convergence speed.

Having the hard time spent in the last computation, I was in hesitation to go to the second algorithm. I wondered how many iterations are required using the second algorithm. To have a feeling, I used optimization theory. Suppose we are searching the maximum on the interval (0, 1) for the accuracy of 10^{-6}. Then, by using the Golden Section Search,

$$10^{-6} = 0.618^{24}.$$

This means that 24 iterations at least are required. By the Bisection Method,

$$10^{-6} = 0.5^{20}.$$

Thus, I did not believe that we can complete the job in 20 iterations. With enough patience and energy, I started my computation again. The result came to me, not enough to say surprisingly, I was shocked indeed.

Example 4 For the same matrix Q and $\tilde{v}_0$ as in Example 1, by RQI, we need two iterations only:

$$z_1 \approx -0.528215, \quad z_2 \approx -0.525268.$$

This shows not only the power of the second method but also the effectiveness of my v_0. For simplicity, from now on, we set $\lambda_j := \lambda_j(-Q)$. In particular $\lambda_0 = -\rho(Q) > 0$.

As usual, "too good" is dangerous. For instance, a too beautiful person may have a lot of trouble. Instead of our previous v_0, we adopt the uniformly distributed one:

$$v_0 = \{1, 1, 1, 1, 1, 1, 1, 1\}/\sqrt{8}.$$

This is somehow fair since we may have no knowledge about g in advance.

Example 5 Let Q be the same as above and use the uniformly distributed v_0. Then

$$(z_1, z_2, z_3, \boldsymbol{z_4}) \approx (4.78557,\ 5.67061,\ 5.91766,\ \mathbf{5.91867}).$$

$$(\lambda_0, \lambda_1, \boldsymbol{\lambda_2}) \approx (0.525268,\ 2.00758,\ \mathbf{5.91867}).$$

The computation becomes stable at the 4th iteration. Unfortunately, it is not what we want λ_0 but λ_2. In other words, the algorithm converges to a pitfall. Very often, there are $n-1$ pitfalls for a matrix having n eigenvalues. This shows once again our initial $\tilde{v}_0$ is efficient.

In the last example, z_0 is chosen in the automatic way: $z_0 = v_0^*(-Q)v_0$. If we keep this v_0 which is not so good, but using a new z_0, then we come back to our result in two iterations.

Example 6 Let Q and v_0 be the same as in the last example. Choose

$$z_0 = 2.05768^{-1} \approx 0.485985.$$

Then $z_1 \approx 0.525313$, $z_2 \approx 0.525268$.

This shows that the new z_0 ($= \delta^{-1}$ to be specified at the end of this section) is efficient.

We have now computed the same example for 4 times. Here is the comparison of different initials.

Table 3 Comparison of different initials

Q	v_0	z_0	# of Iterations
1	$\tilde{v}_0$	Power	10^3
2	$\tilde{v}_0$	Automatic	**2**
3	Uniformly distributed	Automatic	Collapse
4	Uniformly distributed	δ_1^{-1}	**2**

We now come to the following conclusion.

- RQI is much efficient than Power One.
- The initials (v_0, z_0) are very sensitive and our $\tilde{v}_0$ and $z_0 = \delta_1^{-1}$ are efficient.
- It is very hard to handle with the initials. Actually, a large part of mathematics research are devoted to this problem.

Hopefully, everyone here has heard the name *Google's PageRank*. In other words, the Google's search is based on the maximal left-eigenvector (Exactly the same as what used in the Hua's Theorem 2). On this topic, the following book was published 10 years ago:

Langville, A.N. and Meyer, C. D. (2006).

Google's PageRank and Beyond: The Science of Search Engine Rankings.

Princeton University Press.
In this book, the Power Iteration is included but not the RQI.

Up to now, we have discussed only a small size ($8 \times 8\,(N = 7)$) matrix. How about large N? In computational mathematics, one often expects the number of iterations grows in a polynomial way N^α for α greater or equal to 1. In our efficient case, since $2 = 8^{1/3}$, we expect to have $10000^{1/3} = 22$ iterations. The next table subverts completely my imagination.

Table 4 Comparison of RQI for different N

$N+1$	z_0	z_1	$z_2 = \lambda_0$	upper/lower
8	0.523309	0.525268	0.525268	$1+10^{-11}$
100	0.387333	0.376393	0.376383	$1+10^{-8}$
500	0.349147	0.338342	0.338329	$1+10^{-7}$
1000	0.338027	0.327254	0.32724	$1+10^{-7}$
5000	0.319895	0.30855	0.308529	$1+10^{-7}$
7500	0.316529	0.304942	0.304918	$1+10^{-7}$
10^4	0.31437	0.302586	0.302561	$1+10^{-7}$

Here z_0 is defined by

$$z_0 = 7/(8\delta_1) + v_0^*(-Q)v_0/8,$$

where v_0 and δ_1 are computed by our general formulas to be defined very soon below. We computed the matrices of order $8, 100, \ldots, 10^4$ by using MatLab in a notebook, in no more than 30 seconds, the iterations finished at the second step. This means that the outputs starting from z_2 are the same and coincide with λ_0. See the first row for instance, which becomes stable at the first step indeed. We did not believe such a result for some days, so we checked it in different ways. First, since $\lambda_0 = 1/4$ when $N = \infty$, the answers of λ_0 given in the fourth column are reasonable. More essentially, by using the output v_2, we can deduce upper and lower bounds of λ_0 (using [5; Theorem 2.4 (3)]), and then the ratio upper/ lower is presented in the last column. For the first row, by using v_1 instead of v_2, we also have $1 + 10^{-7}$. In each case, the algorithm is significant up to 6 digits.

It is the position to write down the formulas of $\tilde{v}_0$ and δ_1. Then our initial z_0 used in Table 4 is a little modification of δ_1^{-1}: a convex combination of δ_1^{-1} and $v_0^*(-Q)v_0$.

Let us consider the tridiagonal matrix. Fix $N \geqslant 1$ and denote by $E = \{0, 1, \ldots, N\}$ the set of indices. By a shift if necessary, we may reduce A to Q

with negative diagonals: $Q = A - mI$, $m := \max_{i\in E} \sum_{j\in E} a_{ij}$,

$$Q = \begin{pmatrix} -(b_0+c_0) & b_0 & 0 & 0 & \cdots \\ a_1 & -(a_1+b_1+c_1) & b_1 & 0 & \cdots \\ 0 & a_2 & -(a_2+b_2+c_2) & b_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & a_N & -(a_N+c_N)) \end{pmatrix}.$$

Thus, we have three sequences $\{a_i > 0\}$, $\{b_i > 0\}$, and $\{c_i \geqslant 0\}$. Our main assumption here is that the first two sequences are positive. In order to define our initials, we need three new sequences, $\{\mu_k\}$ (speed measure), $\{h_k\}$, and $\{\varphi_k\}$.* The sequence $\{\mu_k\}$ uses $\{a_k\}$ and $\{b_k\}$ only, independent of $\{c_k\}$:

$$\mu_0 = 1, \;\; \mu_n = \mu_{n-1}\frac{b_{n-1}}{a_n}, \qquad 1 \leqslant n \leqslant N.$$

Here and in what follows, our iterations are often of one-step. Next, we define the sequence $\{h_k\}$:

$$h_0 = 1, \;\; h_n = h_{n-1}r_{n-1}, \qquad 1 \leqslant n \leqslant N;$$

here we need another sequence $\{r_k\}$:

$$r_0 = 1 + c_0/b_0, \;\; r_n = 1 + \frac{a_n + c_n}{b_n} - \frac{a_n}{b_n r_{n-1}}, \qquad 1 \leqslant n < N.$$

The boundary of h is defined by

$$h_{N+1} = c_N h_N + a_N(h_N - h_{N-1}).$$

Note that if $c_k \equiv 0$ for $k < N$, then we do not need the sequence $\{h_k\}$, simply set $h_k \equiv 1$. Having $\{\mu_k\}$ and $\{h_k\}$ at hand, we can define $\{\varphi_k\}$ as follows.

$$\varphi_n = \sum_{k=n}^{N} \frac{1}{h_k h_{k+1} \mu_k b_k}, \qquad 0 \leqslant n \leqslant N, \; b_N := 1.$$

We are now ready to define v_0 and δ_1 (or z_0) using the three new sequences.

$$\tilde{v}_0(i) = h_i\sqrt{\varphi_i}, \; i \leqslant N; \qquad v_0 = \tilde{v}_0/\|\tilde{v}_0\|; \quad \|\cdot\| := \|\cdot\|_{L^2(\mu)}$$

$$\delta_1 = \max_{0\leqslant n\leqslant N}\left[\sqrt{\varphi_n}\sum_{k=0}^{n}\mu_k h_k^2\sqrt{\varphi_k} + \varphi_n^{-1/2}\sum_{n+1\leqslant j\leqslant N}\mu_j h_j^2 \varphi_j^{3/2}\right] =: z_0^{-1}.$$

*A modification of the algorithm here is presented in [14; Apendix §4.4].

Note that v_0 and δ_1 are explicitly expressed by these three new sequences. In other words, we have used three new sequences $\{\mu_k\}$, $\{h_k\}$, and $\{\varphi_k\}$ instead of the original three $\{a_i\}$, $\{b_i\}$, and $\{c_i\}$.

Finally, the RQI goes as follows. Solve w_k:

$$(-Q - z_{k-1}I)w_k = v_{k-1}, \qquad k \geqslant 1;$$

and define

$$v_k = w_k/\|w_k\|, \qquad z_k = (v_k, -Q\,v_k)_{L^2(\mu)}.$$

Then

$$v_k \to g \quad \text{and} \quad z_k \to \lambda_0 \qquad \text{as } k \to \infty.$$

Certainly, the next step is going to the general matrix from the tridiagonal one. This is possible once we understand the probabilistic meaning of the sequences $\{\mu_k\}$, $\{h_k\}$, and $\{\varphi_k\}$. This work is done in [12] but omitted here. For more recent progress on this topic, refer to [13, 14].

§2 Unified speed estimation of various stabilities

We are now going to explain the reason why our initials are efficient. The answer comes from the following result about the unified speed estimation of various stabilities. The result is a short summary of a series of the author's papers published during 2010–2014, starting from 1988. Refer also to [17].

Theorem 7 (Informal!) For a tridiagonal matrix Q or a one-dimensional elliptic operator (order 2) with or without killing on a finite or infinite interval, in each of twenty cases, there exist explicit δ, δ_1, δ_1' (and then δ_n, δ_n', recursively) such that $\delta_n' \uparrow$, $\delta_n \downarrow$ and

$$(4\delta)^{-1} \leqslant \delta_n^{-1} \leqslant \lambda_0 \leqslant {\delta_n'}^{-1} \leqslant \delta^{-1}, \qquad n \geqslant 1.$$

Besides, $1 \leqslant {\delta_1'}^{-1}/{\delta_1}^{-1} \leqslant 2$.

The initial δ_1 used in the previous section is taken from here in one specific case. Then the $\tilde{v}_0$ used there was originally used in [5; §3] to deduce δ_1.

Certainly, the notation λ_0 and $\delta_\#$ here may be changed case by case. For instance, for the exponentially ergodic rate (or the exponential decay rate), λ_0 is replaced by α^*. By [5; Theorems 1.5 and 7.4] (discrete case) and [6; Theorem 2.1 and Proposition 6.1] (continuous case), the rate α^* coincides with $\lambda^\#$ to be discussed immediately below and so the study on α^* is omitted here.

We now leave the matrix situation and move to differential operators. First, we consider a special case in parallel to the tridiagonal matrix. Define the operator

$$L^c = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} - c(x), \quad a(x) > 0,\ c(x) \geqslant 0$$

on $(0, N)$ with $N \leqslant \infty$. Certainly, by a shift if necessary, one may relax the condition "$c(x) \geqslant 0$." To study the maximal eigenpair of L^c, instead of the triple (a, b, c) of functions, we introduce three functions $\mathrm{d}\mu/\mathrm{d}x$, h, and φ as follows. Let

$$\frac{\mathrm{d}\mu}{\mathrm{d}x} = \frac{e^C}{a}, \qquad C(x) := \int_0^x \frac{b}{a},$$

where the Lebesgue measure $\mathrm{d}x$ is omitted in the last integral; let h be positive L^c-harmonic: $L^c h = 0$; and let

$$\varphi(x) = \int_0^x \frac{e^{-C}}{h^2}.$$

Having $(\mathrm{d}\mu/\mathrm{d}x, h, \varphi)$ at hand, as in the discrete case, we can define $\tilde{v}_0$ and $z_0 = \delta_1^{-1}$ as follows.

$$\tilde{v}_0 = h\sqrt{\varphi},$$
$$\delta_1 = \sup_{0 \leqslant x \leqslant N} \left[\sqrt{\varphi(x)} \int_0^x h^2 \sqrt{\varphi}\, \mathrm{d}\mu + \varphi(x)^{-1/2} \int_x^N h^2 \varphi^{3/2}\, \mathrm{d}\mu \right].$$

We now go to a more general setup. Consider the space $E = (-M, N)$, $M, N \leqslant \infty$ and the eigenvalue problem:

$$\text{Eigenequation}: \quad Lg = -\lambda g, \quad g \neq 0$$

for some differential operator L. Here we use codes 'D' and 'N' to denote the Dirichlet or Neumann boundary, respectively.

D: (Absorbing) Dirichlet boundary,

N: (Reflecting) Neumann boundary $g'(-M) = 0$,

where $g(-\infty) := \lim_{M \to \infty} g(-M)$. Similarly we have $g'(-\infty)$ and others. Correspondingly, we have four types of eigenvalues.

- λ^{NN}: Neumann boundaries at $-M$ and N.
- λ^{DD}: Dirichlet boundaries at $-M$ and N.
- λ^{DN}: Dirichlet at $-M$ and Neumann at N.
- λ^{ND}: Neumann at $-M$ and Dirichlet at N.

Given an elliptic operator $L = L^0$:

$$L = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x},$$

define the speed measure μ and scale measure $\hat{\nu}$, respectively, as follows

$$\frac{\mathrm{d}\mu}{\mathrm{d}x} = \frac{e^C}{a}, \quad \frac{\mathrm{d}\hat{\nu}}{\mathrm{d}x} = e^{-C}, \qquad C(x) := \int_\theta^x \frac{b}{a},$$

where $\theta \in (-M, N)$ is a reference point. Then the leading eigenvalues $\lambda^{\#}$ defined above describe, respectively, the following $L^2(\mu)$-exponential convergence of the semigroup $\{P_t = e^{tL}\}_{t\geqslant 0}$:

$$\|P_t f\| \leqslant \|f\|\, e^{-\lambda^{\rm NN} t}, \quad \mu(f) := \int_E f \mathrm{d}\mu = 0,$$
$$\|P_t f\| \leqslant \|f\|\, e^{-\lambda^{\#} t}, \quad t \geqslant 0,\ f \in L^2(\mu),\ \text{ if } \# \text{ is not NN}.$$

Thus, $\lambda^{\rm NN}$ describes the L^2-exponentially ergodic rate and the other $\lambda^{\#}$ describe the L^2-exponential decay rate.

Here is our main result in this part of the talk.

Theorem 8 (Chen, 2010) For each # of 4 cases, we have the following unified estimates

$$\left(4\kappa^{\#}\right)^{-1} \leqslant \lambda^{\#} \leqslant \left(\kappa^{\#}\right)^{-1},$$

where

$$\left(\kappa^{\rm NN}\right)^{-1} = \inf_{x<y} \left\{\mu(-M, x)^{-1} + \mu(y, N)^{-1}\right\} \hat{\nu}(x, y)^{-1}$$
$$\left(\kappa^{\rm DD}\right)^{-1} = \inf_{x\leqslant y} \left\{\hat{\nu}(-M, x)^{-1} + \hat{\nu}(y, N)^{-1}\right\} \mu(x, y)^{-1}$$
$$\kappa^{\rm DN} = \sup_{x\in(-M,N)} \hat{\nu}(-M, x)\, \mu(x, N)$$
$$\kappa^{\rm ND} = \sup_{x\in(-M,N)} \mu(-M, x)\, \hat{\nu}(x, N)$$

and $\mu(\alpha, \beta) = \int_\alpha^\beta \mathrm{d}\mu$. In particular, $\lambda^{\#} > 0$ iff $\kappa^{\#} < \infty$.

The beauty of the theorem is displayed in the following aspects.

- Each of the estimates has a universal factor 4.
- Each constant $\kappa^{\#}$ is expressed by μ and $\hat{\nu}$ only.
- In the expressions of $\kappa^{\rm NN}$ and $\kappa^{\rm DD}$, two boundaries are symmetric.
- An intrinsic relation between the four constants $\kappa^{\#}$ can be expressed as follows.

$$\begin{array}{ccc} \kappa^{\rm DD} & \xrightarrow{\text{Remove } \hat{\nu}(y,N)^{-1}} & \kappa^{\rm DN} \\ \Big\updownarrow \text{Rule} & & \Big\updownarrow \text{Rule} \\ \kappa^{\rm NN} & \xrightarrow{\text{Remove } \mu(y,N)^{-1}} & \kappa^{\rm ND} \end{array}$$

Rule:

Exchange of codes D and N in $\lambda^{\#}$

$\Longleftrightarrow$ exchange μ and $\hat{\nu}$ in $\kappa^{\#}$

We remark that the theorem is not as simple as it stands. In the DN case for instance, it was started by G.H. Hardy in 1920 and completed half a century later by B. Muckenhoupt et al around 1970. To obtain the answer in the bilateral cases, one has to wait for another 40 years until 2010. The proofs in the DD- and NN-cases use three advanced mathematical tools (the

coupling and distance method, the dual technique, and the capacitary method) and were completed in five steps (refer to [5]).

There are two ways to generalize the above theorem. The first one is including the potential term c, that is, using L^c instead of L. Again, assume $E = (-M, N),\ M, N \leqslant \infty$. First, we consider the *Poincaré-type inequalities*:

$$\lambda_c^{\#} \|f\|_{\mu,2}^2 \leqslant \|f'\|_{\nu,2}^2 + \|cf\|_{\mu,2}^2,$$

where

$$\nu(\mathrm{d}x) = e^{C(x)}\mathrm{d}x, \qquad \hat{\nu}(\mathrm{d}x) = e^{-C(x)}\mathrm{d}x,$$

and $\|\cdot\|_{\mu,p} = \|\cdot\|_{L^p(\mu)}$. The inequality becomes equality once $f = g$: the eigenfunction corresponding to $\lambda_c^{\#}$. This explains the relationship between the inequality and its corresponding eigenvalue. In particular, when $c \equiv 0$, we return to what we have already studied above:

$$\sqrt{\lambda^{\#}}\, \|f\|_{\mu,2} \leqslant \|f'\|_{\nu,2}.$$

This leads to the second generalization (generalized to the nonlinear situation): the *Hardy-type inequalities*:

$$\|f\|_{\mu,q} \leqslant A^{\#} \|f'\|_{\nu,p}, \qquad p, q \in (1, \infty).$$

We use these inequalities to describe the algebraic convergence $t^{-\alpha}$ for some $\alpha > 0$. Corresponding to ν in such a general setup, we have

$$\hat{\nu}(\mathrm{d}x) = \exp\left[-\frac{C(x)}{p-1}\right]\mathrm{d}x$$

which goes back to the previous one when $p = 2$. Finally, we can generalize the left-hand side of the last inequality to a general normed linear space $\mathbb{B}$:

$$\||f|^q\|_{\mathbb{B}}^{1/q} \leqslant A_{\mathbb{B}}^{\#} \|f'\|_{\nu,p}.$$

A particular use of this class of inequalities is to describe the exponential convergence in entropy. Note that the entropy functional does not belong to any L^q-space:

$$\|f\|_{L^1(\pi)} \leqslant \mathrm{Ent}(f) \leqslant \|f\|_{L^{1+\varepsilon}(\pi)}^{1+\varepsilon}, \qquad \varepsilon > 0.$$

The *normed linear space* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ here means a subset of Borel measurable functions on $(X, \mathscr{X}, \mu)$ having the following norm

$$\|f\|_{\mathbb{B}} = \sup_{g \in \mathscr{G}} \int_X |f|\, g \mathrm{d}\mu,$$

for a given $\mathscr{G} \subset \mathscr{X}/\mathbb{R}_+$. If we set $\mathscr{G} = L^p\,(p > 1)$, then $\mathbb{B} = L^{p^*}$: $\frac{1}{p} + \frac{1}{p^*} = 1$. In the study of logarithmic Sobolev inequality, we use

$$\mathscr{G} = \left\{g \geqslant 0 : \int_X e^g \mathrm{d}\pi \leqslant e^2 + 1\right\}.$$

Here is a summary of 16 criteria included in Theorem 7 (Recall that, as mentioned before, for the omitted 4 cases of $\alpha^{\#}$, we have $\alpha^{\#} = \lambda^{\#}$).

Theorem 9 The optimal constants $\lambda^{\#}$ in the Poincaré-type inequalities, with/without c, satisfy

$$\kappa^{\#} \leqslant \lambda^{\#\,-1} \leqslant 4\kappa^{\#};$$

and the optimal constants $A^{\#}$ in the Hardy-type inequalities, with/without $\mathbb{B}$, satisfy

$$B^{\#} \leqslant A^{\#} \leqslant 2B^{\#},$$

where in the DD case for instance, we have

Table 5 Isoperimetric constants in different cases

$B_{\mathbb{B}}$	$\sup\limits_{x\leqslant y} \dfrac{\|\mathbb{1}_{(x,\,y)}\|_{\mathbb{B}}^{1/q}}{\left\{\hat{\nu}(-M,x)^{1-p}+\hat{\nu}(y,N)^{1-p}\right\}^{1/p}}$
$\mathbb{B}=L^1(\mu)$ B	$\sup\limits_{x\leqslant y} \dfrac{\mu(x,y)^{1/q}}{\left\{\hat{\nu}(-M,x)^{1-p}+\hat{\nu}(y,N)^{1-p}\right\}^{1/p}}$
$q=p=2$ κ	$\sup\limits_{x\leqslant y} \dfrac{\mu(x,y)}{\hat{\nu}(-M,x)^{-1}+\hat{\nu}(y,N)^{-1}}$
Killing c κ_c	$\sup\limits_{x\leqslant y} \dfrac{\mu_c(x,y)}{\hat{\nu}_c(-M,x)^{-1}+\hat{\nu}_c(y,N)^{-1}}$

In details, the first line is the most general case $\mathbb{B}$. Setting $\mathbb{B}$ to be $L^1(\mu)$, we get the second line, that is the Hardy-type inequalities for $q \geqslant p$. Setting $q = p = 2$, we get the Poincaré-type without c. By a change of μ and $\hat{\nu}$, we obtain the last line with c: $\mu_c = h^2\mu$, $\hat{\nu}_c = h^{-2}\hat{\nu}$, and h is L^c-harmonic: $L^c h = 0$.

It is remarkable that the previous proofs for the linear case ($q = p = 2$) do not suitable to the present nonlinear situation. To which, we use new analytic proofs (refer to [7] and [17]).

§3 Original motivation: study on phase transitions

One may be disappointed if I say nothing for the higher dimensional case since up to now we have worked only in dimension one. For this, let us recall the exponential convergence in L^2 or in entropy.

Let π be a probability measure and denote by $\|\cdot\|$ and $(\cdot,\cdot)$ the norm and inner product on $L^2(\pi)$. For a given self-adjoint operator L on $L^2(\pi)$:

$$(f, Lg) = (Lf, g), \qquad f, g \in \mathscr{D}(L) \subset L^2(\pi),$$

denote by $\{P_t = e^{tL}\}_{t\geqslant 0}$ be the semigroup generated by L. We have already seen the *exponential stability in L^2-sense*:

$$\|P_t f - \pi(f)\| \leqslant \|f\| e^{-\varepsilon t}, \qquad t \geqslant 0,\ f \in L^2(\pi),$$

and moreover $\varepsilon_{\max} = \lambda_1 := \lambda^{\mathrm{NN}}$. Here is an often stronger stability, *exponential stability in entropy*:

$$\mathrm{Ent}(P_t f) \leqslant \mathrm{Ent}(f) e^{-2\sigma t}, \qquad t \geqslant 0,$$
$$\mathrm{Ent}(f) := H(\mu\|\pi) = \int_E f \log f \mathrm{d}\pi, \ \text{ if } \ \frac{\mathrm{d}\mu}{\mathrm{d}\pi} = f$$

We now go to an infinite-dimensional model. For each $x : \mathbb{Z}^d \to \mathbb{R}$, the interaction potential is $H(x) = -2J \sum_{\langle i,j\rangle} x_i x_j$ for some $J \geqslant 0$, where $\langle i, j\rangle$ is the nearest neighbors in $\mathbb{Z}^d$. At each site $i \in \mathbb{Z}^d$, we have the spin potential

$$u(x_i) = x_i^4 - \beta x_i^2, \qquad x_i \in \mathbb{R},\ \beta \geqslant 0.$$

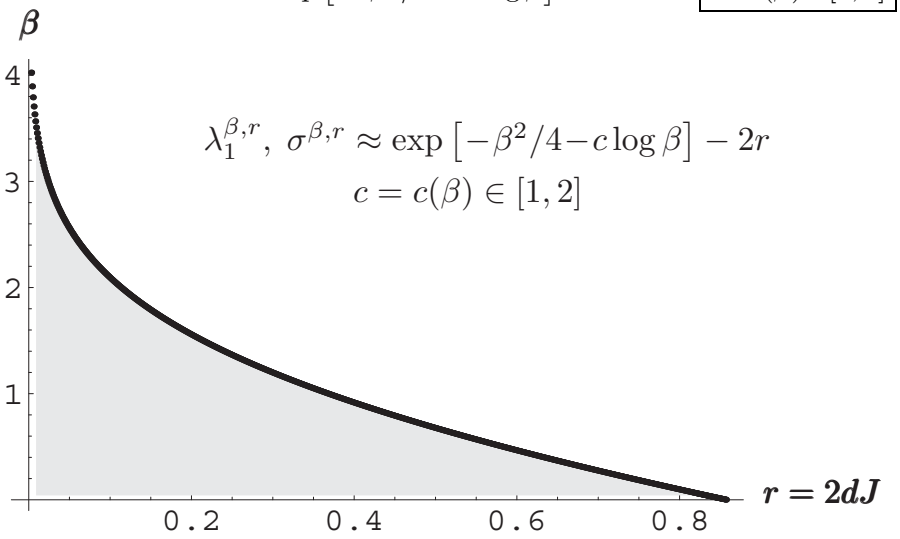
The operator for the whole system is

$$L = \sum_{i\in\mathbb{Z}^d} \big[\partial_{ii} - (u'(x_i) + \partial_i H)\partial_i\big].$$

Here is our main result for this model $\big($the φ^4-model$\big)$.

Theorem 10 (Chen, 2008)

$$\inf_{\Lambda\Subset\mathbb{Z}^d} \inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}} \lambda_1^{\beta,J}(\Lambda,\omega) \approx \inf_{\Lambda\Subset\mathbb{Z}^d} \inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}} \sigma^{\beta,J}(\Lambda,\omega)$$
$$\approx \exp\big[-\beta^2/4 - c\log\beta\big] - 4dJ \qquad \boxed{c = c(\beta) \in [1,2]}$$



Then we proved that the eigenvalue λ_1, as well as the logarithmic Sobolev constant σ have the same leading decay rate $\exp[-\beta^2/4] - 2r$. More precisely, it says that these two constants have locally such a decay rate uniformly in the finite box Λ and the boundary ω. These constants decay from positive to zero rapidly. This shows the phase transitions of the model. We mention that the leading term $-\beta^2/4$ is exact.

The model illustrates our original motivation of the study on the leading eigenvalue, to describe the phase transitions. Note that for infinite-dimensional mathematics, the known mathematical tools are very limited. We need to look for new mathematical tools. The goal of our study is developing a new way to describe the phase transitions in statistical physics. Mathematically, we are looking for a theory of stability speed, an advanced stage of the study on stability. No doubt, such a theory is valuable, as illustrated by Section 1 of the talk.

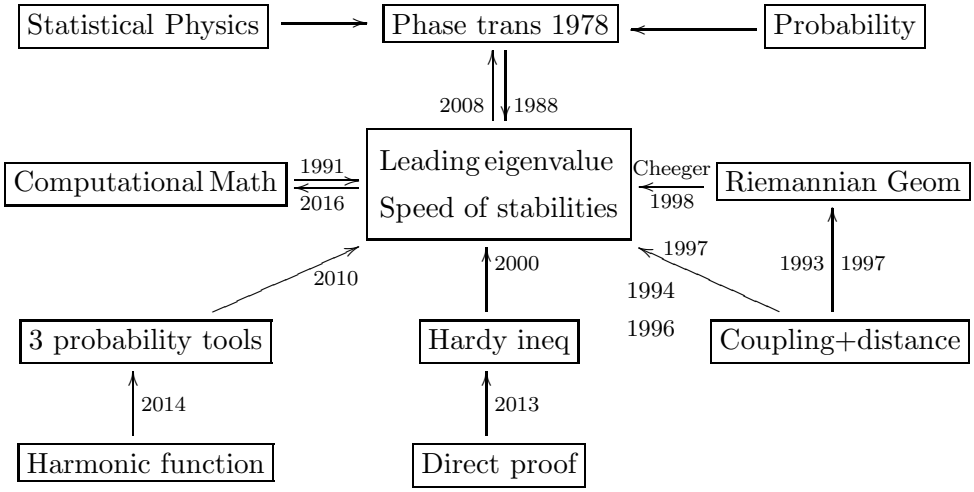
Up to now, we have discussed the easier part of Theorem 7: $(4\delta)^{-1} \leqslant \lambda_0 \leqslant \delta^{-1}$, but have not touched the harder part: $\delta_n^{-1} \leqslant \lambda_0 \leqslant {\delta'_n}^{-1}$. Hence we have not explained the way to construct $\tilde{v}_0$ and δ_1 used in §1. In the present situation, we may assume that $h_i \equiv 1$ (otherwise, use [16, 9] to reduce to this case). Then δ_1 is defined by [5; (3.4)] and $\tilde{v}_0$ is the function f_1 defined in [5; Theorem 3.2 (1)]. Therefore, to understand $(\tilde{v}_0, \delta_1)$, it suffices to have a look at the first three sections of [5]. We are not going to the details here. Instead, we prefer to have a short overview of our story, given below.

Appendix. A brief overview of the research roadmap

Here we introduce our research roadmap of the topic, and to provide some additional survey articles for the developments of the story.

In 1960's, as a product of the interaction between probability theory and statistical physics, new branches of mathematics appeared, first the random fields and then the interacting particle systems, for instance. We came to the interacting field in 1978, emphasized on the mathematical foundation of non-equilibrium particle systems. Our research results were partially collected in [2]. As we know, a central problem in the study of statistical physics is the phase transition phenomenon. Around 1988, we learnt a possible way to describe the phase transition in terms of the spectral gap (i.e. the first non-trivial eigenvalue, or more generally the leading eigenvalue) of its generator of the stochastic process. This led us to a long trip to study the leading eigenvalue or more generally the speed of various stabilities.

The author's first paper on this topic published in 1991. At the time, one could compute precisely the principal eigenvalue of the generator of a Markov Chain in only two or three examples. This was based on the main theorem in the paper: for a birth–death process, the ergodic rate (the probabilistic way to describe the the exponential stability) actually coincides with the first non-trivial eigenvalue of its generator. If you take a look at this paper and compare it with what I talked above, you will see how far we have come since then. Because our knowledge at the beginning on this topic was rather poor, we started to visit other branches of mathematics. The first one we visited is the eigenvalue computation for matrices. In the 1991's paper, we adopted an algorithm to compute the first non-trivial eigenvalue for a class of tridiagonal matrices, without analytic explicit estimation.



The next important event is, we found in 1992 that this topic was well studied in Riemannian geometry. Hence we started to learn the geometric methods, the gradient estimates, in particular. Soon we understood that our probabilistic method — the coupling method, can also be used for studying this problem. Thus, we went to an opposite way: studying the geometric topic using our probabilistic approach. This was done in several joint papers with Feng-Yu Wang. To obtain sharp estimates however, we need to examine not only the couplings but also the closely related distances. Thus, the refined method is sometimes called the coupling and distance method. In a survey article of mine, the story was summarized as "the trilogy of couplings". The same idea was also used for elliptic operators, as well as matrices. The main credit is that some new variational formula for the lower bound of the eigenvalue was discovered which then improves a number of the known sharp estimates. This may be regarded as our contribution to geometry. After 5 years or so, we also came back to the opposite direction: using some geometric approach (the Cheeger's approach, for instance) to handle with our main problem.

The third important event happened around 2000, we learnt that the Hardy-type inequality (an important subject in Harmonic Analysis) can be used in our study to provide a nice criterion for the positivity of the principal eigenvalue. This led us to establish 10 criteria for the positive property of different types of stability (or equivalently, inequalities), using our own technique. At the same time, we established new dual variational formulas for the leading eigenvalues, as well as approximating procedures in computing the eigenvalues. At this stage, a more or less systematic theory was formed. A series of lectures on the theory up to 2003 consist of the book [3].

Having worked for 20 years, in 2008, we returned to our original subject, the interacting particle systems (the φ^4-model in particular as discussed in §3) to justify the power of the results obtained until 2003. Luckily, we obtained the exact leading decay rate of the first non-trivial eigenvalue which describes

more or less the phase transition curve for the model. We recall that the submission of [4] was delayed for 5 years until we were able to figure out the exact coefficient $1/4$ in the leading rate $\beta^2/4$ given in Theorem 10.

In 2010, we present a unified treatment in [5] of the leading eigenvalue in each of the four cases (i.e. with four different boundary conditions). In this unusually long paper, we obtained not only the unified basic estimates (Theorem 8) but also the improved ones (Theorem 7). Note that the improved estimates are essential for our efficient initials as shown at the beginning of the paper. For this, we have used three probabilistic tools: the coupling and distance method, the dual technique, and the capacitary method. The main ideas of the proofs were surveyed in [6]. Unfortunately, these powerful tools in the linear case is not suitable for the non-linear one. This is the reason why, to extend the results given in [5] to the Hardy-type inequality, we have to wait for another 13 years. That is, in 2013 ([7]), we were able to do so by using new direct proofs. Refer to [8, 10] for surveys on [7]. Thus, only after 13 years known the Hardy-type inequality, we were able to make some contribution to the subject of Hardy inequalities.

The final important event happened in 2014. With Xu Zhang, in [15], we were able to treat the tridiagonal matrix with general diagonal elements, using (locally) harmonic functions. This is crucial, otherwise, we can handle only with a smaller class of tridiagonal matrices (i.e. $c_i \equiv 0$ in the last part of §1). This completes the path $2014 \to 2010 \to 2016$ in the roadmap above. Recall that we started at using computational mathematics in 1991, and recently returned to it in 2016, more than 25 years have been passed. All the materials talked here are included in the survey article [11] (from which one may find more original references), except part 1 of the talk which has appeared in [12].

Sometimes, I feel disappointed since so much time have been spent on a single topic, I am worrying to be foolish. I tried several times to leave this area, but I came back, once a new idea appeared, i.e. the meaning of charming used at the title. Actually, I have been very lucky for the choice of this topic, so that I can continue my work for many years, learn much from the other branches of mathematics and make some contributions to them at last. This overview shows the importance of choosing a good research topic/direction, and also shows the globality of mathematics. At this moment, I recall that these two points are actually the main mathematical philosophy presented by D. Hilbert in his famous lecture given in 1900.

Acknowledgments. This paper is based on a series of talks, five of them given in 2015 were listed in [12]. The others are presented at Shandon University (2016/3), Xiamen University (2016/4), Brigham Young University (2016/4, USA), at the conference "Frontier Probability Days" (2016/5, U. of Utah, USA), at "The 8th Intern. Conf. on Stoch. Anal. Appl." (2016/6, Beijing), as a distinguished lecture at "The 10th Cross-Strait Conf. in Stat. & Probab." (abbrev. CSCSP) (2016/8, Chengdu), "The Sixth National Mathematical Culture Forum" (2016/8, Lanzhou), Zhejiang University (2016/9), Henan University (2016/9), Southwest Jiaotong University (2016/11), Center of StatSci, Peking University (2017/3), and Institute of

Mathematics, Academia Sinica (2017/3). The author acknowledge Professors Shi-Ge Peng, Zeng-Jing Chen, Ya-Nan Lin, Huo-Xiong Wu, Ke-Ning Lu, D. Khoshnevisan, E.C. Waymire, Zhen-Qing Chen, Zhi-Ming Ma, the organization committee and the local one for CSCSP, Jia-An Yan, Hu-Sheng Qiao, Gang Bao, Shu-Xia Feng, Ling-Di Wang, Wei-Ping Li, Shang-Yun Chen, Song-Xi Chen, Chii-Ruey Hwang, Tzuu-Shuh Chiang, Yunshyon Chow, Shuenn-Jyi Sheu, and Yuh-Jia Lee for their invitations and hospitality. Research supported in part by National Natural Science Foundation of China (No. 11131003 and 11626245) the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

参考文献

[1] Chen, M.F. (1991). *Exponential L^2-convergence and L^2-spectral gap for Markov processes*. Acta Math. Sin. New Ser. 7(1): 19–37.

[2] Chen, M.F. (2004). *From Markov Chains to Non-Equilibrium Particle Systems*. World Scientific, Singapore, 2nd Ed. (1st Ed., 1992).

[3] Chen, M.F. (2005). *Eigenvalues, Inequalities, and Ergodic Theory*. Springer

[4] Chen, M.F. (2008). *Spectral gap and logarithmic Sobolev constant for continuous spin systems*. Acta Math. Sin. New Ser. 24(5): 705–736.

[5] Chen, M.F. (2010). Speed of stability for birth–death processes. Front. Math. China 5(3), 379–515.

[6] Chen, M.F. (2011/2012). *Basic estimates of stability rate for one-dimensional diffusions*. Chapter 6 in "Probability Approximations and Beyond": 75–99. Lecture Notes in Statistics 205, eds: A. Barbour, H.P. Chan, D. Siegmund.

[7] Chen, M.F. (2013a). *Bilateral Hardy-type inequalities*. Acta Math Sin Eng Ser. 29(1): 1–32.

[8] Chen, M.F. (2013b). *Bilateral Hardy-type inequalities and application to geometry* (in Chinese) Mathmedia Vol. 37 No. 2, 12-32; Bulleton of Math Vol. 52, No. 8/9.

[9] Chen, M.F. (2014). *Criteria for discrete spectrum of 1D operators*. Commu. Math. Stat. 2, 279–309.

[10] Chen, M.F. (2015). *Progress on Hardy-type inequalities* in "Festschrift Masatoshi Fukushima", 131–142. World Sci.

[11] Chen, M.F. (2016a). *Unified speed estimation of various stabilities*. Chin. J. Appl. Probab. Statis. 32(1), 1–22.

[12] Chen, M.F. (2016b). *Efficient initials for computing the maximal eigenpair*. Front. Math. China 11(6): 1379–1418. A package based on the paper is available on CRAN now. One may check it through the link:
https://cran.r-project.org/web/packages/EfficientMaxEigenpair/index.html

[13] Chen, M.F. (2017a). *Efficient algorithm for principal eigenpair of discrete p-Laplacian*. Preprint.

[14] Chen, M.F. (2017b). *Global algorithms for maximal eigenpair*. Preprint.

[15] Chen, M.F. and Zhang, X. (2014). *Isospectral operators*. Commu Math Stat 2, 17–32.

[16] Chen, M.F. and Zhang, Y.H. (2014). *Unified representation of formulas for single birth processes*. Front. Math. China 9(4), 761–796.

[17] Liao, Z.W. (2016). *Variational formulas and basic estimates of the optimal constant in Hardy-type inequalities*. Doctorial Thesis at Beijing Normal Univ.

Mu-Fa Chen
School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.
E-mail: mfchen@bnu.edu.cn
Home page: http://math0.bnu.edu.cn/~chenmf/main_eng.htm

迷人的最大特征对子

陈木法
(北京师范大学，北京，100875)

摘要： “主导特征对子”在不同场合有不同名称. 在矩阵论中称为最大特征对子(最大特征值及其对应的特征向量). 本文首先介绍计算矩阵最大特征对子的十分意外的新结果. 主要贡献是选取一个孰知算法的高效初值. 其想法来源于我们新近关于主导特征值估计的研究. 在第二部分里，我们介绍很幸运得到的关于主导特征值的统一估计. 在第三部分里，我们通过一个特别例子说明此项研究的源头. 最后概述我们关于引领特征值估计和更一般的各种稳定性速度研究的漫长历程.

关键词： 主导特征对子；高效初值；三对角阵；速度估计；Hardy (Poincaré) 型不等式；φ^4 模型.

Front. Math. China 2017, 12(5): 1023–1043.

Global algorithms for maximal eigenpair

Mu-Fa Chen

Assisted by Yue-Shuang Li

(Beijing Normal University)

April 29, 2017

Abstract

This paper is a continuation of our previous paper [Front. Math. China, 2016, 11(6): 1379–1418] where an efficient algorithm for computing the maximal eigenpair was introduced first for tridiagonal matrices and then extended to the irreducible matrices with nonnegative off-diagonal elements. This paper introduces mainly two global algorithms for computing the maximal eigenpair in a rather general setup, including even a class of real (with some negative off-diagonal elements) or complex matrices.

2000 *Mathematics Subject Classification*: 15A18, 65F15, 93E15, 60J27

Key words and phrases. Maximal eigenpair, shifted inverse iteration, global algorithm.

1 Introduction

To compute the maximal eigenpair of the tridiagonal matrices with positive sub-diagonal elements, an efficient algorithm was introduced in [5; §3]. In the tridiagonal case, the construction of the initials for the algorithm is explicit. In some sense, the results are more or less complete (a modified algorithm, Algorithm 17, is included in §4.4). Next, the algorithm was extended to the general case in [5; §4] which is still efficient for tridiagonally dominant matrices. Note that the initial v_0 constructed in [5; §4.2] may not be efficient enough, since the shape of the maximal eigenvector can be rather arbitrary, could be quite far away from v_0 constructed in [5; §4.2]. Thus, we are worrying about the efficiency of the extended algorithm and moreover a global algorithm is still missed in our general setup. This is the aim of this paper. In §3, a part of the off-diagonal elements of the matrices are allowed to be negative. We can even handle with some complex matrices. Let us concentrate on the nonnegative matrices from now on, unless otherwise is stated.

By a shift if necessary, unless otherwise stated, we assume that the given matrix $A = (a_{ij} : 0 \leqslant i, j \leqslant N)$ is irreducible and nonnegative: $a_{ij} \geqslant 0$. We now state our algorithms. To guarantee the convergence of the iterations in

the paper, we assume that the matrix is irreducible having positive trace, or equivalently,

$$A^n > 0 \qquad \text{for each } n \geqslant \text{ some } n_0. \tag{1}$$

We mention that in the present nonnegative case, the condition having positive trace is not serious, otherwise, simply adopt a shift as mentioned at the beginning of [5].

In what follows, we omit, without mention time by time, the trivial case that $\sum_j a_{ij} \equiv$ constant $m > 0$. Since then the maximal eigenpair of A becomes $(m, \mathbb{1})$, where $\mathbb{1}$ is the constant function having components 1 everywhere.

Recall that the choice of the initials is quite essential for the Rayleigh Quotient Iteration (RQI), a special shifted inverse iteration. In general, it seems no hope at the moment to have such explicit analytic formulas as used in [5; §3]. Instead, as suggested in many textbooks, one may use other approach to obtain in a numerical way the required initials, say use the power iteration for instance. The last approach is safe, but rather slow as shown at the beginning of [5]. This leads us to come back to the shifted inverse iterations which is a fast cubic algorithm. The ratio of the numbers of iterations for these two algorithms can be thousands. Throughout this paper, we use varying shifts rather than a fixed one only. Now, a critical point is to avoid the dangerous pitfalls, i.e., the region $(0, \rho(A))$, where $\rho(A)$ is the maximal eigenvalue of A. The answer is given in part (1) of the next two algorithms. At the moment, we are interested in the generality and safety, do not take care much about the convergence speed, perhaps, maybe some price we have to pay here. We will see soon what happen in the next section.

Algorithm 1 (Specific Rayleigh quotient iteration) Let $A = (a_{ij})$ be given.

(1) Define column vectors

$$w^{(0)} = (1, 1, \ldots, 1)^*, \qquad v^{(0)} = w^{(0)}/\sqrt{N+1},$$

where w^* is the transpose of w, and set

$$z^{(0)} = \max_{0 \leqslant i \leqslant N} \left(Aw^{(0)}\right)_i.$$

(2) For given $v := v^{(n-1)}$ and $z := z^{(n-1)}$, let $w := w^{(n)}$ solve the equation

$$(zI - A)w = v. \tag{2}$$

As in step (1), define $v^{(n)} = w/\sqrt{w^*w}$. Next, define

$$x^{(n)} = \min_{0 \leqslant j \leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}}, \quad y^{(n)} = \max_{0 \leqslant j \leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}}, \quad z^{(n)} = {v^{(n)}}^* A v^{(n)}.$$

(3) If at some $n \geqslant 1$, $y^{(n)} - x^{(n)} < 10^{-6}$ (or $|z^{(n)} - z^{(n+1)}| < 10^{-6}$)(say!), then stop the computation. At the same time, regard $\left(z^{(n)}, v^{(n)}\right)$ as an approximation of the maximal eigenpair.

The algorithm was presented in [5; §4.1: Choice I]. The simplest choice v_0 is reasonable in the sense that it enables us to cover the general case. We did not pay enough attention on this algorithm since it looks less efficient. However, as some examples will be illustrated below, this algorithm is actually rather powerful. It is the place to state the main new algorithm of the paper.

Algorithm 2 (Shifted inverse iteration) Everything is the same as in Algorithm 1, except $y^{(n)}$ and $z^{(n)}$ defined in parts (2) and (3) there are exchanged. Moreover, the resulting $z^{(n)}$ (resp., $x^{(n)}$) is decreasing (resp., increasing) in n.

Let us repeat the sequences $z^{(n)}$, $y^{(n)}$ and $x^{(n)}$ defined in Algorithm 2:

$$x^{(n)} = \min_{0 \leqslant j \leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}}, \quad y^{(n)} = {v^{(n)}}^* A v^{(n)}, \quad z^{(n)} = \max_{0 \leqslant j \leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}}.$$

It is obvious that

$$x^{(n)} \leqslant y^{(n)} \leqslant z^{(n)}.$$

In general, Algorithm 1 is often a little effective than Algorithm 2, saving one iteration for instance, but in Algorithm 2, each iteration is safe, never failed into the pitfall. This is based on the following dual variational formula.

Proposition 3 [11; Theorem (8)] For a nonnegative irreducible matrix A, the Collatz–Wielandt formula holds:

$$\sup_{x>0} \min_{i \in E} \frac{(Ax)_i}{x_i} = \rho(A) = \inf_{x>0} \max_{i \in E} \frac{(Ax)_i}{x_i}.$$

Here and in what follows, unless otherwise stated, set $E = \{0, 1, \ldots, N\}$.

Actually, suppose that we have $w^{(n-1)} > 0$ in Algorithm 2. Then by Proposition 3 and step (2) of Algorithm 2, we have $z^{(n-1)} > \rho(A)$ and then the solution $w^{(n)}$ to the equation (2) should be positive: $w^{(n)} > 0$. Otherwise, if $z^{(n-1)} < \rho(A)$, then the solution $w^{(n)}$ is negative. This is the main reason why we choose such a $z^{(n-1)}$ for each $n \geqslant 1$ in Algorithm 2 and in the case of $n = 0$ in Algorithm 1 as our shift, avoiding the change of signs. Note that in Algorithm 1 we adopt $y^{(n)}$ (recall that at the moment, we use the notation given below Algorithm 2) at each step $n \geqslant 1$, hence the solution $w^{(n)}$ changes its sign often. This seems dangerous because $y^{(n)}$ is located in the dangerous region, but up to now, we have not meet serious trouble. Therefore, it is still regarded as one of our two main algorithms. Nevertheless, for large scale matrices, we will introduce a modification of Algorithm 1 in the next section.

A careful comparison of Algorithm 1 and the powerful one introduced in [5; §3] is delayed to the Appendix.

An easier way to see the efficiency of Algorithms 1 and 2 is comparing them with the one given in [5; §4.2]. Suppose that we have used three iterations in computing a model using the method introduced in [5; §4.2], this means on the one hand we have solved the linear equations in three times. On the other hand, we have solved three more times in advance to figure out the initials $v^{(0)}$ and $z^{(0)}$ in terms of the triple (ψ, h, μ). Altogether, we have solved six linear equations. Or in other words, we have used 6 iterations in the computation for the specific model. Thus, Algorithms 1 and 2 should be regarded as efficient one if no more than 6 iterations are used in the computation for the same model. As we will see soon, we are actually in such a successful situation.

To conclude this section, we rewrite Algorithms 1 and 2 to a class of matrices with nonnegative off-diagonal elements and negative diagonal elements: $Q = (q_{ij})$:

$$q_{ij} \geqslant 0, \quad i \neq j; \qquad \sum_{j=0}^{N} q_{ij} \leqslant 0, \quad 0 \leqslant i \leqslant N.$$

In this case, we are studying the maximal eigenpair of Q, or alternatively, the minimal eigenpair of $-Q$. To which, the next two algorithms are devoted.

Again, the trivial case that $\sum_{j=0}^{N} q_{ij}$ equals a constant is ignored throughout the paper.

Algorithm 4 (Specific Rayleigh quotient iteration) Let $Q = (q_{ij})$ be given.

(1) Define column vectors

$$w^{(0)} = (1, 1, \ldots, 1)^*, \qquad v^{(0)} = w^{(0)}/\sqrt{N+1},$$

and set $z^{(0)} = 0$.

(2) For given $v := v^{(n-1)}$ and $z := z^{(n-1)}$, let $w := w^{(n)}$ solve the equation

$$(-Q - zI)w = v. \tag{3}$$

As in step (1), define $v^{(n)} = w/\sqrt{w^*w}$. Next, define

$$x^{(n)} = \min_{0 \leqslant j \leqslant N} \frac{((-Q)w^{(n)})_j}{w_j^{(n)}}, \; y^{(n)} = \max_{0 \leqslant j \leqslant N} \frac{((-Q)w^{(n)})_j}{w_j^{(n)}}, \; z^{(n)} = v^{(n)*}(-Q)v^{(n)}.$$

(3) If at some $n \geqslant 1$, $y^{(n)} - x^{(n)} < 10^{-6}$ (or $|z^{(n)} - z^{(n+1)}| < 10^{-6}$)(say!), then stop the computation. At the same time, regard $\left(z^{(n)}, v^{(n)}\right)$ as an approximation of the minimal eigenpair.

Algorithm 5 (Shifted inverse iteration) Everything is the same as in Algorithm 4, except $x^{(n)}$ and $z^{(n)}$ defined in parts (2) and (3) there are exchanged. Moreover, the resulting $z^{(n)}$ (resp., $x^{(n)}$) is increasing (resp., decreasing) in n.

Algorithms 4 and 5 are based on [5; Corollary 12], a corollary of Proposition 3.

2 Examples

To illustrate the power of the algorithms introduced in the last section, we examine some typical examples in this section.

To go to practical computation for concrete models, our readers are urged to prepare enough patience, one may have a large number of iterations since the initials given in part (1) are quite rough.

The efficient application of Algorithm 1 was illustrated by [5; Examples 13–16]. To have a concrete comparison of the present algorithms with the one introduced in [5; §4.2], let us consider a simple example.

Example 6 [5; Example 21] Let

$$Q = \begin{pmatrix} -3 & 2 & 0 & 1 & 0 \\ 4 & -7 & 3 & 0 & 0 \\ 0 & 5 & -5 & 0 & 0 \\ 10 & 0 & 0 & -16 & 6 \\ 0 & 0 & 0 & 11 & -11 - b_4 \end{pmatrix}.$$

Corresponding to different b_4, the minimal eigenvalue λ_0 of $-Q$ and its approximation are shown in Tables 1 and 2, while the outputs by the algorithm given in [5; §4.2] are shown in Table 3. Here and in what follows, we stop at $z^{(2)}$ once the outputs $z^{(k)} = z^{(2)}$ for every $k \geqslant 2$.

Table 1. The outputs by Algorithm 1

b_4	$z^{(1)}$	$z^{(2)}$	$z^{(3)} = \lambda_{\min}(-Q)$
10^{-2}	0.000278773	0.000278686$= \lambda_{\min}(-Q)$	
10^{0}	0.0251531	0.0245175	
10^{2}	0.191729	0.182822	0.182819
10^{4}	0.201695	0.195019	0.195015

Table 2. The outputs by Algorithm 2

b_4	$z^{(1)}$	$z^{(2)}$	$z^{(3)} = \lambda_{\min}(-Q)$
10^{-2}	0.000278637	$0.000278686 = \lambda_{\min}(-Q)$	
10^{0}	0.0241546	0.0245175	
10^{2}	0.168776	0.18275	0.182819
10^{4}	0.179525	0.194932	0.195015

Table 3. The outputs by the algorithm given in [5]

b_4	$z^{(1)}$	$z^{(2)}$	$z^{(3)} = \lambda_{\min}(-Q)$
10^{-2}	0.000278573	0.000278686$= \lambda_{\min}(-Q)$	
10^{0}	0.0236258	0.0245174	0.0245175
10^{2}	0.200058	0.182609	0.182819

These tables show that the three algorithms are more or less at the same level of effectiveness. However, the first two are actually more economic since the last one requires an extra work computing the initial v_0.

Comparing [5; Example 15] with the corrected version of [5; Example 20] and its improvements given in [5; Tables 11, 12] (see the author's homepage), we see that the extended algorithm introduced in [5; §4.2] can be less efficient than Algorithm 1, it has some limitation for general non-symmetrizable (non-symmetric) matrices. We call a matrix $A = (a_{ji})$ is symmetrizable, if there exists a positive measure (μ_i) such that

$$\mu_i a_{ij} = \mu_j a_{ji}, \qquad i \neq j.$$

A simple necessary condition for the symmetrizability is

$$a_{ij} > 0 \Longleftrightarrow a_{ji} > 0, \qquad i \neq j.$$

Refer to [3; Chapter 7] and references within for the solution to the symmetrizability problem.

Let us start at a class of non-symmetrizable matrices which are taken from the so-called single birth Q-matrix (cf. [3] and references within). Define

$$Q = \begin{pmatrix} -1 & 1 & & & & \\ a_1 & -a_1-2 & 2 & & \mathbf{0} & \\ a_2 & & -a_2-3 & 3 & & \\ \vdots & & & \ddots & \ddots & \\ a_{N-1} & & \mathbf{0} & & -a_{N-1}-N & N \\ a_N & & & & & -a_N-N-1 \end{pmatrix}. \tag{4}$$

For this matrix, we have computed several cases:

$$a_k = 1/(k+1), \;\; a_k \equiv 1, \;\; a_k = k, \;\; a_k = k^2.$$

Among them, the first one is hardest and is hence presented below.

Example 7 Let Q be defined by (4). For different N, the outputs of Algorithm 5 (equivalently, Algorithm 2) are shown in Table 4.

Table 4. The outputs for different N by Algorithm 5

$N+1$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$	$z^{(5)}$	$z^{(6)}$
8	0.276727	0.427307	0.451902	0.452339		
16	0.222132	0.367827	0.399959	0.400910		
32	0.187826	0.329646	0.370364	0.372308	0.372311	
50	0.171657	0.311197	0.357814	0.360776	0.360784	
100	0.152106	0.287996	0.343847	0.349166	0.349197	
500	0.121403	0.247450	0.321751	0.336811	0.337186	
1000	0.111879	0.233257	0.313274	0.334155	0.335009	0.335010
5000	0.0947429	0.205212	0.293025	0.328961	0.332609	0.332635
10^4	0.0888963	0.194859	0.284064	0.326285	0.332113	0.332188

The last line of Table 4 shows that when $N = 10^4$,

$$\lambda_{\min}(-Q) \approx 0.332188.$$

If we use the shifted matrix $A = Q + mI$, then $\rho(A) \approx 9999.67$. From which, we get

$$\lambda_{\min}(-Q) \approx 10^4 + 10^{-4} - 9999.67.$$

Clearly, the second approach has a less precise output. That is the main difference between Algorithms 1, 2 and 4, 5, even though they are equivalent analytically.

It should be meaningful to have a comparison of the present results with those produced by [5; §4.2]. The outputs listed in Table 5 come from the algorithm without using δ_1 defined in that section. For the outputs using δ_1, one more iteration is needed for those N from 16 to 100 listed in the table.

Table 5. The outputs for different N by the algorithm given in [5; §4.2]

$N+1$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.450694	0.452338	0.452339
16	0.399520	0.400910	
32	0.371433	0.372311	
64	0.355722	0.355940	
100	0.349501	0.349197	
500	0.340666	0.337185	0.337186
1000	0.340871	0.335003	0.335010
5000	0.347505	0.332536	0.332635
10^4	0.352643	0.331975	0.332188

Clearly, the general algorithm introduced in [5; §4.2] is efficient for this non-symmetrizable model. We have seen that the present algorithms require more iterations than the earlier one, this is reasonable since the computations of the initials are excluded from Table 5. Actually, the computations of Table 5 cost double time than the previous one.

We now turn to Algorithm 4 which is often fast than Algorithm 5. The number of iterations by Algorithm 4 to this example is given in the first line of Table 6. It follows that the only case which is slower than Algorithm 5 given in Table 4 is the one: $N + 1 = 10000$. This leads a modification of Algorithm 4 as follows.

Algorithm 4_2 At the step $k \geqslant 3$, keep z_k to be the Rayleigh quotient as defined in Algorithm 4; for $k = 0, 1, 2$, choose z_k to be the same as those defined in Algorithm 5.

Table 6. Outputs of Agorithms 4 and 4_2

$N+1$	8	16	32	50	100	500	1000	5000	10000
Algorithm 4	4	4	4	4	4	5	5	6	7
Algorithm 4_2					5	5	5	5	5

Clearly, this algorithm becomes more essential for larger scale matrices, $N \geqslant 500$ for instance. The number of iterations using Algorithm 4_2 starting from $N+1=100$ is given in the second line of Table 6. According to the definition of the modified algorithm, one may relabel the original Algorithm 4 as 4_0, and then we can define Algorithm 4_m, as a mixture of Algorithms 5 and 4. Here we restricted on $m=2$ is based on [5; §4.2] where the initials are computed in three steps. In parallel, we can define Algorithm 1_m. A good way to make a threshold for m goes as follows. Once the components of $v^{(m+1)}$ have different signs, we replace the original m by $m+1$ and do the $(m+1)$th iteration again. However, it is unbelievable to use an m larger than 10. Note that Algorithms 4_m and 5 (similarly, Algorithms 1_m and 2) are suitable for unstructured matrices. However, Example 7 is structured. Hence, we have another way to speed up the convergence: the convex combination. Because $\lambda_{\min}(-Q) \in \big(0, v^{(0)*}(-Q)v^{(0)}\big)$, we choose the convex combination

$$z_0 = \xi\, v^{(0)*}(-Q)v^{(0)} + (1-\xi)\cdot 0 = \xi\, v^{(0)*}(-Q)v^{(0)}$$

for some $\xi \in (0,1)$. To determine ξ, computing $\lambda_{\min}(-Q)$ in the specific cases $(N+1)=8,16,32$, finding out a quadratic approximation in variable $1/x$ of the minimal eigenvalue in $(N+1)$ and then an approximation of $\lambda_{\min}(-Q)$ for $N+1=10000$. The resulting estimate, may be a little smaller, can be used as the required $\xi\,(=0.34189)$. The reason for this choice is as follows: the eigenvalue $\lambda_{\min}(-Q)$ is decreasing in $N+1$, and a smaller z_0 is safer in our iterations. The number of the iterations of the convex combination of Algorithm 4 is given at the first line of Table 7.

Table 7. Convex combination for Algorithms 4 and 5

$N+1$	8	16	32	50	100	500	1000	5000	10000
Alg-4, $\xi=0.34189$	3	3	3	3	3	3	3	3	3
Algorithm 4, $\xi=0.23$	3	3	3	3	3	3	3	4	4
Algorithm 4, $\xi=0.3$	3	3	3	3	3	3	3	3	4
Algorithm 5, $\xi=0.23$	4	4	4	4	4	4	5	5	5
Algorithm 5, $\xi=0.3$	4	4	4	4	4	4	4	4	4

Unfortunately, for large N, this $\xi=0.34189$ does not work for Algorithm 5 since $\xi > \lambda_0 = \lambda_0(N)$ when $N \geqslant 500$ and then $\min_{0\leqslant j\leqslant N}(-Qv^{(1)})_j/v_j^{(1)} < 0$ which means that Algorithm 5 is not meaningful. This becomes more clear if we lift Q to a nonnegative A and then examine the proof of Proposition 16, where the condition $z > \rho(A)$ is used to quarantee the convergence of the

iterations. Thus, a common choice for these two algorithms could be a little smaller $\xi = 0.3$. For which, the outputs are given in Table 7. If one is lazy to compute the quadratic approximation of $\lambda_0(N)$, one may compute only one smaller N, say $N = 7$. At which, the best choice is $\xi \approx 0.452$. Thus, to cover every $N < 10^4$, we may choose $\xi \approx 0.45/2 \approx 0.23$ (the bisection method). Again, the outputs are given in Table 7.

The long analysis on Example 7 not only shows the power of our algorithms, but also indicates a big room for the improvements.

The next example is motivated from the classical branching process. Denote by $(p_k : k \geqslant 0)$ a given probability measure with $p_1 = 0$. Let

$$Q = \begin{pmatrix} -1 & p_2 & p_3 & p_4 & \cdots & p_{N-1} & \sum_{k\geqslant N} p_k \\ 2p_0 & -2 & 2p_2 & 2p_3 & \cdots & 2p_{N-2} & 2\sum_{k\geqslant N-1} p_k \\ & 3p_0 & -3 & 3p_2 & \cdots & 3p_{N-3} & 3\sum_{k\geqslant N-2} p_k \\ & & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & \ddots & \vdots & \vdots \\ & & 0 & & \ddots & -(N-1) & (N-1)\sum_{k\geqslant 2} p_k \\ & & & & & Np_0 & -Np_0 \end{pmatrix},$$

In the original model, the state 0 is an absorbing one. Here we regard it as a killing boundary. Hence it is ruled out from our state space. Thus, the matrix is defined on $E := \{1, 2, \ldots, N\}$. Set $M_1 = \sum_{k\in E} kp_k$. When $N = \infty$, in the subcritical case that $M_1 < 1$, with a little modification at 0, it is known that the process generated by Q is ergodic, and is indeed exponentially ergodic (cf. [8; Theorem 1.4 (iii)]). Hence the exponential convergence rate should be positive. Otherwise, the process is not ergodic and so the convergence rate should be zero.

From now on, fix

$$p_0 = \frac{\alpha}{2},\ p_1 = 0,\ p_k = \frac{2-\alpha}{2^k}\ (k = 2, 3, \cdots), \qquad \alpha \in (0, 2).$$

Then $M_1 = 3(2-\alpha)/2$ and hence we are in the subcritical case iff $\alpha \in (4/3, 2)$.

Example 8 Set $\alpha = 1$. Then the outputs of the approximation for the minimal eigenvalue of $-Q$ by Algorithm 2 (or 5) are shown in Table 8.

Table 8. The outputs in the supercritical case

N	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.0311491	0.0346044	0.0346310
16	0.00256281	0.00260088	

When $N \geqslant 50$, $z^{(1)} < 10^{-6}$. Hence, $z^{(n)}$ decays quite quick to zero when $N \to \infty$ (for $n \geqslant 2$). This is reasonable since we are now away from the subcritical region.

Example 9 Set $\alpha = 7/4$. We are now in the subcritical case and so the maximal eigenvalue should be positive. We want to know how fast the local maximal eigenvalue becomes stable (i.e., close enough to the converge rate at $N = \infty$). The numbers of iterations of Algorithms 4 and 5 are given in Table 9.

Table 9. Number of iterations of Algorithms 4 and 5

N	8	16	50	100	500	1000	5000	10000
Algorithm 4	5	6	7	7	8	8	9	10
Algorithm 5	6	6	7	7	8	9	9	10

Next, with the convex combination

$$z^{(0)} = \xi(v^{(0)})^*(-Q)v^{(0)} + (1-\xi)\max_{0\leqslant j\leqslant N}(-Qw^{(0)})_j.$$

In view of the practice on $N = 8$, we make the choice that $\xi = 0.31$. Then we obtain Table 10.

Table 10. Number of iterations of the Algorithms with convex combination

N	8	16	50	100	500	1000	5000	10000
Algorithm 4	2	3	4	4	4	4	4	4
Algorithm 5	3	3	4	4	4	4	4	4

In particular, the outputs of Algorithm 4 with convex combination is given in detail in Table 11.

Table 11. The outputs in the subcritical case

N	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$
8	0.637800	0.638153		
16	0.621430	0.625490	0.625539	
50	0.609976	0.624052	0.624997	0.625000
100	0.606948	0.623377	0.624991	0.625000
500	0.604409	0.622116	0.624962	0.625000
1000	0.604082	0.621688	0.624944	0.625000
5000	0.603817	0.620838	0.62489	0.625000
10^4	0.603784	0.620511	0.624861	0.625000

From the above table, we see that for N varies from 8 to 10^4, in each case, we need at most 4 iterations only. The computation in each case costs no more than one minute. Besides, starting from $N = 50$, the final outputs are all the same: 0.625, which then can be regarded as a very good approximation of $\lambda_{\min}(-Q)$ at infinity $N = \infty$. Since the convergence of this model becomes stable for small $N \leqslant 50$, the computations become much simpler than the previous one, we use neither Algorithm 4_2 nor the quadratic fit.

Hopefully, we have already shown the power of our algorithms.

3 A class of real or complex matrices

This section is out of the scope of [5] which depends heavily on probabilistic idea. Thanks are given to the extended Perron–Frobenius theory ([10–12]) which makes this section possible.

First, we consider the real case. The special case that all off-diagonal elements of A are negative has been treated above, using $-Q$ instead of A here. Thus, we are now mainly interested in the case that a part of the off-diagonal elements are negative. Again, we are concentrated in the study of the maximal eigenpair.

Proposition 10 Let A be a real matrix. By a shift of A if necessary, assume that (1) holds. Then Algorithms 1 and 2 are available.

Proof. By [10; Theorem 2.2], condition (1) implies that the matrix A possesses the strong Perron–Frobenius property. Hence it has the maximal eigenvalue $\rho(A)$ which is simple, positive and corresponds to a positive eigenvector. Besides, by [10; Theorem 2.6], the Collatz–Wielandt formula given in Proposition 3 holds. These facts are enough to use Algorithms 1 and 2. □

The next simple observation is helpful.

Lemma 11 Condition (1) holds iff

$$A^k > 0 \qquad \text{for } k = n_0, n_0 + 1, \ldots, 2n_0 - 1.$$

Proof. Given $n \geqslant n_0$, write

$$n = rn_0 + s$$

for some integer $r \geqslant 1$ and $s = 0, 1, \ldots, n_0 - 1$. If $r = 1$, then the conclusion holds by assumption. Otherwise, let $r \geqslant 2$. Then express

$$n = (r - 1)n_0 + (n_0 + s).$$

It follows that

$$A^n = \left(A^{n_0}\right)^{r-1} A^{n_0+s} > 0$$

as required. □

We now illustrate our algorithms by a simple example.

Example 12 [11; Example (7)] Let

$$A = \begin{pmatrix} -1 & 8 & -1 \\ 8 & 8 & 8 \\ -1 & 8 & 8 \end{pmatrix}.$$

Then

$$A^2 = \begin{pmatrix} 66 & 48 & 57 \\ 48 & 192 & 120 \\ 57 & 120 & 129 \end{pmatrix} > 0, \qquad A^3 = \begin{pmatrix} 261 & 1368 & 774 \\ 1368 & 2880 & 2448 \\ 774 & 2448 & 1935 \end{pmatrix} > 0.$$

By Lemma 11, condition (1) holds with $n_0 = 2$. The eigenvalues of A are as follows.

$$17.5124, \quad -7.4675, \quad 4.95513.$$

The corresponding maximal eigenvector is

$$(0.486078,\ 1.24981,\ 1)^*$$

which is positive.

Outputs of our algorithms are shown in Table 12. Both algorithms are started at $z^{(0)} = 24$.

Table 12. The outputs for a matrix with more negative elements

n	$z^{(n)}$: Algorithm 1	$z^{(n)}$: Algorithm 2
1	17.3772	18.5316
2	17.5124	17.5416
3		17.5124

Next, we turn to study the complex case. Instead of (1), we assume that

$$\mathrm{Re}\big(A^n\big) > 0 \qquad \text{for } n \geqslant \text{ some } n_0, \tag{5}$$

up to a shift mI of A. Certainly, as usual $\mathrm{Re}(A)$ means the real part of a complex matrix A. This condition is based on [12; Theorems 2.3 and 2.2], from which we know that A has the maximal, simple, positive eigenvalue. Then we have a weak extension of the Collatz–Wielandt formula as follows.

Proposition 13 [12; Theorems 2.3 and 2.4] Let $A^k \neq 0$ for each $k \geqslant 1$ and $\mathrm{Re}(A^n) \geqslant 0$ for every large enough n. Then we have for each $x > 0$

$$\min_{0 \leqslant j \leqslant N} \frac{(\mathrm{Re}(A)x)_j}{x_j} \leqslant \rho(A) \leqslant \max_{0 \leqslant j \leqslant N} \frac{(\mathrm{Re}(A)x)_j}{x_j}.$$

Since for the complex conjugate $\bar{x}^*$ of x, the quantity $\bar{x}^* Ax$ may still be complex, in view of this, Proposition 13 and the positivity of $\rho(A)$ by (5), it seems not reasonable to use $\bar{x}^* Ax/(\bar{x}^* x)$ as a shift. In this sense, we do not have a modified version of Algorithm 1. Fortunately, Algorithm 2 is still meaningful.

Algorithm 14 (Shifted inverse iteration) By a shift of A if necessary, assume that (5) holds.

(1) Define column vectors

$$w^{(0)} = (1, 1, \ldots, 1)^*, \qquad v^{(0)} = w^{(0)}/\sqrt{N+1},$$

and set

$$z^{(0)} = \max_{0 \leqslant i \leqslant N} \left(\mathrm{Re}(A) w^{(0)}\right)_i.$$

(2) For given $v := v^{(n-1)}$ and $z := z^{(n-1)}$, let $w := w^{(n)}$ solve the equation

$$(zI - A)w = v. \tag{6}$$

As in step (1), define $v^{(n)} = w/\sqrt{\overline{w}^* w}$. Next, define

$$z^{(n)} = \max_{0 \leqslant j \leqslant N} \frac{(\mathrm{Re}(A)\mathrm{Re}(w^{(n)}))_j}{\mathrm{Re}(w^{(n)})_j}, \qquad y^{(n)} = (\bar{v}^{(n)})^* A v^{(n)}.$$

(3) If at some $n \geqslant 1$, $|y^{(n+1)} - y^{(n)}| < 10^{-6}$ (say!), then stop the computation. At the same time, regard $\left(y^{(n)}, v^{(n)}\right)$ as an approximation of the maximal eigenpair.

Note that in Algorithm 14, the sequence $\left\{y^{(n)}\right\}_{n \geqslant 0}$, but not $\left\{z^{(n)}\right\}_{n \geqslant 0}$, converges to $\rho(A)$. To illustrate the use of the algorithm, we consider the following example.

Example 15 [12; Example 2.1] Let

$$A = \begin{pmatrix} 0.75 - 1.125\,i & 0.5882 - 0.1471\,i & 1.0735 + 1.4191\,i \\ -0.5 - i & 2.1765 + 0.7059\,i & 2.1471 - 0.4118\,i \\ 2.75 - 0.125\,i & 0.5882 - 0.1471\,i & -0.9265 + 0.4191\,i \end{pmatrix},$$

where the coefficients are all accurate, to four decimal digits. Then A has eigenvalues

$$3, \quad -2 - i, \quad 1 + i$$

with maximal eigenvector

$$(0.408237,\ \ 0.816507,\ \ 0.408237)^*.$$

The outputs of Algorithm 14 are shown in Table 13.

Table 13. The outputs for a complex matrix

$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
$3.03949 - 0.0451599\,i$	$3.00471 - 0.0015769\,i$	3

4 Appendix*

4.1 Proof of the last assertion in Algorithm 2

Proposition 16 The sequence

$$z^{(n)} = \max_{0\leqslant j\leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}} \quad \left(\text{resp., } x^{(n)} = \min_{0\leqslant j\leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}}\right)$$

defined in Algorithm 2 is decreasing (resp., increasing) in n.

Proof. Let $w > 0$ and define

$$\bar{\rho} = \max_{0\leqslant j\leqslant N} \frac{(Aw)_j}{w_j}.$$

Then $(Aw)_j \leqslant \bar{\rho} w_j$ for every j. That is,

$$(A_z w)_j \leqslant \bar{\rho}_z w_j \quad \forall j, \qquad A_z := A/z,\ \bar{\rho}_z = \bar{\rho}/z.$$

Since $A_z \geqslant 0$, it follows that

$$A\sum_{n=0}^{\infty} A_z^n w \leqslant A\left(w + \bar{\rho}_z \sum_{n=0}^{\infty} A_z^n w\right) \leqslant \bar{\rho} w + \sum_{n=1}^{\infty} \bar{\rho} A_z^n w = \bar{\rho} \sum_{n=0}^{\infty} A_z^n w.$$

This means that

$$A(I - A_z)^{-1} w \leqslant \bar{\rho}(I - A_z)^{-1} w$$

since $z > \rho(A)$ by assumption and then $\rho(A_z) < 1$. Hence

$$\max_{0\leqslant j\leqslant N} \frac{(A((I - A_z)^{-1} v))_j}{((I - A_z)^{-1} v)_j} \leqslant \bar{\rho}, \qquad v := w/\sqrt{w^* w}.$$

Regarding $w = w^{(n-1)}$ and $v = v^{(n-1)}$, this gives us

$$z^{(n)} = \max_{0\leqslant j\leqslant N} \frac{(Aw^{(n)})_j}{w_j^{(n)}} \leqslant \bar{\rho} = \max_{0\leqslant j\leqslant N} \frac{(Aw^{(n-1)})_j}{w_j^{(n-1)}} = z^{(n-1)}.$$

Here we have assumed that $z^{(n-1)} > \rho(A)$, otherwise, the computation should be finished at the step $n - 1$. We have thus proved the assertion on $z^{(n)}$. Dually, we have the assertion on $x^{(n)}$. □

*In the published version, the subsections and formulas in this section are relabelled as either A.# or A#. For instance, we have subsection A.1, Proposition A1, Algorithm A2, formula (A2) and so on.

4.2 Proof of the last assertion in Algorithm 5

.

Recall the sequence $\{z^{(n)}\}$ used in Algorithm 2 is given in Proposition 16. Denote by $\{\tilde{z}^{(n)}\}$ the corresponding one in Algorithm 5. Then, by the relation of Q and A used in Algorithm 5: $A = Q + mI$, where $m = \max_i \sum_j a_{ij}$. Hence

$$z^{(0)}I - A = -Q - (m - z^{(0)})I.$$

This means not only $\tilde{z}^{(0)} = 0$, but also

$$w^{(1)} = \big(z^{(0)}I - A\big)^{-1}v^{(0)} = \big(-Q - \tilde{z}^{(0)}I\big)^{-1}v^{(0)} =: \widetilde{w}^{(1)},$$

where $\tilde{w}^{(1)}$ is obtained by the first iteration of Algorithm 5. Furthermore, similar to the proof of [5; Corollary 12], we have

$$\tilde{z}^{(1)} = \min_i \frac{(-Q\widetilde{w}^{(1)})_i}{\widetilde{w}_i^{(1)}} = m - \max_i \frac{(Aw^{(1)})_i}{w_i^{(1)}} = m - z^{(1)}.$$

Recursively, we obtain the required assertion. □

4.3 Comparison of Algorithms 1 and 4 with the one given in [5; §3]

.

Since Algorithms 1 and 4 are equivalent, we need only to compare Algorithm 4 with the one given in [5; §3]. The main difference is their initial $(v^{(0)}, z^{(0)})$. Clearly, the initial $v^{(0)}$ used in [5; §3] is finer than the one used in Algorithm 4. Hence, we need only to compare their $z^{(0)}$.

Next, let $v := v^{(0)}$ be the initial vector used in [5; §3]. Denote by w be the solution of the ordinary inverse iteration (that is the first step of Algorithm 4 or equivalently, Algorithm 1):

$$-Qw = v.$$

Then

$$\frac{(-Qw)_j}{w_j} = \frac{v_j}{((-Q)^{-1}v)_j} = \mathit{\Pi}_j(v)^{-1}. \tag{7}$$

Here in the last equality of (7), we have used the first formula in the proof of [5; Proposition 23]. Hence

$$\inf_j \frac{(-Qw)_j}{w_j} = \inf_j \mathit{\Pi}_j(v)^{-1}. \tag{8}$$

The right-hand side of (8) is just δ_1^{-1} used in [5; §3] as its initial $z^{(0)}$. The left-hand side of (8) should be positive, due to the inverse iteration algorithm, it is certainly bigger than 0 used as the initial $z^{(0)}$ in Algorithm 4. In conclusion, both initials used in [5; §3] are better than those used in Algorithm 4. This completes the comparison of Algorithm 4 and the one given in [5; §3].

Naturally, this comparison leads to the next subsection.

4.4 Modification of the algorithm introduced in [5; §3]

Step 1. By a shift if necessary, we may assume that we are given a matrix Q having the form

$$\begin{pmatrix} -b_0-c_0 & b_0 & & & & \\ a_1 & -a_1-b_1-c_1 & b_1 & & 0 & \\ & a_2 & -a_2-b_2-c_2 & b_2 & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & & -a_{N-1}-b_{N-1}-c_{N-1} & b_{N-1} \\ & & & & a_N & -a_N-c_N \end{pmatrix},$$

where $a_i > 0,\ b_i > 0,\ c_i \geqslant 0$ but $c_i \not\equiv 0$. Note that the maximal eigenvalue of Q is shifted from the original one but the corresponding eigenvector remains the same.

Step 2. Following [5; §3], assume for a moment that some of $c_i\,(i=0,1,\ldots,N-1)$ is positive. Then, define

$$r_0 = 1 + \frac{c_0}{b_0},\ \ r_n = 1 + \frac{a_n + c_n}{b_n} - \frac{a_n}{b_n r_{n-1}}, \qquad 1 \leqslant n < N,$$

$$h_0 = 1,\ \ h_n = h_{n-1} r_{n-1} = \prod_{k=0}^{n-1} r_k, \qquad 1 \leqslant n \leqslant N,$$

and additionally,

$$h_{N+1} = c_N h_N + a_N(h_N - h_{N-1}).$$

We remark that in the special case that

$$c_0 = \cdots = c_{N-1} = 0,$$

by induction, it is easy to check that

$$r_0 = \cdots = r_{N-1} = 1$$

and hence

$$h_0 = \cdots = h_N = 1.$$

Furthermore, $h_{N+1} = c_N$. Thus, in this special case, we simply ignore the sequence $\{h_k\}$ but replace c_N by b_N. Note that here we use all of the three sequence (a_k), (b_k) and (c_k) given in Q but no extra thing. The role of the sequence $\{h_k\}$ is reducing the former case to the last special one and keep the same spectrum, in terms of the H-transform $\widetilde{Q}$:

$$\widetilde{Q} = \mathrm{Diag}(h_i)^{-1} Q\, \mathrm{Diag}(h_i). \tag{9}$$

The maximal eigenpair $(\rho(Q), g)$ is transformed to $\big(\rho(\widetilde{Q}) = \rho(Q), \mathrm{Diag}(h_i)^{-1} g\big)$.

Step 3. In view of Step 2 above, it suffices to consider the following matrix

$$Q = \begin{pmatrix} -b_0 & b_0 & & & \\ a_1 & -(a_1+b_1) & b_1 & \mathbf{0} & \\ & a_2 & -(a_2+b_2) & b_2 & \\ & \ddots & \ddots & \ddots & \ddots \\ & \mathbf{0} & \ddots & -(a_{N-1}+b_{N-1}) & b_{N-1} \\ & & & a_N & -(a_N+b_N) \end{pmatrix}, \tag{10}$$

where $a_i,\ b_i > 0$. This step is changed from the original, where everything we are working here is transfer into the original matrix Q rather than the simpler one here. It seems a direct treatment of the present matrix Q is slightly simpler.

Define the sequence (μ_i) as usual:

$$\mu_0 = 1,\ \ \mu_n = \mu_{n-1}\frac{b_{n-1}}{a_n} = \frac{b_0 b_1 \cdots b_{n-1}}{a_1 a_2 \cdots a_n}, \qquad 1 \leqslant n \leqslant N.$$

Next, define

$$\varphi_n = \sum_{k=n}^{N} \frac{1}{\mu_k b_k}, \qquad 0 \leqslant n \leqslant N. \tag{11}$$

and

$$\delta_1 = \max_{0\leqslant i\leqslant N}\left[\sqrt{\varphi_i}\sum_{j=0}^{i}\mu_j\sqrt{\varphi_j} + \frac{1}{\sqrt{\varphi_i}}\sum_{i+1\leqslant j\leqslant N}\mu_j\varphi_j^{3/2}\right]. \tag{12}$$

Having these preparations at hand, we can now start our iterations.

Step 4. As in [5; §3], choose

$$w^{(0)} = \sqrt{\varphi}, \quad v^{(0)} = w^{(0)}/\|w^{(0)}\|_{\mu,2}, \quad z^{(0)} = \delta_1^{-1}, \tag{13}$$

where $\|\cdot\|_{\mu,2}$ denotes the $L^2(\mu)$-norm. Note that here in the non-symmetric case, the use of the measure (μ_i) cannot be ignored since in this case, we are based on, δ_k for instance, the $L^2(\mu)$ setup.

Step 5. For given $v = v^{(n-1)}$ and $z = z^{(n-1)}$, let $w = w^{(n)}$ solve the linear equation

$$(-Q - zI)w = v \tag{14}$$

and then define $v^{(n)} = w/\|w\|_{\mu,2}$. An explicit solution of this w is now available, refer to [6; Algorithm 3].

Step 6. At the kth ($k \geqslant 1$) iteration, in addition to the one $(v^{(k)}, -Qv^{(k)})_\mu$ used in [5; §3], one may also adopt $z^{(k)} = \delta_k^{-1}$:

$$\delta_k = \max_{0\leqslant i\leqslant N}\frac{1}{v_i^{(k)}}\left[\varphi_i\sum_{j=0}^{i}\mu_j v_j^{(k)} + \sum_{i+1\leqslant j\leqslant N}\mu_j\varphi_j v_j^{(k)}\right]. \tag{15}$$

Certainly, when $k = 1$, the present δ_k reduces to (12). This is the main new point in the modified algorithm. Since [4; Theorems 2.4 (3), 3.2 (1) and (3.6)], we have

$$\delta_k^{-1} \leqslant \lambda_{\min}(-Q) \leqslant (v^{(n)}, -Qv^{(n)})_\mu \text{ for each } k \text{ and } n.$$

By [5; Proposition 23] and [4; Theorem 3.2 (1)], we have known that the sequence $\{\delta_k^{-1}\}$, deduced in the theorem just cited using the approximating eigenvectors obtained by the ordinary inverse iteration (without shift), is increasing to $\lambda_{\min}(-Q)$. It should be clear that the present sequence $\{\delta_k^{-1}\}$ produced by the advanced shifted inverse iteration should converge to $\lambda_{\min}(-Q)$ more faster. Thus the new $z^{(k)}$ $(k \geqslant 1)$ not only avoids the dangerous region but may also accelerate the convergence of the algorithm. Certainly, the computation of δ_k needs more work than the one of $(v^{(k)}, -Qv^{(k)})_\mu$.

The use of the quantity (15) is motivated from the remark above subsection 4.3. The formula (15) is a corollary of [4; Theorem 2.4 (3)] which depends on the form (10) of Q. For general Q as the one in Step 1, we do not have an analog of [4; Theorem 2.4 (3)], and so (15) is not applicable in such a general situation.

Step 7. To go back to the original matrix A, denote its maximal eigenpair by $(\rho(A), g)$. Recall that the matrix Q at the beginning is obtained from A by a shift: $Q = A - mI$, $m := \max_i \sum_j a_{ij}$. Let (z, v) be the output from the last iteration in Step 6. Then we have

$$\rho(A) \approx m - z, \qquad g \approx \mathrm{Diag}(h_i)v. \tag{16}$$

We now summery the above discussions as a modified algorithm.

Algorithm 17 For tridiagonal matrix, the Step 1–Step 7 above consist a modified algorithm of the one introduced in [5; §3].

We are now ready to study a randomly chosen example, introduced to the author by Tao Tang, to justify the power of our algorithms and also to compare their efficiency.

Example 18 Let

$$A = \begin{pmatrix} 2.334 & 0.9962 & & & & \\ 0.5142 & 2.6725 & 0.1111 & & 0 & \\ & 0.2115 & 2.263 & 0.1405 & & \\ & & 0.8442 & 2.8457 & 0.7595 & \\ & 0 & & 0.2347 & 2.2257 & 0.0781 \\ & & & & 0.9837 & 2.1582 \end{pmatrix}.$$

Then the eigenvalues of A are

$$3.26753,\ 3.16247,\ 2.40182,\ 2.12632,\ 1.80416,\ 1.73679.$$

The outputs of our algorithms are given in Table 14.

Table 14. Comparison of four algorithms

Algorithm	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$	$z^{(5)}$
Algorithm 1	3.30193	3.26737	3.26754	3.26753	
Algorithm 2	3.64033	3.32623	3.26937	3.26756	3.26753
Algorithm 17a	3.2618	3.26752	3.26753		
Algorithm 17b	3.27947	3.2685	3.26754	3.26753	

where the algorithms in the last two lines mean that

Algorithm 17a: take $z^{(k)} = (v^{(k)}, -Qv^{(k)})_\mu$ for each $k \geqslant 1$.

Algorithm 17b: take $z^{(k)} = \delta_k^{-1}$ defined by (15) for each $k \geqslant 1$.

Proof. To apply Algorithm 17, take $m = 4.4494$. Then $Q = A - mI$:

$$Q = \begin{pmatrix} -2.1154 & 0.9962 & & & & \\ 0.5142 & -1.7769 & 0.1111 & & 0 & \\ & 0.2115 & -2.1864 & 0.1405 & & \\ & & 0.8442 & -1.6037 & 0.7595 & \\ & 0 & & 0.2347 & -2.2237 & 0.0781 \\ & & & & 0.9837 & -2.2912 \end{pmatrix}.$$

We have $h = (2.12347,\ 29.3339,\ 453.284,\ 924.514,\ 24961)$. The H-transform of Q becomes

$$\widetilde{Q} = \begin{pmatrix} -2.1154 & 2.1154 & & & & \\ 0.242151 & -1.7769 & 1.53475 & & 0 & \\ & 0.0153104 & -2.1864 & 2.17109 & & \\ & & 0.0546316 & -1.6037 & 1.54907 & \\ & 0 & & 0.115072 & -2.2237 & 2.10863 \\ & & & & 0.0364346 & -2.2912 \end{pmatrix}.$$

Then we are ready to use Algorithm 17 for the maximal eigenpair of $\widetilde{Q}$ and finally return to the one for A by (16). $\square$

To explain the word "modified" in detail, we transfer Algorithm 17 to the one presented in [5; §3]. To do so, we keep the notation Q, μ, φ, δ_1 and so on used in [5; §3], but add superscript $\tilde{}$ to those notation used in Steps 3, 4 above. Let $\tilde{\mu} = h^2\mu$ (i.e., $\tilde{\mu}_i = h_i^2\mu_i$). Then, as mentioned in [5; §5], the mapping $f \to \tilde{f} := f/h$ gives us not only an isometry from $L^2(\mu)$ to $L^2(\tilde{\mu})$ $\big($i.e., $\|f\|_{\mu,2} = \|\tilde{f}\|_{\tilde{\mu},2}\big)$, and then also an isospectrum of Q on $L^2(\mu)$ and $\widetilde{Q}$ on $L^2(\tilde{\mu})$:

$$(f, Qf)_\mu = (\tilde{f}, \widetilde{Q}\tilde{f})_{\tilde{\mu}}, \qquad \|f\|_{\mu,2} = 1.$$

Now, from $L^2(\tilde{\mu})$ to $L^2(\mu)$, we have

$$\tilde{\varphi}_n = \sum_{k=n}^{N} \frac{1}{\tilde{\mu}_k \tilde{b}_k} \to \sum_{k=n}^{N} \frac{1}{h_k h_{k+1} \mu_k b_k} = \varphi_n, \qquad 0 \leqslant n \leqslant N.$$

Here the transform $\tilde{\mu}_k\tilde{b}_k \to h_k h_{k+1}\mu_k b_k$ for each $k \leqslant N-1$ is regular, except the last term in the sum $\big(\tilde{\mu}_N\tilde{b}_N\big)^{-1}$, where $\tilde{b}_N$ is actually the element $\tilde{c}_N$ which is obtained from the transform $Q \to \widetilde{Q}$, and h_{N+1} and b_N are specified in [5; §3] to make the unified expression in the second sum. We mention here that h_{N+1} is the original paper [5] should be replaced by

$$h_{N+1} = c_N h_N + a_N(h_N - h_{N-1})$$

since the sequence (c_i) used in [5] and [7] have different sign. Next,

$$\begin{aligned}\tilde{\delta}_1 &= \max_{0\leqslant n\leqslant N}\left[\sqrt{\tilde{\varphi}_n}\sum_{k=0}^{n}\tilde{\mu}_k\sqrt{\tilde{\varphi}_k} + \frac{1}{\sqrt{\tilde{\varphi}_n}}\sum_{n+1\leqslant j\leqslant N}\tilde{\mu}_j\tilde{\varphi}_j^{3/2}\right]\\ \to \delta_1 &= \max_{0\leqslant n\leqslant N}\left[\sqrt{\varphi_n}\sum_{k=0}^{n}\mu_k h_k^2\sqrt{\varphi_k} + \frac{1}{\sqrt{\varphi_n}}\sum_{n+1\leqslant j\leqslant N}\mu_j h_j^2\varphi_j^{3/2}\right].\end{aligned}$$

At the same time,

$$\begin{aligned}&\big(-\widetilde{Q} - \tilde{z}I\big)\tilde{w} = \tilde{v}\\ \Longleftrightarrow\ &\big(-\operatorname{Diag}(h)^{-1}Q\operatorname{Diag}(h) - \tilde{z}I\big)\tilde{w} = \tilde{v}\\ \Longleftrightarrow\ &(-Q - \tilde{z}I)\operatorname{Diag}(h)\tilde{w} = \operatorname{Diag}(h)\tilde{v}\\ \Longleftrightarrow\ &(-Q - zI)w = v.\end{aligned}$$

Here in the last line, $\tilde{z}$ is replaced by z, this is due to the isospectrum: an lower bound of the spectrum of $-\widetilde{Q}$ is also the one of $-Q$. The fact that $\operatorname{Diag}(h)\tilde{w} = w$ comes from the definition of our mapping $f \to \tilde{f}$. Finally, since the isometry, we have $\|w\|_{\mu,2} = \|\tilde{w}\|_{\tilde{\mu},2}$. We have thus deduced the algorithm presented in [5; §3] from the modified one.

4.5 Modification of the algorithm introduced in [5; §4.2]

In parallel to §4.4, we may introduce a modification of the algorithm presented in [5; §4.2]. The main idea is: once we obtain the function h, it can be ignored since we can use the general transform $\widetilde{Q}$ defined in (9) instead of the original Q to continue the procedure of the algorithm constructed in [5; §4.2]. Since this modification is only a mimic of the one for tridiagonal matrix (§4.4), something may be lost. For instance, the sequence $\{\delta_k^{-1}\}$ formally defined by (15) may no longer be the lower bound of $\lambda_{\min}(-Q)$, one has to take care in practice.

To conclude this paper, we remark some possible extension of the algorithms given here to a more general setup. For a larger class of Markov generators, the algorithms are meaningful. Actually, the Perron–Frobenius property as well as the the Collatz–Wielandt formula have been generalized by a number of authors. In particular, the part of the Collatz–Wielandt formula used in Algorithm 5 as $z^{(n)}$ was extended by [9; $\psi_2(V)$ in the Theorem].

See also [13; (1.1) i) and §2] and more recently, [1; Theorem 2.1]. Note the difference: we are working on $\lambda_{\min}(-L)$ here rather than $\lambda_{\max}(L)$ in the cited papers.

In the nonlinear case, the shifted inverse iteration (Algorithms 2 or 5) is more essential, actually Algorithm 1 may no longer be applicable since equation (2) often has no real solution. This point is illustrated in [6] where the shift is based on a generalization of (15). In view of [2; Theorem 2.3 and Corollary 2.5], it seems that Algorithm 2 and its variations could be applied to a more general setup.

Acknowledgments The author thanks Ms Yue-Shuang Li for her assistance in computing the large matrices using MatLab, and also pointed out the error on h_{N+1} mentioned in §4.4. The author also acknowledges Mr Xu Zhu for constructing Example 18 which leads us to find out the error just mentioned. Research supported in part by National Natural Science Foundation of China (Grant Nos. 11626245, 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Arapostathis, A., Borkar, V.S. and Kumar, K.S. (2016). *Risk–sensitive control and an abstract Collatz–Wielandt formula*. J. Theor. Probab. 29(4), 1458–1484.

[2] Chang, K.C. (2014). *Nonlinear extensions of the Perron–Frobenius theorem and the Krein–Rutman theorem*. J. Fixed Point Theory Appl. 15, 433–457.

[3] Chen, M.F. (2004). *From Markov Chains to Non-Equilibrium Particle Systems*. World Scientific, Singapore, 2nd Ed. (1st Ed., 1992).

[4] Chen, M.F. (2010). Speed of stability for birth–death processes. Front. Math. China 5(3), 379–515.

[5] Chen, M.F. (2016). *Efficient initials for computing the maximal eigenpair*. Front. Math. China 11(6): 1379–1418. See also volume 4 in the middle of the author's homepage:

http://math0.bnu.edu.cn/~chenmf

A package based on the paper is available on CRAN now (by X.J. Mao). One may check it through the link:
https://cran.r-project.org/web/packages/EfficientMaxEigenpair/index.html

[6] Chen, M.F. (2017) *Efficient algorithm for principal eigenpair of discrete p-Laplacian*. Preprint.

[7] Chen, M.F. and Zhang, X. (2014) *Isospectral operators*. Commu Math Stat 2, 17–32.

[8] Chen, R.R. (1997). *An Extended Class of Time-Continuous Branching Processes*. J. Appl. Probab. 34(1), 14-23

[9] Donsker, W.D. and Varadhan, S.R.S. (1975). *On a variational formula for the principal eigenvalue for operators with maximum principle*. Proc. Natl. Acad. Sci. 72(3), 780–783.

[10] Noutsos, D. (2006). *On Perron-Frobenius property of matrices having some negative entries*. Linear Algebra Appl. 412, 132–153.

[11] Noutsos, D. (2008). *Perron Frobenius theory and some extensions*. http://www.pdfdrive.net/perron-frobenius-theory-and-some-extensions-e10082439.html

[12] Noutsos, D. and Varga, R.S. (2012). *On the Perron–Frobenius theory for complex matrices*. Linear Algebra and its Applications 437, 1071–1088.

[13] Sheu, S.J. (1984). *Stochastic control and principal eigenvalue*. Stochastics 11(3–4), 191–211.

Mu-Fa Chen
School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.
E-mail: mfchen@bnu.edu.cn
Home page: http://math0.bnu.edu.cn/~chenmf/main_eng.htm

In "Queueing Theory and Network Applications", LNCS 10591: 312–329
Springer 2017 (Updated 2017/12/25)

Trilogy on Computing Maximal Eigenpair

Mu-Fa Chen

(Beijing Normal University)

June 8, 2017

Abstract The eigenpair here means the twins consist of eigenvalue and its eigenvector. This paper introduces the three steps of our study on computing the maximal eigenpair. In the first two steps, we construct efficient initials for a known but dangerous algorithm, first for tridiagonal matrices and then for irreducible matrices, having nonnegative off-diagonal elements. In the third step, we present two global algorithms which are still efficient and work well for a quite large class of matrices, even complex for instance.

2000 *Mathematics Subject Classification*: 15A18, 65F15, 93E15
Key words and phrases. Maximal eigenpair, efficient initial, tridiagonal matrix, global algorithm.

1 Introduction

This paper is a continuation of [4]. For the reader's convenience, we review (with some improvements) shortly the first part of [4]. Especially, we recall the story of the proportion of 1000 and 2 of iterations for two different algorithms.

The most famous result on the maximal eigenpair should be the Perron-Frobenius theorem. For nonnegative (pointwise) and irreducible A, if Trace $(A) > 0$, then the theorem says there exists uniquely a maximal eigenvalue $\rho(A) > 0$ with positive left-eigenvector u and positive right-eigenvector g such that

$$uA = \lambda u, \qquad Ag = \lambda g, \qquad \lambda = \rho(A).$$

These eigenvectors are also unique up to a constant. Before going to the main body of the paper, let us make two remarks.

1) We need to study the right-eigenvector g only. Otherwise, use the transpose A^* instead of A.

2) The matrix A is required to be irreducible with nonnegative off-diagonal elements, its diagonal elements can be arbitrary. Otherwise, use a shift $A+mI$ for large m:

$$(A + mI)g = \lambda g \Longleftrightarrow Ag = (\lambda - m)g, \tag{1}$$

their eigenvector remains the same but the maximal eigenvalues are shifted to each other.

Consider the following matrix:

$$Q=\begin{pmatrix} -1^2 & 1^2 & 0 & 0 & \cdots \\ 1^2 & -1^2-2^2 & 2^2 & 0 & \cdots \\ 0 & 2^2 & -2^2-3^2 & 3^2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & N^2 & -N^2-(N+1)^2 \end{pmatrix}. \tag{2}$$

The main character of the matrix is the sequence $\{k^2\}$. The sum of each row equals zero except the last row. Actually, this matrix is truncated from the corresponding infinite one, in which case we have known that the maximal eigenvalue is $-1/4$ (refer to [2; Example 3.6]).

Example 1 Let $N = 7$. Then the maximal eigenvalue is -0.525268 with eigenvector:

$$g \approx (55.878,\ 26.5271,\ 15.7059,\ 9.97983,\ 6.43129,\ 4.0251,\ 2.2954,\ 1)^*,$$

where the vector v^* = the transpose of v.

We now want to practice the standard algorithms in matrix eigenvalue computation. The first method in computing the maximal eigenpair is the *Power Iteration*, introduced in 1929. Starting from a vector v_0 having a nonzero component in the direction of g, normalized with respect to a norm $\|\cdot\|$. At the kth step, iterate v_k by the formula

$$v_k = \frac{Av_{k-1}}{\|Av_{k-1}\|}, \quad z_k = \|Av_k\|, \qquad k \geqslant 1. \tag{3}$$

Then we have the convergence: $v_k \to g$ (first pointwise and then uniformly) and $z_k \to \rho(Q)$ as $k \to \infty$. If we rewrite v_k as

$$v_k = \frac{A^k v_0}{\|A^k v_0\|},$$

one sees where the name "power" comes from. For our example, to use the Power Iteration, we adopt the ℓ^1-norm and choose $v_0 = \tilde{v}_0/\|\tilde{v}_0\|$, where

$$\tilde{v}_0=(1,\ 0.587624,\ 0.426178,\ 0.329975,\ 0.260701,\ 0.204394, 0.153593, 0.101142)^*.$$

This initial comes from a formula to be given in the next section. In Figure 1 below, the upper curve is g, the lower one is modified from $\tilde{v}_0$, renormalized so that its last component becomes one. Clearly, these two functions are quite different, one may worry about the effectiveness of the choice of v_0. Anyhow, having the experience of computing its eigensystem, I expect to finish the computation in a few of seconds. Unexpectly, I got a difficult time to compute

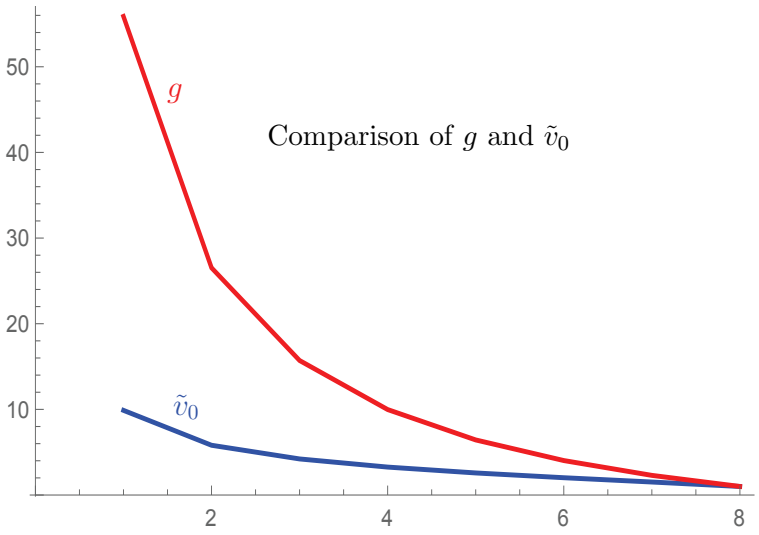


Figure 1: g and $\tilde{v}_0$.

the maximal eigenpair for this simple example. Altogether, I computed it for 180 times, not in one day, using 1000 iterations. The printed pdf-file of the outputs has 64 pages. Figure 2 gives us the outputs.

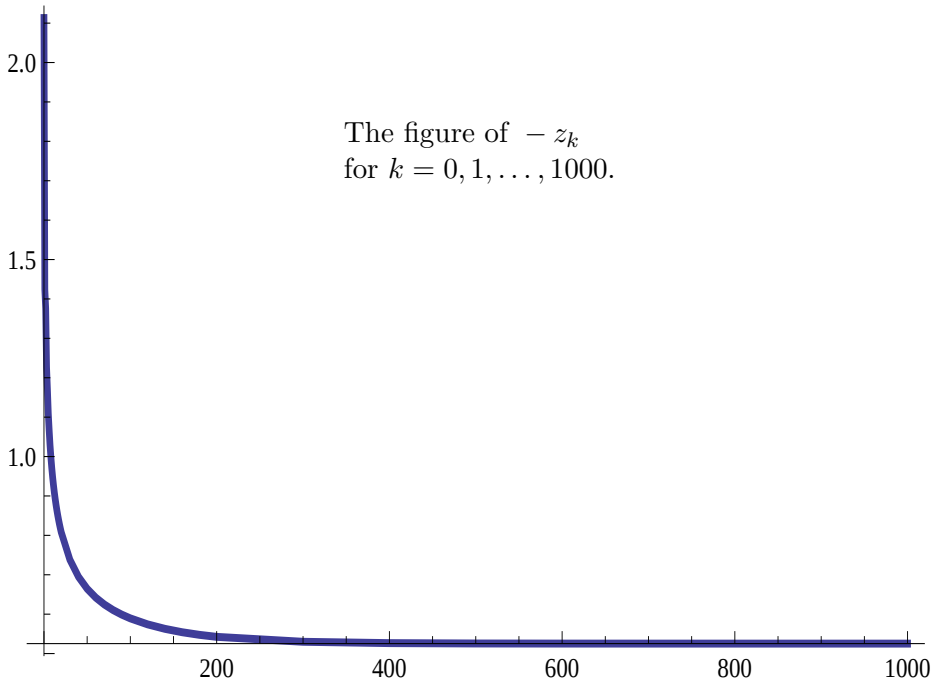


Figure 2: $-z_k$ for $k = 0, 1, \ldots, 1000$.

The figure shows that the convergence of z_k goes quickly at the beginning of the iterations. This means that our initial v_0 is good enough. Then the

convergence goes very slow which means that the Power Iteration Algorithm converges very slowly.

Let us have a look at the convergence of the power iteration. Suppose that the eigenvalues are all different for simplicity. Denote by (λ_j, g_j) the eigenpairs with maximal one (λ_0, g_0). Write $v_0 = \sum_{j=0}^{N} c_j g_j$ for some constants (c_j). Then $c_0 \neq 0$ by assumption and

$$A^k v_0 = \sum_{j=0}^{N} c_j \lambda_j^k g_j = c_0 \lambda_0^k \left[g_0 + \sum_{j=1}^{N} \frac{c_j}{c_0} \left(\frac{\lambda_j}{\lambda_0} \right)^k g_j \right].$$

Since $|\lambda_j/\lambda_0| < 1$ for each $j \geqslant 1$ and $\|g_0\| = 1$, we have

$$\frac{A^k v_0}{\|A^k v_0\|} = \frac{c_0}{|c_0|} g_0 + O\left(\left| \frac{\lambda_1}{\lambda_0} \right|^k \right) \qquad \text{as } k \to \infty,$$

where $|\lambda_1| := \max\{|\lambda_j| : j > 0\}$. Since $|\lambda_1/\lambda_0|$ can be very closed to 1, this explains the reason why the convergence of the method can be very slow.

Before moving further, let us mention that the power method can be also used to compute the minimal eigenvalue $\lambda_{\min}(A)$, simply replace A by A^{-1}. That is the *Inverse Iteration* introduced in 1944:

$$v_k = \frac{A^{-1} v_{k-1}}{\|A^{-1} v_{k-1}\|} \Longleftrightarrow v_k = \frac{A^{-k} v_0}{\|A^{-k} v_0\|}. \tag{4}$$

It is interesting to note that the equivalent assertion on the right-hand side is exactly the the input-output method in economy.

To come back to compute the maximal $\rho(A)$ rather than $\lambda_{\min}(A)$, we add a shift z to A: replacing A by $A - zI$. Actually, it is even better to replace the last one by $zI - A$ since we will often use $z > \rho(A)$ rather than $z < \rho(A)$, the details will be explained at the beginning of Section 4 below. When z is close enough to $\rho(A)$, the leading eigenvalue of $(zI - A)^{-1}$ becomes $(z - \rho(A))^{-1}$. Furthermore, we can even use a variant shift $z_{k-1} I$ to accelerate the convergence speed. Throughout this paper, we use varying shifts rather than a fixed one only. Thus, we have arrived at the second algorithm in computing the maximal eigenpair, the *Rayleigh Quotient Iteration* (RQI), a variant of the *Inverse Iteration*. From now on, unless otherwise stated, we often use the ℓ^2-norm. Starting from an approximating pair (z_0, v_0) of the maximal one $(\rho(A), g)$ with $v_0^* v_0 = 1$, use the following iteration.

$$v_k = \frac{(z_{k-1} I - A)^{-1} v_{k-1}}{\|(z_{k-1} I - A)^{-1} v_{k-1}\|}, \qquad z_k = v_k^* A v_k, \qquad k \geqslant 1. \tag{5}$$

If (z_0, v_0) is close enough to $(\rho(A), g)$, then

$$v_k \to g \quad \text{and} \quad z_k \to \rho(A) \quad \text{as } k \to \infty.$$

Since for each $k \geqslant 1$, $v_k^* v_k = 1$, we have $z_k = v_k^* A v_k / (v_k^* v_k)$. That is where the name "Rayleigh Quotient" comes from. Unless otherwise stated, z_0 is setting to be $v_0^* A v_0$.

Having the hard time spent in the first algorithm, I wondered how many iterations are required using this algorithm. Of course, I can no longer bear 1000 iterations. To be honest, I hope to finish the computation within 100 iterations. What happens now?

Example 2 For the same matrix Q and $\tilde{v}_0$ as in Example 1, by RQI, we need two iterations only:

$$z_1 \approx -0.528215, \quad z_2 \approx -0.525268.$$

The result came to me, not enough to say surprisingly, I was shocked indeed. This shows not only the power of the second method but also the effectiveness of my initial v_0. From the examples above, we have seen the story of the proportion of 1000 and 2.

For simplicity, from now on, we often write $\lambda_j := \lambda_j(-Q)$. In particular $\lambda_0 = -\rho(Q) > 0$. Instead of our previous v_0, we adopt the uniform distribution:

$$v_0 = (1, 1, 1, 1, 1, 1, 1, 1)^*/\sqrt{8}.$$

This is somehow fair since we usually have no knowledge about g in advance.

Example 3 Let Q be the same as above. Use the uniform distribution v_0 and set $z_0 = v_0^*(-Q)v_0$. Then

$$(z_1, z_2, z_3, \boldsymbol{z_4}) \approx (4.78557,\ 5.67061,\ 5.91766,\ \mathbf{5.91867}).$$
$$(\lambda_0, \lambda_1, \boldsymbol{\lambda_2}) \approx (0.525268,\ 2.00758,\ \mathbf{5.91867}).$$

The computation becomes stable at the 4th iteration. Unfortunately, it is not what we want λ_0 but λ_2. In other words, the algorithm converges to a pitfall. Very often, there are $n-1$ pitfalls for a matrix having n eigenvalues. This shows once again our initial $\tilde{v}_0$ is efficient and the RQI is quite dangerous.

Hopefully, everyone here has heard the name *Google's PageRank*. In other words, the Google's search is based on the maximal left-eigenvector. On this topic, the book [8] was published 11 years ago. In this book, the Power Iteration is included but not the RQI. It should be clear that for PageRank, we need to consider not only large system, but also fast algorithm.

It may be the correct position to mention a part of the motivations for the present study.

- Google's search–PageRank.
- Input–output method in economy. In this and the previous cases, the computation of the maximal eigenvector is required.
- Stability speed of stochastic systems. Here, for the stationary distribution of a Markov chain, we need to compute the eigenvector; and for the stability rate, we need to study the maximal (or the fist nontrivial) eigenvalue.

- Principal component analysis for BigData. One choice is to study the so-called five-diagonal matrices. The second approach is using the maximal eigenvector to analysis the role played by the components, somehow similar to the PageRank.
- For image recognition, one often uses Poisson or Toeplitz matrices, which are more or less the same as the Quasi-birth-death matrices studied in queueing theory. The discrete difference equations of elliptic partial differential equations are included in this class: the block-tridiagonal matrices.
- The effectiveness of random algorithm, say Markov Chain Monte Carlo for instance, is described by the convergence speed. This is also related to the algorithms for machine learning.
- As in the last item, a mathematical tool to describe the phase transitions is the first nontrivial eigenvalue (the next eigenpair in general). This is the original place where the author was attracted to the topic.

Since the wide range of the applications of the topic, there is a large number of publications. The author is unable to present a carefully chosen list of references here, what instead are two random selected references: [8] and [11].

Up to now, we have discussed only a small size $8 \times 8\,(N = 7)$ matrix. How about large N? In computational mathematics, one often expects the number of iterations grows in a polynomial way N^α for α greater or equal to 1. In our efficient case, since $2 = 8^{1/3}$, we expect to have $10000^{1/3} \approx 22$ iterations for $N+1=10^4$. The next table subverts completely my imagination.

Table 1 Comparison of RQI for different N

$N+1$	z_0	z_1	$z_2 = \lambda_0$	upper/lower
8	0.523309	0.525268	0.525268	$1+10^{-11}$
100	0.387333	0.376393	0.376383	$1+10^{-8}$
500	0.349147	0.338342	0.338329	$1+10^{-7}$
1000	0.338027	0.327254	0.32724	$1+10^{-7}$
5000	0.319895	0.30855	0.308529	$1+10^{-7}$
7500	0.316529	0.304942	0.304918	$1+10^{-7}$
10^4	0.31437	0.302586	0.302561	$1+10^{-7}$

Here z_0 is defined by

$$z_0 = 7/(8\delta_1) + v_0^*(-Q)v_0/8,$$

where v_0 and δ_1 are computed by our general formulas to be defined in the next section. We compute the matrices of order $8, 100, \ldots, 10^4$ by using MatLab in a notebook, in no more than 30 seconds, the iterations finish at the second step. This means that the outputs starting from z_2 are the same and coincide with λ_0. See the first row for instance, which becomes stable at the first step indeed. We do not believe such a result for some days, so we checked it in

different ways. First, since $\lambda_0 = 1/4$ when $N = \infty$, the answers of λ_0 given in the fourth column are reasonable. More essentially, by using the output v_2, we can deduce upper and lower bounds of λ_0 (using [2; Theorem 2.4 (3)]), and then the ratio upper/ lower is presented in the last column. In each case, the algorithm is significant up to 6 digits. For the large scale matrices here and in 4, the computations are completed by Yue-Shuang Li.

2 Efficient initials: tridiagonal case

It is the position to write down the formulas of v_0 and δ_1. Then our initial z_0 used in Table 1 is a little modification of δ_1^{-1}: a convex combination of δ_1^{-1} and $v_0^*(-Q)v_0$.

Let us consider the tridiagonal matrix (cf. [3; §3] and [6; §4.4]). Fix $N \geqslant 1$, denote by $E = \{0, 1, \ldots, N\}$ the set of indices. By a shift if necessary, we may reduce A to Q with negative diagonals: $Q^c = A - mI$, $m := \max_{i\in E} \sum_{j\in E} a_{ij}$,

$$Q^c = \begin{pmatrix} -b_0 - c_0 & b_0 & 0 & 0 & \cdots \\ a_1 & -a_1 - b_1 - c_1 & b_1 & 0 & \cdots \\ 0 & a_2 & -a_2 - b_2 - c_2 & b_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & a_N & -a_N - c_N \end{pmatrix}.$$

Thus, we have three sequences $\{a_i > 0\}$, $\{b_i > 0\}$, and $\{c_i \geqslant 0\}$. Our main assumption here is that the first two sequences are positive and $c_i \not\equiv 0$. In order to define our initials, we need three new sequences, $\{h_k\}$, $\{\mu_k\}$, and $\{\varphi_k\}$.

First, we define the sequence $\{h_k\}$:

$$h_0 = 1, \quad h_n = h_{n-1} r_{n-1}, \qquad 1 \leqslant n \leqslant N; \tag{6}$$

here we need another sequence $\{r_k\}$:

$$r_0 = 1 + \frac{c_0}{b_0}, \quad r_n = 1 + \frac{a_n + c_n}{b_n} - \frac{a_n}{b_n r_{n-1}}, \qquad 1 \leqslant n < N.$$

Here and in what follows, our iterations are often of one-step. Note that if $c_k = 0$ for every $k < N$, then we do not need the sequence $\{h_k\}$, simply set $h_k \equiv 1$. An easier way to remember this (h_i) is as follows. It is nearly harmonic of Q^c except at the last point N:

$$Q^{c\,\setminus \text{the last row}} h = 0, \tag{7}$$

where $B^{\setminus \text{the last row}}$ means the matrix modified from B by removing its last low.

We now use H-transform, it is designed to remove the sequence (c_i):

$$\widetilde{Q} = \operatorname{Diag}(h_i)^{-1} Q^c \operatorname{Diag}(h_i).$$

Then

$$\widetilde{Q}=\begin{pmatrix} -b_0 & b_0 & 0 & 0 & \cdots \\ a_1 & -a_1-b_1 & b_1 & 0 & \cdots \\ 0 & a_2 & -a_2-b_2 & b_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & a_N & -a_N-c_N \end{pmatrix}$$

for some modified $\{a_i>0\}$, $\{b_i>0\}$, and $c_N>0$. Of course, Q^c and $\widetilde{Q}$ have the same spectrum. In particular, under the H-transform,

$$(\lambda_{\min}(-Q^c),\ g)\to\big(\lambda_{\min}\big(-\widetilde{Q}\big)=\lambda_{\min}(-Q^c),\ \mathrm{Diag}(h_i)^{-1}g\big).$$

From now on, for simplicity, we denote by Q the matrix replacing c_N by b_N in $\widetilde{Q}$.

Next, we define the second sequence $\{\mu_k\}$:

$$\mu_0=1,\ \ \mu_n=\mu_{n-1}\frac{b_{n-1}}{a_n},\qquad 1\leqslant n\leqslant N. \tag{8}$$

And then define the third one $\{\varphi_k\}$ as follows:

$$\varphi_n=\sum_{k=n}^{N}\frac{1}{\mu_k b_k},\qquad 0\leqslant n\leqslant N. \tag{9}$$

We are now ready to define v_0 and δ_1 (or z_0) using the sequences (μ_i) and (φ_i).

$$\tilde{v}_0(i)=\sqrt{\varphi_i},\ i\leqslant N;\qquad v_0=\tilde{v}_0/\|\tilde{v}_0\|;\quad \|\cdot\|:=\|\cdot\|_{L^2(\mu)} \tag{10}$$

$$\delta_1=\max_{0\leqslant n\leqslant N}\left[\sqrt{\varphi_n}\sum_{k=0}^{n}\mu_k\sqrt{\varphi_k}+\frac{1}{\sqrt{\varphi_n}}\sum_{n+1\leqslant j\leqslant N}\mu_j\varphi_j^{3/2}\right]=:z_0^{-1} \tag{11}$$

with a convention $\sum_{\emptyset}=0$.

Finally, having constructed the initials (v_0,z_0), the RQI goes as follows. Solve w_k:

$$(-Q-z_{k-1}I)w_k=v_{k-1},\qquad k\geqslant 1; \tag{12}$$

and define

$$v_k=w_k/\|w_k\|,\qquad z_k=(v_k,\ -Q\,v_k)_{L^2(\mu)}.$$

Then

$$v_k\to g\quad\text{and}\quad z_k\to\lambda_0\qquad\text{as } k\to\infty.$$

Before moving further, let us mention that there is an explicit representation of the solution (w_i) to equation (12). Assume that we are given $v:=v_{k-1}$ and $z:=z_{k-1}$. Set

$$M_{sj}=\mu_j\sum_{k=j}^{s}\frac{1}{\mu_k b_k},\qquad 0\leqslant j\leqslant s\leqslant N. \tag{13}$$

Define two independent sequences $\{A(s)\}$ and $\{B(s)\}$, recurrently:

$$\begin{cases} A(s) = -\sum_{0\leqslant j\leqslant s-1} M_{s-1,j}\big(v(j) + zA(j)\big), \\ B(s) = 1 - z\sum_{0\leqslant j\leqslant s-1} M_{s-1,j}B(j), \qquad 0 \leqslant s \leqslant N. \end{cases} \tag{14}$$

Set

$$x = \frac{\sum_{j=0}^{N} \mu_j\big(v(j) + zA(j)\big) - \mu_N b_N A(N)}{\mu_N b_N B(N) - z\sum_{j=0}^{N} \mu_j B(j)}. \tag{15}$$

Then the required solution $w_k := \{w(s) : s \in E\}$ can be expressed as $w(s) = A(s) + xB(s)\,(s \in E)$.

To finish the algorithm, we return to the estimates of $\big(\lambda_{\min}(-Q^c), g(Q^c)\big)$ $(g(Q^c) = g(-Q^c))$ or further $(\rho(A), g(A))$ if necessary, where $g(A)$, for instance, denotes the maximal eigenvector of A. Suppose that the iterations are stopped at $k = k_0$ and set $(\bar{z}, \bar{v}) = \big(z_{k_0}, v_{k_0}\big)$ for simplicity. Then, we have

$$\big(\lambda_{\min}\big(-Q^c\big),\ \mathrm{Diag}(h_i)^{-1}g(Q^c)\big) = \big(\lambda_{\min}\big(-\widetilde{Q}\big),\ g\big(\widetilde{Q}\big)\big) \approx (\bar{z}, \bar{v}),$$

and so

$$\big(\lambda_{\min}(-Q^c),\ g(Q^c)\big) \approx \big(\bar{z},\ \mathrm{Diag}(h_i)\,\bar{v}\big). \tag{16}$$

Because $\lambda_{\min}(-Q^c) = m - \rho(A)$, we obtain

$$(\rho(A),\ g(A)) \approx \big(m - \bar{z},\ \mathrm{Diag}(h_i)\,\bar{v}\big). \tag{17}$$

Now, the question is the possibility from the tridiagonal case to the general one.

3 Efficient initials: the general case ([3; §4.2] and [6; §4.5])

When we first look at the question just mentioned, it seems quite a long distance to go from the special tridiagonal case to the general one. However, in the eigenvalue computation theory, there is the so-called Lanczos tridiagonalization procedure to handle the job, as discussed in [3; Appendix of §3]. Nevertheless, what we adopted in [3; §4] is a completely different approach. Here is our main idea. Note that the initials v_0 and δ_1 constructed in the last section are explicitly expressed by the new sequences. In other words, we have used three new sequences $\{h_k\}$, $\{\mu_k\}$, and $\{\varphi_k\}$ instead of the original three $\{a_i\}$, $\{b_i\}$, and $\{c_i\}$ to describe our initials. Very fortunately, the former three sequences do have clearly the probabilistic meaning, which then leads us a way to go to the general setup. Shortly, we construct these sequences by solving three linear equations (usually, we do not have explicit solution in such a general setup). Then use them to construct the initials and further apply the RQI-algorithm.

Let $A = (a_{ij} : i, j \in E)$ be the same as given at the beginning of the paper. Set $A_i = \sum_{j\in E} a_{ij}$ and define

$$Q^c = A - \Big(\max_{i\in E} A_i\Big) I.$$

We can now state the probabilistic/analytic meaning of the required three sequences (h_i), (μ_i), and (φ_i).

- (h_i) is the harmonic function of Q^c except at the right endpoint N, as mentioned in the last section.
- (μ_i) is the invariant measure (stationary distribution) of the matrix Q^c removing the sequence (c_i).
- (φ_i) is the tail related to the transiency series, refer to [3; Lemma 24 and its proof].

We now begin with our construction. Let $h = (h_0, h_1, \ldots, h_N)^*$ (with $h_0 = 1$) solve the equation

$$Q^{c\,\setminus\text{the last row}} h = 0$$

and define

$$\widetilde{Q} = \mathrm{Diag}(h_i)^{-1} Q^c \,\mathrm{Diag}(h_i).$$

Then for which we have

$$c_0 = \ldots = c_{N-1} = 0, \quad c_N =: q_{N,N+1} > 0.$$

This is very much similar to the tridiagonal case.

Next, set $Q = \widetilde{Q}$. Let $\varphi = (\varphi_0, \varphi_1, \ldots, \varphi_N)^*$ (with $\varphi_0 = 1$) solve the equation

$$\varphi^{\setminus\text{the first row}} = P^{\setminus\text{the first row}}\, \varphi,$$

where

$$P = \mathrm{Diag}\big((-q_{ii})^{-1}\big) Q + I.$$

Thirdly, assume that $\mu := (\mu_0, \mu_1, \ldots, \mu_N)$ with $\mu_0 = 1$ solves the equation

$$Q^{*\,\setminus\text{the last row}} \mu^* = 0.$$

Having these sequences at hand, we can define the initials

$$\tilde{v}_0(i) = \sqrt{\varphi_i}, \;\; i \leqslant N; \quad v_0 = \tilde{v}_0/\|\tilde{v}_0\|_\mu; \quad z_0 = (v_0, -Qv_0)_\mu.$$

Then, go to the RQI as usual. For $k \geqslant 1$, let w_k solve the equation

$$(-Q - z_{k-1} I) w_k = v_{k-1}$$

and set

$$v_k = w_k/\|w_k\|_\mu, \qquad z_k = (v_k, -Qv_k)_\mu.$$

Then we often have $(z_k, v_k) \to (\lambda_0, g)$ as $k \to \infty$.

We remark that there is an alternative choice (more safe) of z_0:

$$z_0^{-1} = \frac{1}{1-\varphi_1} \max_{0 \leqslant n \leqslant N} \left[\sqrt{\varphi_n} \sum_{k=0}^{n} \mu_k \sqrt{\varphi_k} + \frac{1}{\sqrt{\varphi_n}} \sum_{n+1 \leqslant j \leqslant N} \mu_j \varphi_j^{3/2} \right]$$

which is almost a copy of the one used in the last section.

The procedure for returning to the estimates of $\big(\lambda_{\min}(-Q^c), g(Q^c)\big)$ or further $(\rho(A), g(A))$ is very much the same as in the last section.

To conclude this section, we introduce two examples to illustrate the efficiency of the extended initials for tridiagonally dominant matrices. The next two examples were computed by Xu Zhu, a master student in Shanghai.

Example 4 (Block-tridiagonal matrix) Consider the matrix

$$Q = \begin{pmatrix} A_0 & B_0 & 0 & 0 & \cdots \\ C_1 & A_1 & B_1 & 0 & \cdots \\ 0 & C_2 & A_2 & B_2 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & C_N & A_N \end{pmatrix},$$

where A_k, B_k, C_k are 40×40-matrices, B's and C's are identity matrices, and A's are tridiagonal matrices. For this model, two iterations are enough to arrive at the required results (Table 2).

Table 2 Outputs for Poisson matrix

$N+1$	z_0	z_1	$z_2 = \lambda_0$
1600	7.985026	7.988219	7.988263
3600	7.993232	7.994676	7.994696
6400	7.996161	7.988256	7.987972

Example 5 (Toeplitz matrix) Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & \cdots & n-1 & n \\ 2 & 1 & 2 & \cdots & n-2 & n-1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ n-1 & n-2 & n-3 & \cdots & 1 & 2 \\ n & n-1 & n-2 & \cdots & 2 & 1 \end{pmatrix}.$$

For this model, three iterations are enough to arrive at the required results (Table 3).

Table 3 Outputs for Toeplitz matrix

$N+1$	$z_0\times10^6$	$z_1\times10^6$	$z_2\times10^6$	$z_3=\lambda_0$
1600	0.156992	0.451326	0.390252	0.389890
3600	0.157398	2.30731	1.97816	1.97591
6400	0.157450	7.32791	6.25506	6.24718

As mentioned before, the extended algorithm should be powerful for the tridiagonally dominant matrices. How about more general case? Two questions are often asked to me by specialists in computational mathematics: do you allow more negative off-diagonal elements? How about complex matrices? My answer is: they are too far away from me, since those matrices can not be a generator of a Markov chain, I do not have a tool to handle them. Alternatively, I have studied some more general matrices than the tridiagonal ones: the block-tridiagonal matrices, the lower triangular plus upper-diagonal, the upper triangular plus lower-diagonal, and so on. Certainly, we can do a lot case by case, but this seems still a long way to achieve a global algorithm. So we do need a different idea.

4 Global algorithms

Several months ago, AlphaGo came to my attention. From which I learnt the subject of machine learning. After some days, I suddenly thought, since we are doing the computational mathematics, why can not let the computer help us to find a high efficiency initial value? Why can not we leave this hard task to the computer? If so, then we can start from a relatively simple and common initial value, let the computer help us to gradually improve it.

The first step is easy, simply choose the uniform distribution as our initial v_0:

$$v_0 = (1, 1, \cdots, 1)^*/\sqrt{N+1}.$$

As mentioned before, this initial vector is fair and universal. One may feel strange at the first look at "global" in the title of this section. However, with this universal v_0, the power iteration is already a global algorithm. Unfortunately, the convergence of this method is too slow, and hence is often not practical. To quicken the speed, we should add a shift which now has a very heavy duty for our algorithm. The main trouble is that the usual Rayleigh quotient $v_0^*Av_0/(v_0^*v_0)$ can not be used as z_0, otherwise, it will often lead to a pitfall, as illustrated by Example 3. The main reason is that our v_0 is too rough and so z_0 deduced from it is also too rough. Now, how to choose z_0 and further z_n?

Clearly, for avoiding the pitfalls, we have to choose z_0 from the outside of the spectrum of A (denoted by $\mathrm{Sp}(A)$), and as close to $\rho(A)$ as possible to quicken the convergence speed. For nonnegative A, $\mathrm{Sp}(A)$ is located in a circle with radius $\rho(A)$ in the complex plane. Thus, the safe region should be on the outside of $\mathrm{Sp}(A)$. Since $\rho(A)$ is located at the boundary on the right-hand side of the circle, the effective area should be on the real axis on the right-hand side of, but a little away from, $\rho(A)$.

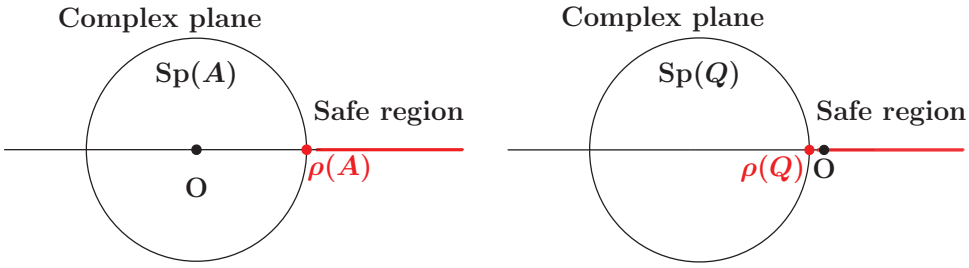


Figure 3: Safe region in complex plane.

For the matrix Q used in this paper, since $\rho(Q) < 0$, its spectrum $Sp(Q)$ is located on the left-hand side of the origin. Then, one can simply choose $z_0 = 0$ as an initial. See Figure 3.

Having these idea in mind, we can now state two of our global algorithms. Each of them uses the same initials:

$$v_0 = \text{uniform distribution}, \qquad z_0 = \max_{0 \leqslant i \leqslant N} \frac{Av_0}{v_0}(i),$$

where for two vectors f and g, $(f/g)(i) = f_i/g_i$.

Algorithm 1 (Specific Rayleigh quotient iteration) At step $k \geqslant 1$, for given $v := v_{k-1}$ and $z := z_{k-1}$, let w solve the equation

$$(zI - A)w = v.$$

Set $v_k = w/\|w\|$ and let $z_k = v_k^* A v_k$.

This algorithm goes back to [3; §4.1 with Choice I].

Algorithm 2 (Shifted inverse iteration) Everything is the same as in Algorithm 1, except redefine z_k as follows:

$$z_k = \max_{0 \leqslant i \leqslant N} \frac{Av_k}{v_k}(i)$$

for $k \geqslant 1$ (or equivalently, $k \geqslant 0$).

The comparison of these algorithms is the following: with unknown small probability, Algorithm 1 is less safe than Algorithm 2, but the former one has a faster convergence speed than the latter one with possibility 1/5 for instance. A refined combination of the above two algorithms is presented in [6; §2], say Algorithm 4_2 for instance.

With the worrying on the safety and convergence speed in mind, we examine two examples which are non-symmetric.

The first example below is a lower triangular plus the upper-diagonal. It is far away from the tridiagonal one, we want to see what can be happened.

Example 6 ([6; Example 7]) Let

$$Q=\begin{pmatrix} -1 & 1 & 0 & 0 & \cdots\cdots & 0 & 0 \\ a_1 & -a_1-2 & 2 & 0 & \cdots\cdots & 0 & 0 \\ a_2 & 0 & -a_2-3 & 3 & \cdots\cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots\cdots & N-1 & 0 \\ a_{N-1} & 0 & 0 & 0 & \cdots & -a_{N-1}-N & N \\ a_N & 0 & 0 & 0 & \cdots\cdots & 0 & -a_N-N-1 \end{pmatrix}. \quad (18)$$

For this matrix, we have computed several cases:

$$a_k = 1/(k+1),\ \ a_k \equiv 1,\ \ a_k = k,\ \ a_k = k^2.$$

Among them, the first one is the hardest and is hence presented below.

For different N, the outputs of our algorithm are given in Table 4.

Table 4. The outputs for different N by our algorithm

$N+1$	z_1	z_2	z_3	z_4	z_5	z_6
8	0.276727	0.427307	0.451902	0.452339		
16	0.222132	0.367827	0.399959	0.400910		
32	0.187826	0.329646	0.370364	0.372308	0.372311	
50	0.171657	0.311197	0.357814	0.360776	0.360784	
100	0.152106	0.287996	0.343847	0.349166	0.349197	
500	0.121403	0.247450	0.321751	0.336811	0.337186	
1000	0.111879	0.233257	0.313274	0.334155	0.335009	0.335010
5000	0.0947429	0.205212	0.293025	0.328961	0.332609	0.332635
10^4	0.0888963	0.194859	0.284064	0.326285	0.332113	0.332188

The next example is upper triangular plus lower-diagonal. It is motivated from the classical branching process. Denote by $(p_k : k \geqslant 0)$ a given probability measure with $p_1 = 0$. Let

$$Q=\begin{pmatrix} -1 & p_2 & p_3 & p_4 & \cdots\cdots & p_{N-1} & \sum_{k\geqslant N} p_k \\ 2p_0 & -2 & 2p_2 & 2p_3 & \cdots\cdots & 2p_{N-2} & 2\sum_{k\geqslant N-1} p_k \\ 0 & 3p_0 & -3 & 3p_2 & \cdots & 3p_{N-3} & 3\sum_{k\geqslant N-2} p_k \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots & -(N-1) & (N-1)\sum_{k\geqslant 2} p_k \\ 0 & 0 & 0 & 0 & \cdots\cdots & Np_0 & -Np_0 \end{pmatrix}.$$

The matrix is defined on $E := \{1, 2, \ldots, N\}$. Set $M_1 = \sum_{k\in E} kp_k$. When $N = \infty$, it is subcritical iff $M_1 < 1$, to which the maximal eigenvalue should be positive. Otherwise, the convergence rate should be zero.

Now, we fix

$$p_0 = \alpha/2,\ p_1 = 0,\ p_2 = (2-\alpha)/2^2,\ \ldots p_n = (2-\alpha)/2^n, \cdots, \qquad \alpha \in (0, 2).$$

Then $M_1 = 3(2-\alpha)/2$ and hence we are in the subcritical case iff $\alpha \in (4/3, 2)$.

Example 7 ([6; Example 9]) Set $\alpha = 7/4$. We want to know how fast the local ($N < \infty$) maximal eigenvalue becomes stable (i.e., close enough to the converge rate at $N = \infty$). Up to $N = 10^4$, the steps of the iterations we need are no more than 6. To quicken the convergence, we adopt an improved algorithm. Then the outputs of the approximation of the minimal eigenvalue of $-Q$ for different N are given in Table 5.

Table 5. The outputs in the subcritical case

N	z_1	z_2	z_3	z_4
8	0.637800	0.638153		
16	0.621430	0.625490	0.625539	
50	0.609976	0.624052	0.624997	0.625000
100	0.606948	0.623377	0.624991	0.625000
500	0.604409	0.622116	0.624962	0.625000
1000	0.604082	0.621688	0.624944	0.625000
5000	0.603817	0.620838	0.62489	0.625000
10^4	0.603784	0.620511	0.624861	0.625000

The computation in each case costs no more than one minute. Besides, starting from $N = 50$, the final outputs are all the same: 0.625, which then can be regarded as a very good approximation of $\lambda_{\min}(-Q)$ at infinity $N = \infty$.

It is the position to compare our global algorithm with that given in the last section. At the first look, here in the two examples above, we need about 6 iterations, double of the ones given in the last section. Note that for the initials of the algorithm in the last section, we need solve three additional linear equations, which are more or less the same as three additional iterations. Hence the efficiency of these two algorithms are very close to each other. Actually, the computation time used for the algorithm in the last section is much more than the new one here.

It is quite surprising that our new algorithms work for a much general class of matrices, out of the scope of [3]. Here we consider the maximal eigenpair only.

The example below allows partially negative off-diagonal elements.

Example 8 ([9; Example (7)], [6; Example 12]) Let

$$A = \begin{pmatrix} -1 & 8 & -1 \\ 8 & 8 & 8 \\ -1 & 8 & 8 \end{pmatrix}.$$

Then The eigenvalues of A are as follows.

$$17.5124, \quad -7.4675, \quad 4.95513.$$

The corresponding maximal eigenvector is

$$(0.486078,\ 1.24981,\ 1)^*$$

which is positive.

Started at $z_0 = 24$, the outputs of our algorithms are given in Table 6.

Table 6. The outputs for a matrix with more negative elements

n	z_n: Algorithm 1	z_n: Algorithm 2
1	17.3772	18.5316
2	17.5124	17.5416
3		17.5124

Furthermore, we can even consider some complex matrices.

Example 9 ([10; Example 2.1], [6; Example 15]) Let

$$A = \begin{pmatrix} 0.75 - 1.125\,i & 0.5882 - 0.1471\,i & 1.0735 + 1.4191\,i \\ -0.5 - i & 2.1765 + 0.7059\,i & 2.1471 - 0.4118\,i \\ 2.75 - 0.125\,i & 0.5882 - 0.1471\,i & -0.9265 + 0.4191\,i \end{pmatrix},$$

where the coefficients are all accurate, to four decimal digits. Then A has eigenvalues

$$3, \quad -2 - i, \quad 1 + i$$

with maximal eigenvector

$$(0.408237, \ \ 0.816507, \ \ 0.408237)^*.$$

The outputs (y_n) (but not (z_n)) of [6; Algorithm 14], a variant of Algorithm 2, are as follows.

Table 7. The outputs for a complex matrix

y_1	y_2	y_3
$3.03949 - 0.0451599\,i$	$3.00471 - 0.0015769\,i$	3

We mention that a simple sufficient condition for the use of our algorithms is the following:

$$\mathrm{Re}(A^n) > 0 \text{ for large enough } n, \ \text{ up to a shift } mI. \tag{19}$$

Then we have the Perron–Frobenius property: there exists the maximal eigenvalue $\rho(A) > 0$ having simple left- and right-eigenvectors.

Hopefully, the reader would now be accept the use of "global" here for our new algorithms. They are very much efficient indeed. One may ask about the convergence speed of the algorithms. Even though we do not have a universal estimate for each model in such a general setup, it is known however that the shifted inverse algorithm is a fast cubic one, and hence should be fast enough in practice. This explains the reason why our algorithms are fast enough in the general setup. Certainly, in the tridiagonal dominate case, one can use the algorithms presented in the previous sections. Especially, in the tridiagonal

situation, we have analytically basic estimates which guarantee the efficiency of the algorithms. See [4] for a long way to reach the present level.

When talking about the eigenvalues, the first reaction for many people (at least for me, 30 years ago) is that well, we have known a great deal about the subject. However, it is not the trues. One may ask himself that for eigenvalues, how large matrix have you computed by hand? As far as I know, 2×2 only in analytic computation by hand. It is not so easy to compute them for a 3×3 matrix, except using computer. Even I have worked on the topic for about 30 years, I have not been brave enough to compute the maximal eigenvector, we use its mimic only to estimate the maximal eigenvalue (or more generally the first nontrivial eigenvalue). The first paper I wrote on the numerical computation is [3]. It is known that the most algorithms in computational mathematics are local, the Newton algorithm (which is a quadratic algorithm) for instance. Hence, our global algorithms are somehow unusual.

About three years ago, I heard a lecture that dealt with a circuit board optimization problem. The author uses the Newton method. I said it was too dangerous and could fall into the trap. The speaker answered me that yes, it is dangerous, but no one in the world can solve this problem. Can we try annealing algorithm? I asked. He replied that it was too slow. We all know that in the global optimization, a big problem (not yet cracked) is how to escape from the local traps. The story we are talking about today seems to have opened a small hole for algorithms and optimization problems, and perhaps you will be here to create a new field.

Acknowledgments. This paper is based on a series of talks: Central South U (2017/6), 2017 IMS-China, ICSP at Guangxi U for Nationalities (2017/6), Summer School on Stochastic Processes at BNU (2017/7), the 9th Summer Camp for Excellent College Students at BNU (2017/7), Sichun U (2017/7), the 12th International Conference on Queueing Theory and Network Applications at Yanshan U (2017/8), the 2nd Sino-Russian Seminar on Asymptotic Methods in Probability Theory and Mathematical Statistics & the 10th Probability Limit Theory and Statistic Large Sample Theory Seminar at Northeast Normal U (2017/9), Workshop on Stochastic Analysis and Statistical Physics at AMSS of CAS (2017//11), Yunnan U (2017/11). The author thanks professors Zhen-Ting Hou, Zai-Ming Liu, Zhen-Qing Chen, Elton P. Hsu, Jing Yang, Xiao-Jing Xu, An-Min Li, Lian-Gang Peng, Qian-Lin Li, Zhi-Dong Bai, Ning-Zhong Shi, Jian-Hua Guo, Zheng-Yan Lin, Zhi-Ming Ma and C. Newman et al, and Nian-Sheng Tang for their invitations and hospitality. The author also thanks Ms Jing-Yu Ma for the help in editing the paper. Research supported in part by National Natural Science Foundation of China (Grant Nos. 11626245, 11771046), the "985" project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Chen, M.F. (2005). *Eigenvalues, Inequalities, and Ergodic Theory.* Springer

[2] Chen, M.F. (2010). Speed of stability for birth–death processes. Front. Math. China 5(3), 379–515.
[3] Chen, M.F. (2016). *Efficient initials for computing the maximal eigenpair*. Front. Math. China 11(6): 1379–1418. A package based on the paper is available on CRAN now. One may check it through the link: https://cran.r-project.org/web/packages/EfficientMaxEigenpair/index.html
[4] Chen, M.F. (2017a). *The charming leading eigenpair*. Adv. Math. (China) 46 (4), 281–297.
[5] Chen, M.F. (2017b). *Efficient algorithm for principal eigenpair of discrete p-Laplacian*. Preprint.
[6] Chen, M.F. (2017c). *Global algorithms for maximal eigenpair*. Front. Math. China 12(5): 1023–1043.
[7] Golub, G.H., van der Vorst, H.A. (2000). *Eigenvalue computation in the 20th century*. J. Comp. Appl. Math. 123, 35C65.
[8] Langville, A.N. and Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.
[9] Noutsos, D. (2008). *Perron Frobenius theory and some extensions*. http://www.pdfdrive.net/perron-frobenius-theory-and-some-extensions-e10082439.html
[10] Noutsos, D. and Varga, R.S. (2012). *On the Perron–Frobenius theory for complex matrices*. Linear Algebra and its Applications 437, 1071–1088.
[11] Solomon, J. (2015). *Numerical Algorithms: Methods for Computer Vision, Machine Learning, and Graphics*. CRC Press, Boca Raton.

Mu-Fa Chen
School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.
E-mail: mfchen@bnu.edu.cn
Home page: http://math0.bnu.edu.cn/~chenmf/main_eng.htm

Front. Math. China 2018, 13(3): 509–524

Efficient algorithm for principal eigenpair of discrete p-Laplacian

Mu-Fa Chen

School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, China

Abstract

This paper is a continuation of the author's previous papers [Front. Math. China, 2016, 11(6): 1379–1418; 2017, 12(5): 1023–1043], where the linear case was studied. A shifted inverse iteration algorithm is introduced, as an acceleration of the inverse iteration which is often used in the non-linear context (the p-Laplacian operators for instance). Even though the algorithm is formally similar to the Rayleigh quotient iteration which is well-known in the linear situation, but they are essentially different. The point is that the standard Rayleigh quotient cannot be used as a shift in the non-linear setup. We have to employ a different quantity which has been obtained only recently. As a surprised gift, the explicit formulas for the algorithm restricted to the linear case ($p = 2$) is obtained, which improves the author's approximating procedure for the leading eigenvalues in different context, appeared in a group of publications. The paper begins with p-Laplacian, and is closed by the non-linear operators corresponding to the well-known Hardy-type inequalities.

2000 *Mathematics Subject Classification*: 34L15, 34G20, 39A12, 65F15
Key words and phrases. Discrete p-Laplacian, principal eigenpair, shifted inverse iteration, approximating procedure.

1 Introduction

Let $E = \{k \in \mathbb{Z}_+(\text{nonnegative intergers}) : k < N + 1\}\,(N \leqslant \infty)$. Given a positive sequence $\{\nu_k : k \in E\}$ with boundary condition $\nu_{-1} = 0$ and $p \in (1, \infty)$, the (weighted) discrete p-Laplacian operator Ω_p is defined as follows:

$$\Omega_p f(k) = \nu_k |f_{k+1} - f_k|^{p-2}\big(f_{k+1} - f_k\big) - \nu_{k-1}|f_k - f_{k-1}|^{p-2}\big(f_k - f_{k-1}\big),$$

or more simply,

$$\Omega_p f = \partial_{\bullet-1}\big(\varphi_\nu(\partial f)\big), \qquad f \in \mathscr{C}_K, \tag{1}$$

where $\mathscr{C}_K$ is the set of functions vanishing out of (M_1, M_2) for some $0 \leqslant M_1 < M_2 < N$ if $N = \infty$,

$$(\partial f)_k = \partial_k f = f_{k+1} - f_k \quad \text{and} \quad \big(\varphi_\nu(f)\big)_k = \nu_k |f_k|^{p-2} f_k = \nu_k |f_k|^{p-1} \operatorname{sgn}(f_k).$$

Throughout this paper, we are interesting in the principal eigenvalue λ_p with the boundary conditions $f_{-1} = f_0$ and $f_{N+1} = 0$ (which means $\lim_{n\to\infty} f_n = 0$ if $N = \infty$). That is, the smallest λ satisfying the following eigenequation

$$\Omega_p g = -\lambda \varphi_\mu(g) \quad \text{for some } \ g \neq 0, \tag{2}$$

where $(\mu_k : k \in E)$ is another given positive measure (weight). Actually, we are working in the $L^p(\mu)$ setup, the principal eigenvalue λ_p has an alternative description in the following classical variational formula

$$\lambda_p = \inf\big\{D_p(f) : \mu\big(|f|^p\big) = 1,\ f \in \mathscr{C}_K\big\}, \tag{3}$$

where $\mu(f) = \sum_{k\in E} \mu_k f_k$ and

$$D_p(f) = \big(-\Omega_p f, f\big) = \sum_{k\in E} \nu_k |f_{k+1} - f_k|^p, \quad p > 1, \qquad f \in \mathscr{C}_K$$

(cf. [6; p. 1263]).

This principal eigenvalue of p-Laplacian was the aim of the first part of [6] where the criterion for the positivity, the basic estimates, and the approximating procedures of the eigenvalue were presented. The main purpose of the present paper is introducing a new iteration algorithm (with efficient initials), which is much more efficient than the known ones. As a byproduct, we obtain an efficient iteration algorithm which is more efficient than the approximating procedures given in [2] in the linear case.

Algorithm 1 (Shifted inverse iteration) Given measures (μ_k), (ν_k) on E, and $p \in (1,\infty)$, define p^* as the conjugate of p: $1/p + 1/p^* = 1$ and set $\hat{\nu}_k = \nu_k^{1-p^*}$ for $k \in E$. Denote by (λ_p, g_p) the principal eigenpair (g_p is the eigenvector corresponding to λ_p). The algorithm is to construct an approximating sequence $\{(z^{(n)}, v^{(n)})\}_{n\geqslant 0}$ of (λ_p, g_p). In part (3) below, assume that $N < \infty$.

(1) Construction of $v^{(0)}$. Let $\tilde{v}^{(0)}$ denote the column vector

$$\bigg(\bigg(\sum_{j=k}^{N} \hat{\nu}_j\bigg)^{1/p^*} : k \in E\bigg).$$

Then define $v^{(0)} = \tilde{v}^{(0)}/\big\|\tilde{v}^{(0)}\big\|_{\mu,p}$, where $\|\cdot\|_{\mu,p}$ is the $L^p(\mu)$-norm.

(2) Construction of $z^{(0)}$. The construction works for general positive v with decreasing components. Set

$$\delta = \max_{i\in E}\left(\frac{1}{v_i}\sum_{j=i}^{N}\hat{\nu}_j\bigg[\sum_{k=0}^{j}\mu_k v_k^{p-1}\bigg]^{p^*-1}\right)^{p-1}, \qquad \bar{\delta} = \frac{\|v\|_{\mu,p}^p}{D_p(v)}. \tag{4}$$

If $1/\bar{\delta} - 1/\delta < 10^{-5}$ (say!), then define $z^{(0)} = \big(1/\delta + 1/\bar{\delta}\big)/2$ and stop the computation. At the same time, regard $\big(z^{(0)}, v^{(0)}\big)$ as an approximation of (λ_p, g_p). Otherwise, define $z^{(0)} = 1/\delta$, and then go to the next step.

Alternatively, without using $\bar{\delta}$, one can stop the computations at the nth iteration once $|z^{(n)} - z^{(n-1)}| < 10^{-5}$.

(3) Given $v := v^{(n-1)}$ and $z := z^{(n-1)}$, we are going to construct $v^{(n)}$. Define $w_s := w_s(x)$ successively:

$$w_s = x - \sum_{0\leqslant k\leqslant s-1} \left\{ \frac{1}{\nu_k} \sum_{j=0}^{k} \mu_j \big(v_j^{p-1} + z w_j^{p-1}\big) \right\}^{p^*-1}, \qquad 0 \leqslant s \leqslant N \tag{5}$$

here and in what follows, we adopt the convention $\sum_{\varnothing} = 0$ (then $w_0 = x$ by (5)), where x is a large enough root to the equation

$$(\nu_N - z\mu_N)\, w_N^{p-1} - z \sum_{j=0}^{N-1} \mu_j w_j^{p-1} = \sum_{j=0}^{N} \mu_j v_j^{p-1}. \tag{6}$$

We will come back to this point after the proof of this algorithm given in Section 3. It is the place we have to restrict ourselves to $N < \infty$, in the cases either $p \neq 2$ or $z^{(n)} \not\equiv 0$. Finally, define $v^{(n)} = w/\|w\|_{\mu,p}$.

(4) Repeat the step (2) (updating the superscript of $(z^{(0)}, v^{(0)})$) and step (3).

We now consider two special cases, for which the algorithm becomes simpler and even allow $N = \infty$ (in which case, one often needs some modification, refer to [2], but we omit the details here). The first one is ignoring "shift", the resulting algorithm is often used in the literature (see [1, 7] for instance), and it indeed coincides with approximating procedure given by [6; Theorem 2.4].

Algorithm 2 (Inverse iteration) Everything is the same as in Algorithm 1 except in part (3) of the algorithm, the parameter z is setting to be zero and the sequence $\{w_k\}_{k\geqslant 0}$ takes the following simple form

$$w_k = \sum_{\ell=k}^{N} \left[\frac{1}{\nu_\ell} \sum_{j=0}^{\ell} \mu_j v_j^{p-1} \right]^{p^*-1}, \qquad 0 \leqslant k \leqslant N. \tag{7}$$

The second special case is the linear one: $p = 2$. For which, the sequence $\{w_k\}$ used in part (3) of Algorithm 1 has an explicit construction.

Algorithm 3 (Shifted inverse iteration in linear case: $p = 2$) Everything is the same as Algorithm 1 except in part (3) of the algorithm, the sequence $\{w_k\}_{k\in E}$ is constructed as follows. Set

$$M_{sj} = \mu_j \sum_{k=j}^{s} \frac{1}{\nu_k}, \qquad 0 \leqslant j \leqslant s \leqslant N.$$

Define two independent sequences $\{A_s\}$ and $\{B_s\}$, recurrently:

$$\begin{cases} A_s = -\displaystyle\sum_{0\leqslant j\leqslant s-1} M_{s-1,j}(v_j + zA_j), \\ B_s = 1 - z \displaystyle\sum_{0\leqslant j\leqslant s-1} M_{s-1,j} B_j, \end{cases} \qquad 0 \leqslant s \leqslant N. \tag{8}$$

Then the required $\{w_k\}_{k\in E}$ can be expressed as $w_k = A_k + xB_k\,(k \in E)$, where

$$x = \left[\sum_{j=0}^{N} \mu_j(v_j + zA_j) - \nu_N A_N\right] \Big/ \left[\nu_N B_N - z\sum_{j=0}^{N} \mu_j B_j\right]. \tag{9}$$

We claim that the sequences $\{A_s\}$ and $\{B_s\}$ defined in Algorithm 3 have a unified explicit representation. For this, we need the following result.

Proposition 4 Given sequences $\{\bar{a}_s : s \in E\}$ and $\{\overline{M}_{sj} : s \in E,\, 0 \leqslant j \leqslant s\}$, the solution $\bar{A}_s := \bar{A}_s\big(\bar{a}, \overline{M}\big)$ to the equation

$$\bar{A}_s = \bar{a}_s + \sum_{0\leqslant j\leqslant s-1} \overline{M}_{s-1,\,j}\bar{A}_j, \qquad 0 \leqslant s \leqslant N \tag{10}$$

can be expressed as

$$\begin{aligned}\bar{A}_s\big(\bar{a}, \overline{M}\big) = \bar{a}_s &+ \sum_{0\leqslant j_1\leqslant s-1} \overline{M}_{s-1,\,j_1}\bar{a}_{j_1} + \sum_{1\leqslant j_1\leqslant s-1} \overline{M}_{s-1,\,j_1} \sum_{0\leqslant j_2\leqslant j_1-1} \overline{M}_{j_1-1,\,j_2}\bar{a}_{j_2} \\ &+ \cdots + \sum_{s-2\leqslant j_1\leqslant s-1} \overline{M}_{s-1,\,j_1} \cdots \sum_{0\leqslant j_{s-1}\leqslant 1} \overline{M}_{1,\,j_{s-1}}\bar{a}_{j_{s-1}} \\ &+ \sum_{s-1\leqslant j_1\leqslant s-1} \overline{M}_{s-1,\,j_1} \cdots \sum_{1\leqslant j_{s-1}\leqslant 1} \overline{M}_{1,\,j_{s-1}} \sum_{0\leqslant j_s\leqslant 0} \overline{M}_{0,\,j_s}\bar{a}_{j_s}. \end{aligned} \tag{11}$$

The right-hand side is a sum of $s+1$ terms, labeled as $k = 0, 1, \ldots, s$. The 0th term is simply $\bar{a}_s$. The $k\,(\geqslant 1)$th term is a k-multiple iterated sums with k parameters $(j_1, \ldots, j_k)$ over the region: $k-1 \leqslant j_1 \leqslant s-1$, $k-2 \leqslant j_2 \leqslant j_1 - 1$, …, $0 \leqslant j_k \leqslant j_{k-1} - 1$. In particular, the last term on the right-hand side is simply equal to

$$\overline{M}_{s-1,\,s-1} \cdots \overline{M}_{1,\,1}\overline{M}_{0,\,0}$$

once $s \geqslant 1$, and $= 0$ if $s = 0$.

To return to Algorithm 3, simply compare (10) with (8). Applying (11) to $\overline{M} = -zM$ and then setting

$$\bar{a}_s \equiv 1 \qquad \text{or} \qquad \bar{a}_s = -\sum_{0\leqslant j\leqslant s-1} M_{s-1,\,j}\, v_j,$$

we obtain, respectively, the sequences $\{B_s\}$ and $\{A_s\}$ as follows.

Corollary 5 For each $s \in E$, we have

$$\begin{aligned} B_s = 1 &+ (-z)\sum_{0\leqslant j_1\leqslant s-1} M_{s-1,\,j_1} + (-z)^2 \sum_{1\leqslant j_1\leqslant s-1} M_{s-1,\,j_1} \sum_{0\leqslant j_2\leqslant j_1-1} M_{j_1-1,\,j_2} \\ &+ \cdots + (-z)^{s-1} \sum_{s-2\leqslant j_1\leqslant s-1} M_{s-1,\,j_1} \cdots \sum_{0\leqslant j_{s-1}\leqslant 1} M_{1,\,j_{s-1}} \\ &+ (-z)^s \sum_{s-1\leqslant j_1\leqslant s-1} M_{s-1,\,j_1} \cdots \sum_{1\leqslant j_{s-1}\leqslant 1} M_{1,\,j_{s-1}} \sum_{0\leqslant j_s\leqslant 0} M_{0,\,j_s} \end{aligned}$$

and

$$A_s = -\sum_{0\leqslant j_1\leqslant s-1} M_{s-1,\,j_1} v_{j_1} - (-z)\sum_{1\leqslant j_1\leqslant s-1} M_{s-1,\,j_1}\sum_{0\leqslant j_2\leqslant j_1-1} M_{j_1-1,\,j_2} v_{j_2}$$
$$-\cdots-(-z)^{s-2}\sum_{s-2\leqslant j_1\leqslant s-1} M_{s-1,\,j_1}\cdots\sum_{0\leqslant j_{s-1}\leqslant 1} M_{1,\,j_{s-1}} v_{j_{s-1}}$$
$$-(-z)^{s-1}\sum_{s-1\leqslant j_1\leqslant s-1} M_{s-1,\,j_1}\cdots\sum_{1\leqslant j_{s-1}\leqslant 1} M_{1,\,j_{s-1}}\sum_{0\leqslant j_s\leqslant 0} M_{0,\,j_s} v_{j_s}.$$

Clearly, the recurrent formulas (8) is more practical in the numerical computation than the explicit ones given in Corollary 5. Similarly, we can represent the construction of $\{w_k\}$ given in Algorithms 1 and 2 in recurrent form. Consider for instance the sequence $\{w_k\}$ defined by (5) with variable $x > 0$. It can be computed in terms of $\{V_k\}$ and $\{W_k\}$, recurrently, as follows. Start at $V_0 = v_0^{p-1}$, $w_0 = x$, and $W_0 = zx^{p-1}$. For $k = 1, 2, \ldots, N$, let

$$\begin{cases} V_k = V_{k-1} + \mu_k v_k^{p-1}, \\ w_k = w_{k-1} - \left[\nu_{k-1}^{-1}(V_{k-1} + W_{k-1})\right]^{p^*-1}, \\ W_k = W_{k-1} + \mu_k w_k^{p-1}. \end{cases}$$

For finite N, the convergence of $(z^{(n)}, v^{(n)})$ to the principal eigenpair in Algorithm 2, in the linear context, was proved in [4; Proposition 23]. The same conclusion holds also for general p, refer to [8]. The shifted algorithm does not disturb but accelerates the convergence.

In the linear case, Algorithm 2 coincides with the author's approximating procedure ([2; Theorem 3.2]), refer also to [4; Proof of Proposition 23]. Algorithm 3 improves considerably the convergence speed in a number of contexts, see for instance [2]. Algorithm 3 also simplifies [4; Proposition 9].

In the next section, we illustrate the application of the results presented in Section 1 by a simple example. The proofs of the main results are given in Section 3. The extension of the results to a more general setup is delayed to the last section of the paper.

2 An example

Consider the following non-trivial example:

$$N = 7,\ \ E = \{0, 1, \ldots, 7\},\ \ \mu_k = 20^k \text{ and } \nu_k = 20^{k+1} \text{ for each } k \in E.$$

This is a particular case of [6; Example 2.6].

Before moving to the details, let us mention a simple fact. Note that if we replace v by cv for some constant $c > 0$, then the outputs $\left(\delta, \bar{\delta}\right)$ given in part (2) of Algorithm 1 remain the same. From the proof of Algorithm 1 given in the next section (equation (13), more precisely), it will be clear that

the resulting w will become cw in the iteration, hence such a change does not make influence to $\left(\delta, \bar{\delta}\right)$. In other words, without loss of generality, we may and will omit in this section the normalization procedure $w/\|w\|_{\mu,p}$. We simply use $v^{(n)}$ to denote the output (the approximation of the eigenvector) after the nth iteration. However, in general, such normalization cannot be ignored, otherwise, in the case that one needs a large number of iterations, the initial $v_0^{(n)}$ of $v^{(n)}$ may increase or decrease rapidly. Then normalization is helpful. Besides, the simplest normalization would be $v^{(n)}/v_0^{(n)}$ (rather than $v^{(n)}/\|v^{(n)}\|_{\mu,p}$) instead of $v^{(n)}$ at the nth (each) step.

We now state the numerical results for this example using the three algorithms presented in Section 1, respectively. We start at the linear case.

Case I (Linear case: $p = 2$). The outputs $1/\delta^{(n)}$ and $1/\bar{\delta}^{(n)}$ at the nth iteration by Algorithm 3 are given in Table 1.

Table 1

n	$z^{(n)} = 1/\delta^{(n)}$	$1/\bar{\delta}^{(n)}$
0	12.0878	13.0275
1	12.4955	12.5679
2	12.5623	$12.5637 = \lambda_p$
3	12.5637	12.5637

Clearly, the iterations can stop at step 3, according to Algorithm 1 (2). In the present linear case actually, when $n \geqslant 1$, we use $1/\bar{\delta}^{(n)}$ as the output $z^{(n)}$ (cf. [4]), and so we can actually stop the computation at step 2. We remark that in the present case $N = 7$, without using shift, we need 16 iterations to achieve at the six precisely significant digits,

For large N, the problem becomes more serious. The following results are taken from [8]. Using the Rayleigh quotient iteration ([4; §3]), we get the result as in Table 2. While using the ordinary inverse iteration, to arrive at the same accuracy as in Table 2, the number n of iterations we need is as in Table 3.

Table 2

$N+1$	$z^{(0)} = 1/\delta_1^{-1}$	$z^{(1)}$	$z^{(2)} = \lambda_p$
50	12.0557	12.0721	12.0719
100	12.0557	12.0600	12.0599
150	12.0557	12.0577	12.0576
200	12.0557	12.0568	12.0568

Table 3

$N+1$	50	100	150	200
n	243	1240	1783	2163

We have thus shown the serious difference of the inverse algorithms with/without shift.

Case II (Inverse iteration). Let $p = 3$, then the output $1/\bar{\delta}^{(n)}$ at the nth iteration by Algorithm 2 is given as in Table 4. Thus, we can stop the computation at the 16th step.

Table 4

n	$1/\bar{\delta}^{(n)}$
0	5.90161
1	5.78915
2	5.75709
3	5.7465
4	5.74274
5	5.74132
6	5.74075
7	5.7405
8	5.74038
9	5.74031
10	5.74028
11	5.74026
12	5.74025
13	5.74024
14	5.74024
15	5.74024
16	$5.74023 = \lambda_p$

Case III (Shifted inverse iteration). Let $p = 3$. Then the outputs $1/\delta^{(n)}$ and $1/\bar{\delta}^{(n)}$ at the nth iteration by Algorithm 1 are given in Table 5. The iterations can stop at step 3.

Table 5

n	$z^{(n)} = 1/\delta^{(n)}$	$1/\bar{\delta}^{(n)}$
0	5.20417	6.42897
1	5.6652	5.74679
2	5.73842	$5.74023 = \lambda_p$
3	5.74023	5.74023

Comparing the last two cases, the inverse algorithm is much easier in practice. The convergence at the beginning is quick enough, but then the convergence speed becomes slower and slower. The shifted inverse iteration converges much quicker but it required a harder computation in looking for a root of a nonlinear equation. This suggests us in practice to use the inverse iteration for a few of steps and then move to the shifted one. Further comments on the shifted inverse iteration are presented in the subsequent sections.

3 Proofs

We begin this section with some general preparations. First, we discuss δ and $\bar{\delta}$ used in part (2) of Algorithm 1. The next result is basic in the present study.

Lemma 6 For each positive v, the quantities δ and $\bar{\delta}$ defined in part (1) of Algorithm 1 provide the following estimates of λ_p:

$$\delta^{-1} \leqslant \lambda_p \leqslant \bar{\delta}^{-1}.$$

Proof. The upper estimate comes from (3), the lower one comes from [6; Theorem 2.1 (2)]. □

From the above lemma, it is clear that a good choice of v is essential for our purpose. For part (1) of Algorithm 1, the specific $v^{(0)}$ is just the function f_1 used in part (1) of [6; Theorem 2.4]. Now, part (2) of Algorithm 1 is based on Lemma 6. To explain the meaning of part (3) of Algorithm 1, we need more preparation.

The ordinary inverse iteration says that at the nth iteration, for given $v^{(n-1)}$, define $w^{(n)}$ by:

$$-\Omega_p w^{(n)} = \varphi_\mu\big(v^{(n-1)}\big).$$

The key point is that, instead of the original operator $-\Omega_p$, we use a shifted one:

$$-\Omega_p w^{(n)} - z^{(n-1)}\varphi_\mu\big(w^{(n)}\big) = \varphi_\mu\big(v^{(n-1)}\big). \tag{12}$$

Comparing this with the eigenequation (2): a shift term is added on the left-hand side and the constant λ is ignored on the right-hand side. Making a sum of the both sides of (12) over the set $\{0, 1, \ldots, k\}$ $(k \leqslant N)$ and using (1) with $\nu_{-1} = 0$, it follows that

$$\begin{aligned} &-\nu_k|\partial_k w^{(n)}|^{p-2}(\partial_k w^{(n)}) - z^{(n-1)}\sum_{j=0}^{k}\mu_j\big|w_j^{(n)}\big|^{p-2}w_j^{(n)} \\ &\quad = \sum_{j=0}^{k}\mu_j\big|v_j^{(n-1)}\big|^{p-2}v_j^{(n-1)}, \qquad w_{N+1}^{(n)} := 0,\ 0 \leqslant k \leqslant N. \end{aligned}$$

By [6; Propositions 3.1 and 3.2], the eigenvector g_p corresponding to λ_p should have positive and decreasing components, hence we can assume so do $v^{(n-1)}$ and $w^{(n)}$. Thus, we can rewrite the last equation as

$$\nu_k\big(-\partial_k w^{(n)}\big)^{p-1} - z^{(n-1)}\sum_{j=0}^{k}\mu_j\big(w_j^{(n)}\big)^{p-1} = \sum_{j=0}^{k}\mu_j\big(v_j^{(n-1)}\big)^{p-1}, \qquad 0 \leqslant k \leqslant N.$$

Omitting the superscript n everywhere for simplicity, we obtain

$$\nu_k(-\partial_k w)^{p-1} - z\sum_{j=0}^{k}\mu_j w_j^{p-1} = \sum_{j=0}^{k}\mu_j v_j^{p-1}, \qquad 0 \leqslant k \leqslant N. \tag{13}$$

That is,

$$-\partial_k w = \left\{\frac{1}{\nu_k}\sum_{j=0}^{k}\mu_j\big(v_j^{p-1} + z w_j^{p-1}\big)\right\}^{p^*-1}, \qquad w_{N+1} := 0,\ 0 \leqslant k \leqslant N. \tag{14}$$

Proof of Algorithm 1

Summing up in k of the both sides of (14) over $\{0, 1, \ldots, s\}$ $(s \leqslant N-1)$, with $x := w_0$, it follows that

$$w_{s+1} = x - \sum_{k=0}^{s} \left\{ \frac{1}{\nu_k} \sum_{j=0}^{k} \mu_j \big(v_j^{p-1} + z w_j^{p-1} \big) \right\}^{p^*-1}, \qquad 0 \leqslant s \leqslant N-1.$$

Or equivalently (5) holds. Starting from (5), one can express w_k $(0 \leqslant k \leqslant N)$, step by step, as a function of x. On the other hand, by (14), we have

$$\nu_N w_N^{p-1} = \sum_{j=0}^{N} \mu_j \big(v_j^{p-1} + z w_j^{p-1} \big),$$

or equivalently (6) holds. Solving equation (6), we obtain a root $x\,(= w_0)$. Then, we obtain the other solutions w_k $(1 \leqslant k \leqslant N)$ in terms of (5). □

Here is a technical point to find a suitable root x to (6). Actually, we do not need all of the roots (there are usually more than one), but what instead is looking for a large enough positive root of (6) only. Now, there is a question about the starting point in seeking for the root (using FindRoot in Mathematica for instance, an incorrect starting point will lead to an unrelated root since the algorithm is local as usual, and then the iterations may go to a pitfall). For this purpose, we should provide a meaningful estimate of $w_0 = x$. To do so, let $\bar{\xi}$ and $\underline{\xi}$ denote the upper and lower bounds of λ_p (which are $1/\bar{\delta}$ and $1/\delta$, respectively, in the original context), respectively, obtained by using Lemma 6. The left-hand side of (12) is approximately $(\lambda_p - z)\mu_k w_k^{p-1}$, where $z = \underline{\xi}$. Thus, according to (12), we have

$$(\lambda_p - z)\mu_k w_k^{p-1} \approx \mu_k v_k^{p-1},$$

Equivalently,

$$w_k \approx \frac{v_k}{(\lambda_p - z)^{p^*-1}}.$$

In particular,

$$w_0 \approx \frac{v_0}{(\lambda_p - z)^{p^*-1}}.$$

Now, the problem is that λ_p is unknown. Of which, an approximation of λ_p is choosing to be $\alpha\bar{\xi} + (1-\alpha)\underline{\xi}$ for some $\alpha \in (0,1)$. We have thus made the choice of the initial point x_0 in searching the solution x to equation (6):

$$x_0 = v_0 \big[\alpha \big(\bar{\xi} - \underline{\xi} \big) \big]^{1-p^*}. \tag{15}$$

For instance, in Case III given in Section 2, at the first iteration, we choose $\alpha = 1/8$ and then use $\alpha = 2/3$ at the subsequent iterations. One may have a look at the result computed for Case III in Section 2.

Generally speaking, the choice of α depends on the model, but not on its size N (see for instance [4; Table 2']). Hence, we can choose α first from a smaller N and then apply it to the other larger N. Actually, there is a room for the choice of α, it can be quite rough. The main point is that one should choose such an α so that the resulting x_0 is large enough, bigger than x, as illustrated in Table 6.

Table 6

n	1	2	3
x_0	1.11459	2.12296	51.0808
x	0.495123	1.77869	41.7535

Our shifted inverse iteration is quite similar to the well-known Rayleigh quotient iteration often used in the linear situation. However, they are essentially different. The point is that, in the linear case, the Rayleigh quotient $D_p(v)/\|v\|_{\mu,p}^p$ used as a shift z in each iteration, cannot be used in the non-linear case (i.e. $p \neq 2$) for constructing $v^{(n)}$ whenever $n \geqslant 1$, since then the equation (6) often has no positive root (its roots are often complex, because of the non-linearity). The root x we need to be positive since the vector w with $w_0 = x$ is regarded as an approximation of the maximal eigenvector which should be positive. This seems the main reason that the shifted inverse iteration algorithm has not appeared in the non-linear context, as far as we know.

Proof of Algorithm 2

Summing up in k of the both sides of (14) with $z = 0$ over $\{s, \ldots, N\}$ with $w_{N+1} = 0$, we obtain the required assertion. □

We remark that (7) coincides with [6; (10)] if (v, w) is replaced by (f, g). In other words, we can rewrite (7) as $w = v\Pi(v)^{p^*-1}$ in terms of the operator Π defined in [6; §2]. This means that the inverse iteration (Algorithm 2) coincides with [6; Theorem 2.4 (1)], as mentioned before. Refer also to [5; §A.4, Step 6] for more details in the linear case.

Proof of Algorithm 3

Applying (5) to $p = 2$ and exchanging the order of the sums, we can rewrite w_s as

$$w_s = x - \sum_{0 \leqslant j \leqslant s-1} M_{s-1,j}(zw_j + v_j), \qquad 0 \leqslant s \leqslant N$$

with the convention $\sum_{\varnothing} := 0$ again. Expressing

$$w_s = A_s + xB_s, \qquad 0 \leqslant s \leqslant N, \tag{16}$$

then we obtain the iteration formulas (8). Having the sequence $\{w_s\}_{s=0}^N$ (with one variable x only) at hand, we can use (6), i.e.,

$$(\nu_N - z\mu_N)w_N - z\sum_{j=0}^{N-1} \mu_j w_j = \sum_{j=0}^{N} \mu_j v_j$$

to get the required solution (9) and then using (8) and (16) to obtain the solution $\{w_k\}$. □

Proof of Proposition 4 In view of (10), we certainly have $\bar{A}_0 = \bar{a}_0$. Recurrently, for $s \geqslant 1$, we have

$$\bar{A}_s = \bar{a}_s + \sum_{0 \leqslant j_1 \leqslant s-1} \overline{M}_{s-1,\, j_1} \Big[\bar{a}_{j_1} + \sum_{0 \leqslant j_2 \leqslant j_1 - 1} \bar{A}_{j_2} \Big]. \tag{17}$$

In particular, if $s = 1$, then

$$\bar{A}_1 = \bar{a}_1 + \overline{M}_{00}\, \bar{a}_0.$$

Otherwise assume that $s \geqslant 2$. Note that if $j_1 = 0$, then in (17), the second sum in j_2 can be ignored. Thus, at the first recurrent step, we indeed have

$$\bar{A}_s = \bar{a}_s + \sum_{1 \leqslant j_1 \leqslant s-1} \overline{M}_{s-1,\, j_1} \Big[\bar{a}_{j_1} + \sum_{0 \leqslant j_2 \leqslant j_1 - 1} \bar{A}_{j_2} \Big].$$

Continuing the recurrent procedure (or by induction), it is not difficult to prove the proposition. □

4 Extension

In this section, we extend Algorithms 1 and 2 to a more general setup. That is studying the principal eigenpair corresponding to the following eigenequation (an extension of (2), ignoring the constant λ):

$$\Omega_p g = -\varphi_{\mu,q}(g) \quad \text{for some} \;\; g \neq 0, \tag{18}$$

where $p, q \in (1, \infty)$ and

$$\big(\varphi_{\mu,q}(f)\big)_k = \mu_k |f_k|^{q-2} f_k = \mu_k |f_k|^{q-1} \mathrm{sgn}\,(f_k).$$

More precisely, the principal eigenvalue we are interested is

$$\lambda_{p,q} = \inf_{f \in \mathscr{C}_K,\, f \neq 0} \frac{\|\partial_\bullet f\|_{\nu,p}}{\|f\|_{\mu,q}}$$

(cf. [3; Proposition 4.7]).

Here is our first main result in this section.

Algorithm 7 (Shifted inverse iteration) Given measures (μ_k), (ν_k) on E, and $p, q \in (1, \infty)$ with $q \geqslant p$, set $\hat{\nu}_k = \nu_k^{1-p^*}$ for $k \in E$. Denote by $(g_{p,q}, \lambda_{p,q})$ the principal eigenpair described by (18). The algorithm is to construct an approximating sequence $\big\{\big(v^{(n)}, z^{(n)}\big)\big\}_{n \geqslant 0}$ of $(g_{p,q}, \lambda_{p,q})$. In part (3) below, assume that $N < \infty$.

(1) Construction of $v^{(0)}$. Let $\tilde{v}^{(0)}$ denote the column vector

$$\left(\left(\sum_{j=k}^{N}\hat{\nu}_j\right)^{1/\tilde{p}^*}: k\in E\right),\qquad \tilde{p}^*:=\frac{p^*}{q}+1.$$

Then define $v^{(0)}=\tilde{v}^{(0)}/\|\tilde{v}^{(0)}\|_{\mu,q}$, where $\|\cdot\|_{\mu,q}$ is the $L^q(\mu)$-norm.

(2) Construction of z for a given general positive v with decreasing components. Set

$$\delta=\max_{i\in E}\left(\frac{1}{v_i}\sum_{j=i}^{N}\hat{\nu}_j\Big[\sum_{k=0}^{j}\mu_k v_k^{q/p^*}\Big]^{p^*/q}\right)^{1/p^*},\qquad \bar{\delta}=\frac{\|v\|_{\mu,q}}{\|\partial_\bullet(v)\|_{\nu,p}}. \tag{19}$$

When $v=v^{(n)}$, write $(\delta,\bar{\delta})$ as $(\delta^{(n)},\bar{\delta}^{(n)})$. We can stop the computation at the nth iteration once

$$1/\bar{\delta}^{(n-1)}-1/\bar{\delta}^{(n)}<10^{-5},$$

then define $z^{(n)}=1/\bar{\delta}^{(n)}$. Otherwise, define $z^{(n)}=\max\{1/\delta^{(n)},z^{(n-1)}\}$, and then go to the next step.

(3) Given $v:=v^{(n-1)}$ and $z:=z^{(n-1)}$, we are going to construct $v^{(n)}$. For this, define $w_s:=w_s(x)$ successively:

$$w_s=x-\sum_{0\leqslant k\leqslant s-1}\left\{\frac{1}{\nu_k}\sum_{j=0}^{k}\mu_j\big(v_j^{q-1}+zw_j^{q-1}\big)\right\}^{p^*-1},\qquad 0\leqslant s\leqslant N, \tag{20}$$

where x is a large enough root of the equation

$$x=\sum_{k=0}^{N}\left\{\frac{1}{\nu_k}\sum_{j=0}^{k}\mu_j\big(v_j^{q-1}+zw_j^{q-1}\big)\right\}^{p^*-1}. \tag{21}$$

Finally, define $v^{(n)}=w/\|w\|_{\mu,q}$.

(4) Repeat the step (2) (updating the superscript of $(z^{(0)},v^{(0)})$) and step (3).

The main difference of Algorithms 7 and 1 is the quantity δ in part (2). Here in Algorithm 7, we do not use the pair (p,q) directly, but instead use a single $\tilde{p}:=q/p^*+1$ (as indicated in part (1) of Algorithm 7), returning to the setup of Section 1. When $q=p$, we go back again to the setup of Section 1. Otherwise, when $q>p$, the story is different as indicated in part (2) of Algorithm 7:

$$z^{(n)}=\max\{1/\delta^{(n)},z^{(n-1)}\}.$$

However, since the shifted operator and the original one have the same eigenvector (up to a constant), the vectors $\{v^{(n)}\}$ constructed by the shifted algorithm do converge to the principal eigenvector $g_{p,q}$ and so at the same time, the sequence $\{1/\bar{\delta}^{(n)}\}$ should converge to $\lambda_{p,q}$. Recall that the use of shifts is

to accelerate the convergence speed, the z more closer to $\lambda_{p,q}$ makes the convergence quicker, hence it is valuable to use a suitable combination of $1/\delta^{(n)}$ and $1/\bar{\delta}^{(n)}$ as an accelerated shift $z^{(n)}$ (has to be located below $\lambda_{p,q}$ in the present non-linear situation), as we did several times in [4]. We will come back to this point in Remark 10 (3) below.

Similar to Algorithm 2, the algorithm becomes much simpler once the "shift" is ignored.

Algorithm 8 (Inverse iteration) Everything is the same as in Algorithm 7 except in part (3) of the algorithm, the parameter z is setting to be zero and the sequence $\{w_k\}_{k\geqslant 0}$ takes the form:

$$w_k = \sum_{\ell=k}^{N} \left[\frac{1}{\nu_\ell} \sum_{j=0}^{\ell} \mu_j v_j^{q-1} \right]^{p^*-1}, \qquad 0 \leqslant k \leqslant N.$$

To illustrate the application of the above results, let us return to the example studied in Section 2.

Example 9 Fix $p = 3$ and $q = 4$. Then to achieve at the six precisely significant digits,

- for the inverse iteration, we need six iterations; and
- for the shifted inverse iteration, we need only three iterations.

The outputs are given in Table 7.

Table 7

n	$1/\bar{\delta}^{(n)}$
0	2.6306
1	2.27309
2	2.2586
3	2.25736
4	2.25726
5	2.25725
6	2.25725

n	$1/\bar{\delta}^{(n)}$
0	2.6306
1	2.27187
2	2.25725
3	2.25725

Here are three remarks on the shifted inverse iteration.

Remark 10 (1) When $q \neq p$, due to the inhomogeneous problem of the shifted operator, unlike the inverse iteration or the algorithms given in Section 1 for the p-homogeneous case, here the normalization procedure $v^{(n)} = w^{(n)}/\|w^{(n)}\|_{\mu,q}$ is necessary for each $n \geqslant 0$.

(2) As in part (3) of Algorithm 1, in part (3) of Algorithm 7, we need to specify the initial point x_0 in finding the suitable root of x. For simplicity, as in §3, set $\bar{\xi} = 1/\bar{\delta}$ and $\underline{\xi} = 1/\delta$. Then, as an analog of (15), we have

$$x_0 = v_0\big[\alpha\big(\bar{\xi} - \underline{\xi}\big)\big]^{1/(1-q)}.$$

We can even ignore v_0 here

$$x_0 = \big[\alpha\big(\bar{\xi} - \underline{\xi}\,\big)\big]^{1/(1-q)}, \tag{22}$$

since for normalized v, its v_0 is close to 1. As usual, the choice of α depends on the model. The specific choice for our model is $\alpha = 0.18$ at the first step and then $\alpha = 0.1$ for subsequent iterations. We will come back to this point again soon.

(3) In part (2) of Algorithm 1, the sequence $\{1/\delta^{(n)}\}$ increases in n to λ_p. However, the increasing property happens in Algorithm 7 only at the first of few steps, and then it goes decreasingly. The reason is as follows. On the one hand, the sequence $\{v^{(n)}\}$ converges to the principal eigenfunction $g_{p,q}$, and hence $\{1/\bar{\delta}^{(n)}\}$ converges to $\lambda_{p,q}$, as $n \to \infty$. On the other hand, if we denote by $\{\tilde{v}^{(n)}\}$ the sequence from the inverse iterations given in Algorithm 2 with $p = \tilde{p}$, and replacing v by $\tilde{v}$ and p by $\tilde{p}$ in the definition of δ in (4), we obtain

$$\begin{aligned}
\tilde{\delta} &:= \max_{i\in E}\left(\frac{1}{\tilde{v}_i}\sum_{j=i}^{N}\hat{\nu}_j\Big[\sum_{k=0}^{j}\mu_k\tilde{v}_k^{\tilde{p}-1}\Big]^{\tilde{p}^*-1}\right)^{\tilde{p}-1}\\
&= \max_{i\in E}\left(\frac{1}{\tilde{v}_i}\sum_{j=i}^{N}\hat{\nu}_j\Big[\sum_{k=0}^{j}\mu_k\tilde{v}_k^{q/p^*}\Big]^{p^*/q}\right)^{q/p^*}\\
&= \left\{\max_{i\in E}\left(\frac{1}{\tilde{v}_i}\sum_{j=i}^{N}\hat{\nu}_j\Big[\sum_{k=0}^{j}\mu_k\tilde{v}_k^{q/p^*}\Big]^{p^*/q}\right)^{1/p^*}\right\}^{q}.
\end{aligned}$$

Once we set $\tilde{v} = v$, we return to (19) except an additional power q which comes from the difference of the definition of λ_p and $\lambda_{p,q}$. Anyhow, we understand where the quantity δ in (19) comes from and the difference between (4) and (19) (cf. [3]). Clearly, except at the initial point $n = 0$, $\{v^{(n)}\}$ is different from $\{\tilde{v}^{(n)}\}$ if $q \neq p$. Remember that $\{v^{(n)}\}$ and $\{\tilde{v}^{(n)}\}$ converge to the eigenvectors $g_{p,q}$ and $g_{\tilde{p}}$, respectively, and have the same initial $v^{(0)} = \tilde{v}^{(0)}$. Hence, up to some n_0, $v^{(k)}$ improves not only $v^{(k-1)}$ but also $\tilde{v}^{(k-1)}$ for $k \leqslant n_0$. Therefore, $\{1/\delta^{(k)}\}$ is increasing up to n_0. Then $v^{(n)}$ becomes close and close to $g_{p,q}$ and hence away from $g_{\tilde{p}}$ step by step. In other words, $\{1/\delta^{(n)}\}$ turns to be decreasing after n_0. Therefore, the sequence $\{1/\delta^{(n)}\}$ is not monotone, but unimodal. This is the reason why we adopt

$$z^{(n)} = \max\{1/\delta^{(n)}, z^{(n-1)}\},$$

but not $1/\delta^{(n)}$ only in Algorithm 7 (2). As proved in [3], the limits of $\big\{(1/\tilde{\delta}^{(n)})^{1/q}\big\}$ and $\{1/\bar{\delta}^{(n)}\}$ are quite closed to each other. Actually, the last two sequences were used in [3; Theorem 2.2] for the upper/lower estimates of $\lambda_{p,q}^{-1}$.

Let us present more details in the computation of the second part of Example 9 to explain the above idea (Table 8).

Table 8

n	$1/\delta^{(n)}$	$z^{(n)}$	$1/\bar{\delta}^{(n)}$	$\alpha^{(n)}$	$x_0^{(n)}$	$x^{(n)}$
0	2.16948	2.16948	2.6306	0.18	2.29248	0.288079
1	2.21445	2.21445	2.27187	0.1	5.58444	5.0765
2	2.17016	2.21445	2.25725	0.1	6.15897	5.07653
3	2.16986		2.25725			

In words, the computation goes as follows. Suppose we are given $v = v^{(n)}$, then we can use the formulas given in the first two parts of Algorithm 7 to compute $1/\delta$, $1/\bar{\delta}$ and determine z (ignoring the upper scripts for simplicity). Then, in order to define the next v, at the beginning step, we choose $\alpha = 0.18$, and in the subsequent steps, choose $\alpha = 0.1$. There is a technical point here. Since the lower estimate $\underline{\xi} = z$ used in (22) is smaller than the real lower bound determined by v, we choose a smaller number α here avoiding a modification of $\underline{\xi}$. It is often meaningful (and there is some freedom) to choose α so that the resulting x_0 is bigger (even two or three times bigger) than the root x we required. Otherwise, we should use a smaller α. Once, α is chosen, we can compute x_0, x, and then the sequence $\{w_k : k \in E\}$. Finally, define $v^{(n+1)} = w/\|w\|_{\mu,q}$ and repeat the above computations. Note that as n increases, we obtain a unimodal sequence $1/\delta^{(n)}$ with maximal point achieved at $n = 1$. The reason was explained in the previous paragraph. Hence we have $z^{(1)} = z^{(2)} > 1/\delta^{(2)}$. Since the output of $1/\bar{\delta}^{(n)}$ at $n = 3$ coincides with it at $n = 2$, we do not need to do anymore, and so the computation is stopped at this step.

From the above discussion, it is clear that we can apply Algorithm 1, replacing (μ, ν, p) by $(\mu, \bar{\nu}, \tilde{p})$ with

$$\tilde{p} = \frac{q}{p^*} + 1, \qquad \bar{\nu}_k = \nu_k^{(p^*-1)/(\tilde{p}^*-1)}, \quad k \in E$$

to compute $\lambda_{\tilde{p}}$, and then set $z^{(n)} \equiv \lambda_{\tilde{p}}^{1/q}$ in the application of Algorithm 7. The reason of the change $\nu \to \bar{\nu}$ is what we used in (19) for δ is $\hat{\nu} = \nu^{1-p^*}$ rather than $\hat{\nu} = \nu^{1-\tilde{p}^*}$ required in Algorithm 1 with $p = \tilde{p}$. For Example 9, $\lambda_{\tilde{p}}^{1/q} \approx 2.23211$. However, such a choice may not be economical in practice.

Proof of Algorithm 7 Recall the eigenequation (18):

$$\Omega_p w^{(n)} = -\varphi_{\mu,q}(v^{(n-1)}).$$

By adding a shifted term, we obtain the so-called shifted inverse iteration as follows.

$$\Omega_p w^{(n)} + z^{(n-1)}\varphi_{\mu,q}(w^{(n)}) = -\varphi_{\mu,q}(v^{(n-1)}).$$

Next, by [9; Proposition 2.1], we may and will assume that both $v = v^{(n-1)}$ and $w := w^{(n)}$ are positive and decreasing. Hence as an analog of (14), we

have

$$-\partial_k w = \left\{ \frac{1}{\nu_k} \sum_{j=0}^{k} \mu_j \big(v_j^{q-1} + z w_j^{q-1} \big) \right\}^{p^*-1}, \qquad w_{N+1} := 0,\ 0 \leqslant k \leqslant N. \tag{23}$$

From this, we obtain (20). Combining (20) and (23) at N, we obtain (21). We have thus completed the construction of the shifted algorithm. □

The proof of Algorithm 8 is almost the same as the one for Algorithm 2.

As usual, the results presented in this paper should be meaningful in the continuous context.

Acknowledgments The author thanks Yue-Shuang Li's contribution in the earlier stage of looking for the new algorithm, especially a lot of work on computer checking. Thanks are also given to Zhong-Wei Liao for his corrections on the earlier version of the paper. The author acknowledges the referees for their careful comments and corrections. Research supported in part by National Natural Science Foundation of China (Nos. 11626245, 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Biezuner, R.J., Ercole, G. and Martins, E.M. (2009). *Computing the first eigenvalue of the p-Laplacian via the inverse power method*. J. Funct. Anal. 257: 243–270.

[2] Chen, M.F. (2010). *Speed of stability for birth–death processes*. Front Math China 5(3): 379–515.

[3] Chen, M.F. (2015). *The optimal constant in Hardy-type inequalities*. Acta Math. Sin., Eng. Ser. 31(5): 731–754.

[4] Chen, M.F. (2016). *Efficient initials for computing the maximal eigenpair*. Front. Math. China 11(6): 1379–1418. A package based on the paper is available on CRAN now. One may check it through the link:
https://cran.r-project.org/web/packages/EfficientMaxEigenpair/index.html
A MatLab package is also available, see the author's homepage.

[5] Chen, M.F. (2017). *Global algorithms for maximal eigenpair*. Front. Math. China 12(5): 1023–1043.

[6] Chen, M.F., Wang, L.D., and Zhang, Y.H. (2014). *Mixed eigenvalues of discrete p -Laplacian*. Front. Math. China, 9(6): 1261–1292.

[7] Ercole, G. (2015). *An inverse iteration method for obtaining q-eigenpairs of the p-Laplacian in a general bounded domain*. Mathematics, 2015.

[8] Li, Y.S. (2017). *The inverse iteration method for discrete weighted p-Laplacian* (in Chinese). Master's thesis at Beijing Normal University.

[9] Liao, Z.W. (2016). *Discrete weighted Hardy inequalities with different kinds of boundary conditions*. Acta Math. Sin. Eng. Ser. 32(9): 993-1013.

Front. Math. China 2018, 13(6): 1267–1311
DOI 10.1007/s11464-018-0716-x

Hermitizable, Isospectral Complex Matrices or Differential Operators

Mu-Fa Chen

(Beijing Normal University)

January 1, 2018

Abstract

The main purpose of the paper is looking for a larger class of matrices which have real spectrum. The first well-known class having this property is the symmetric one, then is the Hermite one. This paper introduces a new class, called Hermitizable matrices. The closely related isospectral problem, not only for matrices but also for differential operators is also studied. The paper provides a way to describe the discrete spectrum, at least for tridiagonal matrices or one-dimensional differential operators. Especially, an unexpected result in the paper says that each Hermitizable matrix is isospectral to a birth–death type matrix (having positive subdiagonal elements, in the irreducible case for instance). Besides, new efficient algorithms are proposed for computing the maximal eigenpairs of these class of matrices.

2000 *Mathematics Subject Classification*: 15A18, 34L05, 35P05, 37A30
Key words and phrases. Real spectrum, symmetrizable, Hermitizable, isospectral, matrix, differential operator.

1 Introduction

In the study of the submaximal eigenvalue computation, we learn that the complex spectrum is harder to handle than the real one. This leads to looking for a larger class of matrices having real spectrum. Certainly, the problem is meaningful not only in several branches of mathematics but also for quantum mechanics. In the context of real matrices, one knows that the class of symmetric matrices has this property. Actually, in probability theory, we have also known that an extended class, called symmetrizable matrices has the same property. For the last class, it was restricted to those matrices having nonnegative off-diagonal elements (the restriction is natural in the scope of probability theory, since A is regarded, for instance, as a formal generator of a Markov chain). Hence, it is up to now still an open question about the symmetrizability for real matrices. Next, in the context of complex matrices, it is again well known that the Hermite matrices have real spectrum. Thus, it is natural to study the so-called Hermitizable class, as an extension of the Hermitian one. At the same time, we study the isospectral problem for matrices. As a byproduct, we update remarkably some algorithms published earlier,

for computing the maximal eigenpair. Besides, we also examine the problems for the second-order differential operators. Combining the results obtained here with [8], it provides a way to describe the discrete spectrum, at least for complex tridiagonal matrices or one-dimensional differential operators.

To have a quick understanding of what is going on in the paper, let us consider the simplest typical case: the tridiagonal matrices

$$A = \begin{pmatrix} -c_0 & b_0 & & & & 0 \\ a_1 & -c_1 & b_1 & & & \\ & a_2 & -c_2 & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-1} & -c_{N-1} & b_{N-1} \\ 0 & & & & a_N & -c_N \end{pmatrix},$$

where N is fixed to be finite at the moment. In the context of real matrices, the symmetry means that $b_k = a_{k+1}$ for each $k : 0 \leqslant k < N$. Without loss of generality, one may assume that $a_k \neq 0$ for each $k : 1 \leqslant k \leqslant N$. In the symmetrizable case (i.e. there exists positive (μ_i) such that $\mu_i b_i = \mu_{i+1} a_{i+1}$ for every i), the subdiagonals (a_k) and (b_k) can be rather arbitrary except both of them were assumed to be positive. In the context of complex matrices, the Hermitizability means that there exists positive (μ_i) such that $\mu_i b_i = \mu_{i+1} \bar{a}_{i+1}$ for every i. This holds iff three conditions hold simultaneously:

1) $a_{k+1} \neq 0$ for $k : 1 \leqslant k \leqslant N$;
2) the ratio $b_k/\bar{a}_{k+1} > 0$ for each $k : 0 \leqslant k < N$;
3) the diagonals $(c_k)_{0 \leqslant k \leqslant N}$ are real.

In each of these three cases (symmetric, symmetrizable, or Hermitizable), the spectrum of A is real. It is much more surprising that one can reduce the Hermitizable case to the symmetrizable one, especially for computing the maximal eigenpair, developed in [9–12]. Refer to [12; §2] for a short survey, in particular. To the last topic, new efficient algorithms are presented, especially for non-symmetric tridiagonal matrices with large size.

The paper is organized as follows. In Section 2, we present an extended and improved circle criterion for the Hermitizability and an invariance of this property under the so-called h-transform, which is an important addition to [15]. In Section 3, we mainly concentrate on tridiagonal matrices, especially on the h-transform. The main result is quite unexpected, it reduces the present complex situation to the well understood symmetrizable case for computing the maximal eigenpair developed in [9–12]. As a continuation of Section 3, new algorithms for the computation of the maximal eigenpairs, especially for non-symmetric matrices with large size are presented in Section 4. In Section 5, we study the symmetrizable and isospectral problems briefly for second-order differential operators.

2 Hermitizable complex matrices

In this section, we study the Hermitizable complex matrices which consist of an important and operable extension of the ordinary Hermite matrices. At the same time, we study in a general setup an isospectral transform (called h-transform) and the invariance of the Hermitizability under the transform.

Let $A = (a_{ij})$ be a given complex matrix on a countable (finite or infinite) set E.

Definition 1 The matrix $A = (a_{ij} : i, j \in E)$ is called **Hermitizable** (or **complex symmetrizable**) if there exists a positive measure $(\mu_i : i \in E)$ such that

$$\mu_i a_{ij} = \mu_j \bar{a}_{ji}, \qquad i, j \in E, \tag{1}$$

where $\bar{a}$ denotes the conjugate of a.

In other words, a complex matrix $A = (a_{ij})$ is Hermitizable if there exists $\mu = (\mu_i)$ such that the matrix $(\mu_i a_{ij} : i, j \in E)$ becomes Hermite, even though A itself may not be so unless μ is the uniformly distributed measure. Thus, the existence of a Hermitizable measure μ is essential in this context. Once the measure μ is at hand, one may call A Hermite, or symmetric, or more often selfadjoint on the Hilbert space $L^2(\mu)$ of complex functions. In the context of real matrix with nonnegative off-diagonal elements, it is usually called symmetrizable (or symmetric on $L^2(\mu)$ once μ is known). See for instance [5; Chapter 7]. The next result is somehow standard.

Proposition 2 (1) A complex matrix A is Hermitizable with respect to μ iff, as an operator, A is selfadjoint (Hermitian, symmetric) on the space $L^2(\mu)$ of complex functions:

$$(Af, g)_\mu = (f, Ag)_\mu, \qquad f, g \in \mathscr{C}_K, \tag{2}$$

where $\mathscr{C}_K$ is the set of functions with finite support and $(\cdot, \cdot)_\mu$ is the usual $L^2(\mu)$-inner product:

$$(f, g)_\mu = \int_E f \bar{g} \mathrm{d}\mu.$$

(2) If so, the spectrum of A in $L^2(\mu)$ is real.

Proof. The following simple proof may be helpful for the new comers to the topic. Denote by $\mathrm{Diag}\,(h)$ the diagonal matrix having (h_k) as its diagonals. Rewrite the pointwise formula (1) as the matrix form

$$\mathrm{Diag}\,(\mu)A = A^H \mathrm{Diag}\,(\mu) \left[= \big(\mathrm{Diag}\,(\mu)A\big)^H \right],$$

where A^H is the conjugate and transpose of A: $\bar{A}^*$. Then

$$A = \mathrm{Diag}\,(\mu)^{-1} A^H \mathrm{Diag}\,(\mu). \tag{3}$$

This means that A and A^H have the same spectrum and hence which should be real. We have thus proved part (2) of Proposition 2. The standard way to prove that part (2) is implied by part (1) of the proposition goes as follows. Let λ be a L^2-eigenvalue of A, then there exists $g \in L^2(\mu)$, $g \neq 0$ such that $Ag = \lambda g$. Hence

$$(Ag, g)_\mu = \lambda(g, g)_\mu, \qquad (g, Ag)_\mu = \bar{\lambda}(g, g)_\mu.$$

Therefore, $\lambda = \bar{\lambda}$ which shows that λ must be real. By the way, in our study on leading eigenvalues, it is more convenient to allow $g \in L^1(\mu)$ rather than $g \in L^2(\mu)$, especially for infinite E. See [6] for more details and additional reference, in particular [6; Proposition 3.5].

By setting $f = \mathbb{1}_{\{j\}}$ and $g = \mathbb{1}_{\{i\}}$, it follows that (2)⇒(1). We now prove that (1)⇒(2), or equivalently, (3)⇒(2). Note that

$$(f, g)_\mu = g^H \mathrm{Diag}\,(\mu) f.$$

We have

$$\begin{aligned}
(Af,\, g)_\mu &= g^H \mathrm{Diag}\,(\mu) Af \\
&= g^H \mathrm{Diag}\,(\mu) \mathrm{Diag}\,(\mu)^{-1} A^H \mathrm{Diag}\,(\mu) f \quad \text{(by (3))} \\
&= g^H A^H \mathrm{Diag}\,(\mu) f \\
&= (Ag)^H \mathrm{Diag}\,(\mu) f \\
&= \big(\mathrm{Diag}\,(\mu) f\big)^* \overline{Ag} \\
&= (f,\, Ag)_\mu.
\end{aligned}$$

We have thus proved that (3)⇒(2) and then (1)⟺(2). □

For infinite E, as an operator on $L^2(\mu)$, one has to take care of the domain of A, a suitable extension of $\mathscr{C}_K$ for instance.

Clearly, each Hermitizable matrix satisfies the following conditions.

Lemma 3 Each complex symmetrizable matrix $A = (a_{ij})$ should have the following properties.

(1) **Co-zero property**: $a_{ij} = 0$ iff $a_{ji} = 0$ for every pair $i, j \in E$, $i \neq j$.
(2) **Positive ratio property**: $a_{ij}/\bar{a}_{ji} > 0$ once $a_{ij} \neq 0$ for each pair $i, j \in E$.

Moreover, these properties can be combined into a single one:

(3) either both a_{ij} and a_{ji} are zero, or $a_{ij}a_{ji} > 0$.

We remark that by the lemma, the diagonal elements $(a_{ii} : i \in E)$ must be real: either $a_{ii} = 0$, or $a_{ii}/\bar{a}_{ii} > 0$. Throughout the paper, the fractions are assumed to be irreducible to avoid confusion.

Definition 4 Let $i \neq j$ and write $i \to j$ if $a_{ij} \neq 0$. If

$$i_0 \to i_1 \to \cdots \to i_n \to i_{n+1},$$

then it is called a **path** from i_0 to i_{n+1}, denoted by $i_0 \rightsquigarrow i_{n+1}$. We say that A is **irreducible** if for every pair (i, j), $i \neq j$, we have $i \rightsquigarrow j$ (and so does $j \rightsquigarrow i$).

Note that if we make a convention: $i \rightsquigarrow i$ for every $i \in E$, then the relation '$\rightsquigarrow$' is an equivalence one, in view of Lemma 3. Hence we can divide the space E into irreducible subclasses (subsets), and then study the symmetrizable problem of the submatrices on each irreducible subset, separately. Thus, for simplicity, in what follows, we may restrict ourselves to one irreducible class. Therefore, each symmetrizable matrix A is endowed with a graph structure with bonds '$i \to j$', and each submatrix owns a subgraph structure.

Now, assume that $i_0 \to i_1 \to \cdots \to i_n \to i_{n+1}$. Obviously, we can rewrite (1) as

$$\mu_{i_0} \frac{a_{i_0 i_1}}{\bar{a}_{i_1 i_0}} = \mu_{i_1}.$$

Multiplying both sides by $a_{i_1 i_2}/\bar{a}_{i_2 i_1}$ and using the equality above replacing (i_0, i_1) by (i_1, i_2), we obtain

$$\mu_{i_0} \frac{a_{i_0 i_1}}{\bar{a}_{i_1 i_0}} \cdot \frac{a_{i_1 i_2}}{\bar{a}_{i_2 i_1}} = \mu_{i_1} \frac{a_{i_1 i_2}}{\bar{a}_{i_2 i_1}} = \mu_{i_2}.$$

Successively, we have

$$\mu_{i_0} \frac{a_{i_0 i_1}}{\bar{a}_{i_1 i_0}} \frac{a_{i_1 i_2}}{\bar{a}_{i_2 i_1}} \cdots \frac{a_{i_n i_{n+1}}}{\bar{a}_{i_{n+1} i_n}} = \mu_{i_{n+1}}. \tag{4}$$

By setting $i_{n+1} = i_0$, we have obtained an extended **Kolomogrov circle theorem** ([19] where the result was proved for transition probabilities of Markov chains with finite space):

$$\frac{a_{i_0 i_1}}{\bar{a}_{i_1 i_0}} \frac{a_{i_1 i_2}}{\bar{a}_{i_2 i_1}} \cdots \frac{a_{i_n i_0}}{\bar{a}_{i_0 i_n}} = 1 \tag{5}$$

for each closed path (circle) $i_0 \to i_1 \to \cdots \to i_n \to i_0$.

Due to the positive ratio property of A, each fraction on the left-hand side of (5) is positive. Note that for the degenerated closed path, the round-trip path $i_0 \to i_1 \to i_2 \to i_1 \to i_0$ for instance, condition (5) is trivial, and so from now on, unless otherwise stated, we ignore it in examining the circle condition (5). We can apply the argument given in [5; §7.1] to obtain the following result.

Theorem 5 (Improved circle theorem) Under the necessary conditions given in Lemma 3, a matrix A is Hermitizable iff (5) holds for each closed path (equivalently, each smallest (non-cross-) closed path). If so, regard i_0 as a

reference point and set $\mu_{i_0} = 1$. Then, for each i_{n+1} belonging to the connect subgraph of A containing i_0, $\mu_{i_{n+1}}$ can be computed by using (4) along a shortest path $i_0 \to \ldots \to i_{n+1}$. The computations for the other μ_k in different sub-graphs are similar.

A simple illustration of the meaning of the present extension goes as follows. Suppose that $A = (a_{ij})$ is a real and Hermitizable (i.e. symmetrizable) matrix. If we change the sign of an arbitrary number of pairs (a_{ij}, a_{ji}) simultaneously, then the resulted matrix is again symmetrizable, even with the same symmetrizing measure (μ_k).

This result goes back to [17] and [23; Chapter 6]. For a matrix A with non-negative off-diagonal elements, Theorem 5 has a lot of powerful applications, refer to [5; §7.2 and Chapter 11] for simple criteria and more original references (even extended into uncountable spaces). Actually, it was the first tool we used to statistical mechanics: distinguishing the equilibrium and non-equilibrium systems. Refer to the survey articles [13, 14] for a long story along this research direction. We recall that Kolmogorov's [19, 20] was motivated from Schrödinger's [24] (1931), who discovered the "wave equation" in 1926. Hence, the present study on the complex operators with real spectrum may be meaningful in quantum mechanics.

The power of Theorem 5 is especially based on its use on the geometric (graphic) structure of the graph. Regarding E as the set of vertexes. Define the edges to be the set of pairs (i, j) with $a_{ij} \neq 0$. Then we have $i \to j$ and $j \to i$ and furthermore paths, as stated in Definition 4. We have thus obtained a graph structure. For simplicity, we may assume that the graph is connected.

Recalling that the round-trip paths can be ignored, the next result is obvious. To be precise, we say that $i_0 \to i_1 \to \cdots \to i_n \to i_0$ is a real circle if the elements in $\{i_0, i_1, \cdots, i_n\}$ are all different.

Corollary 6 If the graph of A contains no real circle, then it is Hermitizable iff the properties listed in Lemma 3 hold.

Next, we introduce a (conservative) potential field structure, the 'work', to the graph: the work done by the graph along the path

$$i_0 \to i_1 \to \cdots \to i_{n+1},$$

denoted by

$$w(i_0 \to i_1 \to \cdots \to i_{n+1}),$$

is defined by

$$\log\left(\frac{a_{i_0 i_1}}{\bar{a}_{i_1 i_0}} \cdots \frac{a_{i_n i_{n+1}}}{\bar{a}_{i_{n+1} i_n}}\right).$$

Clearly, we have the additive property:

$$w(i_0 \to i_1 \to \cdots \to i_{n+1}) = w(i_0 \to i_1 \to \cdots \to i_k) + w(i_k \to i_{k+1} \cdots \to i_{n+1}).$$

Now, (5) means that the work done by the graph along each closed path equals zero:

$$w(i_0 \to i_1 \to \cdots \to i_n \to i_0) = 0. \tag{6}$$

Having the field structure at hand, the proof of Theorem 5 is more or less the same as those presented in [17], [23; Chapter 6], or [5; Chapter 7]. Here we sketch some points only.

Sketch of the proof of Theorem 5 First, we have proved that the conditions listed in Lemma 3 plus the circle condition (5) are necessary for the Hermitizability (1). Next, under these conditions, we have not only the field structure, but also the path-independence (6). Due to this, we can define a potential V on the graph. Assume that the graph is connected for simplicity. Fix a reference point i_0 and set $V_{i_0} = 0$. For each $j \neq i_0$, choose and fix a path from i_0 to j, denoted by $L_{i_0 j}$. Then define V_j to be the work done by the field along the path $L_{i_0 j}$:

$$V_j = V_j - V_{i_0} = w(L_{i_0 j}).$$

We have thus defined a potential function V on the graph, which is actually independent of the specific choice of path $L_{i_0 j}$. By using the path-independence again, for every pair $\{i, j\}\,(i \neq j)$, we have

$$w(L_{i_0 i}) + w(i \to j) = w(L_{i_0 j}).$$

Hence

$$\log \frac{a_{ij}}{\bar{a}_{ji}} = w(i \to j) = w(L_{i_0 j}) - w(L_{i_0 i}) = V_j - V_i.$$

Therefore,

$$e^{V_i} a_{ij} = e^{V_j} \bar{a}_{ji}.$$

This gives us (1) with $\mu_k = e^{V_k}$. We have thus proved that the conditions given in Lemma 3 plus the circle condition (5) are sufficient for (1).

To show that the minimal closed path condition stated in Theorem 5 is indeed sufficient, let us consider a simple (random chosen) example, Figure 1. We check that

$$w(0 \to 3 \to 2 \to 1 \to 0) = 0. \tag{7}$$

This closed circle evolves two smallest closed circles:

$$0 \to 3 \to 2 \to 0 \quad \text{and} \quad 0 \to 2 \to 1 \to 0.$$

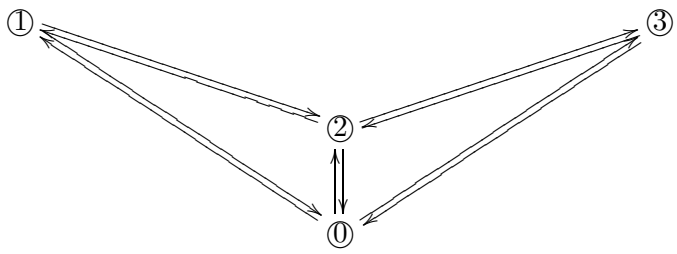


Figure 1 Circles of a connected graph

Theorem 5 says that for having (7), it suffices to assume that

$$w(0 \to 3 \to 2 \to 0) = 0 \quad \text{and} \quad w(0 \to 2 \to 1 \to 0) = 0.$$

By the additive property, these two conditions imply that

$$w(0 \to 3 \to 2 \to 0 \to 2 \to 1 \to 0) = 0. \tag{8}$$

Noting that

$$\frac{a_{02}}{\bar{a}_{20}} > 0 \implies \text{conjugate of } \frac{a_{02}}{\bar{a}_{20}} = \frac{\bar{a}_{02}}{a_{20}} > 0,$$

we have

$$w(2 \to 0 \to 2) = \log \frac{a_{20}}{\bar{a}_{02}} + \log \frac{a_{02}}{\bar{a}_{20}} = 0.$$

Hence by using the additive property again, we can remove the round-trip path $2 \to 0 \to 2$ in (8), and then obtain the required assertion (7). □

Note that for the graph here, even though a closed path can be rather complex, we need to handle with triangles and quadrilaterals only. This is actually meaningful in general (refer to [5; §7.2 and Chapter 11] again).

To illustrate the use of Theorem 5, we consider a particular example.

Example 7 Let

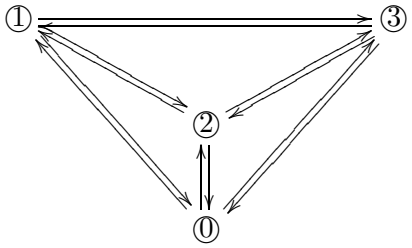
$$A = \begin{pmatrix} -6 & (8-6\,i)/5 & (8+14\,i)/13 & (18+4\,i)/17 \\ 3+9\,i/4 & -55/4 & (-5+40\,i)/13 & (30+35\,i)/17 \\ (12-21\,i)/5 & (-4-32\,i)/5 & -13 & (60-66\,i)/17 \\ (63-14\,i)/10 & (84-98\,i)/15 & (70+77\,i)/13 & -16 \end{pmatrix}$$

This matrix is Hermitizable and so has real eigenvalues:

$$-21.3806,\ -17.7581,\ -9.44576,\ -0.165558.$$

Besides, the Hermitizable measure μ is

$$\mu_0 = 1,\ \mu_1 = \frac{8}{15},\ \mu_2 = \frac{10}{39},\ \mu_3 = \frac{20}{119}.$$



Three smallest circles :
$0 \to 1 \to 2 \to 0$,
$0 \to 3 \to 2 \to 0$,
$1 \to 2 \to 3 \to 1$.

Figure 2 Heart figure

Proof. The co-zero property is obvious since the matrix contains no zero elements. Thus, we need only to check the other two conditions: the positive ratio and the circle conditions.

To do so, we need some simplification. Denote by r_{ij} the ratio $a_{ij}/\bar{a}_{ji}$. Then we need to check only six bonds $\langle i,j\rangle$ since $r_{ij} > 0$ iff $r_{ji} > 0$, even though there are altogether twelve oriented bonds for the graph (see Fig. 2). Note that each bond belongs to one of the circles, we need only to consider the bonds of the circles. Otherwise, the remainder bonds will have to be treated separately for the positivity of ratio. Now,

$$r_{01} = \frac{8}{15},\ r_{12} = \frac{25}{52},\ r_{20} = \frac{39}{10},\ r_{03} = \frac{20}{119},\ r_{32} = \frac{119}{78},\ r_{31} = \frac{238}{75}.$$

Having these ratios at hand, it is rather easy to check the circle condition for the three smallest circles and to compute the measure μ in terms of the formula: $\mu_0 = 1$, $\mu_k = \mu_{k-1} r_{k-1,k}$. □

The second aim of this paper is to study the isospectral operators (matrices, in particular), as a continuation of [15, 8]. For this, we need to copy a result from [15; Lemma 1.3 and Remark 1.4].

Lemma 8 Let $(E, \mathscr{E}, \mu)$ be a measure space and let h be Lebesgue measurable: $E \to \mathbb{C}$, $h \neq 0$, μ-a.e. Then the following assertions hold.

(1) The mapping $\tilde{f} := \mathbb{1}_{[h\neq 0]} f/h$ is an isometry from $L^2(E,\mu)$ to $L^2(E,\tilde{\mu})$ (complex), where $\tilde{\mu} = |h|^2\mu$.

(2) Let L be an operator on $L^2(E,\mu)$ with domain $\mathscr{D}(L)$. Define an operator $\widetilde{L}$ as follows:

$$\widetilde{L}\tilde{f} = \mathbb{1}_{[h\neq 0]}\frac{1}{h}L(\tilde{f}h), \qquad \mathscr{D}(\widetilde{L}) = \{\tilde{f} \in \mathscr{E} : \tilde{f}h \in \mathscr{D}(L)\}. \tag{9}$$

Then the operators $(L, \mathscr{D}(L))$ on $L^2(E,\mu)$ and $(\widetilde{L}, \mathscr{D}(\widetilde{L}))$ on $L^2(E,\tilde{\mu})$ are isospectral (say L and $\widetilde{L}$ are L^2-isospectral, for short) in the following sense:

$$(Lf, f)_\mu = (\widetilde{L}\tilde{f}, \tilde{f})_{\tilde{\mu}}, \qquad f \in \mathscr{D}(L).$$

(3) If additionally, $h \in \mathscr{D}(L)$, then for fixed $B \in \mathscr{E}$, $\widetilde{L}\mathbb{1} = 0$, $\tilde{\mu}$-a.e. on B iff $Lh = 0$, μ-a.e. on B.

Example 7(Continued) As an application of Lemma 8, let h denote the vector having $h_3 = 1$ (i.e. $B = \{0,1,2\}$ in Lemma 8 (3)). Solving the equation

$$A^{\setminus \text{the last row}} h = 0,$$

where $A^{\setminus \text{the last row}}$ is the matrix obtained from A by removing its last row, we obtain

$$h_0 = \frac{9+2\,i}{17},\ h_1 = \frac{6+7\,i}{17},\ h_2 = \frac{10-11\,i}{17},\ h_3 = 1.$$

Next, define (an alternative expression of (9))

$$\tilde{A} := \text{Diag}\,(h)^{-1} A\,\text{Diag}\,(h),$$

where Diag (h) is the diagonal matrix having vector h as its diagonal elements. Then, we have

$$\tilde{A} = \begin{pmatrix} -6 & 2 & 2 & 2 \\ 15/4 & -55/4 & 5 & 5 \\ 3 & 4 & -13 & 6 \\ 7/2 & 14/3 & 7 & -16 \end{pmatrix}. \tag{10}$$

Note that each row of this real matrix is conservative (i.e. the sum of the row equals zero) except the last one. By the above lemma, A and $\tilde{A}$ have the same spectrum but in general the real $\tilde{A}$ is much more convenient in application (refer to [9 – 12] for instance).

Here a new question arrives. Is the transformed matrix $\tilde{A}$ Hermitizable? Note that the measures μ and $\tilde{\mu}$ in the transformation defined by Lemma 8 are often different. The transformation is designed in a general setup, as used a lot in [9; Section 4], not necessary for Hermitizable ones. We remark that Lemma 8 goes from a complicated L (containing a potential or a killing term, for instance) to a simpler $\widetilde{L}$ (without potential or killing term). In this case, we need not only the first two conditions of Lemma 8, but also its third one: assuming h to be somehow L-harmonic. Sometimes, we go to the opposite direction: from $\widetilde{L}$ to L:

$$Lf = h\widetilde{L}\left(\frac{f}{h}\right)$$

as we will see several times subsequently. In this case, we do not need the third condition of Lemma 8. Thus, in what follows, unless otherwise stated, we call the one defined by the first two conditions of Lemma 8 an h-transform. The next result is an important addition to [15].

Theorem 9 The selfadjointness

$$(Lf, g)_\mu = (f, Lg)_\mu, \qquad f, g \in \mathscr{D}(L) \subset L^2(\mu)$$

is invariant under the h-transform.

Proof. (a) Define a multiplying operator H as follows.

$$(Hf)(x) = h(x)f(x) \qquad \text{for every } x$$

(where H plays the same role as Diag (h) used in the discrete context). Then the h-transform defined in Lemma 8 can be expressed by

$$\widetilde{L} = H^{-1}LH.$$

Part (2) of Lemma 8 says, as proved in [15], that with $\tilde{f} = f/h$ (equivalently, $H\tilde{f} = f$), we have

$$(Lf, f)_\mu = \big(\widetilde{L}\tilde{f}, \tilde{f}\big)_{\tilde{\mu}}.$$

Let us repeat the proof of this conclusion here in terms of the operator H.

$$\big(\widetilde{L}\tilde{f}, \tilde{f}\big)_{\tilde{\mu}} = \big(H^{-1}LH\tilde{f}, \tilde{f}\big)_{\tilde{\mu}} = \big(\bar{H}^{-1}H^{-1}LH\tilde{f}, H\tilde{f}\big)_{\tilde{\mu}} = \big(|H|^{-2}Lf, f\big)_{\tilde{\mu}} = (Lf, f)_\mu.$$

(b) By the way, due to the conjugate property of inner product, it is obvious that

$$\big(\widetilde{L}\tilde{f}, \tilde{f}\big)_{\tilde{\mu}} = (Lf, f)_\mu \Longleftrightarrow \big(\tilde{f}, \widetilde{L}\tilde{f}\big)_{\tilde{\mu}} = (f, Lf)_\mu.$$

(c) To go to the selfadjointness, recall the polarization identity

$$(f, g)_\mu = \frac{1}{4}\big(\|f + g\|_\mu^2 - \|f - g\|_\mu^2 + i\|f + ig\|_\mu^2 - i\|f - ig\|_\mu^2\big).$$

Since L is a linear operator, we may set $f_1 = Lf$, $g_1 = Lg$, and write

$$(L(f + g), f + g)_\mu = (f_1 + g_1, f + g)_\mu.$$

Expressing in the same way the other terms on the right-hand side of the next formula, one may check the following identity:

$$\begin{aligned}(Lf, g)_\mu = \frac{1}{4}\big[&(L(f + g), f + g)_\mu - \big(L(f - g), f - g\big)_\mu \\ &+ i\big(L(f + ig), f + ig\big)_\mu - i\big(L(f - ig), f - ig\big)_\mu\big].\end{aligned}$$

Roughly speaking, if we regard (Lf, g) as a new bivariate functional of f and g, say $\langle f, g\rangle$ for instance: linear in f and conjugate linear in g, then the present identity can be read out from the previous one for $(f, g)_\mu$. Similarly, we have the identity for $\big(\widetilde{L}\tilde{f}, \tilde{g}\big)_{\tilde{\mu}}$. Since the right-hand side is expressed by the diagonal elements of the quadratic form $(Lf, g)_\mu$, and correspondingly for $\big(\widetilde{L}\tilde{f}, \tilde{g}\big)_{\tilde{\mu}}$, by proof (a), we obtain

$$(Lf, g)_\mu = \big(\widetilde{L}\tilde{f}, \tilde{g}\big)_{\tilde{\mu}}.$$

By exchanging f and g (correspondingly, $\tilde{f}$ and $\tilde{g}$) and using the property used in proof (b), we also obtain

$$(f, Lg)_\mu = \big(\tilde{f}, \widetilde{L}\tilde{g}\big)_{\tilde{\mu}}.$$

The last two identities are more than enough for the required assertion. □

By the way, we study an extension of Lemma 8 and Theorem 9. As in Lemma 8, let L be a linear operator in $L^2(E, \mu)$ with domain $\mathscr{D}(L)$, and let M be an invertible linear one. Define a new inner product

$$\langle f, g\rangle = (Mf, Mg)_\mu.$$

Then we have an inner product space, and furthermore a Hilbert space $\mathscr{H}(E, \langle\cdot, \cdot\rangle)$.

Theorem 10 Let L, M, and $\mathscr{H}(E,\langle\cdot,\cdot\rangle)$ be defined as above. Then the following assertions hold.

(1) The mapping $\tilde{f} := M^{-1}f$ is an isometry from $L^2(E,\mu)$ to $\mathscr{H}(E,\langle\cdot,\cdot\rangle)$.

(2) Define an operator $\widetilde{L}$ as follows:

$$\widetilde{L}\tilde{f} = M^{-1}LMf, \qquad \mathscr{D}\big(\widetilde{L}\big) = \big\{\tilde{f} \in \mathscr{E} : M\tilde{f} \in \mathscr{D}(L)\big\}. \tag{11}$$

Then the operators $(L,\mathscr{D}(L))$ on $L^2(E,\mu)$ and $\big(\widetilde{L},\mathscr{D}\big(\widetilde{L}\big)\big)$ on $\mathscr{H}(E,\langle\cdot,\cdot\rangle)$ are isospectral in the following sense:

$$(Lf,f)_\mu = \big\langle \widetilde{L}\tilde{f},\tilde{f}\big\rangle, \qquad f \in \mathscr{D}(L).$$

(3) The operator L is selfadjoint on $L^2(E,\mu)$ iff so does $\widetilde{L}$ on $\mathscr{H}(E,\langle\cdot,\cdot\rangle)$.

Proof. The proof is quite similar to the one for Theorem 9. Recall that $f = M\tilde{f}$. First, we have

$$(f,g)_\mu = \big(M\tilde{f},\, M\tilde{g}\big)_\mu = \big\langle \tilde{f},\, \tilde{g}\big\rangle.$$

Next,

$$(Lf,g)_\mu = \big(LM\tilde{f},M\tilde{g}\big)_\mu = \big(MM^{-1}LM\tilde{f},M\tilde{g}\big)_\mu \overset{(11)}{=} \big(M\widetilde{L}\tilde{f},M\tilde{g}\big)_\mu = \big\langle \widetilde{L}\tilde{f},\tilde{g}\big\rangle.$$

We have thus proved the first two parts of the theorem. Then the third one follows by the same idea used in the proof of Theorem 9: from diagonal case to the general one of the quadratic forms. □

By Theorem 9, the transformed matrix $\tilde{A}$ given in (10) is Hermitizable. We have seen that this matrix $\tilde{A}$ has a nice property: its off-diagonal elements are nonnegative. For this, it is necessary that the diagonal elements of A should be enough negative. If not, one may replace the original A by a shift one:

$$A_1 = A - mI, \quad m := \sup_k \left(a_{kk} + \sum_{j\neq k} |a_{kj}|\right)^+, \quad x^+ := \max\{x,0\}. \tag{12}$$

For real A, this was used in [9–12] for computing the maximal eigenpair. For complex A, this is based on the well-known Gershgorin Circle Theorem (cf. [27]) used to bound the spectrum of a square matrix. For finite matrix, there is no problem for $m < \infty$ even for arbitrary (c_k). However, it is often a restriction for infinite matrix. In which case, one may adopt some approximation procedure to avoid $m = \infty$. The quantity in (12) can be sharp in the special case: the off-diagonal elements are nonnegative and

$$\begin{cases} a_{kk} + \sum_{j\neq k} a_{kj} = m, & 0 \leqslant k < N \text{ and} \\ a_{NN} + \sum_{j\neq N} a_{Nj} \leqslant m & \text{if } N < \infty. \end{cases}$$

In this case, the constant m is exactly the shift we required. As will be seen from Example 18 below, there may have a little room for improvement. However, since the estimate m for a necessary shift is often not exact, it should be permitted to have a smaller (generally speaking, for safe, a bigger one is safer) perturbation so that m becomes simpler (integer, for instance).

Example 7(Continued) We now illustrate the effectiveness (12) by the present example. Even though we have seen the given $\{a_{ii}\}$ are negative enough, which is however not known in advance, it is meaningful to have a test about condition (12). For this, since

$$\sup_k \left(a_{kk} + \sum_{j \neq k} |a_{kj}| \right) = 7.0634,$$

we choose $m = 7$. Then

$$A_1 = \begin{pmatrix} -6-7 & (8-6\,i)/5 & (8+14\,i)/13 & (18+4\,i)/17 \\ 3+9\,i/4 & -55/4-7 & (-5+40\,i)/13 & (30+35\,i)/17 \\ (12-21\,i)/5 & (-4-32\,i)/5 & -13-7 & (60-66\,i)/17 \\ (63-14\,i)/10 & (84-98\,i)/15 & (70+77\,i)/13 & -16-7 \end{pmatrix}$$

The corresponding h becomes

$$h_0 = \frac{24102+5356\,i}{163217},\ h_1 = \frac{22620+26390\,i}{163217},\ h_2 = \frac{40360-44396\,i}{163217},\ h_3 = 1.$$

The h-transform of A_1,

$$\tilde{A}_1 := \mathrm{Diag}\,(h)^{-1} A_1\, \mathrm{Diag}\,(h)$$

becomes

$$\tilde{A}_1 = \begin{pmatrix} -13 & 290/103 & 4036/1339 & 9601/1339 \\ 309/116 & -83/4 & 2018/377 & 9601/754 \\ 4017/2018 & 3770/1009 & -20 & 28803/2018 \\ 9373/9601 & 52780/28803 & 28252/9601 & -23 \end{pmatrix}.$$

Clearly, the off-diagonal elements are nonnegative, each of the first three rows is conservative but not the last row. We have arrived at the same structure of real matrix as $\tilde{A}$.

We have seen how the matrix A can be isospectrally transformed to the real one $\tilde{A}$. Originally, we went in an opposite way, we transformed $\tilde{A}$ to A in terms of Lemma 8

$$A = \mathrm{Diag}\,(h)\tilde{A}\,\mathrm{Diag}\,(h)^{-1}$$

by using the (randomly chosen) function h

$$h_0 = 2+i,\ h_1 = 1+2\,i,\ h_2 = 3-2\,i,\ h_3 = 4+i.$$

In conclusion, for a given A, regarding it as an operator L on some $L^2(\mu)$ and using Lemma 8, we can obtain a very large class of isospectral operators $\tilde{A}$ (i.e. operators $\widetilde{L}$); conversely, from each resulting $\tilde{A}$, we can return to the original A, in terms of Lemma 8 again.

To understand more precisely the role played by the h-transform we have used several times above, let us return to the matrix $\tilde{A}$ given in (10). Recall that if we change the sign of some pairs $(a_{ij}, a_{ji})\,(i \neq j)$, the Hermitizable property is invariant. This leads us to consider the function (h_k) taken values ± 1 only. Note that by homogeneous property, the vectors h and $-h$ are equivalent. Thus, up to the equivalence, there are only 7 choices:

$$\text{single negative: } (-,+,+,+),\ (+,-,+,+),\ (+,+,-,+),\ (+,+,+,-);$$

and

$$\text{double negatives: } (-,-,+,+),\ (+,-,-,+),\ (-,+,-,+).$$

However, these transforms do not include the simple one: only one pair (a_{01}, a_{10}) changes its sign, devote by A_1 the resulting matrix for a moment. This is due to the simple fact that

$$\begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{pmatrix} \mathrm{Diag}\,(\alpha,1,1,1) = \begin{pmatrix} \alpha\, a_{00} & a_{01} & a_{02} & a_{03} \\ \alpha\, a_{10} & a_{11} & a_{12} & a_{13} \\ \alpha\, a_{20} & a_{21} & a_{22} & a_{23} \\ \alpha\, a_{30} & a_{31} & a_{32} & a_{33} \end{pmatrix}$$

and

$$\mathrm{Diag}\,(\beta,1,1,1) \begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} \beta\, a_{00} & \beta\, a_{01} & \beta\, a_{02} & \beta\, a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{pmatrix},$$

and moreover, our matrix is not tridiagonal. (By the way, we mention that the formulas above are meaningful for general h since

$$\mathrm{Diag}\,(\alpha,\beta,1,1) = \mathrm{Diag}\,(\alpha,1,1,1)\,\mathrm{Diag}\,(1,\beta,1,1)$$

for instance.) More seriously, A_1 can not be obtained by any similar transform from the original matrix, since they have different spectrum. Furthermore, it is impossible to remove the negative sign $-(a_{01}, a_{10})$ from A_1 in terms of our h-transform. The main key is that the graph of the given matrix contains real circles. Nevertheless, it is known that A_1 can be transformed into a diagonal real matrix by a similar transform since it is symmetrizable.

It may be helpful if we write down the quadratic form used in Lemma 8 explicitly in the context of matrices.

Remark 11 Let $A = (a_{ij})$ be Hermitizable with respect to μ. Then we have

$$-(Af,f)_\mu = \sum_{i\in E}\mu_i \sum_{j>i}\big[a_{ij}\big(|f_i|^2 - \bar f_i f_j\big) + \bar a_{ij}\big(|f_j|^2 - \bar f_j f_i\big)\big] - \sum_{i\in E}\mu_i\Big[a_{ii} + \sum_{j\neq i} a_{ij}\Big]|f_i|^2.$$

In particular, for real A, we have

$$-(Af,f)_\mu = \sum_{i\in E}\mu_i \sum_{j>i} a_{ij}\big|f_i - f_j\big|^2 - \sum_{i\in E}\mu_i\Big[a_{ii} + \sum_{j\neq i} a_{ij}\Big]|f_i|^2.$$

Proof. It suffices to prove the first assertion, then the second one follows in view of

$$|f_i|^2 - \bar f_i f_j + |f_j|^2 - \bar f_j f_i = (\bar f_i - \bar f_j)(f_i - f_j) = |f_i - f_j|^2.$$

First, we have

$$\begin{aligned} -(Af,f)_\mu &= -\sum_i \mu_i \bar f_i \sum_j a_{ij} f_j \\ &= -\sum_i \mu_i \bar f_i \sum_{j\neq i} a_{ij} f_j - \sum_i \mu_i a_{ii}|f_i|^2 \\ &= \sum_i \mu_i \sum_{j\neq i} a_{ij}\big(|f_i|^2 - \bar f_i f_j\big) - \sum_i \mu_i\Big[a_{ii} + \sum_{j\neq i} a_{ij}\Big]|f_i|^2. \end{aligned}$$

Next,

$$\sum_i \mu_i \sum_{j\neq i} a_{ij}\big(|f_i|^2 - \bar f_i f_j\big) = \sum_i \mu_i \sum_{j>i} a_{ij}\big(|f_i|^2 - \bar f_i f_j\big) + \sum_i \mu_i \sum_{j<i} a_{ij}\big(|f_i|^2 - \bar f_i f_j\big).$$

By the Hermitizable property, the second term on the right

$$\begin{aligned} &= \sum_i \sum_{j<i} \mu_j \bar a_{ji}\big(|f_i|^2 - \bar f_i f_j\big) \\ &= \sum_j \mu_j \sum_{i>j} \bar a_{ji}\big(|f_i|^2 - \bar f_i f_j\big) \\ &= \sum_i \mu_i \sum_{j>i} \bar a_{ij}\big(|f_j|^2 - \bar f_j f_i\big). \end{aligned}$$

Combining these facts together, we obtain the required assertion. □

To have a more concrete impression about our approach, let us consider a simple example.

Example 12 Let

$$A = \begin{pmatrix} -1 & -1 & 0 & 0 \\ -1 & -5 & 4 & 0 \\ 0 & 4 & -13 & 9 \\ 0 & 0 & 9 & -25 \end{pmatrix}$$

Since the matrix is symmetric, we have $\mu_k \equiv 1$. Set

$$c_i = a_{ii} + \sum_{j \neq i} a_{ij}.$$

By Remark 11, we have

$$\begin{aligned} -(Af, f)_\mu &= \big[a_{01}(f_1 - f_0)^2 + a_{12}(f_2 - f_1)^2 + a_{23}(f_3 - f_2)^2\big] - \sum_{k=0}^{3} c_k f_k^2 \\ &= \big[-(f_1 - f_0)^2 + 4(f_2 - f_1)^2 + 9(f_3 - f_2)^2\big] + 2f_0^2 + 2f_1^2 + 16f_3^2. \end{aligned}$$

By a rearrangement of the sum on the right-hand side, we finally obtain

$$-(Af, f)_\mu = (f_0 + f_1)^2 + 4(f_1 - f_2)^2 + 9(f_2 - f_3)^2 + 16f_3^2$$

which is clearly nonnegative definite.

The last sentence of the example above is not obvious since there are negative off-diagonal elements: $a_{01} = a_{10} = -1$. However, these negative terms can be removed by an isospectral transform used in Lemma 8 with $h = (-1, 1, 1, 1)^*$ or $h = (1, -1, 1, 1)^*$. This is the main task in the next section.

Here is the final remark about the problem studied in this section.

Remark 13 The co-zero property in Lemma 3 comes from the assumption of the positivity of the symmetrizing measure μ. In the case allowing some of μ_k to be zero, one may divide the space in different subspaces. Hence it is not essential if we assume that $\mu_k \neq 0$ for each k, we can even allow μ to be complex and ignore the positive ratio property in Lemma 3. Then we have to avoid the logarithm function for using the potential theory, but the path-independence (5), and further Theorem 5 are still meaningful. The reason we do not handle with such a general setup is that we are at the moment mainly interested in the real spectrum. For a complex sequence (μ_k), the spectral theory, if exists, may be quite different from what we are studying here.

Having Theorem 5 at hand as a key for the complexification of the theory due to [17] and [23; Chapter 6], it should be natural to extend the results in the papers just cited, as well as those in [5; §7.2 and Chapter 11] to the complex context. This idea is justified in the next section for a very special situation.

3 Tridiagonal matrices

Let $E = \{k \in \mathbb{Z}_+ : k < N + 1\}$ with $N \leqslant \infty$. In this section, we mainly concentrate on the complex tridiagonal matrix defined on E. Such a matrix is described by three sequences $\{a_k\}_{k=1}^N$, $\{c_k\}_{k=0}^N$, and $\{b_k\}_{k=0}^{N-1}$. It takes the following form

$$A = \begin{pmatrix} -c_0 & b_0 & & & & \\ a_1 & -c_1 & b_1 & & 0 & \\ & a_2 & -c_2 & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & b_{N-1} \\ & & & & a_N & -c_N \end{pmatrix}, \tag{13}$$

Here we allow $N = \infty$. Actually, the case of $N < \infty$ is truncated from the infinite one ($N = \infty$), ignoring the rows and column containing the subscript N:

$$A = \begin{pmatrix} -c_0 & b_0 & & & 0 \\ a_1 & -c_1 & b_1 & & \\ & a_2 & -c_2 & b_2 & \\ 0 & & \ddots & \ddots & \ddots \end{pmatrix}.$$

In what follows, we will not mention this point time by time. We may simply write

$$A \sim (a_k, -c_k, b_k)$$

for simplicity.

Throughout this section, assume that $a_{k+1}b_k > 0\,(0 \leqslant k < N)$ and the sequence (c_k) is real. By Corollary 6, these conditions are equivalent to the Hermitizability of A. Thus, in this section, we are concentrated in the isospectral problem of the complex matrix A and a real (having positive subdiagonal elements, in particular) tridiagonal matrix.

From now on, unless otherwise stated, assume that $c_k > 0$ at least for the first finite number of k, otherwise, simply use a shift. In particular, when $N < \infty$, we may assume directly that $c_k > 0$ for each $k : 0 \leqslant k \leqslant N$.

We recall that the important quantity (12) in the present context becomes

$$m = \sup_{k \in E} (-c_k + |a_k| + |b_k|)^+, \tag{14}$$

where we have used the convention that $a_0 = 0$ and $b_N = 0$ if $N < \infty$. The main result in this section is the next algorithm for constructing a valuable isospectral matrix $\tilde{A} \sim (\tilde{a}_k, -\tilde{c}_k, \tilde{b}_k)$ of the original A. This not only extends the earlier results to the complex context but also simplifies an important step of our earlier algorithms, refer to [9–12] for more details.

Algorithm 14 Assume $m < \infty$. Set $u_k = a_k b_{k-1} (= |a_k b_{k-1}|)$.

(1) Let $m = 0$. Then for each k, $c_k \geqslant |a_k| + |b_k| > 0$. Set $\tilde{c}_k = c_k$ and $\tilde{b}_0 = c_0 > 0$. Next, let

$$\begin{cases} \tilde{b}_k = c_k - \dfrac{u_k}{\tilde{b}_{k-1}}, & 1 \leqslant k < N \\ \tilde{a}_k = c_k - \tilde{b}_k, & 1 \leqslant k < N \\ \tilde{a}_N = \dfrac{u_N}{\tilde{b}_{N-1}} & \text{if } N < \infty. \end{cases} \tag{15}$$

More explicitly,

$$\begin{cases} \tilde{b}_0 = c_0, \\ \tilde{b}_k = c_k - \dfrac{u_k}{c_{k-1} - \dfrac{u_{k-1}}{c_{k-2} - \dfrac{u_{k-2}}{\ddots \; c_2 - \dfrac{u_2}{c_1 - \dfrac{u_1}{c_0}}}}}, & 1 \leqslant k < N, \\ \tilde{a}_k = c_k - \tilde{b}_k, & 1 \leqslant k < N, \\ \tilde{a}_N = \dfrac{u_N}{\tilde{b}_{N-1}}, & N < \infty. \end{cases}$$

(2) Let $m > 0$. Then replacing (c_k) by $(\tilde{c}_k := c_k + m)$ and then repeat the procedure in part (1) to compute $\big(\tilde{a}_k, \tilde{b}_k\big)$.

Theorem 15 For $\tilde{A}$ defined in Algorithm 14, the sequences $(\tilde{a}_k)$ and $(\tilde{b}_k)$ are positive, the sum of each row equals zero, except the Nth row which is not positive if $N < \infty$.

The positivity of $(\tilde{a}_k)$ and $(\tilde{b}_k)$ in the theorem are essential for the algorithms introduced in [9–12] for computing the maximal eigenpair. In other words, we have luckily reduced the computation for complex matrices to the one having positive sub-diagonal elements which has been well studied.

To identify the spectrum of the matrix $\tilde{A}$ constructed by the algorithm above and that of the matrix $A^{(m)} := A - mI$, we need more preparation.

In what follows, replace A by $A^{(m)}$ if necessary. Recall that for A, we have the measure (μ_n):

$$\mu_0 = 1, \;\; \mu_n = \mu_{n-1} \frac{b_{n-1}}{\bar{a}_n}, \qquad 1 \leqslant n < N + 1.$$

Alternatively,

$$\mu_0 = 1, \;\; \mu_n = \prod_{j=1}^{n} \frac{b_{j-1}}{\bar{a}_j}, \qquad 1 \leqslant n < N + 1.$$

Since both $(\tilde{a}_k)$ and $(\tilde{b}_k)$ constructed in the algorithm are positive. We can define $(\tilde{\mu}_n)$ using $(\tilde{a}_k,\ \tilde{b}_k)$ instead of (a_k, b_k).

Next, define successively

$$h_0 = 1,\ h_1 = h_0\frac{\tilde{b}_0}{b_0},\ \ldots,\ h_k = h_{k-1}\frac{\tilde{b}_{k-1}}{b_{k-1}}, \qquad 1 \leqslant k < N+1.$$

Alternatively,

$$h_0 = 1,\ \ h_k = \prod_{j=0}^{k-1}\frac{\tilde{b}_j}{b_j}, \qquad 1 \leqslant k < N+1.$$

Since the case of $N = \infty$ is allowed, we have to take care of the domain of the infinite matrices. Let $\mathscr{D}(\tilde{A})$ be the domain of $\tilde{A}$ on $L^2(\tilde{\mu})$. Define the deduced domain of A on $L^2(\mu)$ as follows.

$$\mathscr{D}(A) = \{f \in L^2(\mu) : f/h \in \mathscr{D}(\tilde{A})\}.$$

Due to Corollary 6 and Lemma 8, we have the following result.

Theorem 16 The selfadjoint operators $(A, \mathscr{D}(A))$ and $(\tilde{A}, \mathscr{D}(\tilde{A}))$ have the same real spectrum. More precisely, for each $f \in \mathscr{D}(A)$, with $\tilde{f} = f/h$, we have $(Af, f)_\mu = (\tilde{A}\tilde{f}, \tilde{f})_{\tilde{\mu}}$, where

$$\begin{aligned}-(Af,f)_\mu &= \sum_{i\in E}\mu_i\big[b_i\big(|f_i|^2 - \bar{f}_i f_{i+1}\big) + \bar{b}_i\big(|f_{i+1}|^2 - \bar{f}_{i+1}f_i\big)\big] \\ &\quad + \sum_{i\in E}\mu_i\big(c_i - a_i - b_i\big)|f_i|^2. \\ -(\tilde{A}f,f)_{\tilde{\mu}} &= \sum_{i\in E}\tilde{\mu}_i\tilde{b}_i\big|f_i - f_{i+1}\big|^2 + \mathbb{1}_{\{N<\infty\}}\tilde{\mu}_N\big(\tilde{c}_N - \tilde{a}_N\big)|f_N|^2.\end{aligned}$$

We remark that since $\tilde{A}$, as well as its spectrum, are all real, in the study of the spectrum of $\tilde{A}$, it suffices to use real L^2-space, rather than the complex one. It should be pointed out here that, based on Theorem 16, a large part of the results obtained earlier (see [5–7]) can be extended to the present complex content.

Before going to the proofs, let us look at some examples to show the application of the above results.

Example 17 Let

$$A = \begin{pmatrix} -1 & -1 & & \\ -1 & -1 & 4 & \\ & 4 & -1 & 9 \\ & & 9 & -1 \end{pmatrix}$$

Then A has spectrum

$$-10.8573,\ \ 8.8573,\ \ -1.91303,\ \ 0.08697.$$

clearly, $\tilde{b}_1 = 0$. Hence, we use an arbitrarily shift $A^{(1)} := A - I$ instead of A. Then, by using Algorithm 14, we have

$$\tilde{c}_k \equiv 2;\ \tilde{b}_0 = 2,\ \tilde{b}_1 = 3/2,\ \tilde{b}_2 = -26/3;\ \tilde{a}_1 = 1/2,\ \tilde{a}_2 = 32/3,\ \tilde{a}_3 = -243/26.$$

The resulting tridiagonal matrix $\tilde{A}$ has the same spectrum as $A^{(1)}$ by Theorem 16:

$$-11.8573,\ \ 7.8573,\ \ -2.91303,\ \ -1.08697.$$

In this example the off-diagonal elements of the matrix $\tilde{A}$ are not all positive. This is because the adopted shift is not big enough. It leads to the next example.

Example 18 Let A be the same as in the previous example. By Algorithm 14, we adopt the shift defined by (14): $m = 12$. Then, for $A^{(12)}$, we have

$$\tilde{c}_k \equiv 13;\ \tilde{b}_0 = 13,\ \tilde{b}_1 = \frac{168}{13},\ \tilde{b}_2 = \frac{247}{21};\ \tilde{a}_1 = \frac{1}{13},\ \tilde{a}_2 = \frac{26}{21},\ \tilde{a}_3 = \frac{1701}{247}.$$

Clearly, the off-diagonal elements of the resulting tridiagonal matrix $\tilde{A}$ are positive. It also proposes the following property:

The sum of each row equals zero except the last one which is negative.

Certainly, $A^{(12)}$ and $\tilde{A}$ have the same spectrum:

$$-22.8573,\ \ -13.913,\ \ -12.087,\ \ -3.1427.$$

To keep this property, as mentioned below (12), the shift can often be a little smaller. For this example, $m = 9$ is okay but not $m \in [0, 8.5]$.

Example 19 Let

$$A = \begin{pmatrix} -1 & 2+i & & & 0 \\ 2^2(2-i) & -1 & 2^4(2+i) & & \\ & 6^2(2-i) & -1 & 3^4(2+i) & \\ & & 12^2(2-i) & -1 & 4^4(2+i) \\ 0 & & & 20^2(2-i) & -1 \end{pmatrix}.$$

Since the shift defined by (14) equals $400\sqrt{5} - 1 \approx 893.427$, we choose $m = 899$ for simplifying the computation. Then, by Algorithm 14, we have

$$\tilde{c}_k \equiv 900;\ \tilde{b}_0 = 900,\ \tilde{b}_1 = \frac{40499}{45},\ \tilde{b}_2 = \frac{36319500}{40499};\ \tilde{b}_3 = \frac{168475824}{201775};$$
$$\tilde{a}_1 = \frac{1}{45},\ \tilde{a}_2 = \frac{129600}{40499},\ \tilde{a}_3 = \frac{13121676}{201775},\ \tilde{a}_4 = \frac{6456800000}{10529739} < \tilde{c}_4.$$

Here are the eigenvalues of $\tilde{A}$ $\big($or $A^{(899)}\big)$:

$$-1655.39,\quad -951.031,\quad -900,\quad -848.969,\quad -144.609.$$

From which we obtain the eigenvalues of the original A:

$$-756.391,\quad -52.0308,\quad -1.,\quad 50.0308,\quad 754.391.$$

The most of the remainder of this section, except the last result at the end of the section, is devoted to the proofs of the above three results with some extension.

We review a few of points known from earlier study. In the real context, when $a_k > 0,\ b_k > 0,\ c_k \geqslant a_k + b_k$, and $m = 0$, the h-transform was initially introduced in [15; §2],

$$\tilde{b}_k = b_k \frac{h_{k+1}}{h_k}, \qquad \tilde{a}_k = a_k \frac{h_{k-1}}{h_k}, \tag{16}$$

where h (with $h_0 = 1$) is harmonic on the set $\{k : 0 \leqslant k < N\}$: $Ah = 0$ on this set. That is

$$b_k h_{k+1} + a_k h_{k-1} - c_k h_k = 0, \qquad 0 \leqslant k < N,\ a_0 := 0.$$

Equivalently,

$$c_k = \tilde{b}_k + \tilde{a}_k, \qquad 0 \leqslant k < N,\ \tilde{a}_0 := 0. \tag{17}$$

With $r_k := h_k/h_{k+1}$, in [8; §5], we find first a recursive formula of (r_k), then a solution of (h_k), and finally the pair $\big(\tilde{a}_k, \tilde{b}_k\big)$ in terms of (16). We remark that the discussion from (16) to here, we preassume that $h_k \neq 0$ for each k. Equivalently, $\tilde{b}_k \neq 0$ for each k. This is however not necessary true as we have seen from Example 17. We are going to present more details in the next lemma and its corollaries. Thanks are given to the explicit formulas, it is easy to see that all these formulas remain the same under the extension from real to complex tridiagonal A.

We are now going to study the direct construction of the pair $\big(\tilde{a}_k, \tilde{b}_k\big)$ given in Algorithm 14. The key point is adopting the recursive method used originally for (r_k), now to $\big(\tilde{b}_k\big)$ directly. By the Hermitizable property, $b_k/\bar{a}_{k+1} > 0$, we can write

$$b_k = \beta_k e^{i\theta_k}, \quad a_{k+1} = \alpha_{k+1} e^{-i\theta_k}, \qquad \beta_k := |b_k|,\ \alpha_k := |a_k|.$$

It follows from (16) that the h-transform has an invariance:

$$\tilde{a}_k \tilde{b}_{k-1} = a_k b_{k-1} = \alpha_k \beta_{k-1} > 0, \qquad 1 \leqslant k < N + 1. \tag{18}$$

Thus, we can rewrite (17) as

$$\tilde{b}_k = c_k - \tilde{a}_k \overset{(16)}{=} c_k - \frac{\alpha_k \beta_{k-1}}{\tilde{b}_{k-1}}, \qquad 0 \leqslant k < N \tag{19}$$

provided $\tilde{b}_{k-1} \neq 0$. This is a critical observation which enables us to use the sequence $\big(\tilde{b}_k\big)$ instead of (r_k), ignoring (h_k). It then deduces a direct construction of the sequences $\big(\tilde{b}_k\big)$ and $(\tilde{a}_k)$. From this, it follows that $\tilde{b}_k$ must be real once $\tilde{b}_{k-1} \neq 0$, and then $\tilde{a}_k$ should also be real. Furthermore, $\tilde{a}_k$ and $\tilde{b}_k$ are both real for every k.

Actually, we need to examine the definition of $(\tilde{a}_k)$ and $\big(\tilde{b}_k\big)$ more carefully. In the discussion below, we often allow general real $(c_k : k \in E)$. However, we have assumed that $c_k > 0$ at least for the first finite number of k. In the special case that $\sup_{k\in E}(-c_k) < \infty$ (which is weaker than (14)), by using a shift if necessary, one can even assume that (c_k) is positive, but we do not need this condition at the moment.

Lemma 20 We have the following assertions.

(1) $\tilde{b}_0 = c_0 > 0$.

(2) Suppose that $\tilde{b}_j \neq 0$ for each $j \leqslant k-1$. Then $\tilde{b}_k$ is well defined by (19). Furthermore $\tilde{b}_k = c_k - F_k(c_0, \ldots, c_{k-1})$:

$$F_k(x_0, x_1, \ldots, x_{k-1}) = \cfrac{u_k}{x_{k-1} - \cfrac{u_{k-1}}{x_{k-2} - \cfrac{u_{k-2}}{\ddots \\ x_2 - \cfrac{u_2}{x_1 - \cfrac{u_1}{x_0}}}}}, \tag{20}$$

where $u_k = a_k b_{k-1} (= \alpha_k \beta_{k-1})$. The subscript k of F_k means the function has k variables.

(3) $\tilde{a}_k$ is well-defined once $\tilde{b}_{k-1} \neq 0$: $\tilde{a}_k = u_k/\tilde{b}_{k-1}$. In which case, both $\tilde{a}_k$ and $\tilde{b}_{k-1}$ have the same sign.

Proof. The first assertion is obvious by (17). The third assertion follows from (18).

To prove the second assertion, simply use (19) repeatedly:

$$\tilde{b}_k = c_k - \frac{u_k}{\tilde{b}_{k-1}} = c_k - \cfrac{u_k}{c_{k-1} - \cfrac{u_{k-1}}{\tilde{b}_{k-2}}} = \cdots\cdots$$

plus the fact that $\tilde{b}_0 = c_0$. □

Lemma 20 provides us a direct way to compute $\big(\tilde{a}_k, \tilde{b}_k\big)$ defined by (15), which are real, without using the sequence (r_k) and (h_k).

Corollary 21 For a given Hermitizable tridiagonal matrix $(a_k, -c_k, b_k)$, a direct construction of a real one $\big(\tilde{a}_k, -\tilde{c}_k, \tilde{b}_k\big)$ goes as follows. Keep $\tilde{c}_k = c_k$. Let $u_k = a_k b_{k-1} (= |a_k b_{k-1}|)$ and $\tilde{b}_0 = c_0$. For each $k : 1 \leqslant k < N$, if $\tilde{b}_j \neq 0$ for every $j \leqslant k-1$, then define $\big(\tilde{a}_k, \tilde{b}_k\big)$ by (15). If otherwise $\tilde{b}_{k-1} = 0$ for some $k < N$, then replacing (c_k) by $(c_k + m)$ for some constant $m > 0$, and then repeat the above procedure to compute the new pairs $\big(\tilde{a}_k, \tilde{b}_k\big)$.

At the moment, it is rare to use the shift at the end of the corollary. However, the shift will become much important for constructing positive pairs $\big(\tilde{a}_k, \tilde{b}_k\big)$ for our study on the spectrum of the matrices. The aim now is to look for a condition to guarantee this positivity. Since $\tilde{b}_0 = c_0 > 0$, by (19), it follows that the sequence $\big(\tilde{b}_k\big)$ is positive iff

$$c_k > \frac{\alpha_k \beta_{k-1}}{\tilde{b}_{k-1}} \ \text{(or equivalently } c_k > \tilde{a}_k) \quad \text{for each } k\colon\ 1 \leqslant k < N+1. \tag{21}$$

By using a shift, this can be improved as follows. There exists a constant $m_0 > 0$ such that

$$\begin{aligned} c_k + m_0 > &\frac{\alpha_k \beta_{k-1}}{\tilde{b}_{k-1}} \ \text{(or equivalently } c_k + m_0 > \tilde{a}_k) \\ &\text{for each } k\colon\ 1 \leqslant k < N+1. \end{aligned} \tag{22}$$

This simple observation looks very nice but it is unfortunately not practical. Because $\big(\tilde{a}_k, \tilde{b}_k\big)$ depends on the "harmonic" function h, and then h depends on (c_k) and m_0. The problem is that for a given $A \sim (a_k, -c_k, b_k)$, we do not know in advance how to choose m_0 such that (22) holds for the resulting $\big(\tilde{b}_k\big)$ under an h-transform. What we are going to prove is that $m_0 = m$ (defined by (14)) is sufficient for this purpose. Note that the constant m depends on A only and as we have mentioned in the last section, it is reasonable and can even be sharp. This conclusion is included in the following corollary. For temporary use in the next corollary, we introduce the concept of singular point. We call $y_0 = 0$ a singular point of $F_1(x_0)$. If $x_0 \neq y_0$, we call $y_1 := u_1/x_0$ a singular point of $F_2(x_0, x_1)$. Successively, if $x_0 \neq y_0$, $x_1 \neq y_1$, …, $x_{k-2} \neq y_{k-2}$, we call $y_{k-1} := u_{k-1}/(x_{k-2} - F(x_0, \ldots, x_{k-2}))$, a singular point of $F_k(x_0, \ldots, x_{k-1})$.

Corollary 22 Use the notation given in Lemma 20.

(1) The function F_k is decreasing in each of its components out of the set of its singular points, provided $\tilde{b}_j > 0$ for each $j \leqslant k-1$.

(2) For given (a_k) and (b_k), there exists $(c_k)\,(0 \leqslant k < N)$ such that $\tilde{b}_k > 0$ up to $N-1$ and so does $\tilde{a}_k$ Up to N. A particular choice is $c_k \equiv |a_k| + |b_k|$ for $k < N$ and $c_N \geqslant |a_N|$ if $N < \infty$.

(3) The second assertion still holds if the sequence (c_k) is replaced by $(c_k + m)$, where $m > 0$ is a constant.

Proof. The first assertion is obvious in view of (20) and the fact that $u_k > 0$. From this, one can construct the sequence (c_k) required in part (2) successively starting from an arbitrarily chosen positive c_0. Alternatively, one may check

that the particular choice by a simple computation.

$$\begin{aligned}
&\tilde{b}_0 = c_0 = |b_0| > 0,\\
&\tilde{b}_1 = c_1 - \frac{|a_1 b_0|}{\tilde{b}_0} = c_1 - \frac{|a_1 b_0|}{|b_0|} = |b_1| > 0,\\
&\tilde{b}_2 = c_2 - \frac{|a_2 b_1|}{\tilde{b}_1} = c_2 - \frac{|a_2 b_1|}{|b_1|} = |b_2| > 0,\\
&\quad \cdots\cdots\\
&\tilde{a}_N = \frac{|a_N b_{N-1}|}{\tilde{b}_{N-1}} = \frac{|a_N b_{N-1}|}{b_{N-1}} = |a_N| > 0 \text{ if } N < \infty.
\end{aligned}$$

Hence, $\tilde{b}_k > 0$ for each k and so does $\tilde{a}_k$ by (18). Actually, we have proved in the special case that $N = \infty$, (a_k) and (b_k) are positive, then $\tilde{b}_k \equiv b_k$ and $\tilde{a}_k \equiv a_k$. The third assertion follows from the first two. □

Proofs of Theorems 15 and 16 The second assertion of Theorem 15 about the property of the sum of rows comes from the definition of Algorithm 14. The first assertion of the theorem is the hard part of the algorithm, it follows from Corollaries 21 and 22.

Theorem 16 is a simple application of Lemma 8. □

To conclude this section, we study the complexification of the results obtained in [8] for the discrete spectrum in the context of matrices. Even though we are mainly interested here the case that $N = \infty$ but the next result is meaningful with a slight modification for finite N. For the domain of an operator A, here we are in a simple situation, we use either the maximal or the minimal domain as in [8].

Given a tridiagonal matrix $A \sim (a_k, -c_k, b_k)$ with a convention $a_0 = 0$, suppose that the sequences are all positive and moreover $c_k = a_k + b_k$ for each $k \in E$. Next, let α_k and β_k be arbitrary positive sequences and (θ_k) be arbitrary sequence with $-\pi < \theta_k \leqslant \pi$. Define

$$\tilde{b}_k = b_k \beta_k e^{i\theta_k}, \quad \tilde{a}_{k+1} = a_{k+1}\alpha_{k+1} e^{-i\theta_k}, \quad \tilde{c}_k = b_k\beta_k + a_k\alpha_k, \qquad k \in E$$

and

$$\bar{b}_k = b_k\beta_k, \quad \bar{a}_{k+1} = a_{k+1}\alpha_{k+1}, \quad \bar{c}_k = b_k\beta_k + a_k\alpha_k (= \tilde{c}_k), \qquad k \in E.$$

Theorem 23 The tridiagonal matrices $\tilde{A}$ and $\bar{A}$ have the same real spectrum. In particular, when $\alpha_k \equiv 1$ and $\beta_k \equiv 1$, the resulting $\tilde{A}$ and A have the same real spectrum.

Proof. By Corollary 6, the tridiagonal matrices A, $\tilde{A}$ and $\bar{A}$ are all Hermitizable. The isospectral property of $\tilde{A}$ and $\bar{A}$, as well as the last assertion, follows from Theorem 16. □

We remark that the bounded perturbation of the diagonal elements are permitted in the above results, using a shift if necessary as we used several times before.

Having Theorem 23 at hand, it should be not hard to extend the criteria for discrete spectrum obtained in [8] to the present complex setup, but we omit the details here.

To conclude this section, we return to our general Hermitizable setup.

Theorem 24 The spectrum of each Hermitizable matrix coincides (up to a constant shift) with a union of the spectrums of some irreducible birth–death Q-matrices.

Proof. Let $A = (a_{ij} : i, j \in E)$ be Hermitizable with respect to some positive μ. That is

$$\mathrm{Diag}(\mu)A = A^H \mathrm{Diag}(\mu).$$

From this, it is easy to check that

$$H := \mathrm{Diag}(\mu)^{1/2} A \mathrm{Diag}(\mu)^{-1/2}$$

is Hermitian, which is clearly similar to A. By [25; Theorem 2.4] (see also [22]), H is similar to a real, symmetric tridiagonal matrix T. Certainly, T is Hermitizable. Write $T \sim (a_k, -c_k, b_k)$ as before. Then the space E can be divided uniquely into subsets $\{E_j\}$: $E = \sum_j E_j$, on each of them T is irreducible (i.e. $a_k b_{k-1} > 0$ on each E_j). Replacing T by a shift $T + mI$ (for large enough real constant m) if necessary, and then applying the construction given in Algorithm 14 to $T|_{E_j}$, we obtain for each j an isospectral birth-death Q-matrix. Since each similar transform is isospectral and the transforms are transitive, we have thus proved the required assertion. □

4 New algorithms for tridiagonal matrices

In the last section (Algorithm 14 and Theorem 15 in particular), we have proved that the spectrum of a complex symmetrizable tridiagonal matrix coincides with the one having positive subdiagonal elements. As mentioned before, the last object was more or less well studied in [9–12]. We start this section with a new problem, and then we will introduce new algorithms.

The problem

As pointed out in [26; §3.3 and Example 4.4], in non-symmetric case, we may get trouble for large size matrix. To see this more clearly, we look at a simple

example [26; (4.6)]. Let $E=\{0,1,\dots,N\}$ with $N<\infty$. Consider the matrix

$$Q=\begin{pmatrix} -3 & 2 & & & & \\ 1 & -3 & 2 & & 0 & \\ & 1 & -3 & 2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & 2 \\ & & & & 1 & -3 \end{pmatrix}. \tag{23}$$

Since $c_0>b_0$ and $c_N>a_N$, we need to remove the first one in terms of Algorithm 14, the resulting matrix becomes

$$\widetilde{Q}=\begin{pmatrix} -3 & \dfrac{2^2-1}{2-1} & & & & \\ \dfrac{2^2-2}{2^2-1} & -3 & \dfrac{2^3-1}{2^2-1} & & 0 & \\ & \dfrac{2^3-2}{2^3-1} & -3 & \dfrac{2^4-1}{2^3-1} & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & \dfrac{2^{N+1}-1}{2^N-1} \\ & & & & \dfrac{2^{N+1}-2}{2^{N+1}-1} & -3 \end{pmatrix}.$$

Then, we have

$$\begin{aligned} &\tilde{c}_k=3,\ \tilde{b}_k=\frac{2^{k+2}-1}{2^{k+1}-1}=2+\frac{1}{2^{k+1}-1},\ \tilde{a}_k=\frac{2^{k+1}-2}{2^{k+1}-1}=1-\frac{1}{2^{k+1}-1}, \qquad k\in E; \\ &\tilde{\mu}_0=1,\ \tilde{\mu}_n=2^{n+2}-4+2^{-n}, \quad \tilde{\mu}_n\tilde{b}_n=2^{n+3}-6+2^{-n}, \qquad n\in E \end{aligned} \tag{24}$$

where $\big(\tilde{b}_N:=3-\tilde{a}_N$ and$\big)$

$$\tilde{\varphi}_n=\sum_{j=n}^{N}\frac{1}{\tilde{\mu}_j\tilde{b}_j}=\sum_{j=n}^{N}\frac{1}{2^{j+3}-6+2^{-j}}, \qquad n\in E. \tag{25}$$

The maximal eigenvalue is

$$\lambda_0+3=2\sqrt{2}\cos\frac{\pi}{N+2}\to 2\sqrt{2}\approx 2.82843 \qquad \text{as } N\to\infty.$$

To obtain the 6-bit accuracy 2.82843, it is necessary and sufficient that $N\geqslant 2562$. From now on in the discussion about this model, we fix $N=2562$. We also remark that the maximal eigenvector has exponentially decay. As a mimic of the eigenvector of $\lambda_0\big(-\widetilde{Q}\big)$, we choose

$$w^{(0)}=\sqrt{\tilde{\varphi}}$$

as the initial vector in our algorithm.

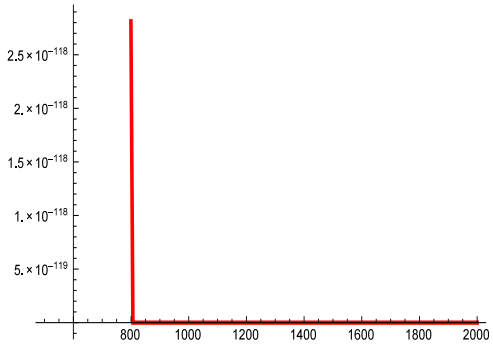


Figure 3: $w^{(0)}$ on $[500, 2000]$

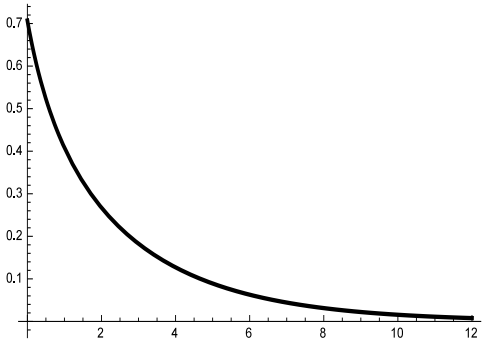


Figure 4: $w^{(0)}$ on $[0, 12]$

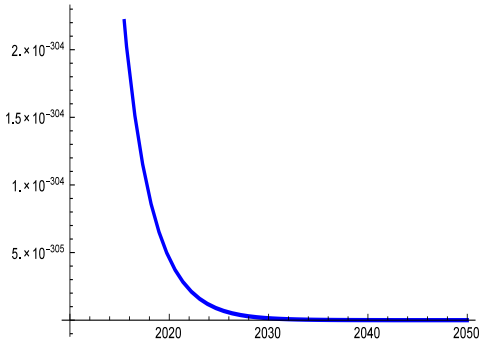


Figure 5: $w^{(0)}$ on $[2010, 2050]$

From Figure 3, one sees that the curve of $w^{(0)}$ on $[500, 2000]$ goes down rapidly from 10^{-118} to 10^{-120} and then stay there. To figure out more clearly, we choose smaller intervals at the beginning and at the end: $[0, 12]$ and $[2010, 2050]$. For the first one in Figure 4, $w^{(0)}$ starts at 0.7 and then goes down quickly. In Figure 5, $w^{(0)}$ goes down very fast, starts at 10^{-303} goes down to 10^{-306}. Thus,

$$\frac{\text{Maximum of } w^{(0)}}{\text{Minimum of } w^{(0)}} \approx \frac{0.7}{10^{-306}} > 10^{305}. \tag{26}$$

The numerical computations for this example are completed by using Mathematica version 10.3. The precision level is made automatically by the software.

On the other hand, in computational mathematics, one often treats the symmetric matrices to which there are a lot of algorithms. Actually, a standard algorithm introduced in the textbooks (refer to [3; pp. 142–146], for instance) was used in the author's first paper [4] (1991) for the study on this topic, which was also included in the first edition (but replaced by analytic results in the second edition) of the book [5] (1992). The symmetrizing matrix of $\widetilde{Q}$

is as follows.

$$Q^{\rm sym}=\begin{pmatrix} -3 & \sqrt{2} & & & & \\ \sqrt{2} & -3 & \sqrt{2} & & 0 & \\ & \sqrt{2} & -3 & \sqrt{2} & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & \sqrt{2} \\ & & & & \sqrt{2} & -3 \end{pmatrix},$$

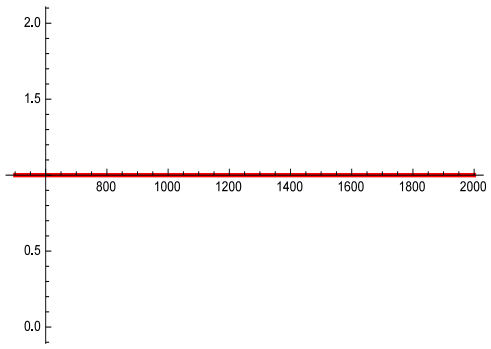


Figure 6: $w^{(0)}$ on $[500, 2000]$

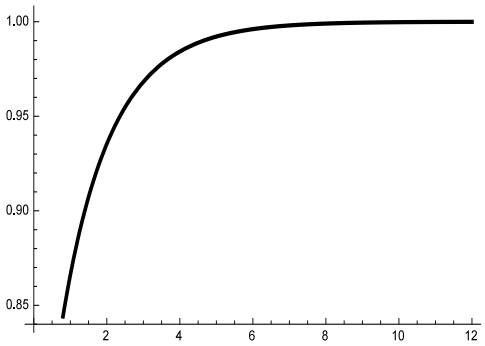


Figure 7: $w^{(0)}$ on $[0, 12]$

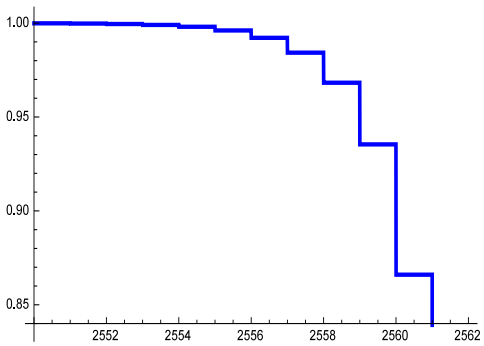


Figure 8: $w^{(0)}$ on $[2550, 2562]$

Note that for this matrix, in contrast with $\widetilde{Q}$, the sum of each row is not zero. Hence we can not use our analytic estimates developed so far as used in [9–12]. That leads to the total loss of martial arts. On the other hand, if we use the isospectral transform again (Algorithm 14), then the resulting matrix should have zero sum for the first $N-1$ rows, but the resulting matrix should be asymmetry. We have thus involved in an unsolvable circulation. This problem has been opened for some years, and luckily we have now found a solution. That is the new algorithm to be stated quite soon. Before going to the details, let us now show the results of our new initial vector $w^{(0)}$ for our new algorithm. Figure 6 is the value of $w^{(0)}$ on $[500, 2000]$. It is simply

lying in the straight line 1. Figure 7 shows that on the initial interval $[0, 12]$, $w^{(0)}$ starts at $1/\sqrt{2} \approx 0.7$, increases to 1. Figure 8 shows on the end interval $[2550, 2562]$, $w^{(0)}$ goes down from 1 to $1/\sqrt{2}$. Therefore, we have

$$\frac{\text{Maximum of } w^{(0)}}{\text{Minimum of } w^{(0)}} \approx \frac{1}{0.7} \approx 1.4. \tag{27}$$

Comparing (27) with (26), it should be clear the difference of the new algorithm with the earlier one.

Non-conservative case

We now start to state our new algorithm.

For a given complex tridiagonal matrix, by Algorithm 14, using shift if necessary, we may assume that the resulting matrix on $E = \{k \in \mathbb{Z}_+ : 0 \leqslant k < N + 1\}$ is as follows:

$$\widetilde{Q} = \begin{pmatrix} -\tilde{c}_0 & \tilde{b}_0 & & & & \\ \tilde{a}_1 & -\tilde{c}_1 & \tilde{b}_1 & & 0 & \\ & \tilde{a}_2 & -\tilde{c}_2 & \tilde{b}_2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & \tilde{b}_{N-1} \\ & & & & \tilde{a}_N & -\tilde{c}_N \end{pmatrix},$$

where $\tilde{a}_k > 0\,(1 \leqslant k < N + 1)$, $\tilde{b}_k > 0\,(0 \leqslant k < N)$, $\tilde{c}_k = \tilde{a}_k + \tilde{b}_k$ (which means that $\widetilde{Q}$ is conservative at k) for each $k < N$ (with $\tilde{a}_0 := 0$), and $\tilde{c}_N \geqslant \tilde{a}_N$. In general, this matrix $\widetilde{Q}$ is non-symmetric. In the particular case that $\tilde{c}_N = \tilde{a}_N$, the matrix is conservative (i.e. conservative at each $k \in E$) and so has trivial maximal eigenvalue 0. We will come back to this case in the third part of this section. From now on, unless otherwise stated, assume that $\tilde{c}_N > \tilde{a}_N$.

In the computation of the maximal eigenpair, this form of tridiagonal matrix is essential for which we have explicit and efficient initials, and strong estimates of the maximal eigenvalue. The initials are expressed by two sequences. The first one is the measure $(\tilde{\mu}_k)$:

$$\tilde{\mu}_0 = 1, \;\; \tilde{\mu}_n = \tilde{\mu}_{n-1}\frac{\tilde{b}_{n-1}}{\tilde{a}_n}, \qquad 1 \leqslant n < N + 1. \tag{28}$$

The second one is

$$\tilde{\varphi}_n = \sum_{k=n}^{N} \frac{1}{\tilde{\mu}_k \tilde{b}_k}, \qquad 0 \leqslant n < N + 1, \tag{29}$$

where $\tilde{b}_N := \tilde{c}_N - \tilde{a}_N$ if $N < \infty$.

Unfortunately, the resulting matrix $\widetilde{Q}$ is often non-symmetric, as mentioned before. Certainly, there is a simple way to symmetrizing $\widetilde{Q}$. That is,

$$Q^{\rm sym} = \mathrm{Diag}(\tilde{\mu})^{1/2}\,\widetilde{Q}\,\mathrm{Diag}(\tilde{\mu})^{-1/2}$$
$$= \begin{pmatrix} -\tilde{c}_0 & \sqrt{\tilde{a}_1\tilde{b}_0} & & & 0 & \\ \sqrt{\tilde{a}_1\tilde{b}_0} & -\tilde{c}_1 & \sqrt{\tilde{a}_2\tilde{b}_1} & & & \\ & \sqrt{\tilde{a}_2\tilde{b}_1} & -\tilde{c}_2 & \sqrt{\tilde{a}_3\tilde{b}_2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \sqrt{\tilde{a}_N\tilde{b}_{N-1}} \\ & 0 & & & \sqrt{\tilde{a}_N\tilde{b}_{N-1}} & -\tilde{c}_N \end{pmatrix},$$

where $\mathrm{Diag}(\mu)$ is the diagonal matrix having diagonal elements (μ_k). Since it happens often that

$$\tilde{c}_k = \tilde{a}_k + \tilde{b}_k \neq \sqrt{\tilde{a}_k\tilde{b}_{k-1}} + \sqrt{\tilde{a}_{k+1}\tilde{b}_k},$$

as an example, $\tilde{a}_k \equiv a$, $\tilde{b}_k \equiv b$, and $a \neq b$, the conservative property can be lost in this symmetrizing procedure. As mentioned before, this may lead to an unsolvable circulation. It is the main reason that we have not used such a symmetrizing procedure for more than two decades. In the special case that $\widetilde{Q}$ is already symmetric, since $\tilde{\mu}_k \equiv 1$, the algorithm to be introduced soon coincides with the original algorithms introduced in [9, 11, 12].

We are now lucky to find a new way to solve the problem. Even though in the non-symmetric case, we can not use either the h-transform or the symmetrizing procedure, individually, but we can couple them together, using them simultaneously. Let us now state our new algorithm, a specific coupling of two algorithms originally designed to these matrices separately. In the other words, the new algorithm couples the advantages of the both algorithms for the different matrices.

Algorithm 25 (1) Let $w_i^{(0)} = \sqrt{\tilde{\mu}_i\tilde{\varphi}_i}$, $i \in E$. Define

$$v^{(0)} = \frac{w^{(0)}}{\sqrt{w^{(0)*}w^{(0)}}}, \qquad z_0 = \frac{1}{\delta_0}, \tag{30}$$

$$\delta_0 = \sup_{0\leqslant n<N+1}\left[\sqrt{\tilde{\varphi}_n}\sum_{i=0}^{n}\tilde{\mu}_i\sqrt{\tilde{\varphi}_i} + \frac{1}{\sqrt{\tilde{\varphi}_n}}\sum_{n+1\leqslant j<N+1}\tilde{\mu}_j\tilde{\varphi}_j^{3/2}\right]. \tag{31}$$

(2) For each $k \geqslant 1$, solve $w^{(k)}$:

$$\big(-Q^{\rm sym} - z_{k-1}I\big)\,w^{(k)} = v^{(k-1)}, \tag{32}$$

and define

$$v^{(k)} = \frac{w^{(k)}}{\sqrt{w^{(k)*} w^{(k)}}}, \qquad z_k = \frac{1}{\delta_k}, \quad (33)$$

$$\delta_k = \sup_{0 \leqslant n < N+1} \frac{\sqrt{\tilde{\mu}_n}}{v_n^{(k)}} \Bigg[\tilde{\varphi}_n \sum_{i=0}^{n} \sqrt{\tilde{\mu}_i}\, v_i^{(k)} + \sum_{n+1 \leqslant j < N+1} \sqrt{\tilde{\mu}_j}\, \tilde{\varphi}_j v_j^{(k)} \Bigg]. \quad (34)$$

Then

$$v^{(k)} \to g \quad \text{and} \quad z_k \to \lambda_0 \qquad \text{as} \ \ k \to \infty,$$

where $(\lambda_0,\ g)$ is the minimal eigenpair of $-Q^{\,\mathrm{sym}}$. Furthermore, the minimal eigenpair of $-\widetilde{Q}$ equals $\big(\lambda_0,\ \mathrm{Diag}(\tilde{\mu})^{-1/2} g\big)$.

We now mention shortly the role played by $\widetilde{Q}$ and $Q^{\,\mathrm{sym}}$ in the algorithm. The initial $v^{(0)}$ and $(z_k)_{k \geqslant 0}$ come from $\widetilde{Q}$. The vectors $(v^{(k)})_{k \geqslant 0}$ are produced by using $Q^{\,\mathrm{sym}}$ plus the shifts (z_k). Next, noting that δ_k defined by (34) is invariant if the vector $v^{(k)}$ is replaced by $c\, v^{(k)}$ for every constant $c > 0$, applying (34) to the vector $w^{(0)}$, we return to δ_0 defined by (31). Hence, in what follows, we may ignore (31) and use (34) (for $k \geqslant 0$) only.

Before moving further, let us make a remark.

Remark 26 One may apply Algorithm 25 in the opposite way. Suppose that we are given a symmetric matrix, say $Q^{\,\mathrm{sym}}$. In the special case that the sum of each of the first $N-1$ rows equals zero, then we do not need the h-transform. Just return to Algorithm 25 by setting $\widetilde{Q} = Q^{\,\mathrm{sym}}$. Otherwise, we can construct another matrix $\widetilde{Q}$ using the h-transform of $Q^{\,\mathrm{sym}}$. With these matrices at hand, we can apply Algorithm 25 to compute the maximal eigenpair of $Q^{\,\mathrm{sym}}$ $\big($as well as $\widetilde{Q}\big)$. This indicates that the algorithm seems to be new for the eigenvalue computation of symmetric matrices.

Example 27 Apply Algorithm 25 to the matrix defined by (23) with $N = 7$, the outputs are as follows.

Table 1 Outputs $\big(z_n, v^{(n)}\big)$ of Algorithm 25 at step $n = 0, \ldots, 3$

z_n	$v^{(n)}$
.304256	(.28755, .351484, .378148, .388302, .388302, .378148, .351484, .28755)*
.340851	(.164291, .305267, .407944, .461955, .461955, .407944, .305267, .164291)*
.342146	(.161233, .303016, .408248, .46424, .46424, .408248, .303016, .161233)*
.342148	(.16123, .303013, .408248, .464243, .464243, .408248, .303013, .16123)*

We have

$$\frac{\text{Maximum of } v^{(0)}}{\text{Minimum of } v^{(0)}} \approx 1.35038, \qquad \frac{\text{Maximum of } v^{(3)}}{\text{Minimum of } v^{(3)}} \approx 2.87939.$$

Proof of Algorithm 25 Roughly speaking, the starting matrix $\widetilde{Q}$ obtained by the h-transform is used to compute the maximal eigenvalue of the original Q (equivalently, of $Q^{\rm sym}$); while the symmetrized matrix $Q^{\rm sym}$ is used to compute the maximal eigenvector $g_{\max}\big(Q^{\rm sym}\big)$ and then

$$g_{\max}\big(Q^{\rm sym}\big) = \mathrm{Diag}(\tilde{\mu})^{1/2} g_{\max}(\widetilde{Q}),$$

where $g_{\max}(Q)$ denotes the maximal eigenvector of Q.

(a) First, it is easy to check that the matrix $Q^{\rm sym}$ is symmetric, due to the symmetrizable property: $\mathrm{Diag}(\tilde{\mu})\widetilde{Q} = \widetilde{Q}^{*}\mathrm{Diag}(\tilde{\mu})$. The proof goes as follows.

$$\begin{aligned}\big(\mathrm{Diag}(\tilde{\mu})^{1/2}\widetilde{Q}\,\mathrm{Diag}(\tilde{\mu})^{-1/2}\big)^{*} &= \mathrm{Diag}(\tilde{\mu})^{-1/2}\widetilde{Q}^{*}\mathrm{Diag}(\tilde{\mu})^{1/2}\\ &= \mathrm{Diag}(\tilde{\mu})^{-1/2}\widetilde{Q}^{*}\mathrm{Diag}(\tilde{\mu})\,\mathrm{Diag}(\tilde{\mu})^{-1/2}\\ &= \mathrm{Diag}(\tilde{\mu})^{-1/2}\mathrm{Diag}(\tilde{\mu})\,\widetilde{Q}\,\mathrm{Diag}(\tilde{\mu})^{-1/2}\\ &= \mathrm{Diag}(\tilde{\mu})^{1/2}\widetilde{Q}\,\mathrm{Diag}(\tilde{\mu})^{-1/2}.\end{aligned}$$

(b) Note that the eigenequation

$$\widetilde{Q}g = -\lambda\, g$$

can be rewritten as

$$\big[\mathrm{Diag}(\tilde{\mu})^{1/2}\,\widetilde{Q}\,\mathrm{Diag}(\tilde{\mu})^{-1/2}\big]\big(\mathrm{Diag}(\tilde{\mu})^{1/2}g\big) = -\lambda\big(\mathrm{Diag}(\tilde{\mu})^{1/2}g\big).$$

That is

$$Q^{\rm sym}\,\big(\mathrm{Diag}(\tilde{\mu})^{1/2}g\big) = -\lambda\big(\mathrm{Diag}(\tilde{\mu})^{1/2}g\big).$$

Hence, the symmetrizing transform produces a transform of the eigenpairs

$$\big(\lambda,\ g\big(\widetilde{Q}\big)\big) \to \big(\lambda,\ \mathrm{Diag}(\tilde{\mu})^{1/2}g\big(\widetilde{Q}\big)\big) = \big(\lambda, g\big(Q^{\rm sym}\big)\big). \tag{35}$$

This is the key point of our algorithm. On the one hand, in the non-symmetric case, since $g\big(\widetilde{Q}\big)$ often decays fast, it leads to the use of $Q^{\rm sym}$. In particular, noting that the initial mimic vector for $\widetilde{Q}$ is $\sqrt{\widetilde{\varphi}}$, we should use $w^{(0)} = \sqrt{\mathrm{Diag}(\tilde{\mu})\tilde{\varphi}}$ as the initial vector for $Q^{\rm sym}$ based on (35). The new initial vector avoid the serious problem mentioned at the beginning of this section (Figures 3–8). On the other hand, this also shows that the matrix $\widetilde{Q}$ plays an important role, that is its initial $\sqrt{\widetilde{\varphi}}$. Which gives us not only the initial vector but also the initial estimate z_0 of $\lambda_{\max}\big(\widetilde{Q}\big) = \lambda_{\max}\big(Q^{\rm sym}\big)$. Similarly, the present δ_n given in Algorithm 25 is translated from [11; (15) in the preprint or (A9) in the published version of the paper]. In conclusion, in Algorithm 25, even though it appears only $Q^{\rm sym}$, but not $\widetilde{Q}$, the initial $v^{(0)}$ and the shifts z_k are all come from $\widetilde{Q}$, they can not be deduced directly from $Q^{\rm sym}$ as far as we know. □

Algorithm 28 (Improved) We mention that there are some ways to improve the computation speed in Algorithm 25. For instance, if we set

$$M_{kk}=1,\ M_{kj}=M_{k,j-1}\frac{\tilde{a}_j}{\tilde{b}_{j-1}}\left[=\frac{\tilde{a}_{k+1}\cdots\tilde{a}_j}{\tilde{b}_k\cdots\tilde{b}_{j-1}}=\frac{\tilde{\mu}_k}{\tilde{\mu}_j}\right],\qquad 1\leqslant k+1\leqslant j<N+1,$$

$$\Phi_k=\tilde{\mu}_k\tilde{\varphi}_k=\frac{1}{\tilde{b}_k}+\sum_{k+1\leqslant j<N+1}\frac{\tilde{a}_{k+1}\cdots\tilde{a}_j}{\tilde{b}_k\cdots\tilde{b}_j}=\sum_{k\leqslant j<N+1}\frac{M_{kj}}{\tilde{b}_j},\qquad 0\leqslant k<N+1,$$

then we can rewrite (34) as

$$\delta_k=\sup_{0\leqslant n<N+1}\frac{1}{v_n^{(k)}}\left[\Phi_n\sum_{0\leqslant i\leqslant n}v_i^{(k)}\sqrt{M_{in}}+\sum_{n+1\leqslant j<N+1}\sqrt{M_{nj}}\,\Phi_j v_j^{(k)}\right].\qquad(36)$$

The improved Algorithm 28 is applied by Y.S. Li to the model (23) successfully up to $N=10^4$ using MatLab. The number of the iterations needed by this algorithm is no more than 3. Actually, for $N\geqslant 4500$, up to the six precisely significant digits, the initial z_0 already coincides with λ_0.

The main computational complexity of Algorithm 25 or 28 comes from the formulas (34) or (36), which is due to the use of the operator II of double summation [7; Theorem 2.4 (3)]. To reduce this complexity, we introduce the following algorithm.

Algorithm 29 (Improved)

(1) In Algorithm 25 or 28, replace δ_k by the new one:

$$\zeta_k=\sup_{0\leqslant n<N+1}\frac{1}{\sqrt{\tilde{b}_n}\,v_n^{(k)}-\sqrt{\tilde{a}_{n+1}}\,v_{n+1}^{(k)}}\sum_{j=0}^{n}v_j^{(k)}\sqrt{\frac{M_{jn}}{\tilde{b}_n}},\qquad k\geqslant 0,$$

where $\tilde{a}_{N+1}=0$ and $v_{N+1}^{(k)}:=0$ if $N<\infty$.

(2) Solve equation (32) by using the Thomas algorithm.

Proof. The proof for part (1) is almost the same as the one for Algorithm 25, except for computing ζ_k, here we use the operator I of single summation instead of the double one II used in Algorithm 25 for computing δ_k, plus an application of [7; Theorem 2.4 (2)]. The advantage is the computation becomes simpler but the price we have to pay is that the convergence speed becomes slower. This is not serious since the convergence speed of the shifted inverse iteration is very fast. We will see this point very soon.

In general, the Thomas algorithm may not be applicable in the present situation which is not diagonal dominant. However, by [7; Theorem 2.4 (2)], we have $\lambda_0\geqslant\zeta_k^{-1}$ for every $k\geqslant 0$. Our computation is stopped at some k_0 if $\lambda_0-\zeta_{k_0}<10^{-6}$ (or $|\zeta_{k_0+1}-\zeta_{k_0}|<10^{-6}$) for instance. Otherwise, for $k<k_0$, we have $\lambda_0>\zeta_k^{-1}=z_k$ and moreover the matrix $-Q^{\rm sys}-z_kI$ is invertible.

Hence there is uniquely a solution to equation (32) which means that Thomas algorithm is applicable. It is well known that Thomas algorithm is of O(N), refer to [16] for detail analysis on this point. Note that in our algorithm, the main quantity one may worried is the array $\{M_{jn}\}$, which requires about $N(N-1)/2$ multiplications. However, the array can be regarded as an input, can be fixed at the beginning, without re-computing in the iterations. Therefore, Algorithm 29 is essentially also O(N). Refer also to [26] for further discussion, in which the algorithm is claimed to be O(1) number of the iterations. The conclusion is true, due to the fact proved in [7; Theorem 3.2 and Corollary 3.3] that the initials used here produce upper and lower basic estimates of the eigenvalue up to a universal factor no more than 4 (in terms of the operator I), and no more than 2 in practice (in terms of the operator II). □

To show the power of Algorithm 29, we return to Example 27. The outputs of the new algorithm are given in Table 2.

Table 2 Outputs of Example 27 by Algorithm 29

n	0	1	2	3
z_n	0.253835	0.33544	0.342107	0.342148

Even through the convergence speed is slower but we arrive at the same result as Algorithm 25 in the same steps.

In the past two decades or so, in the study of the estimation of leading eigenvalues, we have used three operators: except II and I mentioned above, there is one more, called difference (differential) operator R (refer to [6, 7] and references within). Among them, the sharpest estimate is deduced by II, the next one is by I, and the last one is by R. The computational complexity goes in the inverse order. In our recent numerical study on the maximal eigenvalue, we have adopted II only for finest estimates, without consider the computational complexity. It is the first time in Algorithm 29 we use the operator I to keep the balance between the sharpness and the computational complexity. Even though it is the easiest in the computation, we do not want to use R here, since on the one hand, it is more or less covered by a more general algorithm, call global one (cf. [11]); and on the other hand, it does not use much of the advantage of the tridiagonal property. Nevertheless, the initial vector $w^{(0)}$ used in Algorithm 25, as well as in Algorithms 28 and 29, comes from an mimic of the principal eigenvector, closely related to the operator II, refer to [7; Theorem 3.2].

Conservative case

For the remainder of this section, we study the following conservative tridiagonal Q-matrix on $E = \{k \in \mathbb{Z}_+ : 0 \leqslant k < N+1\}$:

$$Q = \begin{pmatrix} -c_0 & b_0 & & & & \\ a_1 & -c_1 & b_1 & & 0 & \\ & a_2 & -c_2 & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & b_{N-1} \\ & & & & a_N & -c_N \end{pmatrix}, \tag{37}$$

where $a_k > 0\,(1 \leqslant k < N+1)$, $b_k > 0\,(0 \leqslant k < N)$ and $c_k = a_k + b_k$ (with $a_0 := 0$ and $b_N = 0$ if $N < \infty$). Hence this is a conservative Q-matrix.

Clearly, the maximal eigenvalue for this matrix is $\lambda_0 = 0$ with constant eigenvector 1. We are now going to compute the submaximal eigenvalue $\lambda_1(Q)$.

To state our algorithm, we need an auxiliary tridiagonal matrix on $E_1 := \{k \in \mathbb{Z}_+ : 1 \leqslant k < N+1\}$: $\widetilde{Q} \sim \big(\tilde{a}_k, -\tilde{c}_k, \tilde{b}_k\big)$ defined as follows. First, set $\tilde{b}_1 = a_1 + b_0$. Next, define

$$\begin{aligned} \tilde{b}_k &= a_k + b_{k-1} - \frac{a_{k-1}b_{k-1}}{\tilde{b}_{k-1}}, \quad \tilde{a}_k = a_k + b_{k-1} - \tilde{b}_k, \qquad 2 \leqslant k < N, \\ \tilde{c}_k &= a_k + b_{k-1}, \qquad 1 \leqslant k < N+1, \\ \tilde{a}_N &= \frac{a_{N-1}b_{N-1}}{\tilde{b}_{N-1}}, \qquad \text{if } N < \infty. \end{aligned} \tag{38}$$

As analog of (28) and (29), define

$$\tilde{\mu}_1 = 1, \;\; \tilde{\mu}_n = \tilde{\mu}_{n-1}\frac{\tilde{b}_{n-1}}{\tilde{a}_n}, \qquad 2 \leqslant n < N+1. \tag{39}$$

and

$$\tilde{\varphi}_n = \sum_{k=n}^{N} \frac{1}{\tilde{\mu}_k \tilde{b}_k}, \qquad 1 \leqslant n < N+1, \tag{40}$$

where

$$\tilde{b}_N = \tilde{c}_N - \tilde{a}_N = a_N + b_{N-1} - \tilde{a}_N$$

if $N < \infty$. Besides, we need also the following symmetrized matrix on E_1.

$$Q^{\rm sym} = \mathrm{Diag}(\tilde{\mu})^{1/2}\, \widetilde{Q}\, \mathrm{Diag}(\tilde{\mu})^{-1/2}$$
$$= \begin{pmatrix} -\tilde{c}_1 & \sqrt{\tilde{a}_2\tilde{b}_1} & & & 0 \\ \sqrt{\tilde{a}_2\tilde{b}_1} & -\tilde{c}_2 & \sqrt{\tilde{a}_3\tilde{b}_2} & & \\ & \sqrt{\tilde{a}_3\tilde{b}_2} & -\tilde{c}_3 & \sqrt{\tilde{a}_4\tilde{b}_3} & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \sqrt{\tilde{a}_N\tilde{b}_{N-1}} \\ & 0 & & & \sqrt{\tilde{a}_N\tilde{b}_{N-1}} & -\tilde{c}_N \end{pmatrix}. \tag{41}$$

The algorithm below is almost the same as Algorithm 25 with very slight modification. Here we repeat it for safe. We use $\tilde{\mu}$, $\tilde{\varphi}$, and $Q^{\rm sym}$ defined by (38) – (41), respectively.

Algorithm 30 (1) Let $w_i^{(0)} = \sqrt{\tilde{\mu}_i\tilde{\varphi}_i}$, $i \in E_1$. Define

$$v^{(0)} = \frac{w^{(0)}}{\sqrt{w^{(0)*}w^{(0)}}}, \qquad z_0 = \frac{1}{\delta_0}, \tag{42}$$

$$\delta_0 = \sup_{1\leqslant k<N+1}\left[\sqrt{\tilde{\varphi}_k}\sum_{i=1}^{k}\tilde{\mu}_i\sqrt{\tilde{\varphi}_i} + \frac{1}{\sqrt{\tilde{\varphi}_k}}\sum_{k+1\leqslant j<N+1}\tilde{\mu}_j\tilde{\varphi}_j^{3/2}\right]. \tag{43}$$

(2) For each $k \geqslant 1$, solve $w^{(k)}$:

$$\big(-Q^{\rm sym} - z_{k-1}I\big)\, w^{(k)} = v^{(k-1)} \qquad \text{on } E_1, \tag{44}$$

and define

$$v^{(k)} = \frac{w^{(k)}}{\sqrt{w^{(k)*}w^{(k)}}}, \qquad z_k = \frac{1}{\delta_k}, \tag{45}$$

$$\delta_k = \sup_{1\leqslant n<N+1}\frac{\sqrt{\tilde{\mu}_n}}{v_n^{(k)}}\left[\tilde{\varphi}_n\sum_{i=1}^{n}\sqrt{\tilde{\mu}_i}\, v_i^{(k)} + \sum_{n+1\leqslant j<N+1}\sqrt{\tilde{\mu}_j}\,\tilde{\varphi}_j v_j^{(k)}\right]. \tag{46}$$

Then

$$v^{(k)} \to g_1 \quad \text{and} \quad z_k \to \lambda_1 \qquad \text{as } k \to \infty,$$

where $(\lambda_1,\ g_1)$ is the minimal eigenpair of $-Q^{\rm sym}$. Furthermore, the sub-minimal eigenvalue of $-Q$ equals λ_1.

We mention that Algorithms 28 and 29 are also meaningful for Algorithm 30.

The next example, closely related to Example 27, is used to illustrate the algorithm.

Example 31 The matrix on E with $N = 7$ is the same as in (23) except $c_0 = 2$ and $c_7 = 1$ for the conservativity. The outputs $\left(z_n, v^{(n)}\right)$ of this example by Algorithm 30 are given in Table 3.

Table 3 Outputs $\left(z_n, v^{(n)}\right)$ of Algorithm 30 at step $n = 0, \ldots, 3$

z_n	$v^{(n)}$
.342108	$(.313645, .38262, .409984, .417473, .409984, .38262, .313645)^*$
.385369	$(.194385, .35518, .460762, .497514, .460762, .35518, .194385)^*$
.386872	$(.191344, .353555, .461939, .499997, .461939, .353555, .191344)^*$
.386874	$(.191342, .353553, .46194, .5, .46194, .353553, .191342)^*$

The last line of the table represents the minimal eigenpair (λ_1, g_1) of $-Q^{\mathrm{sym}}$. While the sub-minimal eigenpair of the original $-Q$ is

$$\lambda_1 = .386874,$$
$$g_1 = (-16,\ -12.905,\ -8.8612,\ -5.12522,\ -2.26582,\ -.397825,\ .613126,\ 1)^*.$$

It is interesting to compare the difference of their amplitudes for these eigenvectors $v^{(3)}$ and g_1: 0.3 and 17, respectively, for such a small size of matrices. Thus, the difference in the computations should be serious for large scale of matrices. From the table, it follows that

$$\frac{\text{Maximum of } v^{(0)}}{\text{Minimum of } v^{(0)}} \approx 1.33103, \qquad \frac{\text{Maximum of } v^{(3)}}{\text{Minimum of } v^{(3)}} \approx 2.61313.$$

This result is quite close to the comparison given in Example 27.

Proof of Algorithm 30 Suppose we are given the tridiagonal matrix (37) on $E = \{k : 0 \leqslant k < N + 1\}$, $N \leqslant \infty$. When $N = \infty$, one can ignore the boundary condition $b_N = 0$. As in [7; (5.1)], define a dual tridiagonal matrix $\widehat{Q} \sim (\hat{a}_k, -\hat{c}_k, \hat{b}_k)$ as follows.

$$\begin{cases} \hat{a}_i = b_{i-1}, \\ \hat{c}_0 = 0,\ \hat{c}_i = \hat{a}_i + \hat{b}_i = a_i + b_{i-1}, \\ \hat{b}_0 = 0,\ \hat{b}_i = a_i, \quad 1 \leqslant i < N + 1. \end{cases}$$

Alternatively,

$$\widehat{Q} = \begin{pmatrix} -\hat{c}_0 & \hat{b}_0 & & & & \\ \hat{a}_1 & -\hat{c}_1 & \hat{b}_1 & & 0 & \\ & \hat{a}_2 & -\hat{c}_2 & \hat{b}_2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & \hat{b}_{N-1} \\ & & & & \hat{a}_N & -\hat{c}_N \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & 0 & & & & \\ b_0 & -(a_1+b_0) & a_1 & & \mathbf{0} & \\ & b_1 & -(a_2+b_1) & a_2 & & \\ & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & \ddots & \ddots & a_{N-1} \\ & & & & b_{N-1} & -(a_N+b_{N-1}) \end{pmatrix}. \tag{47}$$

Next, define

$$M = \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_N \\ & \mu_1 & \mu_2 & \cdots & \mu_N \\ & & \mu_2 & \cdots & \mu_N \\ & & & \ddots & \vdots \\ \mathbf{0} & & & & \mu_N \end{bmatrix} \Rightarrow M^{-1} = \begin{bmatrix} \mu_0^{-1} & -\mu_0^{-1} & & & \mathbf{0} \\ & \mu_1^{-1} & -\mu_1^{-1} & & \\ & & \mu_2^{-1} & -\mu_2^{-1} & \\ & & & \ddots & \ddots \\ \mathbf{0} & & & & \mu_N^{-1} \end{bmatrix}.$$

Then, we have $\widehat{Q} = MQM^{-1}$. To check this, one may avoid the computation of the inverse matrix M^{-1}, but use the dual equation $MQ = \widehat{Q}M$:

$$MQ = \begin{pmatrix} 0 & 0 & & & & \mathbf{0} \\ \mu_0 b_0 & -\mu_0 b_0 & & & & \\ & \mu_1 b_1 & -\mu_1 b_1 & & & \\ & & \ddots & \ddots & & \\ \mathbf{0} & & & \mu_{N-1} b_{N-1} & -\mu_{N-1} b_{N-1} & \end{pmatrix} = \widehat{Q}M.$$

This technique was mentioned in [7; Above Examples 7.5]. By Theorem 10, $\widehat{Q}$ and Q have the same spectrum. In particular, the maximal eigenpair $(0, \mathbb{1})$ of Q corresponds to the one $(0, M\mathbb{1})$ of $\widehat{Q}$. By removing the first row and the first column of $\widehat{Q}$, we obtain a matrix on E_1:

$$\widehat{Q}_1 = \begin{bmatrix} -\hat{c}_1 & \hat{b}_1 & & & \mathbf{0} \\ \hat{a}_2 & -\hat{c}_2 & \hat{b}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \hat{b}_{N-1} \\ \mathbf{0} & & & \hat{a}_N & -\hat{c}_N \end{bmatrix}$$

$$= \begin{bmatrix} -(a_1+b_0) & a_1 & & & & \mathbf{0} \\ b_1 & -(a_2+b_1) & a_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & a_{N-1} & \\ \mathbf{0} & & & b_{N-1} & -(a_N+b_{N-1}) \end{bmatrix}.$$

Now, the spectrum of Q ignoring the trivial eigenvalue 0 coincides with the spectrum of $\widehat{Q}_1$. The sub-minimal eigenvalue of $-Q$ coincides the minimal eigenvalue of $-\widehat{Q}_1$.

Note that the matrix $\widehat{Q}_1$ has bilateral Dirichlet boundaries, at 0 and at $N+1$ if $N<\infty$. In order to apply our technique, we need to adopt the h-transform, removing the killing at 1. This transform is presented by (38). Having constructed the transformed matrix $\widetilde{Q}$, one can easily deduce the symmetrizing one $Q^{\,\mathrm{sym}}$ and complete the Algorithm 30, in parallel to Algorithm 25. □

We remark that for a complex matrix, if it is not real, for keeping the real spectrum, the duality may not be suitable. To see this, simply look at second matrix in (47). Even though we may assume that $a_k b_{k-1}>0$, but then a_k+b_{k-1} may still not be real and so the diagonals in the dual matrix may not be real, In general, for a complex matrix, we can apply the h-transform first, if the deduced matrix is conservative, then we can apply Algorithm 30 to compute the sub-maximal eigenvalue of Q.

5 Differential operators

Throughout this section, denote by $\mathscr{C}^m(\mathbb{R}^d)$ the set of functions on $\mathbb{R}^d$ with continuous derivatives up to order m. In the first part of this section, we study the Hermitizable and isospectral problems for second-order complex differential operators having the form:

$$L=D^*(aD)-c=\sum_{j,k=1}^{d}\big(\partial_j+b_j(x)\big)\big[a_{jk}(x)\big(\partial_k+b_k(x)\big)\big]-c, \tag{48}$$

where $\partial_j=\mathrm{d}/\mathrm{d}x_j$, c is a $\mathscr{B}(\mathbb{R}^d)$-measurable function, a is a $d\times d$ matrix, assumed to be in $\mathscr{C}^1(\mathbb{R}^d)$, and so is the vector $b=(b_j(x))$. Set $D_j=\partial_j+b_j$. For the later use, we now express L into more explicit forms in the order of differentials*:

$$\begin{aligned} L&=\partial^*(a\partial)+b^*(a+a^*)\partial+D^*(ab)-c &(49)\\ &=a\cdot\partial\partial^*+(D^*a+b^*a^*)\partial+D^*(ab)-c, &(50)\end{aligned}$$

*The term $b^*a^*\partial$ was missed in each of the two lines below in the published version. To be careful, we write the details here.

$$\begin{aligned} D^*aDf&=D^*(a\partial f+abf)\\ &=\partial^*(a\partial f)+b^*(a\partial f)+\partial^*\big(\hat{b}f\big)+b^*\hat{b}f\quad\big(\hat{b}:=ab\big)\\ &=(a\cdot\partial\partial^*)f+(\partial^*a)\partial f+(b^*a)\partial f+\hat{b}\cdot\partial f+\big(\partial^*\hat{b}\big)f+b^*\hat{b}f\\ &=(a\cdot\partial\partial^*)f+\big(D^*a+b^*a^*\big)\partial f+D^*(ab)f.\end{aligned}$$

We have thus obtained (50). This correction costs a little change in Theorem 33 below—[2019-06-27]

here for given matrices $a = (a_{ij})$ and $b = (b_{ij})$, the product $a \cdot b$ is defined to be $\sum_{i,j} a_{ij}b_{ij}$, regarded as an analog of the inner product of two vectors.

Theorem 32 Let Ω (maybe unbounded)$\subset \mathbb{R}^d$ with Dirichlet boundary condition on $\partial\Omega$. Then the operator defined in (48) is selfadjoint (formally) on $L^2(\mathrm{d}x)$ iff a is Hermitian: $a^H(:= \bar{a}^*) = a$, b is purely imaginary: $\bar{b} = -b$ and c is real: $\bar{c} = c$. In this case,

$$(-Lf,\, f) = (aDf,\, Df) + (cf, f).$$

Proof. Let $f, g \in \mathscr{C}^2(\mathbb{R}^d)$. For each fixed j, we have

$$\begin{aligned}\Big(\sum_k a_{jk}(\partial_k + b_k)f,\ g\Big) &= \sum_k (D_k f,\ \bar{a}_{jk}g)\\ &= \sum_k [f a_{jk}\bar{g}]\big|_{\partial\Omega} - \sum_k \big[(f,\ \partial_k(\bar{a}_{jk}g)) - (f,\ \bar{b}_k\bar{a}_{jk}g)\big]\\ &= -\sum_k \big(f,\ (\partial_k - \bar{b}_k)(\bar{a}_{jk}g)\big) \text{ (by boundary condition)}\\ &= -\sum_k \big(f,\ (\partial_k - \bar{b}_k)(a^H_{kj}g)\big). \qquad (51)\end{aligned}$$

Next, applying this to the identity matrix a, we obtain

$$(D_j f,\ g) = -\big(f,\ (\partial_j - \bar{b}_j)g\big). \qquad (52)$$

Thus, for fixed j, with $\tilde{f} = \sum_k a_{jk}D_k f$ and $\tilde{g}_j = \big(\partial_j - \bar{b}_j\big)g$, we have

$$\begin{aligned}\big(D_j\tilde{f},\ g\big) &= -\big(\tilde{f},\ (\partial_j - \bar{b}_j)g\big) \quad \text{(by (52))}\\ &= -\Big(\sum_k a_{jk}D_k f,\ \tilde{g}_j\Big) \quad \text{(by definition of } \tilde{f} \text{ and } \tilde{g})\\ &= \sum_k \big(f,\ (\partial_k - \bar{b}_k)\big(a^H_{kj}\tilde{g}_j\big)\big) \quad \text{(by (51))}\\ &= \sum_k \big(f,\ (\partial_k - \bar{b}_k)\big(a^H_{kj}(\partial_j - \bar{b}_j)g\big)\big) \quad \text{(by definition of } \tilde{g}).\end{aligned}$$

Summing up over j and adding the term c, it follows that

$$(Lf, g) = ([(\partial + b)^*(a(\partial + b)) - c]f,\ g) = \sum_{j,k} \big(f,\ \big[(\partial_k - \bar{b}_k)\big(a^H_{kj}(\partial_j - \bar{b}_j)\big) - \bar{c}\big]g\big).$$

Finally, for the selfadjointness:

$$\big(\partial - \bar{b}\big)^*\big[a^H\big(\partial - \bar{b}\big)\big] - \bar{c} = (\partial + b)^*[a(\partial + b)] - c,$$

we obtain the conditions $a^H = a$, $\bar{b} = -b$ and $\bar{c} = c$. $\square$

In the context of Markov processes, as mentioned in Section 2, the symmetrizable problem for Markov chains goes back to [19]. For diffusions, it goes back to [20]. In which, the author adopted geometric approach. A particular result says that if the diffusion coefficient is the identity matrix, then the process is symmetrizable (reversible) iff the drift coefficient should be a gradient of a (conservative) potential. The operator in (48) is a slight extension of [18]. It is less popular than the special case that $b(x) \equiv 0$. However, the latter one is rather restrictive: in dimension one, it has to be real for the selfadjointness. We use "formal" in Theorem 32 since it is only the first step for the property, we have not specified a domain of the operator. To which, some additional conditions on the coefficients of the operator are often required. See [18], [1], and [2] for more details and additional references. This seems not too hard, as mentioned in the matrix case, since we are mainly interested in either maximal or minimal domains as did in [8].

We now study the h-transform in Lemma 8 for the operator having the form (48). The purpose of the next result is to remove the potential term from (50).

Theorem 33 Let L be given by (48) and set

$$L^0 = L - D^*(ab) + c = a \cdot \partial\partial^* + (D^*a + b^*a^*)\partial.$$

Next, let h be L-harmonic: $Lh = 0$, $h \neq 0$ (a.e.). Then the h-transform given in Lemma 8 transfers L to

$$\widetilde{L} = L^0 + \mathbb{1}_{[h\neq 0]}\frac{1}{h}(\partial h)^*(a+a^*)\partial = a\cdot\partial\partial^* + \left[D^*a + b^*a^* + \mathbb{1}_{[h\neq 0]}\frac{1}{h}(\partial h)^*(a+a^*)\right]\partial.$$

In particular, when $h = \exp\psi$,

$$\widetilde{L} = a \cdot \partial\partial^* + \big[D^*a + b^*a^* + (\partial\psi)^*(a + a^*)\big]\partial.$$

Moreover, L and $\widetilde{L}$ are both selfadjoint or not, simultaneously.

Proof. The final assertion comes from Theorem 9.

By (50), we have

$$\begin{aligned}
\frac{1}{h}L(hf) &= \frac{1}{h}\big[(a \cdot \partial\partial^*)(hf) + (D^*a + b^*a^*)\partial(hf) + (D^*(ab) - c)hf\big] \\
&= \big[(a \cdot \partial\partial^*)f + (D^*a + b^*a^*)\partial f\big] \\
&\quad + \frac{f}{h}\big[(a \cdot \partial\partial^*)h + (D^*a + b^*a^*)\partial h + (D^*(ab) - c)h\big] \\
&\quad + \frac{1}{h}\big[(\partial h)^*a\partial f + (\partial f)^*a\partial h\big] \\
&=: I + II + III.
\end{aligned}$$

Note that

$$\begin{aligned} I &= L^0 f, \\ II &= \frac{f}{h} Lh = 0 \text{ (by harmonic assumption)}, \\ III &= \frac{1}{h}(\partial h)^*(a + a^*)\partial f \text{ (since } (\partial f)^* a \partial h = (\partial h)^* a^* \partial f). \end{aligned}$$

Combining these facts together, we obtain the required assertion. □

We now go to the second part of this section. It goes from $\widetilde{L}$ to L, as an analog of [15; Theorem 1.1 (2), Theorem 3.6 and Corollary 3.7].

Theorem 34 Let $\widetilde{L} = \partial^* \tilde{a} \partial - \tilde{c}$ having domain $\mathscr{D}(\widetilde{L}) \subset L^2(\tilde{\mu})$. Then for each complex function $h \in \mathscr{C}^2(\mathbb{R}^d)$, $h \neq 0$, μ-a.e., $\widetilde{L}$ is L^2-isospectral to $L = L^h$:

$$\begin{aligned} L^h &= \widetilde{L} - \frac{1}{h}(\partial h)^*(\tilde{a} + \tilde{a}^*)\partial + \left[\frac{2}{h^2}(\partial h)^* \tilde{a} - \frac{1}{h}\partial^* \tilde{a}\right](\partial h), \\ \mathscr{D}(L^h) &= \{f : f/h \in \mathscr{D}(\widetilde{L})\}. \end{aligned}$$

In particular, if we set $h = \exp[-\psi]$, then[†] $L^h = L_\psi$:

$$\begin{aligned} L_\psi = \widetilde{L} + (\partial\psi)^*(\tilde{a} + \tilde{a}^*)\partial + \Big[(\tilde{a} \cdot \partial\partial^*)\psi + (\partial\psi)^* \tilde{a}(\partial\psi) + \partial^* \tilde{a}(\partial\psi)\Big], \\ \mathscr{D}(L_\psi) = \{f : f \exp[\psi] \in \mathscr{D}(\widetilde{L})\}. \end{aligned}$$

Moreover, L and $\widetilde{L}$ are both selfadjoint or not, simultaneously.

Proof. Again, the final assertion is a consequence of Theorem 9.

Note that

$$(\tilde{a}\partial)\left(\frac{f}{h}\right) = \tilde{a}\left(\frac{1}{h}\partial f + f\partial\left(\frac{1}{h}\right)\right) = \frac{1}{h}\tilde{a}\partial f + f\tilde{a}\partial\left(\frac{1}{h}\right).$$

We have

$$\begin{aligned} (\partial^* \tilde{a} \partial)\left(\frac{f}{h}\right) &= \partial^*\left(\frac{1}{h}\tilde{a}\partial f + f\tilde{a}\partial\left(\frac{1}{h}\right)\right) \\ &= \frac{1}{h}(\partial^* \tilde{a} \partial) f + \left(\partial\left(\frac{1}{h}\right)\right)^* \tilde{a}\partial f + f\partial^* \tilde{a}\partial\left(\frac{1}{h}\right) + (\partial f)^* \tilde{a}\partial\left(\frac{1}{h}\right) \end{aligned}$$

Because

$$(\partial f)^* \tilde{a}\partial\left(\frac{1}{h}\right) = \left(\tilde{a}\partial\left(\frac{1}{h}\right)\right)^* (\partial f) = \left(\partial\left(\frac{1}{h}\right)\right)^* \tilde{a}^* (\partial f),$$

[†]The term $(\tilde{a} \cdot \partial\partial^*)\psi$ was missed in the published version–[2019-06-27]

and then

$$\left(\partial\left(\frac{1}{h}\right)\right)^* \tilde{a}(\partial f) + (\partial f)^* \tilde{a}^* \partial\left(\frac{1}{h}\right) = \left(\partial\left(\frac{1}{h}\right)\right)^* (\tilde{a} + \tilde{a}^*)(\partial f),$$

we obtain

$$(\partial^* \tilde{a} \partial)\left(\frac{f}{h}\right) = \frac{1}{h}(\partial^* \tilde{a} \partial) f + f \partial^* \tilde{a} \partial\left(\frac{1}{h}\right) + \left(\partial\left(\frac{1}{h}\right)\right)^* (\tilde{a} + \tilde{a}^*)(\partial f).$$

Hence

$$h\widetilde{L}\left(\frac{f}{h}\right) = \widetilde{L} f + h\left(\partial\left(\frac{1}{h}\right)\right)^* (\tilde{a} + \tilde{a}^*)\partial f + h\partial^* \tilde{a} \partial\left(\frac{1}{h}\right) f.$$

Next, because

$$\partial\left(\frac{1}{h}\right) = -\frac{1}{h^2}\partial h, \qquad \tilde{a}\partial\left(\frac{1}{h}\right) = -\frac{1}{h^2}\tilde{a}\partial h,$$
$$\partial^* \tilde{a} \partial\left(\frac{1}{h}\right) = \frac{2}{h^3}(\partial h)^* \tilde{a}(\partial h) - \frac{1}{h^2}\partial^* \tilde{a} \partial h,$$

we obtain the expression of

$$L^h f := h\widetilde{L}\left(\frac{f}{h}\right).$$

Then the expression of L_ψ now follows immediately since[‡]

$$\partial^* \tilde{a} \partial h = h\big[-(\tilde{a} \cdot \partial\partial^*)\psi + (\partial\psi)^* \tilde{a}(\partial\psi) - \partial^* \tilde{a}(\partial\psi)\big].$$

Finally, the proof of the isospectrum is almost the same as those of [15; Lemma 1.3]. First, we have

$$\big(\tilde{f}, \tilde{f}\big)_{\tilde{\mu}} = (f/h, f/h)_{\tilde{\mu}} = (f, f)_\mu.$$

Next, we also have

$$\big(\widetilde{L}\tilde{f}, \tilde{f}\big)_{\tilde{\mu}} = \big(\widetilde{L}(f/h), f/h\big)_{\tilde{\mu}} = \big((L^h f)/h, f/h\big)_{\tilde{\mu}} = \big(L^h f, f\big)_\mu.$$

We have thus complete the proof of the theorem. □

Remark 35 By using the multiplying operator H defined in Theorem 9, the h-transform defined in Theorem 34 can be expressed by

$$L^h = H\widetilde{L}H^{-1}.$$

This means that L^h is similar to $\widetilde{L}$ and hence have the same spectrum. Furthermore, the eigenpair $(\lambda, \tilde{g})$ of $\widetilde{L}$:

$$\widetilde{L}\tilde{g} = \lambda\tilde{g}$$

is transferred to the eigenpair $(\lambda, H\tilde{g})$ of L^h:

$$L^h(H\tilde{g}) = H\widetilde{L}H^{-1}(H\tilde{g}) = \lambda(H\tilde{g}).$$

[‡]The term $-(\tilde{a} \cdot \partial\partial^*)\psi$ was missed in the published version below

Applying Theorem 34 to

$$\tilde{a}(x) = \frac{1}{2}e^{-|x|^2}I,$$

where I is the $d \times d$ identity matrix, we obtain the following result.

Corollary 36 For each complex function $\psi \in \mathscr{C}^2(\mathbb{R}^d)$, the operator

$$L_\psi = e^{-|x|^2}\left[\frac{1}{2}\partial^*\partial + (\partial\psi - x)^*\partial + \left[\frac{1}{2}\partial^*\partial\psi + \left(\frac{1}{2}\partial\psi - x\right)^*(\partial\psi)\right]\right]$$

is isospectral to the Ornstein-Uhlenbeck operator $\widetilde{L}$:

$$\widetilde{L} = e^{-|x|^2}\left(\frac{1}{2}\partial^*\partial - x^*\partial\right).$$

Hence L_ψ and $\widetilde{L}$ have the same discrete spectrum.

It is clear that, based on Theorem 9, Corollary 36 gives us a typical example for constructing a large class of Hermitizable complex operators L. We now construct more explicit examples. As an application of Theorem 34 and [8; Example 7.6], we obtain the following result.

Example 37 Consider the operator $\widetilde{L}$ on $\mathbb{R}$:

$$\widetilde{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - \tilde{c}(x), \qquad \tilde{c}(x) = \frac{1}{4}|x|^{2\alpha-2} + \frac{\alpha-1}{2}|x|^{\alpha-2}, \qquad \alpha \in \mathbb{N}$$

with domain $\mathscr{D}\big(\widetilde{L}\big)$. Then for each complex $\psi \in \mathscr{C}^2(\mathbb{R})$, $\widetilde{L}$ is isospectral to the operator

$$L_\psi = \widetilde{L} + 2\psi'(x)\frac{\mathrm{d}}{\mathrm{d}x} + \psi'(x)^2 + \psi''(x),$$
$$\mathscr{D}(L_\psi) = \big\{f : f\exp[\psi] \in \mathscr{D}\big(\widetilde{L}\big)\big\}.$$

Moreover, the spectrum of these operators are both discrete whenever $\alpha > 1$, and are not so if $\alpha = 1$.

Corresponding to the particular $\tilde{c}(x) = x^2$ in Example 37, we have the so-called harmonic oscillator $\widetilde{L}$. Then the next result is an application of [8; Example 7.7].

Example 38 For the harmonic oscillator $\widetilde{L}$, the conclusions of Example 37 hold.

Added in proof After submitting the manuscript, the book [21] caught the attention of the author. Perhaps the Hermitizable tool introduced in this article is useful for further development of matrix mechanics.

Acknowledgments The author thanks Ms Yue-Shuang Li for her assistance. In particular, she has proved a partial answer for the existence of a positive solution to the h-transform in the tridiagonal case. The author also thanks Professor Tao Tang for sending to him (in March, 2018) a preprint of [26] where the problem on non-symmetric tridiagonal matrices arises. The author luckily found (within a week) a solution to the open question, it is now presented in Section 4. The careful corrections by the referees are also acknowledged. The results were presented in four times in January and March, 2018 at our seminar. The author obtained many helpful suggestions from our research group. The results have been reported at Beijing Normal U. (2018/3), Fudan U. (2018/4), Shanghai Jiaotong U. (2018/4), Fujian Normal U. (2018/4), USTC (2018/4), Jiangsu Normal U. (2018/4), Zhejiang U. (2018/5), Southwest Jiaotong U. (2018/7), Sichuan U. (2018/7), Workshop on Stochastic Analysis (FNU) (2018/4), and Workshop on Probability Theory its Applications (HUAS) (2018/7). The author acknowledge the following professors and their institutes for the invitation and financial support: Zeng-Hu Li and Zhong-Wei Tang, Da-Qian Li and Wei-Guo Gao, Dong Han, Huo-Nan Lin and Jian Wang, Jia-Yu Li, Tu-Sheng Zhang, Ying-Chao Xie, Gang Bao, Wei-Ping Li, An-Min Li, Ke-Ning Lu, Lian-Gang Peng, Wei-Nian Zhang, Xu Zhang, Xiang-Qun Yang and Xu-Yan Xiang. Research supported in part by National Natural Science Foundation of China (Grant No. 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Berezanskii, Yu.M. and Samoilenko, V.G. (1981). *On the self-adjointness of differential operators with finitely or infinitely many variables, and evolution equations*. Russian Math. Surveys. 36(5): 1-62.

[2] Brustentsev, A.G. (2004). *Selfadjointness of elliptic differential operators in $L_2(G)$, and correction potentials*. Trans. Moscow Math. Soc. 65, 31–61.

[3] Cao, Z.H. (1983). *Eigenvalue Problem of Matrices* (In Chinese.) Shanghai Press of Sci. & Tech.

[4] Chen, M.F. (1991). *Exponential L^2-convergence and L^2-spectral gap for Markov processes*. Acta Math. Sin., New Series 7(1): 19–37.

[5] Chen, M.F. (2004). *From Markov Chains to Non-Equilibrium Particle Systems*. World Scientific, Singapore, 2nd Ed. (1st Ed., 1992).

[6] Chen, M.F. (2005). *Eigenvalues, Inequalities, and Ergodic Theory*. Springer, London.

[7] Chen, M.F. (2010). *Speed of stability for birth–death processes*. Front Math China, 5(3): 379–515.

[8] Chen, M.F. (2014). *Criteria for discrete spectrum of 1D operators*. Commu. Math. Stat. 2: 279–309.

[9] Chen, M.F. (2016). *Efficient initials for computing the maximal eigenpair*. Front. Math. China 11(6): 1379–1418. See also volume 4 in the middle of the author's homepage:

http://math0.bnu.edu.cn/~chenmf

A package based on the paper is available on CRAN now (by X.J. Mao). One may check it through the link:
https://cran.r-project.org/web/packages/EfficientMaxEigenpair/index.html
A MatLab package is also available, see the author's homepage above.

[10] Chen, M.F. (2017a). *The charming leading eigenpair*. Adv. Math. (China) 46(4): 281–297.

[11] Chen, M.F. (2017b). *Global algorithms for maximal eigenpair*. Front. Math. China 12(5): 1023–1043.

[12] Chen, M.F. (2017c). *Trilogy on computing maximal eigenpair*. In: Yue, W., Li, Q. L., Jin, S., Ma, Z., eds. Queueing Theory and Network Applications. QTNA 2017. Lecture Notes in Comput. Sci., Vol. 10591. Cham: Springer, 312–329

[13] Chen, M.F. (2018a). *Mathematical Topics motivated from statistical physics* (I) (In Chinese). To appear in Sci. Sin. Math.

[14] Chen, M.F. (2018b). *Mathematical Topics motivated from statistical physics* (II)(In Chinese). To appear in Sci. Sin. Math.

[15] Chen, M.F. and Zhang, X. (2014). *Isospectral operators*. Commu Math Stat 2, 17–32.

[16] Frolov, A.V. et al. *Thomas algorithm, pointwise version*. http://algowiki-project.org/en/Thomas_algorithm,_pointwise_version

[17] Hou, Z.T. and Chen, M.F. (1980). *Markov Processes and field theory* (Abstract). Kuoxue Tongbao 25(10): 807–811. Complete version appeared in [23; pages 194–242].

[18] Kato, T. (1981). *Remarks on the selfadjointness and related problems for differential operators.* In I.W. Knowles and R. Lewis (Eds): Spectral Theory of Differential Operators, North Holland, 1981, 253–266.

[19] Kolmogorov, A.N. (1936). *Zur Theorie der Markoffschen Ketten*. Math. Ann., 112: 155–160. English translation: On the theory of Markov ehains. Article 21 in Selected Works of A.N. Kolmogorov, Vol. II: Probability Theory and Mathematical Statistics, 182–187, edited by A.N. Shiryayev. Nauka, Moscow 1986. Translated by G. Undquist. Springer 1992.

[20] Kolmogorov, A.N. (1937). *Zur Umkehrbarkeit der statistischen Naturgesetze*. Math. Ann. 113: 766–772. English translation: On the reversibility of the statistical laws of nature. Article 24 in "Selected Works of A.N. Kolmogorov", Vol. II: 209–215.

[21] Ludyk, G. (2018). *Quantum mechanics in matrix form* (undergraduate lecture notes in physics). Berlin: Springer

[22] Nino, A., Munoz-Caro, C. and Reyes, S. (2011). *A concurrent object-oriented approach to the eigenproblem treatment in shared memory multicore environments*. LNCS 6782, 630–642

[23] Qian, M. and Hou, Z.T. (Eds) (1979). *Reversible Markov Processes* (in Chinese). Changsha: Hunan Sci. Press.

[24] Schrödinger, E. (1931). *Über die Umkehrung der Naturgesetze*. Sitzungsber. Preuss. Akad. Wiss., Phys.-Math. KI., 12 März. 144–153.

[25] Shukuzawa, O., Suzuki, T., Yokota, I. (1996). *Real tridiagonalization of Hermitian matrices by modified Householder transformation*. Proc. Japan. Acad. Ser. A, 72, 102–103.

[26] Tang, T. and Yang, J. (2018). *Computing the maximal eigenpairs of large size tridiagonal matrices with* $\mathcal{O}(1)$ number of iterations. Numer. Math. Theor. Meth. Appl. 11 (4): 877–894.

[27] Varga, R.S. (2004). *Geršgorin and His Circles*. Berlin: Springer.

Mu-Fa Chen
School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems (Beijing Normal University), Ministry of Education, Beijing 100875, The People's Republic of China.
E-mail: mfchen@bnu.edu.cn
Home page: http://math0.bnu.edu.cn/~chenmf/main_eng.htm

Front. Math. China 2019, 14(3), 493–519
https://doi.org/10.1007/s11464-019-0769-5

Development of powerful algorithm for maximal eigenpair

Mu-Fa CHEN, Yue-Shuang LI

School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems(Beijing Normal University), Ministry of Education, Beijing 100875, China

Abstract Based on a series of recent papers, a powerful algorithm is reformulated for computing the maximal eigenpair of self-adjoint complex tridiagonal matrices. In parallel, the same problem in a particular case for computing the sub-maximal eigenpair is also introduced. The key ideas for each critical improvement are explained. To illustrate the present algorithm and compare it with the related algorithms, more than 10 examples are included.

Keywords Powerful algorithm, maximal eigenpair, sub-maximal eigenpair, Hermitizable tridiagonal matrix
MSC 15A18, 65F15, 93E15, 60J27

1 Introduction. A powerful algorithm and a typical example

Matrix eigenvalues play an important role in many areas, not only in mathematics but also in quantum mechanics. In the past 170 years and more, a large number of publications, as well as libraries have been devoted to the study on eigenproblems. Refer to [13, 19] and references therein. An algorithm which is closely related to the aim of the paper is the so-called Householder's decomposition or reduction (with the other two algorithms for matrix eigenproblems: the Krylov subspace iteration methods and the QR algorithm, they were selected into [10] as the 10 top algorithms in the 20th century). About the decomposition technique, six algorithms were surveyed in [17], including Householder transformation [14]. It transforms an Hermitian matrix into a real symmetric tridiagonal one. There are three aspects of contributions in the algorithm stated below. The first one is an extension of the Householder transformation from Hermitian to Hermitizable. This is an easier part ([8; Theorem 24]). The wsecond one is reducing further to a birth-death type Q-matrix (i.e. the tridiagonal matrix having positive sub-diagonals). This enables us to use

Received March 29, 2019; accepted April 30, 2019
Corresponding author: Yue-Shuang LI, E-mail: liyueshuang@bnu.edu.cn

some long cumulated results in the study on probability theory. As will be seen in §4, it takes a long trip to arrive at the present formulation of Algorithm 1 which is very efficient and theoretically complete. An algorithmic program for birth–death type Q-matrix designed based on computational complexity is the third contribution to Algorithm 1. The main aim of the paper is to illustrate the power of the algorithm and to explain the main steps in the development of the algorithm.

To move further, let us define the Hermitizable matrix [8] first. Let

$$E = \{k \in \mathbb{Z} : \ 0 \leqslant k < N + 1\} (N \leqslant \infty).$$

A matrix $A = (a_{ij} : i, j \in E)$ is called Hermitizable if there exists a positive measure $(\mu_i : i \in E)$ such that

$$\mu_i a_{ij} = \mu_j \bar{a}_{ji}, \qquad i, j \in E,$$

where $\bar{a}$ denotes the conjugate of a. With μ at hand, the matrix $(\sqrt{\mu_i} a_{ij} / \sqrt{\mu_j} : i, j \in E)$ becomes Hermitian having the same spectrum as A. As mentioned above, an Hermitizable matrix can be transformed into a real symmetric tridiagonal one with the help of Householder transformation. Besides, a reducible tridiagonal matrix can be decomposed into irreducible sub-matrices. Thus, we consider only irreducible tridiagonal one in this paper. In what follows, we focus on the tridiagonal matrices of the following form on finite space $E = \{0, 1, \cdots, N\} (N < \infty)$:

$$T = \begin{pmatrix} -c_0 & b_0 & & & & \\ a_1 & -c_1 & b_1 & & & \\ & a_2 & -c_2 & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-1} & -c_{N-1} & b_{N-1} \\ & & & & a_N & -c_N \end{pmatrix}. \tag{1}$$

Here we assume that T is Hermitizable([8; §1]):

$$(c_k) \text{ is real}, (a_k) \text{ and } (b_k) \text{ are complex but } a_{k+1} b_k > 0 \ (0 \leqslant k < N). \tag{2}$$

As usual, the 'eigenpair' means the twins consisting of an eigenvalue and its eigenvector. In what follows, we simply write the tridiagonal matrix as

$$T \sim (a_k, -c_k, b_k),$$

since such a matrix is determined by the three sequences

$$\{a_k\}_{k=1}^{N}, \qquad \{-c_k\}_{k=0}^{N}, \qquad \{b_k\}_{k=0}^{N-1}.$$

Now, the powerful algorithm for maximal eigenpair, according to [4-7] and [8; §4], can be stated as follows.

1.1 Main algorithm

Algorithm 1. Suppose that T is an Hermitizable tridiagonal matrix of form (1). Before computing the maximal eigenpair of T, we need prepare three steps as follows.

Step 1. Let

$$m = \sup_{k\in E}(-c_k + |a_k| + |b_k|)^+, \qquad x^+ := \max\{x, 0\},$$

and

$$u_k = a_k b_{k-1}, \qquad k \in E \setminus \{0\}.$$

Set $\widetilde{c}_k = c_k + m\,(k \in E)$ and $\widetilde{b}_0 = \widetilde{c}_0 > 0$. Next, let

$$\widetilde{b}_k = \widetilde{c}_k - \frac{u_k}{\widetilde{b}_{k-1}}, \quad \widetilde{a}_k = \widetilde{c}_k - \widetilde{b}_k, \quad 1 \leqslant k < N,$$

$$\widetilde{a}_N = \frac{u_N}{\widetilde{b}_{N-1}}.$$

More explicitly,

$$\begin{cases} \tilde{b}_0 = \tilde{c}_0, & \\ \tilde{b}_k = \tilde{c}_k - \cfrac{u_k}{\tilde{c}_{k-1} - \cfrac{u_{k-1}}{\tilde{c}_{k-2} - \cfrac{u_{k-2}}{\ddots \cfrac{}{\tilde{c}_2 - \cfrac{u_2}{\tilde{c}_1 - \cfrac{u_1}{\tilde{c}_0}}}}}}, & 1 \leqslant k < N, \\ \tilde{a}_k = \tilde{c}_k - \tilde{b}_k, & 1 \leqslant k < N, \\ \tilde{a}_N = \cfrac{u_N}{\tilde{b}_{N-1}}. & \end{cases}$$

Then the tridiagonal matrix

$$\widetilde{Q} \sim (\widetilde{a}_k, -\widetilde{c}_k, \widetilde{b}_k)$$

possesses the properties: both $(\widetilde{a}_k)$ and $(\widetilde{b}_k)$ are positive, the sum of each row equals zero except the Nth row $(\widetilde{c}_N \geqslant \widetilde{a}_N)$. If $\widetilde{c}_N = \widetilde{a}_N$, then T has the maximal eigenvalue $\lambda_{\max} = m$ with eigenvector $g_{\max} = h$:

$$h_0 = 1, \quad h_k = h_{k-1}\frac{\widetilde{b}_{k-1}}{b_{k-1}} \quad \left[= \prod_{j=0}^{k-1}\frac{\tilde{b}_j}{b_j}\right], \qquad k \in E\setminus\{0\}.$$

Otherwise set $\widetilde{b}_N = \widetilde{c}_N - \widetilde{a}_N$ and go to the next step.

Step 2. Define the symmetric tridiagonal matrix

$$Q^{\mathrm{sym}} \sim (a_k^{\mathrm{sym}}, -c_k^{\mathrm{sym}}, b_k^{\mathrm{sym}})$$

as follows:

$$c_k^{\mathrm{sym}} = \widetilde{c}_k, \qquad k \in E,$$
$$a_k^{\mathrm{sym}} = b_{k-1}^{\mathrm{sym}} = \sqrt{a_k b_{k-1}}, \qquad k \in E \backslash \{0\}.$$

Step 3. Define the upper triangular matrix (M_{kj}) and the vector (Φ_k) as follows:

$$M_{kk} = 1,\ M_{kj} = M_{k,j-1}\frac{\widetilde{a}_j}{\widetilde{b}_{j-1}} \quad \Big[= \frac{\widetilde{a}_{k+1}\cdots\widetilde{a}_j}{\widetilde{b}_k\cdots\widetilde{b}_{j-1}}\Big], \qquad 1 \leqslant k+1 \leqslant j \leqslant N,$$
$$\Phi_k = \frac{1}{\widetilde{b}_k} + \sum_{k+1\leqslant j\leqslant N} \frac{\widetilde{a}_{k+1}\cdots\widetilde{a}_j}{\widetilde{b}_k\cdots\widetilde{b}_j} = \sum_{k\leqslant j\leqslant N} \frac{M_{kj}}{\widetilde{b}_j}, \qquad 0 \leqslant k \leqslant N.$$

With $(\widetilde{a}_k, \widetilde{b}_k)$, Q^{sym}, M and Φ at hand, we can now start our iterations. Note that one may use the parallel computing the next step.

Step 4. For given $v^{(k)}$ $(k \geqslant 0)$, define

$$\zeta_k = \sup_{0\leqslant n\leqslant N} \frac{1}{\sqrt{\widetilde{b}_n}v_n^{(k)} - \sqrt{\widetilde{a}_{n+1}}v_{n+1}^{(k)}} \sum_{j=0}^{n} v_j^{(k)} \sqrt{\frac{M_{jn}}{\widetilde{b}_n}}, \qquad k \geqslant 0, \tag{3}$$

with a convention that $\widetilde{a}_{N+1} = 0$. As in [8; §4], choose

$$w^{(0)} = \sqrt{\Phi}, \quad v^{(0)} = \frac{w^{(0)}}{\sqrt{w^{(0)*}w^{(0)}}}, \qquad z^{(0)} = \frac{1}{\zeta_0},$$

where ζ_0 is defined by (3) with $k = 0$. For each $k \geqslant 1$, solve $w^{(k)}$:

$$\big(-Q^{\mathrm{sym}} - z^{(k-1)}I\big)w^{(k)} = v^{(k-1)}, \tag{4}$$

and define

$$v^{(k)} = \frac{w^{(k)}}{\sqrt{w^{(k)*}w^{(k)}}}, \qquad z^{(k)} = \frac{1}{\zeta_k},$$

where ζ_k is defined again by (3). Then $(v^{(k)}, z^{(k)})$ converges to the maximal eigenpair of Q^{sym}.

Step 5. To go back to the original matrix T, denote its maximal eigenpair by $(\lambda_{\max}(T), g_{\max})$. Then we have

$$\lambda_{\max}(T) = m - \lim_{k\to\infty} z^{(k)}, \qquad g_{\max} = \lim_{k\to\infty} \mathrm{diag}(h^{\mu})v^{(k)},$$

where $\mathrm{diag}(h^{\mu})$ is the diagonal matrix having diagonal elements (h_k^{μ}):

$$h_0^{\mu} = 1,\ h_k^{\mu} = h_{k-1}^{\mu}\frac{\sqrt{u_k}}{b_{k-1}} \quad \Big[= \prod_{j=1}^{k} \frac{\sqrt{u_j}}{b_{j-1}}\Big], \qquad k \in E \setminus \{0\}.$$

Remark 2. Note that the main aim of the algorithm is for the maximal eigenvector, the approximation of the maximal eigenvalue is its by-product. This is different from the algorithms for computing the eigenvalues only. Actually, as can be seen from the examples in the paper, the sequence $\{z^{(k)}\}$ is monotone in k. Refer to the last two paragraphs of §4.3 for more details. Next, we use the following Thomas algorithm to solve equation (4).

Thomas algorithm ([12]). Given a tridiagonal matrix $T \sim (a_k, -c_k, b_k)$, a constant shift z and a vector v, define

$$d_i = \begin{cases} \dfrac{b_0}{z - c_0}, & i = 0, \\ \dfrac{b_i}{z - c_i - a_i d_{i-1}}, & i = 1, 2, \cdots, N-1. \end{cases}$$

Next, define

$$\xi_i = \begin{cases} \dfrac{v_0}{c_0 - z}, & i = 0, \\ \dfrac{v_i + a_i \xi_{i-1}}{c_i - z + a_i d_{i-1}}, & i = 1, 2, \cdots, N. \end{cases}$$

Then, the solution w to the equation

$$(-T - zI)w = v \qquad \text{on} \qquad E$$

is given as follows:

$$\begin{cases} w_N = \xi_N, \\ w_i = \xi_i - d_i w_{i+1}, & i = N-1, N-2, \cdots, 1, 0. \end{cases}$$

Combining the well-known Householder Transformation (see [14], [16]) with Algorithm 1, one can compute the maximal eigenpair of Hermitizable matrices. The next example illustrates the power of this idea.

Example 3. Consider the following matrix:

$$A = \begin{pmatrix} -2 & 2+2i & 1-i & 0 \\ 0.5-0.5i & -3 & 1-0.5i & 3+i \\ 1+i & 4+2i & -4 & 8+2i \\ 0 & 3-i & 2-0.5i & -5 \end{pmatrix}.$$

According to the Improved circle theorem [8; Theorem 5], we obtain the Hermitizing measure $\mu = (1,\ 4,\ 1,\ 4)$ of A: $\mu_k a_{k\ell} = \mu_\ell \bar{a}_{\ell k}$ (refer to [8; Example 7] for more details). Thus A is Hermitizable and then

$$\hat{A} := \operatorname{diag}(\sqrt{\mu}\,) A \operatorname{diag}\left(\frac{1}{\sqrt{\mu}}\right) = \begin{pmatrix} -2 & 1+i & 1-i & 0 \\ 1-i & -3 & 2-i & 3+i \\ 1+i & 2+i & -4 & 4+i \\ 0 & 3-i & 4-i & -5 \end{pmatrix}$$

is an Hermitian matrix. Now, the Householder transformation says that there exist a sequence of extended reflection matrices U_j having the form

$$U_j = I + (\kappa - 1)uu^H,$$

(where κ is a constant with $|\kappa| = 1$ and u is a unit vector) such that for some ℓ,

$$U := \prod_{j=0}^{\ell} U_j$$

is unitary and

$$T := U\hat{A}U^H$$

becomes a real, symmetric tridiagonal matrix:

$$T \approx \begin{pmatrix} -2 & 2 & & \\ 2 & -2.5 & 4.092676 & \\ & 4.092676 & -1.977612 & 2.622282 \\ & & 2.622282 & -7.522388 \end{pmatrix}.$$

Here, we mention that even though the resulting matrix T is real symmetric, the h-transform in Step 1 of Algorithm 1 is still needed to produce the efficient initials. Thus, for the matrix T, we have $m \approx 4.737347$. After Step 1, we get the matrix

$$\widetilde{Q} \approx \begin{pmatrix} -6.737347 & 6.737347 & & \\ 0.5937055 & -7.237347 & 6.643641 & \\ & 2.521208 & -6.714959 & 4.193751 \\ & & 1.639669 & -12.259735 \end{pmatrix}.$$

For the matrix $-\widetilde{Q}$ with the summation of each line being zero except the last line, we have efficient approximation of its minimal eigenpair. Next, to leveling the eigenvector, we turn to compute the eigenvector of Q^{sym} which is obtained by Step 2 of Algorithm 1:

$$Q^{\mathrm{sym}} \approx \begin{pmatrix} -6.737347 & 2 & & \\ 2 & -7.237347 & 4.092676 & \\ & 4.092676 & -6.714959 & 2.622282 \\ & & 2.622282 & -12.259735 \end{pmatrix}.$$

For convenience, we write $A \simeq B$ if A and B are isospectral. Since $\widetilde{Q} \simeq Q^{\mathrm{sym}}$, the efficient initials of $\widetilde{Q}$ can be transformed into the ones of Q^{sym} with the help of M, Φ in Step 3. The efficient initials of Q^{sym} are as follows:

$$z^{(0)} \approx 1.531417, \quad v^{(0)} \approx (0.463553, 0.566223, 0.588314, 0.344088)^*.$$

Now, combining the initials with the iterative equation (4) of Q^{sym}, we obtain an approximation of the minimal eigenvalue of $-Q^{\mathrm{sym}}$. Furthermore, noticing that

$$Q^{\mathrm{sym}} \simeq T - mI \simeq A - mI,$$

we obtain an approximation of the maximal eigenvalue of A (denoted by $\rho(A)$) as follows:

$$z^{(0)}\approx 3.205929,\quad z^{(1)}\approx 2.661892,\quad z^{(2)}\approx 2.628326,\quad z^{(3)}\approx 2.628164\approx\rho(A).$$

From [4-8], one sees a long trip to achieve Algorithm 1. Based on the sharp estimates of maximal eigenvalue given by [2; Theorems 2.4, 3.2], the efficient initials were introduced in [4; §3] for Rayleigh Quotient Iteration. To avoid the dangerous region, the modified algorithm by redefining $z_k=\delta_k^{-1}$ was presented in [6; §A.4]. In the following-up article [7; §2], an explicit representation of the solution to tridiagonal equation was proposed. To balance the sharpness and the complexity, several methods to improve the algorithm were proposed in [8]. We will outline this trip in §4. Besides, based on [4, 6, 7], Tang and Yang [18] simplified the computational complexity and proved that the total cost for computing is $O(N)$.

1.2 Typical example

To illustrate the power of Algorithm 1, we choose the following typical example, taken from [8; §4] and [18; §3.3].

Example 4. Consider the following matrix on E:

$$Q=\begin{pmatrix} -3 & 2 & & & & \\ 1 & -3 & 2 & & & \\ & 1 & -3 & 2 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -3 & 2 \\ & & & 1 & -3 \end{pmatrix}.$$

We are going to use three algorithms to compute the maximal eigenpair of the matrix Q. The first one is Algorithm 1, the others are also recent. One can see the difference of their effectiveness and then understand a part of the development of the algorithm.

(a) For the matrix Q, we have $m=0$. After the preparations of Steps 1–3 in Algorithm 1, we start the iterations at Step 4. The computing result for different N is given in Table 1.

Here, we mention that the numerical experiments in this article are fulfilled on a PC with Intel(R) Core(TM)i5-5200 CPU @2.20 GHz and 4.00 GB RAM using MATLAB (R2014a).

For this example, with the button 'Run and time' in Matlab, we record the running time needed to get the results for larger N and the time 's' denotes seconds. From Table 1, one sees that the six precisely significant digits are achieved with no more than three steps. Actually, when $N\geqslant 4500$, up to six precisely significant digits, the initial $z^{(0)}$ already coincides with $z^{(1)}\approx\lambda_0$ (i.e., the minimal eigenvalue of $-Q$).

Table 1. Outputs for different N by Algorithm 1(Example 4)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.253835	0.33544	0.342107	0.342148
16	0.182046	0.21533	0.219673	0.219732
50	0.171577	0.175993	0.176912	0.176937
100	0.171573	0.172686	0.172934	0.172941
500	0.171573	0.171618	0.171628	
1000	0.171573	0.171584	0.171587	
5000	0.171573	0.171573 (1.597s)		
10000	0.171573	0.171573 (6.578s)		
15000	0.171573	0.171573 (29.160s)		

(b) For comparison, as in [8; Algorithm 27], we take $z^{(k)} = \delta_k^{-1}$ instead of $z^{(k)} = \zeta_k^{-1}$ in Algorithm 1 to compute the same example, where

$$\delta_k = \sup_{0\leqslant n\leqslant N} \frac{1}{v_n^{(k)}} \left[\Phi_n \sum_{0\leqslant i\leqslant n} v_i^{(k)} \sqrt{M_{in}} + \sum_{n+1\leqslant j\leqslant N} \sqrt{M_{nj}} \Phi_j v_j^{(k)} \right], \quad k \geqslant 0, \tag{5}$$

and $\sum_{\emptyset} := 0$. The computing result is given in Table 2.

Table 2. Outputs for different N using δ_k instead of ζ_k(Example 4)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.304256	0.340851	0.342146	0.342148
16	0.195163	0.217878	0.219722	0.219732
50	0.171632	0.17606	0.176916	0.176937
100	0.171573	0.17269	0.172934	0.172941
500	0.171573	0.171618	0.171628	
1000	0.171573	0.171584	0.171587	
5000	0.171573	0.171573 (2.814s)		
10000	0.171573	0.171573 (15.927s)		
15000	0.171573	0.171573 (96.483s)		

Comparing Table 1 with Table 2, we know that both the use of ζ_k and δ_k need no more than three steps to get the expected results. But the use of ζ_k saves much time. The reason is clear: ζ_k uses a single summation and δ_k uses a double summation. Their computational complexity are $O(N)$ and $O(N^2)$, respectively. For further comparison of ζ_k and δ_k, see Example 9 below.

(c) We now compare the use of Q^{sym} in equation (4) of Algorithm 1 with the earlier algorithm presented in [6; §A.4]. The computing result is given in Table 3.

Actually, as the problem mentioned in [8; §4], the curve of the initial vector $w^{(0)}$ goes down rapidly at the beginning smaller intervals, and becomes too small at the end intervals, due to the limitation of computer for calculation accuracy, we can only get the expected result up to 1023 by using the algorithm in [6; §A.4].

Table 3. Outputs using the algorithm in [6; §A.4](Example 4)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.304256	0.340851	0.342146	0.342148
16	0.195163	0.217878	0.219722	0.219732
50	0.171632	0.17606	0.176916	0.176937
100	0.171573	0.17269	0.172934	0.172941
500	0.171573	0.171618	0.171628	
1000	0.171573	0.171584	0.171587	
1023	0.171573	0.171584	0.171586	

(d) It is the position to compare Algorithm 1 with the known ones. As well known, the maximal eigenpair has a very wide application, such as Google's PageRank, the input-output method in economic optimization. Refer to [4-7] for more information. From Wikipedia or textbooks, one may learn that there are mainly two algorithms for computing the maximal eigenpair: Power Iteration and Rayleigh Quotient Iteration. The former one is simpler and has a wide application. But its convergence speed is very slow and hence is less practical. The second one is a cubic algorithm, provided the initials are sharp enough. However, it is a dangerous algorithm, as will be discussed in detail in §4. A related algorithm is the famous QR Algorithm, which is used to compute the eigensystem of a matrix. This algorithm is widely used in practice. The price is the limitation of the scale of the matrix. As an illustration, we now compare Algorithm 1 with the function 'eig' contained in Matlab. Table 4 presents the outputs corresponding to Example 4 for different N using the two methods.

Table 4. Outputs using Algorithm 1 and eig (Example 4)

$N+1$	Algorithm 1	eig
100	0.172941	0.172941
198	0.171925	0.171925
199	0.171922	0.171923
250	0.171794	0.171728
500	0.171628	0.172726

The function 'eig' is mature for computing all the eigenvalues of a matrix. Table 4 shows that when $N \geqslant 199$, the outputs are incorrect. In fact, when $N = 198$, the outputs using 'eig' are already complex (for instance, a complex eigenvalue $5.77283 + 0.0163545i$ appears in the outputs of 'eig'), but this matrix Q only has real eigenvalues because it is symmetrizable. Thus, 'eig' is efficient for medium-size matrices but not for larger matrices.

The typical Example 4 shows that Algorithm 1 is powerful for computing the maximal eigenpair. In the next section, we introduce additional examples to illustrate the power of the algorithm.

2 Examples

In this section, the examples are taken from the papers [2, 4, 6, 8], except Example 6 which is newly added. To compare the results with Algorithm 1, let us look at the following example taken from [6; Appendix A.4] first.

Example 5. [6; Example A3] Let

$$T = \begin{pmatrix} 2.334 & 0.9962 & & & & \\ 0.5142 & 2.6725 & 0.1111 & & & \\ & 0.2115 & 2.263 & 0.1405 & & \\ & & 0.8442 & 2.8457 & 0.7595 & \\ & & & 0.2347 & 2.2257 & 0.0781 \\ & & & & 0.9837 & 2.1582 \end{pmatrix}.$$

Then the eigenvalues of T are

$$3.26753,\ 3.16247,\ 2.40182,\ 2.12632,\ 1.80416,\ 1.73679.$$

For the matrix T, we have $m = 4.4494$. The outputs using Algorithm 1 are given in Table 5.

Table 5. Outputs by Algorithm 1(Example 5)

$m - z^{(0)}$	$m - z^{(1)}$	$m - z^{(2)}$	$m - z^{(3)}$	$m - z^{(4)}$
3.41401	3.28721	3.26957	3.26757	3.26753

Comparing the results here with those in [6; Example A3], we know that Algorithm 1 is as effective as the algorithm presented in [6; §A.4]. Note that this example is somehow dangerous for Rayleigh Quotient Iteration (cf.[4] or the first paragraph of §4.3 below) since the first two eigenvalues of T are very close to each other. Now, the dangerous problem is avoided completely in Algorithm 1.

The next two examples satisfy condition (2) and hence have real spectrum. Then we compute the maximal eigenvalue of the two complex matrices using Algorithm 1.

Example 6. Let

$$T = \begin{pmatrix} -2 & 2+i & & & & \\ 2^2(2-i) & -1 & 3+i & & & \\ & 2^2(3-i) & -3 & 1+2i & & \\ & & 1-2i & -2 & 2+3i & \\ & & & 2-3i & -4 & 2-i \\ & & & & 4+2i & 3 \end{pmatrix}.$$

Then the eigenvalues of T are

$$-10.0244,\ -7.26296,\ -2.65371,\ 0.243075,\ 4.50158,\ 6.19640.$$

For the complex matrix T, we have

$$m = \sqrt{5} + 4\sqrt{10} - 3.$$

The outputs using Algorithm 1 are given in Table 6.

Table 6. Outputs by Algorithm 1(Example 6)

$m - z^{(0)}$	$m - z^{(1)}$	$m - z^{(2)}$	$m - z^{(3)}$	$m - z^{(4)}$
7.31593	6.34282	6.20382	6.19644	6.19640

Example 7. [8; Example 19] Let

$$T = \begin{pmatrix} -1 & 2+i & & & \\ 2^2(2-i) & -1 & 2^4(2+i) & & \\ & 6^2(2-i) & -1 & 3^4(2+i) & \\ & & 12^2(2-i) & -1 & 4^4(2+i) \\ & & & 20^2(2-i) & -1 \end{pmatrix}.$$

Then the eigenvalues of T are

$$-756.391,\ -52.0308,\ -1,\ 50.0308,\ 754.391.$$

For the complex matrix T, we have $m = 400\sqrt{5} - 1$. The outputs using Algorithm 1 are given in Table 7.

Table 7. Outputs by Algorithm 1 (Example 7)

$m - z^{(0)}$	$m - z^{(1)}$	$m - z^{(2)}$
773.836	754.594	754.391

We now compute three examples to explicitly illustrate the power of the difference in using $z^{(k)} = \zeta_k^{-1}$ and $z^{(k)} = \delta_k^{-1}$, respectively, in Algorithm 1. The examples are birth-death matrices on E taken from [2; §3] satisfying that $c_0 = b_0$ and $c_k = a_k + b_k$, $k \in E\backslash\{0\}$. Then, we have $m = 0$.

Example 8. [2; Example 3.5] Let $b_k = 2\,(k \geqslant 0)$, $a_k = k\,(k \geqslant 1)$. Then $\lambda_0 = 1$ for large enough N. For different N, the outputs are given in Tables 8 and 9, using ζ_k and δ_k, respectively.

Table 8. Outputs for different N using ζ_k (Example 8)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$
8	0.836045	0.992135	1.01307	1.01355	
16	0.834044	0.977037	0.999391	1.00011	
32	0.83404	0.976651	0.999245	0.999999	1.0
50	0.83404	0.97665	0.999245	0.999999	1.0
100	0.83404	0.97665	0.999245	0.999999	1.0

Table 9. Outputs for different N using δ_k instead of ζ_k (Example 8)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.943632	1.00843	1.01353	1.01355
16	0.918635	0.99231	1.00004	1.00011
32	0.917966	0.99197	0.999921	1.0
50	0.917965	0.991969	0.999921	1.0
100	0.917965	0.991969	0.999921	1.0

Comparing Table 8 with Table 9, we know that when $N \geqslant 32$, the use of ζ_k needs one more step than that of δ_k to get the expected result. But usually, for larger matrices, the time spent in Table 8 is shorter than that of Table 9. See for instance in the next example.

Example 9. [2; Example 3.7] Let

$$b_k = (k+1)^4, \qquad a_k = k(k-1/2)(k^2+3k+3).$$

Then $\lambda_0 = 1/2$ for large enough N. For different N, the outputs are given in Tables 10 and 11, using ζ_k and δ_k, respectively.

Table 10. Outputs for different N using ζ_k(Example 9)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.416918	0.627353	0.633461	0.633466
50	0.35708	0.541483	0.548162	0.548169
100	0.347301	0.526674	0.533345	0.533353
500	0.335057	0.507877	0.514489	0.514498
1000	0.332284	0.503583	0.510173	0.510182
5000	0.328655	0.497945	0.504504	0.504512 (2.009s)
10000	0.327808	0.496625	0.503175	0.503184 (8.973s)

Table 11. Outputs for different N using δ_k(Example 9)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$
8	0.616418	0.63343	0.633466
50	0.530028	0.548113	0.548169
100	0.515265	0.533295	0.533353
500	0.496589	0.514436	0.514498
1000	0.492332	0.51012	0.510182
5000	0.486749	0.50445	0.504512 (3.383s)
10000	0.485443	0.503121	0.503184 (24.602s)

For this example, we record the time needed to get the result for larger N. From Tables 10 and 11, it follows that the use of ζ_k needs one more step than that of δ_k, but saves much time. The next example exhibits the same phenomenon.

In view of Examples 8 and 9, for matrices whose scale is less than or equal to 1000, one may use δ_k instead of ζ_k. But as we have already seen, the difference between them is slight.

Example 10. [2; Example 3.6], [4; Example 7] Let

$$b_k = (k+1)^2, \qquad a_k = k^2.$$

Then $\lambda_0 = 1/4$ for large enough N. For different N, the outputs are given in Tables 12 and 13, using ζ_k and δ_k, respectively.

Table 12. Outputs for different N using ζ_k(Example 10)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.406762	0.514094	0.525176	0.525268
100	0.304993	0.36995	0.376269	0.376383
500	0.279999	0.333226	0.33823	0.338329
1000	0.273336	0.322412	0.327148	0.32724
5000	0.26322	0.304128	0.308454	0.308529 (2.088s)
7500	0.261484	0.300603	0.304845	0.304918 (4.711s)
10000	0.260397	0.298305	0.302489	0.302561 (8.516s)

Table 13. Outputs for different N using δ_k(Example 10)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.485985	0.52415	0.525267	0.525268
100	0.348549	0.374848	0.376378	0.376383
500	0.310195	0.33686	0.33832	0.338329
1000	0.299089	0.325735	0.327229	0.32724
5000	0.281156	0.306874	0.308514	0.308529 (4.281s)
7500	0.277865	0.303213	0.304903	0.304918 (11.077s)
10000	0.275762	0.300821	0.302545	0.302561 (55.355s)

Examples 8– 10 show the reason for choosing $z^{(k)} = 1/\zeta_k$ from the computational point of view. To conclude this section. we study one more example.

Example 11. [4; Example 22] Let

$$T = \begin{pmatrix} -5 & 5 & & & & \\ 3 & -7 & 4 & & & \\ & 2 & -3 & 1 & & \\ & & 10 & -16 & 6 & \\ & & & 11 & -11-b_4 \end{pmatrix}.$$

Then we have $m = 0$. For different b_4, the outputs using Algorithm 1 are given in Table 14.

Table 14. Outputs for different b_4(Example 11)

b_4	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$
0.01	0.000143394	0.000278683	0.000278686
1	0.0130396	0.0244922	0.0245175
100	0.102368	0.182367	0.182819
10^6	0.109962	0.19468	0.195145

The number of iterations is the same as those given in [4], except when $b_4 = 0.01$, for which the use of Algorithm 1 requires one more step.

From the above examples, it follows that Algorithm 1 is efficient for computing the maximal eigenpair. Furthermore, there is a natural way to study the next to the maximal eigenpair, which is the topic in the next section.

3 Conservative case

In this section, we continue the study on a special case in Algorithm 1. That is $\tilde{c}_N = \tilde{a}_N$. Then the matrix $\widetilde{Q}$ constructed in Algorithm 1 is conservative and so has a trivial eigenvalue 0. The aim of this section is to study its sub-maximal eigenpair. By using a shift mI, if necessary, we deal with the tridiagonal Q-matrix of the following form:

$$Q=\begin{pmatrix} -c_0 & b_0 & & & & \\ a_1 & -c_1 & b_1 & & & \\ & a_2 & -c_2 & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-1} & -c_{N-1} & b_{N-1} \\ & & & & a_N & -c_N \end{pmatrix}, \tag{6}$$

where $a_k > 0$ $(1 \leqslant k \leqslant N)$, $b_k > 0 (0 \leqslant k \leqslant N-1)$ and $c_k = a_k + b_k$ (with $a_0 := 0$ and $b_N := 0$). Clearly, the maximal eigenvalue for this matrix is $\lambda_0 = 0$ with constant eigenvector $\mathbb{1}$. Now, we compute the sub-maximal eigenvalue $\lambda_1(Q)$ by the following algorithm, which is essentially harder but formally the same as Algorithm 1, except the use of a new auxiliary tridiagonal matrix $\widetilde{Q}$ on $E_1 := \{k \in \mathbb{Z} : \ 1 \leqslant k \leqslant N\}$. Recall that $E = E_1 \bigcup \{0\}$. The next result is due to [8; §4].

Algorithm 12. Suppose that Q is a tridiagonal matrix of form (6). Define a measure μ corresponding to Q as follows:

$$\mu_0 = 1, \quad \mu_n = \mu_{n-1}\frac{b_{n-1}}{a_n}, \qquad 1 \leqslant n \leqslant N. \tag{7}$$

Before computing the sub-maximal eigenpair of Q, we need prepare three steps 1–3 as follows.

Step 1. Define $\widetilde{Q}\sim(\widetilde{a}_k,-\widetilde{c}_k,\widetilde{b}_k)$ on E_1 as follows:

$$\begin{cases}\widetilde{b}_1=a_1+b_0, \qquad \widetilde{b}_k=a_k+b_{k-1}-\dfrac{a_{k-1}b_{k-1}}{\widetilde{b}_{k-1}}, \qquad 2\leqslant k<N,\\ \widetilde{a}_k=a_k+b_{k-1}-\widetilde{b}_k, \qquad 2\leqslant k<N, \quad \widetilde{a}_N=\dfrac{a_{N-1}b_{N-1}}{\widetilde{b}_{N-1}},\\ \widetilde{c}_k=a_k+b_{k-1}, \qquad 1\leqslant k\leqslant N.\end{cases}$$

This is essentially different from Step 1 in Algorithm 1, in fact it uses a dual technique. Other steps are similar to those in Algorithm 1, replacing E by E_1 and recall that in Algrithm 1, we made a convention that $\widetilde{b}_N=\widetilde{c}_N-\widetilde{a}_N$.

Step 2. Define the symmetric tridiagonal matrix $Q^{\rm sym}\sim(a_k^{\rm sym},-c_k^{\rm sym},b_k^{\rm sym})$ on E_1 as follows:

$$c_k^{\rm sym}=\widetilde{c}_k, \qquad k\in E_1,$$
$$a_k^{\rm sym}=b_{k-1}^{\rm sym}=\sqrt{a_{k-1}b_{k-1}}, \qquad k\in E_1\backslash\{1\}.$$

Step 3. Define the upper triangular matrix (M_{kj}) and the vector (Φ_k) on E_1:

$$M_{kk}=1,\ M_{kj}=M_{k,j-1}\frac{\widetilde{a}_j}{\widetilde{b}_{j-1}} \quad \left[=\frac{\widetilde{a}_{k+1}\cdots\widetilde{a}_j}{\widetilde{b}_k\cdots\widetilde{b}_{j-1}}\right], \qquad 2\leqslant k+1\leqslant j\leqslant N,$$
$$\Phi_k=\frac{1}{\widetilde{b}_k}+\sum_{k+1\leqslant j\leqslant N}\frac{\widetilde{a}_{k+1}\cdots\widetilde{a}_j}{\widetilde{b}_k\cdots\widetilde{b}_j}=\sum_{k\leqslant j\leqslant N}\frac{M_{kj}}{\widetilde{b}_j}, \qquad 1\leqslant k\leqslant N.$$

With $(\widetilde{a}_k,\widetilde{b}_k)$, $Q^{\rm sym}$, M and Φ at hand, we can now start our iterations.

Step 4. For given $v^{(k)}$ $(k\geqslant 0)$, define

$$\zeta_k=\sup_{1\leqslant n\leqslant N}\frac{1}{\sqrt{\widetilde{b}_n}v_n^{(k)}-\sqrt{\widetilde{a}_{n+1}}v_{n+1}^{(k)}}\sum_{j=1}^{n}v_j^{(k)}\sqrt{\frac{M_{jn}}{\widetilde{b}_n}}, \qquad k\geqslant 0, \tag{8}$$

with a convention that $\widetilde{a}_{N+1}=0$. As in [8; Algorithm 29], choose $w_i^{(0)}=\sqrt{\Phi_i}$, $i\in E_1$. Define

$$v^{(0)}=\frac{w^{(0)}}{\sqrt{w^{(0)*}w^{(0)}}}, \qquad z^{(0)}=\frac{1}{\zeta_0},$$

where ζ_0 is defined by (8) with $k=0$. For each $k\geqslant 1$, solve $w^{(k)}$:

$$\big(-Q^{\rm sym}-z^{(k-1)}I\big)w^{(k)}=v^{(k-1)} \qquad \text{on } E_1, \tag{9}$$

and define

$$v^{(k)}=\frac{w^{(k)}}{\sqrt{w^{(k)*}w^{(k)}}}, \qquad z^{(k)}=\frac{1}{\zeta_k},$$

where ζ_k is again defined by (8). Let $(\lambda_1^{\rm sym}, g^{\rm sym})$ be the minimal eigenpair of $-Q^{\rm sym}$. Then

$$\lambda_1^{\rm sym} = \lim_{k\to\infty} z^{(k)}, \qquad g^{\rm sym} = \lim_{k\to\infty} v^{(k)}.$$

Step 5. Furthermore, denote the sub-maximal eigenpair of Q by $(-\lambda_1, f)$. Then

$$\lambda_1 = \lambda_1^{\rm sym} = \lim_{k\to\infty} z^{(k)}, \qquad f = \mathscr{M}^{-1} g = \lim_{k\to\infty} \big(\mathscr{M}^{-1} g^{(k)}\big),$$

where $\mathscr{M}$ is a matrix on $E \times E$ (recall that $E = E_1 \bigcup \{0\}$) of the following form:

$$\mathscr{M} = \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_N \\ & \mu_1 & \mu_2 & \cdots & \mu_N \\ & & \mu_2 & \cdots & \mu_N \\ & & & \ddots & \vdots \\ 0 & & & & \mu_N \end{pmatrix},$$

μ is defined by (7), g and $\big(g^{(k)}\big)$ are vectors on E satisfying

$$g_0 = 0,\ g|_{E_1} = \mathrm{diag}\big(h^\mu\big) g^{\rm sym},$$

and

$$g_0^{(k)} = 0,\ g^{(k)}|_{E_1} = \mathrm{diag}\big(h^\mu\big) v^{(k)}.$$

Here, $\mathrm{diag}\big(h^\mu\big)$ is a diagonal matrix on E_1 having diagonal elements $\big(h_k^\mu\big)$:

$$h_1^\mu = 1, \quad h_k^\mu = h_{k-1}^\mu \sqrt{\frac{b_{k-1}}{a_{k-1}}} \quad \left[= \prod_{j=2}^{k} \sqrt{\frac{b_{j-1}}{a_{j-1}}} \right], \qquad k \in E_1 \setminus \{1\}.$$

Remark 13. We also use Thomas algorithm to solve equation (9).

For the remainder of this section, we introduce three examples to illustrate the power of Algorithm 12. The next two examples are closely related to Examples 4 and 10, respectively.

Example 14. Consider the matrix on E :

$$Q = \begin{pmatrix} -2 & 2 & & & & \\ 1 & -3 & 2 & & & \\ & 1 & -3 & 2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -3 & 2 \\ & & & & 1 & -1 \end{pmatrix}.$$

The computing result for different N using Algorithm 12 is given in Table 15.

Table 15. Outputs for different N using Algorithm 12(Example 14)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.284681	0.37967	0.386839	0.386874
16	0.184834	0.221395	0.225864	0.225920
50	0.171578	0.176176	0.177128	0.177154
100	0.171573	0.172709	0.172961	0.172969
500	0.171573	0.171618	0.171628	0.171629
1000	0.171573	0.171584	0.171587	
5000	0.171573	0.171573		
10000	0.171573	0.171573		
15000	0.171573	0.171573		

Comparing Table 15 with Table 1, we know that when $N \geqslant 1000$, the outputs are the same, which reflect some probabilistic idea.

Example 15. [4; Example 25] Consider the matrix

$$Q = \begin{pmatrix} -1 & 1 & & & & & & \\ 1 & -5 & 2^2 & & & & & \\ & 2^2 & -13 & 3^2 & & & & \\ & \ddots & \ddots & \ddots & \ddots & & & \\ & & (N-1)^2 & -N^2-(N-1)^2 & N^2 & & & \\ & & & N^2 & -N^2 & & & \end{pmatrix}.$$

For different N, the outputs using Algorithm 12 are given in Table 16.

Table 16. Outputs for different N using Algorithm 12(Example 15)

$N+1$	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.60449	0.80318	0.820402	0.820539
16	0.481611	0.638168	0.650021	0.650141
50	0.379585	0.494848	0.5035	0.503596
100	0.344814	0.444154	0.451895	0.451977
500	0.299131	0.373782	0.380427	0.380497
1000	0.287724	0.354814	0.36116	0.361228
5000	0.270961	0.324518	0.330299	0.330367
7500	0.268149	0.318939	0.32459	0.324659
10000	0.266398	0.315346	0.320907	0.320976

Example 16. [4; Example 26] Let

$$Q = \begin{pmatrix} -5 & 5 & & & \\ 3 & -7 & 4 & & \\ & 2 & -3 & 1 & \\ & & 10 & -16 & 6 \\ & & & 11 & -11 \end{pmatrix}.$$

The eigenvalues of $-Q$ are

$$22.348, \quad 10.6857, \quad 5.92951, \quad 3.03673, \quad 0.$$

The outputs using Algorithm 12 are given in Table 17.

Table 17. Outputs using Algorithm 12(Example 16)

$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
2.4898	2.97585	3.03569	3.03673

Comparing the result here with the outputs given in [4], we need one more iteration using Algorithm 12. This often happens for smaller N. From Examples 15 and 14, it follows that Algorithm 12 is as powerful as Algorithm 1, since they have the same computational complexity $O(N)$.

The story of the development as well as intrinsic points are presented in the next section.

4 Development and proofs

This section sketches some key points of the development of Algorithm 1. A complete exploring would take a hundred of pages and hence is out of the scope of such a survey article.

Let us begin this section with the well-known Rayleigh Quotient Iteration.

Rayleigh Quotient Iteration (RQI) For a given real matrix A defined on $E \times E$, with nonnegative off-diagonal elements, let $(\lambda_{\max}, g_{\max}(A))$ be the maximal eigenpair of A and $(z^{(0)}, v^{(0)})$ be an approximation of $(\lambda_{\max}, g_{\max}(A))$. At the kth step ($k \geqslant 1$), solve the linear equation in $w^{(k)}$:

$$\big(z^{(k-1)}I - A\big)w^{(k)} = v^{(k-1)},$$

where I is the identity matrix and define

$$v^{(k)} = \frac{w^{(k)}}{\sqrt{w^{(k)*}w^{(k)}}}, \qquad z^{(k)} = v^{(k)*}Av^{(k)}.$$

Then $v^{(k)}$ converges to $g_{\max}$ and $z^{(k)}$ converges to $\lambda_{\max}(A)$ as $k \to \infty$, provided $(z^{(0)}, v^{(0)})$ is close enough to $(\lambda_{\max}(A), g_{\max})$.

In what follows, we will often use some probabilistic ideas. Our first object is the Q-matrix:

$$Q = (q_{ij} : i, j \in E),$$

where $q_{ij} \geqslant 0$ for every pair $i \neq j$ and $\sum\limits_{j \in E} q_{ij} \leqslant 0$ for every $i \in E$. For simplicity, we assume at the moment that

$$0 < \lambda_0 < |\lambda_1| \leqslant |\lambda_2| \leqslant \cdots,$$

where $\{\lambda_j\}$ is the sequence of the eigenvalues of $-Q$. Suppose that the orthogonal eigenvectors corresponding to $\{\lambda_j\}$ are $\{g_j\}$. We now introduce RQI for Q-matrix.

RQI for Q-matrix The algorithm is almost the same as the original one except A is replaced by $-Q$, at the same time, the original iteration equation is replaced by

$$(-Q - z^{(k-1)}I)w^{(k)} = v^{(k-1)}.$$

We know that RQI is fast for computing the maximal eigenpair but can be dangerous because there are many pitfalls. To get the maximal eigenpair, there is a strict restriction for the initial $(z^{(0)}, v^{(0)})$. Fortunately, some basic analytic estimates of the maximal eigenvalue and some mimics of its eigenvector were obtained for many years of accumulation in the study of stochastic stability speed. See [1, 2] for instance. Based on this, Algorithm 1 has been developed in [4-8], and then Algorithm 12 is presented in [8]. Fig.1 exhibits the diagram of the development from RQI to Algorithm 1.

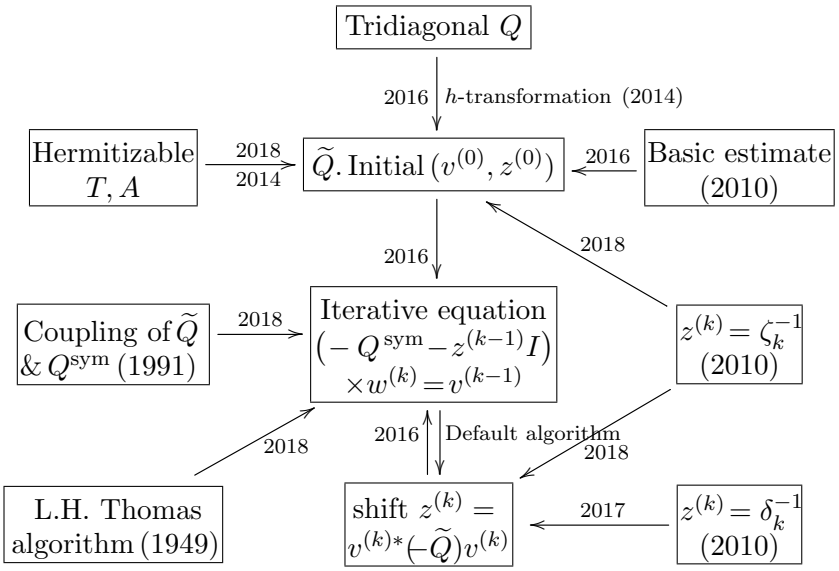


Fig.1 Diagram of development of Algorithm 1.

For the remainder of this section, we are going to explain this diagram by six steps:

- isospectral operators. $Q \overset{2014}{\longrightarrow} \widetilde{Q}, T \overset{2018}{\longrightarrow} \widetilde{Q}$;
- initial $(v^{(0)}, z^{(0)})$;
- Rayleigh quotient $\to \delta_k^{-1}$;
- coupling of non-symmetric matrix Q and symmetrized $Q^{\rm sym}$;
- computational complexity;
- Thomas algorithm.

At the end of this section, we introduce the proof of Algorithm 12.

4.1 Isospectral operators. $Q \xrightarrow{2014} \widetilde{Q}, T \xrightarrow{2018} \widetilde{Q}$.

Before moving further, let us recall the so-called birth-death Q-matrix.

Definition 17. The matrix $Q \sim (a_k, -c_k, b_k)$ is called birth-death Q-matrix defined on $E \times E$ if

$$a_k > 0, \quad b_k > 0, \quad c_k \geqslant a_k + b_k.$$

Let T be an Hermitizable tridiagonal matrix of form (1) and $\widetilde{Q}$ be the one defined by Algorithm 1. Recall that an isospectral transformation (h-transform) was introduced in [9] which allows us to study the Q-matrix having arbitrary diagonals. Then, an explicit construction of isospectral transformation in tridiagonal case (3 steps) was presented in [3], which is the h-transform used in [4]. Nevertheless, there is quite a distance to arrive at the explicit formula of h presented in Algorithm 1. For which, we need to say a little more. In the study on the sub-maximal real part of eigenvalues, a question appears: when a complex matrix has real spectrum? This triggered the study on Hermitizable complex matrix [8; §2]. In which the direct explicit representation of isospectral transformation in tridiagonal case is deduced [8; §3]. This completes the proof of $T \to \widetilde{Q}$, that is quite hard. Furthermore, in terms of a modified Householder transformation, it can be proved that a general Hermitizable complex matrix is also isospectral to a birth-death matrix $\widetilde{Q}$ [8; Theorem 24]. For tridiagonal matrix, the transformation can be explained roughly by Theorem 18.

Theorem 18. Assume T is a complex matrix of form (1). Set $A = T - mI$, where

$$m = \sup_{k \in E}(-c_k + |a_k| + |b_k|)^+.$$

Let h be A-harmonic on $E \setminus \{N\}$ with $h_0 = 1$, i.e.,

$$(Ah)(k) = 0, \qquad k \in \{0, 1, \ldots, N-1\}.$$

Define $\widetilde{Q} = \operatorname{diag}(h)^{-1} A \operatorname{diag}(h)$, where $\operatorname{diag}(h)$ is the diagonal matrix having diagonal elements (h_k). Then $\widetilde{Q}$ is a birth-death matrix of the form $\widetilde{Q} \sim (\widetilde{a}_k, -\widetilde{c}_k, \widetilde{b}_k)$, as presented in Step 1 of Algorithm 1.

Proof. By the explicit representation of the h-transform in [3; §2], we know that the A-harmonic function h is positive. Since $\widetilde{Q} = \operatorname{diag}(h)^{-1} A \operatorname{diag}(h)$, we have

$$\widetilde{c}_k = c_k + m, \qquad \widetilde{a}_k = \frac{h_{k-1}}{h_k} a_k, \qquad \widetilde{b}_k = \frac{h_{k+1}}{h_k} b_k. \tag{10}$$

According to $(Ah)(k) = 0, k \in \{0, 1, \ldots, N-1\}$, we get

$$a_k \frac{h_{k-1}}{h_k} = \widetilde{c}_k - b_k \frac{h_{k+1}}{h_k}.$$

Combining this with (10), we have conservativity:

$$\widetilde{a}_k = \widetilde{c}_k - \widetilde{b}_k, \quad 1 \leqslant k < N,$$

and invariance:

$$\widetilde{a}_k \widetilde{b}_{k-1} = a_k b_{k-1} = |a_k b_{k-1}| = u_k, \quad 1 \leqslant k \leqslant N,$$

where u is the one given in Algorithm 1. Thus,

$$\widetilde{b}_0 = \widetilde{c}_0, \quad \widetilde{b}_k = \widetilde{c}_k - a_k \frac{b_{k-1}}{\widetilde{b}_{k-1}} = \widetilde{c}_k - \frac{u_k}{\widetilde{b}_{k-1}}, \quad 1 \leqslant k \leqslant N,$$

and

$$\widetilde{a}_N = \frac{u_N}{\widetilde{b}_{N-1}}.$$

Furthermore, according to [8; Theorem 15], the sequences $\{\widetilde{a}_k\}$ and $\{\widetilde{b}_k\}$ are positive, provided $m < \infty$. □

Remark 19. For the matrix $\widetilde{Q}$ in Theorem 18, the sum of each row equals zero, except the Nth row which is not positive.

Now, for a given Hermitizable tridiagonal complex matrix T, in terms of Step 1, we can transform its spectrum to the one of a birth-death Q-matrix $\widetilde{Q}$. After completing this procedure, we need only to deal with the maximal eigenpair computation for the birth-death Q-matrix $\widetilde{Q}$. In what follows, unless otherwise stated, we omit '$\sim$' in the notation of $\widetilde{Q}$, $\widetilde{a}_k$, $\widetilde{c}_k$, and $\widetilde{b}_k$.

4.2 Initial $(\boldsymbol{v}^{(0)}, \boldsymbol{z}^{(0)})$.

About the study on maximal eigenpair of Q-matrix, there is a long accumulation on the estimation of stability speed. To illustrate this fact, here, we introduce a couple of results which play a crucial role in the current algorithms. For this, with (μ_k) defined by (7), let

$$\varphi_i = \sum_{j=i}^{N} \frac{1}{\mu_j b_j}.$$

Next, define a sequence of test functions $\{f^{(n)}\}_{n=0}^{\infty}$ as follows.

$$\begin{aligned} f_i^{(0)} &= \sqrt{\varphi}_i, \\ f_i^{(n+1)} &= \sum_{j=i}^{N} \frac{1}{\mu_j b_j} \sum_{k=0}^{j} \mu_k f_k^{(n)}, \qquad n \geqslant 0, \qquad i \in E. \end{aligned} \tag{11}$$

To distinguish δ_n used in this paper, set

$$^o\delta_n = \sup_{i\in E} \frac{f_i^{(n+1)}}{f_i^{(n)}}, \qquad n \geqslant 0,$$

where the superscript 'o' means 'original'. Dually, we also have a sequence $\{^o\delta'_n\}$, but its expression is omitted here (cf. [2; §3]). Now, our basic estimates given in [2; Theorem 3.2] say that

$$^o\delta_k^{-1}\uparrow \leqslant \lambda_0 \leqslant \downarrow {}^o\delta_k'^{-1}.$$

Furthermore, by [2; Corollary 3.3], we have

$$1 \leqslant \frac{^o\delta_0}{^o\delta'_0} \leqslant 4.$$

Moreover, the upper bound is actually no more than 2 in practice. At the beginning of paper [4], we adopted the initial $(w^{(0)}, z^{(0)}) := \big(f^{(0)}, {}^o\delta_0^{-1}\big)$. Based on the basic estimates just mentioned, our initial is good enough in most situations, hence in [4], we simply set $z^{(k)}$ for $k \geqslant 1$ to be the Rayleigh quotient:

$$z^{(k)} = v^{(k)*}(-Q)\, v^{(k)}, \qquad k \geqslant 1.$$

4.3 Rayleigh quotient $\to \delta_k^{-1}$.

RQI can be used in many cases due to its simplicity, the danger going into pitfalls is its limitation. Given the symmetrizable matrix Q for example, we usually have

$$v^{(k)*}(-Q)\, v^{(k)} \geqslant \lambda_0.$$

There may be another eigenvalue $\lambda'(> \lambda_0)$ nearer $v^{(k)*}(-Q)\, v^{(k)}$ than λ_0, then the algorithm will converge to $\lambda' \neq \lambda_0$, i.e. falling into the pitfall λ'. Because of this fact, in [6, 7], when $k \geqslant 1$, the Rayleigh quotient $v^{(k)*}(-Q)\, v^{(k)}$ is replaced by $z^{(k)} = \delta_k^{-1}$:

$$\delta_k = \sup_{0\leqslant n\leqslant N} \frac{1}{v_n^{(k)}} \left[\varphi_n \sum_{0\leqslant i\leqslant n} \mu_i v_i^{(k)} + \sum_{n+1\leqslant j\leqslant N} \mu_j \varphi_j v_j^{(k)}\right], \qquad k \geqslant 0, \tag{12}$$

By [2; Theorem 2.4 (3)], δ_k^{-1} is a lower bound of λ_0. That is, $\delta_k^{-1} \in (0, \lambda_0]$. Now, since there are no other eigenvalues on $(0, \lambda_0)$, our modified algorithm becomes safe, never falling into pitfalls. It is quite remarkable that if we set $f^{(k)} = v^{(k)}$ in (11), then the resulting $^o\delta_k$ coincides with δ_k in (12) (cf. [2; (3.6)]).

The difference of the sequences $\{\delta_k\}$ and $\{^o\delta_k\}$ should be clear now. For $\{\delta_k\}$, the sequence $\{v^{(k)}\}$ comes from the shifted inverse iteration:

$$(-Q - z^{(k-1)}I)w^{(k)} = v^{(k-1)}, \qquad v^{(k)} := \frac{w^{(k)}}{\sqrt{w^{(k)*}w^{(k)}}}.$$

For $\{{}^o\delta_k\}$, the involved sequence $\{f^{(n)}\}$ defined in (11) comes from the ordinary inverse iteration (without shift):

$$(-Q)f^{(k)} = f^{(k-1)}. \tag{13}$$

Actually, for given $f^{(k-1)}$, the solution $f^{(k)}$ to the Poisson equation (13) has an explicit representation as written by (11). Therefore, we can also write

$$f^{(k)} = (-Q)^{-1} f^{(k-1)}.$$

Similarly, for $w^{(k)}$, we have the representation

$$w^{(k)} = (-Q - z^{(k-1)}I)^{-1} v^{(k-1)}.$$

Thus, for given v, these two iterations produce two outputs

$$f = (-Q)^{-1}v, \qquad w = (-Q - zI)^{-1}v,$$

for some $z \in (0, \lambda_0)$. Corresponding to these two functions f and w, we have the following estimates of λ_0, ${}^o\delta$ and δ, respectively:

$$\begin{aligned} {}^o\delta &= \sup_{i\in E} \frac{1}{((-Q)^{-1}v)_i} \sum_{j=i}^{N} \frac{1}{\mu_j b_j} \sum_{k=0}^{j} \mu_k \big((-Q)^{-1}v\big)_k, \\ \delta &= \sup_{i\in E} \frac{1}{((-Q-zI)^{-1}v)_i} \sum_{j=i}^{N} \frac{1}{\mu_j b_j} \sum_{k=0}^{j} \mu_k \big((-Q-zI)^{-1}v\big)_k. \end{aligned}$$

Since the convergence speed of the shifted inverse iteration is much faster than the one of ordinary inverse iteration, we believe that

$$\delta \leqslant {}^o\delta. \tag{14}$$

Perhaps, $\delta = \delta(z)$ is decreasing in z on $(0, \lambda_0)$. But there is still no analytic proof for (14).

Nevertheless, we can prove that the sequence $\{\delta_k\}$ defined by (12) is decreasing in k. The proof will be published elsewhere.

4.4 Coupling of non-symmetric matrix Q and symmetrized $Q^{\rm sym}$.

In [18] and [8; §3], the computational trouble caused by non-symmetric matrix was illustrated in detail. On the one hand, in non-symmetric case, the eigenvector we concerned increases (or decreases) very fast, which is the problem that the computer cannot deal with. On the other hand, if we replace Q by $Q^{\rm sym}$, the spectrum remains, but it often happens that the sum of some of the first N rows is not zero, which makes the use of $\{\delta_k\}$ and $w^{(0)}$ meaningless. If we now use the isospectral transformation in §4.1 on $Q^{\rm sym}$, then we return to the non-symmetric Q. This leads to an unsolvable circulation. To be brief,

the advantage of Q is that for which we have good estimates δ_k and $w^{(0)}$, and the advantage of Q^{sym} is that for which there are many mature algorithms. However, these two are not compatible. This hard problem troubled us for a couple of years. The method we used here is to "let the two get married", which means to use both of their advantages simultaneously. For mathematical details, see [8; §4]. Clearly, this coupling technique should be meaningful for other algorithms of non-symmetric and symmetric ones.

In order to go back to the original matrix T, let $(\lambda_{\max}(T), g_{\max})$ denote its maximal eigenpair. From Steps 1–3 of Algorithm 1, we have taken the following two isospectral transforms:

$$A = T - mI \xrightarrow{h} \widetilde{Q} = \mathrm{diag}(h)^{-1} A\, \mathrm{diag}(h) \xrightarrow{\widetilde{\mu}} Q^{\mathrm{sym}} = \mathrm{diag}(\widetilde{\mu})^{1/2} \widetilde{Q}\, \mathrm{diag}(\widetilde{\mu})^{-1/2}.$$

Noticing that the output (z, v) from the last iteration in Step 4 is an approximation of the minimal eigenpair of $-Q^{\mathrm{sym}}$, we have

$$\lambda_{\max}(T) \approx m - z, \qquad g_{\max} \approx \mathrm{diag}\big(h\widetilde{\mu}^{-1/2}\big)v.$$

By (10), we have

$$h_0 = 1, \;\; h_k = h_{k-1}\frac{\widetilde{b}_{k-1}}{b_{k-1}}.$$

By the definition of $\widetilde{\mu}$, we have

$$\widetilde{\mu}_0 = 1, \;\; \widetilde{\mu}_k = \widetilde{\mu}_{k-1}\frac{\widetilde{b}_{k-1}}{\widetilde{a}_k}.$$

Since

$$\frac{\widetilde{b}_{k-1}}{b_{k-1}}\left(\frac{\widetilde{b}_{k-1}}{\widetilde{a}_k}\right)^{-1/2} = \frac{1}{b_{k-1}}\sqrt{\widetilde{a}_k\widetilde{b}_{k-1}} = \frac{\sqrt{u_k}}{b_{k-1}},$$

we have

$$h^{\mu}_k := \frac{h_k}{\sqrt{\widetilde{\mu}_k}} = h^{\mu}_{k-1}\frac{\sqrt{u_k}}{b_{k-1}}, \qquad h^{\mu}_0 = 1.$$

Then, we get the last assertion of Algorithm 1.

4.5 Computational Complexity

Based on the computational complexity, two improvements are made to arrive at Algorithm 1. The first one is the use of the matrix M and the vector Φ stated in Step 3 of Algorithm 1, which are used to leveling the quantities we required and then fasten the computations.

The second improvement is replacing δ_k in §4.2 and §4.3 by ζ_k appeared in Step 4 of the algorithm. This is somehow strange since in general, the estimates using ζ_k is poorer than the use of δ_k (refer to [2; (2.22)]):

$$\zeta_k^{-1} \leqslant \delta_k^{-1} \leqslant \lambda_0.$$

Hence, we usually do not use ζ_k to estimate λ_0 (refer to [2; Theorem 3.2] for instance). However, such a replacement is reasonable since in the expression of δ_k, one uses double summation, it has complexity $O(N^2)$, but ζ_k uses only a single summation, having complexity $O(N)$. Therefore, the use of ζ_k saves much computation time especially for larger N.

4.6 Thomas algorithm

The algorithm goes back to L.H. Thomas (1949), referring to [15] for a short history. The algorithm is designed especially for tridiagonal systems. It is famous since it is an unusual $O(N)$-algorithm (see [11]). Nevertheless, it is not surprising that the algorithm may be unstable, as pointed in [12]. Very fortunately, our algorithm costs no trouble since at each step of the iterations, it keeps us in the safe region. See [8] for more details.

4.7 Proof of Algorithm 12

From [8; Proof of Algorithm 29], we have $\widehat{Q} = \mathscr{M} Q \mathscr{M}^{-1}$, where

$$\widehat{Q} = \begin{pmatrix} 0 & 0 & 0 & & & & \\ b_0 & -(a_1+b_0) & a_1 & & & & \\ & b_1 & -(a_2+b_1) & a_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b_{N-2} & -(a_{N-1}+b_{N-2}) & a_{N-1} & \\ & & & & b_{N-1} & -(a_N+b_{N-1}) \end{pmatrix}.$$

Then $\widehat{Q}$ and Q have the same spectrum. By removing the first row and the first column of $\widehat{Q}$, we obtain a matrix on $E_1 \times E_1$:

$$\widehat{Q}_1 = \begin{pmatrix} -(a_1+b_0) & a_1 & & & & \\ b_1 & -(a_2+b_1) & a_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & b_{N-2} & -(a_{N-1}+b_{N-2}) & a_{N-1} & \\ & & & b_{N-1} & -(a_N+b_{N-1}) \end{pmatrix}.$$

Now, the spectrum of Q ignoring the trivial eigenvalue 0 coincides with the spectrum of $\widehat{Q}_1$. The sub-minimal eigenvalue of $-Q$ coincides with the minimal eigenvalue of $-\widehat{Q}_1$. Applying Algorithm 1 to the matrix $\widehat{Q}_1$, we get Algorithm 12. Here, we prove the assertion about the sub-maximal eigenvector in Algorithm 12. Let (z, v^{sym}) be the output from the last iteration in Step 4 of Algorithm 12. Note that the development of Algorithm 12 goes as follows:

$$\widehat{Q}_1 \sim (\hat{a}_k, -\hat{c}_k, \hat{b}_k) \longrightarrow \widetilde{Q} \sim (\widetilde{a}_k, -\widetilde{c}_k, \widetilde{b}_k) \longrightarrow Q^{\mathrm{sym}} \sim (a_k^{\mathrm{sym}}, -c_k^{\mathrm{sym}}, b_k^{\mathrm{sym}}).$$

Following the proof given in the last part of §4.4, replacing E by E_1, we have

$$h_1 = 1,\ h_k = h_{k-1}\frac{\tilde{b}_{k-1}}{\hat{b}_{k-1}} = h_{k-1}\frac{\tilde{b}_{k-1}}{a_{k-1}},$$
$$\tilde{\mu}_1 = 1,\ \tilde{\mu}_k = \tilde{\mu}_{k-1}\frac{\tilde{b}_{k-1}}{\tilde{a}_k}.$$

Noting that

$$\frac{\tilde{b}_{k-1}}{a_{k-1}}\left(\frac{\tilde{b}_{k-1}}{\tilde{a}_k}\right)^{-1/2} = \frac{1}{a_{k-1}}\sqrt{\tilde{a}_k\tilde{b}_{k-1}} = \frac{1}{a_{k-1}}\sqrt{\hat{a}_k\hat{b}_{k-1}} = \sqrt{\frac{b_{k-1}}{a_{k-1}}},$$

we obtain

$$h_k^\mu := \frac{h_k}{\sqrt{\tilde{\mu}_k}} = h_{k-1}^\mu\sqrt{\frac{b_{k-1}}{a_{k-1}}},\quad k \geqslant 2;\qquad h_1^\mu = 1.$$

Then, it follows that the eigenvector of $\widehat{Q}_1$ corresponding to λ_1 is

$$\hat{g} \approx \operatorname{diag}\big(h^\mu\big)v^{\,\mathrm{sym}}.$$

The eigenvector of $\widehat{Q}$ corresponding to λ_1 is g:

$$g_0 = 0,\quad g|_{E_1} = \hat{g} \approx \operatorname{diag}\big(h^\mu\big)v^{\,\mathrm{sym}}.$$

Combining this with $\widehat{Q} = \mathscr{M}Q\mathscr{M}^{-1}$, we get the eigenvector of Q corresponding to λ_1 is $\mathscr{M}^{-1}g$.

To conclude this paper, we make two remarks. First, even though we concentrate on tridiagonal matrix here, there are some ways to extend it to the general one, as did in [4; Appendix of §3 or §4], except the general Hermitizable matrices mentioned above. Next, we also have global algorithms presented in [6] to deal with general matrices, especially for those having nonnegative off-diagonal elements. Here is a very simple example to show the importance of the last algorithm.

Example 20. Study the maximal eigenpair of the matrix

$$A = (a_{kj}),\quad a_{kj} = \big(2^{k\wedge j} - 1\big)\,2^{-j},\qquad 1 \leqslant k, j \leqslant N.$$

Here we adopt three different approaches.

(a) By using the package 'Eigensystem' in Mathematica (version 10.3 or 11.3) on PC, it works well up to $N = 11$. Starting from $N = 12$, the output of the components of the vector have different signs. Since A is positive, this is clearly impossible by the well-known Perron-Frobenius theorem.

(b) Next, we use the package 'eig' in MatLab, similar result appears. Starting from $N = 173$, the output of the components of the vector have different signs.

(c) However, when we use [6; Algorithm 5], it is well done up to $N = 2104$ by saving 10^{-6} precision $\big(\max(Aw)_j/w_j - \min(Aw)_j/w_j < 10^{-6}$, where w is the output of the corresponding eigenvector$\big)$. It takes 235 iterations to get the stable approximation of the eigenvalue: 5.82832. The larger number of iterations is due to the fast decay of the approximating eigenvector, up to 10^{-310}. For more larger N, the outputs increase to 5.828335 at $N = 2145$, which seems incorrect since then we have

$$\max(Aw)_j/w_j - \min(Aw)_j/w_j > 10^{-6}.$$

This application of the global algorithm just mentioned leads to some improvements of [6]. The results will be published elsewhere.

Hopefully, this paper proves the value of the algorithms developed in [4-8].

Acknowledgements The authors thank MS Zhou-Jing Wang for providing a program in MatLab on the Householder transformation for Hermitian matrices. This work is supported in part by National Natural Science Foundation of China (Grant No. 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Chen M F. Eigenvalues, Inequalities, and Ergodic Theory. London: Springer, 2005
2. Chen M F. Speed of stability for birth–death processes. Front Math China, 2010, 5(3): 379–515
3. Chen M F. Criteria for discrete spectrum of 1D operators. Commun Math Stat, 2014, 2: 279–309
4. Chen M F. Efficient initials for computing the maximal eigenpair, Front Math China, 2016, 11(6): 1379–1418
   See also Vol 4 in the middle of author's homepage: http://math0.bnu.edu.cn/~chenmf
   A package based on the paper is available on CRAN now (by X.J. Mao). One may check it through the link:
   https://github.com/mxjki/PowerfulMaxEigenpair
   A Matlab package is also available, see the author's homepage above
   The authors' papers cited in this article can be found from Vols 1–4 in the middle of the homepage above.
5. Chen M F. The charming leading eigenpair. Adv Math(China), 2017, 46(4): 281–297
6. Chen M F. Global algorithms for maximal eigenpair. Front Math China, 2017, 12(5): 1023–1043
7. Chen M F. Trilogy on computing maximal eigenpair. In: Yue W, Li Q L, Jin S, Ma Z, eds. Queueing Theory and Network Applications (QTNA 2017). Lecture Notes in Comput Sci, Vol 10591. Cham: Springer, 2017, 312-329
8. Chen M F. Hermitizable, isospectral complex matrices or differential operators. Front Math China, 2018, 13(6): 1267–1311
9. Chen M F, Zhang X. Isospectral operators. Commun Math Stat, 2014, 2: 17–32
10. Cipra B A. The best of the 20th century: Editors name top 10 algorithms. SIAM News, 2000, 33(4): 1–2

11. Frolov A V, Voevodin V V, Teplov A. Thomas algorithm, pointwise version. https://algowiki-project.org/en/$Thomas_$algorithm,_pointwise_version
12. From Wikipedia. Tridiagonal matrix algorithm https://en.wikipedia.org/wiki/Tri-diagonal_matrix_algorithm
13. Golub G H, van der Vorst H A. Eigenvalue computation in the 20th century. J Comput Appl Math, 2000, 123(1–2): 35–65
14. Householder A S. Unitary triangularization of a nonsymmetric matrix. J Assoc Comput Mach, 1958, 5: 339–342
15. Moler C. Llewellyn Thomas. https://en.wikipedia.org/wiki/Llewellyn_Thomas 1996
16. Shukuzawa O, Suzuki T, Yokota I. Real tridiagonalization of Hermitian matrices by modified Householder transformation. Proc Japan Acad Ser A, 1996, 72, 102–103
17. Stewart G W. The decompositional approach to matrix computation. IEEE Comput Sci Eng, 2000, 2(1): 50–59
18. Tang T, Yang J. Computing the maximal eigenpairs of large size tridiagonal matrices with O(1) number of iterations. Numer Math Theory Methods Appl, 2018, 11(4):877–894
19. van der Vorst H A, Golub G H. 150 Years old and still alive: eigenproblems. In: Duff I S, Watson G A, eds. The State of the Art in Numerical Analysis. Oxford: Oxford Univ Press, 1997, 93–119

Front. Math. China 2019, 14(6): 1077–1116
https://doi.org/10.1007/s11464-019-0799-z

Improved global algorithms for maximal eigenpair

Mu-Fa Chen and Yue-Shuang Li

(Beijing Normal University)

May 27, 2019

Abstract

This paper is a continuation of our previous paper [Front. Math. China, 2017, 12(5): 1023–1043] where global algorithms for computing the maximal eigenpair were introduced in a rather general setup. The efficiency of the global algorithms is improved in this paper in terms of a good use of power iteration and two quasi-symmetric techniques. Finally, the new algorithms are applied to Hua's economic optimization model.

2000 *Mathematics Subject Classification*: 65F15, 65F10, 68Q25, 93E15, 60J27
Key words and phrases. Maximal eigenpair, global algorithm, power iteration, shifted inverse iteration, quasi-symmetrization.

1 Introduction.

Let A be a nonnegative matrix. Given v (vector) and z (constant), there are three main points for the main algorithm proposed in [5]. The first one is solving the linear equation:

$$(zI - A)w = v$$

in w. Analytically, one may rewrite it as

$$w = (zI - A)^{-1}v,$$

and hence it is called the shift inverse iteration. The second one is the choice of initial vector $v^{(0)}$ (regarded as a mimic of the maximal eigenvector). For the global use, it was simply chosen to be the uniform one $\mathbb{1}$ (the column

vector having components 1 everywhere). The third point is Collatz–Wielandt formula (variational formula for the maximal eigenvalue):

$$\sup_{x>0}\min_{i\in E}\frac{(Ax)_i}{x_i}=\rho(A)=\inf_{x>0}\max_{i\in E}\frac{(Ax)_i}{x_i}. \tag{1}$$

Due to the fact that the convergence speed of the shift inverse algorithm is faster than the inverse one, the algorithm is often practical. The advantage of the initial $v^{(0)}=\mathbb{1}$ and the formula (1) is that they can be used in common. But they are clearly not as effective as we can imagine.

Clearly, each of the three points mentioned above is actually not new in the field. However, the following example proves the value of [5; Algorithm 2].

Example 1 Let $H=(h_{k\ell})_{k,\ell=1}^{N}$ with

$$h_{k\ell}=\frac{2^{k\wedge\ell}-1}{2^{\ell}}+\mathbb{1}_{\{\ell\leqslant k\}}.$$

In what follows, we adopt three methods to compute the maximal eigenpair of the matrix H.

(a) By using the package 'eig'(eigenvalues and eigenvectors) in MatLab (version R2016b) to compute the maximal eigenpair of H: $(g,\lambda_0)=\mathrm{eigs}(H,1)$. The correct eigenpair can be computed up to $N=45$ in terms of

$$\max\{(Hg)_j/g_j\}-\min\{(Hg)_j/g_j\}<10^{-6}$$

to confirm the result.

(b) Next, when we use the package 'Eigensystem' in Mathematica (version 11.3) on PC, similar result appears. Starting from $N=40$, the components of the output of the vector have different signs.

(c) However, when we use [5; Algorithm 2](which is reviewed right after Algorithm 2 in Section 2), it is well done up to $N=1795$ by saving 10^{-6} precision

$$\max\frac{(Hw)_j}{w_j}-\min\frac{(Hw)_j}{w_j}<10^{-6},$$

where w is the output of the maximal eigenvector approximation. It takes 234 iterations to get the stable approximation of the eigenvalue: 8.99978. The larger number of iterations is due to the fast decay of the approximating eigenvector, up to 10^{-317}. But for $N=1796$, the output seems incorrect since then we have

$$\max\frac{(Hw)_j}{w_j}-\min\frac{(Hw)_j}{w_j}>10^{-6}.$$

Example 1 cshows that [5; Algorithm 2] is meaningful but also illustrates that some skills are still needed to improve [5; Algorithm 2]. It is a challenge for us to find some way to improve the algorithm. Very fortunately, we can now answer partially the question in the present paper.

Since it takes a non-trivial way to obtain the improved algorithms, we use several sections to describe step by step with illustrations, the progress of the improvements. For the reader who is in hurry to see the main result, one may simply have a look at Sections 3 and 4. In Section 2, the improved algorithm with the help of the power iteration is introduced and its effectiveness is illustrated. In Section 3, the main algorithm of the paper is presented. For which, we adopt a quasi-symmetrizing technique to leveling the amplitude of the matrices. Furthermore, a technique to symmetrizing the maximal eigenvector is given in Section 4. In Section 5, an improved algorithm corresponding to sparse matrices is presented and three examples are computed to illustrate the power of the algorithm. In Section 6, we first prove two assertions: the first one is the positivity of the inverse Q-matrices, the second one is the convergence of the algorithms. Besides, the explicit representation of the inverse for a single birth Q-matrix or a single death one is presented. In Section 7, an application of our algorithms to Hua's economic optimization is included.

2 Good use of power iteration

In what follows, let the matrix $A = (a_{ij} : 0 \leqslant i, j \leqslant N)$ be irreducible and nonnegative: $a_{ij} \geqslant 0$. Here, we omit the trivial case that $\sum_j a_{ij} \equiv \text{constant}\, m > 0$. There is a natural improvement of [5; Algorithm 2] using power iteration. Throughout the paper, we often use the ℓ^2-norm $\|v\|$ of a vector v. Certainly, one is free to use other norms.

Algorithm 2 (Improved algorithm with power iteration) Let $A = (a_{ij})$ be given.

(i) Initials. $\big(y^{(0)}, z^{(0)}\big)$: let

$$w^{(0)} = A\left(\frac{\mathbb{1}}{\|\mathbb{1}\|}\right), \quad x^{(0)} = A\left(\frac{w^{(0)}}{\|w^{(0)}\|}\right), \quad u^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}, \quad y^{(0)} = Au^{(0)}.$$

Replacing A with A^*(the transpose of A), we have:

$$\widehat{w} = A^*\frac{\mathbb{1}}{\|\mathbb{1}\|}, \quad \widehat{x} = A^*\frac{\widehat{w}}{\|\widehat{w}\|}, \quad \widehat{u} = \frac{\widehat{x}}{\|\widehat{x}\|}, \quad \widehat{y} = A^*\widehat{u}.$$

Next, define

$$z^{(0)} = \left(\max_{0\leqslant i\leqslant N}\frac{y_i^{(0)}}{u_i^{(0)}}\right) \bigwedge \left(\max_{0\leqslant i\leqslant N}\frac{\widehat{y}_i}{\widehat{u}_i}\right).$$

Here, for given two numbers a and b, $a \wedge b = \min\{a, b\},\ a \vee b = \max\{a, b\}$.

(ii) Iterations. $\left(y^{(n)}, z^{(n)}\right)$ $(n \geqslant 1)$: given $y = y^{(n-1)}$ and $z = z^{(n-1)}$, let $v = y/\|y\|$ and set $w = w^{(n)}$ to be the solution to the equation

$$(zI - A)w = v. \tag{2}$$

Then, define

$$x^{(n)} = A\frac{w}{\|w\|}, \quad u^{(n)} = \frac{x^{(n)}}{\|x^{(n)}\|}, \quad y^{(n)} = Au^{(n)}.$$

$$\widetilde{z}^{(n)} = \min_{0\leqslant j\leqslant N} \frac{y_j^{(n)}}{u_j^{(n)}}, \quad z^{(n)} = \max_{0\leqslant j\leqslant N} \frac{y_j^{(n)}}{u_j^{(n)}}.$$

If at some $n \geqslant 1$, $z^{(n)} - \widetilde{z}^{(n)} < 10^{-7}$ (or $|z^{(n)} - z^{(n+1)}| < 10^{-7}$)(say!), then stop the computation. At the same time, regard $\left(z^{(n)}, y^{(n)}\right)$ as an approximation of the maximal eigenpair of A. Then the sequence $\{\left(z^{(n)}, y^{(n)}\right)\}_{n\geqslant 0}$ converges to the maximal eigenpair of A. Moreover, the resulting $\{z^{(n)}\}$ $\left(\text{resp.}, \{\widetilde{z}^{(n)}\}\right)$ is decreasing (resp., increasing) in n.

For simplicity, in what follows we say that "use row for $z^{(0)}$" in the case that the first term in the definition of $z^{(0)}$ is smaller or equal to the second one, since then we mainly use the rows of A in the products with a fixed column vector. Otherwise, we say that "use column for $z^{(0)}$".

In Algorithm 2, the approximation $\{y^{(n)}\}$ of the maximal eigenvector is different from those in [5]. Here $y^{(n)}$ is not normalized. The precision setting to be 10^{-7} is for safe to a designed precise level of the outputs, but all the examples in this paper are computed with the precision of 10^{-6}. Algorithm 2 is obtained by adding two steps of power iteration to [5; Algorithm 2]. Since the shift inverse iteration depends heavily on the initial $z^{(0)}$ (as well as $v^{(0)}$), a natural way to look for an improvement of the global algorithm is examining the development of the study on the estimation of the maximal eigenvalue for nonnegative matrices. Meanwhile, we found a nice progress has been made in recent years, see for instance [1, 13, 20]. In particular, we found that the estimates provided in [20] does improve our original algorithm of $z^{(0)}$. The three papers just cited started at the same initial estimates, which were given by (1) with the simplest choice $x = \mathbb{1}$, as we did in our algorithm. After working for some weeks along this line, we realized that the method in both [1] and [20] is more or less a modification of the power iteration, since the main tool used in these papers is the power $A^2\mathbb{1}$ or more general $A^m\mathbb{1}$ $(m \in \mathbb{N})$. For this, we simply choose $A^m\mathbb{1}$ instead of $\mathbb{1}$ as a new initial $v^{(0)}$ and adopt once again (1) with $x = v^{(0)}$ to produce new initial $z^{(0)}$. This is actually a round way in our study. From the first example in [4], we learnt that the power iteration is nearly impractical since its convergence speed is too slow. However, we also know from the example that the convergence speed is not so slow at the beginning iterations. This leads us to come back to the power iteration.

For later use, we fix a phase "2 iterations for $v^{(0)}$". That is at the beginning part (i) in Algorithm 2: we use the power iteration twice, in $w^{(0)}$ and $x^{(0)}$, respectively, ignoring the successive $u^{(0)}$. Usually, we use the phrase "2 iterations" only in what follows. Occasionally, we adopt "4 iterations for $v^{(0)}$", which should be now understandable and so we are not going to state the details here. Let us look at some examples to illustrate the effectiveness of power iteration used in Algorithm 2. Examples 3-5 are computed using MATLAB R2016b.

Example 3 (Circulant matrix) Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & \cdots & N-1 & N \\ 2 & 1 & 2 & \cdots & N-2 & N-1 \\ 3 & 2 & 1 & \cdots & N-3 & N-2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ N-1 & N-2 & N-3 & \cdots & 1 & 2 \\ N & N-1 & N-2 & \cdots & 2 & 1 \end{pmatrix}.$$

For this model, A is positive and symmetric, and its maximal eigenvector is symmetric. Tables 1 and 2 present the approximating outputs $(z^{(k)})$ of the maximal eigenvalue of A corresponding to the fixed phases 2 and 4 iterations for $v^{(0)}$, respectively.

Table 1. Approximating outputs $(z^{(k)})$ for different N using Algorithm 2 (2 iterations for $v^{(0)}$)

N	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$
8	29.7241	29.638	
32	388.542	386.326	386.325
100	3594.75	3570.32	
500	87974.7	87334.7	87334.5
1000	350950	348374	348373
1600	897520	890911	890910

Table 2. Approximating outputs $(z^{(k)})$ for different N using Algorithm 2 (4 iterations for $v^{(0)}$)

N	$z^{(0)}$	$z^{(1)}$	N	$z^{(0)}$	$z^{(1)}$
8	29.6396	29.638	500	87355.6	87334.5
32	386.386	386.325	1000	348459	348373
100	3571.08	3570.32	1600	891129	890910

In Tables 1 and 2, the results of $N = 8$ are checked by both Mathematica and MatLab. As a comparison, the outputs using [5; Algorithm 2] are presented in Table 3 below.

Table 3. Approximating outputs $(z^{(k)})$ for different N using [5; Algorithm 2]

N	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$
8	30.519	29.6602	29.638	
32	414.272	387.922	386.331	386.325
100	3883.74	3591.51	3570.43	3570.32
500	95577.5	87927.4	87338.1	87334.5
10^3	381553	350778	348388	348373
1600	976051	897098	890948	890910

Example 4 (Sequential array) Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & \cdots & N-1 & N \\ N+1 & N+2 & N+3 & \cdots & 2N-1 & 2N \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ N^2-N+1 & N^2-N+2 & N^2-N+3 & \cdots & N^2-1 & N^2 \end{pmatrix}.$$

When $N = 4$, it is better to use column than row for the initial $z^{(0)}$. By Algorithm 2, the computation is done in 1 step:

$$\{z^{(k)}\}_{k=0}^{1}: \quad 36.2222, \quad 36.2094.$$

This was done in [4; Example 14] in 4 iterations.

When $N = 50$, by Algorithm 2, the computation is also done in 1 step:

$$\{z^{(k)}\}_{k=0}^{1}: \quad 62938.6, \quad 62938.6.$$

For this example, the results of $N = 4$ and $N = 50$ are checked by both Mathematica and MatLab. For different $N \geqslant 100$, the outputs $\{z^{(n)}\}$ of Algorithm 2 are given in Table 4. All the computations need only 1 steps.

Table 4. Outputs $\{z^{(n)}\}$ for different N using Algorithm 2

N	$z^{(0)}$	$z^{(1)}$	N	$z^{(0)}$	$z^{(1)}$
100	501710.85	501710.82	1000	500167110.85	500167110.82
500	62541888.63	62541888.59	2000	4000667555.3	4000667555.26

Example 5 (Continued) Let H be the same as the one in Example 1. Suppose that $\widetilde{H} = (\widetilde{h}_{k\ell})$ is obtained from H by omitting the term $\mathbb{1}_{\{\ell \leqslant k\}}$:

$$\widetilde{h}_{k\ell} = \frac{2^{k \wedge \ell} - 1}{2^{\ell}}.$$

Using Algorithm 2, let iter denote the iterations needed to get the expected stable result z^{iter}, and v^{iter} denote the expected stable eigenvector approximation. Tables 5 and 6 present the final outputs using Algorithm 2, where iter0 denotes the iterations needed using [5; Algorithm 2] to get the same expected result.

Table 5. Outputs of H for different N using Algorithm 2

N	iter0	iter	z^{iter}	$\max v^{\text{iter}}/\min v^{\text{iter}}$
10	8	4	6.357802	2.239986×10^{-12}
100	28	17	8.934787	7.743369×10^{-9}
500	89	61	8.997203	6.594556×10^{-8}
1000	150	107	8.999295	9.026395×10^{-8}
1500	204	149	8.999686	1.617165×10^{-7}
1795	234	173	8.999781	2.762561×10^{-7}
1796	234	173	8.999781	1.291966×10^{-6}

Table 6. Outputs of $\widetilde{H}$ for different N using Algorithm 2

N	iter0	iter	z^{iter}	$\max v^{\text{iter}}/\min v^{\text{iter}}$
10	6	3	3.950588	3.838872×10^{-9}
100	25	15	5.784423	2.677261×10^{-7}
500	80	55	5.826557	9.008915×10^{-8}
1000	133	96	5.827956	5.365588×10^{-7}
1500	181	134	5.828217	6.120925×10^{-8}
2000	226	169	5.828309	9.745920×10^{-7}
2099	235	176	5.82832	9.718122×10^{-7}
2100	235	176	5.82832	1.727677×10^{-6}

The above three examples show that the power iteration works well in choosing the initial $(v^{(0)}, z^{(0)})$. Now, we want to show that the conclusion may fail in some cases. Examples 6 and 7 are accomplished using Mathematica.

Example 6 ([5; Example 7]) Define

$$Q = \begin{pmatrix} -1 & 1 & & & & \\ a_2 & -a_2-2 & 2 & & & \\ a_3 & & -a_3-3 & 3 & & \\ \vdots & & & \ddots & \ddots & \\ a_{N-1} & & 0 & & -a_{N-1}-N+1 & N-1 \\ a_N & & & & & -a_N-N \end{pmatrix}$$

on $E = \{1, \cdots, N\}$. For simplicity, we specify $a_k = 1/k$. Next, define a positive shift of Q:

$$A = Q + (N + N^{-1})\, I.$$

Then, A is a nonnegative matrix.

First, we want to know the effectiveness of the power iteration for A. To do so, let

$$v^{(0)} = \frac{\mathbb{1}}{\sqrt{N}}, \qquad z^{(0)} = \max_{1 \leqslant j \leqslant N} \frac{(Av^{(0)})_j}{v_j^{(0)}},$$

$$v^{(k)} = A\frac{v^{(k-1)}}{\|v^{(k-1)}\|}, \qquad z^{(k)} = \max_{1 \leqslant j \leqslant N} \frac{(Av^{(k)})_j}{v_j^{(k)}}.$$

Clearly, $v^{(k)}$ and $z^{(k)}$ are the mimic of the maximal eigenvector and the corresponding maximal eigenvalue, respectively. When $N = 50$, the outputs of the power iteration are given by Table 7. The outputs are checked by both Mathematica and MatLab.

Table 7. Outputs of power iteration for Example 6

k	$z^{(k)}$
$1 \leqslant k \leqslant 93$	50.02
$94 \leqslant k \leqslant 98$	50.0199
$99 \leqslant k \leqslant 100$	50.0198

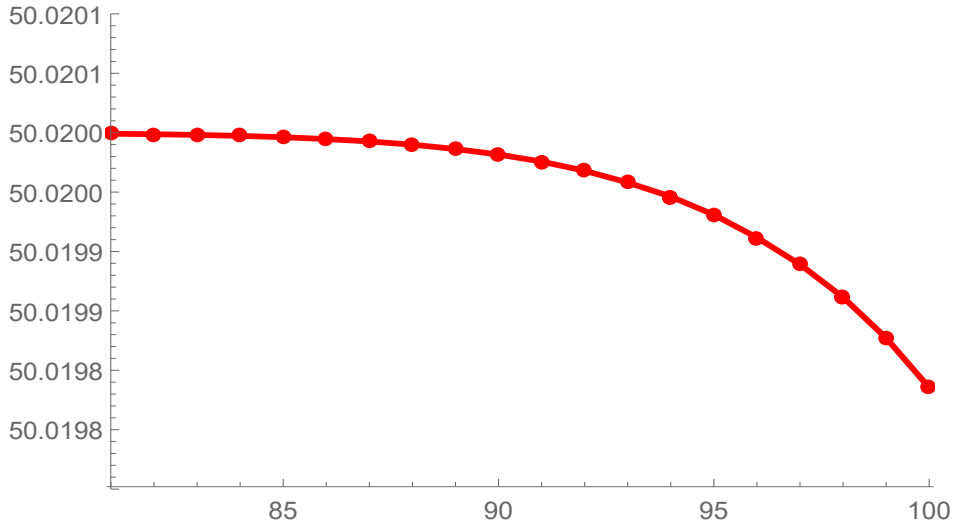


Fig.1 Figure of $z^{(k)}$ for $k = 0, 1, \cdots, 100$

Fig.1 shows that it is too slow for the sequence $(z^{(k)})$ to converge to the maximal eigenvalue $\rho(A) = 49.6592$. Now, what is the reason? To see this, recall that the iteration is mainly designed for the convergence of the approximating sequence $(v^{(k)})$ of the maximal eigenvector. We now look at the outputs of $v^{(5)}$:

$$(7.37395, \cdots, 7.37395, 1.38646, .0825569, .0100788, .00601827, 0.0029484).$$

This says that after 5 iterations, among the 50 components of the vector $v^{(5)}$, only its last 5 components are changed, all the others remain to be the same 7.37395. Next, if we look at the outputs of $v^{(10)}$, then only the last 10 components of $v^{(10)}$ are changed. The numbers of modified components are actually decreasing in k. This is due to the fact that the matrix A is very sparse and its conservativity except its last line. Hence, the resulting $(z^{(k)})$ can keep to be the same 50.02 until $k = 93$.

So far, we have seen that for positive matrices (Examples 4, 5), the power iteration is powerful enough, however, for sparse matrix (Example 6), it is rather poor. To confirm this conclusion, let us give a revised example: reset

$$A_Q = (-Q)^{-1}.$$

Then, as will be proved later (Lemma 21), A_Q is positive. We now apply our algorithm to this A_Q. By Algorithm 2, we arrive at the required result at the second step:

$$\{z^{(k)}\}_{k=0}^{2}: \quad 2.83409, \quad 2.77215, \quad 2.77174. \tag{3}$$

This example was computed in [5; Example 7] with 5 iterations.

Example 7 (A single death model) Let $Q=(q_{ij})_{i,j=1}^{N}$:

$$Q=\begin{pmatrix} -\frac{7}{3^2} & \frac{2}{3^3} & \frac{2}{3^4} & \cdots & \frac{2}{3^{N-1}} & \frac{2}{3^N} & \frac{1}{3^N} \\ \frac{2}{3} & -\frac{7}{3^2} & \frac{2}{3^3} & \cdots & \frac{2}{3^{N-2}} & \frac{2}{3^{N-1}} & \frac{1}{3^{N-1}} \\ & \frac{2}{3} & -\frac{7}{3^2} & \ddots & \ddots & \frac{2}{3^{N-2}} & \frac{1}{3^{N-2}} \\ & & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & -\frac{7}{3^2} & \frac{2}{3^2} & \frac{1}{3^2} \\ & \mathbf{0} & & & \frac{2}{3} & -\frac{7}{3^2} & \frac{1}{3^2} \\ & & & & & \frac{2}{3} & -\frac{2}{3} \end{pmatrix}.$$

and

$$A = Q + \frac{7}{9} I.$$

Clearly, A is nonnegative. As in Example 6, we now apply the power iteration to the present A. When $N = 50$, the outputs are given in Table 8.

Table 8. Outputs of power iteration for Example 7

k	$z^{(k)}$
$1 \leqslant k \leqslant 44$	$7/9 \approx 0.777778$
$k = 45$	0.777728
$k = 46$	0.777308
$k = 50$	0.77263

The critical point here is $k = 45$, smaller than that in Example 6. This is reasonable since the matrix A in Example 7 is much less sparse than the one in Example 6.

Again, following the previous example, we now apply our algorithm to $H(= A_Q) = (-Q)^{-1}$. It seems rather strange that we need 10 iterations to achieve the required result, much more than 2 iterations for the previous example. The outputs of $\{z^{(k)}\}_{k=0}^{10}$ are listed below:

$$\begin{gathered} 27.7745,\ 19.2867,\ 14.3267,\ 11.7687,\ 10.3249,\ 9.49169, \\ 9.03474,\ 8.8256,\ 8.76758,\ 8.7628,\ 8.76276. \end{gathered} \tag{4}$$

The different iteration numbers in (3) and (4) is due to the different amplitudes of A_Q and H in Examples 6 and 7, as shown in Table 9.

Table 9. Amplitude of the matrices in Examples 6 and 7

Example	Amplitude
Example 6	$\min A_Q = 8.16327 \times 10^{-6}, \quad \max A_Q = 1.802$
Example 7	$\min H = 8.88178 \times 10^{-16}, \quad \max H = 2$

In conclusion, our algorithm becomes less efficient if the amplitude of the matrix is bigger. The main task of the next section is to overcome this difficulty.

The next section introduces a method to leveling the amplitude of the bigger matrices. Algorithm 8 given in the next section is the main algorithm of this paper.

3 Quasi-symmetrizing technique for matrices

Recalling that an efficient algorithm for tridiagonal matrices was introduced in [4; §3] to compute the maximal eigenpair of matrices, later, improved algorithms were proposed in the following-up articles [5] and [9]. Refer to paper [10] for the development of the powerful algorithm for maximal eigenpair of Hermitizable matrices in [4; §3], [5; §A.4] and [9; §4]. Among which, the symmetrizing technique plays a crucial role.

The main algorithm of the paper is to use all the techniques mentioned above to improve the efficiency of [5; Algorithm 2]. In what follows, for given vector $u = (u_k)$, $\mathrm{Diag}(u)$ denotes the diagonal matrix having diagonal elements (u_k).

Algorithm 8 Let $A = (a_{ij})$ be given. Define a conservative Q-matrix

$$Q = A - \mathrm{Diag}(A\mathbb{1}),$$

denote by $\mu = (\mu_0, \mu_1, \cdots, \mu_N)$ with $\mu_0 = 1$ the unique solution to the equation

$$\mu Q = 0.$$

Then, define the quasi-symmetric matrix $A^{\rm sym}$:

$$A^{\rm sym} = {\rm Diag}(\mu^{1/2})A{\rm Diag}(\mu^{-1/2}).$$

With the preparations $A^{\rm sym}$ and μ at hand, we can now give the initials and iterations to get the maximal eigenpair of A . In fact, everything is the same as Algorithm 2 except replacing A with $A^{\rm sym}$. If at some $n \geqslant 1$, $z^{(n)} - \widetilde{z}^{(n)} < 10^{-7}$ (or $|z^{(n)} - z^{(n+1)}| < 10^{-7}$)(say!), then stop the computation. At the same time, regard $\big(z^{(n)}, {\rm Diag}(\mu^{-1/2})y^{(n)}\big)$ as an approximation of the maximal eigenpair of A. Then the sequence $\{\big(z^{(n)}, {\rm Diag}(\mu^{-1/2})y^{(n)}\big)\}_{n\geqslant 0}$ converges to the maximal eigenpair of A. Moreover, the resulting $\{z^{(n)}\}$ $\big($resp., $\{\widetilde{z}^{(n)}\}\big)$ is decreasing (resp., increasing) in n.

To have a more concrete impression of the influence of the amplitude in the eigenproblem, let us consider a very simple model first.

Example 9 (Continued) Let

$$\widetilde{H} = (\widetilde{h}_{kj})_{k,j=1}^{N}, \qquad \widetilde{h}_{kj} = \frac{2^{k\wedge j} - 1}{2^j}. \tag{5}$$

Then $\widetilde{H}$ is positive, its elements $\widetilde{h}_{kj}$ have exponential decay in j.

As mentioned in [10; Example 20], to compute the maximal eigenpair of this model, the package 'Eigensystem' in Mathematica is available only up to $N = 11$; and the package 'eig' in MatLab works only up to $N = 50$. However, [5; Algorithm 2] works well up to $N = 2102$. While the result in Example 5 shows that it is well done up to $N = 2099$ using Algorithm 2.

Actually, the matrix $\widetilde{H}$ is symmetrizable in the following sense: there is a positive measure (μ_k) such that

$$\mu_i\widetilde{h}_{ij} = \mu_j\widetilde{h}_{ji}, \qquad 1 \leqslant i, j \leqslant N \text{ (here } \mu_j = 2^{-j}).$$

This is equivalent to the symmetry of the matrix

$$\widehat{H} = {\rm Diag}(\mu^{1/2})\widetilde{H}\,{\rm Diag}(\mu^{-1/2}). \tag{6}$$

Therefore, the computation of the maximal eigenpair of A is reduced to the one of $\widehat{H}$. Using Mathematica, applying Algorithm 2 to the matrix $\widehat{H}$, when $N = 2043$, we obtain the required result at the second step:

$$\{z^{(k)}\}_{k=0}^{2}: \quad 5.82843, \quad 5.82834, \quad z^{(2)} = 5.82831.$$

For the final $v^{(2)}$ (i.e., the approximation of the maximal eigenvector of $\widehat{H}$), rewrite it as v^* for simplicity, we have $\max v^*/\min v^* = 648.912$. The figure of v^* is given in Fig.2.

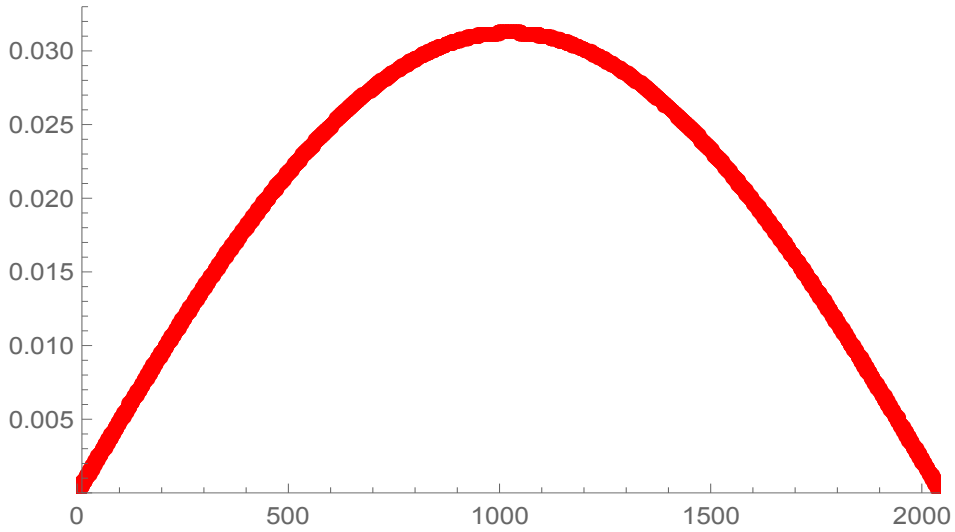


Fig.2 Figure of the final v^*.

Furthermore, it is easy to check that

$$-\widetilde{H}^{-1} = Q = \begin{pmatrix} -3 & 1 & & & & \\ 2 & -3 & 1 & & & \\ & 2 & -3 & 1 & & \\ & \ddots & \ddots & \ddots & & \\ & & 2 & -3 & 1 & \\ & & & 2 & -2 & \end{pmatrix}.$$

Hence, the computation of the maximal eigenpair of $\widetilde{H}$ is reduced to the one for Q, which is the birth–death type and is studied in detail in [10]. Hence, it is not hard now to solve the problem for large N, say $N = 15000$ for instance. But we are not going to the details here.

Anyhow, in the symmetrizable case, by (6), we have already reduced our problem to the symmetric one. Now, a new question arrives. How to handle with the non-symmetrizable case? For this, we introduce a notation.

Definition 10 For a given positive measure μ, the matrix $\widehat{H}$ defined by (6) is called a quasi-symmetrizing of $\widetilde{H}$ with respect to μ.

A basic fact, as already mentioned above, is that $\widetilde{H}$ is symmetrizable with respect to μ if and only if the matrix $\widehat{H}$ defined by (6) is symmetric. Therefore, the quasi-symmetrizing method is an extension of the symmetrizing technique. In our recent study on the same topic for triDiagonal matrices [9, 10], this technique plays an important role.

Now, the problem is that for a given A with nonnegative off-Diagonal elements, how to find a positive measure μ such that the quasi-symmetrized

matrix $\widehat{A}$ becomes as symmetric as possible. For this, we return to the symmetrizable situation. In this case, the analytic solution of μ was presented in [9]. Here, we introduce a computational construction.

Definition 11 (computational construction) Let A_0 be the matrix obtained from A by removing its Diagonal elements. Next, let

$$Q = A_0 - \text{DiagonalMatrix}[A_0\mathbb{1}].$$

Finally, let μ solve the equation

$$\mu Q = 0, \qquad \mu_0 = 1. \tag{7}$$

The measure μ is called the harmonic measure of Q.

Clearly, the matrix Q is a conservative Q-matrix (i.e., $Q\mathbb{1} = 0$), once Q is irreducible, the harmonic/invariant measure μ is positive and unique.

To justify the effectiveness of the quasi-symmetrizing idea, we return to Example 7. For this, we use an expression of $H = (-Q)^{-1}$ (for Q defined by Example 7) as follows.

$$H = (h_{kj})_{k,j=1}^{N}, \qquad h_{kj} = \frac{2^{k\wedge j} - 1}{2^j} + \mathbb{1}_{\{j\leqslant k\}}. \tag{8}$$

This result is obtained from oral communication with Y.H. Zhang and Y.T. Ma, based on [14; Theorem 9.3.3]. See also **Single death processes with DN boundary** at the last part of §6. Comparing (8) with (5) in Example 9, there is a newly added term $\mathbb{1}_{\{j\leqslant k\}}$.

Example 12 (Continued) Corresponding to Example 7, let H be defined by (8). Then by Algorithm 8, when $N = 50$, we obtain the required result in 4 iterations. The outputs are as follows:

$$\{z^{(k)}\}_{k=0}^{4}: \ 9.10864,\ 8.8773,\ 8.78454,\ 8.76351,\ 8.76276.$$

Usually, it should be enough to use the harmonic measure as the quasi-symmetrizing measure. Actually, it is the basis used in the examples in §4. In the symmetrizable case, this is quite natural. However, it is not completely necessary. For instance, the matrix H defined in (8) is closely related to (5), one may simply use the symmetrizing measure $\mu_k = 2^{-k}$ of (5) as the quasi-symmetrizing measure for the matrix defined by (8). Denote the resulting matrix by H_1. Then, our algorithm works well until $N = 523$. If one want to go further for large N, it is natural to look for the harmonic measure for H_1. However, the harmonic equation (7) is solvable if and only if $N \leqslant 303\,(< 523)$. Hence, we need a new idea.

It is a suitable position to have a random test of our main Algorithm 8. We would like to know not only the effectiveness of the algorithm but also to

count the percent of random examples which fail, as illustrated by Example 12 or by Example 13 in §4. We use the package "rand" in MatLab to produce the random data of the elements of the matrices, the data is chosen from $(0, 10)$. In an ordinary Notebook PC [Intel(R) Core(TM)i5-8350U CPU @1.70GHz(8 CPUs), ~1.9 GHz and 8192MB RAM], We have done two tests.

(1) Take $N = 5000$. The total 2,326 of examples are worked out in 7 hours. In average, each example costs less than 11 seconds.
(2) Take $N = 1000$. Then total 36,448 examples are done in two hours. In average, each one costs less than 0.2 second.

The computation goes fast, since we need only one iteration in the second part of the algorithm. In both tests, there is no failed example. The result of the tests is quite unexpected for us.

4 Quasi-symmetrizing method for maximal eigenvector

For the matrices with larger amplitude, the quasi-symmetrizing technique introduced in Section 3 is often practical for computing the maximal eigenpair. However, as illustrated by Example 13, sometimes, additional work is needed. In the following Examples, the initials use columns without mention again. Besides, all the examples in this section are completed using Mathematica.

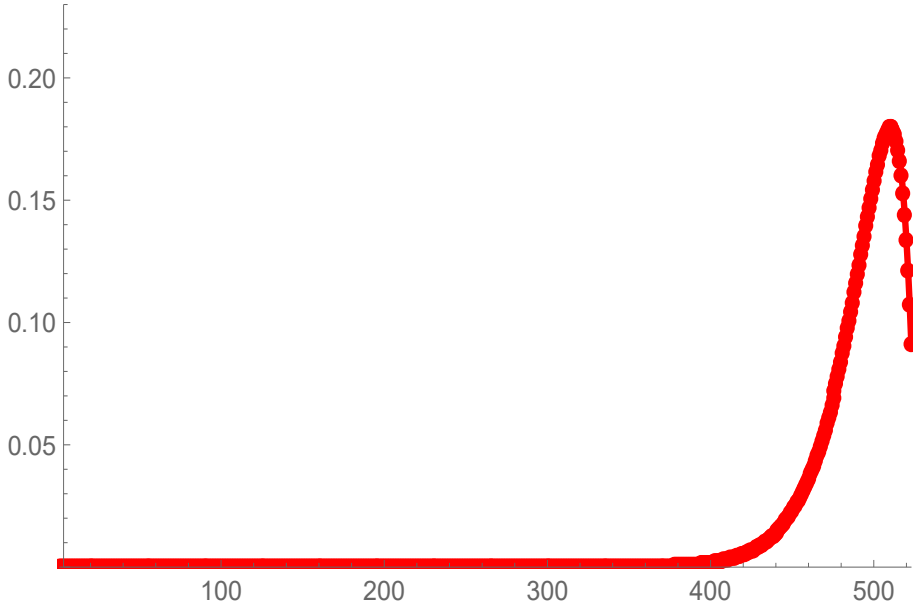


Fig.3 Figure of v^*.

Example 13 (Continued) As mentioned by the remark after Example 12, it is natural to adopt $\mu_k = 2^{-k}$ as the quasi-symmetrizing measure of (8), and the resulting matrix is denoted by H_1. Then, our algorithm takes 18 iterations to arrive at the required approximation v^* of maximal eigenvector (cf. Fig.3) when $N = 523$:

As an approximation of the maximal eigenvector, we have

$$\max v^* / \min v^* \approx 1.79 \times 10^{14}.$$

Hence, the maximal eigenvector of H_1 is far away from being symmetric. This leads to the next quasi-symmetrizing procedure. Recalling that our algorithm is essentially concentrated on the eigenvector, it is certainly important to make the maximal eigenvector to be as symmetric as possible. Even though in general we do not know how to solve the problem in a precise way, it is natural to make the value at the two endpoints of v^* to be as close as possible. In this way, by using optimal sequential search for instance (refer to [8] and references therein), we find a new quasi-symmetrizing measure as follows:

$$\mu_k = \left(2 \times 10^{2/39}\right)^{-k}.$$

We repeat that this μ is designed at $N = 523$ but used for each N listed in Table 10. Applying Algorithm 8 to the matrix H_1 and replacing μ in the Algorithm by $\mu_k = \left(2 \times 10^{2/39}\right)^{-k}$, denote the new quasi-symmetrized matrix by H_2, the outputs of our new quasi-symmetrized algorithm are given in Table 10.

Table 10. Approximation $\{z^{(n)}, \widetilde{z}^{(n)}\}$ of maximal eigenvalue for different N

N	523	800	1000	1500	1700
$\widetilde{z}^{(1)}$	8.96625	8.97632	8.98126	8.98993	8.99111
$z^{(1)}$	8.99793	8.99911	8.99943	8.99975	8.9998
$\widetilde{z}^{(2)}$	8.99734	8.99885	8.99926	8.99967	8.99974
$z^{(2)}$	8.99746	8.99891	8.9993	8.99969	8.99976
$\widetilde{z}^{(3)}$	8.99744	8.9989	8.9993	8.99969	8.99976
$z^{(3)}$	8.99744	8.9989	8.9993	8.99969	8.99976

The figure of the final stable maximal eigenvector approximation v^* of our new quasi-symmetrizing matrix H_2 for $N = 1700$ is given by Fig.4. It is quite surprising that the outputs look very much symmetric, and the figure is very much the same as Fig. 2.

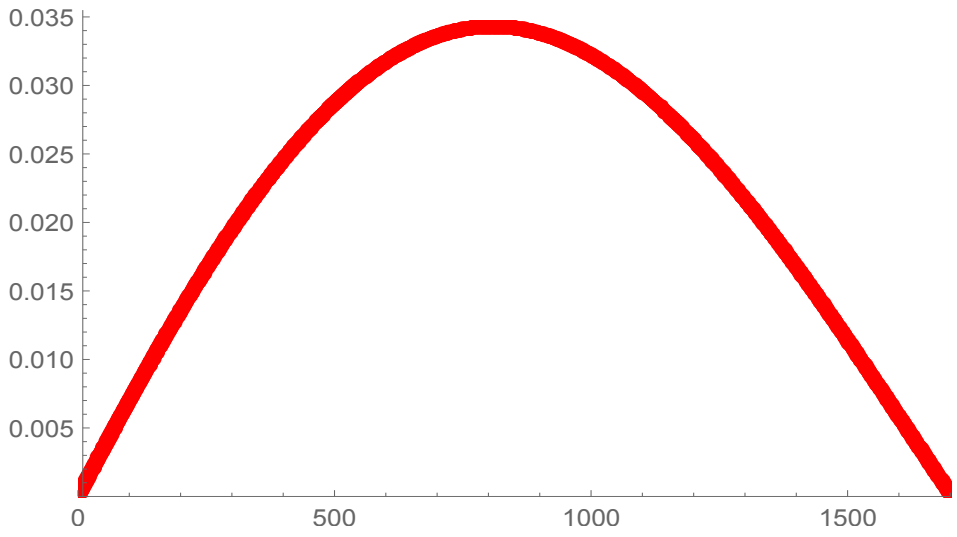


Fig.4 Figure of v^*, $\max v^*/\min v^* \approx 480$

Here, we remark that the amplitude of the second quasi-symmetrized matrix max / min may become worse. Table 11 presents the amplitude of the three matrices: the original matrix H, the one-step quasi-symmetrized matrix H_1 and the two-step quasi-symmetrized matrix H_2.

Table 11. Amplitude for different matrices when $N = 1700$

$\max H \times 10^{-256}/\min H$	1.1×10^{256}
$\max H_1 \times 10^{-256}/\min H_1$	2.12345
$\max H_2 \times 10^{-256}/\min H_2$	2.6×10^{43}

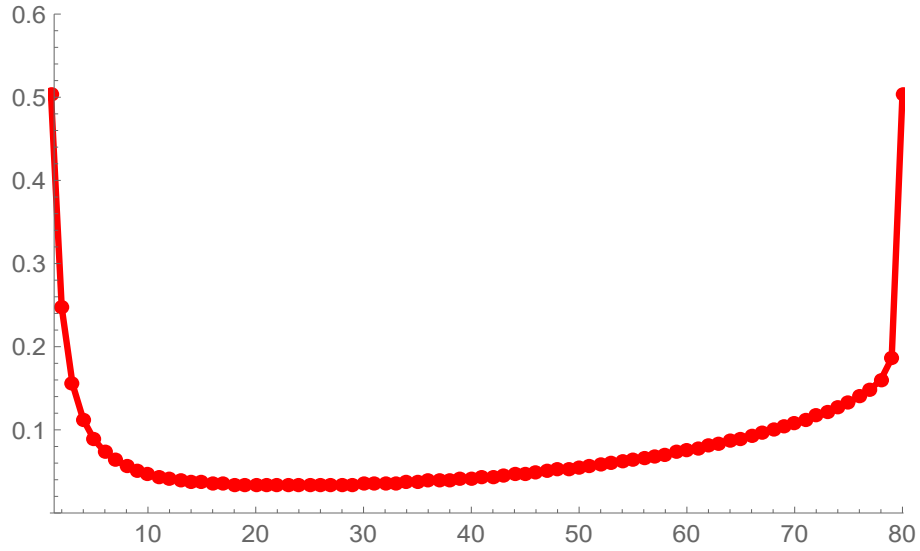


Fig.5 Figure of v^*. Use measure $\mu_k \times \exp[k/9.13]$, $\max v^*/\min v^* \approx 15.2$

We now introduce briefly more examples. In what follows, μ denote the corresponding harmonic measure.

Example 14 (Continued) Returning to Example 6, when $N = 80$, the figure of the final stable maximal eigenvector approximation v^* of the quasi-symmetrized matrix A_Q is given by Fig.5. The approximation of the corresponding maximal eigenvalue is as follows:

$$\{z^{(k)}\}_{k=0}^{2}: \quad 2.98274\,(\text{row}), \quad 2.83914, \quad 2.83862.$$

Here "row" in the last line means "use row for $z^{(0)}$". When $N = 400$, the figure of the final stable maximal eigenvector approximation of the quasi-symmetrized matrix A_Q is given by Fig.6.

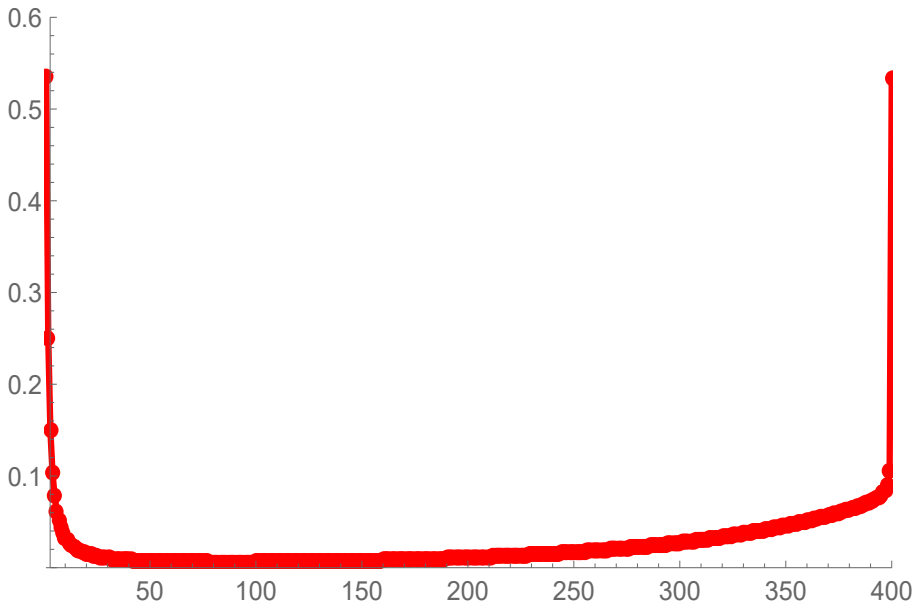


Fig.6 Figure of v^*. Use measure $\mu_k \times \exp[k/35.7]$, $\max v^*/\min v^* \approx 85.86$

The approximation of the corresponding maximal eigenvalue is as follows:

$$\{z^{(k)}\}_{k=0}^{2}: \quad 3.23091\,(\text{row}), \quad 2.95922, \quad 2.95732.$$

Example 15 (Continued) Finally, we return to Example 4. When $N = 80$, the figure of the final stable maximal eigenvector approximation $v*$ of the quasi-symmetrized matrix A is given by Fig.7.

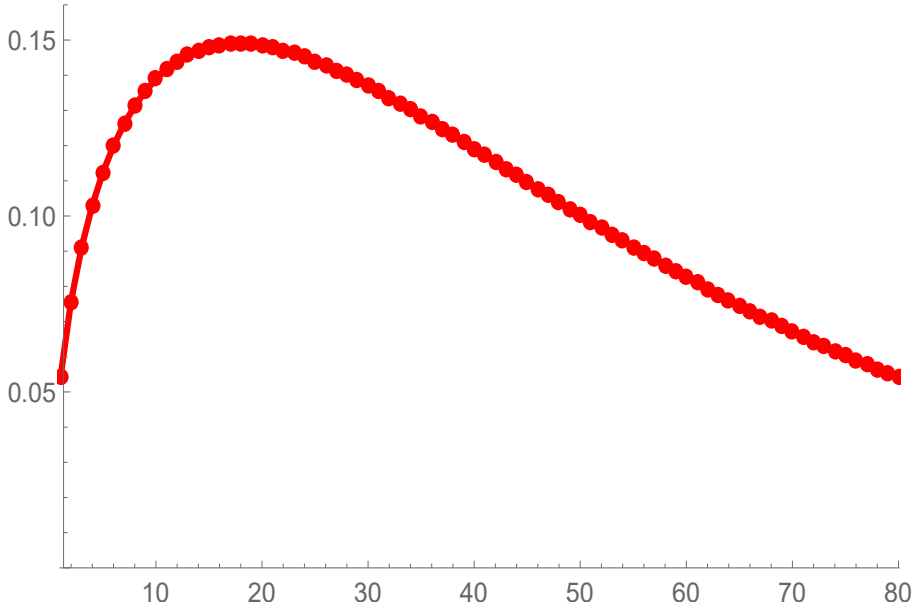


Fig.7 Figure of v^*. Use measure $\mu_k \times \exp[-k/17.288]$

The approximation of the corresponding maximal eigenvalue is as follows:

$$\{z^{(k)}\}_{k=0}^{1}: \quad 257208\,(\text{row}), \quad 257102.$$

For this example, the power iteration works well. When using 2 iterations for $v^{(0)}$, need only 1 iteration; when using 4 iterations for $v^{(0)}$, no further iteration is needed.

5 Improved algorithm for Q-matrix

Now, we have already explained every step of the improved Algorithm 8 in §2-§3. Besides, as can be predicted, for nearly positive matrices, solving the linear equation (2) by machine is harder than the one for sparse matrices. We now introduce another algorithm replacing A in (2) by sparse Q. Recalling the definition of Q-matrix:

$$Q = (q_{ij})_{i,j=0}^{N},$$

where $q_{ij} \geqslant 0$ for every pair $i \neq j$ and $-\infty < \sum\limits_{j=0}^{N} q_{ij} \leqslant 0$ for every $i \geqslant 0$. In what follows, we often assume that Q is irreducible.

Algorithm 16 Let $Q = (q_{ij})$ be a sparse Q-matrix. Set

$$A = (a_{ij}) = (-Q)^{-1}.$$

Everything is the same as Algorithm 8 except equation (2) is replaced by

$$(-Q^{\mathrm{sym}} - zI)w = v, \tag{9}$$

where

$$Q^{\mathrm{sym}} = \mathrm{Diag}(\mu^{1/2})Q\mathrm{Diag}(\mu^{-1/2}).$$

Besides, $z^{(n)}(n \geqslant 0)$ and $\widetilde{z}^{(n)}(n \geqslant 1)$ here are the inverse of those in Algorithm 8, which means at step n,

$$u^{(n)} = \frac{w}{\|w\|}, \quad y^{(n)} = A^{\mathrm{sym}}u^{(n)}.$$

$$\widetilde{z}^{(n)} = \max_{0\leqslant j\leqslant N} \frac{u_j^{(n)}}{y_j^{(n)}}, \quad z^{(n)} = \min_{0\leqslant j\leqslant N} \frac{u_j^{(n)}}{y_j^{(n)}}.$$

If at some $n \geqslant 1$, $\widetilde{z}^{(n)} - z^{(n)} < 10^{-7}$ (or $|z^{(n)} - z^{(n+1)}| < 10^{-7}$)(say!), then stop the computation. At the same time, regard $\left(z^{(n)}, \mathrm{Diag}(\mu^{-1/2})y^{(n)}\right)$ as an approximation of the minimal eigenpair of $-Q$. Moreover, the resulting $\{z^{(n)}\}$ $\left(\mathrm{resp.}, \{\widetilde{z}^{(n)}\}\right)$ is increasing (resp., decreasing) in n.

Examples 17-19 illustrate the power of Algorithm 16 using MATLAB R2016b.

Example 17 (Continued) Let Q be the same as that in Example 6, Table 12 presents the outputs using Algorithm 16 for larger matrices.

Table 12. Outputs using Algorithm 16 for larger matrices

N	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
100	0.344177	0.349006	0.349197	
500	0.330312	0.336506	0.337186	
1000	0.327542	0.333984	0.33501	
5000	0.324294	0.330556	0.332632	0.332635
10^4	0.323673	0.329604	0.332181	0.332188

Example 18 (Continued) Let Q be the same as that in Example 7. Noticing that $(-Q)^{-1} = H$, where H is defined by (8). Table 13 presents the outputs using Algorithm 16.

Table 13. Outputs using Algorithm 16

N	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$
10	0.153758	0.157099	0.157287		
20	0.121057	0.125727	0.126405	0.126417	
40	0.11059	0.11425	0.11546	0.11564	0.115643
55	0.109324	0.112008	0.113244	0.113609	0.113631

Here, the reason we do not go to large matrices is that we meet the same problem in solving μ when $N = 56$, the resulting μ by Definition 11 appears complex. In fact, for this example, even though the matrix Q has nearly half number of zeros, it is still not sparse enough.

For comparison, another example is given below which comes from [5; Example 9].

Example 19 The example is motivated from the classical branching process. Denote by $(p_k : k \geqslant 0)$ a given probability measure with $p_1 = 0$. Let

$$Q = \begin{pmatrix} -1 & p_2 & p_3 & \cdots\cdots & p_{N-2} & p_{N-1} & \sum_{k\geqslant N} p_k \\ 2p_0 & -2 & 2p_2 & \cdots\cdots & 2p_{N-3} & 2p_{N-2} & 2\sum_{k\geqslant N-1} p_k \\ & 3p_0 & -3 & \ddots & \ddots & 3p_{N-3} & 3\sum_{k\geqslant N-2} p_k \\ & & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & -(N-2) & (N-2)p_2 & (N-2)\sum_{k\geqslant 3} p_k \\ & & & & (N-1)p_0 & -(N-1) & (N-1)\sum_{k\geqslant 2} p_k \\ & & \mathbf{0} & & & Np_0 & -Np_0 \end{pmatrix},$$

where

$$p_0 = \frac{\alpha}{2},\ p_1 = 0,\ p_k = \frac{2-\alpha}{2^k},\ k = 2, 3, \cdots, \qquad \alpha \in (0, 2).$$

Using Algorithms 8 and 16, we get the same result. Table 14 presents the approximation of the minimal eigenvalues of $-Q$ when $\alpha = 7/4$.

Table 14. Approximation of the minimal eigenvalue of $-Q$

N	$z^{(0)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$
8	0.607604	0.637006	0.638152	0.638153
16	0.58672	0.623429	0.625536	0.625539
50	0.583721	0.622556	0.624995	0.625
100	0.583719	0.622555	0.624995	0.625
5000	0.583719	0.622555	0.624995	0.625
10000	0.553387	0.620935	0.624987	0.625

In fact, one may regard Algorithm 16 as a generalization of the tridiagonal algorithm in [10]. The next two sections collect the main theoretical part of the paper.

6 Proofs

In this section, we prove the convergence of the iteration sequences in the algorithms and the existence and the positivity of $(-Q)^{-1}$. At the same time, for two classes of Q-matrices often used in practice, the explicit expressions of $(-Q)^{-1}$ are presented.

Some preparations

First, we need a convergence result for the power iteration.

Lemma 20 Let $v^{(0)}$ be a positive column vector and A be a nonnegative and irreducible matrix. For the power iteration

$$v^{(n)} = A\frac{v^{(n-1)}}{\|v^{(n-1)}\|}, \qquad n \geqslant 1,$$

define

$$z^{(n)} = \sup_k \frac{\big(Av^{(n)}\big)_k}{v_k^{(n)}}.$$

Then $\{z^{(n)}\}$ converges decreasingly to the maximal eigenvalue $\rho(A)$ of A as $n \to \infty$.

Proof. First, by assumption, $v^{(0)}$ is positive, and then so is $v^{(n)}$ for each $n \geqslant 1$.

Next, there is nothing to do if $z^{(n)} \equiv \infty$. Otherwise, assume that n_0 is the first integer among $\{0, 1, \cdots, n_0\}$ for which $z^{(n_0)} < \infty$. For simplicity, assume that $n_0 = 0$.

For nonnegative sequence $\{\alpha_j\}$, real $\{a_j\}$ and $\{b_j\}$, if there is a real constant z such that

$$a_j \leqslant zb_j, \qquad \forall j,$$

then we first have

$$\sum_k \alpha_k a_k \leqslant z \sum_k \alpha_k b_k.$$

If furthermore $\sum_k \alpha_k b_k > 0$, then

$$\frac{\sum_k \alpha_k a_k}{\sum_k \alpha_k b_k} \leqslant z.$$

Noticing that in the definition of $z^{(n)}$, the normalizing constant $\|v^{(n-1)}\|$ plays no role, hence we may rewrite the power iteration as

$$v^{(n)} = Av^{(n-1)}, \qquad n \geqslant 1.$$

Therefore, by the elementary fact just proved, we have

$$\frac{\big(Av^{(n)}\big)_j}{v_j^{(n)}} = \frac{\big(Av^{(n)}\big)_j}{\big(Av^{(n-1)}\big)_j} = \frac{\sum_k a_{jk}\big(Av^{(n-1)}\big)_k}{\sum_k a_{jk}v_k^{(n-1)}} \leqslant \sup_k \frac{\big(Av^{(n-1)}\big)_k}{v_k^{(n-1)}} = z^{(n-1)}.$$

Making supremum with respect to j, we obtain $z^{(n)} \leqslant z^{(n-1)}$ as required. □

Next, we also need a conclusion about the inverse of a Q-matrix.

Lemma 21 Let Q be a Q-matrix (not necessarily conservative) on a countable set E, corresponding to some semigroup $\{P(t)\}_{t\geqslant 0}$. Then $G := \int_0^\infty P(t)dt \in [0,\infty]$ (pointwise). If moreover, Q is irreducible, $\{P(t)\}$ is transient and satisfying both the backward and forward Kolmogorov equations, then $(-Q)^{-1} = G$ is finite and positive (pointwise).*

Proof. In the weak topology (i.e., the pointwise convergence), we have the resolvent $R(\lambda)$ of the semigroup $\{P(t)\}$:

$$R(\lambda) = \int_0^\infty e^{-\lambda t} P(t)dt \geqslant 0.$$

According to [19; §4.3, Theorem 5] or [12; Theorem 1.7], in the weak topology, we have $R(\lambda) = (\lambda I - Q)^{-1}$. By the monotone convergence, we obtain $G = \lim_{\lambda\downarrow 0} R(\lambda) = \int_0^\infty P(t)dt$. Now, the irreducibility implies that $G > 0$, and the same property plus the transiency gives us $G < \infty$. One may refer to [11; §2.2]. To check that $(-Q)^{-1} = G$, it suffices to show that

$$\lim_{\lambda\downarrow 0}(-Q)R(\lambda) = \lim_{\lambda\downarrow 0} R(\lambda)(-Q) = I.$$

This follows immediately by the Kolmogorov equations:

$$(-Q)R(\lambda) = R(\lambda)(-Q) = I - \lambda R(\lambda)$$

(cf. [2; §2.2]) plus the transiency: $\lim_{\lambda\downarrow 0} \lambda R(\lambda) = 0$. □

The following result contains a shift z. Note that the minimal eigenvalue of $-Q$ can be zero in the case of $N = \infty$. Then the shift z in $-Q - zI$ is meaningless except $z = 0$. In which case, the conclusion we required is included in Lemma 21.

Lemma 22 Let Q be a Q-matrix on E. Assume that the minimal eigenvalue λ_0 of $-Q$ is positive. Then for every $z \in (0, \lambda_0)$, $(-Q - zI)$ is invertible. If moreover, $(-Q)^{-1}$ is finite and positive, then so is $(-Q - zI)^{-1}$.

Proof. Because $\lambda_0 > z > 0$, the module of each eigenvalue of $-Q - zI$ is bigger than or equal to $\lambda_0 - z > 0$. It follows that $(-Q - zI)$ is invertible. Now, since the eigenvalues $\{\lambda_j\}$ of $(-Q)$ satisfy

$$\frac{z}{|\lambda_j|} \leqslant \frac{z}{\lambda_0} < 1,$$

it follows that $\|z(-Q)^{-1}\| < 1$. Hence,

$$\big(I - z(-Q)^{-1}\big)^{-1} = \sum_{n=0}^{\infty} \big[z(-Q)^{-1}\big]^n. \tag{10}$$

*A little modification is made to the published version

If moreover, $(-Q)^{-1}$ is finite and positive, then so is $\big(I-z(-Q)^{-1}\big)^{-1}$ by (10). Now, the required assertion holds since

$$(-Q-zI)^{-1}=(-Q)^{-1}\big(I-z(-Q)^{-1}\big)^{-1}. \qquad \square$$

Finally, to prove the monotonicity of $\{z^{(n)}\}$, we need a convergence result for the shift inverse iteration. Before moving on, let us give some notation first. Under the assumptions of Lemma 21, let v be a given positive vector. Set

$$z=\min_{0\leqslant\ell\leqslant N}\frac{v(\ell)}{(-Q)^{-1}v(\ell)}. \tag{11}$$

By the Collatz-Wielandt formula corresponding to Q-matrix(c.f. [4; Corollary 12]), for each positive function g, we obtain

$$\lambda_0\geqslant\inf_{0\leqslant\ell\leqslant N}\frac{(-Q)g(\ell)}{g(\ell)}.$$

Replacing g by $(-Q)^{-1}v$, we obtain $\lambda_0\geqslant z$. If $\lambda_0=z$, then there is nothing to do (which is actually ruled out in our algorithms with restriction $z^{(n)}-\tilde{z}^{(n)}<10^{-7}$). Thus, without loss of generality, assume $z<\lambda_0$. Next, set $w=(-Q-zI)^{-1}v$ which is positive by Lemma 22. Define

$$z_1=\min_{0\leqslant\ell\leqslant N}\frac{w(\ell)}{(-Q)^{-1}w(\ell)}. \tag{12}$$

For the same reason above, assume $z_1<\lambda_0$. We have the following assertion.

Lemma 23 For z and z_1 defined by (11) and (12), respectively, we have

$$0<z\leqslant z_1<\lambda_0.$$

Proof. By the definition of z, we have

$$z(-Q)^{-1}v(\ell)\leqslant v(\ell). \tag{13}$$

Hence

$$0<w(\ell)=(-Q-zI)^{-1}v(\ell)\overset{(10)}{=}(-Q)^{-1}\sum_{n=0}^{\infty}\big[z(-Q)^{-1}\big]^n v(\ell), \tag{14}$$

and

$$\begin{aligned}(-Q)^{-1}\sum_{n=0}^{\infty}\big[z(-Q)^{-1}\big]^n v(\ell)&\overset{(13)}{\leqslant}\frac{v(\ell)}{z}+(-Q)^{-1}\sum_{n=1}^{\infty}\big[z(-Q)^{-1}\big]^{n-1}v(\ell)\\&\overset{(10)}{=}\frac{1}{z}(-Q)(-Q-zI)^{-1}v(\ell).\end{aligned}$$

Combining this with (14), we get

$$w \leqslant \frac{1}{z}(-Q)w.$$

According to Lemma 21, $(-Q)^{-1}$ is a positive operator, multiplying both sides by $z(-Q)^{-1}$, it follows that

$$z(-Q)^{-1}w \leqslant w.$$

Dividing both sides by the positive $(-Q)^{-1}w$ (pointwise), we obtain

$$z \leqslant \frac{w(\ell)}{(-Q)^{-1}w(\ell)}.$$

Making the infimum in both sides with respect to ℓ, it follows that

$$z \leqslant \inf_{\ell} \frac{w(\ell)}{(-Q)^{-1}w(\ell)} = z_1. \qquad \square$$

For $\{\delta_k\}$ used in [5; A.4] and [9; Algorithms 25 and 28], noticing that $vII(v) = (-Q)^{-1}v$ and $\delta_k = \max_j II_j(v^{(k)})$, the monotonicity of $\{\delta_k\}$ follows from Lemma 23.

Proof of the convergence in the algorithms

With Lemmas 20-23 at hand, we can prove the monotonicity of the algorithms. In fact, Algorithm 8 is equivalent with Algorithm 2 by adding the quasi-symmetry procedure. Thus, we just need to prove the monotonicity of $\{z^{(n)}\}$ and $\{\widetilde{z}^{(n)}\}$ in Algorithm 2. The next result describes a tight relation between Algorithms 8 and 16.

Remark 24 The sequences $(z^{(n)}, v^{(n)})$ obtained by Algorithms 8 and 16, respectively, are determined each other, as shown in (15) below.

Proof. To show the differences of Algorithms 8 and 16, we need some notation: for a given Q-matrix Q, set $A = (-Q)^{-1}$. Given initials (z_Q, v) and (z_A, v) satisfying that $z_A = 1/z_Q$, solve the equations in w_Q and w_A :

$$(-Q - z_Q I)w_Q = v \qquad \text{and} \qquad (z_A I - A)w_A = v,$$

respectively. Define $\widetilde{w}_Q = Aw_Q$,

$$z_Q^{(1)} = \min_k \frac{w_Q(k)}{\widetilde{w}_Q(k)}, \qquad v_Q^{(1)} = \frac{\widetilde{w}_Q}{\|\widetilde{w}_Q\|},$$

and write $w = Aw_A,\ \widetilde{w}_A = Aw$,

$$z_A^{(1)} = \max_k \frac{\widetilde{w}_A(k)}{w(k)}, \qquad v_A^{(1)} = \frac{\widetilde{w}_A}{\|\widetilde{w}_A\|}.$$

Then

$$z_Q^{(1)} = \frac{1}{z_A^{(1)}}, \qquad v_Q^{(1)} = v_A^{(1)}. \tag{15}$$

In fact, Lemmas 21 and 22 ensure the existence, the nonnegative and finite properties of $(-Q)^{-1}$ and $(-Q - z_Q I)^{-1}$. According to the equations above, we have

$$w_Q = (-Q - z_Q I)^{-1} v = (-Q)^{-1}\big(I - z_Q(-Q)^{-1}\big)^{-1} v = A(I - z_Q A)^{-1} v,$$

and

$$w_A = (z_A I - A)^{-1} v = z_A^{-1}(I - z_A^{-1} A)^{-1} v = z_Q(I - z_Q A)^{-1} v$$

since $z_A = (z_Q)^{-1}$ by assumption.

Hence, we have the identity

$$z_Q w_Q = A w_A.$$

By the definition of $\widetilde{w}_Q$ and $\widetilde{w}_A$, we have

$$z_Q \widetilde{w}_Q = \widetilde{w}_A.$$

Making normalizing in ℓ^2, it follows that

$$v_Q^{(1)} = \frac{\widetilde{w}_Q}{\|\widetilde{w}_Q\|} = \frac{z_Q \widetilde{w}_Q}{\|z_Q \widetilde{w}_Q\|} = \frac{\widetilde{w}_A}{\|\widetilde{w}_A\|} = v_A^{(1)}$$

by definition. Furthermore, we have

$$\frac{w_Q(k)}{\widetilde{w}_Q(k)} = \frac{w_Q(k)}{(A w_Q)(k)} = \frac{(A w_A)(k)}{(A(A w_A))(k)} = \left[\frac{\widetilde{w}_A(k)}{w(k)}\right]^{-1}.$$

Making infimum on both sides of the equality, we obtain $z_Q^{(1)} = 1/z_A^{(1)}$. □

As mentioned in the proof of Lemma 20, we can ignore the normalizing procedure at each step for the approximating eigenvectors. Combining Lemmas 20 with 23, one can easily understand that inserting several power iterations into the shift inverse iteration can accelerate the convergence of $z^{(n)}$. This idea is explained as follows.

Under the assumptions of Lemma 21, for a given positive vector w, let

$$A = (-Q)^{-1}, \quad \widehat{w} = A^m w (m \geqslant 1), \quad v = \frac{\widehat{w}}{\|\widehat{w}\|}.$$

Next, make a convention that $A^0 = I$ (the identity matrix), set

$$z = \min_{0 \leqslant \ell \leqslant N} \frac{\big(A^{m-1} w\big)(\ell)}{(A^m w)(\ell)}, \tag{16}$$

According to the explanation before Lemma 23, we assume $z < \lambda_0$. Set $w_Q = (-Q - zI)^{-1}v$, which is positive by Lemma 22. Define

$$z_1 = \min_{0 \leqslant \ell \leqslant N} \frac{\left(A^{m-1}w_Q\right)(\ell)}{\left(A^m w_Q\right)(\ell)}, \tag{17}$$

and assume $z_1 < \lambda_0$ again. Then we have Proposition 25 below, which is a generalization of Lemma 23.

Proposition 25 For z and z_1 defined by (16) and (17), respectively, we have

$$0 < z \leqslant z_1 < \lambda_0.$$

Proof. Applying power iteration to the positive eigenvector w_Q, by Lemma 20, it follows that

$$z_1 \geqslant \min_{0 \leqslant \ell \leqslant N} \frac{w_Q(\ell)}{\left(Aw_Q\right)(\ell)}.$$

According to Lemma 23, we have

$$\min_{0 \leqslant \ell \leqslant N} \frac{w_Q(\ell)}{\left(Aw_Q\right)(\ell)} \geqslant \min_{0 \leqslant \ell \leqslant N} \frac{v(\ell)}{(Av)(\ell)} = \min_{0 \leqslant \ell \leqslant N} \frac{\left(A^m w\right)(\ell)}{A\left(A^m w\right)(\ell)}.$$

Thus,

$$z_1 \geqslant \min_{0 \leqslant \ell \leqslant N} \frac{\left(A^m w\right)(\ell)}{A\left(A^m w\right)(\ell)}. \tag{18}$$

Again, by Lemma 20,

$$\min_{0 \leqslant \ell \leqslant N} \frac{\left(A^m w\right)(\ell)}{A\left(A^m w\right)(\ell)} \geqslant \min_{0 \leqslant \ell \leqslant N} \frac{\left(A^{m-1}w\right)(\ell)}{\left(A^m w\right)(\ell)} = z.$$

Combining this with (18) and Perron-Frobenius theorem, we obtain the final conclusion. □

From the proof above, we know that Lemmas 20 and 23 play a crucial role. By setting $m = 1$ in Proposition 25, the monotonicity of $\{z^{(n)}\}$ in Algorithm 16 is proved. Combining the proposition with Remark 24, the monotonicity of $\{z^{(n)}\}$ in Algorithm 8 is also obtained. Besides, the monotonicity of $\{\widetilde{z}^{(n)}\}$ in the Algorithms 2, 8 and 16 can also be obtained in a similar way.

Explicit formulas of inverse matrices

In what follows, we show that the inverse of some sparse Q-matrices, such as single birth and single death, can be written explicitly.

Single birth processes with ND boundary Let

$$E = \{k \in \mathbb{Z} : 0 \leqslant k < N + 1\}, \qquad N \leqslant \infty.$$

The matrix $Q=(q_{ij})_{i,j\in E}$ is called single-birth Q-matrix if

$$\begin{cases} q_{i,i+1}>0, \quad q_i:=-q_{ii}>0, & i\in E,\\ q_{ij}\geqslant 0, \quad 0\leqslant j<i, & i\geqslant 1. \end{cases}$$

Now, suppose that the given Q-matrix $Q=(q_{ij})$ is irreducible (i.e., connected matrix), totally stable ($0<q_i<\infty$) and satisfy the ND boundary:

$$\begin{cases} q_0=q_{01},\\ q_\ell=\sum_{k=0}^{\ell-1}q_{\ell k}+q_{\ell,\ell+1}, & 1\leqslant \ell<N,\\ q_N=\sum_{k=0}^{N-1}q_{Nk}+c_N & c_N>0, N<\infty,\\ \sum_{j=0}^{N}F_j^{(0)}<\infty, & N=\infty. \end{cases}$$

Here, $F_j^{(0)}$ is defined by (19) below. In what follows, we make a convection that $q_{N,N+1}=c_N$ whenever $N<\infty$. Define two sequences as follows:

$$q_n^{(k)}=\sum_{j=0}^{k}q_{nj}, \quad 0\leqslant k<n<N+1,$$

$$F_k^{(k)}=1, \quad F_n^{(k)}=\frac{1}{q_{n,n+1}}\sum_{i=k}^{n-1}q_n^{(i)}F_i^{(k)}, \quad 0\leqslant k<n. \tag{19}$$

Proposition 26 Let Q be the irreducible and totally stable single-birth Q-matrix defined above. Then, with ND boundary, the inverse of $-Q$ can be represented as $H=(h_{k\ell})$ with

$$h_{k\ell}=\frac{1}{q_{\ell,\ell+1}}\sum_{j=\ell\vee k}^{N}F_j^{(\ell)}, \qquad 0\leqslant k,\ell<N+1.$$

Dually, one can give the inverse of the single death Q-matrices. The details are given as follows.

Single death processes with DN boundary Let

$$E=\{k\in\mathbb{Z}: 1\leqslant k<N+1\}, \qquad N\leqslant\infty.$$

The matrix $Q=(q_{ij})_{i,j\in E}$ is called single-death Q-matrix if

$$\begin{cases} q_{i,i-1}>0, \quad q_i:=-q_{ii}>0,\\ q_{ij}\geqslant 0, \quad i,j\in E,\ j>i\geqslant 0. \end{cases}$$

Now, suppose that the given Q-matrix $Q=(q_{ij})$ is irreducible (i.e., connected matrix), totally stable ($0<q_i<\infty$) and satisfy the DN boundary:

$$\begin{cases} q_1=\sum_{k=2}^{N}q_{1k}+c_1, & c_1>0,\\ q_\ell=\sum_{k=\ell+1}^{N}q_{\ell k}+q_{\ell,\ell-1}, & \ell\in E\backslash\{1\},\\ q_N=q_{N,N-1}, & N<\infty,\\ \mathbb{P}_k[\tau_0<\infty]<1, & k\in E. \end{cases}$$

Here, τ_0 is the hitting time of the process $\{X_t\}$ corresponding to the above matrix $Q = (q_{ij})$:

$$\tau_0 = \inf\{t > 0 : X_t = 0\}.$$

In what follows, we make a convection that $q_{10} = c_1$. Define two sequences as follows:

$$q_n^{(k)} = \sum_{j=k}^{N} q_{nj}, \quad 1 \leqslant n < k,$$

$$G_i^{(i)} = 1, \quad G_n^{(i)} = \frac{1}{q_{n,n-1}} \sum_{k=n+1}^{i} q_n^{(k)} G_k^{(i)}, \quad n, k, i \in E, 1 \leqslant n < i.$$

At the moment, we do not have a criterion for $\mathbb{P}_k[\tau_0 < \infty] = 1$, but there is a sufficient condition, due to [21; Theorem 5.1], as follows.

Assume that the above single death Q-matrix is regular and irreducible. Suppose in addition that

$$G_n := \varliminf_{m \to +\infty} \frac{G_n^{(m)}}{G_1^{(m)}} < \infty, \qquad n \geqslant 1$$

and

$$\sum_{n \geqslant 1} G_n = +\infty,$$

then

$$\mathbb{P}_k[\tau_0 < \infty] = 1, \qquad k \geqslant 1.$$

Proposition 27 Let Q be the irreducible and totally stable single-death Q-matrix defined above. Then, with DN boundary, the inverse of $-Q$ can be represented as $H = (h_{k\ell})$ with

$$h_{k\ell} = \frac{1}{q_{\ell,\ell-1}} \sum_{j=1}^{k \wedge \ell} G_j^{(\ell)}, \quad 1 \leqslant k, \ell < N + 1.$$

One can prove the above two propositions directly by checking

$$H(-Q) = (-Q)H = I,$$

where I is the identity matrix. Certainly, to find out these two formulas, more work is needed but we are not going to the details here.

7 Application to economic optimization

In this section, we first make a remark on two algorithms for L.K. Hua's economic optimization model. Then, based on the algorithms presented here, we specify an algorithm for the economic model (or two models more precisely). A theoretical analysis on the effectiveness of the algorithm is included, and then illustrated by examples.

A remark on the application to economic optimization

Remark 28 As mentioned several times before ([4; §2], [6; §1], [7; §1]), one of the motivations for the study on maximal eigenpair is L.K. Hua's economic optimization model (cf. [15],[16],[18] or [3; Chapter 10] and references therein for details). We now show that this paper can be considered as a complement to the important Hua's theory, especially on the computational aspect.

The key point of the model is using the maximal eigenpair (especially the maximal eigenvector) of an nonnegative matrix. Theoretically, if one knows either the maximal eigenvalue or its eigenvector, then one can compute the other. Certainly, it is easier to compute the eigenvalue first. For this, Hua (1984) introduced the following formula for computing the maximal eigenvalue:

$$\rho(A) = \lim_{\ell\to\infty} \left(\frac{\mathrm{Tr}(A^\ell)}{N}\right)^{1/\ell}, \qquad N := \text{Order of } A.$$

Of course, for a fixed A, it is the same as

$$\rho(A) = \lim_{\ell\to\infty} \big(\mathrm{Tr}(A^\ell)\big)^{1/\ell}.$$

In [18], Hua introduced a nice improved algorithm

$$\rho(A) = \lim_{k\to\infty} \big(\mathrm{Tr}(A^{2^k})\big)^{1/2^k},$$

and claimed that the number of the matrix products is $O(k)$. The point goes as follows. To get A^{2^k}, we need only k iterations:

$$A\times A = A^2, \quad A^2\times A^2 = A^{2^2}, \quad A^{2^2}\times A^{2^2} = A^{2^3}, \cdots$$

Actually, it is rather easy to see that the convergence speed is at least 2^{-k}, as shown by the next result which strengthens the above formula of $\rho(A)$.

Proposition 29 We have

$$\Big|\big(\mathrm{Tr}(A^m)\big)^{1/m} - \rho(A)\Big| \leqslant \frac{\text{Constant}}{m}, \qquad m\to\infty.$$

Proof. Let

$$\rho(A) =: \lambda_1 > |\lambda_2| \geqslant \cdots \geqslant |\lambda_N|$$

denote the eigenvalues of A (maybe repeated). If k is large enough, then

$$\big(\mathrm{Tr}(A^m)\big)^{1/m} = \left[\sum_{j=1}^N \lambda_j^m\right]^{1/m} = \lambda_1\left[1+\sum_{j=2}^N \left(\frac{\lambda_j}{\lambda_1}\right)^m\right]^{1/m}.$$

Noticing that

$$\sum_{j=2}^{N}\left(\frac{\lambda_j}{\lambda_1}\right)^m \leqslant (N-1)\left(\frac{|\lambda_2|}{\lambda_1}\right)^m,$$

there exists a large enough m_0 such that

$$x_0 := (N-1)\left(\frac{|\lambda_2|}{\lambda_1}\right)^{m_0} < 1.$$

Here for simplicity, we use x_0 to bound the exponentially decay term on the right-hand side above. Thus, we have

$$\lambda_1(1-x_0)^{1/m} \leqslant \big(\mathrm{Tr}(A^m)\big)^{1/m} \leqslant \lambda_1(1+x_0)^{1/m}, \qquad m \geqslant m_0.$$

Equivalently,

$$\lambda_1\big[(1-x_0)^{1/m}-1\big] \leqslant \big(\mathrm{Tr}(A^m)\big)^{1/m} - \lambda_1 \leqslant \lambda_1\big[(1+x_0)^{1/m}-1\big], \qquad m \geqslant m_0.$$

Because

$$\frac{(1\pm x_0)^\alpha - 1}{\pm\alpha x_0} \to \frac{\log(1\pm x_0)}{\pm x_0}, \quad \alpha \to 0,$$

we obtain the required assertion. □

Comparing with the direct approach just discussed, the algorithms introduced in the paper are rather involved. The reason is that here we are concentrated on the computation of the maximal eigenvector, the maximal eigenvalue is simply a by-product of the eigenvector. The opposite way may not work: with the eigenvalue at hand, it may not be practical for computing its eigenvector, as we have seen at the end of §3 (Example 12, which shows that some high-order linear equations cannot be solved). The speed of our algorithms are based on the shift inverse iteration, which are usually very fast. Anyhow, good estimates of the maximal eigenvalue are still very important, since they can be used as the shifts in the algorithms and accelerate the convergence speed of the algorithms. For this, we need to examine the computational complexity more carefully, not only the iteration number k but also the size N of the matrix A.

Because for vectors u and v with order N, the product $u\cdot v$ has N products (of numbers). Thus, for each matrix B with order N, the product $B\times B$ requires

$$N^2 \times N = N^3 \qquad \text{products.}$$

Therefore, the above algorithm requires

$$O(k\times N^3) \qquad \text{products.}$$

Clearly, once the size N is ignored, we return to what discussed before: we have a very good asymptotic behavior of the convergence to $\rho(A)$ as $k\to\infty$.

However, even though the number of the computation for the matrix products is $O(k)$, its elements of the products can be exponentially increasing or decreasing. Refer to the last part of Example 30 below, for instance. Roughly speaking, we are working on the direction of $k \ll N$ rather than $N \ll k$ discussed above.

Recall that our estimate on the maximal eigenvalue is mainly based on the Collatz–Wielandt formula, which is quite rough. Our improvement is due to the use of the power iteration. As discussed before, for a given vector u, the computation Au requires N^2 products (of numbers). Thus, in k iterations, we need

$$O(k \times N^2) \qquad \text{products.}$$

Here, we emphasize that the computational complexity for $(A^2)u$ is $O(N^3)$ but the one for $A(Au)$ is $O(N^2)$. This shows the serious difference between the analytic mathematics and the computational one. As we have seen from the examples presented in the paper, the number of iterations by our algorithms is usually no more than 5. Here, the leading order of the complexity is $O(N^2)$, but not $O(N^3)$. This is especially meaningful for the economic system since for which N is often rather large.

Finally, based on the Collatz–Wielandt formula, by Lemma 20, our $z^{(0)}$ chosen from a sequence produced by the power iteration is always located in the safe region $[\rho(A), \infty)$. However, the sequence

$$\xi_n := \left(\mathrm{Tr}(A^{2^n})\right)^{1/2^n}$$

is not necessary monotone and can even locate in the dangerous region: $[0, \rho(A))$ (see Example 30 below). Therefore, choosing the initial $z^{(0)}$ from the sequence $\{\xi_n\}$ is dangerous, for which our algorithms may fall into a trap.

Example 30 (Continued) Consider Example 6 again. When $N = 50$, $\rho(A) = 49.6592$ and

$$\{\xi_k\}_{k=1}^{10}: \quad 200.558,\ 87.525,\ 61.0271,\ 52.735,\ 50.0956,$$
$$\mathbf{49.5633},\ \mathbf{49.6321},\ 49.6597,\ 49.6592,\ 49.6592.$$

When $N = 150$, $\rho(A) = 149.662$ and

$$\{\xi_k\}_{k=1}^{10}: \quad 1054.92,\ 349.387,\ 212.247,\ 171.075,\ 156.532,$$
$$151.312,\ 149.735,\ \mathbf{149.564},\ \mathbf{149.649},\ 149.662.$$

Next, if we replace ξ_n by ξ'_n:

$$\xi'_n = \left(\frac{\mathrm{Tr}(A^{2^n})}{N}\right)^{1/2^n},$$

then for $N = 50$, we have

$$\{\xi'_k\}_{k=1}^{10}: \quad 28.3632,\ 32.9147,\ 37.4241,\ 41.2965,\ 44.3309,$$
$$46.6244,\ 48.1381,\ 48.9066,\ 49.2812,\ 49.4699.$$

Each of them is smaller than $\rho(A)$. Similarly, for $N = 150$, we have

$$\{\xi'_k\}_{k=1}^{10}: \quad 86.1335,\ 99.8353,\ 113.457,\ 125.078,\ 133.844,$$
$$139.918,\ 143.986,\ 146.665,\ 148.191,\ 148.931.$$

Thus, generally speaking, neither $\{\xi_k\}$ nor $\{\xi'_k\}$ can be used as the shifts in our algorithms. By the way, we mention that when $N = 150$, the numerical computation using Mathematica shows that

$$\text{the minimal element of } A^{2^{10}} = 8.96436939807476 \times 10^{2220},$$
$$\text{the maximal element of } A^{2^{10}} = 1.127062180214639 \times 10^{2227}.$$

In general, the computational softwares would refuse to handle with such a huge number. Recall that this model was treated in [5; Example 7] up to $N = 10^4$. Refer also to Example 17 above.

The detailed analysis given above shows that the nice formula for $\rho(A)$ in terms of the trace of A seems less practical in computations. Hence, in the next part of this section, we will return to the algorithms developed in the previous sections.

Application to Hua's economic optimization

It is the position to apply our new improved algorithms to Hua's economic optimization more carefully. Recall that $A = (a_{ij} : 1 \leqslant i, j \leqslant d)$ is a given nonnegative, irreducible matrix. To avoid the trivial case, we often assume that A^{-1} is not nonnegative. The optimization starts at a positive row vector x_0, and then there are two different ways to move on.

I. *The ideal model* (*without consumption*). At the nth step, let

$$x_n = x_0 A^{-n}, \qquad n \geqslant 1.$$

II. *The practical model* (*with consumption*). The consumption coefficient is denoted by $\alpha \in (0, 1)$. Set

$$B = (1 - \alpha)A^{-1} + \alpha I.$$

Then the economy goes as follows.

$$x_n = x_0 B^n, \qquad n \geqslant 1.$$

The main difference of the two models is that the first one has a faster increasing speed but is easier to collapse, that is, the collapsing time

$$T := \inf\big\{n : \text{there is a } j \text{ such that the component } x_n^{(j)} \leqslant 0\big\}$$

can be finite, even quite small. The second one has a slower increasing speed in the case of $\rho(A) < 1$ which is meaningful in economics, but it is more stable in the sense that the collapsing time becomes larger: the bigger α, the larger T. Refer to [15, 16] and [3; Chapter 10] for more details. Obviously, in the extremal case that $\alpha = 0$, Model II returns to Model I.

The lucky point of these two models is as follows. Look at the transforms:

$$A \to A^{-1} \to B;$$

the maximal left-eigenpair $(\rho(A),\ g)$ of A

$\to$ the minimal left-eigenpair $(\rho(A)^{-1},\ g)$ of A^{-1}

$\to$ the eigenpairs of B are the set

$$\{((1-\alpha)\lambda^{-1} + \alpha,\ g) : (\lambda, g) \text{ is an eigenpair of } A\},$$

with the assumption that A is invertible.

Since $|\lambda| \leqslant \rho(A)$ for each eigenvalue λ of A, it is clear that $(1-\alpha)\rho(A)^{-1} + \alpha$ is the minimal eigenvalue of B among the set

$$\{(1-\alpha)|\lambda|^{-1} + \alpha : \lambda \text{ is an eigenvalue of } A\}.$$

In general, for the eigenvalue λ of A, the comparison of $|(1-\alpha)\lambda^{-1} + \alpha|$ and $(1-\alpha)\rho(A)^{-1} + \alpha$ is not trivial, as we will see soon below. Thus, even though, the leading eigenvalues are transformed from one to the other, the corresponding eigenvectors remain the same. Therefore, we have the following algorithm.

Algorithm 31 There are two steps.

(1) Compute the maximal left-eigenpair using the algorithms given in Sections 3–5.
(2) Use the final stable vector output $y^{(\mathrm{iter})}$ in part (1) as the input x_0 of Models I and II.

The main new part of Algorithm 31 is Model II, comparing with increasing the consumption rate, it turns out that the precise level of part (1) is more essential which means a well-designed input is much more important for the economic stability. This is realistic: the economy model is sensitive to the precise level of initial input.

For the algorithm, some additional theoretical analysis on the stability is helpful. We begin with a detailed analysis on a condition to be used subsequently.

Lemma 32 Let $r > 0$ and $z \in \mathbb{C}$ satisfy $|z| < r$.

(1) Assume $\alpha \in (0, 1)$. Then the inequality

$$\left|\frac{1-\alpha}{z} + \alpha\right| < \frac{1-\alpha}{r} + \alpha, \tag{20}$$

does not hold for each $z \geqslant 0$.

(2) Otherwise, let $z \not\geqslant 0$, then (20) holds if and only if

$$0 < \Phi(z) := \frac{r^2 - |z|^2}{r^2(1 - 2\,\mathrm{Re}\,(z)) + |z|^2(2\,r - 1)} < \alpha. \tag{21}$$

Proof. Note that

$$\frac{1-\alpha}{z} + \alpha = \left[\frac{(1-\alpha)\mathrm{Re}(z)}{|z|^2} + \alpha\right] - i\,\frac{(1-\alpha)\mathrm{Im}(z)}{|z|^2}.$$

Thus,

$$(20) \Leftrightarrow \frac{(1-\alpha)^2}{|z|^2} + \alpha^2 + \frac{2\alpha(1-\alpha)\mathrm{Re}(z)}{|z|^2} < \frac{(1-\alpha)^2}{r^2} + \alpha^2 + \frac{2\alpha(1-\alpha)}{r}.$$

Collecting the terms in α, it follows that the above inequality at the right-hand side is equivalent to

$$\alpha\left[\frac{2r-1}{r^2} - \frac{2\mathrm{Re}(z)-1}{|z|^2}\right] > \frac{1}{|z|^2} - \frac{1}{r^2}\ (>0).$$

This means that the term $\left[\dfrac{2r-1}{r^2} - \dfrac{2\mathrm{Re}(z)-1}{|z|^2}\right]$ must be positive and so the above inequality is indeed equivalent to

$$\alpha > \left(\frac{1}{|z|^2} - \frac{1}{r^2}\right)\left[\frac{2r-1}{r^2} - \frac{2\mathrm{Re}(z)-1}{|z|^2}\right]^{-1} > 0.$$

We have thus proven the required assertion. □

We remark that the function $\Phi(z)$ defined in (21) is independent of α. When $\Phi(z) \in (0,1)$, we rewrite $\Phi(z)$ as $\alpha_c(z)$ to indicate a *critical point* of α. Then each $\alpha \in (\alpha_c(z), 1)$ satisfies (20). It is obvious that $\Phi(z) = \Phi(\bar{z})$ and hence $\alpha_c(z) = \alpha_c(\bar{z})$. This simplifies the computations below.

The first part of the next result is taken from [17; Theorem 2], the second part is then a consequence of Lemma 32.

Lemma 33 Let $r = \rho(A)$. Then we have the following results.

(1) The free-degree of x_0 (for Model II) equals

$m =$#{$\lambda_j : z = \lambda_j$ satisfies (20), repeated for the eigenvalue with multiplicity}.

(2) Alternatively,

$m =$ #{λ_j: either $z := \lambda_j < 0$ or z is not real, which satisfies $\Phi(z) \in (0,1)$, repeated for the eigenvalue with multiplicity}.

As usual, $\#\varnothing = 0$.[†]

[†]

The "free-degree" in Lemma 33 (1) means that the initial x_0 can be selected freely from an m-dimensional linear space (with some additional linearly independent vectors if necessary in the case that some eigenvalues have the smaller geometric multiplicity than the algebraic one), plus a term of the maximal left-eigenvector (up to a positive constant).

To understand Lemma 33 better, look at the following example.

Example 34 Let

$$A = \begin{pmatrix} 7 & 2 & 2 & 10 \\ 0 & 5 & 0 & 0 \\ 0 & 1 & 5 & 0 \\ 1 & 0 & 1 & 5 \end{pmatrix}.$$

Then A has eigenvalues

$$\{\lambda_j\}_{j=1}^4 : \ \left(6+\sqrt{11},\ 5,\ 5,\ 6-\sqrt{11}\right) \approx (9.31662,\ 5,\ 5,\ 2.68338)$$

with the corresponding left-eigenvectors

$$(0.431662, 0.3, 0.431662, 1), \quad (0, 1, 0, 0.),$$
$$(0, 0, 0, 0), \quad (-0.231662, 0.3, -0.231662, 1).$$

It follows that A has an eigenvalue 5 with algebraic dimension (or multiplicity) 2 but its geometric one is 1. Hence the eigenspace dimension is 3 but not 4. To obtain a basis of the 4-dimensional space, one may simply replace the above zero vector by a newly added one choosing from the kernel space $\mathrm{Ker}(A)$ of A. Denote by $\{g_k\}_{k=1}^4$ the resulting independent vectors. Then, every positive x_0 can be expressed as

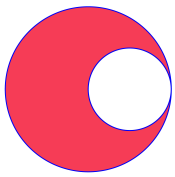
$$x_0 = \sum_{k=1}^4 \gamma_k g_k \qquad \text{with } \gamma_1 > 0.$$

Therefore

$$x_0 \left(\frac{A}{\lambda_1}\right)^n = \sum_{k\neq 3} \gamma_k \left(\frac{\lambda_k}{\lambda_1}\right)^n g_k \to \gamma_1 g_1, \qquad n \to \infty.$$

In other words, here we have free-degree 3 but not 2.

Here we introduce some additional remarks about Φ given in (21). First, it is helpful to rewrite Φ as follows: $\Phi(z) = \left[1+\frac{2r\phi(z)}{r^2-|z|^2}\right]^{-1} (|z|<r)$, where $\phi(z) = |z|^2 - r\mathrm{Re}z$. Then, with $r = \lambda_1$, it is obvious that $\Phi(z) < 1$ iff $\phi(z) > 0$ and hence $\Phi(z) \in (0,1)$ iff $\phi(z) > 0$. Thus, one can make a modification for the assertion: replacing "$\Phi(z) \in (0,1)$" by "$\phi(z) > 0$". This simplifies the computation of $\alpha_c(\lambda_\#)$: it suffices to compute $\Phi(\lambda_\#)$ for those $\lambda_\#$ with $\phi(\lambda_\#) > 0$.



The geometric explanation of the effective eigenvalues goes as follows (cf. [16]). The eigenvalues are located at the crescent shape produced by two circles, Circle((0,0), λ_1) and Circle(($\lambda_1/2, 0$)), $\lambda_1/2$)) (which comes from the equation $\phi(z) = 0$), where Circle((a, b), r) denotes the circle with center (a, b) and radius r. [Chen, 2021/01/16]

Certainly, the above example is simpler than Model II we are working on. However, once the eigenvalues have multiplicity, the conclusion should be parallel, on the free-degree counting the multiplicities of eigenvalues.

Lemma 33 shows that we have more freedom in choosing x_0 once $m \geqslant 1$. In which case, model II is more stable. Note that in computing the collapsing time, taking Model II for instance, it is reasonable to use the normalizing procedure

$$x_n = x_0\left(\frac{B}{(1-\alpha)\,\lambda_{\max}(A)^{-1}+\alpha}\right)^n, \qquad n \geqslant 1,$$

where $\lambda_{\max}(A)$ is the maximal eigenvalue of A. To have a concrete impression, let us look at Hua's original example [15, 16] with a little correction.

Example 35 Let

$$A = \frac{1}{100}\begin{pmatrix}25 & 14\\ 40 & 12\end{pmatrix}.$$

Then we have eigenvalues

$$\lambda_1 = \frac{1}{200}\big(37+\sqrt{2409}\big) \approx 0.430408, \quad \lambda_2 = \frac{1}{200}\big(37-\sqrt{2409}\big) \approx -0.0604078.$$

The maximal left-eigenvector is

$$\left(\frac{5}{7}\big(13+\sqrt{2409}\,\big),\ 20\right) \approx (44.3440,\ 20).$$

Starting at $x_0 = (44, 20)$, the different α and its corresponding collapsing time T are listed in Table 15.

Table 15. The parameter α and its T with $x_0 = (44, 20)$

α	0	0.4	0.6	0.8	0.8768
T	3	4	5	9	> 500

Note that by Lemma 32 (2), with $r = \lambda_1$, here $\alpha_c(\lambda_2) = 185/211 \approx 0.8768$. Hence, when $\alpha > \alpha_c(\lambda_2)$, the free-degree is 1 and so the corresponding T should be almost infinity.

Example 35 shows the influence of α for the stability of economics. The next example is quite interesting.

Example 36 Let

$$A = \begin{pmatrix}1 & 2 & 0 & 1 & 0 & 0\\ 2 & 2 & 2 & 2 & 1 & 0\\ 0 & 2 & 2 & 1 & 2 & 0\\ 1 & 2 & 1 & 1 & 2 & 1\\ 0 & 1 & 2 & 2 & 2 & 2\\ 0 & 0 & 0 & 1 & 2 & 1\end{pmatrix}.$$

Then we have eigenvalues

$$\begin{aligned}\{\lambda_j\}_{j=1}^6\colon\ & 3+2\sqrt{5}, \quad 3, \quad \frac{1}{2}\big(1+\sqrt{5}\big), \quad 3-2\sqrt{5}, \quad -1, \quad \frac{1}{2}\big(1-\sqrt{5}\big)\\ \approx\ & 7.47214, \quad 3, \quad 1.61803, \quad -1.47214, \quad -1, \quad -0.618034.\end{aligned}$$

The maximal left-eigenvector is

$$(1,\ 2.23607,\ 2,\ 2,\ 2.23607,\ 1).$$

Using Lemma 32 (1) with $r=\lambda_1$ and $z=\lambda_\#$, it follows that λ_2 and λ_3 do not satisfy (20), and then by Lemma 32 (2) we have

$$\alpha_c(\lambda_4)\approx 0.214286, \quad \alpha_c(\lambda_5)\approx 0.302205, \quad \alpha_c(\lambda_6)\approx 0.425981.$$

Therefore, we have the free-degree corresponding to the parameter α in Table 16.

Table 16. The free-degree corresponding to the parameter α

Interval of α	$(\alpha_c(\lambda_4),\ \alpha_c(\lambda_5)]$	$(\alpha_c(\lambda_5),\ \alpha_c(\lambda_6)]$	$(\alpha_c(\lambda_6),\ 1)$
Free-degree	1	2	3

With initials $x_0=(1,2,2,2,2,1)$ and $x_0=(1,2.23607,2,2,2.23607,1)$, respectively, some α and its corresponding collapsing time T are listed in Table 17.

Table 17. The parameter α and its T with different x_0

	$x_0=(1,2,2,2,2,1)$					$x_0=(1,2.23607,2,2,2.23607,1)$					
α	0	0.2	0.4	0.6	0.8	α	0	0.2	0.4	0.6	0.8
T	2	26	77	138	323	T	9	30	78	140	330

Clearly, the increase of α improves the stability rapidly for this example. Besides, the free-degree here is $3\,(<5)$. Note that for smaller α, the precise level of x_0 makes a serious effectiveness on the collapsing time for this model.

The next example has a bigger free-degree. Let

$$A_0=\begin{pmatrix}1&2&3&4&5&6&7&8\\9&10&11&12&13&14&15&16\\17&18&19&20&21&22&23&24\\25&26&27&28&29&30&31&32\\33&34&35&36&37&38&39&40\\41&42&43&44&45&46&47&48\\49&50&51&52&53&54&55&56\\57&58&59&60&61&62&63&64\end{pmatrix}. \quad \text{(sequential array)} \tag{22}$$

Since this matrix is not invertible, for our purpose, a modification is necessary.

Example 37 Let $u = \{1, 2, 2, 2, 2, 2, 2, 2\}$, and $A = A_0 - \mathrm{Diag}(u)$. Then, the eigenvalues of A are given as follows:

$$\{\lambda_j\}_{j=1}^{8}: \ 269.409, \ -11.7803, \ -4.34532, \ -2, \ -2, \ -2, \\ -1.14163 + 0.434068\,i, \ -1.14163 - 0.434068\,i.$$

The maximal left-eigenvector is

$$(0.814386,\ 0.874183,\ 0.865275,\ 0.920913,\ 0.919165,\ 0.94611,\ 0.973055,\ 1)$$

By Lemma 32 (2) with $r = \lambda_1$, we obtain

$$\alpha_c(\lambda_2) \approx 0.0390049, \ \alpha_c(\lambda_3) \approx 0.101697, \\ \alpha_c(\lambda_4) = \alpha_c(\lambda_5) = \alpha_c(\lambda_6) \approx 0.19881, \\ \alpha_c(\lambda_7) = \alpha_c(\lambda_8) \approx 0.303548.$$

Starting at a rather good initial

$$x_0 = (0.814386, 0.874183, 0.865275, 0.920913, 0.919165, 0.94611, 0.973055, 1),$$

some α and its corresponding collapsing time T are listed in Table 18.

Table 18. The parameter α and its T

α	0	0.1	0.2	0.3	0.4
T	3	9	18	469	> 500

Note that the left-eigenvector corresponding to λ_4, λ_5 and λ_6 are

$$(0,\ 2,\ 0,\ -3,\ 0,\ 0,\ 0,\ 1),\ (0,\ 1,\ 0,\ -2,\ 0,\ 1,\ 0,\ 0),\ (0,\ 1,\ 0,\ -3,\ 2,\ 0,\ 0,\ 0),$$

respectively. They are linearly independent. Hence the free-degree for this example is the whole dimension 7. From this, it should be obvious that the model goes to stable rapidly when α increases.

In general, we may be not so lucky to expect the free-degree to be $\geqslant 1$. The following corollary can be deduced from Lemma 32 (1) immediately since for which the eigenvalues are all nonnegative.

Corollary 38 Let $A = (-Q)^{-1}$, where Q satisfies the conditions of Lemma 21 and is moreover symmetrizable with respect to a positive measure. Then the free-degree equals 0.

Proof. By Lemma 21, A is finite and positive. Next, by the symmetrizable assumption, the spectrum of $-Q$ is positive. The assertion now follows from Lemma 32 (1). □

The next two examples are not symmetrizable, we again have zero free-degree. The first one below is the same as Example 6.

Example 39 (Continued) Let $A = (-Q)^{-1}$, where Q is defined in Example 6. Then $-Q$ has eigenvalues

$$8.237127,\ 0.452339,\ 7.031610 + 0.779594\,i,$$
$$7.031610 - 0.779594\,i,\ 4.885853 + 1.465494\,i,\ 4.885853 - 1.4654941\,i,$$
$$2.596732 + 1.251562\,i,\ 2.596732 - 1.251562\,i.$$

Their inverse give us the eigenvalues $\{\lambda_j\}_{j=1}^8$ of A. Except the complex ones $\{\lambda_j\}_{j=3}^8$, the others are positive, which do not satisfy (20) by Lemma 32 (1). Next, by (21), we have

$$\Phi(\lambda_3) = \Phi(\lambda_4) = 1.358659 > 1,$$
$$\Phi(\lambda_5) = \Phi(\lambda_6) = 1.523203 > 1,$$
$$\Phi(\lambda_7) = \Phi(\lambda_8) = 2.123888 > 1$$

which do not satisfy condition (21). From Lemma 33 (2), it follows that the free-degree $m = 0$.

Example 40 (Continued) Let $A = (-Q)^{-1}$, where Q is defined in Example 19 for $\alpha = 7/4,\ N = 8$. The eigenvalues of $-Q$ are all positive:

$$8.84941,\ 7.46811,\ 6.07934,\ 4.73839,\ 3.4828,\ 2.35115,\ 1.39264,\ 0.638153,$$

and so are of A. By Lemma 33 (2), we have once again the free-degree $m = 0$.

In view of the last two examples, as well as Corollary 38, it is clear that one cannot expect in general a positive free-degree. In this case, the improvement of the stability by increasing α goes quite slowly, as shown by Example 35. This is somehow reasonable since the algorithm used in these two models is essentially the power iteration which often converges very slowly. Thus, a good way to improve the algorithm is improving the initial x_0. We are going to illustrate this by the following example.

Example 41 Set $A = A_0 + I_8$, where I_8 is the identity matrix with order 8, A_0 is the matrix of (22). Then A has eigenvalues

$$\{\lambda_j\}_{j=1}^8 :\ 272.391,\ -9.17325,\ -1.17606,\ 1.95873,\ 1,\ 1,\ 1,\ 1.$$

The maximal left-eigenvector is

$$(16.2289,\ 17.4847,\ 17.3063,\ 18.419,\ 18.3838,\ 18.9225,\ 19.4613,\ 20).$$

We have

$$\alpha_c(\lambda_2) = 0.050035, \quad \alpha_c(\lambda_3) = 0.297413.$$

Denote by T_1, T_2, and T_3 the collapsing time with respect to the initial x_0 in three different cases

$x_0 = (16,\ 17,\ 17,\ 18,\ 18,\ 19,\ 19,\ 20)$,
$x_0 = (16.23,\ 17.48,\ 17.3,\ 18.41,\ 18.38,\ 18.92,\ 19.46,\ 20)$,
$x_0 = (16.2289,\ 17.4847,\ 17.3063,\ 18.419,\ 18.3838,\ 18.9225,\ 19.4613,\ 20)$,

respectively. Corresponding to different α, the outputs of $\{T_j\}_{j=1}^3$ are listed below.

Table 19. The parameter α and its $\{T_j\}_{j=1}^3$

α	0	0.2	0.4	0.6	0.8	0.9
T_1	1	2	3	5	11	22
T_2	2	5	8	14	32	68
T_3	3	7	13	22	50	105

Clearly, the precision level makes a serious influence on the collapsing time.

Let us mention that even though Lemmas 32 and 33 are crucial in theoretical analysis, they are not practical in application since one cannot compute the whole spectrum of a very large matrix. As pointed out in [16], to check the effectiveness of an initial x_0, one may simply use a direct optimal search. Actually, we may avoid this heavy job, by increasing the precise level of the output in part (1) of Algorithm 31, as just illustrated by Example 41.

Note that if we denote by $A_\#$ the matrix used in Example #, then we have

$$A_{41} - A_{37} = \mathrm{Diag}(\{2, 3, 3, 3, 3, 3, 3, 3\}).$$

Thus, Examples 37 and 41 are quite close to each other. However, as we have seen that the stability for them are very different. This shows that the economic models are very sensitive, and moreover, the improvement of the economic structure (i.e., the matrix A) is quite effective to improve the stability. Besides, the examples in this section justify again the importance of the global algorithms proposed in [5] and the improved ones in this paper.

Acknowledgments This work was supported in part by National Natural Science Foundation of China (Grant No. 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Butler B K , Siegel P H. Sharp bounds on the spectral radius of nonnegative matrices and digraphs. Linear Algebra Appl, 2013, 439(5): 1468–1478

[2] Chen, M.F. (2004). *From Markov Chains to Non-Equilibrium Particle Systems.* World Scientific, Singapore, 2nd Ed. (1st Ed., 1992).

[3] Chen M F. Eigenvalues, Inequalities, and Ergodic Theory. London: Springer, 2005

[4] Chen M F. Efficient initials for computing the maximal eigenpair, Front Math China, 2016, 11(6): 1379–1418.
See also Vol 4 in the middle of author's homepage:
http://math0.bnu.edu.cn/~chenmf

A package based on the paper is available on CRAN now (by X.J. Mao). One may check it through the link: https://github.com/mxjki/PowerfulMaxEigenpair
A Matlab package is also available, see the author's homepage above
The authors' papers cited in this article can be found from Vols 1–4 in the middle of the homepage above.

[5] Chen M F. Global algorithms for maximal eigenpair. Front Math China, 2017, 12(5): 1023–1043

[6] Chen M F. The charming leading eigenpair. Adv Math(China), 2017, 46(4): 281–297

[7] Chen M F. Trilogy on computing maximal eigenpair. In: Yue W, Li Q L, Jin S, Ma Z, eds. Queueing Theory and Network Applications (QTNA 2017). Lecture Notes in Comput Sci, Vol 10591. Cham: Springer, 2017, 312-329

[8] Chen M F. The optimal search problem — begins with the loss of Malaysia Airlines. Math Media, 2017, 41(3): 13–25(in Chinese)

[9] Chen M F. Hermitizable, isospectral complex matrices or differential operators. Front Math China, 2018, 13(6): 1267–1311

[10] Chen M F, Li Y S. Development of powerful algorithm for maximal eigenpair. Front Math China, 2019, 14(3):493-519

[11] Chen M F, Mao Y H. Introduction to Stochastic Processes. Beijing: Higher Education Press, 2007 (in Chinese); Singapore: World Sci, 2019 (English Ed).

[12] Dynkin E B. Markov Processes (Vol 1). New York: Springer, 1965.

[13] Hall C A, Porsching T A. Bounds for the maximal eigenvalue of a nonnegative irreducible matrix. Duke Math J, 1969, 36(1): 159–164

[14] Hou Z T, Guo Q F. Time-Homogeneous Markov Processes with Countable State Space. Beijing: Science Press, 1978 (in Chinese); Beijing/Berlin: Science Press/Springer, 1988 (English Ed)

[15] Hua L K. Mathematical theory of large-scale optimization of planned economy (I). Chin Sci Bull, 1984, 12: 705–709 (in Chinese)

[16] Hua L K. Mathematical theory of large-scale optimization of planned economy (X). Chi Sci Bull, 1985, 9: 641–645 (in Chinese)

[17] Hua L K, Hua S. Study on real square matrix with positive left- and right-eigenvectors simultaneous . Mathematics Bull, 1985, 8: 30–33 (in Chinese)

[18] Hua L K. Mathematical theory on economic optimal balance. In: Yang D Z, ed. Collected Works by L K Hua (Applied Math Ⅱ). Beijing: Sci. Press, 2010, 39-53 (in Chinese)

[19] Wang Z K. General Theory of Stochastic Processes, Vol I . Beijing: Beijing Normal Univ Press, 1996 (in Chinese)

[20] You L H, Shu Y J, Yuan P Z. Sharp upper and lower bounds for the spectral radius of a nonnegative irreducible matrix and its applications. Linear and Multilinear Algebra, 2017, 65(1): 113–128

[21] Zhang Y H. Criteria on ergodicity and strong ergodicity of single death processes. Front Math China, 2018, 13(5):1215-1243

Front. Math. China 2020, 15(2): 285–303
https://doi.org/10.1007/s11464-020-0832-2

On Spectrum of Hermitizable Tridiagonal Matrices

Mu-Fa Chen

(RIMS, Jiangsu Normal University, Xuzhou, 221116;
Sch. Math. & LMCS, Beijing Normal Univ., Beijing 100875)

January 1, 2020

Abstract

This paper is devoted to the study on the spectrum of Hermitizable tridiagonal matrices. As an illustration of the application of the author's recent results on Hermitizable matrices, an explicit criterion for discrete spectrum of the matrices is presented, with a slight and technical restriction. The problem is well-known, but from the author's knowledge, it has been largely opened for quite a long time. It is important in various application, in quantum mechanics for instance. The main tool to solve the problem is the isospectral technique developed a few years ago. Two alternative constructions of the isospectral operator are presented, they are helpful in theoretical analysis and in numerical computations, respectively. Some illustrated examples are included.

1 Introduction

Let

$$E = \{k \in \mathbb{Z}_+ : k < N + 1\}$$

with $N \leqslant \infty$. Throughout this paper, we consider the tridiagonal matrix, denoted by T or Q, having the following form

$$\begin{matrix} T \\ Q \end{matrix} = \begin{pmatrix} -c_0 & b_0 & & & & \\ a_1 & -c_1 & b_1 & & 0 & \\ & a_2 & -c_2 & b_2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & a_{N-1} & -c_{N-1} & b_{N-1} \\ & & & & a_N & -c_N \end{pmatrix}, \tag{1}$$

Received January 1, 2020; accepted April 14, 2020
2000 *Mathematics Subject Classifications.* 15A18, 15B57, 60J27, 81Q10.
Key words and phrases. Hermitizable, birth–death matrix, isospectral matrices, discrete spectrum.

where for T, the sequences $\{b_j\}_{j=0}^{N-1}$ and $\{a_j\}_{j=1}^{N}$ are complex and $\{c_j\}_{j=0}^{N}$ is real. The matrix Q denotes a particular T for which the sequences $\{b_j\}_{j=0}^{N-1}$ and $\{a_j\}_{j=1}^{N}$ are positive and moreover $c_j = a_j + b_j$ for each j ($a_0 = 0$ and $b_N = 0$ if $N < \infty$ by convention), except $c_N \geqslant a_N$ if $N < \infty$.

Here we allow $N = \infty$. In which case, the matrix takes a simpler form:

$$\underset{Q}{T} = \begin{pmatrix} -c_0 & b_0 & & & 0 \\ a_1 & -c_1 & b_1 & & \\ & a_2 & -c_2 & b_2 & \\ 0 & & \ddots & \ddots & \ddots \end{pmatrix}.$$

In what follows, we will not mention this point time by time. Since the matrix is determined by these three sequences $\{a_k\}$, $\{-c_k\}$ and $\{b_k\}$ only, we may simply write

$$T \text{ (or } Q) \sim (a_k, -c_k, b_k)$$

for simplicity. The matrix Q is called birth–death Q-matrix which corresponds an important and basic class of stochastic processes, the birth–death processes (refer to [2] for details).

Recall that a complex matrix $A = (a_{ij})$ is called Hermitizable [4]: if there exists a positive measure $\mu = (\mu_i : i \in E)$ such that

$$\mu_i a_{ij} = \mu_j \bar{a}_{ji} \qquad \forall i, j, \tag{2}$$

where $\bar{a}$ denotes the conjugate of a. For the tridiagonal case, the criterion for the Hermitizability is quite simple: $\{c_k\}$ is real, either $a_{k+1} = 0$ and $b_k = 0$ simultaneously, or $a_{k+1}b_k > 0$ for each k (cf. [4; Corollary 6]). In the first situation, the matrix can be divided into two submatrices and so they can be treated separately. Hence, in what follows, we will ignore this reducible situation. Since $a_{k+1}b_k > 0$ iff $a_{k+1} \neq 0$ and $b_k/\bar{a}_{k+1} > 0$, under this condition, we can define a positive measure μ:

$$\mu_0 = 1,\ \mu_k = \mu_{k-1}\frac{b_{k-1}}{\bar{a}_k}, \qquad k \geqslant 1. \tag{3}$$

Therefore, we have the complex L^2-space: $L^2(\mu)$. The Hermitizability defined by (2) is simply saying that as an operator, the matrix A (or T, in particular) is self-adjoint on $L^2(\mu)$, provided a domain $\mathscr{D}(A) \subset L^2(\mu)$ of A is specified.

Having the self-adjoint operator T at hand, by [4; Remark 16], we can consider the quadratic form

$$\begin{aligned} D_T(f,f) &:= -(Tf, f)_\mu \\ &= \sum_{i\in E} \mu_i \big[b_i\big(|f_i|^2 - \bar{f}_i f_{i+1}\big) + \bar{b}_i\big(|f_{i+1}|^2 - \bar{f}_{i+1} f_i\big)\big] \end{aligned}$$

$$+\sum_{i\in E}\mu_i\big(c_i-a_i-b_i\big)|f_i|^2. \tag{4}$$

For real T, the first sum on the right-hand side of (4) becomes

$$\sum_i \mu_i b_i|f_i-f_{i+1}|^2.$$

This is a standard first-order difference operator, the second sum of (4) comes from the potential function $V_i := c_i-a_i-b_i$. Hence this is indeed a Schrödinger operator once $V=(V_i)\neq 0$. It is known that in the study of spectral analysis, the latter is often much harder than the former if the potential V does not vanish out of the boundary. To overcome this difficulty, we introduce an isospectral birth–death Q-matrix $\widetilde{Q}\sim\big(\tilde{a},-\tilde{c},\tilde{b}\big)$ having the properties mentioned below (1), for which, the potential function V vanishes except at the boundary N when $N<\infty$. The quadratic form defined by (4) for Q defined by (1) is as follows.

$$D_Q(f,f):=-\big(Qf,f\big)_\mu=\sum_{i\in E}\mu_i b_i\big|f_{i+1}-f_i\big|^2+\mathbb{1}_{\{N<\infty\}}\mu_N\big(c_N-a_N\big)|f_N|^2. \tag{5}$$

Replacing Q by $\widetilde{Q}$, we obtain the quadratic form $D_{\widetilde{Q}}$ for $\widetilde{Q}$.

Before moving further, let us introduce the construction of $\widetilde{Q}$. Set

$$m=\sup_{k\in E}(|a_k|+|b_k|-c_k)^+. \tag{6}$$

For finite N, we certainly have $m<\infty$. Otherwise, it may be a restriction.

Definition 1 Assume that $m<\infty$. Set $u_k=a_kb_{k-1}\,(>0),\ 1\leqslant k\in E$. Define $\widetilde{Q}\sim\big(\tilde{a},-\tilde{c},\tilde{b}\big)$ as follows.

$$\begin{cases}\tilde{c}_k=m+c_k\,(k\in E)\\ \tilde{b}_0=\tilde{c}_0,\quad \tilde{b}_k=\tilde{c}_k-\dfrac{u_k}{\tilde{b}_{k-1}},\quad 1\leqslant k<N\\ \tilde{a}_k=\tilde{c}_k-\tilde{b}_k,\quad 1\leqslant k<N,\quad \tilde{a}_N=\dfrac{u_N}{\tilde{b}_{N-1}}\text{ if }N<\infty.\end{cases} \tag{7}$$

In the *simplest case* that

$$c_k=|b_k|+|a_k|,\qquad k\in E,$$

the construction becomes

$$\big(\tilde{b}_k,-\tilde{c}_k,\tilde{a}_k\big)=(|b_k|,\ -c_k,\ |a_k|),\qquad k\in E.$$

Clearly, for each n, $\tilde{b}_n$ depends only on $\{u_j\}_{j=1}^n$ and $\{\tilde{c}_j\}_{j=0}^n$ and so does $\tilde{a}_n$, except $\tilde{a}_0 = 0$. It is quite easy to check the last assertion in the definition above, as we will do so subsequently.

One may get some feeling about the construction of $\widetilde{Q}$ given in (7) from the proof of Theorem 2 in Section 3. Furthermore, in the second paragraph of the proof of Theorem 3, we will explain more on the sequence $\{h_k : k \in E\}$ below:

$$h_0 = 1, \quad h_k = h_{k-1}\frac{\tilde{b}_{k-1}}{b_{k-1}}, \qquad 1 \leqslant k < N+1. \tag{8}$$

For a given domain $\mathscr{D}\big(D_{\widetilde{Q}}\big)$, we adopt a domain for $\mathscr{D}\big(D_T\big)$ induced from $\mathscr{D}\big(D_{\widetilde{Q}}\big)$ by the function h:

$$\mathscr{D}\big(D_T\big) = \big\{f \in L^2(\mu) : h^{-1}f \in \mathscr{D}\big(D_{\widetilde{Q}}\big)\big\}. \tag{9}$$

We can now state a result taking from [4; Theorem 16] which is a key for the present study. Define a measure $\tilde{\mu}$ by (3) replacing Q by $\widetilde{Q}$. From now on, we often write $\mathrm{Diag}\,(u)$ for the diagonal matrix having the vector u as its diagonal elements.

Theorem 2 Let $m < \infty$. Then the operator $T - mI$ on $L^2(\mu)$ with domain $\mathscr{D}\big(D_T\big)$ is isospectral to $\widetilde{Q}$ on $L^2(\tilde{\mu})$ with domain $\mathscr{D}\big(D_{\widetilde{Q}}\big)$. Furthermore,

$$\widetilde{Q} = \mathrm{Diag}(h)^{-1}(T - mI)\,\mathrm{Diag}(h). \tag{10}$$

To study discrete spectrum, we need more notation. Since every compact operator has discrete spectrum, we need only consider the infinite case that $N = \infty$. Then for Q defined by (1), we have $c_k = a_k + b_k$ for every k, and so the quadratic form defined by (5) takes a simpler form:

$$D_Q(f) = \sum_{k\geqslant 0} \mu_k b_k (f_{k+1} - f_k)^2.$$

Define two domains of D_Q as follows.

$$\begin{aligned}
&\mathscr{D}_{\max}\big(D_Q\big) = \{f \in L^2(\mu) : D_Q(f) < \infty\},\\
&\mathscr{D}_{\min}\big(D_Q\big) = \text{smallest closure of } \{f \in L^2(\mu) \colon f \text{ has a finite support}\}\\
&\qquad\qquad \text{w.r.t. } \|\cdot\|_{D_Q}, \text{ where } \|f\|_{D_Q}^2 = \|f\|_{L^2(\mu)}^2 + D_Q(f).
\end{aligned}$$

For simplicity, in what follows, we write $\mathrm{Spec}\,(Q_{\max})$ to denote the spectrum of $\big(D_Q,\ \mathscr{D}_{\max}\big(D_Q\big)\big)$. Similarly, we have $\mathrm{Spec}\,(T_{\max})$, $\big(D_T,\ \mathscr{D}_{\max}\big(D_T\big)\big)$, $\mathrm{Spec}\,\big(\widetilde{Q}_{\max}\big)$ and $\big(D_{\widetilde{Q}},\ \mathscr{D}_{\max}\big(D_{\widetilde{Q}}\big)\big)$ and so on. Let us mention an alternative form of the measure $\tilde{\mu}$ defined above (will be checked in Proof of Theorem 3 given in Section 3):

$$\tilde{\mu}_0 = 1, \quad \tilde{\mu}_n = \tilde{\mu}_{n-1}\frac{\tilde{b}_{n-1}^2}{u_n} = \tilde{\mu}_{n-1}\frac{|b_{n-1}|}{|a_n|}\left(\frac{\tilde{b}_{n-1}}{|b_{n-1}|}\right)^2, \qquad n \geqslant 1. \tag{11}$$

Define another measure $\tilde{\nu}$ as follows.

$$\tilde{\nu}_0=\frac{1}{\tilde{c}_0},\quad \tilde{\nu}_n=\begin{cases}\tilde{\nu}_{n-1}\left(\dfrac{\tilde{c}_n}{\tilde{b}_n}-1\right), \text{ or alternatively}\\ \tilde{\nu}_{n-1}\dfrac{|a_n|}{|b_n|}\left[\dfrac{|b_{n-1}|}{\tilde{b}_{n-1}}\dfrac{|b_n|}{\tilde{b}_n}\right] \qquad n\geqslant 1.\end{cases} \tag{12}$$

Note that the sequences $\{\tilde{\mu}_n\}$ and $\{\tilde{\nu}_n\}$ can be expressed in terms of the sequences $\{u_n\}$ and $\{\tilde{c}_n\}$ only, since so does $\{\tilde{b}_n\}$. In the simplest case (Definition 1), they have the following form:

$$\tilde{\mu}_0=1,\quad \tilde{\mu}_n=\tilde{\mu}_{n-1}\frac{|b_{n-1}|}{|a_n|},\qquad n\geqslant 1;$$
$$\tilde{\nu}_0=\frac{1}{|b_0|},\quad \tilde{\nu}_n=\tilde{\nu}_{n-1}\frac{|a_n|}{|b_n|},\qquad n\geqslant 1.$$

Combining Theorem 2 with [3; Theorem 2.1] (see also [5]), we easily obtain the following criterion for the discrete spectrum of matrix T.

Theorem 3 Assume $m<\infty$. Write $\tilde{\mu}[0,n]=\sum_{j=0}^{n}\tilde{\mu}_j$ and similar for $\tilde{\nu}[0,n]$.

(1) Let $\tilde{\nu}[0,\infty)<\infty$. Then $\operatorname{Spec}(T_{\min})$ is discrete iff $\lim\limits_{n\to\infty}\tilde{\mu}[0,n]\,\tilde{\nu}[n,\infty)=0$.

(2) Let $\tilde{\mu}[0,\infty)<\infty$. Then $\operatorname{Spec}(T_{\max})$ is discrete iff $\lim\limits_{n\to\infty}\tilde{\nu}[0,n]\,\tilde{\mu}[n+1,\infty)=0$.

(3) Let $\tilde{\mu}[0,\infty)=\infty=\tilde{\nu}[0,\infty)$. Then $\operatorname{Spec}(T_{\min})=\operatorname{Spec}(T_{\max})$ is not discrete.

In particular, if $\sum_{k=0}^{\infty}\tilde{\nu}_k\,\tilde{\mu}[0,k]=\infty$, then the $\widetilde{Q}$-process is unique ([1; Corollary 3.18]), hence $\mathscr{D}_{\max}\big(D_{\widetilde{Q}}\big)=\mathscr{D}_{\min}\big(D_{\widetilde{Q}}\big)$ and furthermore $T_{\min}=T_{\max}$.

Theorem 3 is on a quantitative property rather than the qualitative one, usually the former is easier than the latter. Besides, the discrete spectrum depends only the boundary at infinity since on a finite space, each operator is compact and hence the property is automatic. Thus, a local modification of the operator does not make influence to the property. Even though, the infinite matrix makes some difficulty for the study, but only the asymptotic behavior is required. This makes the study easier, as shown in the next section of the paper.

To go further, we introduce two alternative algorithms of (7). The next result should be helpful for the asymptotic analysis of sequences $\{\tilde{b}_k\}$, $\{\tilde{c}_k\}$ and $\{\tilde{a}_k\}$, such as in the application of Theorem 3.

Lemma 4 We may re-express $\tilde{b}_n$ as

$$\tilde{b}_n=\tilde{c}_n-\frac{p_n}{q_n},\qquad 0\leqslant n<N,$$

where

$$\begin{pmatrix} p_0 \\ q_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 0 & |a_n| \\ -\dfrac{1}{|b_{n-1}|} & \dfrac{\tilde{c}_{n-1}}{|b_{n-1}|} \end{pmatrix} \begin{pmatrix} p_{n-1} \\ q_{n-1} \end{pmatrix}, \qquad 1 \leqslant n < N+1. \quad (13)$$

Alternatively, we have the explicit expression:

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 0 & |a_n| \\ -\dfrac{1}{|b_{n-1}|} & \dfrac{\tilde{c}_{n-1}}{|b_{n-1}|} \end{pmatrix} \begin{pmatrix} 0 & |a_{n-1}| \\ -\dfrac{1}{|b_{n-2}|} & \dfrac{\tilde{c}_{n-2}}{|b_{n-2}|} \end{pmatrix} \cdots \begin{pmatrix} 0 & |a_1| \\ -\dfrac{1}{|b_0|} & \dfrac{\tilde{c}_0}{|b_0|} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Note that the expressions for $\{p_n, q_n\}$ here depend only on the given three positive sequences $\{|a_n|\}$, $\{|b_n|\}$ and $\{\tilde{c}_n\}$. In the simplest case (see Definition 1), we have

$$p_n = |a_n|, \ \ q_n = 1, \qquad 1 \leqslant n < N+1.$$

Proof. By definition, we have

$$\frac{p_n}{q_n} = \frac{u_n}{\tilde{b}_{n-1}} = \frac{u_n}{\tilde{c}_{n-1} - u_{n-1}/\tilde{b}_{n-2}} = \frac{u_n}{\tilde{c}_{n-1} - p_{n-1}/q_{n-1}} = \frac{u_n q_{n-1}}{\tilde{c}_{n-1} q_{n-1} - p_{n-1}}.$$

Hence

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 0 & u_n \\ -1 & \tilde{c}_{n-1} \end{pmatrix} \begin{pmatrix} p_{n-1} \\ q_{n-1} \end{pmatrix}.$$

Since we are interested only in the ratio p_n/q_n, we may divide the last matrix by $|b_{n-1}|$, which then leads to the explicit solution stated in the lemma. □

The choice $|b_{n-1}|^{-1}$ in Lemma 4 avoids the extra factor $|b_{n-1}|$ in u_n and using the explicit known $|a_n|$, $|b_{n-1}|$ and $\tilde{c}_{n-1}$ only in the resulting matrix. This has an advantage in theoretical analysis as will be seen several times subsequently. Noting that such a factor can improve the blowing up problem only if $|b_{n-1}| > 1$ (see Example 8), it can be quite poor in numerical computation (see Example 11). Next, there is a problem of the cumulative errors in the iteration of (7), especially for large matrices. The next result is a little bit more complicated than the iteration (7) or Lemma 4, but it not only avoids the blowing up problem but also provides a possible way to decrease the cumulative errors. See the illustrated Example 11 in the next section for more information.

Lemma 5 Except the simplest case (see Definition 1), we may re-express $\tilde{b}_n$ as

$$\tilde{b}_n = \tilde{c}_n - \frac{p_n}{q_n}, \qquad 0 \leqslant n < N,$$

where

$$\begin{pmatrix} p_0 \\ q_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 0 & u_n v_n \\ -v_n & \tilde{c}_{n-1} v_n \end{pmatrix} \begin{pmatrix} p_{n-1} \\ q_{n-1} \end{pmatrix}, \qquad 1 \leqslant n < N+1,$$

and

$$v_n = \begin{cases} \dfrac{1}{2u_n}\Big[\tilde{c}_{n-1} - \sqrt{\tilde{c}_{n-1}^2 - 4u_n}\Big] & \text{if } \tilde{c}_{n-1}^2 \geqslant 4u_n; \\ \dfrac{1}{\sqrt{u_n}} & \text{if } \tilde{c}_{n-1}^2 < 4u_n, \end{cases} \qquad 1 \leqslant n < N+1.$$

In the first case that $\tilde{c}_{n-1}^2 \geqslant 4u_n$, there is a good numerical approximation of v_n:

$$v_n \approx \frac{2}{\tilde{c}_{n-1}}\big(0.615411 - 0.286195\, z^2 + 0.660784\, z^4\big), \quad z := \frac{4u_n}{\tilde{c}_{n-1}^2}.$$

Alternatively, for each n: $1 \leqslant n < N+1$, we have the explicit formula:

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 0 & u_n v_n \\ -v_n & \tilde{c}_{n-1} v_n \end{pmatrix} \begin{pmatrix} 0 & u_{n-1} v_{n-1} \\ -v_{n-1} & \tilde{c}_{n-2} v_{n-1} \end{pmatrix} \cdots \begin{pmatrix} 0 & u_1 v_1 \\ -v_1 & \tilde{c}_0 v_1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Again, the matrices here depend on the sequences $\{u_n\}$ and $\{\tilde{c}_n\}$ only.

In practice, one may simplify the matrices used in Lemma 5. For instance, in the case that $\tilde{c}_{n-1}^2 \geqslant 4u_n$, we have

$$u_n v_n = \frac{\tilde{c}_{n-1} - \sqrt{\tilde{c}_{n-1}^2 - 4u_n}}{2} = \frac{\tilde{c}_{n-1}\big(1 - \sqrt{1-z}\,\big)}{2},$$
$$\tilde{c}_{n-1} v_n = \frac{2\big(1-\sqrt{1-z}\big)}{z}, \qquad z := \frac{4u_n}{\tilde{c}_{n-1}^2}.$$

We will come back to this point in Remark 12, at the end of the paper.

In the next section, we will illustrate the applications of Theorems 2 and 3, as well as Lemmas 4 and 5 by some examples. The proofs of these results are delayed to Section 3.

2 Examples

A simple example to illustrate the use of Lemma 4 is the following one which was used several time before, see [4; §4] for example.

Example 6 Consider

$$T \sim (1, -3, 2)$$

Then, we have

$$\tilde{b}_n = \frac{2^{n+2} - 1}{2^{n+1} - 1}, \qquad n \geqslant 0.$$

Proof. By Lemma 4, we have

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}^n \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad n \geqslant 1.$$

Because

$$\begin{pmatrix} 0 & 1 \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}^n = \begin{pmatrix} -1+\frac{1}{2^{n-1}} & 2-\frac{1}{2^{n-1}} \\ -1+\frac{1}{2^n} & 2-\frac{1}{2^n} \end{pmatrix}, \qquad n \geqslant 1,$$

it follows that

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} 2-\frac{1}{2^{n-1}} \\ 2-\frac{1}{2^n} \end{pmatrix}, \qquad n \geqslant 1.$$

Therefore, applying the lemma again, we obtain

$$\tilde{b}_n = c_n - \frac{p_n}{q_n} = \frac{2^{n+2}-1}{2^{n+1}-1}, \qquad n \geqslant 0$$

as required. □

The next example is an extension of [3; Example 2.5].

Example 7 Let $\{\beta_n\}_{n=0}^{\infty}$ be a given arbitrarily real sequence,

$$b_n = n^4 e^{i\beta_n}, \ \ a_{n+1} = \big(n(n+1)\big)^2 e^{-i\beta_n}, \ \ c_n = |b_n| + |a_n|, \qquad n \geqslant 0,$$
$$a_0 = 0, \ \ a_1 = 1, \ \ b_0 = 1,$$

Then for the matrix $T \sim (a_k, -c_k, b_k)$, both $\mathrm{Spec}\,(T_{\min})$ and $\mathrm{Spec}\,(T_{\max})$ are discrete.

Proof. Note that in the present simplest case,

$$\tilde{c}_k = c_k = |a_k| + |b_k|.$$

By Theorem 2, T and $\widetilde{Q} \sim (|a_k|, -c_k, |b_k|)$ have the same spectrum. Then, the assertions follow by the first two parts of Theorem 3, as known from [3; Example 2.5]. □

Example 8 (Continued) Everything is the same as in Example 7, except the sequence $\{c_n\}$ is replaced by $\{\tilde{c}_n := c_n + n^2\}$.

Proof. Note that the sequence $\{u_n\}$ is the same as in the last example. Since $\tilde{c}_n > c_n$ for every n, by (7) and induction, we have $\tilde{b}_n > |b_n|$ every n. Thus, $\tilde{\nu}_n < \nu_n$ for every n by (12) plus induction, and then

$$\tilde{\nu}[0,\infty) < \nu[0,\infty) < \infty.$$

To estimate $\tilde{\mu}[0,\infty)$, it is necessary to estimate $\tilde{b}_n/b_n$. To do so, we study the ratio p_n/q_n. From the proof of Theorem 2 given in Section 3, we will see that the ratio is in general bounded from above by $|a_n|$. However, what we need now is the harder part, a lower bound of the ratio. For this, we employ first a numerical test using Lemma 5, from which, we guess that $p_n/q_n \geqslant |a_n|(1-1/n)$. For this example, we are in the second case "$\tilde{c}_{n-1}^2 < 4u_n$" in the lemma. Besides, we mention that when $n = 10^3$, the the outputs of (p_n, q_n) are approximately $(10^{14}, 10^2)$ and $(10^{11}, 10^{-1})$ by Lemmas 4 and 5, respectively. The renormalizing factor used in these two lemmas are $|b_{n-1}|^{-1} = (n-1)^{-4}$ and $u_n^{-1/2} = n^{-1}(n-1)^{-3}$, respectively. The former is a little bigger than the latter. This result comes with no surprising since Lemma 5 can avoid the blowing up trouble but may not Lemma 4. We are now going to prove the conjectured lower bound. First, we have

$$\left\{\frac{p_n}{|a_n|q_n}\right\}_{n=1}^{7} = \{1,\ 0.5,\ 0.727273,\ 0.811475,\ 0.855765,\ 0.88315,\ 0.901775\}.$$

Clearly, our conjecture holds at $n = 3, 4, \cdots, 7$. Suppose it already holds at $n-1$ $(n \geqslant 4)$:

$$p_{n-1} \geqslant \left(1 - \frac{1}{n-1}\right)|a_{n-1}|\, q_{n-1} = (n-2)^3(n-1)\, q_{n-1}.$$

By (13), we have

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} |a_n|\, q_{n-1} \\ \dfrac{\tilde{c}_{n-1}}{|b_{n-1}|}\, q_{n-1} - \dfrac{p_{n-1}}{|b_{n-1}|} \end{pmatrix}.$$

Hence

$$q_n \leqslant \left(1 + \frac{|a_{n-1}| + (n-1)^2}{|b_{n-1}|} - \frac{(n-2)^3}{(n-1)^3}\right) q_{n-1} = \frac{n^3 - 2n^2 + 2}{(n-1)^3} q_{n-1}.$$

Furthermore,

$$\frac{p_n}{|a_n|q_n} \geqslant \frac{(n-1)^3}{n^3 - 2n^2 + 2} = 1 - \frac{n^2 - 3n + 3}{n^3 - 2n^2 + 2} = 1 - \frac{1}{n}\frac{n^2 - 3n + 3}{n^2 - 2n + 2/n}.$$

The right-hand side is clearly $\geqslant 1 - 1/n$ once $n \geqslant 3$.

Combining the above estimate with Lemma 4, we obtain

$$(|b_n| \leqslant)\, \tilde{b}_n \leqslant \tilde{c}_n - |a_n|\left(1 - \frac{1}{n}\right) = |b_n| + n^2 + \frac{|a_n|}{n} = |b_n| + n^2 + n(n-1)^2.$$

This means that

$$\frac{\tilde{b}_n}{|b_n|} = 1 + O\left(\frac{1}{n}\right).$$

Then by (11), we get

$$\frac{\tilde{\mu}_n}{\tilde{\mu}_{n-1}} = \frac{|b_{n-1}|^2}{u_n}\frac{\tilde{b}_{n-1}^2}{|b_{n-1}|^2} = \frac{1}{n^2}\left[1 + O\left(\frac{1}{n}\right)\right]^2 = \frac{1}{n^2}\left[1 + O\left(\frac{1}{n}\right)\right].$$

We arrived at $\tilde{\mu}[0,\infty) < \infty$. Again, the required assertion follows by the first two parts of Theorem 3. □

In view of the proof above, it seems that much more perturbation of $(\tilde{c}_n)$ is allowed, not only n^2, to keep the same conclusion.

The next example is an extension of [3; Example 2.6].

Example 9 Let $\gamma \geqslant 0$,

$$b_n = n^\gamma e^{i\beta_n}, \quad a_{n+1} = n^\gamma e^{-i\beta_n}, \quad c_n = |b_n| + |a_n|, \qquad n \geqslant 0,$$
$$a_0 = 0, \quad a_1 = 1, \quad b_0 = 1,$$

where $\{\beta_n\}_{n=0}^\infty$ is again a given arbitrarily real sequence. Then for the matrix $T \sim (a_k, -c_k, b_k)$, $\mathrm{Spec}\,(T_{\min})$ is discrete iff $\gamma > 2$. In particular, if $\gamma \in [0,1]$, then $\mathrm{Spec}\,(T_{\min}) = \mathrm{Spec}\,(T_{\max})$ is not discrete.

Proof. As noted in Example 7, in the present simple case, $m = 0$, and so

$$\tilde{c}_k = c_k = |a_k| + |b_k|, \qquad k \geqslant 0.$$

By Theorem 2, T and $\widetilde{Q} \sim (|a_k|, -c_k, |b_k|)$ have the same spectrum. Hence the conclusion follows from [3; Example 2.6]. □

Example 10 (Continued) Consider the special case of Example 9 with specific $\gamma = 4$, and replacing $\{c_n\}$ by $\{\tilde{c}_n := c_n + n^2\}$. Then $\mathrm{Spec}\,(T_{\min})$ is discrete.

Proof. The proof here is quite closed to the one of Example 8. Because of $\tilde{c}_k < 2\sqrt{u_k}$, $k \geqslant 3$, we are in the second case in using Lemma 5. Next, since $|b_{n-1}| = \sqrt{u_n}$, the iteration of Lemma 5 coincides with Lemma 4. By a test of numerical computation using Lemma 4, we guess again

$$\frac{p_n}{|a_n|\, q_n} \geqslant 1 - \frac{1}{n}.$$

Note that

$$\left\{\frac{p_n}{|a_n|\, q_n}\right\}_{n=1}^{7} = (1,\ 0.5,\ 0.780488,\ 0.866197,\ 0.905112,\ 0.926899,\ 0.940706).$$

Our guest is true at $n = 3$. Assume that the guest is also true at $n-1\,(n \geqslant 4)$. Then we have

$$p_{n-1} \geqslant \left(1 - \frac{1}{n-1}\right)|a_{n-1}|\, q_{n-1} = \frac{(n-2)^5}{n-1}\, q_{n-1}.$$

By Lemma 4, it follows that

$$\frac{p_n}{|a_n|\, q_n} \geqslant \left[1 + \left(\frac{n-2}{n-1}\right)^4 - \left(\frac{n-2}{n-1}\right)^5 + \frac{1}{(n-1)^2}\right]^{-1}.$$

Thus, it suffices to show that

$$\left[1 + \left(\frac{n-2}{n-1}\right)^4 - \left(\frac{n-2}{n-1}\right)^5 + \frac{1}{(n-1)^2}\right]^{-1} \geqslant 1 - \frac{1}{n} = \frac{n-1}{n},$$

or equivalently,

$$1 > \left(\frac{n-2}{n-1}\right)^4 + \frac{1}{n-1}, \qquad n \geqslant 3.$$

By Lemma 4, we have thus arrived at

$$(|b_n| \leqslant)\, \tilde{b}_n = \tilde{c}_n - \frac{p_n}{q_n} = |b_n| + |a_n|\left[1 - \frac{p_n}{|a_n| q_n}\right] + n^2 \leqslant |b_n| + \frac{1}{n}|a_n| + n^2, \quad n \geqslant 3.$$

Then

$$\frac{\tilde{b}_n}{|b_n|} = 1 + \frac{1}{n}\frac{|a_n|}{|b_n|} + \frac{n^2}{|b_n|} = 1 + \frac{(n-1)^4}{n^5} + \frac{1}{n^2} = 1 + O\left(\frac{1}{n}\right).$$

From here, as an application of Theorem 3, it is rather easy to prove the required assertion. □

The next example exhibits both p_n and q_n in Lemma 4 can blow up very fast. It also compares the algorithms given in (7) with Lemmas 4 and 5 in numerical computation.

Example 11 (Continued) Consider the special case of Example 9 with $\gamma = 2$, and replacing $\{c_n\}$ by $\{\tilde{c}_n := c_n + n^2\}$. When $n = 1000$, the output is

$$(p_n,\ q_n) = \begin{cases} (3.383125991265680 \times 10^{421},\ 8.869971727917625 \times 10^{415}) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Lemma 4)}, \\ (3.26965 \times 10^{6},\ 8.57245) \qquad\qquad\qquad\qquad\qquad\qquad \text{(by Lemma 5)}. \end{cases}$$

Then, we have the ratio

$$\frac{p_n}{q_n} = \begin{cases} 381413.3905993774 \qquad\qquad \text{(by Lemma 4)}, \\ 381414 \qquad\qquad\qquad\qquad\quad \text{(by Lemma 5)}. \end{cases}$$

Hence, we have

$$\tilde{b}_n = \begin{cases} 2.6166 \times 10^6 & \text{(by Lemma 4)}, \\ 2.61659 \times 10^6 & \text{(by Lemma 5 or (7))}. \end{cases}$$

By using the algorithms given by (7) or Lemma 5, for $n = 10^4$, the output is

$$\tilde{b}_n = 2.61789 \times 10^8.$$

Proof. When applying Lemma 5 to this example, for $n = 1$, we are in the second case of the lemma. For $n \geqslant 2$, we are in the first one. Note that we use Mathematica v.11.3 in the computation, which has an automatical control for the precision level and so the different algorithms produce very close outputs. Because of this, the cumulative error may be avoided. Since there is a limitation on the number of the iterations by the software, in the last step, when $N = 10^4$, we actually separate the computation into ten parts, at each of them, we adopt 10^3 iterations only. □

The advantage in theoretical analysis and the shortcoming in numerical computation of Lemma 4 should be clear now. See also the alternative proof of Theorem 2 in Section 3 for an illustration of its advantage. We are now at the position to analyse the algorithms given in (7) and Lemma 5 more carefully.

First, we analyse their computational complexity. In the computation of $\{\tilde{b}_k\}_{k=0}^{N-1}$, having the known $\{u_k\}$ and $\{\tilde{c}_k\}$ at hand, at each iteration by (7), only one multiplication (division) is needed. Thus, for $\{\tilde{b}_k\}_{k=0}^{N-1}$, only N multiplications are required.

Next, at each iteration by Lemma 5, the work is done in three steps.

(a) First, we have three multiplications for the product

$$\begin{pmatrix} 0 & u_n v_n \\ -v_n & \tilde{c}_{n-1} v_n \end{pmatrix} \begin{pmatrix} p_{n-1} \\ q_{n-1} \end{pmatrix} = \begin{pmatrix} p_n \\ q_n \end{pmatrix}.$$

(b) Next, for the first case in Lemma 5, to compute v_n, three multiplications are required. Here we count the square-root as one multiplication, in view of the numerical approximation, it may cost 9 multiplications.

(c) Finally, to arrive at $\tilde{b}_n$, one more multiplication (division) is required. Therefore, for $\{\tilde{b}_k\}_{k=0}^{N-1}$, the algorithm of Lemma 5 needs $(7+)N$ multiplications.

In conclusion, the algorithm by (7) has a simpler computational complexity and so is faster than the one given by Lemma 5. However, they are at the same level of complexity: $O(N)$.

We now turn to compare the cumulative errors of the algorithms of (7) and Lemma 5. For the first algorithm, the problem is obvious, the error at step $n-1$ makes influence to the next step n immediately. Hence, we are worrying

about the possible cumulative errors, especially when deal with large matrices. Fortunately, such problem does not appear up to $n = 10^4$ in Example 11. It indicates that the algorithm is safe in the most cases. Let us now look at the errors produced by Lemma 5 at each step of an iteration. At the last step (c) to compute $\tilde{b}_n$ in terms of p_n/q_n, even though there would have an error as usual, but the computation here is independent of $\{\tilde{b}_k\}_{k\neq n}$, and so this step does not make cumulative errors, which is essentially different from the algorithm of (7). One may worry about step (b), which may make more errors. Actually, these errors can be simply ignored, even though they do make influence to the pair (p_n, q_n), but do not interfere the ratio p_n/q_n. Therefore, we need only to study the cumulative errors produced by the following iterations (from step (a))

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = H_n \begin{pmatrix} p_{n-1} \\ q_{n-1} \end{pmatrix},$$

where H_n is an explicitly given matrix having spectral radius $\rho(H_n) = 1$ (due to the use of the renormalization procedure in terms of v_n for avoiding blowing up, here a small perturbation is allowed), independent of $\begin{pmatrix} p_k \\ q_k \end{pmatrix}$ for $k \leqslant n-1$. It seems the cumulative errors made here could be less serious than what made by (7) since only simple products of matrices and vectors are used here. Sometimes, the errors may influence the pair (p_n, q_n), but not its ratio p_n/q_n.

Refer to [3] for more illustrated examples and for a partial history of the study on discrete spectrum. The author is regretted for being unable to find supplementary literature related closely to the complex context of this paper.

3 Proofs

Proof of Theorem 2 For simplicity, throughout this proof, we assume that $m = 0$ and so $\tilde{c}_k \equiv c_k$. Otherwise, simply replace c_k by $\tilde{c}_k$ everywhere in the proof. The proof consists of three parts.

(a) $\{\tilde{b}_k\}$ is positive,

(b) $\{\tilde{a}_k\}$ is positive and an invariant,

(c) T and $\widetilde{Q}$ are isospectral.

(a) *Prove that* $\{\tilde{b}_k\}$ *is positive.* First, we consider the simplest case mentioned in Definition 1. $c_k \equiv |a_k| + |b_k|$. Then the required assertion can be checked step by step as follows. Recall that

$$0 < u_k = a_k b_{k-1} = |a_k b_{k-1}|.$$

Then by definition of $\{\tilde{b}_k\}$, we have

$$\begin{aligned}
&\tilde{b}_0 = c_0 = |b_0| > 0 \text{ (since } a_0 = 0 \text{ by assumption)},\\
&\tilde{b}_1 = c_1 - \frac{u_1}{\tilde{b}_0} = c_1 - \frac{|a_1 b_0|}{|b_0|} = c_1 - |a_1| = |b_1| > 0,\\
&\tilde{b}_2 = c_2 - \frac{u_2}{\tilde{b}_1} = c_2 - \frac{|a_2 b_1|}{|b_1|} = c_2 - |a_2| = |b_2| > 0,\\
&\quad \cdots\cdots\\
&\tilde{b}_{N-1} = c_{N-1} - \frac{u_{N-1}}{\tilde{b}_{N-2}} = c_{N-1} - \frac{|a_{N-1} b_{N-2}|}{|b_{N-2}|} = |b_{N-1}| > 0 \qquad \text{if } N < \infty.
\end{aligned}$$

Hence the assertion holds in this special case. The proof in this part shows that even though the choice of $\{\tilde{b}_k\}$ is not unique, our choice is rather natural and economic.

Next, consider the general case that $c_k \geqslant |a_k| + |b_k|$. Again, we start our study at the simplest situation that $\bar{c}_0 > c_0$ but $\bar{c}_k = |a_k| + |b_k|$ for $k \geqslant 1$. For a moment, denote by $\{\tilde{b}_k\}$ the sequence used in the last paragraph and denote by $\{\bar{b}_k\}$ the sequence produced by the new triple $(a_k, -\bar{c}_k, b_k)$. Then

$$\bar{b}_0 = \bar{c}_0 > c_0 = \tilde{b}_0.$$

Furthermore, by induction, we obtain

$$\bar{b}_k > \tilde{b}_k > 0, \qquad k \geqslant 1.$$

In general, let

$$k_0 = \min\{k : \bar{c}_k > |a_k| + |b_k|\}.$$

Then we have $\bar{b}_k = \tilde{b}_k$ for $0 \leqslant k \leqslant k_0 - 1$ and by induction, $\bar{b}_k > \tilde{b}_k$ for $k \geqslant k_0$. In conclusion, the sequence $\{\tilde{b}_k\}$ in the theorem is increasing in (c_k).

(b) *Prove that $\tilde{a}_k$ is positive and an invariant.* By definition of $\{\tilde{a}_k\}$ and $\{\tilde{b}_k\}$, we have

$$\tilde{a}_k = c_k - \tilde{b}_k = c_k - \left(c_k - \frac{u_k}{\tilde{b}_{k-1}}\right) = \frac{u_k}{\tilde{b}_{k-1}}, \qquad 1 \leqslant k < N.$$

Therefore,

$$\tilde{a}_k = \frac{u_k}{\tilde{b}_{k-1}} > 0, \qquad 1 \leqslant k < N.$$

By definition of $\tilde{a}_N$, this assertion also holds at $k = N$ whenever $N < \infty$. We have thus proved not only the positivity of $\{\tilde{a}_k\}$, but also the nice invariant:

$$\tilde{a}_k \tilde{b}_{k-1} = u_k = a_k b_{k-1}, \qquad 1 \leqslant k < N + 1. \tag{14}$$

(c) *Prove that T and $\widetilde{Q}$ are isospectral.* Recall the sequence $\{h_k\}_{k=0}^{N}$ defined by (8), we have

$$\tilde{b}_k = b_k \frac{h_{k+1}}{h_k}. \tag{15}$$

Hence

$$a_{k+1}\frac{h_k}{h_{k+1}} \overset{(15)}{=} \frac{a_{k+1}b_k}{\tilde{b}_k} \overset{(14)}{=} \frac{\tilde{a}_{k+1}\tilde{b}_k}{\tilde{b}_k} = \tilde{a}_{k+1}. \tag{16}$$

Finally, it is easy to check that

$$(15) \text{ and } (16) \Longleftrightarrow \widetilde{Q} = \mathsf{Diag}(h)^{-1} T\, \mathsf{Diag}(h),$$

and so the required assertion follows. From this, we obtain the last assertion of the theorem.

To finish the proof, we mention a technical point. The function h defined by (8) is an isospectral mapping from $L^2(\mu)$ to $L^2(\tilde{\mu})$, it maps the measure μ to $|h|^2\mu$. We claim that the last measure coincides with $\tilde{\mu}$ defined by (3). This is somehow simple since the mapping keeps the Hermitizability and the Hermitizing measure for $\widetilde{Q}$ should be unique, up to a constant. Since $\{\mu_k\}$, $\{\tilde{\mu}_k\}$, and $\{h_k\}$ are all explicit (see (3) and (8)), a direct check for the identity $\tilde{\mu} = |h|^2\mu$ is also easy. To show this, it suffices to check that

$$\left(\frac{|h_j|^2\mu_j}{|h_{j-1}|^2\mu_{j-1}} =\right) \frac{\tilde{b}_{j-1}^2}{b_{j-1}\bar{b}_{j-1}} \cdot \frac{b_{j-1}}{\bar{a}_j} = \frac{\tilde{b}_{j-1}}{\tilde{a}_j}\left(= \frac{\tilde{\mu}_j}{\tilde{\mu}_{j-1}}\right), \qquad 1 \leqslant j < N+1.$$

Or equivalently, $\tilde{a}_j\tilde{b}_{j-1} = \bar{a}_j\bar{b}_{j-1}$. This is obvious by the invariant (14). □

Alternative proof of Theorem 2 Having Lemma 4 at hand, we can now introduce an alternative proof of Theorem 2. Actually, here, we prove only the positivity of $\{\tilde{b}_k\}$, the other parts of the proof remain the same as in the previous proof.

First, it is easy to check that $\tilde{c}_k \geqslant |a_k| + |b_k|$ for each k. Hence, we set

$$\frac{\tilde{c}_k}{|b_k|} = 1 + \frac{|a_k|}{|b_k|} + \beta_k \quad \text{for some } \beta_k \geqslant 0, \qquad 0 \leqslant k < N. \tag{17}$$

Then

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} 0 & |a_1| \\ -\dfrac{1}{|b_0|} & 1+\beta_0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} |a_1| \\ 1+\beta_0 \end{pmatrix} = \begin{pmatrix} |a_1|(1+\gamma_0) \\ 1+\gamma_1 \end{pmatrix},$$

$$\begin{pmatrix} p_2 \\ q_2 \end{pmatrix} = \begin{pmatrix} 0 & |a_2| \\ -\dfrac{1}{|b_1|} & 1+\dfrac{|a_1|}{|b_1|}+\beta_1 \end{pmatrix} \begin{pmatrix} |a_1|(1+\gamma_0) \\ 1+\gamma_1 \end{pmatrix} = \begin{pmatrix} |a_2|(1+\gamma_1) \\ 1+\gamma_2 \end{pmatrix},$$

where

$$\gamma_1 := \beta_0 \geqslant 0 =: \gamma_0, \tag{18}$$

$$\gamma_2 = (1+\beta_1)(1+\gamma_1) + \frac{|a_1|}{|b_1|}(\gamma_1 - \gamma_0) - 1 \geqslant \gamma_1 + \beta_1(1+\gamma_1) \geqslant \gamma_1. \tag{19}$$

From here, by induction, it should be easy to prove that

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} |a_n|(1+\gamma_{n-1}) \\ 1+\gamma_n \end{pmatrix} \qquad \text{for some } \gamma_n \geqslant \gamma_{n-1} \geqslant 0.$$

Hence

$$\frac{p_n}{q_n} = |a_n| \frac{1+\gamma_{n-1}}{1+\gamma_n} \in (0,\ |a_n|].$$

Therefore, by Lemma 4 and (17), we have

$$\tilde{b}_n = |a_n| + (1+\beta_n)|b_n| - \frac{p_n}{q_n} \geqslant (1+\beta_n)|b_n| \geqslant |b_n| > 0,$$

as required. □

From the proof above, it is also easy to figure out the following monotone property. For a new sequence $\{\tilde{c}'_n\}$ with $\tilde{c}'_n \geqslant \tilde{c}_n$ for every n, then the corresponding $p'_n/q'_n \leqslant p_n/q_n$ and furthermore $\tilde{b}'_n \geqslant \tilde{b}_n$ for every n.

Proof of Theorem 3 To prove the theorem, we need to check (11) and (12). The first one is easy:

$$\tilde{\mu}_n \overset{(3)}{=} \tilde{\mu}_{n-1} \frac{\tilde{b}_{n-1}}{\tilde{a}_n} \overset{(14)}{=} \tilde{\mu}_{n-1} \frac{\tilde{b}^2_{n-1}}{u_n} = \tilde{\mu}_{n-1} \frac{|b_{n-1}|}{|a_n|} \left(\frac{\tilde{b}_{n-1}}{|b_{n-1}|} \right)^2.$$

For (12), recall the usual definition (used in [3] in particular) is

$$\tilde{\nu}_n := \frac{1}{\tilde{\mu}_n \tilde{b}_n} \quad \text{and so} \quad \tilde{\nu}_0 = \frac{1}{\tilde{b}_0} = \frac{1}{\tilde{c}_0}.$$

Furthermore,

$$\tilde{\nu}_n \overset{(3)}{=} \tilde{\nu}_{n-1} \frac{\tilde{a}_n}{\tilde{b}_n} = \begin{cases} \tilde{\nu}_{n-1} \left(\dfrac{\tilde{c}_n}{\tilde{b}_n} - 1 \right) \text{(by (7))}, \text{ or alternatively} \\ \tilde{\nu}_{n-1} \dfrac{u_n}{\tilde{b}_{n-1} \tilde{b}_n} \overset{(14)}{=} \tilde{\nu}_{n-1} \dfrac{|a_n|}{|b_n|} \left[\dfrac{|b_{n-1}|}{\tilde{b}_{n-1}} \dfrac{|b_n|}{\tilde{b}_n} \right], \end{cases}$$

as required. As mentioned Section 1, the expressions of the sequences $\{\tilde{\mu}_k\}$ and $\{\tilde{\nu}_k\}$ can be expressed by the sequences $\{u_k\}$ and $\{\tilde{c}_k\}$ only, which come from the original matrix T. Unlike the coefficients $\{a_k\}$ and $\{b_k\}$ in T, which are complex, here $\{u_k\}$ and $\{\tilde{c}_k\}$ are positive. The expressions of the sequences $\{\tilde{\mu}_k\}$ and $\{\tilde{\nu}_k\}$ in terms of $\{u_k\}$ and $\{\tilde{c}_k\}$ are clearly quite complicated, this may be the main reason such a result has not been appeared before. However, the criterion depends only on the behavior of the sequences at infinity, and hence is applicable in practice, as illustrated in Section 2.

Before we really go into the proof of Theorem 3, let us explain the function h defined by (8) and the related Definition 1. Even though it looks very simple, but each $\tilde{b}_k$ is actually a continued fraction (cf. [4; Algorithm 14]). Hence, the

expression of h is indeed quite complicated. In fact, this is the main difficulty of the story. To explain roughly the trip of the story, let us denote temporarily by Q_0 be a real matrix on $E = \{k \in \mathbb{Z}_+ : k < N+1\}$ having the following two properties (for saving notation, in what follows, we simply write $[0, N)$ instead of $[0, N) \cap \mathbb{Z}_+$):

(i) the off-diagonal elements of $Q_0 \geqslant 0$, pointwise.
(ii) $Q_0\mathbb{1} \leqslant \mathbb{0}$, where $\mathbb{1}$ is the column vector having components 1 everywhere.

Next, let h be Q_0-harmonic: $Q_0 h = 0$ on $[0, N)$ (noting that if $N < \infty$, then the endpoint on the right is excluded). Then, by [5; Theorems 2.1 and 2.5], we can construct an isospectral new matrix Q_1 satisfying the above property (i) but replace property (ii) by

(ii)' $Q_1\mathbb{1} = \mathbb{0}$ except at N: $Q_1\mathbb{1}(N) \leqslant 0$ if $N < \infty$.

In other words, the potential function $V := Q_1\mathbb{1}$ vanishes except at N if $N < \infty$. For tridiagonal Q_0, a preliminary solution of the harmonic h was presented in [3; above Theorem 2.1]. Recall that for the complex tridiagonal T, the harmonic equation ($Th = 0$ on $[0, N)$) corresponds to a second-order differential equation with complex variable coefficients, and its general solution is unknown, even for the operator with real coefficients. Correspondingly, the general solution to equation $Th = 0$ is unknown. Therefore, what we need is to find a particular solution. As we all know, there is no common practical way to do this. Anyway, we are lucky to find the solution h, as shown in (8), which is a trick to be expressed in terms of the sequence $\{\tilde{b}_k\}$ given in (7). For the current general T, the solution was obtained until [4; §3], four years later than [3].

Once having h at hand, the current Theorem 3 is actually a copy of [3; Theorem 2.1] (by setting $h_k \equiv 1$). To be more precise, Theorem 3 is only a modification of [3; Theorem 2.1]: the Q-matrix used in the last theorem is replaced by $\widetilde{Q}$ used in the previous one. According to Theorem 2, this means that the conclusion given in Theorem 3 can be applied to the operator $\widetilde{Q}$, as well as to the operator $T - mI$. Because the shift mI does not interfere with the characteristics of discrete spectrum, hence $T - mI$ and T have or do not have discrete spectrum simultaneously. In view of this conclusion, the condition "$m < \infty$" seems to be technical and can be avoided by using a limiting procedure. So far, we have proved Theorem 3. □

To finish the proofs, we make a remark on an approximation of v_n.

Remark 12 Noting that in the case of $\tilde{c}_{n-1}^2 \geqslant 4u_n$, we have

$$v_n = \frac{1}{2u_n}\Big[\tilde{c}_{n-1} - \sqrt{\tilde{c}_{n-1}^2 - 4u_n}\Big] = \frac{2(1-\sqrt{1-z}\,)}{\tilde{c}_{n-1}z}, \qquad z := \frac{4u_n}{\tilde{c}_{n-1}^2}.$$

The idea is approximating the function

$$f(x) = 1 - \sqrt{1-x}$$

by

$$g(x) := 0.615411x - 0.286195\,x^3 + 0.660784\,x^5.$$

Figure 1 shows that these two functions f and g are quite close each other. More precisely, their difference (< 0.08515) is shown by Figure 2. Two curves are crossed at four points: $0,\ 1/2,\ 3/4$ and 1.

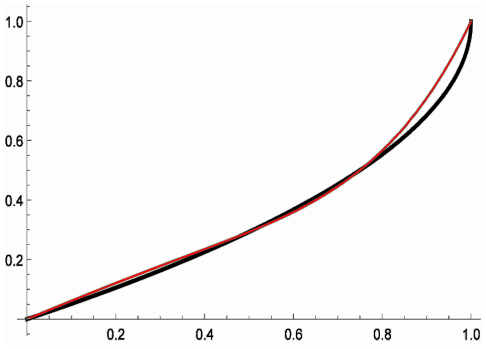


Figure 1: curves of f and g on $[0, 1]$.

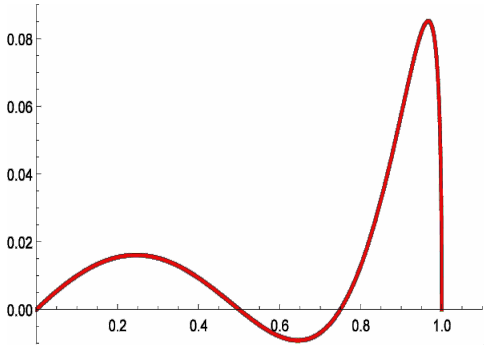


Figure 2: curve of $g - f$ on $[0, 1]$.

To conclude the paper, we mention that for finite matrix, one can replace the tridiagonal matrix T in Theorem 2 by a general Hermitizable one in terms of the Householder transformation (see [4; Theorem 24]). However, at the moment, the extension to the general infinite Hermitizable matrix is still unknown.

Acknowledgments The author would like to acknowledge to the anonymous referees for their careful reading and corrections. Thanks are also given to Professors Hong-Kui Pang and Zhi-Gang Jia for their fruitful discussions, and also for their tests on the comparison of the algorithms given by (7) and Lemma 5, applied to Examples 6 and 11 by using MatLab. Up to $N = 10^4$, the outputs show that these algorithms are very close to each other, as those given in Example 11. This work was supported in part by National Natural Science Foundation of China (Grant No. 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Chen, M.F. (2004). *From Markov Chains to Non-Equilibrium Particle Systems*. World Scientific, Singapore, 2nd Ed. (1st Ed., 1992).

[2] Chen, M.F. (2005) *Eigenvalues, Inequalities, and Ergodic Theory*. London: Springer.

[3] Chen, M.F. (2014). *Criteria for discrete spectrum of 1D operators*. Commu. Math. Stat. 2: 279–309.

[4] Chen, M.F. (2018). *Hermitizable, isospectral complex matrices or differential operators*. Front Math China 13(6): 1267–1311.

[5] Chen, M.F. and Zhang, X. (2014). *Isospectral operators*. Commu Math Stat 2, 17–32.

Mu-Fa Chen
Research Inst. Math. Sci., Jiangsu Normal Univ., Xuzhou, 221116.
School of Math. Sci., Beijing Normal Univ., Laboratory of Math. and Complex Systems (BNU), Ministry of Edu., Beijing 100875, PRC.
E-mail: mfchen@bnu.edu.cn
Home page: http://math0.bnu.edu.cn/~chenmf/main_eng.htm

Front. Math. China 2020, 15(5): 867–889

Hermitizable, isospectral complex second-order differential operators

Mu-Fa Chen[1,2,3] and Jin-Yu Li[2]

([1]RIMS, Jiangsu Normal University, Xuzhou, 221116;
[2]Sch. Math. & [3]LMCS, Beijing Normal Univ., Beijing 100875)

April 26, 2020

Abstract. The first aim of the paper is to study the Hermitizability of second-order differential operators, and then the corresponding isospectral operators. The explicit criteria for the Hermitizable or isospectral properties are presented. The second aim of the paper is to study a non-Hermitian model, which is now well known. In a regular sense, the model does not belong to the class of Hermitizable operators studied in this paper, but we will use the theory developed in the past years, to present an alternative and illustrated proof of the discreteness of its spectrum. The harmonic function plays a critical role in the study of spectrum. Two constructions of the function are presented. The required conclusion for the discrete spectrum is proved by some comparison technique.

Keywords. Hermitizable, isospectral, differential operators, non-Hermitian model, discrete spectrum.
MSC 2020 47B15, 47B91, 47B95.

1 Introduction

Denote by $\mathscr{C}^m(\mathbb{R}^d)$ the set of functions on $\mathbb{R}^d$ with continuous derivatives up to order m. Let $a = (a_{ij}(x))_{i,j=1}^d$ and $b = (b_i(x))_{i=1}^d$ be given complex matrix and vector on $\mathbb{R}^d$, respectively. Assume that $a_{ij} \in \mathscr{C}^1(\mathbb{R}^d, \mathbb{C})$ for each i, j. Next, let $V \in \mathscr{C}^1(\mathbb{R}^d, \mathbb{R})$ and $\mathrm{d}\mu = e^V \mathrm{d}x$. Thus, the first part of the paper is an extension of [5; §5] replacing the Lebesgue measure by μ. Consider the following complex second-order differential operator

$$L = \sum_{i,j} \partial_i (a_{ij} \partial_j) + \sum_i b_i \partial_i - c, \tag{1}$$

where $\partial_i = \mathrm{d}/\mathrm{d}x_i$ and $c \in L^2(\mu)$. We say that L with domain $\mathscr{D}(L)$ is Hermitizable with respect to the measure μ if L is a self-adjoint operator on the complex space $L^2(\mu)$ with inner product $(f, g)_\mu = \int f \bar{g} \mathrm{d}\mu$:

$$(Lf, g)_\mu = (f, Lg)_\mu, \qquad f, g \in \mathscr{D}(L) \subset L^2(\mu).$$

For vectors $F = \{f_k\}_{k=1}^m$ and $G = \{g_k\}_{k=1}^m$, set

$$\langle F,\ G\rangle_\mu = \sum_{k=1}^m (f_k,\ g_k)_\mu.$$

We have the following result for the Hermitizability. The main part of the result is given in [9]. An alternative proof of the result is delayed to Section 4 of the paper.

Theorem 1 The operator L is Hermitizable with respect to the measure μ iff a is Hermitian (i.e., $a^H := \bar{a}^* = a$) and

$$\operatorname{Re} b = (\operatorname{Re} a)(\partial V), \tag{2}$$

$$2\operatorname{Im} c = -\big((\partial V)^* + \partial^*\big)\big((\operatorname{Im} a)(\partial V) + \operatorname{Im} b\big), \tag{3}$$

where x^* denote the transpose of x. If so, its quadratic form is as follows.

$$(-Lf, g)_\mu = \langle a\partial f, \partial g\rangle_\mu + \sqrt{-1}\,\big\langle \big(\text{Im b} + (\text{Im a})(\partial V)\big)f,\ \partial g\big\rangle_\mu + (\bar{c}f,\ g)_\mu,$$
$$f, g \in \mathscr{D}(L) \subset L^2(\mu), \tag{4}$$

where $\operatorname{Im} c$ satisfies (3).

We now make a remark on the ordinary form of the second-order differential operator. Noting that

$$\partial_i(a_{ij}\partial_j f) = a_{ij}\partial_{ij}^2 f + (\partial_i a_{ij})\partial_j f,$$

we have

$$\sum_{i,j}\partial_i(a_{ij}\partial_j f) = \sum_{i,j} a_{ij}\partial_{ij}^2 f + \sum_i\Big(\sum_j \partial_j a_{ji}\Big)(\partial_i f). \tag{5}$$

Hence

$$\sum_{i,j} a_{ij}\partial_{ij}^2 f + \sum_i b_i(\partial_i f) = \sum_{i,j}\partial_i(a_{ij}\partial_j f) + \sum_i\Big(b_i - \sum_j \partial_j a_{ji}\Big)(\partial_i f).$$

Noting that as a product of the row vector ∂^* and matrix a, $\sum_j \partial_j a_{j\bullet} = \partial^* a$ is a row vector, we can write the drift coefficient on the right-hand side as a row vector:

$$b^* - \partial^* a =: \tilde{b}^*\ \big(\text{equivalently, in column } \tilde{b} = b - (\partial^* a)^*\big).$$

Then, as an application of Theorem 1: keeping (a_{ij}) to be the same but replace b by $\tilde{b}$ we obtain the following result.

Corollary 2 Under the assumption of Theorem 1, the operator

$$L = \sum_{i,j} a_{ij}\partial_{ij}^2 + \sum_i b_i\partial_i - c$$

is Hermitizable with respect to the measure μ iff a is Hermitian and

$$\operatorname{Re} b = (\operatorname{Re} a)(\partial V) + (\partial^*(\operatorname{Re} a))^*, \tag{6}$$

$$2\operatorname{Im} c = -\big((\partial V)^* + \partial^*\big)\big((\operatorname{Im} a)(\partial V) + \operatorname{Im} b - (\partial^*(\operatorname{Im} a))^*\big). \tag{7}$$

The main advantage of the Hermitizable operators is having the real spectrum. This is essential for quantum mechanics. However, the potential term c in the operators makes much difficulty for studying their spectrum. The goal of the next result is removing the potential term using a modified operator. The idea goes back to [7].

Theorem 3 Let

$$L = \nabla(a\nabla) + b \cdot \nabla - c$$

be the operator given by (1) with domain $\mathscr{D}(L)$ and let $h \neq 0$ be μ-a.e. harmonic of L: $Lh = 0$, μ-a.e. Then L is isospectral to the operator $\big(\widetilde{L},\ \mathscr{D}(\widetilde{L})\big)$:

$$\begin{aligned}\widetilde{L} &= \nabla(a\nabla) + \big(b + I_{[h\neq 0]}h^{-1}(a + a^*)\nabla h\big) \cdot \nabla,\\ \mathscr{D}(\widetilde{L}) &= \big\{\tilde{f} \in L^2(\tilde{\mu}) : \tilde{f}h \in \mathscr{D}(L)\big\}, \quad \tilde{\mu} := |h|^2\mu.\end{aligned}$$

In particular, in the Hermitizable case, $a + a^* = 2\operatorname{Re} a$.

In Section 4, Theorems 1 and 3 are extended to a more general class of operators including the so-called ferromagnetic potential.

We now go to an opposite direction: from $\widetilde{L}$ to L, as an analog of [7; Theorem 1.1 (2), Theorem 3.6 and Corollary 3.7]. It is very close to [5; Theorem 34]. In this way, we can construct a lot of models which have the same spectrum as those of the given one. The operator $\widetilde{L}$ used below has the same form as given in Theorem 3.

Theorem 4 Let $\widetilde{L} = \partial^*\tilde{a}\partial + \tilde{b}^*\partial$ with domain $\mathscr{D}\big(\widetilde{L}\big) \subset L^2(\tilde{\mu})$. Then for each complex function $h \in \mathscr{C}^2(\mathbb{R}^d)$, $h \neq 0$, $\tilde{\mu}$-a.e., $\widetilde{L}$ is L^2-isospectral to $L = L^h$:

$$\begin{aligned}L^h &= \widetilde{L} - \frac{1}{h}(\partial h)^*(\tilde{a} + \tilde{a}^*)\partial + \left[\frac{2}{h^2}(\partial h)^*\tilde{a} - \frac{1}{h}\big(\partial^*\tilde{a} + \tilde{b}^*\big)\right](\partial h)\\ &= \partial^*\tilde{a}\partial + \left[\tilde{b}^* - \frac{1}{h}(\partial h)^*(\tilde{a} + \tilde{a}^*)\right]\partial + \left[\frac{2}{h^2}(\partial h)^*\tilde{a} - \frac{1}{h}\big(\partial^*\tilde{a} + \tilde{b}^*\big)\right](\partial h),\\ \mathscr{D}(L^h) &= \big\{f \in L^2(\mu_h) : f/h \in \mathscr{D}\big(\widetilde{L}\big)\big\}, \quad \mu_h := |h|^{-2}\tilde{\mu}.\end{aligned}$$

Moreover, L and $\widetilde{L}$ are both selfadjoint or not, simultaneously.

Applying Theorem 4 to $\widetilde{L}$ with

$$\tilde{a}(x) = I, \quad \tilde{b}(x) = -x \implies \mathrm{d}\tilde{\mu}(x) = e^{-|x|^2/2}\mathrm{d}x,$$

where I is the $d \times d$ identity matrix, we obtain the following result.

Corollary 5 For each complex function $h \in \mathscr{C}^2(\mathbb{R}^d)$, $h \neq 0$, $\tilde{\mu}$-a.e., the operator

$$\begin{aligned}L^h &= \partial^*\partial - \left(x^* + \frac{2}{h}(\partial h)^*\right)\partial + \frac{1}{h}\left(x^* + \frac{2}{h}(\partial h)^*\right)(\partial h),\\ \mathscr{D}(L^h) &= \big\{f \in L^2(\mu_h) : f/h \in \mathscr{D}\big(\widetilde{L}\big)\big\}, \quad \mu_h := |h|^{-2}\tilde{\mu}\end{aligned}$$

is isospectral to the Ornstein-Uhlenbeck operator $\widetilde{L}$:

$$\begin{aligned}\widetilde{L} &= \partial^*\partial - x^*\partial,\\ \mathscr{D}\big(\widetilde{L}\big) &= \{f \in L^2(\tilde{\mu}) : |\nabla f| \in L^2(\tilde{\mu})\}.\end{aligned}$$

Hence L^h and $\widetilde{L}$ have the same discrete spectrum.

The proofs of the above results are delayed to Section 4. We are now going to study the discreteness of spectrum for a non-Hermitian model, which is not Hermitizable in the sense we have studied so far but it is so in a different sense. Hence, we have a chance to look at a slight different story.

The operator is

$$L = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} - \left(\sqrt{-1}\,x\right)^3$$

defined on the real line. For which, the eigenequation becomes

$$\left[-\frac{\mathrm{d}^2}{\mathrm{d}x^2} - \left(\sqrt{-1}\,x\right)^3\right] g_k(x) = \lambda_k g_k(x).$$

The serious problem is the complex Hamiltonian $\sqrt{-1}\,x^3$. By Theorem 1, corresponding to this Hamiltonian, the operator is not Hermitizable and hence it is not clear whether it has real spectrum or not, and also about the discreteness of its spectrum. Actually, this is the original model (cf. [2]) leading to the study on non-Hermitian quantum mechanics. There are now quite a number of publications in this field, see for instance [8, 10, 11, 1] and reference therein. Certainly, it is impossible for us to review the details of the study on the topic, what instead, is to show that its spectrum should be real and discrete.

As shown in [8; Appendix B], the first step is a simple change of the variable: $z = \sqrt{-1}\,x$. Then the eigenequation becomes

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}z^2} - z^3\right] g_k(z) = \lambda_k g_k(z).$$

Now, the variable z varies on the line $\sqrt{-1}\,\mathbb{R}$. Since we are interested in the real spectrum $\{\lambda_k\}$, one may regard the operator on the left-hand side as a real one. Thus, the original complex problem is reduced to the real one. The latter is a standard Schrödinger operator, which is clearly symmetrizable/Hermitizable by Theorem 1, and so has real spectrum. Thus, we only need to prove the discreteness of its spectrum, which is down in the next section, and more details are given in Section 3 below.

Now, let us rewrite our equation in the real context as follows.

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} - x^\gamma\right] g_k(x) = \lambda_k g_k(x) \qquad x \in \mathbb{R}. \tag{8}$$

According to [8; Appendix B], for the original complex model mentioned above, there is a restriction for γ: $\gamma \in (2, 4)$, and $\gamma = 3$ in particular. The problem was solved in the cited paper with some extension on the potential function. The exact solutions of the eigenpairs are constructed there in terms of some special functions including the Bessel ones. Certainly, such a nice solution can be worked out only for some special models. Here, we concentrate on the qualitative rather than quantitative aspect: the discreteness of the spectrum. For the specific model given in (8) with $c(x) = x^3$, fixing $\beta \in \mathbb{R}$, and replacing $c(x)$ with $c(x) - c(\beta) \geqslant 0$, then the modified operator on the interval (β, ∞) having discrete spectrum iff the following Molchanov's criterion (cf. [4; above Example 7.7]) holds:

$$\text{for each } \theta > 0, \qquad \int_x^{x+\theta} c \to \infty \quad \text{as } x \to \infty.$$

Here, we present an alternative proof to illustrate an application of our general result presented in [4].

2 Examples

This section is mainly devoted to prove that the spectrum of the real operator given in (8) is discrete, see Example 11.

To begin with, we state a general criterion that we already have in dimension one. For which, we need some notation. For a given elliptic differential (diffusion) operator

$$L = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x} - c(x), \qquad a(x) > 0,\ b(x) \in \mathbb{R},\ c(x) \geqslant 0$$

on $E := (0, \infty)$ or $\mathbb{R}$. Define three measures

$$\mu(\mathrm{d}x) = \frac{e^{C(x)}}{a(x)}\mathrm{d}x, \quad \nu(\mathrm{d}x) = e^{C(x)}\mathrm{d}x, \quad \hat{\nu}(\mathrm{d}x) = e^{-C(x)}\mathrm{d}x, \tag{9}$$

where $C(x) = \int_\theta^x (b/a)(y)\mathrm{d}y$ and θ is a reference point. As usual, write $\mu(f) = \int_E f\mathrm{d}\mu$. Our criterion is based on the μ-a.e. harmonic function h: $Lh = 0$, μ-a.e. having property $h \neq 0$, μ-a.e. For simplicity, we fix $E = [0, \infty)$. Then the other part $(-\infty, 0]$ may be handled in parallel (cf. [4; Theorem 7.13]). The next result is taken from [4; Part (1) of Theorem 7.1]. There are two more parts in the cited theorem but are omitted here for simplicity.

Lemma 6 Let $\hat{\nu}(E) < \infty$. Then the spectrum of L is discrete iff

$$\lim_{x\to\infty} \mu\big(h^2\mathbb{1}_{(0,x)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(x,\infty)}\big)\bigg[= \lim_{x\to\infty} \int_0^x h^2\mathrm{d}\mu \int_x^\infty \frac{1}{h^2}\mathrm{d}\hat{\nu}\bigg] = 0. \tag{10}$$

Let

$$G(x) = \begin{pmatrix} 0 & e^{-C} \\ c\,e^{C}/a & 0 \end{pmatrix}, \qquad F = \begin{pmatrix} f \\ e^{C} f' \end{pmatrix}.$$

Then the function h can be obtained by the following result. Note that $G \geqslant 0$ and $F(\theta)$ given below is nonnegative, the solution given in the next lemma is meaningful.

Lemma 7 Let F^* be the minimal nonnegative solution to the equation:

$$F = F(\theta) + \int_\theta^x GF, \qquad x \in E, \qquad F(\theta) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{11}$$

Then

$$F^* = \begin{pmatrix} h \\ e^{C} h' \end{pmatrix}.$$

In the present real context, the parameter γ is allowed to be bigger or equal to 2, especially, $\gamma > 2$. In the special situation we are working, $a = 1$, $b = 0$, and $c(x) = x^\gamma$. Hence μ, ν and $\hat{\nu}$ are simply the Lebesgue measure on the line. Having Lemma 7 at hand, the function h follows by the second successive approximation scheme of F^* given in [4; Theorem 7.4], plus a use of the induction. From which we obtain first the sequence $\big\{\widetilde{F}^{(k)}\big\}_{k=1}^\infty$, and then the sequence of the first components of $\big\{\widetilde{F}^{(k)}\big\}_{k=1}^\infty$:

$$\tilde{h}^{(2k)}(x) \equiv 0, \qquad \tilde{h}^{(1)}(x) \equiv 1, \quad \tilde{h}^{(2k-1)}(x) = \frac{x^{\alpha(k-1)}}{(k-1)!\,\alpha^{2(k-1)}(1-1/\alpha)_{k-1}},$$

where $\alpha = \gamma + 2$, $(z)_k = z(z+1)\cdots(z+k-1)$, $(z)_0 := 1$. We can write down the function h explicitly:

$$h(x) = 1 + \sum_{k=1}^{\infty} \frac{x^{\alpha k}}{k!\,\alpha^{2k}(1-1/\alpha)_k} = \sum_{k=0}^{\infty} \frac{\left(x^{\alpha}/\alpha^2\right)^k}{k!\,(1-1/\alpha)_k}. \tag{12}$$

Thus,

$$h(x) = {}_0F_1\left(1 - \frac{1}{\alpha},\ \frac{x^{\alpha}}{\alpha^2}\right) = {}_0F_1\left(\frac{\gamma+1}{\gamma+2},\ \frac{x^{\gamma+2}}{(\gamma+2)^2}\right)$$

where ${}_0F_1$ is the confluent hypergeometric function:

$${}_0F_1(\alpha, z) = \sum_{k=0}^{\infty} \frac{z^k}{(\alpha)_k k!}.$$

Since $(1-1/\alpha)_k \geqslant (1-1/\alpha)^k$, by(12), it follows that

$$h(x) \leqslant \sum_{k=0}^{\infty} \frac{x^{\alpha k}}{k!\,\alpha^{2k}(1-1/\alpha)^k} = \exp\frac{x^{\alpha}}{\alpha(\alpha-1)} = \exp\frac{x^{\gamma+2}}{(\gamma+2)(\gamma+1)} < \infty, \qquad x \in \mathbb{R}_+.$$

In the next section, we will introduce a simplified construction of the harmonic function h and a proof of (12).

Before moving further, let make a remark about Lemma 7. Let f be L-harmonic:

$$af'' + bf' = cf.$$

Equivalently,

$$f'' + \frac{b}{a}f' = \frac{c}{a}f. \tag{13}$$

Define

$$F = \begin{pmatrix} f \\ e^C f' \end{pmatrix},$$

then we can lift equation (13) as

$$F' = GF, \qquad F(\theta) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{14}$$

which is clearly equivalent to (11). We mention that the harmonic function constructed by solving the equation (13) can be different from those obtained by Lemma 7, in terms of some successive approximation schemes, since in the former case we do not require $f' \geqslant 0$. The essential condition we need is that $f \neq 0$, μ-a.e. (cf. [5; Lemma 8]), as mentioned before.

Applying (13) to $f = \exp\psi$, it reduces to solve the equation:

$$\psi'' + \psi'^2 + \frac{b}{a}\psi' = \frac{c}{a}. \tag{15}$$

Since $a = 1, b = 0$ here, if we reset $y = \psi'$, then we arrive at the Riccati equation

$$y'_x = -y^2 + c.$$

In the special case that $c(x) = x^{\gamma}$ for arbitrary real number $\gamma \neq -2$, the equation is solvable (refer to [12; 1.2.2-1, case 4]). Moreover, the solution can be expressed in two Bessel functions (infinite series). Clearly, such a solution, as well as the one given in (12) are not convenient to apply our criterion (Lemma 6). Therefore, we will look for some simpler solutions and illustrate the use of this idea.

Example 8 The operator

$$L = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - \left(x^\gamma + \frac{\gamma}{2(1+x)}x^{\gamma/2-1} + \frac{\gamma(\gamma+4)}{16(1+x)^2}\right), \qquad \gamma \geqslant 2$$

on $\mathbb{R}_+$ has discrete spectrum.

The main task in the proof is the construction of a required harmonic function h. To apply Lemma 6, we still need an elementary lemma.

Lemma 9 Let $\varphi \in \mathscr{C}^2(0,\infty)$.

(1) Assume that

$$\lim_{x\to\infty}\int_0^x e^\varphi = \infty = \lim_{x\to\infty} {\varphi'}^{-1}e^\varphi \quad \text{and} \quad \lim_{x\to\infty}\frac{\varphi''}{{\varphi'}^2} \neq 1.$$

Then

$$\int_0^x e^\varphi \sim {\varphi'}^{-1}e^\varphi \quad \text{as } x\to\infty.$$

(2) Assume that

$$\lim_{x\to\infty}\int_x^\infty e^\varphi = 0 = \lim_{x\to\infty} {\varphi'}^{-1}e^\varphi \quad \text{and} \quad \lim_{x\to\infty}\frac{\varphi''}{{\varphi'}^2} \neq 1.$$

Then

$$\int_x^\infty e^\varphi \sim {\varphi'}^{-1}e^\varphi \quad \text{as } x\to\infty.$$

Proof. It suffices to prove part (1) only. By assumption, we have

$$\frac{\int_0^x e^\varphi}{{\varphi'}^{-1}e^\varphi} \sim \frac{1}{1+(\varphi'^{-1})'} = \frac{1}{1-\varphi''/(\varphi')^2}. \quad \square$$

The next result is helpful for using (10).

Corollary 10 Under the assumptions of Lemma 9,

$$\lim_{x\to\infty}\int_0^x e^\varphi \int_x^\infty e^{-\varphi} = 0$$

iff $\lim_{x\to\infty}{\varphi'}^{-1} = 0$. In particular, for a fixed φ, the conclusion is independent of lower order perturbations.

Proof. Under the assumptions of Lemma 9, we have

$$\int_0^x e^\varphi \int_x^\infty e^{-\varphi} \sim {\varphi'}^{-1}e^\varphi {\varphi'}^{-1}e^{-\varphi} \sim {\varphi'}^{-2} \sim 0 \quad \text{as } x\to\infty. \quad \square$$

We are now ready to prove the conclusion of Example 8.

Proof of Example 8 First, we construct the required harmonic function. Let $\alpha \geqslant 2$ and $\beta > 0$ to be specified later. Define

$$\psi_2(x) = \frac{1}{\alpha}x^\alpha - \beta\log(1+x). \tag{16}$$

Then

$$\psi_2'(x) = x^{\alpha-1} - \frac{\beta}{1+x}, \qquad \psi_2''(x) = (\alpha-1)x^{\alpha-2} + \frac{\beta}{(1+x)^2}.$$

Hence

$$\begin{aligned}\psi_2'(x)^2 + \psi_2''(x) &= x^{2(\alpha-1)} + \left(\alpha - 1 - \frac{2\beta x}{1+x}\right)x^{\alpha-2} + \frac{\beta(\beta+1)}{(1+x)^2} \\ &= x^{2(\alpha-1)} + \frac{\alpha-1}{1+x}x^{\alpha-2} + \frac{\alpha^2-1}{4(1+x)^2} \quad \big(\text{set } \beta = (\alpha-1)/2\big) \\ &=: c_2(x).\end{aligned}$$

Combining this with (15), by setting $\alpha = \gamma/2 + 1$, it follows that the function $h_2 := \exp\psi_2$ is L-harmonic.

Setting $\varphi = -2\psi_2$ in part (2) of Lemma 9, it follows that

$$\hat{\nu}(E) = \int_E e^{-2\psi_2} \sim \int_E e^{-2\,x^\alpha/\alpha} < \infty \quad (\text{since } \alpha > 1).$$

Then (10) holds by Corollary 10 whenever $\alpha \geqslant 2$. □

Recall that we are mainly interested in the special case that $\gamma = 3$ in the next model.

Example 11 The operator

$$L = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - x^\gamma, \qquad \gamma \geqslant 2$$

on $\mathbb{R}_+$ has discrete spectrum.

The potential here $c_0(x) := x^\gamma$ is simpler than what given in Example 8. The problem is that for this potential, even though the explicit solution of the corresponding harmonic function h is founded out, given by (12), but it seems not practical for the use of Lemma 6. More seriously, it is usually impossible to get an explicit harmonic function h for a given general potential $c(x)$. Hence, we do need a different approach, independent of the exact solution of h. The approach we adopt is similar to what used in [6; §2]: regarding the given model as perturbation of some solvable models. Let us start at a simple choice: $f = \exp\psi_1$ with $\psi_1(x) = x^\alpha/\alpha$, then

$$\big(\psi_1'^2 + \psi_1''\big)(x) = x^{2(\alpha-1)} + (\alpha-1)x^{\alpha-2} =: c_1(x).$$

By setting $\alpha = \gamma/2 + 1$, we can rewrite the above term as

$$c_1(x) = x^\gamma + \frac{\gamma}{2}x^{\gamma/2-1}. \tag{17}$$

This means that the function $\exp\psi_1$ is harmonic of the operator L with potential $c_1(x)$, for which the function h is solvable. One may regard the original model with potential $c_0(x) = x^\gamma$ is a perturbation of the solvable one with potential $c_1(x)$. Next, we have a solvable model at hand, Example 8, for which its potential $c_2(x)$ is more complicated than $c_1(x)$. This is due to the reason we required: the solvable models should be as close as possible to the model we are interested. Comparing the model given in (17) with the one given in Example 8:

$$c_1(x) - x^\gamma = \frac{\gamma}{2}x^{\gamma/2-1}; \qquad c_2(x) - x^\gamma = \frac{\gamma}{2(1+x)}x^{\gamma/2-1} + \frac{\gamma(\gamma+4)}{16(1+x)^2},$$

it follows that the growth rate of the latter is one order less than the former.

As an accompany, we now introduce another solvable model with $h = \exp\psi_3$:

$$\psi_3(x) = \frac{1}{\alpha}x^\alpha - (\alpha-1)\log(1+x). \tag{18}$$

Correspondingly, we have

$$\big({\psi_3'}^2 + \psi_3''\big)(x) = x^{2(\alpha-1)} + (\alpha-1)\bigg[\frac{2}{1+x} - 1\bigg]x^{\alpha-2} + \frac{\alpha(\alpha-1)}{(1+x)^2} =: c_3(x). \tag{19}$$

Denote by h_0 the function given in (12). Then we can set $\psi_0 = \log h_0$. By, Lemma 7, we have $\psi_0 > 0$ and $\psi_0' > 0$. We now have four pairs $(\psi_k, c_k)_{k=0}^3$, combining them together, we obtain Table 1.

Table 1 Functions ψ_k and c_k

k	$\psi_k(x)$	$c_k(x)$
0	$\psi_0(x) = \log h_0(x)$	$x^{2(\alpha-1)} = x^\gamma$
1	$\dfrac{x^\alpha}{\alpha}$	$x^{2(\alpha-1)} + (\alpha-1)x^{\alpha-2}$
2	$\dfrac{x^\alpha}{\alpha} - \dfrac{\alpha-1}{2}\log(1+x)$	$x^{2(\alpha-1)} + \dfrac{\alpha-1}{1+x}x^{\alpha-2} + \dfrac{\alpha^2-1}{4(1+x)^2}$
3	$\dfrac{x^\alpha}{\alpha} - (\alpha-1)\log(1+x)$	$x^{2(\alpha-1)} + (\alpha-1)\bigg[\dfrac{2}{1+x}-1\bigg]x^{\alpha-2} + \dfrac{\alpha(\alpha-1)}{(1+x)^2}$

When $\alpha \geqslant 2$, we have

$$\{\psi_k\}_{k=0}^3 > 0, \quad \{\psi_k'\}_{k=0}^3 > 0, \quad \text{and } \{c_k\}_{k=0}^3 > 0 \text{ for all } \ x > 0.$$

Besides, we also have

$$c_2(x) \geqslant c_0(x) \geqslant c_3(x),$$

except a small neighborhood of x around 0, say $(0, 1.8)$ when $\alpha \in [2,3]$ for instance for the second inequality $c_0 \geqslant c_3$. Thus, in terms of the comparison theorem (cf. Lemma 13 (1) below) for the minimal nonnegative solution, up to a constant, we should have

$$\exp\psi_2 \geqslant h_0 = \exp\psi_0.$$

Similarly, we should also have $h_0 \geqslant \exp\psi_3$, up to a constant. Note that the local problem does not interfere our goal since we are interested in only the asymptotic behavior at infinity.

We are now ready to prove the assertion of the example.

Proof of Example 11 Note that the case of $\gamma = 2$ is the classical harmonic oscillator model, for which the answer is well known, see for instance [4; Example 7.7]. Thus, in what follows, we may assume that $\gamma > 2$, or equivalently $\alpha > 2$. Since ψ_2 and ψ_3 have the same leading order and so do ψ_2' and ψ_3', we have

$$\begin{aligned}
\mu\big(h^2\mathbb{1}_{(0,x)}\big)\hat{\nu}\big(h^{-2}\mathbb{1}_{(x,\infty)}\big) &\leqslant C(\alpha)\int_0^x \exp\big[2\psi_2\big]\int_x^\infty \exp\big[-2\psi_3\big] \\
&\qquad \text{(for some } C(\alpha) \text{ independent of } x) \\
&\sim (\psi_2')^{-1}\exp\big[2\psi_2\big](\psi_3')^{-1}\exp\big[-2\psi_3\big] \\
&\sim (\psi_2')^{-2}\exp\big[2(\psi_2-\psi_3)\big] \\
&\sim x^{-2(\alpha-1)}x^{\alpha-1} \\
&\sim 0 \qquad \text{as } x \to \infty,
\end{aligned}$$

once $\alpha > 1$. $\square$

Note that in the proof we adopt upper and lower bounds ψ_2 and ψ_3 of ψ_0, we need not only the same leading terms of ψ_2 and ψ_3 and also the second leading terms of them. That is the reason we have to choose ψ_2 and ψ_3 carefully.

Finally, we have proved the discreteness of the operator given in (8) for $\gamma \geqslant 2$ and furthermore for the corresponding complex operator with $\gamma \in (2, 4)$ mentioned in Section 1.

The approach used in this section is taken from [4; §7]. A new simplified one is given in the next section.

3 New simplified approach

First, we present a simplified construction of the harmonic function h given by Lemma 7. The observation goes to the special structure of the matrix G:

$$G(x) = \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix} \quad \text{(say!)}.$$

Then we have

$$G \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \beta x \end{pmatrix},$$

and in the next step, we obtain

$$G \begin{pmatrix} 0 \\ \beta x \end{pmatrix} = \begin{pmatrix} \alpha\beta x \\ 0 \end{pmatrix}.$$

Since we are only interested in the first component of the resulting vector, this suggests us to combining the first two steps into one:

$$G^2 = \begin{pmatrix} \alpha\beta & 0 \\ 0 & \alpha\beta \end{pmatrix}, \quad G^2 \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} \alpha\beta x \\ 0 \end{pmatrix}.$$

This leads us to use one-dimensional function h instead of the two dimensional one F. To do so, we compute the double integration of the original G given in Lemma 7:

$$\begin{aligned} \int_0^x \mathrm{d}y\, G(y) \int_0^y \mathrm{d}z\, G(z) &= \int_0^x \mathrm{d}z \int_z^x \mathrm{d}y\, G(y)G(z) \\ &= \int_0^x \mathrm{d}z \int_z^x \mathrm{d}y \begin{pmatrix} \dfrac{c}{a}(z)e^{C(z)-C(y)} & 0 \\ 0 & \dfrac{c}{a}(y)e^{C(y)-C(z)} \end{pmatrix}. \end{aligned}$$

We need the first component at the first line of the matrix on the right-hand side:

$$\int_0^x \mathrm{d}z \int_z^x \mathrm{d}y\, \frac{c}{a}(z)e^{C(z)-C(y)} = \int_0^x \mathrm{d}z \left[\frac{c}{a}(z)e^{C(z)} \int_z^x \mathrm{d}y\, e^{-C(y)} \right].$$

Therefore, we define the following kernel:

$$k(x, z) = \frac{c}{a}(z)e^{C(z)} \int_z^x \mathrm{d}y\, e^{-C(y)}, \qquad x \geqslant z \geqslant 0. \tag{20}$$

Then, we have the following successful approximating procedure:

$$\tilde{f}^{(n+1)}(x) = \int_0^x \mathrm{d}z\, k(x,z)\tilde{f}^{(n)}(z), \quad n \geqslant 1, \qquad \tilde{f}^{(1)}(x) \equiv 1. \tag{21}$$

Therefore, the required harmonic function h^* is given by

$$h^*(x) = \sum_{n=1}^{\infty} \tilde{f}^{(n)}(x), \qquad x \in \mathbb{R}_+. \tag{22}$$

Here is the new construction of h^*.

Lemma 12 Let k be defined by (20). Then h^* is the minimal nonnegative solution to the equation

$$f(x) = \int_0^x \mathrm{d}z\, k(x,z)f(z) + g(x), \qquad x \in \mathbb{R}_+, \tag{23}$$

with $g(x) \equiv 1$. It can be obtained either by using the first successive approximation scheme $h^* = \lim_{n\to\infty} \uparrow f^{(n)}$ (pointwise):

$$f^{(n+1)}(x) = \int_0^x \mathrm{d}z\, k(x,z)f^{(n)}(z) + 1, \quad n \geqslant 1, \qquad f^{(1)}(x) \equiv 1. \tag{24}$$

or by the second successive approximation scheme defined in (21).

Before moving further, let us mention that the construction of h given in Lemma 12 corresponds [5; Algorithm 14] for $(\tilde{b}_k)$ in the discrete case; while the construction in Lemma 7 corresponds [6; Lemmas 4 and 5].

The proof of Lemma 12 is given in the next section. We are now ready to prove (12).

Proof of (12) In this case, we have $a = 1, b = 0$ and $c(x) = x^\gamma$. Then $C(x) \equiv 0$. Hence $k(x,z) = z^\gamma(x-z)$. By (21), we have $\tilde{f}^{(1)}(x) \equiv 1$ and

$$\tilde{f}^{(n+1)}(x) = \int_0^x \mathrm{d}z\, z^\gamma(x-z)\tilde{f}^{(n)}(z) = x\int_0^x z^\gamma \tilde{f}^{(n)}(z) - \int_0^x z^{\gamma+1}\tilde{f}^{(n)}(z), \quad n \geqslant 1.$$

In particular,

$$\tilde{f}^{(2)}(x) = x\int_0^x z^\gamma - \int_0^x z^{\gamma+1} = \frac{x^{\gamma+2}}{\gamma+1} - \frac{x^{\gamma+2}}{\gamma+2} = \frac{x^{\gamma+2}}{(\gamma+1)(\gamma+2)} = \frac{x^\alpha}{\alpha^2(1-\alpha)}, \ \alpha := \gamma+2$$

which is the special case of $n = 2$ in the following formula:

$$\tilde{f}^{(n)}(x) = \frac{x^{\alpha(n-1)}}{(n-1)!\,\alpha^{2(n-1)}(1-1/\alpha)_{n-1}}.$$

In view of this, it is easy to prove the required assertion by induction. □

To make additional details to the proof of Example 11, we need some simple properties of the minimal solution to equation (23).

Lemma 13

(1) (*Comparison*). Let $\tilde{k} \geqslant k$ and $\tilde{g} \geqslant g$ (pointwise). If

$$\tilde{f}(x) \geqslant \int_0^x \mathrm{d}z\, \tilde{k}(x,z)\tilde{f}(z) + \tilde{g}(x), \qquad x \in \mathbb{R}_+,$$

then $\tilde{f} \geqslant f^*$, where f^* is the minimal solution to (23).

(2) (*Linear combination*). Let $\beta_k \geqslant 0$ and $g_k \geqslant 0$ and f_k^* be the minimal solution to (23) with $g = g_k$, $k = 1, 2$, then $f^* := \beta_1 f_1^* + \beta_2 f_2^*$ is the minimal solution to (23) with $g = \beta_1 g_1 + \beta_2 g_2$.

(3) (*Localization*). Let f^* the minimal solution to (23), $x_0 \in (0, \infty)$, and $\{\tilde{f}^*(x) : x \geqslant x_0\}$ be the minimal solution to the equation

$$f(x) = \int_{x_0}^x \mathrm{d}z\, k(x,z)f(z) + \int_0^{x_0} \mathrm{d}z\, k(x,z)f^*(z) + g(x), \qquad x \geqslant x_0.$$

Then $\tilde{f}^* = f^*$ on $[x_0, \infty)$.

Refer to [3; Theorem 2.6, Corollary 2.8 and Theorem 2.13], respectively, for the proofs of these assertions. Actually, the proofs are rather elementary.

We are now ready to come back to the proof of Example 11. Note that as mentioned in [4; Theorem 7.4 (1)], the solution to (11) and then (23) is indeed unique, hence we simply write h^* as h if there is no confusion. Because $c_2 \geqslant c_0$, it follows that $h_2 \geqslant h_0$, or equivalently, $\psi_2 \geqslant \psi_0$ by Lemma 13 (1). Then we have

$$\mu\big(h_0^2 \mathbb{1}_{(0,x)}\big) \leqslant \int_0^x \exp\big[2\psi_2\big], \qquad x \geqslant 0. \tag{25}$$

It is not so easy to control the other part $\hat{\nu}\big(h_0^{-2}\mathbb{1}_{(x,\infty)}\big)$. For this, we need more work. Corresponding to each c_j, we have a k_j defined by (20). Then we have the solution h_j to equation (23). What we need is a lower bound of h_0 by h_3. Unfortunately, it seems not trivial to prove the inequality $h_0 \geqslant h_3$ since c_0 may be smaller than c_3 on a small neighborhood of the origin, see Figure 1: $\epsilon_1(\alpha, x) = c_0(x) - c_3(x)$. On the right, from top to button, three curves correspond to $\alpha = 3,\ 2.5,\ 2$, respectively.

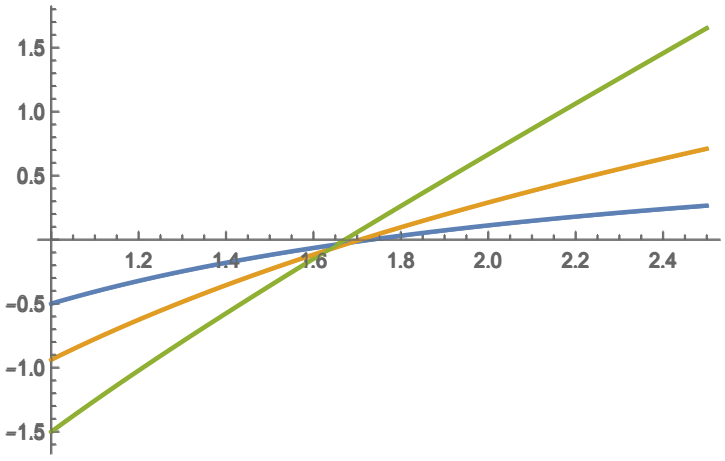


Figure 1 $\epsilon_1(\alpha, x)$ on $[1,\ 2.5]$

Hence we look for a weaker estimate that $h_0 \geqslant \varepsilon h_3$. To do so, recall that by Lemma 13 (3), $(h_j(x) : x \in [1, \infty))$ is the solution to the equation

$$h_j(x) = \int_1^x \mathrm{d}z\, k_j(x,z) h_j(z) + \int_0^1 \mathrm{d}z\, k_j(x,z) h_j^*(z) + 1 \text{ (i.e. } g(x) \equiv 1), \quad x \geqslant 1,\ j = 0, 3.$$

Noting that even though we do not have $c_0 \geqslant c_3$ in a neighborhood of origin, but we do have $C(\alpha)c_0 > c_3$, where $C(\alpha) = \alpha^2$. The resulting

$$\epsilon_2(\alpha, x) := \frac{\alpha^2 c_0(x) - c_3(x)}{\alpha - 1}$$

are shown by Figures 2 and 3. Again, the three curves from top to button correspond to $\alpha = 3,\ 2.5,\ 2$ respectively. The minimum of $\epsilon_2(2, x) = 0.838312$ achieved at $x = 0.171663$. We remark that there is a lot of freedom in choosing a suitable $C(\alpha)$, which is the advantage of the present approach. Mainly, two conditions are required: to guarantee the difference ϵ_2 defined above to be positive; easier for computation. This is based on the fact such a modification does not interfere our conclusion.

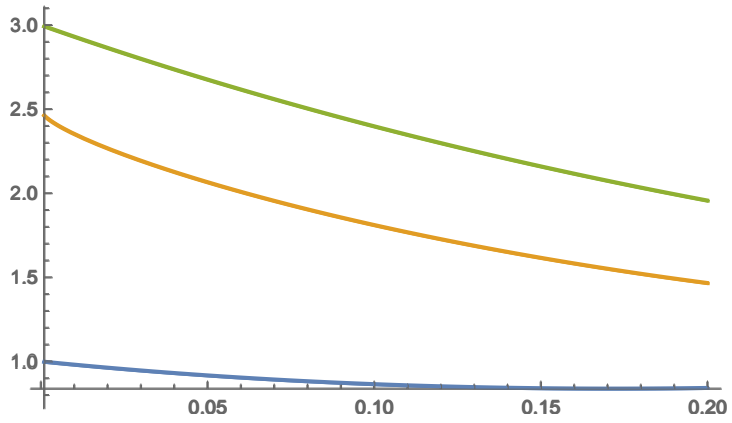


Figure 2 $\epsilon_2(\alpha, x)$ on $[0.01, 0.2]$

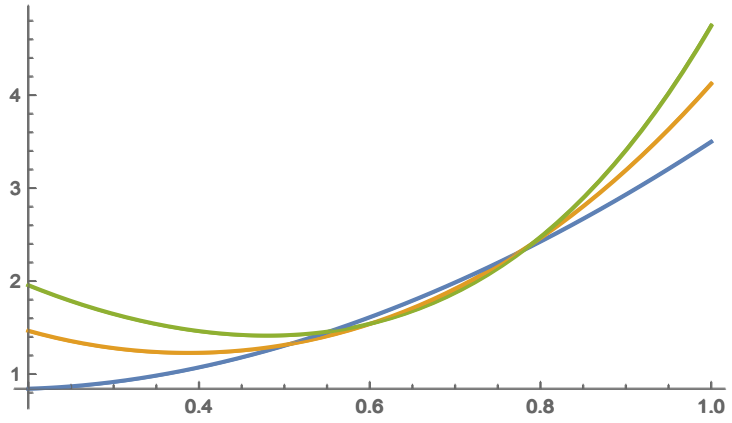


Figure 3 $\epsilon_2(\alpha, x)$ on [0.2, 1]

Let

$$g^{(1)} = C(\alpha)\bigg[\int_0^1 \mathrm{d}z\, k_0(x,z)h_0^*(z) + g(x)\bigg].$$

Note that here not only k_0 but also g is enlarged by the factor $C(\alpha)$. Regarding $g^{(1)}$ as a new g and consider the minimal solution, denoted by $h_0^{(1)}$, to the equation

$$h_0(x) = \int_1^x \mathrm{d}z\, k_0(x,z)h_0(z) + g^{(1)}(x), \qquad x \geqslant 1.$$

Because $k_0 \geqslant k_3$ and $g^{(1)} \geqslant 1$ on $[1,\infty)$, by Lemma 13 (1), we have $h_0^{(1)} \geqslant h_3$ on on $[1,\infty)$. Moreover, By Lemma 13 (2), we also have $h_0^{(1)} = C(\alpha)h_0$ on $[1,\infty)$. Hence $h_0 \geqslant h_3/C(\alpha)$ on $[1,\infty)$. Therefore, we have

$$\hat{\nu}\big(h_0^{-2}\mathbb{1}_{(x,\infty)}\big) \leqslant C(\alpha)^2\hat{\nu}\big(h_3^{-2}\mathbb{1}_{(x,\infty)}\big), \qquad x \geqslant 1. \tag{26}$$

Combining this with (25), we arrived at the first step in the last formula of the original proof at the end of Section 2.

To conclude this section, we mention there is actually a simpler but rough way to make the comparison given in the last paragraph. For this, let

$$g_j = \int_0^1 \mathrm{d}z\, k_j(\cdot,z)h_j^*(z) + g, \qquad j = 0,\ 3.$$

Since g_0 and g_3 are finite and positive, there exists a large enough $\widetilde{C}(\alpha)$ depending on α only such that

$$\tilde{g}_0 := \widetilde{C}(\alpha)g_0 \geqslant g_3.$$

Replacing $C(\alpha)$ by the unexplicit $\widetilde{C}(\alpha)$ in the last paragraph, we obtain the required estimate.

4 Proofs and extensions

In this section, we prove the theorems introduced in Section 1, as well as Lemma 12 given in Section 3. Besides, we extend Theorems 1 and 3 to a more general class of operators.

Proof of Theorem 1 Note that

$$((L+c)f,g)_\mu = (\partial^*(a\partial)f,g)_\mu + (b^*\partial f,g)_\mu =: \mathrm{I} + \mathrm{II}.$$

Under suitable condition at infinity, using the integration by parts formula, we obtain

$$\mathrm{I} = -\int_{\mathbb{R}^d}\sum_{i,j} a_{ij}(\partial_j f)\partial_i(\bar{g}e^V) = -\langle a\partial f,\ \partial g\rangle_\mu - ((\partial V)^* a\partial f,\ g)_\mu.$$

Hence

$$\mathrm{I} + \mathrm{II} = -\langle a\partial f,\ \partial g\rangle_\mu + \big(\big[b^* - (\partial V)^* a\big]\partial f,\ g\big)_\mu.$$

For the convenience of the following, we denote $\hat{b}^* = b^* - (\partial V)^* a$, that is, $\hat{b} = b - a^*(\partial V)$. We obtain

$$((L+c)f,\ g)_\mu = -\langle a\partial f,\ \partial g\rangle_\mu + \big(\hat{b}^*\partial f,\ g\big)_\mu =: \mathrm{III} + \mathrm{X}. \tag{27}$$

Similarly, we have

$$(f,\ (L+c)g)_\mu = -\langle \partial f,\ a\partial g\rangle_\mu + \big(f,\ \hat{b}^*\partial g\big)_\mu =: \mathrm{IV} + \mathrm{V}. \tag{28}$$

Clearly, the first terms on the right-hand side of (27) and (28) are coincided iff a is Hermitian. We assume this in what follows. However, the second terms on the right-hand side in (27) and (28) are not easy to compare since the factors $(\partial_i f)\bar{g}$ and $f(\partial_i \bar{g})$ are different. Note that the term X in (27) is equal to

$$\begin{aligned}\mathrm{X} &= -\sum_i \int_{\mathbb{R}^d} \hat{b}_i f(\partial_i \bar{g}) e^V - \sum_i \int_{\mathbb{R}^d} \partial_i\big(\hat{b}_i e^V\big) f\bar{g} \\ &= -\langle \hat{b}f,\ \partial g\rangle_\mu - \big(\big[((\partial V)^* + \partial^*)\hat{b}\big]f,\ g\big)_\mu.\end{aligned}$$

Thus, we can rewrite (27) as

$$\begin{aligned}((L+c)f,\ g)_\mu &= -\langle a\partial f,\ \partial g\rangle_\mu - \langle \hat{b}f,\ \partial g\rangle_\mu - \big(\big[((\partial V)^* + \partial^*)\hat{b}\big]f,\ g\big)_\mu \\ &=: \mathrm{III} + \mathrm{VI} + \mathrm{VII}.\end{aligned} \tag{29}$$

Keeping $a^H = a$ in mind and combining the last three formulas together, since III = IV, by (28) and (29), it follows that

$$\begin{aligned}&-(Lf,\ g)_\mu + (f,\ Lg)_\mu \\ &\quad = \mathrm{V} - \mathrm{VI} - \mathrm{VII} + ((c-\bar{c})f,\ g)_\mu \\ &\quad = \langle (\bar{\hat{b}} + \hat{b})f, \partial g\rangle_\mu + \big(\big[c - \bar{c} + ((\partial V)^* + \partial^*)\hat{b}\big]f,\ g\big)_\mu \\ &\quad = \langle 2\,(\mathrm{Re}\,\hat{b})f,\ \partial g\rangle_\mu + \big(\big[2\sqrt{-1}\,\mathrm{Im}\,c + ((\partial V)^* + \partial^*)\big(\mathrm{Re}\,\hat{b} + \sqrt{-1}\,\mathrm{Im}\hat{b}\big)\big]f,\ g\big)_\mu.\end{aligned} \tag{30}$$

From the first term on the right-hand side, we obtain (2) since $\mathrm{Re}\,a^* = \mathrm{Re}\,a$ and $\hat{b} = b - a^*(\partial V)$, then (3) follows from the last term on the right-hand side, since

$$\mathrm{Im}\,a^* = \mathrm{Im}\,\bar{a} = -\mathrm{Im}\,a.$$

Finally, we compute the quadratic form of the operator L. Using (2) and (3) and the analysis in (30), from (29), we deduce that

$$(-Lf,\ g)_\mu = \langle a\partial f,\ \partial g\rangle_\mu + \sqrt{-1}\,\big\langle \big(\mathrm{Im\ b} + (\mathrm{Im\ a})(\partial V)\big)f,\ \partial g\big\rangle_\mu + (\bar{c}f,\ g)_\mu, \tag{31}$$

In more details, the second term on the right-hand side of (31) comes from the one of (29):

$$-\mathrm{VI} = \big\langle \hat{b}f,\ \partial g\big\rangle \overset{(2)}{=} \sqrt{-1}\,\big\langle \big(\mathrm{Im\ b} + (\mathrm{Im\ a})(\partial V)\big)f,\ \partial g\big\rangle_\mu.$$

The third term on the right-hand side of (31) comes from the one of (29) plus the term containing c:

$$\begin{aligned}(cf,\ g)_\mu) - \mathrm{VII} &= \big(\big[c + ((\partial V)^* + \partial^*)(b - a^*(\partial V))\big]f,\ g\big)_\mu \\ &\overset{(2)}{=} \big(\big[c + \sqrt{-1}\,((\partial V)^* + \partial^*)(\mathrm{Im}\,b + (\mathrm{Im}\,a)(\partial V))\big]f,\ g\big)_\mu \\ &\overset{(3)}{=} \big(\big[\mathrm{Re}\,c + \sqrt{-1}\,\mathrm{Im}\,c - 2\sqrt{-1}\,\mathrm{Im}\,c\big]f,\ g\big)_\mu \\ &= (\bar{c}\,f,\ g)_\mu.\end{aligned}$$

This proves the third term on the right-hand side of (31). At the same time, we have proved (4).

Similarly, from (28), we obtain

$$(f,\ -Lg)_\mu = \langle \partial f,\ a\partial g\rangle_\mu - \sqrt{-1}\,\big\langle \big(\operatorname{Im}\bar{b} + (\operatorname{Im}\bar{a})(\partial V)\big) f,\ \partial g\big\rangle_\mu + (\bar{c}f,\ g)_\mu. \tag{32}$$

Because

$$\text{Im b} + (\text{Im a})(\partial V) = -\big(\operatorname{Im}\bar{b} + (\operatorname{Im}\bar{a})(\partial V)\big),$$

the second term on the right-hand side of (32) coincides with the one of (31), and so we checked again $(-Lf,\ g)_\mu = (f,\ -Lg)_\mu$ by (31)and (32). □

Proof of Theorem 3 First, we have

$$\begin{aligned}\nabla(a\nabla(hf)) &= \nabla(ha\nabla f + fa\nabla h)\\ &= h\nabla(a\nabla f) + f\nabla(a\nabla h) + (a + a^*)(\nabla h)\cdot(\nabla f).\end{aligned}$$

Hence

$$\begin{aligned}I_{[h\neq 0]}\frac{1}{h}L(hf) &= (L+c)f + I_{[h\neq 0]}\frac{f}{h}Lh + I_{[h\neq 0]}\frac{1}{h}(a+a^*)\nabla h\cdot\nabla f\\ &= (L+c)f + I_{[h\neq 0]}\frac{1}{h}(a+a^*)\nabla h\cdot\nabla f \quad (\text{since } Lh = 0,\ \mu\text{-a.e.}).\end{aligned}$$

The assertion now follows from [5; Lemma 8]. □

In parallel to Theorems 1 and 3, respectively, we have two theorems below. Let D denote the column of operators $\{\partial_j + \gamma_j\}_{j=1}^d$, where $\gamma = (\gamma_j)$ is the ferromagnetic potential. In view of (34) below, one may deduce the required result from Theorem 1, but we prefer to present a direct proof given below for a quadratic form different from Theorem 1.

Theorem 14 The operator

$$L = D^*(aD) + b^*\partial - c \tag{33}$$

$$= \partial^*(a\partial) + (\gamma^*(a+a^*) + b^*)\partial + D^*(a\gamma) - c \tag{34}$$

$$= a\cdot\partial\partial^* + (D^*a + \gamma^*a^* + b^*)\partial + D^*(a\gamma) - c \quad (\text{by (5)}) \tag{35}$$

on $\mathscr{C}^2(\mathbb{R}^d)$ is Hermitizable with respect to μ iff $a^H = a$, $\bar{\gamma} = -\gamma$, and

$$\operatorname{Re} b = (\operatorname{Re} a)(\partial V), \tag{36}$$

$$\operatorname{Im} c = -(\operatorname{Im}\gamma)^*(\operatorname{Re} a)(\partial V) - \frac{1}{2}\big[(\partial V)^* + \partial^*\big]\big[\operatorname{Im} b + (\operatorname{Im} a)(\partial V)\big]. \tag{37}$$

In this case, we have the symmetric form:

$$(-Lf, g)_\mu = \langle aDf,\ Dg\rangle_\mu + \sqrt{-1}\,\big\langle \big(\text{Im b} + (\text{Im a})(\partial V)\big) f,\ \partial g\big\rangle_\mu + (\bar{c}f,\ g)_\mu,$$
$$f, g \in \mathscr{D}(L) \subset L^2(\mu),$$

where $\operatorname{Im} c$ satisfies (37).

By setting $\gamma = 0$ in Theorem 14, we return to Theorem 1. Similarly, we have the following isospectral operator for the operator L defined in Theorem 14.

Theorem 15 The operator $(L, \mathscr{D}(L))$ on $L^2(\mu)$ defined in Theorem 14 and let h be L-harmonic $Lh = 0$, μ-a.e. Set

$$L^0 = a \cdot \partial\partial^* + (D^*a + \gamma^* a^* + b^*)\partial = \partial^*(a\partial) + (\gamma^*(a + a^*) + b^*)\partial \quad \text{(by (35))}.$$

Then the operator $(L, \mathscr{D}(L))$ is isospectral to the following one:

$$\begin{aligned}\widetilde{L} &= L^0 + I_{[h\neq 0]}\frac{1}{h}(\partial h)^*(a + a^*)\partial \\ &= \partial^*(a\partial) + \left[b^* + \left(\gamma^* + I_{[h\neq 0]}\frac{1}{h}(\partial h)^*\right)(a + a^*)\right]\partial, \\ \mathscr{D}(\widetilde{L}) &= \left\{\tilde{f} \in L^2(\tilde{\mu}) : \tilde{f}h \in \mathscr{D}(L)\right\}, \quad \tilde{\mu} := |h|^2\mu.\end{aligned}$$

Proof of Theorem 14 Noting that under suitable boundary condition, using the integration by parts formula, we obtain

$$\begin{aligned}&\big((\partial + \gamma)^*(a(\partial + \gamma))f, g\big)_\mu \\ &= -\langle a(\partial + \gamma)f,\ (\partial V + \partial)g\rangle_\mu + \langle a(\partial + \gamma)f,\ \bar{\gamma}g\rangle_\mu \\ &= -\langle a(\partial + \gamma)f,\ (\partial - \bar{\gamma})g\rangle_\mu - ((\partial V)^* a\partial f, g)_\mu - ((\partial V)^* a\gamma f,\ g)_\mu.\end{aligned}$$

Hence, we have

$$\begin{aligned}-(Lf,\ g)_\mu &= \langle a(\partial + \gamma)f,\ (\partial - \bar{\gamma})g\rangle_\mu - \big((b^* - (\partial V)^* a)\partial f,\ g\big)_\mu + \big((c + (\partial V)^* a\gamma)f,\ g\big)_\mu \\ &=: \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Thus, for some type of symmetriy of the first term I, we should have $a^H = a$ and $\bar{\gamma} = -\gamma$, which are assumed to be held in what follows. Then

$$I = \langle aDf,\ Dg\rangle_\mu = \langle Df,\ aDg\rangle_\mu.$$

For the second term II, we need more work. For the convenience of the following, we let again $\hat{b}^* = b^* - (\partial V)^* a$. By integration by parts formula, we have

$$\mathrm{II} = -\left\langle \partial f,\ \bar{\hat{b}}g\right\rangle_\mu = \left(f,\ \bar{\hat{b}}^*\partial g\right)_\mu + \left(f,\ ((\partial V)^* + \partial^*)\bar{\hat{b}}\, g\right)_\mu.$$

Next, we have

$$-(f,\ Lg)_\mu = \langle(\partial - \bar{\gamma})f,\ a(\partial + \gamma)g\rangle_\mu - \big(f,\ \hat{b}^*\partial g\big)_\mu + \big(f, (c + (\partial V)^* a\gamma)g\big)_\mu.$$

Recall that $a^H = a, \bar{\gamma} = -\gamma$, that is, $\gamma = \sqrt{-1}\,\mathrm{Im}\,\gamma$. Hence

$$\begin{aligned}&-(Lf,\ g)_\mu + (f,\ Lg)_\mu \\ &= \left(f,\ \left(\hat{b}^* + \bar{\hat{b}}^*\right)\partial g\right)_\mu + \left(\left[c - \bar{c} + (\partial V)^*(a + \bar{a})\gamma + ((\partial V)^* + \partial^*)\hat{b}\right]f,\ g\right)_\mu \\ &= \big(f,\ 2(\mathrm{Re}\,\hat{b})^*\partial g\big)_\mu + \Big(\Big[2\sqrt{-1}\left(\mathrm{Im}\,c + (\partial V)^*(\mathrm{Re}\,a)(\mathrm{Im}\,\gamma)\right) \\ &\qquad + ((\partial V)^* + \partial^*)(\mathrm{Re}\,\hat{b} + \sqrt{-1}\,\mathrm{Im}\,\hat{b})\Big]f,\ g\Big)_\mu.\end{aligned}$$

The required condition (36) follows from the first term $\mathrm{Re}\,\hat{b} = 0$ on the right-hand side, and then (37) follows from the first assertion plus the second term there

$$2\sqrt{-1}\left(\mathrm{Im}\,c + (\partial V)^*(\mathrm{Re}\,a)(\mathrm{Im}\,\gamma)\right) + ((\partial V)^* + \partial^*)(\mathrm{Re}\,\hat{b} + \sqrt{-1}\mathrm{Im}\,\hat{b}) = 0,$$

since $\operatorname{Im}\hat{b} = \operatorname{Im} b + (\operatorname{Im} a)(\partial V)$. □

Proof of Theorem 15 Similar to the proof of [5; Theorem 33], replacing the original b by γ, we have

$$\begin{aligned}
L(hf) &= (a\cdot\partial\partial^*)(hf) + (D^*a + \gamma^*a^* + b^*)\partial(hf) + (D^*(a\gamma) - c)hf \\
&= h\big[(a\cdot\partial\partial^*)f + (D^*a + \gamma^*a^* + b^*)(\partial f)\big] \\
&\quad + f\big[(a\cdot\partial\partial^*)h + (D^*a + \gamma^*a^* + b^*)(\partial h) + (D^*(a\gamma) - c)h\big] \\
&\quad + \big[(\partial h)^*a(\partial f) + (\partial f)^*a(\partial h)\big].
\end{aligned}$$

Hence

$$\begin{aligned}
&I_{[h\neq 0]}\frac{1}{h}L(hf) \\
&= \big[(a\cdot\partial\partial^*)f + (D^*a + \gamma^*a^* + b^*)(\partial f)\big] \\
&\quad + I_{[h\neq 0]}\frac{f}{h}\big[(a\cdot\partial\partial^*)h + (D^*a + \gamma^*a^* + b^*)(\partial h) + (D^*(a\gamma) - c)h\big] \\
&\quad + I_{[h\neq 0]}\frac{1}{h}\big[(\partial h)^*a(\partial f) + (\partial f)^*a(\partial h)\big] \\
&=: \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

Because

$$\begin{aligned}
\mathrm{I} &= L^0 f, \\
\mathrm{II} &= I_{[h\neq 0]}\frac{f}{h}Lh = 0 \text{ (by harmonic assumption)}, \\
\mathrm{III} &= I_{[h\neq 0]}\frac{1}{h}(\partial h)^*(a + a^*)(\partial f) \text{ (since } (\partial f)^*a(\partial h) = (\partial h)^*a^*(\partial f)).
\end{aligned}$$

Combining these facts together, we obtain the required assertion. □

Proof of Theorem 4 The final assertion is a consequence of [5; Theorem 9].

Following the proof of [5; Theorem 34], we have

$$(\tilde{a}\partial)\left(\frac{f}{h}\right) = \tilde{a}\left(\frac{1}{h}(\partial f) + f\partial\left(\frac{1}{h}\right)\right) = \frac{1}{h}\tilde{a}(\partial f) + f\tilde{a}\partial\left(\frac{1}{h}\right).$$

Hence

$$\begin{aligned}
(\partial^*\tilde{a}\partial)\left(\frac{f}{h}\right) &= \partial^*\left(\frac{1}{h}\tilde{a}(\partial f) + f\tilde{a}\partial\left(\frac{1}{h}\right)\right) \\
&= \frac{1}{h}(\partial^*\tilde{a}\partial)f + \left(\partial\left(\frac{1}{h}\right)\right)^*\tilde{a}(\partial f) + f\partial^*\tilde{a}\partial\left(\frac{1}{h}\right) + (\partial f)^*\tilde{a}\partial\left(\frac{1}{h}\right)
\end{aligned}$$

Because

$$(\partial f)^*\tilde{a}\partial\left(\frac{1}{h}\right) = \left(\tilde{a}\partial\left(\frac{1}{h}\right)\right)^*(\partial f) = \left(\partial\left(\frac{1}{h}\right)\right)^*\tilde{a}^*(\partial f),$$

and then

$$\left(\partial\left(\frac{1}{h}\right)\right)^*\tilde{a}(\partial f) + (\partial f)^*\tilde{a}^*\partial\left(\frac{1}{h}\right) = \left(\partial\left(\frac{1}{h}\right)\right)^*(\tilde{a} + \tilde{a}^*)(\partial f),$$

we obtain

$$(\partial^*\tilde{a}\partial)\left(\frac{f}{h}\right) = \frac{1}{h}(\partial^*\tilde{a}\partial)f + f\partial^*\tilde{a}\partial\left(\frac{1}{h}\right) + \left(\partial\left(\frac{1}{h}\right)\right)^*(\tilde{a} + \tilde{a}^*)(\partial f).$$

Hence

$$h\widetilde{L}\left(\frac{f}{h}\right) = \widetilde{L}f + h\left(\partial\left(\frac{1}{h}\right)\right)^*(\tilde{a}+\tilde{a}^*)(\partial f) + h\big(\partial^*\tilde{a}+\tilde{b}^*\big)\partial\left(\frac{1}{h}\right)f.$$

Next, because

$$\partial\left(\frac{1}{h}\right) = -\frac{1}{h^2}(\partial h), \qquad \tilde{a}\partial\left(\frac{1}{h}\right) = -\frac{1}{h^2}\tilde{a}(\partial h),$$
$$\partial^*\tilde{a}\partial\left(\frac{1}{h}\right) = \frac{2}{h^3}(\partial h)^*\tilde{a}(\partial h) - \frac{1}{h^2}\partial^*\tilde{a}(\partial h),$$

we obtain the expression of

$$L^h f := h\widetilde{L}\left(\frac{f}{h}\right).$$

The proof of the theorem is now completed. □

Proof of Lemma 12 First, we prove that the function h^* defined by (22) is a harmonic one we required. Note that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\tilde{f}^{(n)}(x) &= k(x,x)\tilde{f}^{(n-1)}(x) + e^{-C(x)}\int_0^x \mathrm{d}z\,\frac{c}{a}(z)e^{C(z)}\tilde{f}^{(n-1)}(z)\\ &= e^{-C(x)}\int_0^x \mathrm{d}z\,\frac{c}{a}(z)e^{C(z)}\tilde{f}^{(n-1)}(z) \quad (\text{since } k(x,x)=0);\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\tilde{f}^{(n)}(x) &= -C'(x)e^{-C(x)}\int_0^x \mathrm{d}z\,\frac{c}{a}(z)e^{C(z)}\tilde{f}^{(n-1)}(z) + \frac{c}{a}(x)\tilde{f}^{(n-1)}(x)\\ &= -\frac{b}{a}(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\tilde{f}^{(n)}(x) + \frac{c}{a}(x)\,\tilde{f}^{(n-1)}(x).\end{aligned}$$

We obtain

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}x^2}h^*(x) &= \frac{\mathrm{d}^2}{\mathrm{d}x^2}\sum_{n=2}^{\infty}\tilde{f}^{(n)}(x) \quad \big(\text{since } \tilde{f}^{(1)}=1\big)\\ &= -\frac{b}{a}(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\sum_{n=1}^{\infty}\tilde{f}^{(n)}(x) + \frac{c}{a}(x)\sum_{n=1}^{\infty}\tilde{f}^{(n)}(x)\\ &= -\frac{b}{a}(x)\,\frac{\mathrm{d}}{\mathrm{d}x}h^*(x) + \frac{c}{a}(x)\,h^*(x).\end{aligned}$$

Therefore, we arrive at the harmonic equation

$$a\,\frac{\mathrm{d}^2}{\mathrm{d}x^2}h^* + b\,\frac{\mathrm{d}}{\mathrm{d}x}h^* - c\,h^* = 0$$

as required.

The remainders of the assertions are standard, see for instance [3; §2.1]. For example, to show that the two successive approximation schemes given in the lemma lead to the same solution h^*, simply check that

$$f^{(n)} = \sum_{j=1}^{n}\tilde{f}^{(n)}$$

for each $n \geqslant 1$, by induction. □

Acknowledgments Thanks are given to the referees for their careful reading of an earlier version of the paper. This work was supported in part by National Natural Science Foundation of China (Grant No. 11771046), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Bagarello F., Passante, R. and Trapani, C. (2015) *Non-Hermitian Hamiltonians in Quantum Physics*. Springer, Switzerland.

[2] Bender, C.M. and Boettcher, S. (1998) *Real spectra in non-Hermitian Hamiltonians having $\mathscr{P}\mathscr{T}$ symmetry*. Phys. Rev. Lett. 80:4243–5246

[3] Chen, M.F. (2004). *From Markov Chains to Non-Equilibrium Particle Systems*. World Scientific, Singapore, 2nd Ed. (1st Ed., 1992).

[4] Chen, M.F. (2014). *Criteria for discrete spectrum of* 1D *operators*. Commu. Math. Stat. 2: 279–309.

[5] Chen M.F. (2018) *Hermitizable, isospectral complex matrices or differential operators*. Front Math China, 2018, 13(6): 1267–1311.

[6] Chen M.F. (2020) *On spectrum of Hermitizable tridiagonal matrices*. Front. Math. China 2020, 15(2): 285–303.

[7] Chen, M.F. and Zhang, X. (2014). *Isospectral operators*. Commu Math Stat 2, 17–32.

[8] Doey, P. Dunning, C. and Tateo1, R. (2001). *Spectral equivalences, Bethe ansatz equations, and reality properties in $\mathscr{P}\mathscr{T}$-symmetric quantum mechanics*. J. Phys. A: Math. Gen. 34:5679–5704.

[9] Li, J.Y. (2020). *Hermitizability of complex elliptic operators* (in Chinese). Master's thesis at Beijing Normal University.

[10] Moiseyev, N. (2011). *Non-Hermitian Quantum Mechanics*. Cambridge University Press.

[11] Mostafazadeh, A. (2015). *Physics of Spectral Singularities*. In "Geometric Methods in Physics", 145–165, edited by P. Kielanowski et al, Springer.

[12] Polyanin, A.D. and Zaitsev, V.F. (2003). *Handbook of Exact Solutions for Ordinary Differential Equations,* 2nd Ed.. Chapman & Hall/CRC, Boca Raton, London, New York, Washington, D.C.

Front. Math. China 2021, 16 (2): 345–379
https://doi.org/10.1007/s11464-021-0909-6

Computing top eigenpairs of Hermitizable matrix

Mu-Fa Chen[1,2,3], Zhi-Gang Jia[1,4], Hong-Kui Pang[1,4]

[1]RIMS, Jiangsu Normal Univ., Xuzhou, 221116;
[2]Sch. Math. & [3]LMCS, Beijing Normal Univ., Beijing 100875;
[4]Sch. Math. & Statis., Jiangsu Normal Univ., Xuzhou, 221116

October 10, 2019

Abstract

The top eigenpairs at the title mean the maximal, the submaximal, or a few of the subsequent eigenpairs of an Hermitizable matrix. Restricting on top ones is to handle with the matrices having large scale, for which only little is known up to now. This is different from some mature algorithms, that are clearly limited only to medium-sized matrix for calculating full spectrum. It is hoped that a combination of this paper with the earlier works, to be seen soon, may provide some effective algorithms for computing the spectrum in practice, especially for matrix mechanics.

This paper is a continuation of [6] which surveys partially the results (algorithms) presented in [3–5], plus some additional materials. The main context in [6] is on real tridiagonal matrix, except few comments on the complex situation. In the real context, the theoretical study on the leading spectrum of the infinitesimal matrix operator is reviewed in [2]. This paper starts at a computational technique for checking the Hermitizability and then goes to study the Householder transformation, and furthermore the submaximal eigenpair for Hermitizable matrices. The algorithms can also be used to compute a few number of the other subsequent eigenpairs. The price we have to pay is mainly for the Householder transformation (Algorithm 3) which is a famous algorithm having complexity $O(N^3)$. The other algorithms in the paper are mainly

Received July 26, 2020; accepted December 4, 2020
Corresponding author: Mu-Fa CHEN, E-mail: mfchen@bnu.edu.cn
2000 *Mathematics Subject Classifications.* 15A18(A57), 60J27, 65F10 (F15, F30).
Key words and phrases. Hermitizable, Householder transformation, birth–death matrix, isospectral matrices, top eigenpairs, algorithm.

$O(N)$ algorithm. In Section 4 of the paper, except some remarks on our algorithms, a proof of a key result, an isospectral property of the Hermitizable matrix and a Jacobi (birth-death) one, originally given in [5], is presented. The last section of the paper is devoted to the practical implementation of the results obtained in the previous sections on large scale matrices. Some additional analysis and the programs in MatLab of the algorithms, as well as a number of tests in comparison with the known programs are presented.

1 Checking the Hermitizability

Let $A = (a_{ij})_{i,j=0}^{N}$ be a given complex matrix. We are going to check by computer its Hermitizability introduced in [5]: there exists a positive measure μ such that

$$\mu_i a_{ij} = \mu_j \bar{a}_{ji} \qquad \forall i, j, \tag{1}$$

where $\bar{a}$ denotes the conjugate of a. Note that we have a very simple necessary condition for the property (1): for each pair (i, j), either $a_{ij} = 0$ and $a_{ji} = 0$ simultaneously, or $a_{ij}a_{ji} > 0$ (cf. [5]). In particular, $(a_{ii})_{i=0}^{N}$ must be real. However, for the criterion of the Hermitizability, one more condition is essential: the so-called circle condition. The analytic method for checking the circle condition was given in [5; Theorem 5]. Here we introduce an algorithm for checking the condition by computer. Define a column vector $\mathbb{1}$ having elements 1 everywhere and denote by $\mathrm{Diag}(u)$ the diagonal matrix with vector u as its diagonal elements. For simplicity, let $B = A - \mathrm{Diag}\big(\bar{A}\mathbb{1}\big)$. Denote by $B^{\setminus\{\text{last line}\}}$ the matrix obtained from B by removing its last line.

The checking procedure consists of three steps.

Algorithm 1 (1) *Computing the harmonic measure.* Consider the (row-) harmonic equation: $\mu B = 0$ with $\mu_0 \neq 0$. Assume that there exists at least one non-zero solution μ. Equivalently, the equation

$$B^{*\setminus\{\text{last line}\}}\mu^* = 0$$

has at least one solution $(\mu_0, \cdots, \mu_N)$ with fixed boundary condition, say $\mu_0 = 1$ for instance. Actually, in the Hermitizable case, the resulting measure μ must be positive (and is indeed unique under the irreducible condition, cf. [5]), then we can go to the next step. Otherwise, the matrix A is not Hermitizable.

(2) Define the *quasi-Hermitizing matrix* as follows.

$$\hat{A} = \mathrm{Diag}(\mu^{1/2}) A \,\mathrm{Diag}(\mu^{-1/2}), \qquad \hat{a}_{ij} = \sqrt{\mu_i}\, a_{ij}/\sqrt{\mu_j} \qquad \forall\, i, j. \tag{2}$$

(3) *Hermitizability criterion.* Now, A is Hermitizable iff $\hat{A} = \hat{A}^H$, where the superscript H means the conjugate transpose.

The next example illustrates an application of Algorithm 1.

Example 2 ([6; Example 3]) Let

$$A = \begin{bmatrix} -2 & 2+2\,i & 1-i & 0 \\ 1/2-i/2 & -3 & 1-i/2 & 3+i \\ 1+i & 4+2\,i & -4 & 8+2\,i \\ 0 & 3-i & 2-i/2 & -5 \end{bmatrix}.$$

Then $\mu_1 = 1,\ \mu_2 = 4,\ \mu_3 = 1$, and $\mu_4 = 4$. Furthermore,

$$\mathrm{Diag}(\mu)^{1/2} A\, \mathrm{Diag}(\mu)^{-1/2} = \begin{bmatrix} -2 & 1+i & 1-i & 0 \\ 1-i & -3 & 2-i & 3+i \\ 1+i & 2+i & -4 & 4+i \\ 0 & 3-i & 4-i & -5 \end{bmatrix}$$

which is clearly Hermitian. Its eigenvalues are as follows.

$$-9.1026,\ -5.75255,\ 2.62816,\ -1.77301.$$

Proof. Note that

$$B = \begin{bmatrix} -3+i & 2+2\,i & 1-i & 0 \\ 1/2-i/2 & -9/2 & 1-i/2 & 3+i \\ 1+i & 4+2\,i & -13+5\,i & 8+2\,i \\ 0 & 3-i & 2-i/2 & -5-3\,i/2 \end{bmatrix}.$$

and then

$$B^{*\backslash\{\text{last line}\}} = \begin{bmatrix} -3+i & 1/2-i/2 & 1+i & 0 \\ 2+2\,i & -9/2 & 4+2\,i & 3-i \\ 1-i & 1-i/2 & -13+5\,i & 2-i/2 \end{bmatrix}.$$

Now, the conclusion follows from Algorithm 1. □

2 Reducing Hermite matrix to tridiagonal one

We have in the last section reduced the Hermitizable matrix to an isospectral Hermitian matrix $\hat{A}$ given in (2). In this section, we further reduce a Hermitian matrix to some isospectral symmetric tridiagonal matrix with nonnegative subdiagonal elements, in terms of Householder transformation. Thus, throughout this section, we fix a Hermitian matrix $A = (a_{ij})_{i,j=1}^{N}$. We are going to use some unitary similar transformation, making the off-tridiagonal elements to be zero. The algorithm is running column by column. Let $A_{k-1} = \left(a_{ij}^{(k-1)}\right)$ ($A_0 := A$) and $b^{(k)}$ be the kth column of A_{k-1} given in Fig. 1. Replacing the

first k components of $b^{(k)}$ by zero, we obtain the vector $x^{(k)}$. Next, define $y^{(k)}$ by the following procedure: replacing each component by 0 in $x^{(k)}$, except the element $b_{k+1}^{(k)}$ is replaced by $s_k := \sqrt{x^{(k)H} x^{(k)}}$.

$$b^{(k)} = \begin{bmatrix} b_1^{(k)} \\ \vdots \\ b_k^{(k)} \\ b_{k+1}^{(k)} \\ b_{k+2}^{(k)} \\ \vdots \\ b_N^{(k)} \end{bmatrix} \longrightarrow x^{(k)} := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b_{k+1}^{(k)} \\ b_{k+2}^{(k)} \\ \vdots \\ b_N^{(k)} \end{bmatrix} \longrightarrow y^{(k)} := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ s_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Figure 1 Construction of two vectors: $x^{(k)}$ and $y^{(k)}$

Algorithm 3 At the $k\,(\geqslant 1)$th step, suppose that the matrix obtained after $k-1$ transformations is A_{k-1}. Then, we want to transform $x^{(k)}$ into $y^{(k)} = (s_k \delta_{i,k+1:\,1\leqslant i\leqslant N})$ by a unitary transformation defined by using $x^{(k)}$ and $y^{(k)}$:

$$U_k = I + uu^H/\alpha, \qquad u := x^{(k)} - y^{(k)}, \quad \alpha := s_k\big(b_{k+1}^{(k)} - s_k\big),$$

or in pointwise form:

$$U_k(i,j) = \delta_{ij} + \frac{1}{s_k\big(b_{k+1}^{(k)} - s_k\big)}\big(x_i^{(k)} - s_k\delta_{i,k+1}\big)\big(\bar{x}_j^{(k)} - s_k\delta_{j,k+1}\big).$$

Furthermore, we obtain the transformed matrix A_k at step k:

$$A_k = U_k A_{k-1} U_k^H.$$

Note that in the special case that $s_k = 0$ or $s_k = b_{k+1}^{(k)} > 0$, the U_k defined above is meaningless, we can simply ignore this step (or reset $U_k = I$) and jump to the next step at $k+1$. At the last step $k = N-1$ (at most), we obtain the required real symmetric tridiagonal matrix A_{N-1}.

We mention that the unitary matrix $I + uu^H/\alpha$ is Hermitian iff α is real, or equivalently, so is $b_{k+1}^{(k)}$. If $\alpha \neq 0$, then

$$\begin{aligned} u^H u &= \sum_{j\neq k+1} \bar{x}_j^{(k)} x_j^{(k)} + (\bar{x}_{k+1}^{(k)} - s_k)(x_{k+1}^{(k)} - s_k) \\ &= 2s_k^2 - s_k(x_{k+1}^{(k)} + \bar{x}_{k+1}^{(k)}) \\ &= 2s_k^2 - s_k(b_{k+1}^{(k)} + \bar{b}_{k+1}^{(k)}) \\ &= -(\alpha + \bar{\alpha}). \end{aligned}$$

Hence,

$$U_k^H U_k = I + \frac{uu^H}{\bar{\alpha}\alpha}(\alpha + \bar{\alpha} + u^H u) = I.$$

Equivalently, $U_k U_k^H = I$ in view of the operation $H : A \to A^H$. Thus U_k is surely unitary.

The following algorithm is for computing the maximal eigenvector $g_{\max}(A)$, which can be run in parallel to Algorithm 3 above.

Algorithm 4 Starting at $V_1 = U_1^H$, update V_j step by step in parallel to Algorithm 3:

$$V_k = V_{k-1} U_k^H, \qquad k = 2, 3, \ldots, N-1.$$

Denote by $(\lambda_{\max}(T),\ g_{\max}(T))$ the maximal eigenpair of $T := A_{N-1}$. Then the maximal eigenpair of A can be expressed by

$$(\lambda_{\max}(A),\ g_{\max}(A)) = (\lambda_{\max}(T),\ V_{N-1}\, g_{\max}(T)).$$

Similarly, one can compute the other eigenpairs of A using the ones of T with the same transform V_{N-1}.

Alternatively, $g_{\max}(A) =: g^{(0)}$ can be obtained by the following procedure:

$$g^{(k-1)} = U_k^H g^{(k)}, \qquad k = N-1, N-2, \ldots, 1.$$

The first method in Algorithm 4 does not need to store, step by step, the whole sequence $\{V_j\}_{j=1}^{N-1}$, but it requires about $N(N+1)(N+1/2)$ times of multiplications. Here is a careful analysis on the complexity of $g_{\max}(A)$ of the method. First, we compute the complexity of V_k. Note that $U_k = I + uu^H/\alpha$, we have $U_k^H = I + uu^H/\bar{\alpha}$. As usual, we count only the multiplications. Since at the kth step, the first k components of u are zero, and so are $u/\bar{\alpha}$ and u^H, it follows that

$$\begin{aligned} z &:= \frac{u}{\bar{\alpha}} \ \text{ requires } N-k \text{ times of multiplications} \\ Z &:= V_{k-1} z \ \text{ requires } N(N-k) \text{ times of multiplications} \\ Zu^H &\ \text{ requires } N(N-k) \text{ times of multiplications.} \end{aligned}$$

The last step needs a little explanation. As a product of the column vector Z and the row vector u^H, one often requires N^2 times of multiplications. Here the first k columns of the resulting matrix are zero and so can be ignored, since the first k components of u^H are zero. Hence the total multiplications are reduced to be $N(N-k)$ as given above. Thus, it means that $V_k = V_{k-1} U_k^H$ requires $(2N+1)(N-k)$ times of multiplications. Next, for k varying from 1 to $N-1$, we obtain V_{N-1}, which requires

$$\sum_{k=1}^{N-1}(2N+1)(N-k) = (2N+1)\left[N(N-1) - \frac{N(N-1)}{2}\right] = N(N-1)\left(N+\frac{1}{2}\right)$$

times of multiplications. Finally, for $g_{\max}(A) = V_{N-1}g_{\max}(T)$, it requires additionally N^2 times of multiplications. Therefore, for $g_{\max}(A)$, it requires totally

$$N(N-1)\left(N+\frac{1}{2}\right)+N^2 = N\left(N^2+\frac{N}{2}-\frac{1}{2}\right) = N(N+1)\left(N-\frac{1}{2}\right)$$

times of multiplications.

Comparing the first method just discussed above, the second one (given at the end of the algorithm) goes on the opposite direction: we have to store the sequence $\{x^{(j)}\}$ (or plus $\{s_j = \|x^{(j)}\|\}$) which generates the sequence $\{U_j\}$, but the iterative computations require only $N(N-1)/2$ multiplications. Thus, the second method is more effective than the first one, at least for large matrices.

We will come back to this topic in Algorithm 7 below.

The Householder transformation goes back to [10]. The representation here is taken from Wang [17] which is based on [16]. Since $s_k = y^{(k)H}u$, the expression of U_k here fits [8; p. 2375, the formula right above part III].

We now illustrate the algorithm by some examples.

Example 5 (Continued) Let A be the Hermitian matrix given at the end of Example 2. Then, we have

$$A_3 = \begin{bmatrix} -2 & 2 & & 0 \\ 2 & -\frac{5}{2} & \frac{\sqrt{67}}{2} & \\ & \frac{\sqrt{67}}{2} & -\frac{265}{134} & \frac{2\sqrt{7717}}{67} \\ 0 & & \frac{2\sqrt{7717}}{67} & -\frac{504}{67} \end{bmatrix}.$$

Notice that the sub-diagonal elements of the symmetric tridiagonal matrix A_3 are positive. We have thus reduced the computation of the maximal eigenpair of Hermitian A to the real tridiagonal one A_3. Furthermore, we have

$$\lambda_{\max}(A) = 2.62816$$
$$g_{\max}(A) = (.51569 + .137426\,i,\ 1.07178 + .0943814\,i,\ .969716 + .439587\,i,\ 1)^*.$$

Proof. At first step, we have

$$x^{(1)} = (0,\ 1-i,\ 1+i,\ 0)^*,$$
$$V_1 = U_1^H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2-i/2 & 1/2+i/2 & 0 \\ 0 & 1/2+i/2 & 1/2-i/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} -2 & 2 & 0 & 0 \\ 2 & -5/2 & 2+i/2 & 7/2+i/2 \\ 0 & 2-i/2 & -9/2 & 7/2+3\,i/2 \\ 0 & 7/2-i/2 & 7/2-3\,i/2 & -5 \end{bmatrix}.$$

At the second step, we have

$$x^{(2)} = (0,\ 0,\ 2-i/2,\ 7/2-i/2)^*,$$

$$V_2 = V_1 U_2^H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .5-.5\,i & .305424+.183254\,i & .106015+.601577\,i \\ 0 & .5+.5\,i & .183254-.305424\,i & .601577-.106015\,i \\ 0 & 0 & .855186-.122169\,i & -.38067-.329881\,i \end{bmatrix},$$

$$A_2 = \begin{bmatrix} -2 & 2 & & 0 \\ 2 & -\dfrac{5}{2} & \dfrac{\sqrt{67}}{2} & \\ & \dfrac{\sqrt{67}}{2} & -\dfrac{265}{134} & \dfrac{(81-34\,i)(\sqrt{67}-4+i)^2}{134\,(21-2\sqrt{67})} \\ 0 & & \dfrac{(81+34\,i)(\sqrt{67}-4-i)^2}{134\,(21-2\sqrt{67}\,)} & -\dfrac{504}{67} \end{bmatrix}.$$

Finally, at the third step, we have

$$x^{(3)} = \left(0,\ 0,\ 0,\ \frac{(81+34\,i)(\sqrt{67}-4-i)^2}{134\,(21-2\sqrt{67}\,)}\right)^*,$$

$$V_3 = V_2 U_3^H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .5-.5\,i & .305424+.183254\,i & .148807+.592445\,i \\ 0 & .5+.5\,i & .183254-.305424\,i & .592445-.148807\,i \\ 0 & 0 & .855186-.122169\,i & -.403308-.301785\,i \end{bmatrix},$$

and then A_3 given in the example.

Because $\lambda_{\max}(A_3) = 2.62816$ and

$$g_{\max}(A_3) = (1.60558,\ 3.71545,\ 3.87088,\ 1)^*.$$

We have $\lambda_{\max}(A) = 2.62816$ and

$$\begin{aligned} &g_{\max}(A) \\ &= V_3\, g_{\max}(A_3) \\ &= (.51569 + .137426\,i,\ 1.07178 + .0943814\,i,\ .969716 + .439587\,i,\ 1)^*. \end{aligned}$$

The conclusion is checked by

$$Ag_{\max}(A)/\lambda_{\max}(A) = g_{\max}(A). \quad \square$$

For the computation of the maximal eigenpair of tridiagonal matrix, refer to [6], and see Section 4 of the paper for analytic details. The next example shows a blocking phenomenon which seems not treated before carefully.

Example 6 Let

$$A = \begin{bmatrix} \frac{732}{289} & -\frac{81}{289} + \frac{27\,i}{289} & -\frac{50}{289} - \frac{50\,i}{289} & -\frac{70}{289} - \frac{60\,i}{289} \\ -\frac{81}{289} - \frac{27\,i}{289} & \frac{813}{289} & -\frac{20}{289} - \frac{40\,i}{289} & -\frac{30}{289} - \frac{50\,i}{289} \\ -\frac{50}{289} + \frac{50\,i}{289} & -\frac{20}{289} + \frac{40\,i}{289} & \frac{648}{289} & \frac{91}{289} - \frac{7\,i}{289} \\ -\frac{70}{289} + \frac{60\,i}{289} & -\frac{30}{289} + \frac{50\,i}{289} & \frac{91}{289} + \frac{7\,i}{289} & \frac{41}{17} \end{bmatrix}.$$

Then the deduced tridiagonal matrix is divisible:

$$A_3 = \begin{bmatrix} \frac{732}{289} & \frac{3\sqrt{2310}}{289} & & 0 \\ \frac{3\sqrt{2310}}{289} & \frac{713}{289} & & \\ & & \frac{64}{27} & \frac{\sqrt{170}}{27} \\ 0 & & \frac{\sqrt{170}}{27} & \frac{71}{27} \end{bmatrix}.$$

Proof. At the first step, we have

$$\begin{aligned} x^{(1)} &= \left(0,\ -\frac{81}{289} - \frac{27\,i}{289},\ -\frac{50}{289} + \frac{50i}{289},\ -\frac{70}{289} + \frac{60\,i}{289}\right)^*, \\ V_1 &= U_1^H \\ &= -\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .561769 + .187256\,i & .254965 + .418919\,i & .373346 + .519098\,i \\ 0 & .346771 - .346771\,i & -.84819 + .018202\,i & .199173 + .00848164\,i \\ 0 & .485479 - .416125\,i & .195533 + .0388436\,i & -.741923 + .0309434\,i \end{bmatrix}, \end{aligned}$$

$$A_1 = \begin{bmatrix} \frac{732}{289} & \frac{3\sqrt{2310}}{289} & & 0 \\ \frac{3\sqrt{2310}}{289} & \frac{713}{289} & & \\ & & \frac{64}{27} & \frac{13}{27} - \frac{i}{27} \\ 0 & & \frac{13}{27} + \frac{i}{27} & \frac{71}{27} \end{bmatrix}.$$

The second step can be ignored since for which we have $x^{(2)} = 0$. Then we have $U_2 = I$ and so $V_2 = V_1$. We now go to the third step.

$$x^{(3)} = (0,\ 0,\ 0,\ 13/27 + i/27)^*,$$

$$V_3 = V_2 U_3^H$$

$$= - \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .561769 + .187256\,i & .254965 + .418919\,i & .332433 + .546204\,i \\ 0 & .346771 - .346771\,i & -.84819 + .018202\,i & .197936 + .0237325\,i \\ 0 & .485479 - .416125\,i & .195533 + .0388436\,i & -.742111 - .0260506\,i \end{bmatrix},$$

and A_3 given in the example. Clearly, the matrix A_3 can be reduced to two 2×2 matrices and so it has two repeated pairs of eigenvalues $\{3, 2\}$. The reason is as follows. First, for an irreducible (or unreduced) tridiagonal matrix, its eigenvalues are distinct. This classical result is included in many textbooks, see for instance [1; p.97, Theorem 3.3], or [11; p.36, Theorem 2.2], or [14; p.134, Lemma 7.7.1]), or [18; pages 300–302]. In this case, the block decomposition for the matrix can be ignored. Hence, A_3 should have multiple eigenvalues and the multiplicity should be less than or equal to 2. Otherwise, there would have more blocks, not only two. If there are three distinct eigenvalues, then there would have two submatrices with size 3×3 and 1×1, respectively. Hence we are not in this situation. The conclusion can be easily checked by computing the eigenvalues of these two 2×2 submatrices separately.

To compute the maximal eigenvector of A in the above example. Let

$$A_3^{(1)} = \frac{1}{289} \begin{bmatrix} 732 & 3\sqrt{2310} \\ 3\sqrt{2310} & 713 \end{bmatrix}, \qquad A_3^{(2)} = \frac{1}{27} \begin{bmatrix} 64 & \sqrt{170} \\ \sqrt{170} & 71 \end{bmatrix}.$$

Then, with the same maximal eigenvalue 3, the maximal eigenvectors for them are

$$g_3^{(1)} = \left(\frac{1}{3}\sqrt{\frac{154}{15}},\ 1 \right)^* \text{ and } g_3^{(2)} = \left(\sqrt{\frac{10}{17}},\ 1 \right)^*,$$

respectively. Thus, the matrix A_3 has the maximal eigenvalue 3 with multiplicity 2 and independent eigenvectors as follows.

$$g^{(1)} = \left(\frac{1}{3}\sqrt{\frac{154}{15}},\ 1,\ 0,\ 0\right)^* \text{ and } g^{(2)} = \left(0,\ 0,\ \sqrt{\frac{10}{17}},\ 1\right)^*.$$

By Algorithm 4, the similar assertion holds for the original A with independent eigenvectors

$$\begin{aligned} V_3 g^{(1)} &= (1.06805,\ -.561769 - .187256\,i,\ -.346771 + .346771\,i,\\ &\quad -.485479 + .416125\,i)^* \quad \text{and}\\ V_3 g^{(2)} &= (0,\ -.527982 - .8675\,i,\ .452596 - .0376928\,i,\ .592145 - .00374103\,i)^*, \end{aligned}$$

respectively. Clearly, these two vectors are linear independent. Let us mention that each N-dimensional Hermite matrix has N linear independent eigenvectors, and so does each Hermitizable one. □

It is the position to come back to the computation of the maximal eigenvector. From the last two examples, we have seen that the computation of the sequence $\{V_j\}$ costs heavier work (actually has a higher complexity). We now show that in some cases (the matrix has a smaller size or is rather sparse, for instance), it is possible to use directly the shift inverse iteration.

Algorithm 7 Let $A = (a_{ij})_{i,j=1}^N$ be a given Hermitizable matrix. Set $z = \lambda_{\max}(A) + \varepsilon$ with small $\varepsilon > 0$ (say 10^{-8} for instance) and choose a suitable vector w_0. For a given vector w (may have subscript), here we fix the normalization of w in terms of its first $w(1)$ or last components $w(N)$:

$$v = w/w(1) \quad \text{or} \quad w/w(N).$$

Now, for given $v := v_{k-1}$, the shift inverse iteration goes as follows. Let $w = w_k$ solve the equation

$$(zI - A)w = v.$$

Then define v_k as the normalization of w_k just defined. Continue the iterations until the solutions become the same up to six digits of precision.

The reason we add a small constant ε in Algorithm 7 is to avoid the singularity of the matrix $zI - A$. The main price we have to pay is for the linear equation involved in the algorithm. Thus, once there is an effective algorithm for solving the equation (when A is symmetric, or sparse, or having smaller size, for instance), the algorithm should work well.

Example 8 (Continued) We now apply Algorithm 7 to Example 6. Set $z = 3 + 10^{-8}$. We have chosen three initials for v_0:

(1) $(1, 1, 0, 0)^*$.

(2) $(0,0,1,1)^*$.
(3) $(1,1,1,1)^*$.

Since the eigenspace of the maximal eigenvalue 3 has dimension 2, there is some freedom for the solution of the algorithm. After one iteration, the outputs of the vector v_1 in the corresponding cases are as follows:

(1) $(1, 1.73539 - 1.01172\,i, -.44239 + 1.3965\,i, -.714757 + 1.77121\,i)^*$.
(2) $(-.588235 - .504202i, -.252101 - .420168i, .764706 - .0588235i, 1)^*$.
(3) $(-.577029 - .140796i, -.0872077 - .970879i, .764706 - .0588235i, 1)^*$.

It is checked by the eigenequation $Av_1/3 = v_1$ so that these three solutions are all the eigenvectors of A. The first two are quite different from those produced by V_3 as seen above. Each pair of them are linear independent.

Originally, the next example was designed as a complex matrix. Since it is too big for the output, we choose a real one. However, the story is very much similar.

Example 9 Due to the multiplicity of the eigenvalues, the tridiagonalized matrix of the next symmetric one

$$A = \begin{bmatrix}
\frac{8074}{2601} & \frac{271}{5202} & \frac{373}{1734} & \frac{746}{2601} & \frac{1355}{5202} & \frac{169}{867} & \frac{676}{2601} & -\frac{959}{5202} \\
\frac{271}{5202} & \frac{31483}{10404} & \frac{373}{3468} & \frac{373}{2601} & \frac{1355}{10404} & \frac{169}{1734} & \frac{338}{2601} & -\frac{959}{10404} \\
\frac{373}{1734} & \frac{373}{3468} & \frac{2787}{1156} & \frac{475}{867} & \frac{1865}{3468} & \frac{271}{578} & \frac{542}{867} & -\frac{245}{3468} \\
\frac{746}{2601} & \frac{373}{2601} & \frac{475}{867} & \frac{7102}{2601} & \frac{1865}{2601} & \frac{542}{867} & \frac{2168}{2601} & -\frac{245}{2601} \\
\frac{1355}{5202} & \frac{1355}{10404} & \frac{1865}{3468} & \frac{1865}{2601} & \frac{37987}{10404} & \frac{845}{1734} & \frac{1690}{2601} & -\frac{4795}{10404} \\
\frac{169}{867} & \frac{169}{1734} & \frac{271}{578} & \frac{542}{867} & \frac{845}{1734} & \frac{1223}{289} & \frac{268}{867} & -\frac{1673}{1734} \\
\frac{676}{2601} & \frac{338}{2601} & \frac{542}{867} & \frac{2168}{2601} & \frac{1690}{2601} & \frac{268}{867} & \frac{11476}{2601} & -\frac{3346}{2601} \\
-\frac{959}{5202} & -\frac{959}{10404} & -\frac{245}{3468} & -\frac{245}{2601} & -\frac{4795}{10404} & -\frac{1673}{1734} & -\frac{3346}{2601} & \frac{46123}{10404}
\end{bmatrix}$$

has a block decomposition with three blocks. Moreover, the distinguished eigenvalues are all included in the largest block of the decomposition.

Proof. At Steps 1–3, the outputs are standard. Here is the output in Step 3.

$$x^{(3)} = (0,\ 0,\ 0,\ .122312,\ -.565458,\ -.20606,\ -.274746,\ -.781817)^*,$$

$$A_3 = \begin{bmatrix}
3.10419 & .581915 & 0 & 0 & & & & \\
.581915 & 6.43398 & 1.2267 & 0 & & & 0 & \\
0 & 1.2267 & 3.6022 & 1.03145 & & & & \\
0 & 0 & 1.03145 & 2.85964 & & & & \\
 & & & & 2.72146 & -.20848 & -.277973 & -.283237 \\
 & & & & -.20848 & 3.48396 & -.688054 & -.211994 \\
 & 0 & & & -.277973 & -.688054 & 3.08259 & -.282659 \\
 & & & & -.283237 & -.211994 & -.282659 & 2.71199
\end{bmatrix}.$$

Note that here and in what follows, once the output is less than 10^{-14}, we simply reset it as zero, because it usually comes from the computational errors. Step 4 can be omitted since $x^{(4)} = 0$. Here is the output of Step 5.

$$x^{(5)} = (0,\ 0,\ 0,\ 0,\ 0,\ -.20848,\ -.277973,\ -.283237)^*,$$

$$A_5 = \begin{bmatrix} 3.10419 & .581915 & & & & & & \\ .581915 & 6.43398 & 1.2267 & & & & & \\ & 1.2267 & 3.6022 & 1.03145 & & & & \\ & & 1.03145 & 2.85964 & 0 & & & \\ & & & 0 & 2.72146 & .448281 & & \\ & & & & .448281 & 2.27854 & 0 & \\ & & & & & 0 & 3.88097 & .323827 \\ & & & & & & .323827 & 3.11903 \end{bmatrix}.$$

We do not need to move further since A_5 has already arrived at the required tridiagonal form. Actually, it reduces into three submatrices:

$$A_5^{(1)} = \begin{bmatrix} 3.10419 & .581915 & & 0 \\ .581915 & 6.43398 & 1.2267 & \\ & 1.2267 & 3.6022 & 1.03145 \\ 0 & & 1.03145 & 2.85964 \end{bmatrix},$$

$$A_5^{(2)} = \begin{bmatrix} 2.72146 & .448281 \\ .448281 & 2.27854 \end{bmatrix},$$

$$A_5^{(3)} = \begin{bmatrix} 3.88097 & .323827 \\ .323827 & 3.11903 \end{bmatrix}.$$

Clearly, the eigenvalues of $A_5^{(2)}$ and $A_5^{(3)}$ are all included in those of $A_5^{(1)}$. Thus, in the case we are only interested in distinct eigenvalues, ignoring their multiplicity, it suffices to study the eigenvalues of the largest submatrix $A_5^{(1)}$. More explicitly, the eigenvalues of $A_5^{(1)}$, $A_5^{(2)}$ and $A_5^{(3)}$ are $\{7, 4, 3, 2\}$, $\{3, 2\}$ and $\{4, 3\}$, respectively. Corresponding to each distinct eigenvalue, one eigenvector can be deduced from those of the largest $A_5^{(1)}$. Note that the eigenvalue 3 has multiplicity 3, this is the reason why we have 3 submatrices and also the reason we can omit several steps in the computation. □

In the past two examples, we have seen that the Householder transformation transfers an irreducible matrix into a reducible tridiagonal one. This is reasonable as explained before, due to the multiplicity of eigenvalues. Note that the Householder transformation is a unitary one and hence keeps the

Hermitizability (symmetrizability), and furthermore the spectrum. By [5; Corollary 6], for the Hermitizability of a tridiagonal matrix, the matrix can be reducible. This property is remarkable since then for which, the Perron–Frobenius theorem may not be true. Therefore, the same theorem may also not true for Hermitian matrix.

To conclude this section, we mention that in practice, for improving the efficiency of computations, one may adopt some artistic design for Algorithms 3 and 4. See for instance [15; pages 106–108]. See also [15; pages 582 and 583, in particular], for some analysis and algorithms on Householder transformation. Besides, refer to [13] for concurrent algorithms.

3 Sub-maximal eigenpair

In this section, we introduce two approaches to compute the submaximal (or its next) eigenpair. For the first one, we need the following result.

Lemma 10 Let A be Hermitian. Denote by (λ_0, g_0) its maximal eigenpair $(\lambda_{\max}, g_{\max})$ with $\lambda_0 > 0$ and $g_0^H g_0 = 1$. Define

$$A_1 = (I - g_0 g_0^H)A. \tag{3}$$

Then A_1 is also Hermitian.

Proof. We need to prove that $A_1 = A_1^H$. Equivalently,

$$g_0 g_0^H A = A^H g_0 g_0^H.$$

First, we have

$$\begin{aligned} g_0 g_0^H A &= g_0 g_0^H A^H \quad (\text{since } A = A^H) \\ &= \bar{\lambda}_0 g_0 g_0^H \quad (\text{since } Ag_0 = \lambda_0 g_0 \Rightarrow g_0^H A^H = \bar{\lambda}_0 g_0^H) \\ &= \lambda_0 g_0 g_0^H \quad (\text{since the spectrum of } A \text{ is real}). \end{aligned}$$

Next,

$$\begin{aligned} A^H g_0 g_0^H &= A g_0 g_0^H \quad (\text{since } A = A^H) \\ &= \lambda_0 g_0 g_0^H \quad (\text{since } Ag_0 = \lambda_0 g_0). \end{aligned}$$

We have thus proved the required assertion. □

We now consider a simple example. Let

$$Q = \begin{bmatrix} -1 & 1 & & & & & & \\ 2 & -3 & 1 & & & & 0 & \\ & 2 & -3 & 1 & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & \ddots & \ddots & & 1 & & \\ & 0 & & 2 & & -3 & 1 & \\ & & & & & 2 & -3 & \end{bmatrix} \in \mathbb{R}^8 \times \mathbb{R}^8.$$

Since the symmetrizing measure of Q is $(\mu_k = 2^{-k+1},\ k = 1, \cdots, 8)$, we have

$$Q^{\text{sys}} = \text{Diag}(\mu)^{1/2} Q\, \text{Diag}(\mu)^{-1/2}$$
$$= \begin{bmatrix} -1 & \sqrt{2} & & & & & & \\ \sqrt{2} & -3 & \sqrt{2} & & & 0 & & \\ & \sqrt{2} & -3 & \sqrt{2} & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & \ddots & \ddots & \sqrt{2} & & & \\ & 0 & & \sqrt{2} & -3 & \sqrt{2} & & \\ & & & & \sqrt{2} & -3 & & \end{bmatrix} \in \mathbb{R}^8 \times \mathbb{R}^8.$$

Example 11 Define $A = Q^{\text{sys}} + 3\,I$. Then the eigenvalues of A are as follows.

$$\lambda_{\max} = 2.99799,\ -2.63352,\ 2.50514,\ -2.07511,$$
$$1.79552,\ -1.22867,\ .847221,\ -.208572$$

and the maximal eigenvector g_0 is as follows.

$$g_{\max} = (.715152,\ .504673,\ .354704,\ .247264,$$
$$.169471,\ .111997,\ .0679521,\ .0320544)^*.$$

We are now going to study the sub-maximal eigenpair of A. Actually, there are at least two approaches to do the work: the deflation technique and the optimal search approach, to be studied below.

Deflation technique

First, we use the known deflation technique introduced in [12; Theorem 2.2]. That is, studying the matrix A_1 defined by (3). With g_0 just obtained above, by (3), the matrix A_1 takes the following form

$$\begin{bmatrix} .466701 & .332185 & -.760492 & -.530139 & -.363349 & -.240124 & -.145691 & -.0687252 \\ .332185 & -.763572 & .877545 & -.374111 & -.25641 & -.169452 & -.102812 & -.0484984 \\ -.760492 & .877545 & -.377192 & 1.15127 & -.180215 & -.119098 & -.0722602 & -.0340866 \\ -.530139 & -.374111 & 1.15127 & -.183296 & 1.28859 & -.083023 & -.0503726 & -.0237618 \\ -.363349 & -.25641 & -.180215 & 1.28859 & -.0861034 & 1.35731 & -.0345246 & -.0162859 \\ -.240124 & -.169452 & -.119098 & -.083023 & 1.35731 & -.0376049 & 1.3914 & -.0107628 \\ -.145691 & -.102812 & -.0722602 & -.0503726 & -.0345246 & 1.3914 & -.0138432 & 1.40768 \\ -.0687252 & -.0484984 & -.0340866 & -.0237618 & -.0162859 & -.0107628 & 1.40768 & -.00308039 \end{bmatrix}$$

which is symmetric by Lemma 10. The eigenvalues of A_1 are as follows.

$$-2.63352,\ 2.50514,\ -2.07511,\ 1.79552,$$
$$-1.22867,\ 0.847221,\ -0.208572,\ -1.33596 \cdot 10^{-15}.$$

Hence $\lambda_{\max}(A_1) = 2.50514$. In other words, the submaximal eigenvalue of A now becomes the maximal one of A_1 (cf. [12]). We remark that at this point, the condition $\lambda_{\max}(A_1) > 0$ is necessary. Otherwise, a shift of A is required.

By Lemma 10, we now come back to the starting point treated in §2. By using the Householder transformation, it follows that A_1 is isospectral to the following tridiagonal matrix

$$T = \begin{bmatrix} .466701 & 1.08875 & & & & & & \\ 1.08875 & .852645 & 1.49569 & & & & & \\ & 1.49569 & -.727517 & 1.21056 & & & & \\ & & 1.21056 & .38967 & 1.44937 & & & \\ & & & 1.44937 & -.592352 & 1.12888 & & \\ & & & & 1.12888 & -.147293 & 1.25861 & \\ & & 0 & & & 1.25861 & -1.0249 & .333798 \\ & & & & & & .333798 & -.214945 \end{bmatrix}.$$

We have thus reduced our problem to the setup treated in [6].

From the discussion above, it is clear that the Householder transformation can also be used to study the submaximal eigenpair, provided the given matrix is not sparse. However, at the moment, we are not in such a case. Hence it is too heavy to use the transformation again for computing the submaximal eigenpair, since then the sparse property of A is lost. Fortunately, we have a different approach given in the next part.

Optimal search approach

Actually, in the present situation, there is an easier way to handle with this problem, which uses the main advantage of the (symmetric) tridiagonal matrix. Following [1; p. 146], we adopt the method of bisection. To fix the notation, as before, let $A \sim (a_k, c_k, b_k)_{k=1}^N$. Throughout the remainder of this section, assume that A is irreducible (i.e., $a_{k+1}b_k > 0$).

Method of bisection

To state the method, we need some notation. Define

$$s_1(\alpha) = c_1 - \alpha$$

$$s_2(\alpha) = \begin{cases} c_2 - \alpha - b_1^2/s_1(\alpha) & \text{if } s_1(\alpha) \neq 0 \\ -\varepsilon\,(0 < \varepsilon \ll 1) & \text{if } s_1(\alpha) = 0, \end{cases}$$

$$s_k(\alpha) = \begin{cases} c_k - \alpha - b_{k-1}^2/s_{k-1}(\alpha) & \text{if } s_{k-1}(\alpha) \neq 0 \text{ and } s_{k-2}(\alpha) \neq 0 \\ c_k - \alpha & \text{if } s_{k-2}(\alpha) = 0 \\ -\varepsilon\,(0 < \varepsilon \ll 1) & \text{if } s_{k-1}(\alpha) = 0. \end{cases}$$

Next, define

$$\gamma(\alpha) = \#\{s_k(\alpha) \geqslant 0, \;\; k = 1, \cdots, N\}.$$

Then, by [1; p. 142, (2.16) and Theorem 2.1; p.146, (2.23) and the remark below it], we have

Lemma 12 The number of eigenvalues of A on $[\alpha, \infty)$ equals $\gamma(\alpha)$.

The application of the bisection method, given in the remainder of this section, is mainly based on Lemma 12. To state an algorithm, we need a little preparation. Define

$$\eta = \max\Big\{\min_{1\leqslant i\leqslant N}(c_i - a_i - b_i),\ \min_{1\leqslant i\leqslant N}(c_i - a_{i+1} - b_{i-1})\Big\}, \qquad b_1 := 0,\ a_{N+1} := 0.$$

In the present setup, there are N distinct eigenvalues, listed as

$$\lambda_1 > \lambda_2 > \cdots > \lambda_N.$$

By Gershgorin Circle Theorem for the spectral radius, we have $\lambda_N \geqslant \eta$. To study the top eigenpairs of A, we start at the top eigenvalues. The computation goes one by one of the eigenvalues. The computation of (λ_1, g_1) was done in [6]. Starting from which, we study the subsequent top eigenvalues. For instance, based on λ_1 and η, we can compute λ_2 by using the bisection method. The algorithm given below consists of two parts. The first one is constructing an initial interval (β_2, β_1) for the bisection method. The second one is the standard sequent search of the method.

Algorithm 13 Let η be defined as above. Suppose that $\lambda_{k-1}\,(k \geqslant 2)$ is given and we are going to compute λ_k.

Step 1 Define

$$\begin{aligned}
\xi_0 &= \lambda_{k-1},\\
\xi_1 &= \frac{1}{N-k+1}\big[(N-k)\xi_0 + \eta\big],\\
\xi_m &= \max\{3\,\xi_{m-1} - 2\,\xi_{m-2},\ \eta\}, \qquad m \geqslant 2.
\end{aligned}$$

Compute $\gamma(\xi_m)$ successively until for the first $m = m_0$ such that $\gamma\big(\xi_{m_0}\big) \geqslant k$. Then take $[\beta_2, \beta_1] = [\xi_{m_0}, \xi_{m_0-1}]$ as the initial interval for using the method of bisection.

Step 2 Set $z = (\beta_2 + \beta_1)/2$ and compute $\gamma(z)$. We update the test interval by making the following changes, step by step. In details, if $\gamma(z) \geqslant k$, then replace β_2 by z; and otherwise replace β_1 by z.

$\gamma(z)$	Change
$\geqslant k$	$\beta_2 \to z$
$= k-1$	$\beta_1 \to z$

Repeating the recursive procedure until the outputs are the same up to six digits of precision, and γ takes value k at the final tested point.

Because

$$2^{-24} \approx 5.96046 \times 10^{-8},$$

for the six/seven digits of precision, the method of bisection requires about 24 times of steps (tests), independent of the matrix size N. Of course, at each step, in computing $\gamma(\alpha)$, it requires N times of computations. Hence, the complexity for using the method of bisection is $O(N)$.

The second (submaximal) eigenpair

(a) We now return to the matrix A given by Example 11. Keeping $\lambda_{\max}(A) = 2.99799$ in mind and using the method of bisection, after 16 steps, the outputs are the same up to six digits of precision, we obtain the submaximal eigenvalue ≈ 2.50514.

(b) Next, we compute the submaximal eigenvector using the shift inverse iteration. Recall that in the present tridiagonal situation, we have the more or less explicit Thomas algorithm for solving the required linear equation (cf. [6]). By (a), we can fix the shift to be $z = 2.50514 - \varepsilon$ ($\varepsilon = 10^{-8}$ for instance). Here and in what follows, the small modification ε is for avoiding the degeneration in using the shift inverse iteration, and also for avoiding the next eigenvalue. Note that the maximal eigenvector of A is

$$g_{\max} = (22.3106,\ 15.7443,\ 11.0657,\ 7.71389,\ 5.28698,\ 3.49398,\ 2.1199,\ 1)^*.$$

Its normalized vector is

$$g = (.715152, .504673, .354704, .247264, .169471, .111997, .0679521, .0320544)^*.$$

Our initial v_0 is chosen to be

$$\begin{aligned} v_0 &:= g - \mathbb{1}/(g^*\mathbb{1}) \ \ (\text{then } v_0 \in \operatorname{Span}(g)^\perp) \\ &= (.261281,\ .0508014,\ -.0991675,\ -.206607,\ -.284401,\ -.341874, \\ &\qquad -.385919,\ -.421817)^*. \end{aligned}$$

After one iteration, the output of the computation is as follows:

$$\begin{aligned} g_2 := (&.341037,\ \ .121815,\ \ -.125255,\ \ -.343691, \\ &-.483561,\ \ -.512889,\ \ -.424972,\ \ -.239907)^*. \end{aligned}$$

This is checked by using the second iteration, its output is exactly the same. A simpler way to check this conclusion is simply using the eigenequation: $Ag_2 = \lambda_2 g_2$, where $\lambda_2 = 2.50514$. Therefore, g_2 can be regarded as the submaximal eigenvector as we required.

The third (next to the submaximal) eigenpair

To conclude this section, we remark that the same method can also be used to compute the subsequent eigenpairs. Here we mention shortly the computation for the next to the submaximal eigenpair for the same example as above. Rewrite $\lambda_1 = \lambda_{\max}(A)$. Then, we have known the eigenpairs (λ_1, g_1) and (λ_2, g_2). We are now going to compute the next one (λ_3, g_3). By using the method of bisection above, we obtain $\lambda_3 = 1.79552$. Now, to apply the shift inverse iteration, choose shift $z = 1.79552 - \varepsilon$. For the initial vector v_0, we choose the form

$$v_0 = (1,\ x,\ x,\ x,\ y,\ y,\ y,\ 1.1)^*$$

with x and y determined by the conditions $v_0 \perp g_1$ and $v_0 \perp g_2$:

$$x = -.753323, \qquad y = .238242.$$

Here, the components 1 and 1.1 in v_0 are chosen randomly. If these random numbers are replaced by x and y, respectively, then the homogeneous equations have only trivial solution $x = y = 0$. In one step (iteration), we obtain the required eigenvector

$$g_3 = (-.350163,\ .0506296,\ .414444,\ .475559,\ .189337,\ -.235171,\\ -.487917,\ -.3843)^*.$$

To conclude this section, we mention that Algorithm 13 can be naturally extended to concurrent computing, simply replacing the bisection method by the equisection one, which is a generalization of the method of bisection in the optimization theory.

4 Remarks and Proofs

This section is mainly devoted to the analytical aspect of the algorithms introduced in the previous sections. Besides, it also provides a detailed exploration on [5; Theorem 24]. At the end of this section, we present a simplified proof for a key result concerning with the isospectral property of a Hermitizable tridiagonal matrix and a birth–death one (see Theorem 14 below).

Remark on Algorithm 1

Before moving to the main text, let us make a remark on Algorithm 1 (1). Let $A = (a_{ij})$ be a given complex matrix. By (1), we have

$$\sum_i \mu_i a_{ij} = \mu_j \sum_i \bar{a}_{ji}.$$

Equivalently, $(\mu A)(j) = \big(\mu \mathrm{Diag}\,(\bar{A}\mathbb{1})\big)(j)$. From this, we obtain $\mu B = 0$ as stated in the algorithm.

Proof of Algorithm 4

Let $g^{(k)} \neq 0$ be the eigenvector of A_k $(0 \leqslant k < N)$ with $A_0 = A$, corresponding to a fixed eigenvalue λ:

$$A_k g^{(k)} = \lambda\, g^{(k)} \Longleftrightarrow \big(U_k A_{k-1} U_k^H\big) g^{(k)} = \lambda\, g^{(k)} \Longleftrightarrow A_{k-1}\big(U_k^H g^{(k)}\big) = \lambda\big(U_k^H g^{(k)}\big).$$

This implies that $g^{(k-1)} = U_k^H g^{(k)}$. Then the last assertion in the algorithm follows. Furthermore,

$$g^{(0)} = U_1^H g^{(1)} = U_1^H U_2^H g^{(2)} \cdots\cdots = U_1^H U_2^H \cdots U_{N-1}^H g^{(N-1)}.$$

In other words, the eigenvector $g = g^{(0)}$ of A corresponding to the eigenvalue λ can be expressed by

$$g = \Big(\prod_{k=1}^{N-1} U_k^H\Big) g^{(N-1)}.$$

Thus, the recursive formula of $\{V_k\}$ given in Algorithm 4 is an alternative algorithm of the product above:

$$V_1 = U_1^H,\ V_2 = U_1^H U_2^H = V_1 U_2^H,\ V_{N-1} = \prod_{k=1}^{N-1} U_k^H = V_{N-2} U_{N-1}^H. \quad \square$$

Remark on Algorithm 13

First, we explain the main idea in the special case that $k = 2$. By assumption, $\xi_0 = \lambda_1$ and $\lambda_N \geqslant \eta$ give us the upper and lower bounds of the eigenvalues $\{\lambda_j\}_{j=1}^N$, respectively. The left-endpoint ξ_1 of the test interval should be an approximation of λ_2, even a rough approximation is still okay since the convergence of the bisection method is quite fast. Actually, a little smaller one is better since then we can ignore a subinterval on the left. If so, we do not need the double extension of the test interval on the left-hand side. For this, we may assume that $\lambda_N = \eta$. Then, there are exact $N-2$ eigenvalues located inside of the interval $[\eta,\ \xi_0]$. Suppose that these eigenvalues are located on $N-2$ equal points. Because the length of each equal division is

$$\frac{1}{N-1}(\xi_0 - \eta).$$

Hence, the submaximal eigenvalue λ_2 should be located around

$$\xi_0 - \frac{1}{N-1}(\xi_0 - \eta) = \frac{1}{N-1}\big((N-2)\xi_0 + \eta\big).$$

That is the ξ_1 given in the algorithm corresponding to $k = 2$. Very often, ξ_1 is enough as the left-end point of the initial interval. If not, we need the

subsequence $\{\xi_m\}_{m\geqslant 2}$. To which, the construction is as follows. Suppose that we are at ξ_{m-1}. Then choose ξ_m so that

$$\xi_{m-1} - \xi_m = 2(\xi_{m-2} - \xi_{m-1}).$$

This gives us ξ_m as in part (1) of the algorithm, in terms of the natural control by η. The factor 2 used above is optimal, simply based on the method of bisection.

For general $k \geqslant 2$, suppose that λ_{k-1} is known and we are going to compute λ_k. In this case, the first $k-2$ eigenvalues $\lambda_1, \ldots, \lambda_{k-2}$ play no role. The number of the eigenvalues decreases. Therefore, we need to make the following change:

$$N \to N - k + 2, \quad \lambda_1 \to \lambda_{k-1},$$

in the formula just obtained for $k = 2$. We have thus arrived at the expression of ξ_1 presented in part (1) of the algorithm. The change from $k = 2$ to general $k \geqslant 2$ stated in part (2) of the algorithm is now obvious.

Proof of reduction to tridiagonal matrix

We now come back to the main text. In the matrix form, the Hermitizability of A can be expressed as

$$\mathrm{Diag}\,(\mu)\,A = A^H\,\mathrm{Diag}\,(\mu) \Longleftrightarrow \mathrm{Diag}\,(\mu)\,A\,\mathrm{Diag}\,(\mu)^{-1} = A^H \tag{4}$$

(actually, it is better to rewrite $\mathrm{Diag}\,(\mu)^\alpha$ as $\mathrm{Diag}\,(\mu^\alpha)$ in computation). As used in parts (2) and (3) in Algorithm 3, it is easy to see that

$$A \text{ is Hermitizable } \Longleftrightarrow H := \mathrm{Diag}\,(\mu^{1/2})\,A\,\mathrm{Diag}\,(\mu^{-1/2}) \quad \text{is Hermitian.} \tag{5}$$

The assertion is clearly important since then every property and algorithm for the Hermitian matrix can be transferred to the Hermitizable one.

Next, for the Hermitian H, as shown in §2, there exists a unitary matrix U, as a product of some Householder transformations $\{U_k\}$ (unitary), such that

$$T := UHU^H \text{ becomes a tridiagonal, real and symmetric matrix.} \tag{6}$$

As also shown in §2, the tridiagonal matrix T can be blocked. In other words, T may be reducible, it can be divided into several irreducible blocks, say $T = \mathrm{Diag}(\{T_j\})$, where T_j is irreducible tridiagonal, real, symmetric matrix. In general, the sum of some row of T can be positive and so is not a Q-matrix. Let $m = \sup_k(T\mathbb{1})(k) < \infty$ (which may be a condition when the matrix is infinite). Then the sum of each row of $T - m\,I$ is not positive, so is each row of $T_j - m\,I$. Fix j for a moment. By [5; Theorems 15 and 16], there exists a positive nearly $(T_j - m\,I)$-harmonic function h_j such that

$$Q_j := \mathrm{Diag}(h_j^{-1})(T_j - m\,I)\mathrm{Diag}(h_j)$$

is a birth–death Q-matrix. In particular, $T_j - m\,I$ and Q_j are isospectral. To return to the original setup, combining the family $\{h_j\}$ into a single vector h with the same ordering as the blocking of $\{T_j\}$. Then, we can combine the family $\{Q_j\}$ into a single one with the same ordering just mentioned, denoted by $Q = \mathrm{Diag}(\{Q_j\})$. Furthermore, from the last formula, we obtain

$$Q = \mathrm{Diag}(h^{-1})(T - m\,I)\mathrm{Diag}(h).$$

Equivalently,

$$\mathrm{Diag}(h^{-1})T\,\mathrm{Diag}(h) = Q + m\,I =: Q^{(m)}. \tag{7}$$

Combining (5)–(7) together, we obtain the following similar transformation of A:

$$\mathrm{Diag}\,(h^{-1})\,U\,\mathrm{Diag}\,(\mu^{1/2})\,A\,\mathrm{Diag}\,(\mu^{-1/2})\,U^H\,\mathrm{Diag}\,(h) = Q^{(m)}.$$

Set

$$M = \mathrm{Diag}\,(\mu^{-1/2})\,U^H\,\mathrm{Diag}\,(h) \Longleftrightarrow M^{-1} = \mathrm{Diag}\,(h^{-1})\,U\,\mathrm{Diag}\,(\mu^{1/2}). \tag{8}$$

It follows that

$$M^{-1}AM = Q^{(m)} \Longleftrightarrow A = MQ^{(m)}M^{-1}.$$

Therefore,

$$Ag = \lambda\,g \Leftrightarrow MQ^{(m)}M^{-1}g = \lambda\,g \Leftrightarrow Q^{(m)}\,(M^{-1}g) = \lambda\,(M^{-1}g).$$

By [5; Theorem 10], with the inner product

$$\langle f, g\rangle := (Mf, Mg)_\mu$$

and $\tilde{f} := M^{-1}f$, $L^2(E,\mu) \to L^2(E, \langle\cdot,\cdot\rangle)$, we have

$$(f,\ g)_\mu = \langle \tilde{f},\ \tilde{g}\rangle \quad \text{and} \quad (Af, g)_\mu = \langle Q^{(m)}\tilde{f}, \tilde{g}\rangle. \tag{9}$$

For this specific M defined by (8), we have here more explicit formulation. Note that

$$\begin{aligned}(Mf, Mg)_\mu &= \big(\mathrm{Diag}\,(\mu^{-1/2})\,U^H\,\mathrm{Diag}\,(h)f,\ \mathrm{Diag}\,(\mu^{-1/2})\,U^H\,\mathrm{Diag}\,(h)g\big)_\mu\\ &= \big(U^H\,\mathrm{Diag}\,(h)f,\ U^H\,\mathrm{Diag}\,(h)g\big)_{\mathrm{d}x}\\ &= \big(\mathrm{Diag}\,(\bar{h})\mathrm{Diag}\,(h)f, g\big)_{\mathrm{d}x} \quad \big(\text{Since } UU^H = I\big)\\ &= (f,\ g)_{\tilde{\mu}},\end{aligned}$$

where $\tilde{\mu} = |h|^2\mathrm{d}x$ and $\mathrm{d}x$ means the uniform measure in the discrete case: $\mu_k \equiv 1$. More precisely, here in the first equality above, we have used the fact that

$$\begin{aligned}\big(\mathrm{Diag}\,(\mu^{-1/2})\,f,\ \mathrm{Diag}\,(\mu^{-1/2})\,g\big)_\mu &= \big(\mathrm{Diag}\,(\mu^{1/2})f,\ \mathrm{Diag}\,(\mu^{-1/2})\,g\big)_{\mathrm{d}x}\\ &= g^H\mathrm{Diag}\,(\mu^{-1/2})\,\mathrm{Diag}\,(\mu^{1/2})\,f\\ &= (f,\ g)_{\mathrm{d}x},\end{aligned}$$

and in the second equality, we have used the fact that

$$(U^H f,\ U^H g)_{\mathrm{d}x} = g^H U U^H f = (f,\ g)_{\mathrm{d}x}.$$

Hence, we have

$$\langle \tilde{f},\ \tilde{g}\rangle = \big(M\tilde{f}, M\tilde{g}\big)_{\mu} = \big(\tilde{f},\ \tilde{g}\big)_{\tilde{\mu}}.$$

Therefore, we indeed have $L^2(E, \langle\cdot,\cdot\rangle) = L^2(E, \tilde{\mu})$. Thus, for the mapping $\tilde{f} := M^{-1}f$, $L^2(E,\mu) \to L^2(E,\tilde{\mu})$, by (9), we have the isometry:

$$(f,g)_{\mu} = \big(\tilde{f},\ \tilde{g}\big)_{\tilde{\mu}},$$

and furthermore, the isospectral property:

$$(Af, g)_{\mu} = \big\langle Q^{(m)}\tilde{f},\ \tilde{g}\big\rangle = \big(Q^{(m)}\tilde{f},\ \tilde{g}\big)_{\tilde{\mu}}.$$

Here is our final conclusion.

Theorem 14 For given Hermitizable A, define M by (8). Then the mapping $\tilde{f} := M^{-1}f$ from the complex $L^2(E,\mu)$ to the real $L^2(E,\tilde{\mu})$ is an isometry: $(f,g)_{\mu} = \big(\tilde{f},\ \tilde{g}\big)_{\tilde{\mu}}$. Furthermore, it owns the isospectral property: $(Af,g)_{\mu} = \big(Q^{(m)}\tilde{f},\ \tilde{g}\big)_{\tilde{\mu}}$.

Remark on the computational complexity

Now we mention the computational complexity of the algorithms used in the paper. The quasi-Hermitizing procedure in (2) requires $2N^2$ multiplications. The computation for the maximal eigenpair of the tridiagonal matrix using the method given in [6], as well as the one of bisection plus Thomas' algorithm for computing the other eigenpairs requires only $O(N)$ multiplications. The main work we need is Householder transformation which requires $2N^3/3$ multiplications. Refer to [18; page 244].

5 Practical implementation on large scale matrices

In this section, we present some practical implementation of our algorithms on large scale matrices. For counting the number of computational operations, the additions and the multiplications are all collected together, and denote it by "flops". All experiments were performed by MatLab on a personal computer with the configuration: Intel(R) Xeon(R) CPU E5-2630 v3 @2.40 GHz and 32 GB of RAM.

Householder-based tridiagonalization

Assuming that the Hermitized $\hat{A}$ of the Hermitizable matrix A is at hand, we are going to propose the details of the Householder tridiagonalization of the Hermitian matrix $\hat{A}$.

We use the following notation: Let N, k ($\leqslant N$) and j ($\leqslant N$) be given three positive integers. Denote by I_N the identity matrix of order N, and by e_k the kth column of the identity matrix. Next, for given $0 < k < N$, an N-dimensional vector x and an $N \times N$ matrix A, set

- $x(k+1:N)$: the vector of order $N-k$ obtained from x by deleting its first k entries;
- $a(k+1,j)$: the entry of A at the $(k+1)$th row and the jth column;
- $a(k+1:N,j)$: the vector of order $N-k$ obtained from A by deleting its first k entries of the jth column;
- $A(k+1:N,k+1:N)$: the $(N-k)\times(N-k)$ submatrix of A by deleting its first k rows and columns.

Suppose that Householder matrices $U_1,\cdots,U_{k-1}$ have been determined such that if

$$\hat{A}_{k-1} = (U_{k-1}\cdots U_1)\hat{A}(U_{k-1}\cdots U_1)^H,$$

then

$$\hat{A}_{k-1} = \begin{bmatrix} B_{11} & B_{12} & 0 \\ B_{21} & B_{22} & B_{23} \\ 0 & B_{32} & B_{33} \end{bmatrix},$$

where $B_{11} \in \mathbb{C}^{(k-1)\times(k-1)}$, $B_{12} \in \mathbb{C}^{(k-1)\times 1}$, $B_{21} = B_{12}^H$, $B_{22} \in \mathbb{C}$, $B_{23} \in \mathbb{C}^{1\times(N-k)}$, $B_{32} = B_{23}^H$, and $B_{33} = B_{33}^H \in \mathbb{C}^{(N-k)\times(N-k)}$. The kth Householder transformation is computed by

$$\widehat{U}_k = I_{N-k} + vv^H/\alpha, \qquad \alpha = s_k(b_{k+1}^{(k)} - s_k), \tag{10a}$$

$$0 \neq v = x^{(k)}(k+1:N) - y^{(k)}(k+1:N). \tag{10b}$$

Define

$$U_k = \begin{bmatrix} I_k & 0 \\ 0 & \widehat{U}_k \end{bmatrix}.$$

Then

$$\hat{A}_k = U_k\hat{A}_{k-1}U_k^H = \begin{bmatrix} B_{11} & B_{12} & 0 \\ B_{21} & B_{22} & B_{23}\widehat{U}_k^H \\ 0 & \widehat{U}_kB_{32} & \widehat{U}_kB_{33}\widehat{U}_k^H \end{bmatrix}.$$

Notice that it is not necessary to compute $B_{23}\widehat{U}_k^H$ and $\widehat{U}_kB_{32}$, since

$$\widehat{U}_kB_{32} = \|B_{32}\|_2 e_1 = s_k e_1 \in \mathbb{R}^{N-k}, \qquad B_{23}\widehat{U}_k^H = (\widehat{U}_kB_{32})^H.$$

Thus the leading k-by-k principal submatrix of $\hat{A}_k$ is a tridiagonal matrix. In the calculation of $\hat{A}_k$, it is important to exploit the Hermitian structure during the formation of the matrix. Note that

$$\begin{aligned}\widehat{U}_k B_{33} \widehat{U}_k^H &= \left(I_{N-k} + \frac{vv^H}{\alpha}\right) B_{33} \left(I_{N-k} + \frac{vv^H}{\alpha}\right)^H \\ &= B_{33} + \frac{vv^H}{\alpha} B_{33} + \frac{B_{33} vv^H}{\bar{\alpha}} + \frac{vv^H}{\alpha} B_{33} \frac{vv^H}{\bar{\alpha}}.\end{aligned}$$

Replacing the last term on the right-hand side by the sum of half and half of it, we obtain

$$\begin{aligned}\widehat{U}_k B_{33} \widehat{U}_k^H &= B_{33} + v\left(\frac{v^H B_{33}}{\alpha} + \frac{v^H B_{33} v}{2\alpha\bar{\alpha}} v^H\right) + \left(\frac{B_{33} v}{\bar{\alpha}} + v \frac{v^H B_{33} v}{2\alpha\bar{\alpha}}\right) v^H \\ &= B_{33} + vw^H + wv^H, \end{aligned} \tag{11}$$

where

$$w = \frac{B_{33} v}{\bar{\alpha}} + v \frac{v^H B_{33} v}{2\alpha\bar{\alpha}}.$$

Since only the upper triangular portion of this matrix needs to be calculated, the transition from $\hat{A}_{k-1}$ to $\hat{A}_k$ can be accomplished in only about $4(N-k)^2$ flops.

Algorithm 15 (Householder-based Tridiagonalization) Given a Hermitian matrix $\hat{A} \in \mathbb{C}^{N \times N}$, the following algorithm overwrites $\hat{A}$ with $T = U\hat{A}U^H$, where T is tridiagonal and $U = U_1 \cdots U_{N-1}$ is the product of Householder transformations.

Step 1 **for** $k = 1 : N-1$

Step 2 $x = \hat{a}(k+1 : N, k)$, $s = \sqrt{x^H x}$

Step 3 $v = x$, $v(1) = v(1) - s$

Step 4 $\alpha = s(x(1) - s)$

Step 5 $p = \hat{A}(k+1 : N, k+1 : N)v/\bar{\alpha}$

Step 6 $w = p + (p^H v/(2\bar{\alpha}))v$

Step 7 $\hat{a}(k+1, k) = s$; $\hat{a}(k, k+1) = \hat{a}(k+1, k)$

Step 8 $\hat{A}(k+1 : N, k+1 : N) = \hat{A}(k+1 : N, k+1 : N) + vw^H + wv^H$

Step 9 **end**

The transition from $\hat{A}_{k-1}$ to $\hat{A}_k$ totally costs about $\sum_{k=1}^{N-2} 4(N-k)^2$ flops. This is $O(N^3)$ flops. This is implemented in Algorithm 15 by

$$\hat{A}(k+1 : N, k+1 : N) = \hat{A}(k+1 : N, k+1 : N) + vw^H + wv^H.$$

This main step makes the total cost of Algorithm 15 to be $O(N^3)$. This algorithm requires $4N^3/3$ flops when Hermitian is exploited in calculating the rank-2 updating as in equation (11). The matrix U is stored in factored form in the subdiagonal portion of $\hat{A}$. If U is explicitly required, then it can be formed with an additional $4N^3/3$ flops.

Remark that the Householder transformation here defined by (10) is different to the traditional one given in [9, Pages 234-237, 243-244],

$$\widehat{U}_k = I_{N-k} - \beta v v^H, \qquad \beta = 2/(v^H v), \tag{12a}$$

$$0 \neq v = x \pm e^{i\theta}\|x\|_2 e_k, \quad x = x^{(k)}(k+1:N). \tag{12b}$$

The differences stay at two aspects: (a) β in (12) is always set to be real, while α in (10) may be real or complex; (b) the resulted vector of (12) is real or complex, while the resulted vector of (10) is set real. If we use alternatively the traditional Householder transformation (12) in Algorithm 15, then we get a complex tridiagonal matrix. Thus, we need to generate not more than $N-1$ necessary rotations to transform conjugate complex (not real) entries on upper and down subdiagonals to real numbers.

Example 16 In this example, we reduce a Hermitian matrix to a real symmetric tridiagonal matrix by Algorithm 15. We compare the Householder transformation (10) with the traditional Householder transformation (12), denoted by HR and HC, respectively. The testing matrix is produced by the command "rand" in MatLab and the elements of the matrices are chosen from $[0, 10]$. More precisely, let $A_1 = 10 \cdot \texttt{rand(N)}$ and $A_2 = 10 \cdot \texttt{rand(N)}$, where N is the matrix size. Then we take

$$A = \frac{A_1 + A_1^T}{2} + i \cdot \frac{A_2 - A_2^T}{2}.$$

Proof. We apply the Householder-based tridiagonalization on A in three cases: $N = 1200$, 2500, and 5000. For each case, the whole procedure is carried out for 100 times and the average values of results are output. The numerical results are given in Tables 1 and 2. Note that the output display format is chosen as the scaled fixed point format with 5 digits. For instance, 4.5937e+2 represents 4.5937×10^2. Table 1 presents the *average* CPU times to work out the Householder transformation. The symbols "CPU-U" and "CPU" refer to the CPU times in seconds with and without constructing the unitary matrix U, respectively. Both of them are tested by the MatLab command `cputime`. Table 2 displays the accuracy of the methods, in which we adopt some *average* errors:

- err-U: $\|UU^H - I\|_\infty / \|U\|_\infty$,
- err-UAU-T: $\|UAU^H - T\|_\infty / \|A\|_\infty$,
- err-A-UTU: $\|U^H T U - A\|_\infty / \|A\|_\infty$,

where U is a unitary matrix arising in the Householder transformation and I is the identity matrix. Here, for vector $x = [x_1, x_2, \cdots, x_N]^*$ and matrix $A = [a_{ij}] \in \mathbb{C}^{N\times N}$, we define

$$\|x\|_\infty = \max_{i=1,\cdots,N} |x_i| \quad \text{and} \quad \|A\|_\infty = \max_{i=1,\cdots,N} \sum_{j=1}^{N} |a_{ij}|.$$

Table 1: Comparison of HC and HR with respect to the CPU time for Example 16.

N	CPU-U		CPU	
	HC	HR	HC	HR
1200	2.8727e+2	2.8064e+2	1.1235e+2	1.1122e+2
2500	2.8840e+3	2.8689e+3	1.1577e+3	1.1514e+3
5000	2.1321e+4	2.1264e+4	9.2617e+3	9.2376e+3

Table 2: Comparison of HC and HR with respect to the accuracy for Example 16.

N	err-U		err-UAU-T		err-A-UTU	
	HC	HR	HC	HR	HC	HR
1200	2.5599e-15	2.5140e-15	4.1728e-14	3.9616e-14	2.3152e-14	2.0504e-14
2500	3.3557e-15	3.2936e-15	7.9696e-14	7.4699e-14	4.6707e-14	3.8897e-14
5000	4.4415e-15	4.4061e-15	1.5283e-13	1.4406e-13	9.0779e-14	7.7541e-14

From the numerical results in Table 1 and Table 2, we see that the performances of HR and HC are comparable with each other. HR saves a little bit of CPU times costed by HC. For instance, when $N = 1200$ HR saves 2.31% CPU times of computing T and U together, and 1.00% CPU times of computing T, respect to HC. Two average residual errors of HR are smaller than those of HC. Taking $N = 1200$ for instance, HR improve "err-U", "err-UAU-T" and "err-A-UTU" by 1.79%, 5.06% and 11.4%, respectively. These numerical results successfully indicate the high efficiency of the proposed Householder transformation (HR), which directly transforms a Hermitian matrix into a real tridiagonal matrix with nonnegative subdiagonals. □

Top k eigenpairs

Now we concentrate on computing top k eigenpairs of large scale matrices. After Hermitizing and tridiagonalizing, we reduce a Hermitizable matrix A to an isospectral symmetric tridiagonal matrix T. The core work of computing top k eigenpairs of A becomes the calculation of top k eigenpairs of T.

Note that for any tridiagonal matrix T, we can find a shift m such that the diagonal entries of $T - mI$ are negative. Without loss of generality, we build our algorithm on the following irreducible tridiagonal matrix,

$$T = \begin{bmatrix} -c_0 & b_0 & & & & 0 \\ a_1 & -c_1 & b_1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & b_{N-1} & \\ 0 & & & a_N & -c_N & \end{bmatrix}, \tag{13}$$

where the sequences $\{b_j\}_{j=0}^{N-1}$ and $\{a_j\}_{j=1}^{N}$ are positive and $\{c_j\}_{j=0}^{N}$ is nonnegative. Let

$$E = \{j \in \mathbb{Z} : \ 0 \leqslant j < N+1\} \ (N \leqslant \infty).$$

We may write $T \sim (a_j, -c_j, b_j)$ for simplicity. If $a_j = b_{j-1}$ $(j \in E\backslash\{0\})$, then T is symmetric.

We now present a new two-stage method of computing top k eigenpairs of the irreducible tridiagonal matrix T. In the first stage, we compute the largest eigenpair, denoted by (λ_1, g_1), of T, for instance, applying Algorithm 1 in [6]. In the second stage, we compute the other top $k-1$ eigenpairs, denoted by $(\lambda_2, g_2), \ \cdots, \ (\lambda_k, g_k)$, applying Algorithm 13 in this paper and the inverse iteration. These two stages are not separated strictly, but are jointed tightly. For convenience, we gather above steps into Algorithm 17 with a subroutine in Algorithm 18.

Algorithm 17 (Computing Top k Eigenpairs) Suppose $T \sim (a_j, -c_j, b_j)$ is an irreducible tridiagonal matrix of the form (13). The following algorithm computes top k eigenpairs of T, denoted by $(\lambda_1, g_1), \ \cdots, \ (\lambda_k, g_k)$.

Step 1 Let $a_0 = 0, \ b_N = 0,$

$$m = \sup_{j \in E}(a_j + b_j - c_j)^+, \qquad x^+ = \max\{x, \ 0\},$$

and

$$u_j = a_j b_{j-1}, \qquad j \in E\backslash\{0\}.$$

Step 2 *Specific Isospectral transformation.* Set $\widetilde{c}_j = c_j + m$ $(j \in E)$ and $\widetilde{b}_0 = \widetilde{c}_0$. Let

$$\widetilde{b}_j = \widetilde{c}_j - \frac{u_j}{\widetilde{b}_{j-1}}, \qquad \widetilde{a}_j = \widetilde{c}_j - \widetilde{b}_j, \qquad 1 \leqslant j < N,$$

$$\widetilde{a}_N = \frac{u_N}{\widetilde{b}_{N-1}}.$$

Then the tridiagonal matrix

$$\widetilde{T} \sim (\widetilde{a}_j, -\widetilde{c}_j, \widetilde{b}_j)$$

possesses the properties: both $(\widetilde{a}_j)$ and $(\widetilde{b}_j)$ are positive, the sum of each row equals zero except the $(N+1)$ th row $(\widetilde{c}_N \geqslant \widetilde{a}_N)$.

Step 3 *Symmetrizing.* Define the symmetric tridiagonal matrix

$$T^{\mathrm{sym}} \sim (a_j^{\mathrm{sym}}, -c_j^{\mathrm{sym}}, b_j^{\mathrm{sym}})$$

as follows:

$$c_j^{\mathrm{sym}} = \widetilde{c}_j \ (j \in E), \qquad a_j^{\mathrm{sym}} = b_{j-1}^{\mathrm{sym}} = \sqrt{u_j} \quad (j \in E \backslash \{0\}).$$

Step 4 *Computing the maximal eigenpair.* If $\widetilde{c}_N = \widetilde{a}_N$, then T^{sym} has the maximal eigenvalue $\lambda_1^{\mathrm{sym}} = 0$ with eigenvector $g_1^{\mathrm{sym}} = \sqrt{\widetilde{\mu}}$:

$$\widetilde{\mu}_0 = 1, \qquad \widetilde{\mu}_j = \widetilde{\mu}_{j-1} \frac{\widetilde{b}_{j-1}}{\widetilde{a}_j}, \qquad j \in E \backslash \{0\}. \tag{14}$$

More economically,

$$g_1^{\mathrm{sym}}(0) = 1, \ g_1^{\mathrm{sym}}(j) = g_1^{\mathrm{sym}}(j-1) \frac{\tilde{b}_{j-1}}{\sqrt{u_j}}, \qquad j \in E \setminus \{0\}. \tag{15}$$

Otherwise, set $\widetilde{b}_N = \widetilde{c}_N - \widetilde{a}_N$. The jth approximation of the maximal eigenpair is computed by Algorithm 18. Then $(-z_j, v^{(j)})$ converges to the maximal eigenpair of T^{sym}:

$$\lambda_1^{\mathrm{sym}} = -\lim_{j \to \infty} z_j, \qquad g_1^{\mathrm{sym}} = \lim_{j \to \infty} v^{(j)}.$$

Step 5 *Computing the subsequent eigenpairs.* Compute $k-1$ eigenvalues, $\lambda_2^{\mathrm{sym}}, \ \cdots, \ \lambda_k^{\mathrm{sym}}$, of T^{sym} by the bisection method (Algorithm 13).

For each j $(2 \leqslant j \leqslant k)$, compute the eigenvector g_j^{sym} corresponding to λ_j^{sym} of T^{sym}, by the inverse iteration with the shift λ_j^{sym} and an initial vector $v^{(0)}$. The initial vector $v^{(0)}$ is generated as follows: choose a vector $x^{(j)} \notin \mathtt{span}\{g_1^{\mathrm{sym}}, \cdots, g_{j-1}^{\mathrm{sym}}\}$, and compute $v^{(0)}$:

$$w^{(0)} = x_j - \sum_{i=1}^{j-1} [(g_i^{\mathrm{sym}})^H x_j] g_i, \qquad v^{(0)} = \frac{w^{(0)}}{\sqrt{(w^{(0)})^* w^{(0)}}}.$$

The modified version of the Gram-Schmidt method [9, Pages 254-255] is used in the practical implementation.

Step 6 *Returning to the original top k eigenpairs.* To go back to the original matrix T, the top k eigenpairs are

$$\lambda_j = \lambda_j^{\mathrm{sym}} + m, \qquad g_j = \mathrm{diag}(h^{\mu}) g_j^{\mathrm{sym}},$$

where $\mathrm{diag}(h^\mu)$ is the diagonal matrix having diagonal elements (h_j^μ):

$$h_0^\mu = 1, \qquad h_j^\mu = h_{j-1}^\mu \frac{\sqrt{u_j}}{b_{j-1}}, \qquad j \in E\backslash\{0\}. \tag{16}$$

Algorithm 18 (The jth approximation of the maximal eigenpair) With computed $\widetilde{a}_j$'s and $\widetilde{b}_j$'s, this algorithm computes the jth approximation of the maximal eigenpair: $(z_j,\ v^{(j)})$.

Step 1 Define the upper triangle matrix (M_{ij}) and the vector (Φ_i) as follows:

$$M_{ii} = 1, \quad M_{ij} = M_{i,j-1}\frac{\widetilde{a}_j}{\widetilde{b}_{j-1}} = M_{i,j-1}\frac{u_j}{\widetilde{b}_{j-1}^2}, \qquad 1 \leqslant i+1 \leqslant j \leqslant N,$$

$$\Phi_i = \sum_{i \leqslant j \leqslant N} \frac{M_{ij}}{\widetilde{b}_j}, \qquad 0 \leqslant i \leqslant N.$$

Step 2 Choose

$$w^{(0)} = \sqrt{\Phi}, \qquad v^{(0)} = \frac{w^{(0)}}{\sqrt{w^{(0)*} w^{(0)}}}.$$

Step 3 For a computed vector $v^{(j)}$ $(j \geqslant 0)$, compute ζ_j:

$$\zeta_j = \sup_{0 \leqslant n \leqslant N} \frac{1}{\sqrt{\widetilde{b}_n} v_n^{(j)} - \sqrt{\widetilde{a}_{n+1}} v_{n+1}^{(j)}} \left(\sum_{i=0}^{n} v_i^{(j)} \sqrt{\frac{M_{in}}{\widetilde{b}_i}} \right), \qquad j \geqslant 0,$$

with assuming $\widetilde{a}_{N+1} = 0$.

Step 4 With setting $z_j = \frac{1}{\zeta_j}$, solve $w^{(j+1)}$:

$$(-T^{\mathrm{sym}} - z^{(j)} I) w^{(j+1)} = v^{(j)}.$$

Step 5 Compute $v^{(j+1)}$:

$$v^{(j+1)} = \frac{w^{(j+1)}}{\sqrt{w^{(j+1)*} w^{(j+1)}}}.$$

Remark that if $\widetilde{c}_N = \widetilde{a}_N$, the way of computing the maximal eigenpair, $(\lambda_1,\ g_1)$, in Algorithm 17, is different from that in Algorithm 1 in [6]. But their results are the same: if $\widetilde{c}_N = \widetilde{a}_N$, then Algorithm 17 shows that T has the maximal eigenvalue $\lambda_1 = m$ with eigenvector $g_1 = h$:

$$h_0 = 1, \qquad h_j = h_{j-1} \frac{\widetilde{b}_{j-1}}{b_{j-1}}, \qquad j \in E\backslash\{0\}. \tag{17}$$

These two algorithms are coincided since

$$h_j^{\mu} = \frac{h_j}{\sqrt{\widetilde{\mu}_j}}, \qquad j \in E\backslash\{0\}.$$

With applying equations (16) and (17), one can simplify the computation of g_1^{sym} in (14) into the economical formula in (15).

One important point is that the non-symmetric matrix $\widetilde{T}$ and the symmetric matrix T^{sym} are coupled together in Algorithm 17. This coupling idea has been introduced in [5, Section 4] and [6, Section 4.4]. For a short explanation, the spectrum of $T - mI$ is transformed (in step 2) to the one of $\widetilde{T}$, which is a birth-death Q-matrix (see [6, Definition 7]); and then the non-symmetric matrix $\widetilde{T}$ is symmetrized to a symmetric matrix T^{sym} in step 3. The maximal pair is computed by coupling $\widetilde{T}$ and T^{sym}; see Algorithm 18 and step 4 of Algorithm 17. The subsequent eigenpairs are computed by applying Algorithm 13 and the iteration method on T^{sym}. We have two reasons: one is that T^{sym} is symmetric and has eigenvectors which are orthogonal to each other; another one is that T^{sym} has been generated during the computation of the maximal eigenpairs, and it does not rise extra computational flops. At last, the top k eigenpairs are generated for the original matrix T.

Clearly, we don't have to go to the last step 6 every time when using Algorithm 17. Actually, for very large N, this step is risky. Especially for non-symmetric matrices, the overflow happens very often. This is due to the limitation of the accuracy of the machine. At this time, if you still want to calculate, then one may use the above iterative formula (17). Because the sequences $\{\tilde{b}_k\}$ and $\{b_k\}$ are known, and so is the ratio $\tilde{b}_j/b_j$ which does not overflow (see [7] for more related details on Hermitizable complex tridiagonal matrices).

Example 19 In this example, we compute top k eigenpairs of a symmetric tridiagonal matrix of the form (13),

$$T \sim (a_j, -c_j, b_j) \in \mathbb{R}^{N\times N},$$

where a_j and c_j are random positive integers which are less than or equal to N, generated by the command "randi" in MatLab, and $b_{j-1} = a_j$ for $j = 1, \cdots, N$.

Proof. We apply Algorithm 17 and the MatLab functions, `eig` and `eigs`, to compute top k eigenvalues and corresponding eigenvectors of T, respectively. For comparison, we set $k = 3$, and $N = 5000,\ 10000,\ 15000,\ 20000$. We run the whole process of each case for 1000 times and take the average numerical results. Note that `eig` computes all eigenpairs, and `eigs` is called by the command `eigs(T,k,'LA')`. For the iterative methods (Algorithm 17 and `eigs`), the convergence tolerance is set as $\mathtt{tol} = 10^{-10}$ and the maximal number of iterations is 100.

Let V denote the matrix consisting of eigenvectors and D the diagonal matrix with eigenvalues on the diagonal. The numerical results are given in Table 3, in which "CPU" refers to the CPU time in seconds, "Res" represents the relative residual error $\|TV-VD\|_\infty/\|T\|_\infty$, and "NaN" refers to no answer.

Table 3: Numerical results of Example 19: CPU times and residual errors.

N	Algorithm 17		`eigs`		`eig`	
	CPU	Res	CPU	Res	CPU	Res
5000	2.2367	4.2206e-11	6.4821e-1	3.3668e-12	2.1670	6.5837e-15
10000	6.8966	7.7877e-11	9.3459e-1	3.4931e-12	5.1046	7.1151e-15
15000	7.9176e+1	5.8389e-11	2.4787	3.5428e-12	4.5517e+1	7.1183e-15
20000	3.6226e+1	9.0249e-11	3.3654e-1	NaN	2.2656e+1	6.7109e-15

From the numerical results, it follows that Algorithm 17 and the MatLab function `eig` correctly compute the top 3 eigenpairs in all cases, and they cost almost same CPU times. The MatLab function `eigs` successfully obtains the top 3 eigenpairs in the first three cases. But when $N = 20000$ it fails to converge at one time of 1000 tests and no result is obtained. Thus we can conclude that the proposed new method is more reliable than the MatLab function `eigs`. Moreover, Algorithm 17 also has the comparable level of accuracy with the MatLab function `eigs` in the first three cases. □

Example 20 In this example, we compute top k eigenpairs of a Hermitizable tridiagonal matrix of the form (13),

$$T \sim (a_j, -c_j, b_j) \in \mathbb{R}^{N\times N},$$

where $a_j \equiv a$, $b_j \equiv b$ and $c_j \equiv c$ are positive. We test the following two cases:

Case 1 $a = 2$, $b = 1$, and $c = 3$;
Case 2 $a = 1$, $b = 2$, and $c = 3$.

Proof. Let $(\lambda_j^{\text{exact}}, g_j^{\text{exact}})$ and (λ_j, g_j) denote the exact and computed eigenpairs of $T \sim (a, -c, b)$, respectively. They are defined by

$$\lambda_j^{\text{exact}} = 2\sqrt{ab}\cos\left(\frac{j\pi}{N+1}\right) - c,$$

$$g_j^{\text{exact}}(i) = \left(\sqrt{\frac{a}{b}}\right)^i \sin\left(\frac{ij\pi}{N+1}\right), \qquad i = 1, \cdots, N.$$

For comparison, we define the error of the computed eigenvalues by

$$\text{ERR} = \max_{j=1,\cdots,k} |\lambda_j - \lambda_j^{\text{exact}}|.$$

Let s_j denote the sign changing time of the jth eigenvector and the vector $\text{SCT} = [s_1, \cdots, s_k]$ denote the sign changing times of each of the top k eigenvalues. Remember that the jth exact eigenvector has $j - 1$ times of sign changing.

Table 4: Numerical results of Example 20: Errors of eigenpairs in **Case 1**.

N	Algorithm 17		`eigs`		`eig`	
	ERR	SCT	ERR	SCT	ERR	SCT
51	4.4409e-16	[0,1,2]	3.4233e-7	[0,1,2]	6.7230e-11	[0,1,2]
52	3.6082e-16	[0,1,2]	5.8039e-6	[0,1,2]	1.4893e-10	[0,1,2]
83	1.8874e-15	[0,1,2]	5.0430e-5	[0,1,2]	5.4506e-7	[0,1,2]
84	1.4710e-15	[0,1,2]	1.0336e-2	[0,1,1]	1.0819e-6	[0,1,2]
103	1.3323e-15	[0,1,2]	–	–	1.0916e-4	[0,1,2]
104	1.9706e-15	[0,1,2]	–	–	5.6538	[102,102,103]
10000	3.3307e-16	[0,1,2]	–	–	–	–
20000	1.6037e-13	[0,1,2]	–	–	–	–

Table 5: Numerical results of Example 20: Errors of eigenpairs in **Case 2**.

N	Algorithm 17		`eigs`		`eig`	
	ERR	SCT	ERR	SCT	ERR	SCT
44	1.6653e-16	[0,1,2]	1.8505e-7	[0,1,2]	4.1633e-16	[0,1,2]
45	2.7756e-16	[0,1,2]	5.6594e-6	[0,1,2]	6.6613e-16	[0,1,2]
104	1.9706e-15	[0,1,2]	1.0453e-3	[0,1,2]	1.0112e-11	[0,1,2]
105	1.8041e-15	[0,1,2]	7.7746e-3	[0,2,2]	3.3805e-11	[0,1,2]
106	1.5821e-15	[0,1,2]	–	–	1.2713e-11	[1,3,2]
160	9.9920e-16	[0,1,2]	–	–	3.2369e-9	[30,33,36]
161	7.4940e-16	[0,1,2]	–	–	5.6561	[160,160,157]
10000	3.3307e-16	[0,1,2]	–	–	–	–
20000	1.6037e-13	[0,1,2]	–	–	–	–

In the two cases listed above, we set $k = 3$ and apply Algorithm 17 and the MatLab functions, `eig` and `eigs`, to compute the top k eigenvalues and corresponding eigenvectors of T, respectively. Note that `eig` computes all eigenpairs and `eigs` is called by the command `eigs(T,k,'LR')`. For the iterative methods, Algorithm 17 and `eigs`, the convergence tolerance is set as `tol`$=10^{-8}$ and the maximal number of iterations is 300.

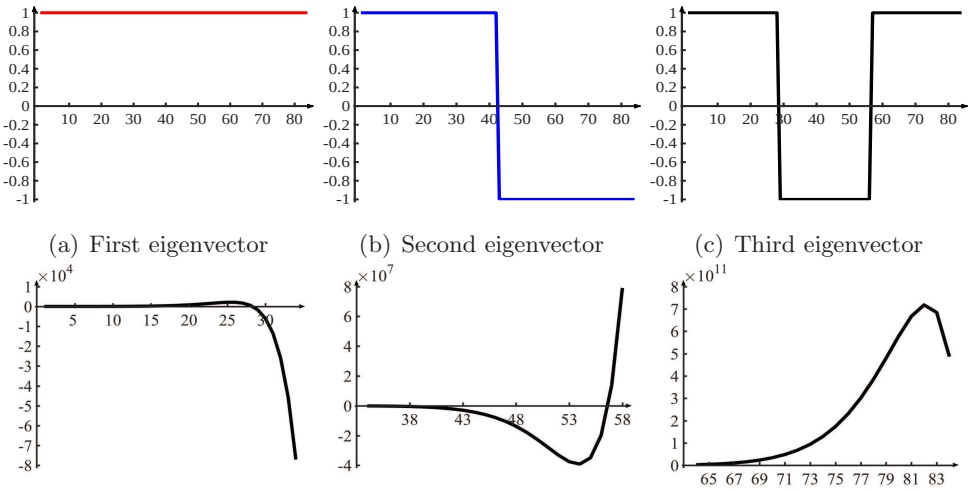
From the numerical results in Tables 4 and 5, we see that Algorithm 17 correctly computes the top k eigenpairs in all cases; but both `eig` and `eigs` fail in most of cases, even when the size of T is very small. We underline the

values of ERR which are larger than 10^{-6}, and also underline the values of SCTs if they are not correct. The notation "–" means that it is not necessary to compute.

Now we give a detailed analysis as follows.

- `Algorithm 17` computes correctly the top k eigenpairs in both **Case 1** and **Case 2**. The maximal errors of the computed top k eigenvalues are always less than 10^{-6}, and the times of sign changing of the computed eigenvectors are also correct, that is, SCT $= [0, 1, 2]$.
- `eigs` computes correctly the top k eigenpairs when $N \leqslant 51$ in **Case 1** (or $N \leqslant 44$ in **Case 2**). The maximal errors of the computed top k eigenvalues become larger than 10^{-6} when $N \geqslant 52$ in **Case 1** (or $N \geqslant 45$ in **Case 2**). Moreover, the times of sign changing of the computed eigenvectors become wrong when $N \geqslant 84$ in **Case 1** (or $N \geqslant 105$ in **Case 2**).
- `eig` computes correctly the top k eigenpairs when $N \leqslant 83$ in **Case 1** (or $N \leqslant 105$ in **Case 2**). The maximal errors of the computed top k eigenvalues become larger than 10^{-6} when $N \geqslant 84$ in **Case 1** (or $N \geqslant 161$ in **Case 2**). Moreover, the times of sign changing of the computed eigenvectors become wrong when $N \geqslant 104$ in **Case 1** (or $N \geqslant 106$ in **Case 2**).

For **Case 1** with $N = 84$, the signs of the top $k = 3$ exact and computed eigenvectors are shown in Figures 2-5. In Figure 2(a)-(c), the signs of the first, second and third maximal exact eigenvectors (from left to right) are drawn; and the times of sign changing are 0, 1 and 2, respectively. That is SCT $= [0, 1, 2]$. In Figure 2 (d)-(f), three parts of the third eigenvector are



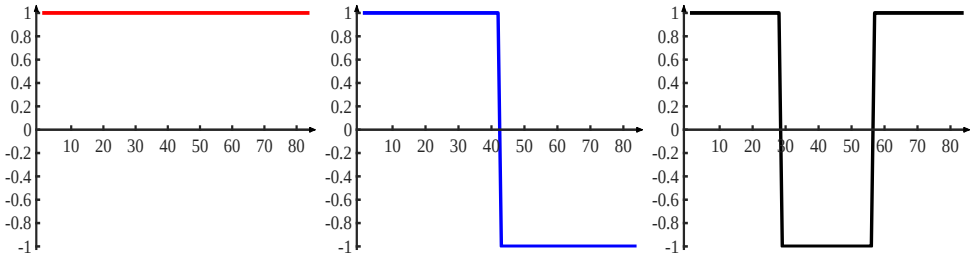
(a) First eigenvector (b) Second eigenvector (c) Third eigenvector

(d) First part from 1 to 34 (e) Second part from 34 to 58 (f) Third part from 64 to 84

Figure 2: (a)-(c): The sign of the top $k = 3$ exact eigenvectors of $T \sim (2, -3, 1)$ with $N = 84$. (d)-(f): Three parts of the third exact eigenvector.

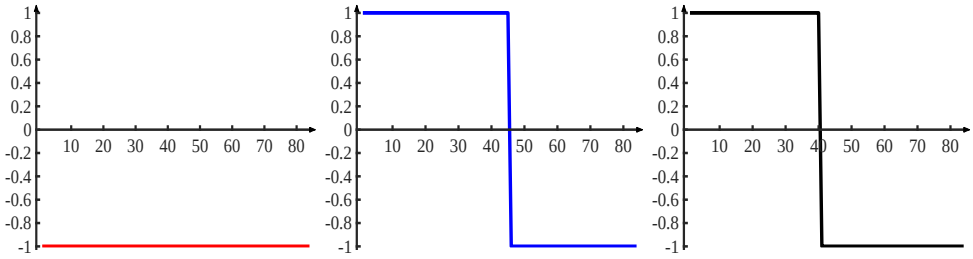
drawn. The two points at which the signs of the third eigenvector change are shown in Figure 2(d) and Figure 2(e).

Figures 3-5 indicate the signs of the top 3 eigenvectors, computed by Algorithm 17, `eigs` and `eig`, respectively; and the SCTs are $[0, 1, 2]$, $[0, 1, 1]$ and $[0, 1, 2]$. From Figure 4, it follows that the sign of the third eigenvector computed by `eigs` is not correct.



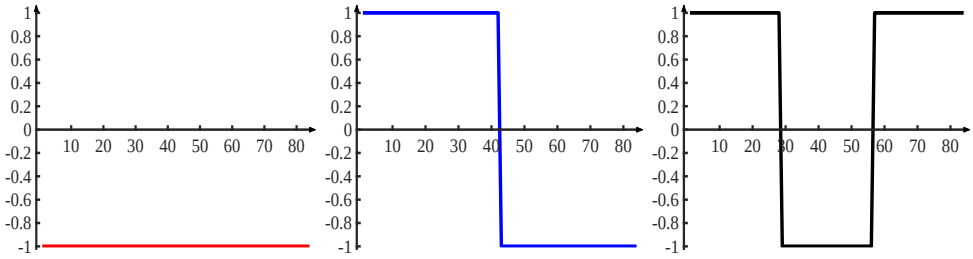
(a) First eigenvector (b) Second eigenvector (c) Third eigenvector

Figure 3: The sign of the top $k = 3$ eigenvectors computed by Algorithm 17 of $T \sim (2, -3, 1)$ with $N = 84$.



(a) First eigenvector (b) Second eigenvector (c) Third eigenvector

Figure 4: The sign of the top $k = 3$ eigenvectors computed by `eigs` of $T \sim (2, -3, 1)$ with $N = 84$.



(a) First eigenvector (b) Second eigenvector (c) Third eigenvector

Figure 5: The sign of the top $k = 3$ eigenvectors computed by `eig` of $T \sim (2, -3, 1)$ with $N = 84$.

As in Example 19, we list the CPU times ("CPU") and the relative residual errors ("Res") of **Case 1** and **Case 2** in Tables 6 and 7, respectively.

Comparing these two tables with Tables 4 and 5, respectively, it follows that the error estimates by using "ERR" plus "SCT" are more precise than using "Res". Note that the CPU times are obtained by running the programs for 100 times and taking the average values. The relative residual errors of Algorithm 17 are less than those of `eigs` and `eig`. Algorithm 17 costs less CPU times than `eigs` and has comparable speed with `eig`.

From above numerical results, we conclude that Algorithm 17 performs better than `eigs` and `eig`, and is very feasible and reliable to compute the top k eigenpairs of large scale matrices. □

Table 6: Numerical results of Example 20: CPU times and residual errors in **Case 1**.

N	Algorithm 17		`eigs`		`eig`	
	CPU	Res	CPU	Res	CPU	Res
51	1.5200e-2	2.7062e-16	3.0200e-2	3.2479e-11	4.3000e-3	1.1148e-15
52	8.9000e-3	3.1456e-16	4.6400e-2	7.6735e-11	4.7000e-3	1.1125e-15
83	1.5500e-2	6.0137e-16	6.9300e-2	3.4396e-11	1.5700e-2	1.5400e-15
84	1.5800e-2	6.1987e-16	4.7600e-2	1.9611e-10	1.8100e-2	1.6283e-15
103	3.0800e-2	7.0777e-16	–	–	2.4100e-2	2.6738e-15
104	3.4000e-2	6.5688e-16	–	–	2.7500e-2	7.5496e-15

Table 7: Numerical results of Example 20: CPU times and residual errors in **Case 2**.

N	Algorithm 17		`eigs`		`eig`	
	CPU	Res	CPU	Res	CPU	Res
44	1.2900e-2	1.7810e-16	2.8100e-2	3.1130e-12	3.6000e-3	1.8735e-15
45	8.9000e-3	3.1456e-16	4.6400e-2	7.6735e-11	4.7000e-3	1.1125e-15
104	3.4500e-2	6.8695e-16	9.9700e-2	7.5717e-11	2.4300e-2	3.2867e-15
105	3.0000e-2	6.1525e-16	6.7300e-2	2.5779e-11	2.0000e-2	1.7139e-15
106	3.2800e-2	6.2913e-16	–	–	2.6600e-2	1.5451e-15
160	3.9500e-2	4.1402e-16	–	–	1.2860e-1	1.7278e-15
161	3.9900e-2	4.0246e-16	–	–	1.2930e-1	2.5905e-15

Acknowledgements We are grateful to anonymous two referees for providing many useful comments and suggestions. Acknowledges are given to Ying-Chao Xie and Yue-Shuang Li for fruitful discussions and valuable suggestions. For the algorithms in the paper, a package in MatLab is preparing by Jia and Pang and should appear soon. Research supported in part by National Natural Science Foundation of

China (Grant Nos. 12090011, 11771046, 11771188, 11771189), National Key R & D Program of China (No. 2020YFA0712900), the Natural Science Foundation of Jiangsu Province (Grant No. BK20171162), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Cao, Z.H. (1983). *Matrix Eigenvalue Problem* (in Chinese). Shanghai Press of Sci. Tech.
[2] Chen, M.F. (2005) *Eigenvalues, Inequalities, and Ergodic Theory*. London: Springer.
[3] Chen, M.F. (2016). *Efficient initials for computing the maximal eigenpair*. Front. Math. China 11(6): 1379–1418.
[4] Chen, M.F. (2017). *Global algorithms for maximal eigenpair*. Front Math China 12(5): 1023–1043.
[5] Chen, M.F. (2018). *Hermitizable, isospectral complex matrices or differential operators*. Front Math China 13(6): 1267–1311.
[6] Chen, M.F., Li, Y.S. (2019). *Development of powerful algorithm for maximal eigenpair*. Front Math China 14(3): 493-519.
[7] Chen, M.F. (2020). *On spectrum of Hermitizable tridiagonal matrices*. Front Math China 15(2): 285-303.
[8] Chung, K.L., Yan, W.M. (1997). *The complex Householder transform*. IEEE transactions on signal processing 45(9): 2374–2376.
[9] Golub, G.H., Van Loan, C.F. (2013). *Matrix Computations*, 4th ed. The Johns University Press.
[10] Householder, A.S. (1958). *Unitary triangularization of a nonsymmetric matrix*. J. Assoc. Comput. Mach. 5:339–342.
[11] Jiang, E.X. (1984). *Symmetric Matrix Computation* (in Chinese). Shanghai Press of Sci. Tech.
[12] Min, C. (2010). *A new understanding of the QR method*. J. KSIAM 14 (1), 29–34.
[13] Niño, A., Muñoz-Caro, C., Reyes, S.(2011). *A concurrent object-oriented approach to the eigenproblem treatment in shared memory multicore environments*. Lecture Notes in Computer Science, 6782, 630-642.
[14] Parlett, B.N. (1998). *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, PA.
[15] Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. (2007). *Numerical Recipes. The Art of Scientific Computing,* 3rd ed. Cambridge Univ. Press, Cambridge.
[16] Shukuzawa O, Suzuki T, Yokota I. (1996). *Real tridiagonalization of Hermitian matrices by modified Householder transformation*. Proc. Japan Acad. Ser. A. 72:102-103.
[17] Wang, Z.J. (2018). *Householder transformation for Hermitizable matrix*. Master thesis at Beijing Normal Univ.
[18] Wilkinson, J.H. (1965). *The Algebraic Eigenvalue Problem*. Oxford University Press, Oxford.

CSIAM Trans. Appl. Math. Vol. 3, No. 1, pp. 1–25
doi: 10.4208/csiam-am.2021-0005 March 2022

Top Eigenpairs of Large Scale Matrices

Mu-Fa Chen[1,2,*] and Rong-Rong Chen[3]

[1]RIMS, Jiangsu Normal University, Xuzhou, 221116, China;
[2]School of Mathematics and Key Laboratory of Mathematics, Beijing Normal University, Beijing 100875, China;
[3]Department of Electrical and Computter Engineering, University of Utah, UT 84112, USA.

December 20, 2020

Abstract

This paper is devoted to the study of an extended global algorithm on computing the top eigenpairs of a large class of matrices. Three versions of the algorithm are presented that includes a preliminary version for real matrices, one for complex matrices, and one for large scale sparse real matrix. Some examples are illustrated as powerful applications of the algorithms. The main contributions of the paper are two localized estimation techniques, plus the use of a machine learning inspired approach in terms of a modified power iteration. Based on these new tools, the proposed algorithm successfully employs the inverse iteration with varying shifts (a very fast "cubic algorithm") to achieve a superior estimation accuracy and computation efficiency to existing approaches under the general setup considered in this work.

1 Introduction. Extended global algorithm

The top eigenpairs of matrix play an important role in many fields. In particular, for the maximal eigenpair for instance, there are well-known algorithms in several different fields. For web-search, it is called PageRank. For economic optimization, there is so called left-positive eigenvector method (cf. [1; Chapter 10]). For statistics, there is principal component analysis (abbrev. PCA) which is also used in quantum mechanics computation (quantum chemistry in

Received 22 December, 2021; Accepted 16 January, 2022
2020 *Mathematics Subject Classifications.* 15A18, 65F15.
Key words and phrases. matrix eigenpair, extended global algorithm, localized estimation technique, top eigenpair, large sparse matrix.
*Corresponding author. *Email address:* mfchen@bnu.edu.cn (M.-F. Chen), rchen@utah.edu (R.-R. Chen)

particular) and AI. In the last case, one needs not only the maximal one, but also a couple of the subsequent eigenpairs. Certainly, for such a well-developed field, there are some powerful algorithms in common use, the "singular value decomposition" (abbrev. SVD) for PCA for example. However, as mentioned at the beginning of [9; p.65, §2.6]: "In some cases, SVD will not only diagnose the problem, it will also solve it, in the sense of giving you a useful numerical answer, although, as we shall see, *not necessarily 'the' answer that you thought you should get*." This happens for a number of known algorithms (see [7; Example 1] for instance) and so more careful study is valuable.

This paper is motivated by the study on the global algorithms given in [3, 7], where some effective algorithms were presented for computing the maximal eigenpair of a rather larger class of matrices. Roughly speaking, two approaches are adopted there: the power iteration (abbrev. PI) and the inverse power iteration with varying/fixed shifts (abbrev. $\mathrm{IPI}_v/\mathrm{IPI}_f$). The PI has only a little restriction on the initial vector and so has a wide range of applications. It is also economical (having lower computational complexity), but has a quite slow convergence speed, especially near the target eigenvalue. The fast convergence speed of the algorithms given in [3, 7] is mainly due to the use of IPI_v (having higher computational complexity). It is however quite dangerous if the initial is not close enough (from above) to the target eigenvalue. The last problem was avoided in [3, 7] mainly due to the assumption: the off-diagonal elements of the matrix are all nonnegative. This is essential: it implies the existence of the maximal eigenpair (as an application of the Perron–Frobenius theorem, by a shift if necessary). Then we have some important variational formulas for the upper/lower bounds of the maximal eigenvalue, i.e., the Collatz–Wielandt (abbrev. C-W) formula (cf. [2; §1 and Corollary 12]). For nonnegative matrix, the formula takes the following form:

$$\sup_{x>0} \min_k \frac{(Ax)(k)}{x(k)} = \lambda = \inf_{x>0} \max_k \frac{(Ax)(k)}{x(k)},$$

where λ is the maximal eigenvalue of the matrix A and $x(k)$ is the kth component of the vector x. The upper bound in the formula is very important in using IPI_v for avoiding the pitfalls (cf. [4; §4]). Now, a challenge appears:

Question: What can we do without the assumption of the nonnegative property of the off-diagonal elements?

A typical model led to the question is PCA, for which some of the off-diagonal elements can be negative. The question is quite serious since almost each advantage introduced in the previous paragraph is lost. We do not have the Perron–Frobenius theorem; more seriously, we do not have the C-W formula; and furthermore, the IPI_v is not practical.

Certainly, the answer to the above question is not obvious. If you have luckily produced enough courage, you may look for a way to find a substitute of the C-W formula. Assume that the given matrix A is real. Assume also

for a moment that the maximal eigenvalue λ we are working is positive. Of course, at the present case, the corresponding eigenvector g is not necessarily positive, and it may have negative or zero components. Because we are now bare-handed, to find an exit from the darkness, we have to go back to the original position: all we know is the eigenequation:

$$Ag = \lambda g. \tag{1}$$

That is, g is an eigenvector corresponding to the eigenvalue λ of A. It follows that once $g(k) \neq 0$, we must have $(Ag)(k)/g(k) > 0$, here we have preassumed that $\lambda > 0$. If a vector x produced by our iterative method (either PI or IPI) is close enough to g, then in one iteration, we have

$$x \approx g \Longrightarrow Ax \approx Ag = \lambda g.$$

We now arrive at the first localized estimation technique: *check sign and locally bilateral estimates* (abbrev. CS-LBE). Due to the property given above, on the set

$$\mathscr{N}_x := \{k : |x(k)| > 0\}, \tag{2}$$

we should have

$$\frac{Ax}{x}(k) := \frac{Ax(k)}{x(k)} > 0, \qquad k \in \mathscr{N}_x, \tag{3}$$

since $\lambda > 0$ by assumption. As usual, here "$k \in \mathscr{N}_x$" means "for each $k \in \mathscr{N}_x$". The procedure checking (3) is called "*check sign*" (abbrev. CS). Once this holds for a few of iterations, then we do not need to check it again, just continue the PI until the *relative difference* (abbrev. RD)

$$1 - \min_{k\in\mathscr{N}_x} \frac{Ax}{x}(k) \Big/ \max_{k\in\mathscr{N}_x} \frac{Ax}{x}(k) < \varepsilon \tag{4}$$

for some sufficiently small ε. Under condition $\lambda > 0$, assertions (3) and (4) are actually due to the convergence of PI, assuming for a moment that the maximal eigenvalue coincides with the maximal one in modulus. In practice, one has to take care for the initial vector in using PI to guarantee its convergence. Next, using condition $\lambda > 0$ again, by (3), we have

$$\text{either} \quad (0 <) \ \frac{Ax}{x}(k) \leqslant \lambda \quad \text{or} \quad \frac{Ax}{x}(k) \geqslant \lambda \quad \text{for each} \quad k \in \mathscr{N}_x.$$

Hence under condition (4) with $\varepsilon \ll 1$, we obtain the following *locally bilateral estimates* (abbrev. LBE):

$$(0 \leqslant) \ \min_{k\in\mathscr{N}_x} \frac{Ax}{x}(k) \leqslant \lambda \leqslant \max_{k\in\mathscr{N}_x} \frac{Ax}{x}(k), \tag{5}$$

the equalities in (5) hold once x is taken to be an eigenvector of the corresponding eigenvalue λ. We now regard (5) as a substitute of the C-W upper/lower estimates, and adopt

$$z := \max_{k\in\mathscr{N}_x} \frac{Ax}{x}(k) \tag{6}$$

as an upper bound of λ for the use in IPI_v or IPI_f. Condition (4) guarantees the validity of LBE (5) and then (6). Thus, in (3)–(6), we use only those x in a small neighborhood of the eigenvector of λ in the corresponding vector space. That is the meaning of "locally" used above.

In the above paragraph, we preassume that the maximal eigenvalue coincides with the one in modulus and is positive. This is important not only in computing the ratios above but also an essential point in the use of PI, since for which, the leading term in the algorithm is determined by the maximal eigenvalue in modulus, one cannot ignore the point "in modulus" here. Certainly, if one has known in advance that the spectrum (at least the top six eigenvalues) of A has satisfied the assumption, then the step we are working can be ignored. Otherwise, to remove the assumption, we simply use a *shift operator*: replacing A by

$$A_1 := A + \bar{\theta}\, I, \tag{7}$$
$$\bar{\theta} = \begin{cases} \theta & \text{if the order of } A \text{ is bigger than 6 and } \theta \text{ is an integer,} \\ \lceil \theta \rceil & \text{otherwise} \end{cases}$$

where $\lceil x \rceil$ denotes the minimal integer that is greater or equal to x, and the constant θ is an upper bound of the spectral radius. Here, the use of $\bar{\theta}$ instead of θ is to simplify the computation. Clearly, the spectrum of A_1 is nonnegative. Therefore, working on A_1, the assumption just mentioned holds automatically. The reason that we choose the top six eigenpairs is to compare with the "eigs" package of MatLab, which is designed for the same aim (See section 4 below). Certainly, one can continue the algorithm for additional subsequent eigenpairs.

There are two ways to obtain an upper bound of the spectral radius of general complex matrix A without additional restriction. The first one is a theoretic result, deduced by the Gershgorin Circle Theorem (cf. [10]):

$$\theta = \min\big\{\|A\|_\infty,\ \|A\|_1\big\}, \qquad \|A\|_\infty := \sup_i \sum_j |a_{ij}|,\ \|A\|_1 = \|A^*\|_\infty$$

where A^* denotes the transpose of A. In the symmetric case, the two terms in $\{\cdots\}$ are the same. The disadvantage of this method is that the result is usually quite rough. We now introduce the second numerical method which is similar to the technique deducing (1) – (6) above. Since we are now interested only in the modulus of the eigenvalue λ, instead of (1), we should start at

$$|Ag| = |\lambda|\,|g|.$$

Next, we follow the analysis between (1) and (6). The output x produced by PI, with suitable initial and after enough iterations, should have the following property. With the same $\mathscr{N}_x$ defined by (2), replacing

$$\frac{Ax}{x} \quad \text{by} \quad \left|\frac{Ax}{x}\right|, \quad \lambda \quad \text{by} \quad |\lambda|, \quad \text{and} \quad z \quad \text{by} \quad \theta,$$

we obtain the analogs of (4) – (6) as follows:

$$1 - \min_{k\in\mathscr{N}_x}\left|\frac{Ax}{x}\right|(k)\Big/\max_{k\in\mathscr{N}_x}\left|\frac{Ax}{x}\right|(k) < \varepsilon,$$
$$\min_{k\in\mathscr{N}_x}\left|\frac{Ax}{x}\right|(k) \leqslant |\lambda| \leqslant \max_{k\in\mathscr{N}_x}\left|\frac{Ax}{x}\right|(k),$$
$$\theta := \max_{k\in\mathscr{N}_x}\left|\frac{Ax}{x}\right|(k). \tag{8}$$

Starting from $\mathbb{1}/\|\mathbb{1}\|$, where $\mathbb{1}$ is the constant column vector having its component 1 everywhere and $\|x\|$ is the L_2-norm of x. The resulting θ defined by (8) is what we need for (7). This method is especially good for PI, it converges economically to λ^*, the maximal eigenvalue in modulus, but not the real maximum, effective enough unless it is too close to λ^*. Hence this method is good enough for our purpose. The value of θ is noticeable since a larger θ makes the lower convergence speed of PI:

$$\lambda_1 > \lambda_2 > 0 \Longrightarrow (0,1) \ni \frac{\lambda_2+\alpha}{\lambda_1+\alpha}\uparrow \quad \text{as} \quad \alpha\,(>0)\uparrow .$$

We emphasize that the constant θ defined by (8) is used only in (7) for producing a matrix with nonnegative spectrum having positive six top eigenvalues. In the subsequent estimation of the eigenpairs, one does not use it again. In the special case that the given matrix already has the required property just mentioned above, one can simply ignore this shift procedure.

Usually, one needs to run the IPI_v only for a few of iterations since its convergence speed is very fast. Otherwise, the calculation will overflow quickly. The computation can be finished once the output arrives at the required precision level:

$$\max_{k\in\mathscr{N}_x}\frac{Ax}{x}(k) - \min_{k\in\mathscr{N}_x}\frac{Ax}{x}(k) < \varepsilon, \tag{9}$$

the left-hand part above is called the *amplitude of LBE*. If we do not want to compute the next eigenpair, then we can stop the computations here. If otherwise, one has to improve the precise level of the output of the eigenvector. For this, one should continue the work, using IPI_f instead of IPI_v. This is important since for computing the next eigenpairs, we will go to the subspace which is orthogonal to this eigenvector. The computation of orthogonalization often requires a higher level of precision. Failure to achieve such a precision often leads to error propagation and thus incorrect final results.

We now discuss the construction of the initial vector used by PI. First, for the maximal eigenpair, simply choose the *initial vector*

$$x_0 = \mathbb{1}/\|\mathbb{1}\|. \tag{10}$$

Once the computation of the maximal eigenpair is done, we obtain the first (maximal) eigenvector, say g_1. After $k-1$ steps, we have $k-1$ eigenvectors

$\{g_1, \cdots, g_{k-1}\}$ (normalized with respect to their L_2-norm, respectively). Then the *initial vector* for computing the kth eigenpair can be chosen to be the projection vector of x_0 defined by (10) on the space which is orthogonal to Span $\{g_1, \cdots, g_{k-1}\}$. In general, for a given linear space $\mathscr{L}$, let $\mathscr{L}^{\perp}$ denote its orthogonal space. Then, the projection Proj (x, k) of a vector x on the space Span $\{v_1, \cdots, v_k\}^{\perp}$ is defined by

$$\text{Proj}\,(x,k) = x - \sum_{j=1}^{k} (v_j^* x) v_j \tag{11}$$

for normalized orthogonal family $\{v_1, \ldots, v_k\}$, where v^* (row vector) is the transpose of v (column vector).

To study several eigenpairs, one may assume that the matrix A has real spectrum. Otherwise, for a complex eigenpair, one may have a conjugate one. This poses some difficulty.

At the last step, return to the original matrix:

$$\text{Eigenpair } (\lambda,\ g) \text{ of } A_1 \to \text{Eigenpair } (\lambda - \bar{\theta},\ g) \text{ of } A. \tag{12}$$

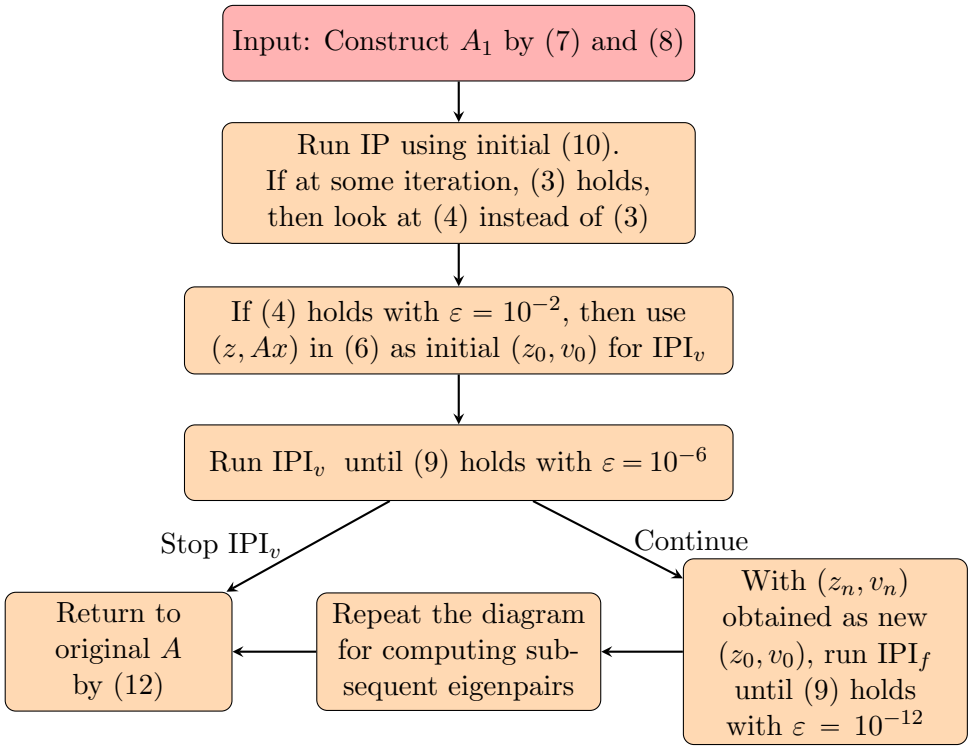


Figure 1: Flowchart of the preliminary version of the extended global algorithm

We now make some additional analysis on the preliminary version of the extended global algorithm in Fig. 1, as well as on the three algorithms used there: PI, IPI$_v$ and IPI$_f$. While the localized estimation technique "check sign and locally bilateral estimates"(CS-LBE) mentioned above looks rather simple, the simplicity is precisely its biggest advantage – it can be applied to a rather wide range of applications, as we will see soon in the subsequent sections. The CS-LBE presents new opportunities to use techniques from a variety of fields such as optimization theory, machine learning, etc., since almost no theoretical results are available in this general setup. What we propose here is the (modified) PI. One may see a concrete example in the next section. Note that the choice of ε used for (4) or (9) in Fig. 1 may be changed according to different types of matrices used in various applications. Roughly speaking, one may use $\varepsilon \in [0.01, 0.1]$ instead of $\varepsilon = 0.01$ in (4) for medium size matrices. At this beginning step, we have used the main advantages of PI: it is safe and allows quite general initial vector, it has a good enough convergence and computing speed, except close too much to the target eigenvector.

Having the initial vector v_0 produced by the CS-LBE technique and the initial shift given by (7) at hand, we are ready to apply IPI$_v$ to accelerate the computing speed. Under the conditions (3) and (4), instead of (5), we have

$$\min_{k\in\mathscr{N}_x} \frac{Ax}{x}(k) \leqslant \frac{x^*Ax}{x^*x} \leqslant \max_{k\in\mathscr{N}_x} \frac{Ax}{x}(k).$$

Replacing the term z given on the right-hand side by the middle one $\frac{x^*Ax}{\|x\|^2}$, the IPI$_v$ becomes the so-called Rayleigh Quotient Iteration (abbrev. RQI), which is well-known a cubic algorithm (i.e., the iterative solutions generated by the algorithm converge cubically). Note that RQI is practical only if x is close enough to the target eigenvector, and hence is also a local algorithm. In particular, it is actually in a dangerous region once

$$\frac{x^*Ax}{x^*x} \in \left[\min_{k\in\mathscr{N}_x} \frac{Ax}{x}(k),\ \lambda\right).$$

However, since the precise local region over which the RQI is effective is not known, practical use of RQI often runs into the issue of converging to other eigenvectors that are close to the target ones. The last point is the main difference between our IPI$_v$ and RQI. The proposed IPI$_v$ ensures the algorithm robustness and allows convergence to the target eigenvector by adapting the shifts automatically. As verified by the practice in [4, 7] and the subsequent sections, the difference given in (4) goes to zero very fast. then so is the difference

$$\max_{k\in\mathscr{N}_x} \frac{Ax}{x}(k) - \frac{x^*Ax}{\|x\|^2}.$$

Hence, it is believable that IPI$_v$ and RQI should have the same order of convergence speed, once RQI works.

For IPI_f, the initial vector v is similar to those of PI; the initial shift z of IPI_f should be bigger than the target λ. Otherwise, the algorithm becomes dangerous. Certainly, IPI_f is more effective if the initial pair (z, v) is closer to the target one. In Fig. 1, IPI_f is used in the last step to improve the target eigenvector. For which, IPI_v may no longer be practical since the inverse matrix would be degenerated too fast. The convergence by IPI_f can be faster than PI whenever the shift is close enough to the target λ from above.

We now summarize roughly the comparison the three algorithms: PI, IPI_v and IPI_f. Let

$$\begin{aligned}\mathscr{D}(U) &= \text{Domain of suitable initial (vector, shift) of algorithm } U,\\ s(U) &= \text{Convergence speed of algorithm } U,\\ t(U) &= \text{Computational complexity of algorithm } U.\\ &\quad \text{From low to high is ordered by ``}\prec\text{''}.\end{aligned}$$

Certainly, for PI, the shift variable is free in $\mathscr{D}(U)$. Then, roughly speaking, we have the following comparison

$$\begin{aligned}\mathscr{D}(\mathrm{PI}) \supset \mathscr{D}(\mathrm{IPI}_f) &\supset \mathscr{D}(\mathrm{IPI}_v),\\ s(\mathrm{PI}) \leqslant s(\mathrm{IPI}_f) &\leqslant s(\mathrm{IPI}_v),\\ t(\mathrm{PI}) \prec t(\mathrm{IPI}_f) &\prec t(\mathrm{IPI}_v).\end{aligned}$$

A mixed algorithm of PI and IPI_v was used in [4, 7]. In the present paper, we introduce some extended algorithms which have more mixture of the above three algorithms, making best use of the advantage and bypassing the disadvantage of each of these three algorithms.

The next section is an exception where the algorithm is applied to the so-called Hermitizable complex matrix, not the real one treated in most part of the paper, to illustrate the wide use of the algorithm. Certainly, from the preliminary version to more general situation, additional work is required, as shown in §3 by the algorithm for large scale sparse matrix. The powerful algorithm is then illustrated by two examples in §4. If a reader is eager to take a look at the power of the proposed algorithm introduced in the paper, he or she can skip Sections 2, 3, and go directly to section 4.

2 Application to Hermitizable matrix

Consider the following complex matrix (cf. [5; Example 7])

$$A_0 = \begin{pmatrix} -6 & \frac{8}{5} - \frac{6\,i}{5} & \frac{8}{13} + \frac{14\,i}{13} & \frac{18}{17} + \frac{4\,i}{17} \\ 3 + \frac{9\,i}{4} & -\frac{55}{4} & -\frac{5}{13} + \frac{40\,i}{13} & \frac{30}{17} + \frac{35\,i}{17} \\ \frac{12}{5} - \frac{21\,i}{5} & -\frac{4}{5} - \frac{32\,i}{5} & -13 & \frac{60}{17} - \frac{66\,i}{17} \\ \frac{63}{10} - \frac{7\,i}{5} & \frac{28}{5} - \frac{98\,i}{15} & \frac{70}{13} + \frac{77\,i}{13} & -16 \end{pmatrix}.$$

A complex matrix $A = (a_{ij})$ is called Hermitizable if there exists a positive measure $\mu = (\mu_k)$ such that $\mu_i a_{ij} = \mu_j \bar{a}_{ji}$ for each pair (i, j) (due to [5]). It is called symmetrizable in the real context. It is easy to check that A_0 is Hermitizable with respect to μ:

$$\mu_0 = 1, \ \ \mu_1 = \frac{8}{15}, \ \ \mu_2 = \frac{10}{39}, \ \ \mu_3 = \frac{20}{119}.$$

In general, from the proof of [5; Theorem 20], it is known that a complex matrix $A = (a_{ij})$ is Hermitizable w.r.t. measure $\mu = (\mu_k)$ iff

$$A = \mathrm{Diag}\,(\mu)^{-1} A^H \mathrm{Diag}\,(\mu) \quad [A^H := \bar{A}^*]. \tag{13}$$

Equivalently,

$$\hat{A} := \mathrm{Diag}(\mu)^{1/2} A\, \mathrm{Diag}\,(\mu)^{-1/2} \tag{14}$$

is Hermitian. Clearly, the transformation of the eigenpair (λ, g) of A to the one $(\lambda, \hat{g})$ of $\hat{A}$ goes as follows.

$$(\lambda, g) \to \big(\lambda, \ \hat{g} = \mathrm{Diag}(\mu)^{1/2} g\big). \tag{15}$$

At the moment,

$$\hat{A}_0 = \begin{pmatrix} -6 & (4-3i)\sqrt{\frac{3}{10}} & (4+7i)\sqrt{\frac{6}{65}} & (9+2i)\sqrt{\frac{7}{85}} \\ (4+3i)\sqrt{\frac{3}{10}} & -\frac{55}{4} & -\frac{2-16\,i}{\sqrt{13}} & (6+7i)\sqrt{\frac{14}{51}} \\ (4-7i)\sqrt{\frac{6}{65}} & -\frac{2+16\,i}{\sqrt{13}} & -13 & (10-11i)\sqrt{\frac{42}{221}} \\ (9-2i)\sqrt{\frac{7}{85}} & (6-7i)\sqrt{\frac{14}{51}} & (10+11i)\sqrt{\frac{42}{221}} & -16 \end{pmatrix}.$$

Due to (15), for computing the eigenpair of A_0, it suffices to study the one for $\hat{A}_0$. Hence, from now on, we need only to consider the matrix $\hat{A}_0$.

The maximal eigenpair

We now start the algorithm given in Fig. 1. The computation in this section is done by using Mathematica (version 11.3) on PC.

Step 1. *Construct* A_1. The upper bound produced by the first method given in §1 is $\theta\big(\hat{A}_0\big) = 29.957$. We now consider the second method.

Starting at $w_0 = \mathbb{1}/\|\mathbb{1}\|$ (cf. (10)) and use the following PI:

$$w_n = \hat{A}_0 v_{n-1}, \quad n \geqslant 1, \qquad v_n := w_n/\|w_n\|, \quad n \geqslant 0.$$

Let

$$\begin{cases} \mathscr{N}(w) = \{k : |w(k)| > 0\}, \\ x_n = \left\{ \left| \dfrac{\hat{A}_0 w_n}{w_n} \right| (k), \ k \in \mathscr{N}(w_n) \right\}, \\ y_n = \min\limits_{k \in \mathscr{N}(w_n)} x_n(k), \qquad z_n = \max\limits_{k \in \mathscr{N}(w_n)} x_n(k). \end{cases}$$

Then in 5 iterations, the outputs are as follows.

$$\begin{aligned}&\{z_n, y_n\}_{n=1}^5\colon (21.2379, 5.26626),\ (27.0853, 17.7591),\ (27.2742, 17.6156),\\ &\qquad\qquad (21.9304, 17.52740),\ (21.6953, 17.4757);\\ &\{1 - y_n/z_n\}_{n=1}^5\colon \quad .752035,\ .344325,\ .354132,\ .200772,\ .194493.\end{aligned}$$

Clearly, PI converges very well. Since z_4 and z_5 are closed each other, for them we have the same $\bar\theta = 22$ which is an upper bound of the spectral radius of $\hat A_0$ and is obviously smaller than the one obtained by the first method. Actually, if we continue PI for more iterations,

$$z_5 = 21.6953,\ z_{10} = 21.5148,\ z_{20} = 21.7481,\ z_{30} = 21.4567,\ z_{40} = 21.3927,$$

then we get the same $\bar\theta$, since the convergence becomes rather slow when z_n is close to the modulus of the maximal eigenvalue $\lambda^* = -21.3806$. Thus by (7), we have

$$\begin{aligned}A_1 &= \hat A_0 + \bar\theta\, I\\ &= \begin{pmatrix} 16 & (4-3i)\sqrt{\frac{3}{10}} & (4+7i)\sqrt{\frac{6}{65}} & (9+2i)\sqrt{\frac{7}{85}}\\ (4+3i)\sqrt{\frac{3}{10}} & \frac{33}{4} & -\frac{2-16\,i}{\sqrt{13}} & (6+7i)\sqrt{\frac{14}{51}}\\ (4-7i)\sqrt{\frac{6}{65}} & -\frac{2+16\,i}{\sqrt{13}} & 9 & (10-11i)\sqrt{\frac{42}{221}}\\ (9-2i)\sqrt{\frac{7}{85}} & (6-7i)\sqrt{\frac{14}{51}} & (10+11i)\sqrt{\frac{42}{221}} & 6 \end{pmatrix}.\end{aligned}$$

To justify the effectiveness of the shift used here, let us compute the eigenvalues of A_1:

$$21.8344,\ 12.5542,\ 4.24189,\ 0.619429.$$

It follows that there is only a little room (about 0.6) for the improvement of the shift $\bar\theta = 22$ to keep the positivity of the spectrum of A_1. The transformation of the maximal eigenpair $(\lambda_1(A_1),\ g_1(A_1))$ of A_1 to the one $(\lambda_1, g_1) := (\lambda_1(A_0), g_1(A_0))$ of the original A_0 is as follows.

$$\lambda_1 = \lambda_1(A_1) - \bar\theta, \qquad g_1 = \mathrm{Diag}(\mu)^{-1/2} g_1(A_1). \tag{16}$$

Step 2. *Run PI.* As in Step 1, we use the following PI:

$$w_n = A_1 v_{n-1}, \quad n \geqslant 1, \qquad v_n := w_n/\|w_n\|, \quad n \geqslant 0.$$

However, the original initial $\mathbb{1}/\|\mathbb{1}\|$ is replaced by $w_0 = (1+i)\mathbb{1}/\big(\sqrt{2}\|\mathbb{1}\|\big)$. The reason is that for non-real A_1, since the eigenvalues are all real, the eigenvectors should be non-real and so as a mimic, it is better to choose w_0 to be non-real. However, this is useless in Step 1, since a nonzero constant factor α can be ignored in the equation

$$|A(\alpha v)| = |\lambda|\,|\alpha v|.$$

We now come to the essential different point from the real case. Actually, for non-real A_1, instead of the single equation (1), we have two:

$$\operatorname{Re}(A_1 g) = \lambda \operatorname{Re}(g), \qquad \operatorname{Im}(A_1 g) = \lambda \operatorname{Im}(g).$$

Thus, it is naturally to split the original vector x (corresponding to g in the eigenequation) into two: x^R and x^I (corresponding to $\operatorname{Re} g$ and $\operatorname{Im} g$, respectively). Similarly we have $\mathscr{N}_R$ and $\mathscr{N}_I$ defined as follows.

$$\begin{cases} \mathscr{N}_R(w) = \{k : |\operatorname{Re} w(k)| > 0\}, \ \ \mathscr{N}_I(w) = \{k : |\operatorname{Im} w(k)| > 0\}; \\ p_n = A_1 w_n; \\ x_n^R = \left\{ \dfrac{\operatorname{Re} p_n}{\operatorname{Re} w_n}(k),\ k \in \mathscr{N}_R(w_n) \right\}, \quad x_n^I = \left\{ \dfrac{\operatorname{Im} p_n}{\operatorname{Im} w_n}(k),\ k \in \mathscr{N}_I(w_n) \right\}; \\ \text{weak} \begin{cases} y_n = \Big(\bigwedge_{k\in\mathscr{N}_R(w_n)} x_n^R(k) \Big) \bigwedge \Big(\bigwedge_{k\in\mathscr{N}_I(w_n)} x_n^I(k) \Big), \\ z_n = \Big(\bigvee_{k\in\mathscr{N}_R(w_n)} x_n^R(k) \Big) \bigvee \Big(\bigvee_{k\in\mathscr{N}_I(w_n)} x_n^I(k) \Big); \end{cases} \\ \text{strong} \begin{cases} y_n = \Big(\bigwedge_{k\in\mathscr{N}_R(w_n)} x_n^R(k) \Big) \bigvee \Big(\bigwedge_{k\in\mathscr{N}_I(w_n)} x_n^I(k) \Big), \\ z_n = \Big(\bigvee_{k\in\mathscr{N}_R(w_n)} x_n^R(k) \Big) \bigwedge \Big(\bigvee_{k\in\mathscr{N}_I(w_n)} x_n^I(k) \Big); \end{cases} \end{cases} \tag{17}$$

where $\alpha \wedge \beta = \min\{\alpha, \beta\}$ and $\alpha \vee \beta = \max\{\alpha, \beta\}$ for real α and β. The last two parts "weak" and "strong" need some explanation. First, the only difference is exchanging the "$\wedge$" and "$\vee$" in the middle of definition of (y_n, z_n). To understand its essential difference, recall that condition (5) is now split into two:

$$\begin{aligned} (\operatorname{Re}_x): &\qquad \min_{k\in\mathscr{N}_R(x)} \frac{\operatorname{Re}(A_1 x)}{\operatorname{Re} x}(k) \leqslant \lambda \leqslant \max_{k\in\mathscr{N}_R(x)} \frac{\operatorname{Re}(A_1 x)}{\operatorname{Re} x}(k), \\ (\operatorname{Im}_x): &\qquad \min_{k\in\mathscr{N}_I(x)} \frac{\operatorname{Im}(A_1 x)}{\operatorname{Im} x}(k) \leqslant \lambda \leqslant \max_{k\in\mathscr{N}_I(x)} \frac{\operatorname{Im}(A_1 x)}{\operatorname{Im} x}(k). \end{aligned}$$

Now, for the "weak" case in (17) we simply adopt a weaker estimate of (y_n, z_n) from $(\operatorname{Re}_{x_n})$ and $(\operatorname{Im}_{x_n})$. And then the "strong" case should be clear. The weaker version of (y_n, z_n) plays the main role for the safety of converging to the required eigenpair, but makes a little slower convergence. While the stronger version makes a faster convergence but it requires that we are at the position close enough to the target eigenpair. Keeping these ideas in mind, one may adopt a mixture of these choices in designing the algorithms.

To fix the idea, throughout this section, the weak version of (y_n, z_n) is adopted at the first use of PI only in the computation of each eigenpair. For the other steps, we adopt the strong version.

It is the position to start the PI. In 6 iterations, the outputs are as follows.

$\{z_n, y_n\}_{n=1}^6$: (22.6771, −8.15858), (92.2205, 21.1287), (25.9135, 20.2681), (23.4485, 18.5274), (22.6331, 19.0585), (22.2652, 19.8867);

$\{z_n - y_n\}_{n=1}^6$: 30.8357, 71.0918, 5.64541, 4.92104, 3.57468, 2.37847;

$\{1 - y_n/z_n\}_{n=1}^6$: 1.35977, .770889, .217856, .209866, .15794, .106825.

Note that here $y_1 < 0$. The outputs show that not only the components of $\operatorname{Re} w_n$ and $\operatorname{Re} w_{n-1}$ have the same sign once $n \geqslant 2$, but also the sequence of relative difference decreases quite quickly. We choose $n = 6$ ($\varepsilon \sim .1$) as the final iteration. Then, we have

$$v_6 = (.363237 + .491209\,i,\ -.00786884 + .44046\,i,\ .488441 - .0516616\,i,\\ .326973 + .290776\,i)^*.$$

Step 3. *Run IPI$_v$.* Starting at $(z_0, v_0) = (z_6,\ v_6)$ obtained in the last step, run IPI$_v$. Here we adopt a little different notation. Let w_n solve the equation

$$(z_{n-1}I - A_1)w_n = v_{n-1}, \qquad n \geqslant 1.$$

and set $v_n = w_n/\|w_n\|$ again. Next, define $\mathscr{N}_R(w)$, $\mathscr{N}_I(w)$ and $\{x_n^R, x_n^I, y_n, z_n\}$ by (17) with the strong version of (y_n, z_n).

Note that z_n and $1 - y_n/z_n$ are analogs of (6) and (4) in the complex context, respectively. Then, in 2 iterations, we obtain

$$(z_n,\ y_n)_{n=1}^2 \colon (21.8358,\ 21.8324),\ (21.8344, 21.8344),\\ (z_n - y_n)_{n=1}^2 \colon .00346973,\ 5.22045 \cdot 10^{-7};\\ \{1 - y_n/z_n\}_{n=1}^2 \colon\ .000158901,\ 2.39093 \cdot 10^{-8},\\ v_2 = (.359825 + .494092\,i,\ -.0061931 + .44037\,i,\ .488017 - .054044\,i,\\ .328093 + .289324\,i)^*.$$

Moreover, $1 - y_2/z_2 \sim 10^{-8}$. This is not too small for the use of IPI$_f$ in the next step.

Step 4. *Run IPI$_f$.* In the case we want to improve the above result furthermore, we adopt the IPI$_f$. Now, we take (z_2, v_2) from the last step as our new initial (z_0, v_0). The only change to the last IPI$_v$ is using the fixed $z_n \equiv z_0$. In 3 iterations, if we adopt the same precise digits as the last step, then we get the same outputs of (z_n, y_n) as the last one:

$$\{(z_n,\ y_n)\}_{n=1}^3 \colon \text{ the same pair } (21.8344,\ 21.8344);\\ \{y_n - z_n\}_{n=1}^3 \colon \{10.6581,\ 0,\ 3.55271\} \cdot 10^{-15};\\ \{1 - y_n/z_n\}_{n=1}^3 \colon \{5.55112,\ 1.11022,\ 2.22045\} \cdot 10^{-16}.\\ v_3 = (.359825 + .494092\,i,\ -.00619309 + .44037\,i,\ .488017 - .054044\,i,\\ .328093 + .289324\,i)^*.$$

In what follows, we rewrite (z_3, v_3) as $(\lambda_1(A_1), g_1(A_1))$ which is regarded as the maximal eigenpair of A_1.

The submaximal eigenpair

From the last part, we have obtained the maximal eigenpair $(\lambda_1(A_1),\ g_1(A_1))$, at the machine level of precision, as follows.

$$\begin{aligned}\lambda_1(A_1) &= 21.834441785286337,\\ g_1(A_1) &= (.35982503686976175 + .49409186313969483\,i,\\ &\quad -.006193088194633169 + .44037016603620777\,i,\\ &\quad .48801737987976945 - .054043998846425696\,i,\\ &\quad .3280927162424674 + .28932402046371486\,i)^*.\end{aligned}$$

Step 1. *Run modified PI.* As an analog (11), the projection of the vector w on the space $\mathrm{Span}\,(g_1(A_1))^\perp$ is as follows.

$$w - \frac{g_1(A_1)^H w}{(g_1(A_1))^H g_1(A_1)} g_1(A_1) \qquad [g^H := \bar{g}^*].$$

The modified PI means the use of the usual PI with the modification by the projection above at each step. That is

$$\begin{aligned} w_0 &= \frac{(1+i)\mathbb{1}}{\sqrt{2}\,\|\mathbb{1}\|},\\ u_n &= w_n - \frac{g_1(A_1)^H w_n}{(g_1(A_1))^H g_1(A_1)} g_1(A_1), \quad n \geqslant 0,\\ w_n &= A_1 \frac{u_{n-1}}{\sqrt{u_{n-1}^H u_{n-1}}}, \quad n \geqslant 1.\end{aligned}$$

Next, similar to (17), replacing w and w_n by u and u_n respectively, we can define $\mathscr{N}_R$, $\mathscr{N}_I$, p_n, x_n^R, x_n^I, and the weak version of (y_n, z_n). Starting at w_0 and running the modified PI, in 5 iterations, we obtain

$$\begin{aligned}\{(z_n, y_n)\}_{n=1}^5 &: (13.3067,\ 1.07981),\ (12.7854,\ -32.9231),\ (18.4212,\ 10.2147),\\ &\quad (13.9055,\ 11.5768),\ (12.9665,\ 12.1958);\\ \{z_n - y_n\}_{n=1}^5 &: \ 12.2269,\ 45.7085,\ 8.20655,\ 2.32871,\ .770683;\\ \{1 - y_n/z_n\}_{n=1}^5 &: \ .918852,\ 3.57505,\ .445495,\ .167467,\ .0594367.\end{aligned}$$

Note that here $y_2 < 0$. We stop at $n = 5$ since $1 - y_5/z_5$ is small enough, even though it is bigger than 10^{-2}. Then, we have

$$\begin{aligned} z_5 &= 12.9665,\\ v_5 &= (.0677311 - .786181\,i,\ -.190409 + .21529\,i,\ .356539 + .247925\,i,\\ &\quad .0428153 + .322965\,i)^*.\end{aligned}$$

Step 2. *Run IPI$_v$.* Taking (z_5, v_5) from the last step as new (z_0, v_0), run IPI$_v$. Let w_n solve the equation

$$(z_{n-1} I - A_1) w_n = v_{n-1}, \qquad n \geqslant 1,$$

and define first $v_n = w_n/\|w_n\|$, and then $\mathscr{N}_R(w)$, $\mathscr{N}_I(w)$ and $\{x_n^R, x_n^I, y_n, z_n\}$ by (17) with the strong version of (y_n, z_n). Now, in 3 iterations, we obtain

$$\begin{aligned}
&\{(z_n, y_n)\}_{n=1}^3 : (12.5546,\ 12.5507),\ (12.5542,\ 12.5542),\ (12.5542,\ 12.5542);\\
&\{z_n - y_n\}_{n=1}^3 : .00385406,\ 1.50454\cdot 10^{-7},\ 3.55271\cdot 10^{-15};\\
&\{1 - y_n/z_n\}_{n=1}^3 :\ \ .000306985,\ 1.19843\cdot 10^{-8},\ 3.33067\cdot 10^{-16}.\\
&v_2 = (.604525 - .508517\,i,\ -.301145 + .0168374\,i,\ .0761289 + .417867\,i,\\
&\qquad -.196423 + .2569\,i)^*.\\
&v_3 = (.604525 - .508517\,i,\ -.301145 + .0168374\,i,\ .0761289 + .417867\,i,\\
&\qquad -.196423 + .2569\,i)^*.
\end{aligned}$$

In the case we do not want to go further, we can stop here at $n = 3$ since $1 - y_3/z_3 \sim 10^{-16}$ is sufficiently small. It is actually too smaller to go to the next step, otherwise it would cost some computational error.

Step 3. *Run IPI$_f$.* To have a test, setting (z_0, v_0) to be (z_2, v_2) obtained in the last step, run IPI$_f$ also in 3 iterations, we obtain the same output $z_n = y_n = 12.5542$ for $n = 1, 2, 3$, and

$$\begin{aligned}
&\{z_n - y_n\}_{n=1}^3 :\ \{3.55271,\ 3.55271,\ 8.88178\}\cdot 10^{-15};\\
&\{1 - y_n/z_n\}_{n=1}^3 :\ \{3.33067,\ 3.33067,\ 6.66134\}\cdot 10^{-16}.
\end{aligned}$$

Moreover

$$\begin{aligned}
v_3 = (&.6045251632662887 - .5085174051419706\,i,\\
&-.3011448284487476 + .016837350902488956\,i,\\
&.07612884589652998 + .4178669662768421\,i,\\
&-.19642273529356236 + .2568995483366027\,i)^*.
\end{aligned}$$

The present v_3 has a much higher precise level than v_2 obtained in Step 2. We now regard (z_3, v_3) as the submaximal eigenpair $(\lambda_2(A_1),\ g_2(A_1))$ of A_1.

Similarly, one can compute the other eigenpairs of A_1 but we are not going to the details here.

Finally, we return to the original eigenpairs of A_0 by (16):

$$\begin{aligned}
&\lambda_1 = \lambda_1(A_1) - 22 = -.165558,\\
&g_1 = \mathrm{Diag}(\mu)^{-1/2} g_1(A_1) = (.359825 + .494092\,i,\ -.00848024 + .603002\,i,\\
&\qquad\qquad .963757 - .106728\,i,\ .800304 + .705737\,i)^*\\
&\lambda_2 = \lambda_2(A_1) - 22 = -9.44576,\\
&g_2 = \mathrm{Diag}(\mu)^{-1/2} g_2(A_1) = (.604525 - .508517\,i,\ -.41236 + .0230555\,i,\\
&\qquad\qquad .150342 + .825221\,i,\ -.479127 + .626645\,i)^*
\end{aligned}$$

It is nice chance to learn some thing from the above computation.

1) In the earlier papers [3] and [7], the sequence $\{z_n\}$ should control the maximal eigenvalue from above, due to the Perron-Frobenius theorem and the C-W formula mentioned in §1. However, this may not be true in the present general setup, as can be seen from Step 1 of computing the maximal eigenpair,

$$z_1 < |\lambda^*(\hat{A}_0)| = 21.3806 < z_2 < z_3 > z_4 > z_5 > |\lambda^*(\hat{A}_0)|,$$

the sequence $\{z_n\}$ arrives its maximum at z_3. In Step 2, we have similarly,

$$\lambda_1(A_1) = 21.8344 < z_1 < z_2 > z_3 > \cdots > z_6 > \lambda_1(A_1).$$

In this case, it follows that the sequence $\{z_n\}$ arrives its maximum at z_2, and then it goes down. In both cases, the sequence $\{1 - y_n/z_n\}$ is decreasing in n quickly.

Step 1 in computing the submaximal eigenpair is much more interesting. It illustrates the unstable property of $\{z_n\}$ at the beginning. Here we adopt the modified PI. We have

$$z_1 > z_2 > \lambda_2(A_1) = 12.5542 < z_3 > z_4 > z_5 > \lambda_2(A_1).$$

Correspondingly, for $\xi_n := 1 - y_n/z_n$, we have

$$\{\xi_n\}_{n=1}^5 : \quad .918852,\ 3.57505,\ .445495,\ .167467,\ .0594367.$$

A big jump happens at z_3 since as mentioned earlier, $y_2 < 0$ and so the check sign (CS) is necessary. At $n = 5$, even though $\xi_5 \sim .059 > .01$, but $z_5 = 13.0168 > \lambda_2(A_1)$, and so the use of IPI_v in the subsequent step is safe. Roughly speaking, one can stop PI at the mth iteration, if starting from z_m, the sequence $\{z_n\}_{n \geqslant m}$ converges decreasingly. It is the case if the matrix has nonnegative off-diagonals, as studied in [3, 7], or the examples given in §4. In view of this point, one may reduce the number of iterations in using PI at the beginning of the computation for the maximal/submaximal eigenpair. More precisely, the PI (*Step* 2) for computing the maximal eigenpair needs only $6 - 2$ iterations and for submaximal one, it requires only $5 - 1$ iterations. For subsequent IPI_v or IPI_f, the number of iterations remains the same as the original in the both cases.

2) All the computations above show that the sequence $\{1 - y_n/z_n\}_n$, may be except a few of terms at the beginning, is monotone decreasing and converges, much stable than the other sequences, $\{z_n\}$ or $\{|1 - z_n/z_{n-1}|\}$, in the present general setup. Among the computations above, the exceptional part of the sequence $\{\xi_n = 1 - y_n/z_n\}$ appears mainly in the last case just discussed above. For which, the first 2 terms are unstable, especially the second one is bigger than 1 since $y_2 < 0$ as mentioned before. The stability starts at the third term. It follows that the use of the sequence $\{1 - y_n/z_n\}$ is more practical and is actually adopted in the preliminary version of the algorithm given in Fig. 1. For this reason, it seems more precise to rename the "CS-LBE" technique

by adding the relative difference (RD): CS-RD-LBE technique in the general situation.

It is hoped that the algorithm given here could be used in the quantum mechanics computation (cf. [6]).

For the remainder of the paper, we return to real matrices for which Hermitizable becomes symmetrizable. By (15), we can reduce a symmetrizable matrix to a symmetric one. Then, by using (16), we can assume that the given symmetric matrix has a nonnegative spectrum.

3 A version of the global algorithm for large scale matrices

As remarked at the end of the last section, we need only to study the symmetric matrix having nonnegative eigenvalues. In this section, we describe the extended global (or global for short) algorithm for computing the top eigenpairs of a large sparse matrix. This algorithm computes the eigenpairs sequentially, starting from the top eigenpair and then uses the previously computed $(i-1)$ eigenpairs to compute the next ith eigenpair. The flowchart of the algorithm for computing the ith eigenpair is shown in Fig. 2.

We first summarize the key points of the proposed algorithm as follows.

- The inputs to the algorithm are the first $i-1$ eigenpairs $\{(\lambda_j, v_j), j = 1, \cdots, i-1\}$ that have already been computed using the same algorithm. Here, λ_j denotes the jth largest eigenvalue and v_j denotes the jth eigenvector.
- At the initial iteration $n = 0$, we initialize with y_0 given by (10).
- Starting from the initial vector y_0, run a procedure called "Check sign with locally bilateral estimates (CS-LBE)" to determine initial shift z_0 and the corresponding eigenvector estimate x_0. This procedure involves running multiple power iterations with projection and check sign, and estimating z_0 based on the locally bilateral estimates (an analog of (5)). Details of the CS-LBE procedure will be described later.
- Given x_n and z_n, determined by the CS-LBE procedure, we then perform one iteration of the IPI$_v$: $(z_n I - A)y_n = x_n$ to solve for the updated eigenvector estimate y_n.
- Given y_n, we will run the CS-LBE procedure to determine the next shift z_{n+1} and the corresponding eigenvector estimate x_{n+1}.
- Given x_{n+1}, we will check whether the accuracy of x_{n+1} has improved compared to that of earlier iterations. Detailed criterion used to evaluate the accuracy of the eigenvector (which corresponds to the amplitude of LBE given in §1) will be described later.
- If the accuracy of x_{n+1} has not improved compared to earlier iterations, then the algorithm has converged. It then outputs the ith eigenpair

$\lambda_i = z_{n+1}$, $v_i = x_{n+1}$ and proceeds with the computation of the $(i+1)$th eigenpair. On the other hand, if the accuracy of x_{n+1} has improved compared to earlier iterations, then the algorithm proceeds with the next iteration of IPI_v.

- Note that for the nth iteration of the IPI_v, if the condition $|z_n - z_{n-1}| < 10^{-8}$ is met, then we stop updating the shift and set $z_n = z_{n-1}$ instead. That is, we turn to IPI_f.

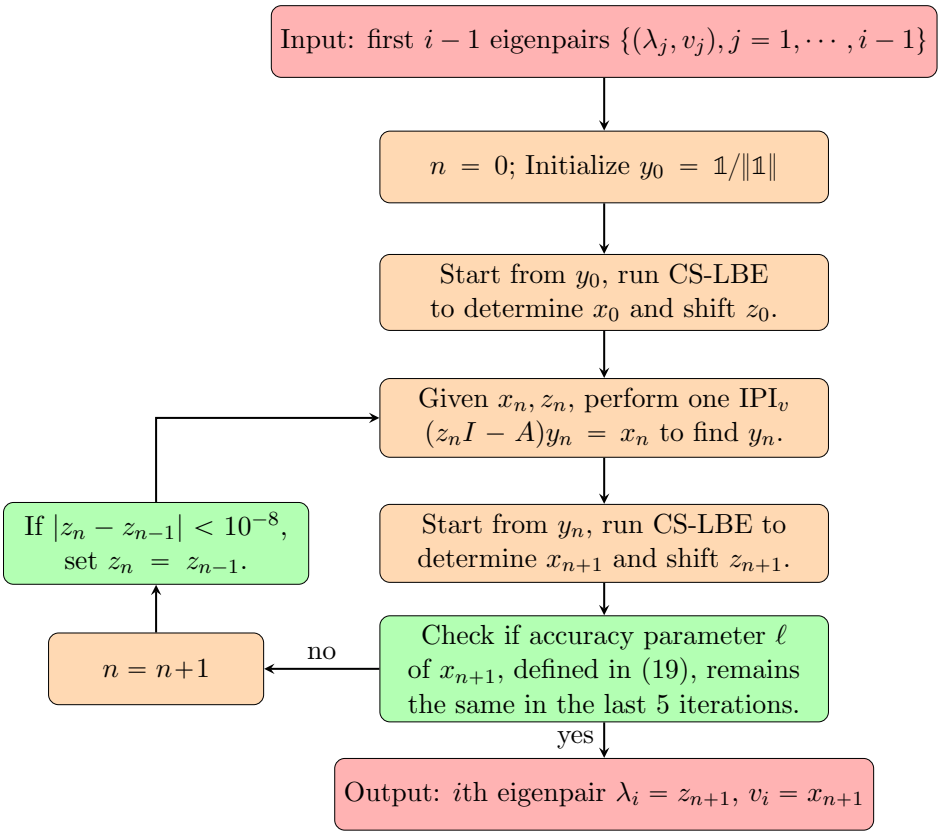


Figure 2: Flowchart of the main algorithm for computing the ith eigenpair. Assume that the previous $i-1$ eigenpairs have been computed.

Next, we provide more details of the global algorithm. We will first define the CS-LBE procedure. This procedure requires the following two basic operations.

Projection operator This is defined by (11). It ensures that after projection, the vector $\text{Proj}(x, k)$ is orthogonal to the linear space $\text{Span}\,(\{v_j, j = 1, \cdots, k\})$.

Shift Evaluation Given a current estimate of the eigenvector x, we aim to determine a proper shift based on the locally bilateral estimates. For large

sparse matrices, the components of x may decay to zero very quickly. Thus, estimation of the shift using all components of x can be unreliable, and sensitive to the estimation errors of those components of x with very small amplitudes. In our algorithm, we propose to calculate the shift based on only the principal components of x such that $|x(i)| \geqslant t(x)$, where $t(x)$ is a threshold value to be determined. Estimating the shift based on only principal components with larger amplitude improves the estimate of the shift. Let x be a unit vector in the L_2-space of dimension N. Given x, we define the shift evaluation function, denoted by $z(x)$, as follows. Let x_a denote the sorted vector of $|x|$ in the descending order. Let n' be the smallest integer such that $\sum_{i=1}^{n'} x_a(i)^2 \geqslant \epsilon_0$. Typically, we set $\epsilon_0 = 0.9$. This means that the first n' components of vector x_a captures 90% of the energy of vector x. Let $t(x) = |x_a(n')|$. Given x and $y = Ax$, we define the shift evaluation function $z(x)$ by considering only the major components of x such that $|x(i)| \geqslant t(x)$:

$$z(x) = \max_{\{i:|x(i)|\geqslant t(x)\}} \frac{y}{x}(i) \qquad \left[\frac{y}{x}(i) := \frac{y(i)}{x(i)}\right]. \tag{18}$$

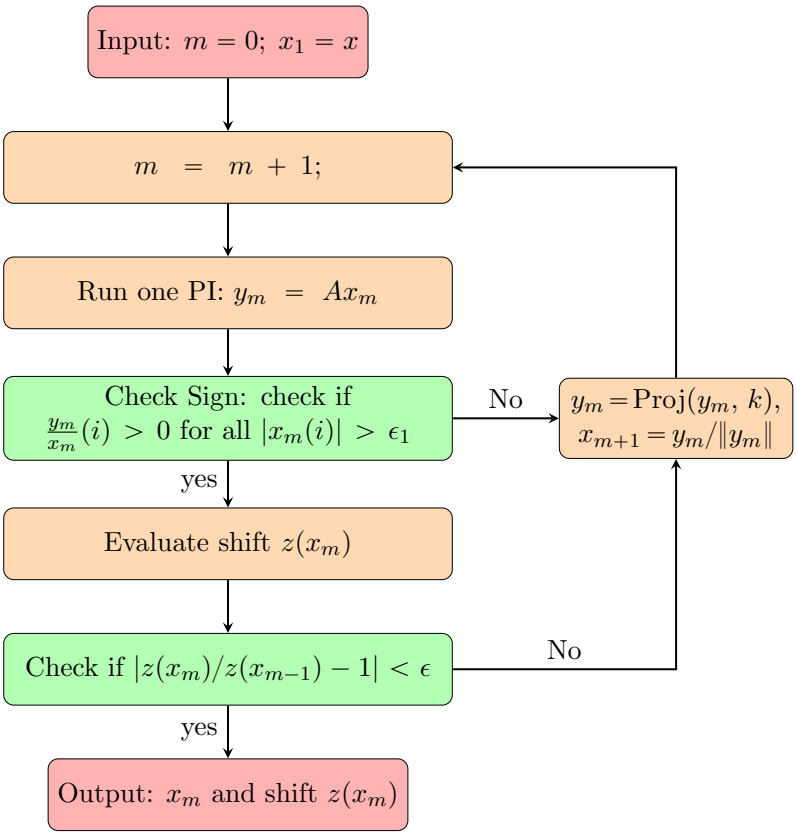


Figure 3: Flowchart of the compute-shift with locally bilateral estimates (CS-LBE) procedure.

This is a modification of (6) for the large scale matrix. Note that in (18), we adaptively determine the principal components of the estimated eigenvector x over iterations. This is important to obtain good estimates of the shift $z(x)$.

In Fig. 3, we show the flow-diagram of the CS-LBE procedure in using the *modified PI* (cf. Step 1 of computing the submaximal eigenpair given in §2). The input to this procedure is an initial estimate of the eigenvector x. The subscript m is the index of the PI. At the mth PI, we calculate $y_m = Ax_m$. This is followed by a check sign step in which we check whether the condition that

$$\frac{y_m}{x_m}(i) > 0 \text{ is satisfied for all } |x_m(i)| > \epsilon_1 \qquad \text{(analog of (3))}.$$

If check sign fails, then we conduct a projection step on y_m to make sure that the resulting vector is orthogonal to the linear space generated by the first $k-1$ eigenvectors. Then we set $x_{m+1} = y_m/\|y_m\|$ and then proceeds to the next PI. If the check sign is successful, then we compute the shift $z(x_m)$ in the next step. The shift evaluation function z is defined as in (18). We will compare the newly computed shift $z(x_m)$ with the previous shift $z(x_{m-1})$ to see whether the shift values have converged. If so, we will finish the procedure and output the updated estimate of the eigenvector x_m and the shift $z(x_m)$. Otherwise, the algorithm will proceed with the next PI.

Check eigenvector accuracy Most works in the literature use L_2 norm of the error vector between the true eigenvector and the estimated eigenvector to evaluate the accuracy of the eigenvector estimation. However, since L_2 norm is obtained by summing over all components of the error vector, it can not accurately describe the accuracy of the individual components. In this work, we adopt a different metric by examining the accuracy of component-wise ratios of $y = Ax$ and x. By the definition of the eigenvector, for each component $x(k) \neq 0$, then the ratio $y(k)x(k)^{-1}$ should equal the eigenvalue λ. This is a challenging task for the setting of large matrices due to the high matrix dimension and the rapid decay of the eigenvectors. Typically, when the amplitude of a component $x(k)$ is large, the estimation tends to be more accurate, and thus the ratio $y(k)x(k)^{-1}$ will be closer to the eigenvalue λ. For small $x(k)$, the ratio $y(k)x(k)^{-1}$ tends to deviate away from λ due to estimation inaccuracy. Hence, it is meaningful to consider the amplitude range of $x(k)$ over which all $y(k)x(k)^{-1}$ are close to λ.

We now arrive at the second localized estimation technique: *Accuracy of the principal components of the approximating eigenvector.*

Consider an estimated eigenvector x of dimension N. Let I denote a permutation of $\{1, 2, \cdots, N\}$ obtained by sorting the components of $|x|$ in the descending order. Given I, we define $\tilde{x}$ as $\tilde{x}(i) = x(I(i)),\ i = 1, \cdots, N$. Given the same I, we let $y = Ax$ and define $\tilde{y}$ as $\tilde{y}(i) = y(I(i)),\ i = 1, \cdots, N$.

- Let $m' = \max_i \Big\{ i : |\tilde{x}(i)| > 0 \Big\}$.

- The accuracy parameter ℓ of the estimated eigenvector specifies the number of reliable components of $\tilde{x}$. It is defined as

$$\ell = \max_{1 \leqslant i \leqslant m'} \left\{ i : \max_{1 \leqslant j \leqslant i} \frac{\tilde{y}}{\tilde{x}}(j) - \min_{1 \leqslant j \leqslant i} \frac{\tilde{y}}{\tilde{x}}(j) < 10^{-6} \right\}. \tag{19}$$

- Based on the definition of ℓ in (19), we see that the estimated eigenvector x achieves a high accuracy for the largest ℓ components (in absolute value). In other words, the components of x have high accuracy for all components $x(i)$ such that $|x(i)| \geqslant |\tilde{x}(\ell)|$.

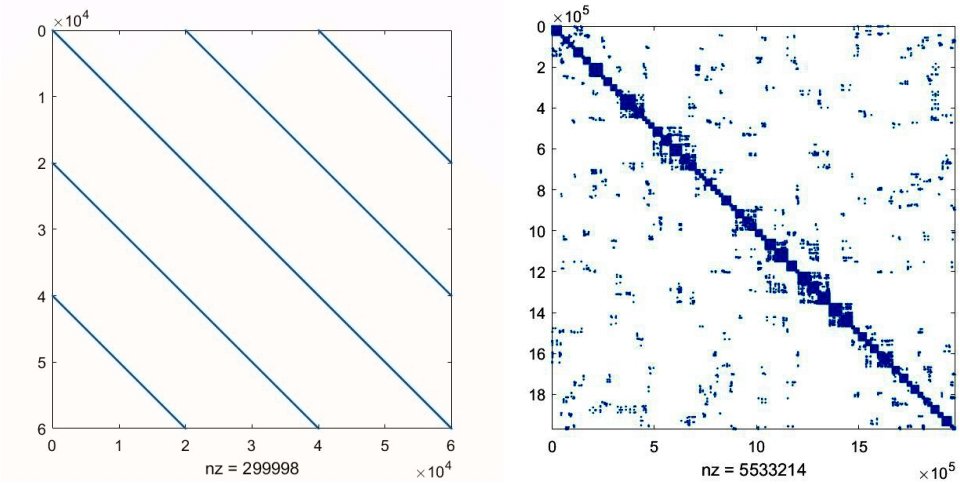
Note that in the proposed algorithm shown in Fig. 2, we calculate the accuracy parameter ℓ for the estimated eigenvector x_{n+1} according to 19. As the algorithm proceeds, ℓ will increase over iterations. We terminate the algorithm if ℓ no longer increases over five consecutive iterations.

4 Application to large scale sparse matrices

In this section, we provide two examples of using the global algorithm to compute the top 6 eigenpairs. The two large matrices come from the SuiteSparse Matrix Collection, publicly available at `https://sparse.tamu.edu`. We will compare the proposed algorithm with two other methods. One is the Matlab Eigs function, which computes the top six eigenpairs of large, sparse matrices. The other is the modified power iteration method, where we perform the standard power iteration together with the projection step to compute the top six eigenpairs. All the experiments presented in this section are executed on an AMD Ryzen 5 2600 Six-Core Processor with single core CPU speed 3.85 GHz, Memory 32 GB. Matlab version is R2015b Windows 10. Related work on computing the top eigenpair for large sparse matrices include [12], [11], [8]. In particular, [12], [11] consider the use of inverse iterations using fixed shifts. This work differs from [12], [11] in the use of the proposed (CS-LBE) procedure to adaptively compute the shifts. Furthermore, the estimated eigenvector accuracy considered in [12], [11], [8] (for the largest eigenpair only) is similar to that of the Matlab Eigs function, which only guarantees the accuracy of a small number of large principal components. In comparison, the proposed global algorithm achieves a high accuracy for even eigenvector components with an exceedingly small magnitude.

dixmaanl dataset

This matrix has a dimension of $N = 60000$. *The number of nonzero elements* (abbrev. nz) is 299998, This matrix is nonnegative, symmetric, and the range of the elements is between 0 and 154.8089. The sparsity pattern of this matrix is shown in Fig. 4(a).



(a) (b)

Figure 4: Sparsity of the two datasets. (a) dixmaanl (b) roadNet-CA

Table 1: dixmaanl dataset. Computed top 6 eigenvalues using the Global algorithm, eigs, and modified PI.

	global	eigs	PI
λ_1	317.0152899359881	317.0152899359666	317.0152899359881
λ_2	317.0058090659085	317.0058090659162	317.0058090659074
λ_3	316.9980633932568	316.9980633932683	316.9980633932562
λ_4	316.9912300516546	316.9912300516576	316.9912300516548
λ_5	316.9849936226963	316.9849936226929	316.9849936226971
λ_6	316.9791911040992	316.9791911040974	316.9791911040990

In Table 1, we show the estimated top 6 eigenvalues obtained by each method. We see that all three methods provide similar eigenvalue estimates that agree with each other up to 10 decimal points.

In Table 2, we provide detailed comparisons of the three methods in terms of the accuracy of the eigenvector, the complexity, and the running time. Each row corresponds to results associated with the ith eigenpair. For instance, the row corresponds to λ_1 reads as follows. The global algorithm estimates the largest (in magnitude) $\ell = 56515$ components of the eigenvector v_1 accurately (see (19)). This represents accurate estimation of all components of v_1 with a magnitude that is greater or equal to $|v_1(\ell)| = 8.1 \cdot 10^{-316}$. The triple $(288, 40, 5)$ means that in order to achieve this accuracy, the global algorithm took a total of 288 power iterations, including 40 iterations for inverse power iterations (35 of IPI_f and 5 of IPI_v). The global algorithm took 5.1 seconds to compute the first eigenpair while achieving this high level of accuracy. In

Table 2: dixmaanl dataset. Results and complexity using the Global algorithm, eigs, and modified PI.

	Global				eigs		
	ℓ	$\lvert\tilde{x}(\ell)\rvert$	# iteration	time	ℓ	$\lvert\tilde{x}(\ell)\rvert$	time
1st	56515	8.1e-316	(288, 40, 5)	5.1	3311	3.7e-08	30
2nd	57294	8.7e-316	(319, 45, 6)	5.8	3883	3.2e-08	
3rd	57936	9.2e-316	(244, 40, 5)	4.6	4306	4.1e-08	
4th	58515	8.7e-316	(274, 45, 7)	5.2	4599	8.6e-08	
5th	59020	1.2e-315	(276, 45, 5)	5.4	5138	2.9e-08	
6th	59536	9.1e-316	(312, 45, 6)	6.3	5401	5.2e-08	

	Modified PI			
	ℓ	$\lvert\tilde{x}(\ell)\rvert$	time	# PI
1st	56460	1.9e-315	894	1.5e+06
2nd	57246	1.8e-315	1173	1.5e+06
3rd	57896	1.7e-315	1442	1.5e+06
4th	58472	1.7e-315	1718	1.5e+06
5th	58988	2.0e-315	1998	1.5e+06
6th	59480	2.0e-315	2292	1.5e+06

comparison, the Matlab eigs function, which computes all 6 top eigenpairs all at once, has a much inferior eigenvector accuracy. Only the largest $\ell = 3311$ components of estimated v_1 achieve the desired accuracy of (19) and these components are at least $|v_1(\ell)| = 3.7 \cdot 10^{-8}$ in magnitude. The total computation time of the eigs function for all 6 eigenpairs is 30 seconds. This is comparable with the total computation time of the global algorithm, however, with a significantly lower level of eigenvector accuracy. For the modified PI, we see that it can achieve an accuracy that is comparable to that of the global algorithm. However, the computation time is significantly longer. Due to its slow convergence, it takes 894 seconds and a total of $1.5 \cdot 10^6$ PIs in order to attain a similar accuracy as that of the global algorithm. Similar observations are made for the estimations of the other 5 eigenpairs. The proposed global algorithm achieves the best accuracy with the shortest computation time. We note that the main difference between the Global algorithm and the modified PI is that the former uses inverse power iteration with adaptive shifts, whereas the latter uses standard power iterations. Our results shown that the proposed CS-LBE procedure for computing the variable shifts is crucial in accelerating the convergence speed of the algorithm.

In Table 3, for each eigenpair, we show the value of the shifts used in the Global algorithm. The shifts are generated using the CS-LBE procedure. For instance, the column labeled as "1st" lists 5 values of the shifts $z_i,\ i = 1, \cdots, 5$,

used in the estimation of the 1st eigenpair. We see that z_i approaches the true λ value (shown in the last row) quickly. For the first eigenpair, only 5 different shifts are needed. In comparison, for the 4th and the 6th eigenpair, more shifts 7, and 6, respectively, are needed.

Table 3: dixmaanl dataset. Shifts used by the Global algorithm.

	1st	2nd	3rd
z_1	317.2018759831095	317.0149029206981	317.0054220600994
z_2	317.0220587013249	317.0412999365110	317.0056140237707
z_3	317.0165531440067	317.0183499796456	316.9974443732343
z_4	317.0152788610227	317.0044840756761	316.9980602821128
z_5	317.0152899359775	317.0057627487807	316.9980633932562
z_6		317.0058090643057	
λ	317.0152899359881	317.0058090659085	316.9980633932568

	4th	5th	6th
z_1	316.9976763951934	316.9908430604246	316.9846066377027
z_2	317.0174070334262	316.9924907259373	317.0002253316557
z_3	316.9879937432648	316.9843539028472	316.9767242599030
z_4	316.9896903234464	316.9849885385036	316.9784124921462
z_5	316.9910317369073	316.9849936226933	316.9791679223474
z_6	316.9912298325970		316.9791911036812
z_7	316.9912300516546		
λ	316.9912300516546	316.9849936226963	316.9791911040992

roadNet-CA dataset

For this dataset, the dimension of the matrix is $N = 1971281$. This matrix corresponds to a graph of the road network of California. Each element is either 0 or 1. The sparsity pattern of this matrix is shown in Fig. 4(b). The number of nonzero elements in the matrix is nz = 5533214, see Fig. 4 (b). In Table 4, we show detailed comparisons of the three methods in terms of the accuracy of the eigenvector, the complexity, and the running time. We see that the Global algorithm reaches very good accuracy in terms of ℓ and $|\tilde{x}(\ell)|$ for all 6 eigenpairs. Due to the increased matrix dimension, the computation time increases compared to that of the dixmaanl dataset. The eigs function can compute the top 6 eigenpairs quickly, using only a total of 38 seconds, but with a much inferior accuracy in ℓ and $|\tilde{x}(\ell)|$. The modified PI algorithm can achieve a very good accuracy for the top 3 eigenpairs, despite a longer computation time for using a high number of PI. The accuracy of the remaining 3 eigenpairs is much worse for the given number of PI.

Table 4: roadNet-CA dataset. Results and complexity using the Global algorithm, eigs, and modified PI.

	Global				eigs		
	ℓ	$\lvert\tilde{x}(\ell)\rvert$	# iterations	time	ℓ	$\lvert\tilde{x}(\ell)\rvert$	time
1st	1933344	2.8e-317	(244, 35, 3)	309	1543	8.7e-10	38
2nd	1926704	3.0e-317	(245, 35, 3)	322	1413	1.1e-09	
3rd	1957027	2.8e-295	(226, 30, 3)	276	2004	1.0e-09	
4th	1948213	2.2e-317	(243, 30, 3)	285	2190	5.3e-10	
5th	1956156	2.3e-317	(242, 30, 3)	293	2409	2.9e-10	
6th	1923583	2.7e-317	(282, 30, 2)	296	1648	7.8e-10	

	Modified PI			
	ℓ	$\lvert\tilde{x}(\ell)\rvert$	# PI	time
1st	1933452	2.0e-317	1.0e+04	381
2nd	1926900	2.0e-317	2.5e+04	1432
3rd	1957027	2.8e-295	2.7e+04	2062
4th	49653	1.6e-33	5e+04	4700
5th	62901	1.8e-33	7e+03	776
6th	1923767	1.9e-317	5.0e+04	6671

Table 5: roadNet-CA dataset. Computed top 6 eigenvalues using the Global algorithm, eigs, and modified PI.

	global	eigs	PI
λ_1	4.638361867351406	4.638361867351387	4.638361867351406
λ_2	4.527027931848926	4.527027931848909	4.527027931848924
λ_3	4.451588326941737	4.451588326941750	4.451588326941737
λ_4	4.390275021532836	4.390275021532792	4.390275021532837
λ_5	4.383736144475813	4.383736144475774	4.383736144475815
λ_6	4.325729176980614	4.325729176980572	4.325729176980615

In Table 5, we show the estimated top 6 eigenvalues using the three algorithms. They all find similar eigenvalues.

In Table 6, we show the shifts produced by the CS-LBE procedure. We observe that, despite the higher dimension of this dataset, the shift values converge to the eigenvalues quickly. Up to 3 shift values are sufficient to approach the eigenvalues.

Acknowledgments The first author thanks Professors Jia, Z.G. and Pang, H.K. and their teams for the fruitful discussions, verifying and suggestions which improve the quality of the paper. The teams are now continuously working on applications of the proposed method, such as computing the sparse principal components of medical

Table 6: roadNet-CA dataset. Shifts used by the Global algorithm.

	1st	2nd	3rd
z_1	4.651095152690492	4.541091827266276	4.457490006778257
z_2	4.638369301398686	4.527034278350056	4.451618990253862
z_3	4.638361867350882	4.527027931841913	4.451588326915770
λ	4.638361867351406	4.527027931848926	4.451588326941737

	4th	5th	6th
z_1	4.390768119815626	4.384210430746412	4.325729209088518
z_2	4.390275047542154	4.383736186751131	4.325729176980588
z_3	4.390275021532815	4.383736144475802	
λ	4.390275021532836	4.383736144475813	4.325729176980614

images and others. Special thanks to Professor Xie, Y.C. for his support and great help during the period the author worked at the university. Research supported in part by National Natural Science Foundation of China (Grant Nos. 12090011, 11771046), National Key R & D Program of China (No. 2020YFA0712900), the project from the Ministry of Education in China, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] Chen, M.F. (2005). *Eigenvalues, Inequalities, and Ergodic Theory*. London: Springer.

[2] Chen, M.F. (2016). *Efficient initials for computing maximal eigenpair*. Front. Math. China 11(6): 1379–1418.

[3] Chen, M.F. (2017a). *Global algorithms for maximal eigenpair*. Front Math China 12(5): 1023–1043.

[4] Chen, M.F. (2017b). *Trilogy on computing maximal eigenpair*. In: Yue, W., Li, Q. L., Jin, S., Ma, Z., eds. "Queueing Theory and Network Applications". QTNA 2017. Lecture Notes in Comput. Sci., Vol. 10591. Cham: Springer, 312–329.

[5] Chen, M.F. (2018). *Hermitizable, isospectral complex matrices or differential operators*. Front Math China 13(6): 1267–1311.

[6] Chen, M.F., Jia, Z.G. and Pang, H.K. (2021). *Computing top eigenpairs of Hermitizable matrix*. Front. Math. China 16(2): 345–379.

[7] Chen, M.F., Li, Y.S. (2019). *Improved global algorithms for maximal eigenpair*. Front. Math. China 14(6): 1077–1116.

[8] Lei, Q., Zhong, K., Dhillon, I.S. (2016). *Coordinate-wise Power Method.* Proc. Advances in Neural Information Processing Systems, 2056–2064.

[9] Press, W.H. et al. (2007). *Numerical Recipes—The Art of Scientific Computing,* 3rd ed. Cambridge Univ. Press.

[10] Varga, R.S. (2004). *Geršgorin and His Circles*. Springer.

[11] Wang, J.L., Wang, W.R., Garber, D., Srebro, N. (2018). *Efficient coordinate-wise leading eigenvector computation*. Proc. Algorithmic Learning Theory, PMLR, 806–820.

[12] Xu, Z.Q. (2018). *Gradient descent meets shift-and-invert preconditioning for eigenvector computation*. Proc. Advances in Neural Information Processing Systems, 31: 2825–2834.

CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2023-0002

Vol. **6**, No. 1, pp. 96-1682
March 2025

Economic ProductRank and Quantum Wave Probability

Mu-Fa Chen*

Research Institute of Mathematical Science, Jiangsu Normal University, Xuzhou, 221116, PRC
School of Mathematical Sciences, Beijing Normal University, Beijing 100875, P.R. China

Abstract. This note discusses a long debated question: whether there is randomness in quantum mechanics or not? A. Einstein's view on the question is "God does not throw dice". Our starting point for the discussion is the classification of products in the economic system, called ProductRank, which seems an analog of the "principal component analysis" in statistics. But the former is much more elaborate than the latter. Interestingly, we find an intrinsic common point among economic system, statistics and quantum mechanics, which then leads to a successful classification of the products in economy, as well as a mathematical view of "wave probability" in quantum mechanics. An application to the algorithm for eigenpair is included.

AMS subject classifications: 91B, 15B57, 81S.

Key words: Economics, ProductRank, Hermitizable, Quantum mechanics.

The problem mentioned above was motivated from M. Born's suggestion (1926) saying that the Schrödinger's wave function describes "waves of probability": "the square of the amplitude (of the wave function) represents the probability density of finding the particle in a certain place at a certain time"(cf. [9; p. 114]). Refer also to [4] for a survey and references therein. Here we introduce a mathematical view of Born's annotation based on our recent study on the classification of the products in economic system. For which an advanced probabilistic tool — Markov chains is adopted. However, as can be seen very soon that there is essentially no randomness in the story.

The main results of the note are stated in the next two sections. First on economics (§1), then on quantum, and finally on algorithm (§2). Their proofs are delayed to the last section (§3).

*Corresponding author. *Email address:* mfchen@bnu.edu.cn

1 Ranking the products in an economic system

Denote by $x=\left(x^{(1)},x^{(2)},\ldots,x^{(d)}\right)$ the vector of products we are interested in the economic system. Then the evolution of the system is mainly determined by its structure matrix $A=(a_{ij}:i,j=1,2,\cdots,d)$ which means that to produce one unit of the ith product, one requires a_{ij} units of the jth product. Thus, once we have an input x_0, then the output x_1 in one year satisfies the equation $x_0=x_1A$. In general, we have $x_0=x_nA^n$ and then

$$x_n=x_0A^{-n},\qquad n\geqslant 1,$$

assuming that A is nonnegative, irreducible and invertible. This is the well-known input-output method.

Denote by $\rho(A)$ the maximal eigenvalue of A, the corresponding left- or right-eigenvectors are denoted by u (row) and v (column), respectively.

From the above simplest idealized model, one can already see the main points of L.K. Hua's optimization of global economic system (the result was appeared firstly in 1984; refer to [5,8] for a short history on the topic):

- For fastest growing rate of the system, the optimal solution of the initial input is $x_0=u$, for which we have $x_n=x_0\rho(A)^{-n}, n\geqslant 1$.
- If A has at least one positive diagonal element, then to keep x_n to be positive for each $n\geqslant 1$, the optimal solution (actually the only one) is again $x_0=u$.

The first assertion above is not so surprising, simply an application of the Perron-Frobenius theorem plus the min-max strategy. The second one is the main contribution of Hua, never appeared before as far as we know. It is even more serious that the system will be collapsed exponentially fast once $x_0\neq u$. Therefore, it is important to know the classification of the products in the economic system: the pillar products, the intermediate products and the disadvantaged products, since the system can often be collapsed at some disadvantaged products.

Following Google's PageRank (appeared in 1998), a natural way to ordering the products is using the maximal left-eigenvector u of A. However, since the matrix A in economy is quite far away from the matrix used in the network, where one has a nice graphic structure. Especially, it is far way to be a transition probability matrix. More seriously, the economic system is very sensitive, much more precise computations are required, and thus we should examine the corresponding ProductRank more carefully than the PageRank. Recall by Perron-Frobenius theorem, every nonnegative irreducible matrix A has three characteristics:

- Its maximal eigenvalue $\rho(A)$ is positive and simple.
- Its maximal left-eigenvector u is positive and one-dimensional.
- Its maximal right-eigenvector v is also positive and one-dimensional.

Note that the eigenvector u owns two of the above characteristics only, not three of them. To go further, we adopt a key transform: transforming A to a transition probability matrix P (which means that the elements of P are nonnegative and the sum of each row of P equals one).

Lemma 1. † ([1, 2, 5]) Given a positive vector w, denote by D_w the diagonal matrix with w as its diagonal elements. Next, define

$$A_w = D_w^{-1} \frac{A}{\rho(A)} D_w.$$

Then, we have

- A_w becomes a transition probability matrix P iff $w = v$.
- The maximal left-eigenvector of P is equal to $\mu := u \odot v$ (the vector consists of the products of the components of u and v). The normalized measure $\pi := \mu/(uv)$ is the stationary distribution of P: $\pi = \pi P^n, n \geqslant 1$.

The second assertion of Lemma 1 shows that the left-eigenvector (or the invariant measure) μ has combined the three characteristics of A together, and hence is more essential to describe the ProductRank of A, for which we adopt μ (or equivalently π) instead of the use of u mentioned above. Moreover, $u \odot v$ owns an important economic meaning: u represents the vector of the amount of each product, v represents the vector of the true value of each product in per unit [8; Chapter 1, § 7] (often different from the price in market). Thus, $u \odot v$ gives us the vector of the total true value of each product. Hence we now have the unified unit for different products. This shows that the ProductRank here is reasonable. Furthermore, from probabilistic point of view, the stationary distribution π, as the normalized one of μ has a very important property: it is the only stationary distribution of P. For A, we do have similar stationary property that $\mu = \mu(A/\rho(A))$, but not $\mu = \mu A$ except in the unusual case that $\rho(A) = 1$. Furthermore, for P, we have the ergodic theorem:

$$\lim_{n\to\infty} P^n = \mathbb{1}\pi, \tag{1}$$

where $\mathbb{1}$ is the column vector having constant 1 everywhere. The matrix on the right simply means that each row is the same vector π. However,

$$\lim_{n\to\infty} A^n = \lim_{n\to\infty} \rho(A)^n \left(\frac{A}{\rho(A)}\right)^n = \begin{cases} \infty & \text{if } \rho(A) > 1, \\ 0 & \text{if } \rho(A) < 1, \end{cases}$$

since as an application of (1), it is not difficult to check that

$$\lim_{n\to\infty} \left(\frac{A}{\rho(A)}\right)^n = vu \quad \text{(a finite positive matrix)}.$$

Therefore, P^n and A^n have completely different limiting behavior and hence have completely different stability. This is essential in the study on economy.

---

†Different to the published version, here all the propositions (lemmas, examples, et al) are laberated by unified single code. Similaely for formulas.

To show our ProductRank is meaningful, let us examine some practical examples. The first figure below is the ProductRank of 42 products produced by the input-output tables of China in 2017 (red), 2012 (blue) and 2007 (black). It covers 15 years of the economy in the country. For details, refer to Chapter 4 in the monograph [7]. We produce in our country one table in each of 5 years. Surprisingly, the shapes of the curves are quite closed each other. Here is a remark about the input-output tables. Keeping the 2012's one at hand, the others are slightly modified for their consistence. Thus, the 24th product is missed in 2007, which is somehow reasonable since in the earlier period, the statistical data may be missed or less completed. Hence there is a dotted black line between 23th and 25th products. According to the blue curve, the top 6 products are marked with blue circled numbers. It is clear that the blue and black curves have the same top 6 ranks among them. The main difference to them is the red curve, for which the top product is the 20th one (communication, computer, etc.), but not the 12th (chemical products). The reason is clear that the mobile phone was rapidly developed during 2012–2017. The ranks 30, 33, 34, 35 are increasing in the three period.

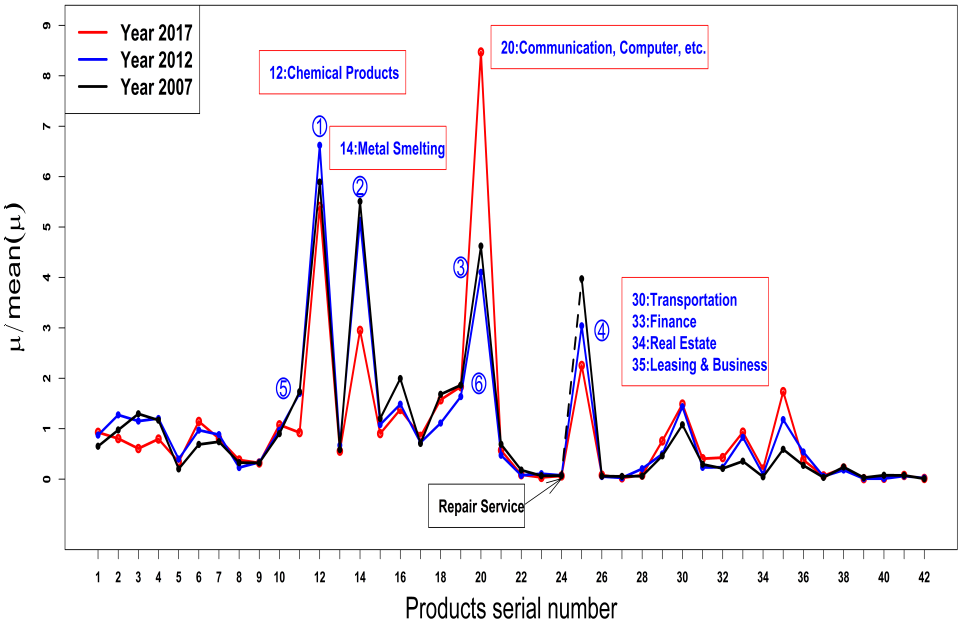


Figure 1 ProductRank by μ of 42 products in 2017, 2012, 2007

The next two figures are the cumulative distribution function produced from π. We order the components $(p_k : k = 1,\ldots,42)$ of π in increasing order $p_1 < p_2 < \ldots < p_{42}$. Then we obtain the discrete cumulative distribution function as $F(n)$: $F(0) = 0$, $F(n) = \sum_{j=1}^{n} p_j$, $F(42) = 1$. From Figure 2, one sees that the top 6 products occupy the above half of the probabilistic distribution. This is reasonable since they are the pillar products. On the other hand, from Figure 3, one may choose the first 17 or 10 products as the disadvantaged

products. These figures show the value of ProductRank for understanding the economic systems.

The Figure 3 is a local part of the above one.

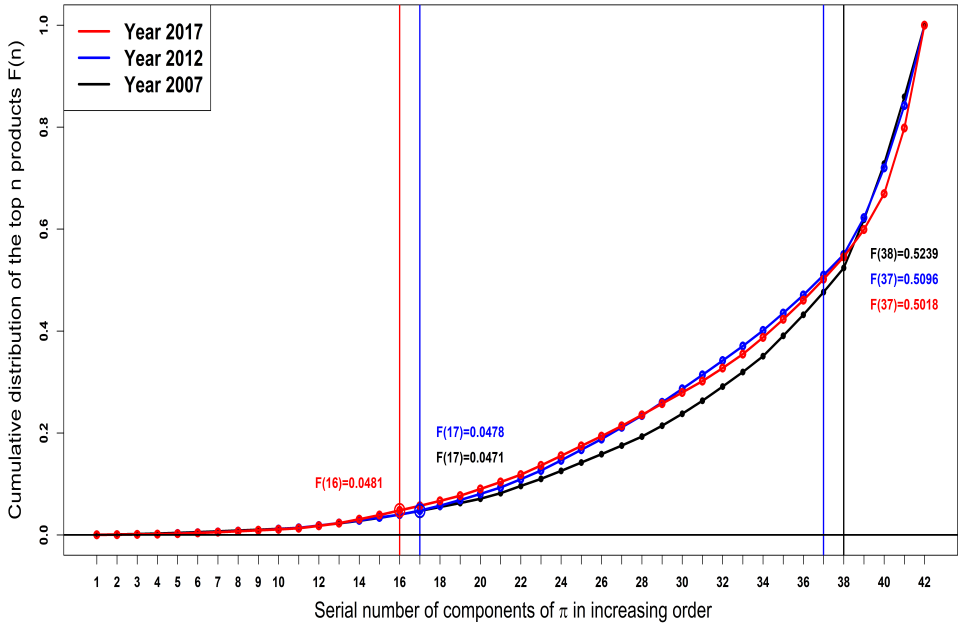


Figure 2 Cumulative distribution function of products in 2017, 2012, 2007

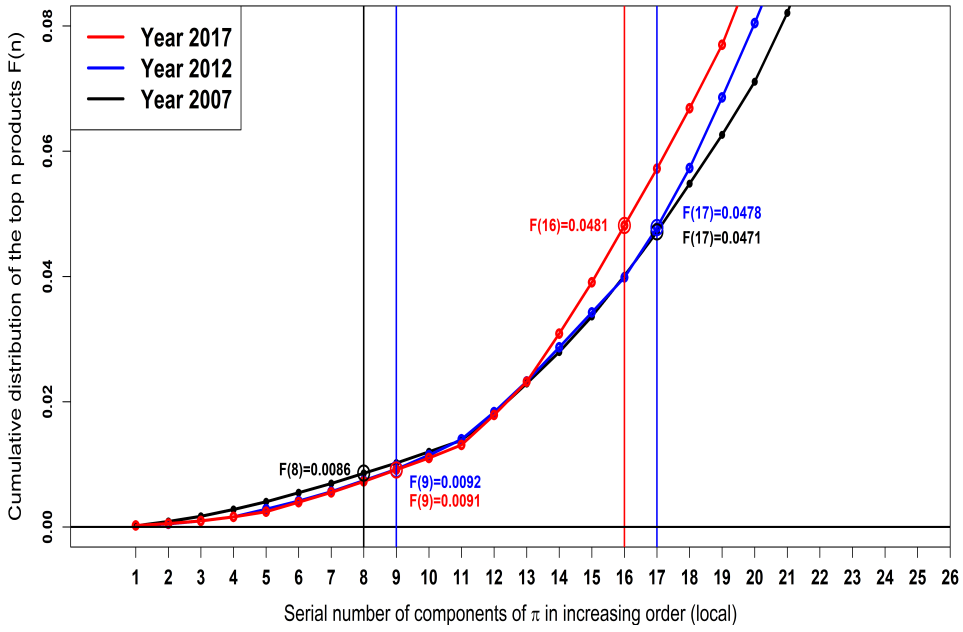


Figure 3 Cumulative distribution function of products in 2017, 2012, 2007

We have seen the application of the transform $A \to P$ to the ProductRank. Actually, the technique was firstly used in [2] to prove the Hua's collapse theorem for economic system. Actually, it has much more application to the analysis on economics, including stability analysis, forecast and adjustment, algorithms of eigenpair, optimization of economic structure, and so on. Refer to [1, 5] and references therein for details. The three figures used here are taken from [10], updated partially from [1]. A new theory of the economic optimization is presented in the monograph [7].

2 Hermitian and hermitizable matrix

We now go to complex matrix. Certainly, in such a general setup, we may have some generalized version of the Perron-Frobenius theorem, but the known results are quite restricted. However, a key point in the last section: the transform from A to P still has a meaning.

Definition 2. A complex matrix A is called SR1-matrix, if $A\mathbb{1}=\mathbb{1}$. That is, the sum of each row of A equals one.

Before moving further, we note that by using a shift if necessary, we can assume that the eigenvalue λ in the study is not zero (cf. Lemma 8 (1) in §3). Thus, in what follows we assume that $\lambda \neq 0$. We may also assume if necessary that λ is simple, for instance in sorting the ProductRank.

Now, as an analog of Lemma 1, we have the following result.

Lemma 3. Suppose that the matrix A has the right-eigenpair $(\lambda, v): Av=\lambda v$ with no zero component of v. For given vector w with no zero component, define

$$R_w = D_w^{-1} \frac{A}{\lambda} D_w. \tag{2}$$

Then, we have

(1) R_w is a SR1-matrix iff $w=v$.

(2) R_v has the left-eigenpair $(\lambda, u\odot v): (u\odot v)R_v = \lambda(u\odot v)$.

Applying Lemma 3 to a Hermite matrix A, since for which, the corresponding $u=\bar{v}$ (the conjugate of v) and so $u\odot v=\bar{v}\odot v$, we obtain the following result.

Corollary 4. Let A be Hermitian satisfying the hypotheses of Lemma 3, then the corresponding left-eigenvector of R_v equals $\bar{v}\odot v$.

In words, the vector $\bar{v}\odot v$ combined the three characters of the Hermitian A, its eigenvalue λ and the corresponding left- and right-eigenvectors u and v. Since $\bar{v}\odot v$ represents the square of the amplitude of the wave function (equivalently, the eigenvector) v of A, we have explained the key reason why we should use "the square of the amplitude" rather than "the amplitude" only for the Born's annotation in the context of matrix mechanics. The vector $\bar{v}\odot v$ describes the ProductRank which provides not only the sort of the products but also a suitable value to each of the product, up to a factor. Equivalently, one can replace $\bar{v}\odot v$ by its normalized probability measure π and talk about the probability of a product (particle) appears, which is the same as Born's suggestion cited at the beginning of the note. Anyhow, it is just an interpretation of the same thing in two different languages, there is no objective randomness here. It is just like Einstein said, "God does not throw dice". Note that in the special case that the matrix A is nonnegative, symmetric, and $\lambda=\rho(A)$, even the "the square of the amplitude" or "the amplitude" provide the same ordering but they often have very different amplitudes. Next, since the equivalence of the matrix or wave mechanics, it follows that the same conclusion holds for the wave mechanics.

To conclude this section, we study an extension of Corollary 4.

Definition 5. A complex matrix A is said to be Hermitizable if there is a positive measure μ such that $D_\mu A = A^H D_\mu$ (A^H is the conjugate and transpose of A). Equivalently, $\hat{A} := D_\mu^{1/2} A D_\mu^{-1/2}$ is Hermitian.

Refer to [3] for a criterion for the Hermitizability and for the construction of the measure μ, or refer to [6] for a short review on the topic.

Lemma 6. Given a Hermitizable matrix A, we have $\hat{A}$ as shown in Definition 5. Corresponding to A and $\hat{A}$, define R_v and $\hat{R}_{\hat{v}}$ (where $\hat{v}$ is the right-eigenvector of $\hat{A}$), respectively, as in Lemma 3. Then, we have $R_v = \hat{R}_{\hat{v}}$. Hence both of them have the same left-eigenvector $\mu\odot\bar{v}\odot v$. Furthermore, the left-eigenvector of A equals $\mu\odot\bar{v}$.

It is the position to remark that since the eigenvectors u and v of A are symmetric in the vector $u\odot v$ given in Lemmas 1 and 3, as well as in the one $\mu\odot\bar{v}\odot v$ (where $\bar{v}=u$) given in Lemma 6, our ProductRank is invariant under the transform: $A\to$ transpose of A.

The last result of the note is an application of the approach introduced above to the computation of eigenpairs. Actually, this is the original way in the note we come to quantum from economy. The classical approach to the goal is the power iteration (PI) and the inverse power iteration (IPI). Both are iterations of the (right-)eigenvector. However, as we have seen from Figure 1 for instance, the eigenvector is usually very complex, oscillating. The question is: is it possible to reduce the original matrix to the one having nearly constant eigenvector? If so, then the computation should be much easier,

simply start at the trivial initial vector $\mathbb{1}$. Once again, the main problem, as those discussed above, is that one does not know at the beginning a way to solve such a simple question. However, once walked up, the solution becomes quite simple: recall that we have a tool to reduce the eigenvector to be a constant $\mathbb{1}$ for a new matrix, that is Lemma 3. The next lemma is an extension of [5; Lemma 16 in §6] where it is called the second quasi-symmetrizing technique.

Lemma 7. As a modification of (2), define

$$Q_w = D_w^{-1} A D_w \tag{3}$$

where $v = \big(v^{(1)}, v^{(2)}, \ldots, v^{(d)}\big)$ is the right-eigenvector of A corresponding to the eigenvalue λ and $w = \big(w^{(1)}, w^{(2)}, \ldots, w^{(d)}\big)$ is a vector having no zero components. Then, once

$$\max_k \left| \frac{w^{(k)}}{v^{(k)}} - 1 \right| < \varepsilon, \tag{4}$$

for sufficient small ε, then the right-eigenvector $w^{-1} \odot v$ of Q_w is nearly a constant vector.

Recall that the transform: either $A \to P$ or $A \to R_v$, both need to compute the right-eigenvector of A, and often require high precision. This is one of the typical applications of the above lemma. Its main function is to accelerate the convergence speed of the iterative method. This is especially important for large matrices, because the inverse power iteration used for acceleration may fail. More seriously, one may meet too large/small numbers which can not be handled by machine directly or ignored by software. Note that Q_w defined in (3) has only $2d^2$ pointwise products, very low computational complexity. The main steps of usage of the lemma are as follows:

- First use PI or IPI to iterate enough times or use software to compute an approximation of the eigenvector, which is recorded as w_0. In the case that w_0 is very close to a constant, then one can terminate the computation. Otherwise, go to the next step.
- Compute Q_{w_0} by (3). Take $\mathbb{1}$ as the initial value, and then use PI or IPI iteration for Q_{w_0} to get w_1. Again, if w_1 is very close to a constant, then one can terminate the computation.
- Repeat the above steps until the resulting vector is very close to the constant vector. Suppose we have stopped the computation at w_m, say w_3, then by Lemma 7, we can compute the required eigenvector v of A by the formula:

$$v = w_3 \odot w_2 \odot w_1 \odot w_0. \tag{5}$$

For more details, refer to [5; §6] and [1; Example 9].

To conclude this section, we mention that in [3], we have proved that the spectrum of a Hermitizable matrix can be described by the spectrum of a special class of tridiagonal transition probability matrices. Once again, we have used the probabilistic language to describe the conclusion. However, there is no randomness at all, as mentioned in [3, 6]. For information along this direction, refer to the papers just cited and references therein.

3 Proofs of the results

Let us start at an elementary result.

Lemma 8. Consider the right-eigenpair only.

(1) **Shift transformation**: The transform $\hat{A}=A+\gamma I$ (γ is a constant) makes the eigenpair (λ,g) of A to the eigenpair $(\lambda+\gamma,g)$ of $\hat{A}$. That is, the eigenvalue is changed from λ to $\lambda+\gamma$ but the eigenvector becomes the same.

(2) **Similar transformation**: $\hat{A}=B^{-1}AB$, where B is invertible. The eigenpair (λ,g) of A is transformed to $(\lambda,B^{-1}g)$ of $\hat{A}$. That is, the eigenvalue remains the same but the eigenvector is transformed to $B^{-1}g$.

Proof. For the first assertion, note that

$$Ag=\lambda g \Longleftrightarrow (A+\gamma I)g=(\lambda+\gamma)g \Longleftrightarrow \hat{A}g=(\lambda+\gamma)g.$$

For the second one, note that

$$\hat{A}\hat{g}=\lambda\hat{g} \Longleftrightarrow B^{-1}AB\hat{g}=\lambda\hat{g} \Longleftrightarrow A(B\hat{g})=\lambda(B\hat{g}). \quad \square$$

Proof of Lemma 3 and Lemma 1.
Without loss of generality, assume that $\lambda=1$ for simplicity. Then for Lemma 3, we have

$$R_w\mathbb{1}=D_w^{-1}AD_w\mathbb{1}=D_w^{-1}Aw\stackrel{?}{=}\mathbb{1} \Longleftrightarrow Aw\stackrel{?}{=}D_w\mathbb{1}=w.$$

The last part of the line above gives us the required assertion: $w=v$. Having proved part (1) of the lemma, the part (2) follows by a simple computation.

Now, by Lemma 3, we obtain Lemma 1 immediately. $\square$

Proof of Lemma 6.
Applying Lemma 8 (2) to $B=D_\mu^{-1/2}$, it follows that for fixed eigenvalue λ, the right-eigenvector v deduces the one of $\hat{A}$: $\hat{v}=D_\mu^{1/2}v=\mu^{1/2}\odot v$. By Corollary 4, the corresponding left-eigenvector of $\hat{A}$ is $\hat{v}^H=\bar{v}D_\mu^{1/2}$. Hence, the first assertion of the lemma comes from

$$R_v=D_v^{-1}\frac{A}{\lambda}D_v=D_v^{-1}\frac{D_\mu^{-1/2}\hat{A}D_\mu^{1/2}}{\lambda}D_v=D_{\mu^{1/2}\odot v}^{-1}\frac{\hat{A}}{\lambda}D_{\mu^{1/2}\odot v}=D_{\hat{v}}^{-1}\frac{\hat{A}}{\lambda}D_{\hat{v}}=\hat{R}_{\hat{v}}.$$

Since $\hat{A}$ is Hermitian, by Corollary 4, the left-eigenvector of $\hat{R}_{\hat{v}}$ equals $\hat{v}^H \odot \hat{v} = \mu \odot \bar{v} \odot v$. This proves the second assertion of the lemma. Then the last assertion follows, which can be also verified directly:

$$(\bar{v}D_\mu)A = \left(\hat{v}^H \hat{A}\right) D_\mu^{1/2} = \left(\lambda \hat{v}^H\right) D_\mu^{1/2} = \lambda(\bar{v}D_\mu). \quad \square$$

Proof of Lemma 7.
By (3), we have $Q_{\mathbb{1}} = A$. Since the eigenvalue λ of A is fixed, for each w, Q_w has the same λ. In the proof below, we consider only the right-eigenvector of Q_w, denoted by g_w. Clearly, we have $g_{\mathbb{1}} = v$. Applying Lemma 8 (2) to $B = D_w$, it follows that the eigenvector of Q_w equals $D_w^{-1} v = w^{-1} \odot v$, which is a nearly constant vector by condition (3). $\square$

Acknowledgements: The author thanks Professors Ai-Hui Zhou, Wei-Hai Fang, Ying-Chao Xie, Zhi-Gang Jia, Zhong-Wei Liao, Ting Yang, and Qin Zhou (who made the figures used in the note) for their discussions, suggestions, and corrections of the earlier version of the note. This study is supported by the National Natural Science Foundation of China (Project No.: 12090011), National key R & D plan (No. 2020YFA0712900), the "double first class" construction project of the Ministry of education (Beijing Normal Univ.), and the advantageous discipline construction project of Jiangsu Universities.

References

[1] Chen, B., Chen, M.F., Xie, Y.C., Yang, T., Zhou,Q. (2022). *Ordering of products and optimization of structure matrix in economy* (in Chinese). Chin. J. Appl. Prob. Stat. 2022, 38(4): 475–504.

[2] Chen, M.F. (1992). *Stochastic model of economic optimization* (in Chinese). Chin. J. Appl. Probab. and Statis., (I): 8(3), 289–294; (II): 8(4), 374–377. The paper was received by the journal in 1989, but published only in 1992.

[3] Chen, M.F. (2018). *Hermitizable, isospectral complex matrices or differential operators*. Front Math China 13(6): 1267–1311.

[4] Chen, M.F. (2021). *A new mathematical perspective of quantum mechanics* (In Chinese). Adv. in Math. (China) 50(3): 321–334.

[5] Chen, M.F. (2022a). *New progress on L.K. Hua's optimization theory of economics* (in Chinese). Chin. J. Appl. Prob. Stat. 2022, 38(2): 159–178.

[6] Chen, M.F. (2022b). *Hermitizable, isospectral matrices or differential operators*. Chaper 3 in Dirichlet Forms and Related Topics— The Festschrift in Honor of Masatoshi Fukushima's Beiju 2022. eds. Chen, Z.Q. et al, 45–55. Springer Proc. Math. & Statis., vol. 394.

[7] Chen, M.F., Xie, Y.C., Chen, B., Zhou,Q., Yang, T. (2024). *Demonstration on the New Theory of L. K. Hua's Economic Optimization* (in Chinese). Beijing Normal Univ. Press, Beijing.

[8] Hua, L.K. (1987). *On the Mathematical Theory of Globally Optimal Planned Economic Systems* (in Chinese). China Finan. & Econ. Publ. House.

[9] Montwill, A. & Breslin, A. (2012). *The Quantum Adventure: Does God Play Dice?*. Imperial College Press & World Sci. Publ.

[10] Yang, T. Chen, B., Zhou, Q. (2024). Demonstrational examples based on the new theory of L. K. Hua's economic optimization (in Chinese). Chin. J. Appl. Prob. Stat. Vol. 40, No. 4, 663–683.