Research Report

KSTS/RR-17/001
January 27, 2017
(Revised July 24, 2017)

Inexact Shift-invert Rational Krylov Method for Evolution Equations

by

Yuka Hashimoto
Takashi Nodera

Yuka Hashimoto
School of Fundamental Science and Technology
Keio University

Takashi Nodera
Department of Mathematics
Keio University

Department of Mathematics
Faculty of Science and Technology
Keio University

3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522 Japan

KSTS/RR-17/001
January 27, 2017
(Revised July 24, 2017)

Inexact Shift-invert Rational Krylov Method for Evolution Equations

Yuka Hashimoto* Takashi Nodera†

July 24, 2017

Abstract

Linear and nonlinear evolution equations have been formulated to address problems in various fields of science and technology. Recently, a method called exponential integrator has been attracting some attention for solving these equations. It requires the computation of matrix functions repeatedly. For this computation, a new method called the Inexact Shift-invert Rational Krylov method is explored. This method provides the shifts which realizes a faster convergence than the existing method called Shift-invert Arnoldi method. Furthermore, it realizes efficient computation, while guaranteeing accuracy.

Key Words. Inexact Shift-invert Rational Krylov, ϕ-function, exponential integrator
AMS(MOS) subject classifications. 65F60, 65M22

1 Introduction

Evolution equations are used in various fields of science and technology, e.g., the heat equation in building physics [26] and the Burgers equation in fluid mechanics [19]. Let $\Omega \subseteq \mathbb{R}^d$ be an open set, $T > 0$, $l \in \mathbb{N}$, and $\mathcal{V} \subseteq L^2(\Omega)$ be the Hilbert space. Let $\mathcal{D}$ be a linear or nonlinear differential operator on $\mathcal{V}$. Explore $u \in C^l((0,T]) \times \mathcal{V}$ which satisfies

$$\frac{\partial^l u}{\partial t^l} = \mathcal{D}u \tag{1}$$

with some appropriate initial and boundary conditions. A different algebraic equation is derived from a spatial discretization with the finite element method or finite difference method:

$$\begin{cases} M\dot{y}(t) = F(y(t)), \\ y(0) = v, \end{cases} \tag{2}$$

where $M \in \mathbb{R}^{n \times n}$, and F is a vector valued function. It is assumed that M is invertible.

*School of Fundamental Science and Technology, Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa, 223-8522, Japan.
yukahashimoto@math.keio.ac.jp

†Department of Mathematics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa, 223-8522, Japan.
nodera@math.keio.ac.jp

If $\mathcal{D}$ is linear and does not depend on t, F is represented as $F(y) = Ly + c$, where $L \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$. In this case, equation (2) is the linear ordinary differential equation of the first order, and its analytical solution is represented as:

$$y(t) = \phi_0(tM^{-1}L)(v + L^{-1}c) - L^{-1}c, \tag{3}$$

where $\phi_0(z) := e^z$ [13]. On the other hand, if $\mathcal{D}$ is nonlinear or depends on t, time discretization is also needed for integrating $M^{-1}F(y(t))$ and finding solution $y(t)$. The exponential integrator [15, 17, 18] is currently the popular method for time integration [13]. In general, at each step, F is rearranged as $F(y) = L_i y(t) + n_i(y)$. For the 1-step method, the scheme is computed as follows:

$$\begin{aligned} Y_{ik} &= \phi_0(c_k \Delta t M^{-1} L_{i+1}) y_i + \Delta t \sum_{l=1}^{k-1} a_{kl}(\Delta t M^{-1} L_{i+1}) M^{-1} n_i(Y_{il}), \\ y_{i+1} &= \phi_0(\Delta t M^{-1} L_{i+1}) y_i + \Delta t \sum_{k=1}^{s} b_k(\Delta t M^{-1} L_{i+1}) M^{-1} n_i(Y_{ik}), \end{aligned} \tag{4}$$

where $v \in \mathbb{R}^n$, Δt is the step size of time, and a_{kl}, b_k are coefficients which consist of ϕ-functions. ϕ-function is defined as

$$\begin{aligned} &\phi_0(z) := e^z, \\ &\phi_k(z) := \frac{\phi_{k-1}(z) - \frac{1}{(k-1)!}}{z}, \qquad k = 1, 2, \ldots. \end{aligned}$$

For the r-step method, the scheme is computed as follows:

$$y_{i+1} = \phi_0(\Delta t M^{-1} L_{i+1}) y_i + \Delta t \sum_{k=1}^{r-1} \gamma_k(\Delta t M^{-1} L_{i+1}) M^{-1} \nabla^k N_i \,, \tag{5}$$

where $N_i := n_i(y_i)$, and $\nabla^k N_i$ and $\gamma_k(z)$ are defined recursively by

$$\begin{aligned} &\nabla^0 N_i := N_i \,, \qquad \nabla^{k+1} N_i := \nabla^k N_i - \nabla^k N_{i-1} \,, \\ &\gamma_0(z) = \phi_1(z), \quad z\gamma_k(z) + 1 = \sum_{l=0}^{k-1} \frac{1}{k-l} \gamma_l(z). \end{aligned}$$

Various methods for computing matrix ϕ-functions have been developed [4, 7, 15, 20, 21]. The Krylov subspace methods are efficient, because the matrices resulting from the spatial discretization of problem (1) usually become large. The most simple and well-known method is the Arnoldi method. According to Hochbruck and Lubich [15, Theorem 5], Arnoldi method may require a number of iterations if the numerical range of $\Delta t M^{-1} L_{i+1}$ is widely distributed. The matrices coming from the spatial discretization of problem (1) often have a wide numerical range, so the Arnoldi method may takes a lot of time until convergence for computing ϕ-functions in the exponential integrator. In order to resolve this issue, the Shift-invert Arnoldi method (SIA) was proposed by Novati [21], and the Rational Krylov method (RK) was proposed by Beckermann and Reichel [1]. RK is a generalization of SIA, and it was also proposed by Güttel [11] and Göckler [9]. For

SIA and RK, it can be shown that their convergences are independent of the width of the numerical range of $\Delta t M^{-1} L_{i+1}$ [9, 10, 21]. However, the SIA and RK have drawbacks. Firstly, solving a linear equation in each step is necessary. The computation cost of solving this linear equation is significant in the SIA and RK. For matrix exponential, to address this issue, Eshof and Hochbruck [27] proposed the stopping criterion for solving the linear equations. Gang et al. [8] also proposed the stoppping criterion for Toeplitz matrix exponential. Hashimoto and Nodera [13] proposed Inexact Shift-invert Arnoldi method (ISIA) and gave the stopping criterion for general matrix ϕ-function with different approach using the decay property of the elements of the matrix. Using these stopping criteria, we can solve the linear equations efficiently while guaranteeing the accuracy of the solution. However, these methods are only for SIA or for matrix exponential.

The second shortcoming of the SIA and the RK is the difficulty of choosing the appropriate shifts. Since RK needs different shifts in every step of the Krylov process, choosing the appropriate shifts is integral. The methods for choosing the appropriate shifts in SIA and RK for ϕ_0 and other functions have been discussed at length, for example [6, 12, 24]. However, the optimization problem must be solved for each shift, or they are only suitable for ϕ_0, and not for general ϕ functions. Göckler [9] proposed a simple way of choosing the shifts for general ϕ-functions of nonsymmetric matrices. However, this shifts involved complex values. Thus, if matrices M and L are real, we must treat complex values due to the shifts. This results in increasing the computational cost needlessly.

To resolve these issues, a new method called the Inexact Shift-invert Rational Krylov method (ISIRK) is proposed in this study. The Shift-invert Rational Krylov method (SIRK) is used to solve the second problem. The appropriate shifts for ϕ-functions in real value are determined in a simple way, and this choice of shifts results in a faster convergence than SIA. In addition, the Inexact Shift-invert Arnoldi method (ISIRK) is used to solve linear equations in the SIRK efficiently. The similar discussion for the ISIA is also valid for the SIRK. ISIRK makes the computation of ϕ-functions efficient.

1.1 Notation

The norm is defined as $\|\cdot\| = \|\cdot\|_2$, and the 2-norm condition number of matrix A is defined as $\kappa(A)$. e_j represents the jth column of identity matrix I. The $n \times n$ identity matrix is also represented as I_n when its dimension is emphasized. Let $\mathbb{C}^+ := \{z \in \mathbb{C} \mid \Re(z) > 0\}$, and $W(A) := \{u^* A u \mid u \in \mathbb{C}^n,\ \|u\| = 1\}$ be the numerical range of matrix A.

2 Krylov subspace methods for computing ϕ-functions

In this paper, $\phi_k(A)v$ is computed to simplify the notation. Throughout this and the next section, it is assumed that $W(\gamma_m I - A) \subseteq \mathbb{C}^+$ for all m, and $W(\gamma I - A) \subseteq \mathbb{C}^+$.

2.1 Shift-invert Arnoldi method (SIA)

Let $\beta = \|v\|$, and $v_1 = v/\beta$ be the initial vector. The m-step Shift-invert Arnoldi or Restricted-denominator (RD) Rational Arnoldi process is:

$$\begin{aligned} h_{j+1,j} v_{j+1} &= (\gamma I - A)^{-1} v_j - \sum_{k=1}^{j} h_{k,j} v_k \,, \\ h_{k,j} &= v_k^* \left[(\gamma I - A)^{-1} v_j - \sum_{l=1}^{k-1} h_{l,j} v_l \right] \qquad (k = 1, \ldots, j), \\ h_{j+1,j} &= \left\| (\gamma I - A)^{-1} v_j - \sum_{l=1}^{j} h_{l,j} v_l \right\| \qquad (j = 1, \ldots, m), \end{aligned}$$

where $\gamma \in \mathbb{C}$ is a shift. This relation is expressed with matrices as:

$$V_m^* (\gamma I - A)^{-1} V_m = H_m, \tag{6}$$

where $V_m = [v_1 \ \cdots \ v_m]$ is an $n \times m$ matrix whose columns are orthonormal, and H_m is an $m \times m$ upper Hessenberg matrix. $\{v_1, \ \ldots, \ v_m\}$ is the orthonormal basis of the Shift-invert Krylov subspace which satisfies:

$$\begin{aligned} \mathrm{Span}\left\{v_1, \ \ldots, \ v_m\right\} &= \mathrm{Span}\left\{v, \ (\gamma I - A)^{-1} v, \ \ldots, \ (\gamma I - A)^{-m+1} v\right\} \\ &= \left\{ r(A) v \mid \ r \in \mathcal{P}_{m-1} / (\gamma - z)^{m-1} \right\}, \end{aligned}$$

where $\mathcal{P}_m$ is the set of polynomials of a degree less than or equal to m. $\phi_k(A)v$ can be regarded as $\tilde{f}_\gamma\left((\gamma I - A)^{-1}\right) v$, the function of $(\gamma I - A)^{-1}$, where $\tilde{f}_\gamma(z) := \phi_k\left(\gamma - z^{-1}\right)$. Therefore, if H_m is invertible, then the matrix function is:

$$\phi_k(A) v \approx V_m \tilde{f}_\gamma \left(V_m^* (\gamma I - A)^{-1} V_m \right) V_m^* v = V_m \tilde{f}_\gamma(H_m) V_m^* v = r(A) v. \tag{7}$$

for some $r \in \mathcal{P}_{m-1}/(\gamma - z)^{m-1}$.

The different approximation method is also discussed [9–11]. This method uses the matrix $V_{m+1}^* A V_{m+1}$ insted of $V_m^*((\gamma I - A))^{-1} V_m = H_m$, and $\phi_k(A)$ is approximated as:

$$\phi_k(A) v \approx V_{m+1} \phi_k(V_{m+1}^* A V_{m+1}) V_{m+1}^* v. \tag{8}$$

Novati [21, Proposition 12] shows that the error bound of approximation eq. (7) does not depend on the width of $W(A)$ when A is the sectorial operator. Grimm [10, Corollary 5.1] and Göckler [9, Theorem 5.9] show the same fact for eq. (8) in the case of $W(A) \subseteq \mathbb{C}^-$.

2.2 Inexact Shift-invert Arnoldi method (ISIA)

SIA requires solving the equation to compute $(\gamma I - A)^{-1} v_j$ in every step of the Krylov process. Hashimoto and Nodera [13] proposed a method for solving this linear equation efficiently while guaranteeing that the generalized residual [16] would become smaller than the arbitrary tolerance. This method is called the Inexact Shift-invert Arnoldi method (ISIA), and the exactness needed for solving the linear equation decreases with each iteration.

2.3 Rational Krylov method (RK)

Let β and v_1 be the same vectors as section 2.1. The m-step Rational Krylov process is:

$$\begin{aligned} h_{j+1,j}v_{j+1} &= (\gamma_j I - A)^{-1}V_j t_j - \sum_{k=1}^{j} h_{k,j}v_k, \\ h_{k,j} &= v_k^* \left[(\gamma_j I - A)^{-1}V_j t_j - \sum_{l=1}^{k-1} h_{l,j}v_l \right] \qquad (k = 1, \ldots, j), \\ h_{j+1,j} &= \left\| (\gamma_j I - A)^{-1}V_j t_j - \sum_{l=1}^{j} h_{l,j}v_l \right\| \qquad (j = 1, \ldots, m), \end{aligned}$$

where $\gamma_j \in \mathbb{C}$ $(1 \le j \le m)$ is a different shift in every step, and $t_j \in \mathbb{C}^j$ is an arbitrary vector. This results in the orthonormal basis $\{v_1,\ \ldots,\ v_{m+1}\}$ of the Rational Krylov subspace which satisfies:

$$\begin{aligned} \mathrm{Span}\,\{v_1,\ \ldots,\ v_{m+1}\} &= \mathrm{Span}\left\{v,\ (\gamma_1 I - A)^{-1}v,\ \ldots,\ (\gamma_m I - A)^{-1}v\right\} \\ &= \left\{r(A)v \mid\ r \in \mathcal{P}_m/q_m,\ q_m(z) = (\gamma_1 - z)\ldots(\gamma_m - z)\right\}. \end{aligned}$$

Let $V_m = [v_1\ \cdots\ v_m]$. $\phi_k(A)v$ is approximated as

$$\phi_k(A)v \approx V_{m+1}\phi_k(V_{m+1}^* A V_{m+1})V_{m+1}^* v = r(A)v, \tag{9}$$

for some $r \in \mathcal{P}_m/q_m, \quad q_m(z) = (\gamma_1 - z)\cdots(\gamma_m - z)$.

Göckler [9, Theorem 7.8] shows that under the appropriate choice of shifts γ_j, the error bound of approximation (9) does not depend on $W(A)$.

The advantage of RK over SIA is the possibility of parallelization. The linear equation $(\gamma_j I - A)x = V_j t_j$ can be solved in parallel [4, 11, 25].

3 Shift-invert Rational Krylov method (SIRK)

We consider extending ISIA to the rational approximation with more than one pole. However, before the extention, the shifts for the approximation, is considered. The new method, SIRK, addresses the issue of the shifts.

The m-step Rational Krylov process in the same manner as illustrated in section 2.3 derives the following relations:

$$\begin{aligned} V_m T_m &= V_m H_m D_m - A V_m H_m + (\gamma_m I - A)h_{m+1,m}v_{m+1}e_m^*, \\ V_m^*(\gamma_m I - A)^{-1}V_m &= H_m(T_m - H_m D_m + \gamma_m H_m)^{-1} =: K_m, \end{aligned} \tag{10}$$

where $D_m := \mathrm{diag}\{\gamma_1, \cdots, \gamma_m\}$, T_m is the upper triangular matrix whose jth column is $[t_j^*\ \mathbf{0}]^* \in \mathbb{C}^m$. The methods of setting t_j for eigenvalue problem are discussed in, for example, [4, 22]. The methods for computing matrix function are discussed by Güttel [11]. In this study, we set

$$t_j = e_{P\lfloor j/P \rfloor} \in \mathbb{R}^j, \tag{11}$$

where P is the number of linear equations solved in parallel. In the SIRK, the shifts $\gamma_j = N - hj \in \mathbb{R}$, where $h > 0$, and $N > 0$ satisfies $\gamma_j > 0$ $(1 \leq \forall j \leq m)$ are used. If H_m is invertible, the matrix function $\phi_k(A)v$ is approximated as:

$$\begin{aligned}\phi_k(A)v &= f_m((\gamma_m I - A)^{-1})v \\ &\approx V_m f_m(V_m^*(\gamma_m I - A)^{-1}V_m)V_m^* v \\ &= V_m f_m(K_m)V_m^* v \\ &= V_m \phi_k\left(\gamma_m I - (T_m - H_m D_m + \gamma_m H_m)H_m^{-1}\right)V_m^* v \\ &= V_m \phi_k\left((H_m D_m - T_m)H_m^{-1}\right)V_m^* v, \end{aligned} \tag{12}$$

where $f_m(z) := \phi_k\left(\gamma_m - z^{-1}\right)$. Approximation (12) is for the function depending on m with the matrix depending on m.

The next consideration is the Rational Krylov subspace constructed by the SIRK. Let $X_j := (\gamma_j I - A)^{-1}$ $(1 \leq j \leq m)$. γ_j is defined as $\gamma_j = N - hj$, thus X_j is denoted as:

$$X_j = (\gamma_j I - A)^{-1} = (I - (\gamma_m - \gamma_j)X_m)^{-1}X_m = (I + h(m-j)X_m)^{-1}X_m. \tag{13}$$

From relation (13), the Rational Krylov subspace generated by the m-step SIRK is represented as:

$$\begin{aligned}&\mathrm{Span}\{v,\ X_1 v,\ \ldots,\ X_{m-1}v\} \\ &\quad = \mathrm{Span}\{v,\ (I + h(m-1)X_m)^{-1}X_m v,\ \ldots,\ (I + hX_m)^{-1}X_m v\} \\ &\quad = \{r(X_m)v \mid\ r \in \mathcal{P}_{m-1}/q_{m-1},\ q_m(z) = (1+hmz)\ldots(1+hz)\}. \end{aligned} \tag{14}$$

The following proposition shown by Beckermann and Reichel [1] is valid from relation (14), and the following theorem regarding the convergence of SIRK is deduced:

Proposition 3.1 *Let $q_m(z) := (1+hmz)\cdots(1+hz)$ and $\mathcal{P}_m$ be the set of polynomials with a degree less than or equal to m. Furthermore, let $\mathcal{P}_{m-1}/q_{m-1} := \{p/q_{m-1} \mid\ p \in \mathcal{P}_{m-1}\}$. Then, for $\forall r \in \mathcal{P}_{m-1}/q_{m-1}$,*

$$r(X_m)v = V_m r(K_m)V_m^* v. \tag{15}$$

Theorem 3.1 *Let $\mathcal{H}(\Pi)$ be the set of $\mathbb{C}$-valued holomorphic functions on a closed and bounded set $\Pi \subseteq \mathbb{C}$. Let $\hat{f}_N(z) = \phi_0(N - z^{-1})$ for $k = 0$, and $\hat{f}_N(z) := \int_0^1 e^{N - sz^{-1}}(1-s)^{k-1}/(k-1)!ds$ for $k \geq 1$. It is possible to choose the closed and bounded set Σ satisfying $\bigcup_{j=1}^{N-1} W(X_j) \subseteq \Sigma \subseteq \mathbb{C}^+$. With this Σ, there exists $1 \leq C \leq 11.08$ such that the error bound of SIRK is*

$$\|\phi_k(A)v - V_m f_m(K_m)V_m^* v\| \leq 2C\|v\|e^{-hm} \min_{r \in \mathcal{P}_{m-1}/q_{m-1}} \|\hat{f}_N - r\|_\Sigma, \tag{16}$$

for $1 \leq m \leq N-1$, where $\|\cdot\|_\Sigma$ is the norm of $\mathcal{H}(\Sigma)$, which is defined as $\|g\|_\Sigma = \sup_{z \in \Sigma}|g(z)|$.

Proof : Since $W(X_j) \subseteq \mathbb{C}^+$ is satisfied for all j in $1 \leq j \leq N-1$, and $W(X_j)$ are bounded, it is possible to choose a closed and bounded set $\Sigma \subseteq \mathbb{C}^+$ which contains $\bigcup_{j=1}^{N-1} W(X_j)$. From the fact $\phi_k(A) = f_m(X_m)$ and Proposition 3.1,

$$\|\phi_k(A)v - V_m f_m(K_m)V_m^* v\| \tag{17}$$

$$= \|f_m(X_m)v - r(X_m)v - V_m f_m(K_m)V_m^* v + V_m r(K_m)V_m^* v\|,$$

is derived for $\forall r \in \mathcal{P}_{m-1}/q_{m-1}$. Since all the poles of functions in $\mathcal{P}_{m-1}/q_{m-1}$ are real and negative, $\mathcal{P}_{m-1}/q_{m-1} \subseteq \mathcal{H}(\Sigma)$. In addition, $f_m,\ \hat{f}_N \in \mathcal{H}(\Sigma)$. From equation (10), $W(K_m) \subseteq W(X_m)$, and from Crouzeix [5], there is $1 \le C \le 11.08$ such that:

$$\begin{aligned} \|f_m(X_m) - r(X_m)\| &\le C\|f_m - r\|_\Sigma, \\ \|f_m(K_m) - r(K_m)\| &\le C\|f_m - r\|_\Sigma. \end{aligned} \tag{18}$$

Since $r \in \mathcal{P}_{m-1}/q_{m-1}$ is arbitrary, it is deduced that:

$$\begin{aligned} &\|\phi_0(A)v - V_m f_m(K_m)V_m^* v\| \\ &\le \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} [\|f_m(X_m) - r(X_m)\|\,\|v\| + \|f_m(K_m) - r(K_m)\|\,\|v\|] \qquad (\because (17)) \\ &\le 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|f_m - r\|_\Sigma \qquad (\because (18)) \\ &= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} |e^{N-hm-z^{-1}} - r(z)| \\ &= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} e^{-hm} \sup_{z\in\Sigma} \left|e^{N-z^{-1}} - e^{hm} r(z)\right| \\ &= 2C\|v\| e^{-hm} \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left|e^{N-z^{-1}} - r(z)\right|. \end{aligned}$$

If $k \ge 1$, ϕ_k is represented as $\phi_k(z) = \int_0^1 e^{sz}(1-s)^{k-1}/(k-1)!ds$. Therefore,

$$\begin{aligned} &\|\phi_k(A)v - V_m f_m(K_m)V_m^* v\| \\ &\le 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|f_m - r\|_\Sigma \qquad (\because (18)) \\ &= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} |\phi_k(N - hm - z^{-1}) - r(z)| \\ &= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left| \int_0^1 e^{s(N-hm-z^{-1})} \frac{(1-s)^{k-1}}{(k-1)!} - e^{s(N-hm)} \left(e^{-s(N-hm)} r(z)\right) ds \right| \\ &\le 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left| e^{N-hm} \left\{ \int_0^1 e^{-sz^{-1}} \frac{(1-s)^{k-1}}{(k-1)!} ds - \int_0^1 e^{-s(N-hm)} r(z) ds \right\} \right| \\ &= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left| e^{-hm} \left\{ \int_0^1 e^{N-sz^{-1}} \frac{(1-s)^{k-1}}{(k-1)!} ds - \int_0^1 e^{N-s(N-hm)} ds\, r(z) \right\} \right| \\ &= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} e^{-hm} \sup_{z\in\Sigma} \left| \int_0^1 e^{N-sz^{-1}} \frac{(1-s)^{k-1}}{(k-1)!} ds - r(z) \right|. \end{aligned}$$

□

Choosing γ_j as $N - hj$ results in the space $\mathcal{P}_{m-1}/q_{m-1}$ expanding with each iteration, because q_m has the form $q_m(z) = (1+hmz)\cdots(1+hz)$. Therefore, $\min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|f-r\|_\Sigma$ in error bound (16) becomes smaller as m becomes larger.

Remark 3.1 *The value N in* SIRK *plays the similar role with the shift γ in* SIA. *In fact, we can deduce the following error bound for* SIA*:*

$$\|\phi_k(A)v - V_m \tilde{f}_\gamma(H_m)V_m^* v\| \le 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/(\gamma-z)^{m-1}} \|\tilde{f}_\gamma - r\|_{W((\gamma I - A)^{-1})}, \tag{19}$$

where $\tilde{f}_\gamma(z) := \phi_k(\gamma - z^{-1})$. *Comparing the bound (16) with (19), we expect that the iteration number for convergence of* SIRK *is the same level or smaller than that of* SIA *with* $\gamma = N$ *due to the factor* e^{-hm}. *In fact, in the case of* $k = 0$, *the functions* $\hat{f}_N$ *and* $\tilde{f}_\gamma$ *are exactly the same, and the iteration number of* SIRK *is smaller. We confirm this fact of reducing the iteration number in the section 5 of numerical experiment.*

4 Inexact Shift-invert Rational Krylov method (ISIRK)

At this point, it is possible to extend the ISIA to the rational approximation with SIRK. It will be shown that a similar discussion for ISIA is also valid for SIRK, and an Inexact Shift-invert Rational Krylov method (ISIRK) will be proposed.

For $j = 1 \dots m$, let $\tilde{x}_j$ be the inexact solution of the linear equation $(\gamma_j I - A)x_j = v_j$, and $f_j^{\text{sys}} := x_j - \tilde{x}_j$ be the error vector for solving the linear equation, and let $R_m^{\text{sys}} := [r_1^{\text{sys}} \ \cdots \ r_m^{\text{sys}}]$, where $r_j^{\text{sys}} := v_j - (\gamma_j I - A)\tilde{x}_j$ is the residual vector for solving the linear equation. The following relation is derived by computing the m-step SIRK process in the same way as section 3. However, in this case, the linear equations must be solved inexactly at every step.

$$
\begin{aligned}
&(\gamma_j I - A)^{-1} V_j t_j - f_j^{\text{sys}} = \sum_{k=1}^{j+1} h_{k,j} v_k, \\
&V_j t_j = \sum_{k=1}^{j+1} h_{k,j}(\gamma_j I - A)v_k + r_j^{\text{sys}}, \\
&V_m T_m = V_m H_m D_m - A V_m H_m + h_{m+1,m}(\gamma_m I - A)v_{m+1} e_m^* + R_m^{\text{sys}}, \\
&(\gamma_m I - A)V_m = V_m K_m^{-1} - h_{m+1,m}(\gamma_m I - A)v_{m+1} e_m^* H_m^{-1} - R_m^{\text{sys}} H_m^{-1},
\end{aligned} \tag{20}
$$

where V_m is the $n \times m$ matrix with orthonormal columns, H_m is an $m \times m$ upper Hessenberg matrix, and $K_m = H_m(T_m - H_m D_m + \gamma_m H_m)^{-1}$. For the approximation, the same fomula used by the SIRK is employed:

$$
\phi_k(A)v \approx V_m f_m(K_m) V_m^* v. \tag{21}
$$

Let $\tilde{f}_m(z) = f_m(z^{-1})$. The error of this approximation, using Cauchy's integral formula, is

$$
\begin{aligned}
E_m &= \tilde{f}_m(\gamma_m I - A)v - V_m \tilde{f}_m(K_m^{-1}) V_m^* v \\
&= \frac{1}{2\pi i} \int_\Gamma \tilde{f}_m(\lambda) \left[(\lambda I - \gamma_m I + A)^{-1} v - V_m (\lambda I - K_m^{-1})^{-1} V_m^* v \right] d\lambda \\
&= \frac{1}{2\pi i} \int_\Gamma \tilde{f}_m(\lambda) e_m^{\text{lin}} \, d\lambda,
\end{aligned} \tag{22}
$$

where Γ is a contour enclosing the eigenvalues of $\gamma_m I - A$ and K_m^{-1}, $e_m^{\text{lin}} = [(\lambda I - \gamma_m I + A)^{-1} v - V_m(\lambda I - K_m^{-1})^{-1} V_m^* v]$. Let $\hat{f}(z) = 1/(\lambda - z^{-1})$. Then, $(\lambda I - \gamma_m I + A)^{-1} = \hat{f}\left((\gamma_m I - A)^{-1}\right)$. If SIRK is applied to function $\hat{f}$, $V_m(\lambda I - K_m^{-1})^{-1} V_m^* v = V_m \hat{f}(K_m) V_m^* v$ is the approximation of $\hat{f}\left((\gamma_m I - A)^{-1}\right) v$. The error bound of this approximation is represented in the same manner as Theorem 3.1:

$$
\| \hat{f}((\gamma_m I - A)^{-1}) v - V_m \hat{f}(K_m) V_m^* v \|
$$

$$\leq \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \left[\|\hat{f}(X_m) - r(X_m)\| \, \|v\| + \|\hat{f}(K_m) - r(K_m)\| \, \|v\| \right]$$
$$\leq 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|\hat{f} - r\|_\Sigma, \tag{23}$$

where Σ is the same set as Theorem 3.1, and $1 \leq C \leq 11.08$. In this case, $\hat{f}$ does not depend on m, so the upper bound (23) decreases as m becomes large. Therefore, this approximation converges. Using equation (20), the residual $r_m^{\rm lin}$ of this approximation for the linear equation is represented as

$$\begin{aligned}
r_m^{\rm lin} &= v - (\lambda I - \gamma_m I + A)V_m \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v \\
&= v - \lambda V_m \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v + (\gamma_m I - A)V_m(\lambda I - K_m^{-1})^{-1} V_m^* v \\
&= v - \lambda V_m \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v \\
&\quad + \left[V_m K_m^{-1} - h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1} - R_m^{\rm sys} H_m^{-1}\right] \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v \\
&= v - V_m(\lambda I - K_m^{-1}) \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v \\
&\quad - h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1} \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v \\
&\quad - R_m^{\rm sys} H_m^{-1} \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v \\
&= \left[-h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1} - R_m^{\rm sys} H_m^{-1}\right] \left(\lambda I - K_m^{-1}\right)^{-1} V_m^* v.
\end{aligned}$$

Replacing $e_m^{\rm lin}$ with $r_m^{\rm lin}$ in equation (22), the generalized residual $r_{\phi,m}^{\rm real}$ [16] of the approximated $\phi_k(A)v$ is

$$\begin{aligned}
r_{\phi,m}^{\rm real} &= -h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1} \tilde{f}_m(K_m^{-1}) V_m^* v - R_m^{\rm sys} H_m^{-1} \tilde{f}_m(K_m^{-1}) V_m^* v \\
&= -h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1} \phi_k\left((H_m D_m - T_m)H_m^{-1}\right) V_m^* v \\
&\qquad\qquad - R_m^{\rm sys} H_m^{-1} \phi_k\left((H_m D_m - T_m)H_m^{-1}\right) V_m^* v \\
&= -\beta h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* \phi_k\left(D_m - H_m^{-1}T_m\right) H_m^{-1} e_1 \\
&\qquad\qquad - \beta R_m^{\rm sys} \phi_k\left(D_m - H_m^{-1}T_m\right) H_m^{-1} e_1
\end{aligned} \tag{24}$$

In order to evaluate equation (24), the following lemma and propositions are used.

Lemma 4.1 (see [13, Proposition 2]) *Let f be the holomorphic function in $\mathbb{C}^+$. If the sequence of the upper Hessenberg matrices $\{H_m \in \mathbb{R}^{m\times m}\}_{m=1}^n$ satisfies*

$$W(H_m) \subseteq \mathbb{C}^+ \quad (1 \leq \forall m \leq n), \tag{25}$$

then there exists $\alpha > 0$ and $0 < \lambda < 1$ which do not depend on m and satisfy

$$\left|[f(H_m)]_{i,j}\right| \leq \alpha\lambda^{i-j} \quad (i \geq j). \tag{26}$$

If f is an entire function, then inequality (26) is satisfied for all the sequence of the upper Hessenberg matrices $\{H_m \in \mathbb{R}^{m\times m}\}_{m=1}^n$.

The proof of Lemma 4.1 is based on Benzi and Boito [2].

Proposition 4.1 *Let* $\{K_m \in \mathbb{R}^{m\times m}\}_{m=1}^n$ *be the sequence of matrices satisfying*

$$|(K_m)_{i,j}| \leq \alpha\lambda^{i-j} \ (i \geq j), \tag{27}$$

where $\alpha > 0$ *and* $0 < \lambda < 1$ *which do not depend on* m*. If*

$$\lambda \leq \frac{1}{\sqrt{2}}, \tag{28}$$

then, there exist a sequence of unitary matrices $\{Q_m \in \mathbb{R}^{m\times m}\}_{m=1}^n$ *and a sequence of upper Hessenberg matrices* $\{H_m \in \mathbb{R}^{m\times m}\}_{m=1}^n$ *such that* $K_m = Q_m^* H_m Q_m$ *and*

$$|(Q_m)_{i,j}| \leq \alpha'\lambda^{|i-j|} \ (i,j \leq m),$$

with $\alpha' > 0$ *which does not depend on* m*.*

Proof : The Householder reflectors for transforming K_m into the upper Hessenberg matrix are applied. Let $k_{i_1:i_2,j}$ be the vector consisting of elements from (i_1, j) to (i_2, j) of K_m, $\eta_j = \|k_{j+1:m,j}\|$, $u_j = (k_{j+1:m,j} - \eta_j e_1)/\|k_{j+1:m,j} - \eta_j e_1\|$. Then, $\tilde{Q}_{j+1} = -2u_j u_j^*$ is defined. $I_{m-j} + \tilde{Q}_{j+1}$ is a unitary matrix and satisfies $(I_{m-j} + \tilde{Q}_{j+1})k_{j+1:m,j} = \eta_j e_1$. Thus, the matrix Q_m defined as $Q_m = (I_m + \hat{Q}_{m-1}) \ldots (I_m + \hat{Q}_2)$, where $\hat{Q}_{j+1} = \mathrm{diag}\{O_j, \ \tilde{Q}_{j+1}\}$, is a unitary matrix, and there exists an upper Hessenberg matrix H_m such that $Q_m K_m Q_m^* = H_m$. The vectors u_j and $k_{j+1:m,j} - \eta_j e_1$ remain the same up to the constant. Vector $k_{j+1:m,j}$ satisfies condition (27), and all the elements except for the first element of $\eta_j e_1$ are 0. In addition, η_j satisfies:

$$\eta_j \leq \sqrt{\sum_{k=1}^{\infty}(\alpha\lambda^k)^2} = \alpha\sqrt{\frac{1}{1-\lambda^2}}\lambda.$$

For these reasons, $|u_{i,j}| < \hat{\alpha}\lambda^i$ is satisfied for some $\hat{\alpha} > 0$, where $u_{i,j}$ is the ith element of $u_j \in \mathbb{C}^{m-j}$. It is deduced that:

$$|[u_j u_j^*]_{k,l}| = |u_{k,j}u_{l,j}| \leq \hat{\alpha}^2\lambda^{k+l}.$$

Let $\check{\alpha} = 2\hat{\alpha}^2$. For $l \geq 2$ and $l < k_1 < k_2 < \ldots < k_r$, it is deduced that

$$\begin{aligned} &|(\hat{Q}_{k_r} \ldots \hat{Q}_{k_1}\hat{Q}_l)_{i,j}| \leq \frac{\check{\alpha}^{r+1}\lambda^{-2(l-r-1)}}{(1-\lambda^2)^r}\lambda^{i+j} = \check{\alpha}^{r+1}\alpha''(l,r)\lambda^{i+j} \quad (i,j \leq k_r), \\ &|(\hat{Q}_{k_r} \ldots \hat{Q}_{k_1}\hat{Q}_l)_{i,j}| = 0 \quad (i > k_r \text{ or } j > k_r), \end{aligned} \tag{29}$$

where $\alpha''(l,r) = \lambda^{-2(l-r-1)}/(1-\lambda^2)^r$. Inequality (29) is proved by the induction of r. For $r = 1$, we have:

$$\begin{aligned} |(\hat{Q}_{k_1}\hat{Q}_l)_{i,j}| &\leq \sum_{a=k_1}^{m} \check{\alpha}\lambda^{i-k_1+1+a-k_1+1}\check{\alpha}\lambda^{a-l+1+j-l+1} \\ &= \check{\alpha}^2\lambda^{i+j}\lambda^{-2(k_1+l-2)}\sum_{a=k_1}^{m}\lambda^{2a} \\ &\leq \frac{\check{\alpha}^2\lambda^{-2(l-2)}}{1-\lambda^2}\lambda^{i+j} \quad (i,j \leq k_1). \end{aligned}$$

For $r \geq 2$, if inequality (29) is satisfied with $r-1$, then we have:

$$\begin{aligned}|(\hat{Q}_{k_r} \dots \hat{Q}_{k_1}\hat{Q}_l)_{i,j}| &\leq \sum_{a=k_r}^{m} \check{\alpha}\lambda^{i-k_r+1+a-k_r+1}\frac{\check{\alpha}^r\lambda^{-2(l-r)}}{(1-\lambda^2)^{r-1}}\lambda^{a+j} \\ &\leq \lambda^{i+j}\frac{\check{\alpha}^{r+1}\lambda^{-2(k_r+l-r-1)}}{(1-\lambda^2)^{r-1}}\sum_{a=k_r}^{m}\lambda^{2a} \\ &= \frac{\check{\alpha}^{r+1}\lambda^{-2(l-r-1)}}{(1-\lambda^2)^r}\lambda^{i+j} \quad (i,j \leq k_r).\end{aligned}$$

This is the proof of inequality (29). In inequality (29), if $\lambda \leq 1/\sqrt{2}$, then $\alpha''(l,r+1) \leq \alpha''(l,r)$ for all l. This results in $\alpha''(l,r) \leq \alpha''(l,1)$ for all r and l.

Q_m is represented as:

$$\begin{aligned}Q_m &= (I_m + \hat{Q}_{m-1})\dots(I_m+\hat{Q}_2) \\ &= I_m + \sum_{k=3}^{m-1}\sum_{l=2}^{k-1}\sum_{(a_1,a_2,\dots,a_{k-l-1})\in\{0,1\}^{k-l-1}} \hat{Q}_k\hat{Q}_{k-1}^{a_1}\hat{Q}_{k-2}^{a_2}\dots\hat{Q}_{l+1}^{a_{k-l-1}}\hat{Q}_l + \sum_{k=2}^{m-1}\hat{Q}_k. \qquad (30)\end{aligned}$$

As a result, for $2 \leq \min\{i,j\} \leq m-1$, equation (30) and inequality (29) shows that:

$$\begin{aligned}|(Q_m - I_m)_{i,j}| &\leq \check{\alpha}^2 \sum_{k=3}^{\min\{i,j\}}\sum_{l=2}^{k-1}\left[1+(k-l-1)\check{\alpha}+\begin{pmatrix}k-l-1\\2\end{pmatrix}\check{\alpha}^2+\dots+\check{\alpha}^{k-l-1}\right] \\ &\quad\times\alpha''(l,1)\lambda^{i+j} + \check{\alpha}^2\sum_{k=2}^{\min\{i,j\}}\alpha''(k,1)\lambda^{i+j} \\ &\leq \check{\alpha}^2\sum_{k=3}^{\min\{i,j\}}\sum_{l=2}^{k-1}(1+\check{\alpha})^{k-l-1}\alpha''(l,1)\lambda^{i+j} + \check{\alpha}^2\sum_{k=2}^{\min\{i,j\}}\alpha''(k,1)\lambda^{i+j}\end{aligned}$$

Therefore, for $2 \leq \min\{i,j\} \leq m-1$ and $i \leq j$, under the assumption of (28), it is deduced that:

$$\begin{aligned}&|(Q_m - I_m)_{i,j}| \\ &\leq \sum_{k=3}^{i}\frac{(1+\check{\alpha})^{k-1}\check{\alpha}^2\lambda^4}{1-\lambda^2}\lambda^{i+j}\frac{((1+\check{\alpha})\lambda^2)^{-2}}{((1+\check{\alpha})\lambda^2)^{-1}-1}[((1+\check{\alpha})\lambda^2)^{-k+2}-1] \\ &\qquad + \frac{\check{\alpha}^2\lambda^4}{1-\lambda^2}\lambda^{i+j}\frac{\lambda^{-4}}{\lambda^{-2}-1}[(\lambda^{-2})^{i-1}-1] \\ &\leq \frac{(1+\check{\alpha})\check{\alpha}^2\lambda^4((1+\check{\alpha})\lambda^2)^{-1}}{(1-\lambda^2)|1-(1+\check{\alpha})\lambda^2|}\lambda^{i+j}\left[\sum_{k=3}^{i}(\lambda^{-2})^{k-2}+i\right] + \frac{\check{\alpha}^2\lambda^4}{(1-\lambda^2)^2}\lambda^{i+j}\lambda^{-2i} \\ &\leq \frac{\check{\alpha}^2\lambda^2}{(1-\lambda^2)|1-(1+\check{\alpha})\lambda^2|}\left[\lambda^{i+j}\frac{\lambda^{-2}}{\lambda^{-2}-1}(\lambda^{-2})^{i-2}+i\lambda^{2i}\lambda^{j-i}\right] + \frac{\check{\alpha}^2\lambda^4}{(1-\lambda^2)^2}\lambda^{i+j}\lambda^{-2i} \\ &= \left[\frac{\check{\alpha}^2\lambda^2}{(1-\lambda^2)|1-(1+\check{\alpha})\lambda^2|}\left(\frac{\lambda^4}{1-\lambda^2}+\frac{1}{2}\right)+\frac{\check{\alpha}^2\lambda^4}{(1-\lambda^2)^2}\right]\lambda^{j-i} \quad (\because eq.\ (28))\end{aligned}$$

$$=: \alpha' \lambda^{j-i},$$

where the sum $\sum_{k=3}^{i}$ becomes 0 for $k = 2$. In a similar manner, it is deduced that $|(Q_m - I_m)_{i,j}| \leq \alpha' \lambda^{i-j}$ for $i > j$. If $\min\{i, j\} = m$, then we have $i = j = m$ and

$$|(Q_m - I_m)_{m,m}| \leq \check{\alpha}^2 \sum_{k=3}^{m-1} \sum_{l=2}^{k-1} (1+\check{\alpha})^{k-l+1} \alpha''(l,1) \lambda^{i+j} + \check{\alpha}^2 \sum_{k=2}^{m-1} \alpha''(k,1) \lambda^{i+j} \leq \alpha'.$$

For $\min\{i, j\} = 1$,

$$\begin{cases} |(Q_m)_{1,1}| = 1 \\ |(Q_m)_{i,j}| = 0 \quad (i \neq 1 \text{ or } j \neq 1) \end{cases}$$

is followed by the definition of Q_m. Since I_m is a diagonal matrix, redefining α' as the sum of 1 and the previous α' completes the proof.

□

Proposition 4.2 *Let $\{H_m \in \mathbb{R}^{m \times m}\}_{m=1}^{n}$ be the sequence of the upper Hessenberg matrices and $\{D_m \in \mathbb{R}^{m \times m}\}_{m=1}^{n}$ be the sequence of diagonal matrices which consist of shifts of the* ISIRK*, $D_m = \mathrm{diag}\{N-1, \ldots, N-m\}$. If the matrix H_m satisfies*

$$W(H_m) \subset \mathbb{C}^+ \quad (1 \leq \forall m \leq n), \tag{31}$$

then there exist $\alpha \hat{>} 0$ and $0 < \hat{\lambda} < 1$ such that $|[D_m - H_m^{-1} T_m]_{i,j}| \leq \hat{\alpha} \hat{\lambda}^{i-j}$ for $i \geq j$. If such $\hat{\lambda}$ satisfies $\hat{\lambda} \leq 1/\sqrt{2}$, then there exist $\alpha > 0$ and $0 < \lambda < 1$ which do not depend on m such that:

$$\left| \left[\phi_k \left(D_m - H_m^{-1} T_m \right) H_m^{-1} \right]_{i,1} \right| \leq \frac{1}{2} \alpha (i+1) i \lambda^{i-1}. \tag{32}$$

Proof : Since H_m is the upper Hessenberg matrix and satisfies condition (31), setting $f(z) = z^{-1}$ and using Lemma 4.1 derives that there exist $\hat{\alpha} > 0$ and $0 < \hat{\lambda} < 1$ such that:

$$|[H_m^{-1}]_{i,j}| \leq \hat{\alpha} \hat{\lambda}^{i-j} \quad (i \geq j). \tag{33}$$

D_m is a diagonal matrix, and T_m is defined as equation (11). Therefore, redefining $\hat{\alpha}$ as the sum of $\|D_m\| = N - 1$ and the previous $\hat{\alpha}$ leads to:

$$|[D_m - H_m^{-1} T_m]_{i,j}| \leq \hat{\alpha} \hat{\lambda}^{i-j} \quad (i \geq j), \tag{34}$$

where $\hat{\alpha}$ and $\hat{\lambda}$ do not depend on m. Let $G^{\exp}(\alpha, \lambda) = \{A : \text{ square matrix } | \ |(A)_{i,j}| \leq \alpha \lambda^{|i-j|} \quad (\forall i, j)\}$. From Proposition 4.1, there exists a unitary matrix Q_m and an upper Hessenberg matrix $\tilde{H}_m$ such that $D_m - H_m^{-1} T_m = Q_m^* \tilde{H}_m Q_m$ and $Q_m \in G^{\exp}(\alpha', \hat{\lambda})$ with $\alpha' > 0$ which does not depend on m. Thus, it is deduced that:

$$\begin{aligned} \phi_k \left(D_m - H_m^{-1} T_m \right) H_m^{-1} e_1 &= \phi_k (Q_m^* \tilde{H}_m Q_m) H_m^{-1} e_1, \\ &= Q_m^* \phi_k (\tilde{H}_m) Q \hat{H}_m e_1 \qquad (\exists \hat{H}_m \in G^{\exp}(\hat{\alpha}, \hat{\lambda})). \end{aligned}$$

The second equality is held, because from inequality (33), there exists $\hat{H}_m \in G^{\exp}(\hat{\alpha}, \hat{\lambda})$ which satisfies $H_m^{-1} e_1 = \hat{H}_m e_1$. From Benzi and Boito [3, Theorem 9.2], there exist $\alpha'' > 0$

and λ'' which do not depend on m and satisfy $Q_m\hat{H}_m \in G^{\exp}(\alpha'', \lambda'')$. In addition, setting $f = \phi_k$, the entire function, and using Lemma 4.1 derive that there exist $\check{\alpha} > 0$ and $0 < \check{\lambda} < 1$ such that $|[\phi_k(\tilde{H}_m)]_{i,j}| \leq \check{\alpha}\check{\lambda}^{i-j}$ $(i \geq j)$. Let $\Sigma = \bigcup_{m=1}^{n}(D_m - H_m^{-1}T_m)$. Σ is closed and bounded, and $\tilde{H}_m$ satisfies

$$W(\tilde{H}_m) = W\left(Q_m(D_m - H_m^{-1}T_m)Q_m^*\right) = W(D_m - H_m^{-1}T_m) \subseteq \Sigma. \tag{35}$$

Therefore, $|e^z| \leq C'$ for some $C' > 0$ is satisfied when $z \in W(\tilde{H}_m)$, and $|\phi_k(z)|$ is bounded as

$$|\phi_k(z)| = \left|\int_0^1 e^{sz}\frac{(1-s)^{k-1}}{(k-1)!}ds\right| \leq |e^z|\left|\int_0^1 \frac{(1-s)^{k-1}}{(k-1)!}ds\right| \leq \frac{C'}{k!} \quad (z \in W(\tilde{H}_m)). \tag{36}$$

Using the theorem by Crouzeix [5, Theorem 2], condition (35) and inequality (36), there exists $1 \leq C \leq 11.08$ such that

$$\|\phi_k(\tilde{H}_m)\| \leq C \sup_{z \in W(\tilde{H}_m)} |\phi_k(z)| \leq \frac{CC'}{k!}.$$

Redefining $\check{\alpha}$ as the sum of $CC'/(k!)$ and the previous $\check{\alpha}$ leads to:

$$\left|\left[\phi_k(\tilde{H}_m)\right]_{i,j}\right| \leq \check{\alpha}\check{\lambda}^{i-j} \quad (i \geq j), \tag{37}$$

$$\left|\left[\phi_k(\tilde{H}_m)\right]_{i,j}\right| \leq \|\phi_k(\tilde{H}_m)\| \leq \check{\alpha} \quad (i < j). \tag{38}$$

From the upper bounds (37) and (38), it is deduced that:

$$\begin{aligned}
\left|\left[\phi_k(\tilde{H}_m)Q_m\hat{H}_m\right]_{i,1}\right| &\leq \sum_{k=1}^{i}\check{\alpha}\check{\lambda}^{i-k}\alpha''\lambda''^{k-1} + \sum_{k=i+1}^{m}\check{\alpha}\alpha''\lambda''^{k-1} \\
&\leq i\check{\alpha}\alpha''\bar{\lambda}^{i-1} + \check{\alpha}\alpha''\frac{\lambda''^{i}}{1-\lambda''} \\
&\leq i\check{\alpha}\alpha''\left(1 + \frac{\lambda''}{1-\lambda''}\right)\bar{\lambda}^{i-1} \\
&= i\bar{\alpha}\bar{\lambda}^{i-1}.
\end{aligned} \tag{39}$$

where $\bar{\alpha} := \check{\alpha}\alpha''/(1-\lambda'')$, $\bar{\lambda} := \max\{\check{\lambda}, \lambda''\} < 1$. As a result, using fact $Q_m \in G^{\exp}(\alpha', \hat{\lambda})$ and the upper bound (39), it is deduced that:

$$\begin{aligned}
\left|\left[\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}\right]_{i,1}\right| &= \left|\left[Q_m^*\phi_k(\tilde{H}_m)Q_m\hat{H}_m\right]_{i,1}\right| \\
&\leq \sum_{k=1}^{i}\alpha'\hat{\lambda}^{i-k}k\bar{\alpha}\bar{\lambda}^{k-1} + \sum_{k=i+1}^{m}\alpha'\hat{\lambda}^{k-i}k\bar{\alpha}\bar{\lambda}^{k-1} \\
&\leq \frac{1}{2}(i+1)i\alpha'\bar{\alpha}\lambda^{i-1} + \alpha'\bar{\alpha}\frac{i+1}{(1-\lambda^2)^2}\lambda^{i+1} \\
&\leq \frac{1}{2}(i+1)i\alpha'\bar{\alpha}\left(1 + \frac{2\lambda^2}{(1-\lambda^2)^2}\right)\lambda^{i-1}
\end{aligned}$$

$$= \frac{1}{2}(i+1)i\alpha\lambda^{i-1}.$$

where $\alpha := \alpha'\bar{\alpha}(1+2\lambda^2/(1-\lambda^2)^2)$ and $\lambda := \max\{\hat{\lambda}, \bar{\lambda}\} < 1$.

□

If the residual of solving the linear equation satisfies $\|r_m^{\rm sys}\| \le \delta$ for some $\delta > 0$, then there exist $\alpha > 0$ and $0 < \lambda < 1$ such that the first term of equation (24) becomes:

$$\begin{aligned}
&\beta \left| h_{m+1,m} e_m^* \phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1 \right| \, \|(\gamma_m I - A)v_{m+1}\| \\
&\quad \le \beta |h_{m+1,m}| \, \left| \left[\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}\right]_{m,1} \right| \, \|\gamma_m I - A\| \, \|v_{m+1}\| \\
&\quad \le \beta |h_{m+1,m}| \, \|\gamma_m I - A\| \frac{1}{2}\alpha m(m+1)\lambda^{m-1} \quad (\because (32)) \\
&\quad \le \frac{\beta}{2}\|(\gamma_m I - A)^{-1}v_m - f_m^{\rm sys} - h_{1,m}v_1 - \ldots - h_{m,m}v_m\| \\
&\qquad\qquad \times \|\gamma_m I - A\|\alpha m(m+1)\lambda^{m-1} \\
&\quad \le \frac{\beta}{2}(\|(\gamma_m I - A)^{-1}v_m\| + \|f_m^{\rm sys}\|)\|\gamma_m I - A\|\alpha m(m+1)\lambda^{m-1} \\
&\quad \le \frac{\beta}{2}(1 + \|r_m^{\rm sys}\|)\|(\gamma_m I - A)^{-1}\| \, \|\gamma_m I - A\|\alpha m(m+1)\lambda^{m-1} \\
&\quad \le \frac{\beta}{2}(1+\delta)\kappa(\gamma_m I - A)\alpha m(m+1)\lambda^{m-1}. \qquad (40)
\end{aligned}$$

Since $0 < \lambda < 1$, the upper bound (40) implies that under the assumption of (28), if $\kappa(\gamma_m I - A)$ does not increase as m becomes larger, the first term of equation (24) decreases as m becomes larger.

Concerning the second term of equation (24), the following theorem is deduced:

Theorem 4.1 *Let* $[\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}]_{i,j} =: g_{i,j}^m$. *Moreover, let* $\mathrm{tol}_\phi > 0$ *be the tolerance for computing the* ϕ*-function and* $m^{\max}$ *be the maximum number of iterations. Under the assumptions about* H_m *and* $\hat{\lambda}$ *in Proposition 4.2, If*

$$\|r_1^{\rm sys}\| \le \frac{\mathrm{tol}_\phi}{2m^{\max}\beta\|\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1\|}, \qquad (41)$$

$$\|r_j^{\rm sys}\| \le \frac{|g_{1,1}^m|}{|g_{j-1,1}^m|}\|r_1^{\rm sys}\| \quad (2 \le j \le m), \qquad (42)$$

then

$$\beta\|R_m^{\rm sys}\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1\| \lesssim \mathrm{tol}_\phi\,.$$

Proof : Based on the above assumptions (41), (42) and Proposition 4.2, the upper bound is derived:

$$\begin{aligned}
&\beta\|R_m^{\rm sys}\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1\| \\
&\quad \le \beta(|g_{1,1}^m| \, \|r_1^{\rm sys}\| + |g_{2,1}^m| \, \|r_2^{\rm sys}\| + \ldots + |g_{m,1}^m| \, \|r_m^{\rm sys}\|) \\
&\quad \le \beta \left(|g_{1,1}^m| \, \|r_1^{\rm sys}\| + |g_{2,1}^m|\frac{|g_{1,1}^m|}{|g_{1,1}^m|}\|r_1^{\rm sys}\| + |g_{3,1}^m|\frac{|g_{1,1}^m|}{|g_{2,1}^m|}\|r_1^{\rm sys}\| + \right.
\end{aligned}$$

$$\ldots + |g^m_{m,1}| \frac{|g^m_{1,1}|}{|g^m_{m-1,1}|} \|r^{\mathrm{sys}}_1\| \Bigg) \quad (\because (42))$$
$$= \beta |g^m_{1,1}| \, \|r^{\mathrm{sys}}_1\| \left(1 + \frac{|g^m_{2,1}|}{|g^m_{1,1}|} + \frac{|g^m_{3,1}|}{|g^m_{2,1}|} + \ldots + \frac{|g^m_{m,1}|}{|g^m_{m-1,1}|} \right)$$
$$\approx \beta \|\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1\| \, \|r^{\mathrm{sys}}_1\| \left(1 + 3\lambda + 2\lambda + \ldots + \frac{m(m+1)}{(m-1)m}\lambda \right) \quad (\because (32))$$
$$\leq \beta \|\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1\| \, \|r^{\mathrm{sys}}_1\| \cdot 2m^{\max}$$
$$\leq \mathrm{tol}_\phi \quad (\because (41)).$$

□

The right-hand side of inequality (42) becomes larger as m becomes larger because of Proposition 4.2. Thus, Theorem 4.1 implies that the larger m becomes, the solution of the linear equation $(\gamma_m I - A)x_m = V_m t_m$ becomes more inexact, and the computational cost decreases compared to the SIRK. However, if the linear equations are solved, satisfying inequalities (41) and (42), then the second term of equation (24) is no longer an issue. In this scenario, the first term of equation (24), $r^{\mathrm{comp}}_{\phi,m}$, is used as the stopping criterion for the convergence of ISIRK.

Remark 4.1 *In practical computation, the values depending on m in inequalities (41) and (42) are unavailable in advance. Thus, for the exponential integrator at the $(i+1)$th step, $\phi_k(D_m - H_m^{-1}T_m)H_m^{-1}e_1$ is replaced with the ones in the largest Krylov subspace at the ith step. For the computation of equation (3):*

$$V_m^*(\gamma_m I - A)V_m \approx (T_m - H_m D_m + \gamma_m H_m)H_m^{-1}$$
$$H_m^{-1}T_m \approx H_m D_m H_m^{-1} - V_m^* A V_m,$$

since $[(\gamma_m I - A)^{-1}]^{-1} V_m e_l \approx V_m K_m^{-1} e_l$ for all $1 \leq l \leq m$. From Lemma 4.1, we have $H_m^{-1}T_m e_1 = H_m^{-1}e_1 \approx (H_m^{-1})_{1,1}e_1$. Thus,

$$H_m^{-1}e_1 \approx H_m d_{1,1}(H_m^{-1})_{1,1}e_1 - V_m^* A V_m e_1$$
$$\approx H_m(\gamma_1 I)H_m^{-1}e_1 - V_m^* A V_m e_1$$
$$= V_m^*(\gamma_1 I - A)V_m e_1$$

Similarly, from Proposition 4.2, $\phi_k((H_m D_m - T_m)H_m^{-1})e_1 \approx [\phi_k((H_m D_m - T_m)H_m^{-1})]_{1,1}e_1$ is deduced. Therefore,

$$\|H_m^{-1}\phi_k((H_m D_m - T_m)H_m^{-1})e_1\| \approx \|H_m^{-1}[\phi_k((H_m D_m - T_m)H_m^{-1})]_{1,1}e_1\|$$
$$\approx \|V_m^*(\gamma_1 I - A)V_m \phi_k((H_m D_m - T_m)H_m^{-1})e_1\|$$
$$\approx \|(\gamma_1 I - A)y(t)\|$$
$$\approx \|(\gamma_1 I - A)y(0)\|. \tag{43}$$

Approximation (43) is employed for inequality (41) in the computation of equation (3). Moreover, since α and λ do not depend on m, the following approximation is used for $P = 1$:

$$|g^m_{1,1}| \approx |g^{j-1}_{1,1}|, \quad |g^m_{1,j-1}| \approx |g^{j-1}_{1,j-1}| \qquad (2 \leq j \leq m).$$

Algorithm 4.1 ISIRK method for exponential integrator

Require: $A \in \mathbb{C}^{n\times n}$, $v \in \mathbb{C}^n$, $\delta > 0$, $\mathrm{tol}_\phi > 0$, $m^{\max} \in \mathbb{N}$, $h > 0$, $N > hm^{\max}$
Ensure: $\beta V_m \phi_k((H_m D_m - T_m)H_m^{-1})e_1$ such that $\|r_m^{\mathrm{real}}\| \le \mathrm{tol}_\phi$
1: $\beta = \|v\|$, $v_1 = v/\beta$
2: $\mathrm{tol}_1^{\mathrm{sys}} = \mathrm{tol}_\phi / (m^{\max}\beta\|f_{m(i)}^i\|)$
3: **for** $m = 1, 2, \ldots$ **do**
4: $d_{m,m} = N - hm$
5: Compute $\tilde{x}$ such that $\|V_m t_m - (d_{m,m}I - A)\tilde{x}\| \le \mathrm{tol}_m^{\mathrm{sys}}$
6: **for** $l = 1, 2, \ldots, m$ **do**
7: $h_{l,m} = \tilde{x}^* v_l$, $\tilde{x} = \tilde{x} - h_{l,m} v_l$
8: **end for**
9: $h_{m+1,m} = \|\tilde{x}\|$, $v_{m+1} = \tilde{x}/h_{m+1,m}$
10: $f_m^{i+1} = H_m^{-1}\phi_k((H_m D_m - T_m)H_m^{-1})e_1$
11: $r = |h_{m+1,m}(f_m^{i+1})_m| \, \|(\gamma_m I - A)v_{m+1}\|$
12: $\mathrm{tol}_{m+1}^{\mathrm{sys}} = \min\{\mathrm{tol}_1^{\mathrm{sys}} \, |(f_m^{i+1})_1|/|(f_m^{i+1})_m|, \delta\}$
13: **if** $r \le \mathrm{tol}_\phi$ **then**
14: $m(i+1) = m$
15: $y_m(t) = \beta V_m \phi_k((H_m D_m - T_m)H_m^{-1})e_1$, break
16: **end if**
17: **end for**

In the case of $P > 1$, if

$$\|r_j^{\mathrm{sys}}\| \le \frac{|g_{1,1}^m|}{|g_{P\lfloor j/P\rfloor,1}^m|}\|r_1^{\mathrm{sys}}\|, \tag{44}$$

then the condition (41) is satisfied for j. Thus, we solve the linear equation $(\gamma_j I - A)x_j = V_j t_j$ with the same level of exactness with $(\gamma_{j_0} I - A)x_{j_0} = V_{j_0} t_{j_0}$, where $j_0 = P\lfloor j/P\rfloor$.

Remark 4.2 *It is easy to check the assumption of Theorem 4.1, since both of them are checked through H_m, and $m \ll n$. In fact, if $P = 1$ and $\gamma_j = \gamma$ for some $\gamma \in \mathbb{C}$ and $1 \le j \le m$, assumption (31) is always satisfied for* SIRK *[13].*

In summary, Algorithm 4.1 is proposed for ϕ-functions in the exponential integrator of the ith step, where $(f_m)_j$ is the jth element of f_m. For the computation of equation (3), the second line is replaced by $tol_1^{\mathrm{sys}} = \mathrm{tol}_\phi / [2m^{\max}\beta\|(\gamma_1 I - A)y(0)\|]$. The linear equation in the fifth line of the algorithm is solved by an iterative method, and the convergence of its solution is judged by its residual. This facilitates ensuring that the residual of the solution of the linear equation satisfies the required conditions. Any iterative methods, for example, the BiCGSTAB [28] or the GMRES [23], are viable options. $(H_m D_m - T_m)H_m^{-1}$ in the tenth line is a small matrix, and it can be computed via a direct method inexpensively. After computing $(H_m D_m - T_m)H_m^{-1}$, $\phi_k((H_m D_m - T_m)H_m^{-1})$ is also computed using a direct method, such as the scaling and squaring method [14].

5 Numerical experiments

A few typical numerical experiments have been implemented in this section. These experiments were in a collection of problems to illustrate the effectiveness of ISIRK. All

numerical computations of these tests were executed with C on an Intel(R) Xeon(R) X5690 3.47GHz processor with an Ubuntu14.04LTS operating system. LAPACK and BLAS were used with ATLAS for the computations. In addition, Open MPI was used for parallel computations.

The Galerkin method with unstructured first order triangle elements and linear weight functions, were used to discretize the problems. After the discretization, the GMRES algorithm [23] with an ILU(0) preconditioner were applied to solve the linear equation in the fifth line of Algorithm 4.1 and in other algorithms. For SIA, RK and SIRK, the linear equation was solved with a residual tolerance of 10^{-14}.

Example 1

In order to show the advantages of the SIRK and ISIRK, the convection diffusion equation in region $\Omega = ((-1.5, 1.5) \times (-1, 1)) \setminus ([-0.5, 0.5] \times [-0.25, 0.25]) \subseteq \mathbb{R}^2$ was implemented:

$$\begin{cases} \rho c_v \dfrac{\partial u}{\partial t} = \lambda \Delta u - 5 \dfrac{\partial u}{\partial x_1} & \text{in } (0, T] \times \Omega, \\ u = 0 & \text{on } \{0\} \times \Omega, \\ u = 10 & \text{on } (0, T] \times \partial\Omega_1, \\ -\lambda \dfrac{\partial u}{\partial n} = 0 & \text{on } (0, T] \times \partial\Omega_2, \end{cases} \tag{45}$$

where $\partial\Omega_1 = \{-1.5\} \times [-0.5, 1]$, $\partial\Omega_2 = \partial\Omega \setminus \partial\Omega_1$, $\rho = 1.3$, $c_v = 1000$ and $\lambda = 0.025$. After the discretization, equation (2) with $F(y) = Ly + c$ was obtained with $n = 390256$. In this example, the differential operator $\mathcal{D} = 1/(\rho C_v)(\lambda\Delta - 5\frac{\partial u}{\partial x_1})$ was linear and did not depend on t. Thus, the solution was obtained through computing equation (3). Equation (3) was computed with the SIA and SIRK. For SIRK, we tried $P = 1$ and $P = 2$. The CPU times and iteration numbers were compared. The results are shown in Table 1, Figure 1 and Figure 2. The relative error tolerance for computing $\phi_0(tM^{-1}L)(v + L^{-1}c)$ was 10^{-6}, and $t = 270$. Here, we used the error instead of the residual in order to consider about the bound (16). Concerning the shift in the SIA and SIRK, $N = \gamma = 100/t$ and $N = \gamma = 200/t$. In order to treat A instead of tA, γ_j/t was used instead of γ_j. For $N = 200/t$, we compared $h = 1/t$ and $h = 2/t$. The results show that SIRK reduces the iteration numbers compared with SIA with $\gamma = N$. On the other hand, the shifted matrices $(N - j)I - A$ in SIRK are more ill-conditioned than $\gamma I - A$ in SIA, since $\gamma > N - j$, and γ realizes the larger transformation of the singular values of A. Thus, SIRK takes more time for solving the linear equations than SIA. As a result, SIRK with $P = 1$ takes a bit more time than SIA. However, the linear equation in SIRK can be solved in parallel while those in SIA cannot. The speed of SIRK with $P = 2$ is about twice as fast as that of SIA. As for the choice of $N = \gamma$ and h, the result provides the following observation. A larger N results in larger factors $\min_{r \in \mathcal{P}_{m-1}/q_{m-1}} \|\hat{f}_N - r\|_\Sigma$ and $\min_{p \in \mathcal{P}_{m-1}} \|\tilde{f}_\gamma - p\|_{W((\gamma I - A)^{-1})}$ in bounds (16) and (19). Thus, the iteration number for convergence increases. In addition, a larger h results in a smaller acceleration factor e^{-hm} in SIRK. Thus, the iteration number for convergence decreases.

Next, the ISIA, which were previously proposed by Hashimoto and Nodera [13], and ISIRK were compared. The CPU times and iteration numbers are shown in Table 2 and Table 3. The residual tolerance for computing $\phi_0(tM^{-1}L)(v + L^{-1}c)$, tol_ϕ was 10^{-6}, and $m^{\max} = 100$, $\delta = 0.01$. The results show that the ISIRK with $P = 2$ is efficient. Figure 3 shows the residual tolerance for solving linear equations at each Krylov step of the ISIRK

Table 1: Example 1, The number of iterations and CPU time of SIA and SIRK.

		SIRK $(P=1)$		SIRK $(P=2)$		SIA	
		Itr.	time	Itr.	time	Itr.	time
$N=\gamma=100/t$	$h=1/t$	52	129.76	52	65.73	57	117.51
$N=\gamma=200/t$	$h=1/t$	74	113.27	74	57.30	84	112.24
	$h=2/t$	65	112.50	66	64.34		

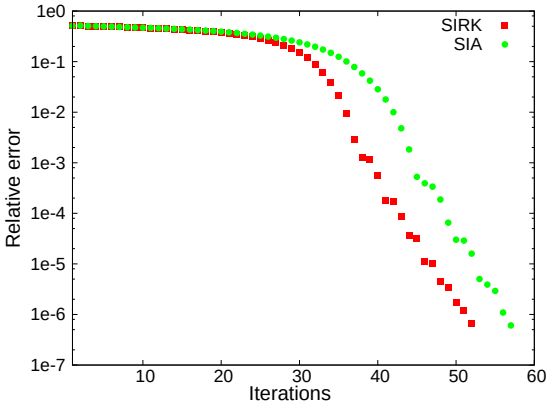

Figure 1: Example 1, Iterations versus relative error ($N=\gamma=100/t$, $h=1/t$, $P=2$).

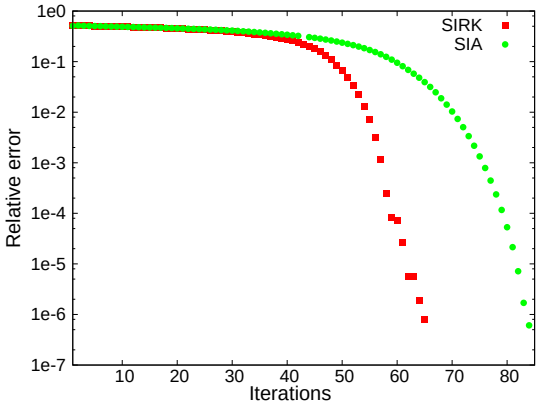

Figure 2: Example 1, Iterations versus relative error ($N=\gamma=200/t$, $h=2/t$, $P=2$).

Table 2: Example 1, The number of iterations and CPU time of SIA and ISIA.

	ISIA		SIA	
	Itr.	time	Itr.	time
$\gamma=100/t$	63	54.64	63	129.20
$\gamma=200/t$	90	48.53	90	119.80

Table 3: Example 1, The number of iterations and CPU time of SIRK and ISIRK.

	ISIRK$(P=1)$		SIRK$(P=1)$		ISIRK$(P=2)$		SIRK$(P=2)$	
	Itr.	time	Itr.	time	Itr.	time	Itr.	time
$\gamma=100/t,\ h=1/t$	60	71.85	60	158.80	60	37.67	60	80.42
$\gamma=200/t,\ h=2/t$	72	55.75	72	132.91	72	29.88	72	67.24

with $N=100/t$, $h=1/t$ and $P=1$. It is observed that the exactness needed to obtain a solution for the linear equation decreases as m becomes larger. The solutions computed with the ISIRK are tabulated in Figure 4. Problem (45) represents the flow of heat coming from boundary $\partial\Omega_1$. The temperature in region Ω is $0°C$ at $t=0$, but at this point, the heat begins to flow toward the right edge of Ω. The accuracy of the ISIRK is illustrated here.

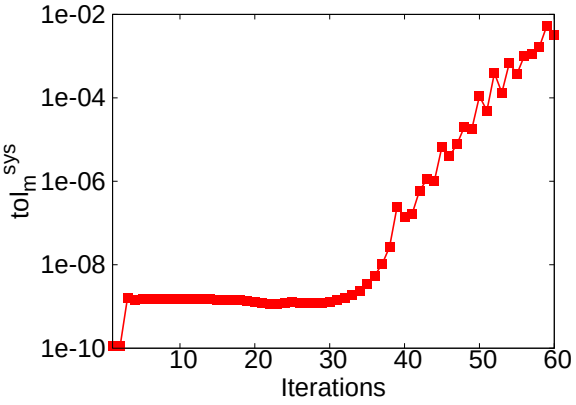

Figure 3: Example 1, Iterations versus $\mathrm{tol}_m^{\mathrm{sys}}$.

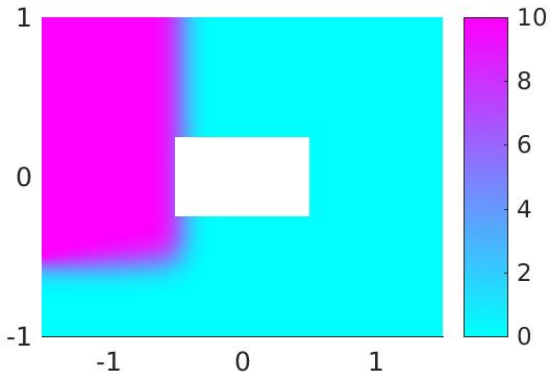

Figure 4: Example 1, Numerical solutions of ISIRK.

Example 2

The next problem is a wave equation in region $(-1.5, 1.5) \times (-1, 1) \subseteq \mathbb{R}^2$:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \Delta u = f(x,t) & \text{in } (0,T] \times \Omega, \\ u = e^{-10(x_1-0.5)^2-10(x_2-0.5)^2} & \text{on } \{0\} \times \Omega, \\ \dfrac{\partial u}{\partial t} = 0 & \text{on } \{0\} \times \Omega, \\ u = 0 & \text{on } (0,T] \times \partial\Omega_1, \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } (0,T] \times \partial\Omega_2, \end{cases}$$

where $f(x,t) = -10^4 \sin(t) e^{(x_1-0.8)^2+(x_2-0.8)^2}$, $c = \sqrt{0.1}$, $\partial\Omega_1 = [-1.5, 1.5] \times \{1, -1\}$, and $\partial\Omega_2 = \partial\Omega \setminus \partial\Omega_1$. After the discretization:

$$\begin{cases} \tilde{M}\ddot{\tilde{y}}(t) = \tilde{L}\tilde{y}(t) + \tilde{b}(t), \\ \tilde{y}(0) = \tilde{v}. \end{cases} \tag{46}$$

Equations (46) were transformed into equations (2), where:

$$M = \begin{bmatrix} \tilde{M} & \\ & I \end{bmatrix}, \quad L = \begin{bmatrix} & \tilde{L} \\ I & \end{bmatrix}, \quad b = \begin{bmatrix} \tilde{b} \\ \mathbf{0} \end{bmatrix}, \quad y = \begin{bmatrix} \dot{\tilde{y}} \\ \tilde{y} \end{bmatrix}, \quad v = \begin{bmatrix} \tilde{v} \\ \mathbf{0} \end{bmatrix},$$
$$F(y) = Ly + b(t).$$

In this example, the dimension of the matrices were $n = 237378$. We used the 1-step exponential integrator [18] whose scheme was:

$$y_{i+1} = y_i + \Delta t \phi_1(\Delta t M^{-1} L_{i+1}) M^{-1} F(y_i).$$

Table 4 shows the CPU times and iteration numbers of SIRK and SIA for computing $\phi_1(\Delta t M^{-1} L) M^{-1} F(y(0))$, where $\Delta t = 0.1$ and the relative error tolerance is 10^{-6}. We used $N = \gamma = 40/\Delta t,\ 50/\Delta t,\ 60/\Delta t$. For SIRK, we set $h = 1/\Delta t$ and $P = 2$. In this example, the iteration numbers of SIRK and SIA are almost the same. This is because in the case of ϕ_1, the functions $\hat{f}_N$ and $\tilde{f}_\gamma$ in bounds (16) and (19) are different, and

Table 4: Example 2, The number of iterations and CPU time of SIRK and SIA.

	SIRK		SIA	
	Itr.	time	Itr.	time
$N = \gamma = 40/\Delta t$	28	5.66	25	6.05
$N = \gamma = 50/\Delta t$	24	3.65	24	5.33
$N = \gamma = 60/\Delta t$	26	3.62	25	5.23

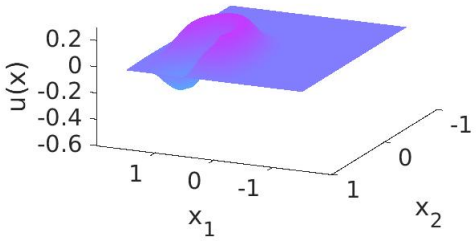

Figure 5: Example 2, Numerical solution of $t = 1$ with EI and SIRK.

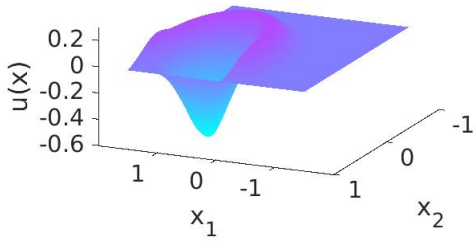

Figure 6: Example 2, Numerical solution of $t = 2$ with EI and SIRK.

$\hat{f}_N(z) > \tilde{f}_\gamma(z)$ for all z. Thus, the factor $\min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|\hat{f}_N - r\|_\Sigma$ in bound (16) is larger than $\min_{p\in\mathcal{P}_{m-1}} \|\tilde{f}_\gamma - p\|_{W((\gamma I - A)^{-1})}$ in bound (19). However, the factor e^{-hm} cancels the magnitude of $\min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|\hat{f}_N - r\|_\Sigma$ in SIRK. Moreover, since the linear equations in SIRK can be solved in parallel, its computation is faster than that of SIA.

Figure 5 and Figure 6 show the numerical solution of $t = 1$ and $t = 2$ computed with SIRK in the exponential integrator. It shows the vibration of the wave, and we see the exactness of the computation of SIRK.

Example 3

The third test problem was a Burgers equation in region $\Omega = (-1.5, 1.5) \times (-1, 1) \subseteq \mathbb{R}^2$ for confirming the effectiveness of ISIRK:

$$\begin{cases} \dfrac{\partial u}{\partial t} = u\dfrac{\partial u}{\partial x_1} + v\dfrac{\partial u}{\partial x_2} + \dfrac{1}{\mathrm{Re}}\Delta u & \text{in } (0,T]\times\Omega, \\ \dfrac{\partial v}{\partial t} = u\dfrac{\partial v}{\partial x_1} + v\dfrac{\partial v}{\partial x_2} + \dfrac{1}{\mathrm{Re}}\Delta v & \text{in } (0,T]\times\Omega, \\ u = 0, \quad v = 0 & \text{on } (0,T]\times\partial\Omega_1, \\ \dfrac{\partial u}{\partial n} = 0, \quad \dfrac{\partial v}{\partial n} = 0 & \text{on } (0,T]\times\partial\Omega_2, \\ u = f, \quad v = -f & \text{on } \{0\}\times\Omega, \end{cases}$$

where $\mathrm{Re} = 10^6$ and $f(x) = e^{-10(x_1-0.5)^2-10(x_2-0.5)^2}$. After the discretization, equation (2) was obtained with $F(y) = Ly + Q(y)y$, where $L \in \mathbb{R}^{n\times n}$ and Q is the matrix valued function of $\mathbb{R}^n \mapsto \mathbb{R}^{n\times n}$ with $n = 29649$, 118689. We set $L_i = L + Q(y_{i-1})$ and $n_i(y) = F(y) - L_i y = Q(y)y - Q(y_{i-1})y$, then used the 2-step exponential integrator [18]. The

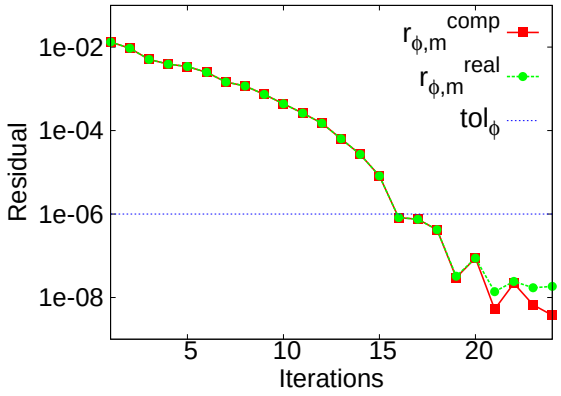

Figure 7: Example 3, Iterations versus $\|r^{\mathrm{real}}_{\phi,m}\|$ and $\|r^{\mathrm{comp}}_{\phi,m}\|$ with $\mathrm{tol}_\phi = 10^{-6}$.

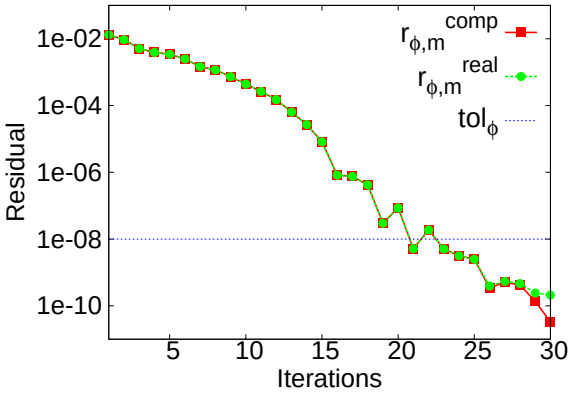

Figure 8: Example 3, Iterations versus $\|r^{\mathrm{real}}_{\phi,m}\|$ and $\|r^{\mathrm{comp}}_{\phi,m}\|$ with $\mathrm{tol}_\phi = 10^{-8}$.

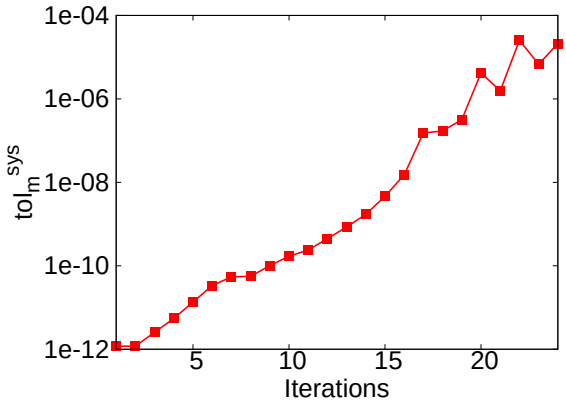

Figure 9: Example 3, Iterations versus $\mathrm{tol}^{\mathrm{sys}}_m$.

scheme was:

$$y_{i+1} = y_i + \Delta t \phi_1(\Delta t M^{-1} L_{i+1}) F(y_i) - \Delta t \frac{2}{3} \phi_2(\Delta t M^{-1} L_{i+1})[n_i(y_i) - n_i(y_{i-1})].$$

The computations of $\phi_2(\Delta t M^{-1} L_{i+1})[n_i(u_i) - n_i(u_{i-1})]$ in the third time step with $\Delta t = 0.1$ with ISIRK were observed. The residual tolerance tol_ϕ for computing ϕ-functions was 10^{-6} or 10^{-8} and $m^{\max} = 50$, $\delta = 0.01$. We also set $N = 50/\Delta t$, $h = 1/\Delta t$ and $P = 1$. Figure 7 and Figure 8 show the relationship between the number of iterations and the residuals of ISIRK with $n = 118689$. The real residual $r^{\mathrm{real}}_{\phi,m}$ decreases until it reaches tol_ϕ, but it stops decreasing after this point. This means that the linear equation is solved efficiently at each Krylov step. On the other hand, the computing residual $r^{\mathrm{comp}}_{\phi,m}$ decreased even after it had reached tol_ϕ. Moreover, the behavior of $r^{\mathrm{real}}_{\phi,m}$ and $r^{\mathrm{comp}}_{\phi,m}$ were the same before they reached tol_ϕ. Thus, $r^{\mathrm{comp}}_{\phi,m}$ was an appropriate stopping criterion for ISIRK. Figure 9 shows the residual tolerance for solving linear equations at each Krylov step with $\mathrm{tol}_\phi = 10^{-6}$ and $n = 118689$. The exactness for a solution of the linear equation decreased as m becames larger. Table 5 shows the CPU time of ISIRK and SIRK with $\mathrm{tol}_\phi = 10^{-6}$. ISIRK was faster than SIRK.

Table 5: Example 3, Comparison of SIRK and ISIRK.

Algorithm	n	Time(s)	Itr.
ISIRK	29649	0.42	14
SIRK	29649	0.59	14
ISIRK	118689	2.4	16
SIRK	118689	3.6	16

6 Conclusion

SIRK and ISIRK were explored in this paper. The advantage of SIRK is that it uses the real shifts and these shifts reduces the iteration numbers of SIRK compared to SIA. This makes SIRK the effective method for computing ϕ-functions. Furthermore, the computation cost of solving linear equations in SIRK can be improved using ISIRK. ISIRK solves linear equations efficiently while guaranteeing that the generalized residual becomes lower than the arbitrary tolerance. The exactness needed for solving a linear equation decreased as the Krylov step progressed, and the stopping criterion for the convergence of SIRK was also valid for that of the ISIRK.

References

[1] Beckermann, B. and Reichel, L., Error Estimates and Evaluation of Matrix Functions via the Faber Transform, *SIAM Journal on Numerical Analysis*, 47(5):3849–3883, 2009, https://doi.org/10.1137/080741744.

[2] Benzi, M. and Boito, P., Decay Properties for Functions of Matrices over C^*-algebras, *Linear Algebla and its Apprications*, 456(1):174–198, 2014, https://doi.org/10.1016/j.laa.2013.11.027.

[3] Benzi, M., Boito, P., and Razouk, N., Decay Properties of Spectral Projectors with Appricatons to Electronic Structure, *SIAM Review*, 55(1):3–64, 2013, https://doi.org/10.1137/100814019.

[4] Berljafa, M. and Güttel, S., Parallelization of the Rational Arnoldi Algorithm, *MIMS EPrint*, 32, 2016, http://eprints.ma.man.ac.uk/2503/01/covered/MIMS_ep2016_32.pdf.

[5] Crouzeix, M., Numerical Range and Functional Calculus in Hilbert Space, *Journal of Functional Analysis*, 244:668–690, 2007, https://doi.org/10.1016/j.jfa.2006.10.013.

[6] Druskin, V., Lieberman, C., and Zaslavsky, M., On Adaptive Choice of Shifts in Rational Krylov Subspace Reduction of Evolutionary Problems, *SIAM Journal on Scientific Computing*, 32(5):2485–2496, 2010, https://doi.org/10.1137/090774082.

[7] Gallopoulos, E. and Saad, Y., Efficient Solution of Parabolic Equations by Krylov Approximation Methods, *SIAM Journal on Scientific Statistics*, 13(5):1236–1264, 1992, https://doi.org/10.1137/0913071.

[8] Gang, W., Feng, T., and Yimin, W., An Inexact Shift-and-invert Arnoldi Algorithm for Toeplitz Matrix Exponential, *Numerical Linear Algebra with Applications*, 22(4):777–792, 2015, https://doi.org/10.1002/nla.1992.

[9] Göckler, T., *Rational Krylov Subspace Methods for 'ϕ'-functions in Exponential Integrators*, PhD thesis, Karlsruher Instituts für Technologie, 2014, http://d-nb.info/1060425408/34.

[10] Grimm, V., Resolvent Krylov subspace approximation to operator functions, *BIT Numerical Mathematics*, 52:639–659, 2012, https://doi.org/10.1007/s10543-011-0367-8.

[11] Güttel, S., *Rational Krylov Methods for Operator Functions*, PhD thesis, Technischen Universität Bergakademie Freiberg, 2010, http://www.qucosa.de/fileadmin/data/qucosa/documents/2764/diss_guettel.pdf.

[12] Güttel, S., Rational Krylov Approximation of Matrix Functions: Numerical Methods and Optimal Pole Selection, *GAMM-Mitteilungen*, 38(1):8–31, 2013, https://doi.org/10.1002/gamm.201310002.

[13] Hashimoto, Y. and Nodera, T., Inexact Shift-invert Arnoldi Method for Evolution Equations, *ANZIAM Journal*, 58(E):E1–E27, 2016, https://doi.org/10.21914/anziamj.v58i0.10766.

[14] Higham, N. J., The Scaling and Squaring Method for the Matrix Exponential Revisited, *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, 2005, https://doi.org/10.1137/04061101X.

[15] Hochbruck, M. and Lubich, C., On Krylov Subspace Approximations to the Matrix Exponential Operator, *SIAM Journal on Numerical Analysis*, 34(5):1911–1925, 1997, https://doi.org/10.1137/S0036142995280572.

[16] Hochbruck, M., Lubich, C., and Selhofer, H., Exponential Integrators for Large Systems of Differential Equations, *SIAM Journal on Scientific Computing*, 19(5):1552–1574, 1997, https://doi.org/10.1137/S1064827595295337.

[17] Hochbruck, M. and Ostermann, A., Exponential Runge-Kutta Methods for Parabolic Problems, *Applied Numerical Mathematics*, 53(2–4):323–339, 2005, https://doi.org/10.1016/j.apnum.2004.08.005.

[18] Hochbruck, M. and Ostermann, A., Exponential Integrators, *Acta Numerica*, 19:209–286, 2010, https://doi.org/0.1017/S0962492910000048.

[19] Hongqing, Z., Huazhong, S., and Meiyu, D., Numerical Solutions of Two-dimensional Burgers' Equations by Discrete Adomian Decomposition Method, *Computers and Mathematics with Applications*, 60(3):840–848, 2010, https://doi.org/10.1016/j.camwa.2010.05.031.

[20] Moler, C. and Van Loan, C. F., Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-five Years Later, *SIAM Review*, 45(1):3–49, 2003, https://doi.org/10.1137/S00361445024180.

[21] Novati, P., Using the Restricted-denominator Rational Arnoldi Method for Exponential Integrators, *SIAM Journal on Numerical Analysis and Applications*, 32(4):1537–1558, 2011, https://doi.org/10.1137/100814202.

[22] Ruhe, A., Rational Krylov, A Practical Algorithm for Large Sparse Nonsymmetric Matrix Pencils, Technical Report UCB/CSD-95-871, EECS Department, University of California, Berkeley, Apr 1995, http://www2.eecs.berkeley.edu/Pubs/TechRpts/1995/5203.html.

[23] Saad, Y. and Schultz, M. H., GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems, *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1983, https://doi.org/10.1137/0907058.

[24] Saff, E. G., Schönhage, A., and Varga, R. S., Geometric Convergence to e^{-z} by Rational Functions with Real Poles, *Numerische Mathematik*, 25(3):307–322, 1975, https://doi.org/10.1007/BF01399420.

[25] Skoogh, D., A Parallel Rational Krylov Algorithm for Eigenvalue Computations, *Applied Parallel Computing Large Scale Scientific and Industrial Problems*, pp. 521–526, 1998, https://doi.org/10.1007/BFb0095377.

[26] Svoboda, Z., The Convective-diffusion Equation and Its Use in Building Physics, *International Journal on Architectural Science*, 1(2):68–79, 2000, http://www.bse.polyu.edu.hk/researchCentre/Fire_Engineering/summary_of_output/journal/IJAS/V1/p.68-79.pdf.

[27] Van den Eshof, J. and Hochbruck, M., Preconditioning Lanczos Approximations to the Matrix Exponential, *SIAM Journal on Scientific Computing*, 27(4):1438–1457, 2006, https://doi.org/10.1137/040605461.

[28] Van der Vorst, H. A., Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems, *SlAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992, https://doi.org/10.1137/0913035.

KSTS/RR-17/001
January 27, 2017
(Revised July 24, 2017)

Department of Mathematics
Faculty of Science and Technology
Keio University

Research Report

2016

[16/001] Shiro Ishikawa,
Linguistic interpretation of quantum mechanics: Quantum Language [*Ver. 2*],
KSTS/RR-16/001, January 8, 2016

[16/002] Yuka Hashimoto, Takashi Nodera,
Inexact shift-invert Arnoldi method for evolution equations,
KSTS/RR-16/002, May 6, 2016

[16/003] Yuka Hashimoto, Takashi Nodera,
A Note on Inexact Rational Krylov Method for Evolution Equations,
KSTS/RR-16/003, November 9, 2016

[16/004] Sumiyuki Koizumi,
On the theory of generalized Hilbert transforms (*Chapter V: The spectre analysis and synthesis on the N.Wiener class S*), KSTS/RR-16/004, November 25, 2016

[16/005] Shiro Ishikawa,
History of Western Philosophy from the quantum theoretical point of view,
KSTS/RR-16/005, December 6, 2016

2017

[17/001] Yuka Hashimoto, Takashi Nodera,
Inexact Shift-invert Rational Krylov Method for Evolution Equations,
KSTS/RR-17/001, January 27, 2017 (Revised July 24, 2017)

[17/002] Dai Togashi, Takashi Nodera,
Convergence analysis of the GKB-GCV algorithm,
KSTS/RR-17/002, March 27, 2017

[17/003] Shiro Ishikawa,
Linguistic solution to the mind-body problem,
KSTS/RR-17/003, April 3, 2017

[17/004] Shiro Ishikawa,
History of Western Philosophy from the quantum theoretical point of view; Version 2,
KSTS/RR-17/004, May 12, 2017

[17/005] Sumiyuki Koizumi,
On the theory of generalized Hilbert transforms (*Chapter VI: The spectre analysis and synthesis on the N.Wiener class S* (2)), KSTS/RR-17/005, June 8, 2017