Research Report

KSTS/RR-17/001
January 27, 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

Inexact Shift-invert Rational Krylov Method for Evolution Equations

Yuka Hashimoto* Takashi Nodera†

January 27, 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 determines the appropriate shifts in the simple way. 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 [24] and the Burgers equation in fluid mechanics [17]. Let $\Omega \subseteq \mathbb{R}^d$ be an open set, $\partial\Omega = \partial\Omega_1 \bigcup \partial\Omega_2$ be the boundary of Ω, and n_b be the unit normal vector of $\partial\Omega_2$. In addition, the time space is defined as $[0,T]$, where $T > 0$ is the maximum time we are interested in. $l \in \mathbb{N}$ is defined as the order of the time derivative. The problem is defined in $[0,T] \times \overline{\Omega}$, and its solution is defined in $\mathcal{V}$. $\mathcal{V}$ is the Hilbert space contained by $L^2([0,T] \times \overline{\Omega})$. Let $\mathcal{D}$ be a linear or nonlinear differential operator on $\mathcal{V}$, and ξ, η, τ_1, τ_2 be known functions. The following initial boundary value problems are explored:

$$\begin{cases} \dfrac{\partial^l u}{\partial t^l} = \mathcal{D}u & \text{in } (0,T] \times \Omega, \\ u = \xi & \text{on } \{0\} \times \overline{\Omega}, \\ u = \eta & \text{on } (0,T] \times \partial\Omega_1, \\ \dfrac{\partial u}{\partial n_b} = \tau_1 u + \tau_2 & \text{on } (0,T] \times \partial\Omega_2. \end{cases} \tag{1}$$

*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

A different algebraic equation is derived from a spatial discretization with the finite element method or finite difference method:

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

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

If $\mathcal{D}$ is linear and does not depend on t, equation (2) is the linear ordinally 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 $L \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$ and $\phi_0(z) := e^z$ [10]. On the other hand, if $\mathcal{D}$ is nonlinear or depends on t, time discretization is also needed for integrating $M^{-1}F(t, y)$ and finding solution $y(t)$. The exponential integrator [12, 13, 15, 16] is currently the popular method for time integration [10]. 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

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

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

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

Various methods for computing matrix ϕ-functions have been developed [6, 13, 18–21]. 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 [13, 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 is not the best choice 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 [8] and Göckler [7]. According to Göckler [7], SIA and RK converge independently of the width of the numerical range of $\Delta t M^{-1} L_{i+1}$. 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. To address this issue, the Inexact Shift-invert Arnoldi method (ISIA) was proposed by Hashimoto and Nodera [10]. This method solves the linear equations efficiently while guaranteeing the accuracy of the solution. The computational time can be reduced using the ISIA, but this requires a shift, and choosing the appropriate shift is difficult. This situation also occurs in the SIA and the RK; this is the second shortcoming of the SIA and the RK. RK needs different shifts in every step of the Krylov process, so choosing the appropriate shifts is integral. The ways choosing the appropriate shifts in SIA and RK for ϕ_0 and other functions have been discussed at length, for example [5, 9, 23]. 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 [7] proposed a simple way of choosing the shifts for general ϕ-functions of nonsymmetric matrices. According to his paper, the optimal shift for SIA changes at every iteration, although only one shift is permitted for the SIA. He also proposed a method for the RK, but this 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. In summary, the existing methods for choosing the shifts are not realistic in this scenario. 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. 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 $\mathbb{C}^- := \{z \in \mathbb{C} \mid \Re(z) < 0\}$, $\mathbb{C}^+ := \{z \in \mathbb{C} \mid \Re(z) > 0\}$, 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. The method for computing $\phi_k(\Delta t M^{-1} L_{i+1}) M^{-1} v$ is based on the method developed by Hashimoto and Nodera [10]. Throughout this and the next section, it is assumed that $W(A) \subseteq \mathbb{C}^-$.

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 process is:

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

where $\gamma > 0$ 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:

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

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

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

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

Göckler showed that the error bound of approximation (7) does not depend on $W(A)$ [7, Theorem 5.9].

2.2 Inexact Shift-invert Arnoldi method (ISIA)

SIA requires solving the linear equation to compute $(\gamma I - A)^{-1}v_j$ in every step of the Krylov process. Hashimoto and Nodera [10] proposed a method for solving this linear equation efficiently while guaranteeing that the generalized residual [14] would become smaller than the arbitrary tolerance. This method is called the Inexact Shift-invert Arnoldi method (ISIAP), 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:

$$h_{j+1,j}v_{j+1} = (\gamma_j I - A)^{-1}v_j - \sum_{k=1}^{j} h_{k,j}v_k,$$

$$h_{k,j} = v_k^*(\gamma_j I - A)^{-1} v_j,$$
$$h_{j+1,j} = \left\| (\gamma_j I - A)^{-1} v_j - \sum_{k=1}^{j} h_{k,j} v_k \right\| \qquad (j = 1, \ldots, m),$$

where $\gamma_j > 0$ $(1 \leq j \leq m)$ is a different shift in every step. 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

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

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

Göckler shows that under the appropriate choice of shifts γ_j, the error bound of approximation (8) does not depend on $W(A)$ [7, Theorem 7.8].

3 Shift-invert Rational Krylov method (SIRK)

We consider extending ISIA to the rational approximation with more than one poles. 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 derives its relations in the same manner as illustrated in section 2.3:

$$\begin{aligned}V_m &= V_m H_m D_m - A V_m H_m + (\gamma_m I - A) h_{m+1,m} v_{m+1} e_m^*, \\ V_m^*(\gamma_m I - A)^{-1} V_m &= H_m (I - H_m D_m + \gamma_m H_m)^{-1} =: K_m,\end{aligned} \tag{9}$$

where $D_m := \mathrm{diag}\{\gamma_1, \cdots, \gamma_m\}$. However, in the SIRK, the shifts $\gamma_j = N - j \in \mathbb{R}$, where $N \in \mathbb{N}$ satisfies $\gamma_j > 0$ $(1 \leq^{\forall} j \leq m)$ are used. The simplest way of determining N is setting $N = m^{\max} + 1$, where $m^{\max}$ is the maximum iteration number. If H_m is invertible, the matrix function $\phi_k(A)v$ is approximated as:

$$\begin{aligned}\phi_k(A)v &= f_m((\gamma_m I - A)^{-1})v \\ &\approx V_m f_m(V_m^*(\gamma_m I - A)^{-1} V_m) V_m^* v \\ &= V_m f_m(K_m) V_m^* v \\ &= V_m \phi_k\left(\gamma_m I - (I - H_m D_m + \gamma_m H_m) H_m^{-1}\right) V_m^* v \\ &= V_m \phi_k\left((H_m D_m - I) H_m^{-1}\right) V_m^* v,\end{aligned} \tag{10}$$

where $f_m(x) := \phi_k\left(\gamma_m - x^{-1}\right)$. Approximation (10) 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 $\gamma_j = N - j$, so X_j is denoted as:

$$X_j = (\gamma_j I - A)^{-1} = (I - (\gamma_m - \gamma_j) X_m)^{-1} X_m = (I + (m - j) X_m)^{-1} X_m. \tag{11}$$

From relation (11), the Rational Krylov subspace generated by the m-step SIRK is represented as:

$$\begin{aligned}&\mathrm{Span}\{v,\ X_1 v,\ \ldots,\ X_{m-1}v\}\\ &= \mathrm{Span}\{v,\ (I+(m-1)X_m)^{-1}X_m v,\ \ldots,\ (I+X_m)^{-1}X_m v\}\\ &= \{r(X_m)v \mid\ r \in \mathcal{P}_{m-1}/q_{m-1},\ q_m(z) = (1+mz)\ldots(1+z)\}\,.\end{aligned} \tag{12}$$

The following proposition shown by Beckermann and Reichel [1] is valid from relation (12), and the following theorem regarding the convergence of SIRK is deduced:

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

$$r(X_m)v = V_m r(K_m)V_m^* v. \tag{13}$$

Theorem 3.1 *Let $\mathcal{H}(\Pi)$ be the set of holomorphic functions on a closed and bounded set $\Pi \subseteq \mathbb{C}$ to $\mathbb{C}$. Let $1 \le C \le 11.08$, and $f(z) := \int_0^1 e^{N-sz^{-1}}(1-s)^{k-1}/(k-1)!ds$. 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, for $1 \le m \le N-1$, the error bound of SIRK is estimated as*

$$\|\phi_k(A)v - V_m f_m(K_m)V_m^* v\| \le 2C\|v\|e^{-m} \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|f-r\|_\Sigma, \tag{14}$$

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

Proof : Since $W(A) \subseteq \mathbb{C}^-$ and $\gamma_j = N - j > 0$, $W(X_j) \subseteq \mathbb{C}^+$ is satisfied for all j in $1 \le j \le N-1$. In addition, $W(X_j)$ are bounded. Thus, 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,

$$\begin{aligned}&\|\phi_k(A)v - V_m f_m(K_m)V_m^* v\|\\ &\quad = \|f_m(X_m)v - r(X_m)v - V_m f_m(K_m)V_m^* v + V_m r(K_m)V_m^* v\|,\end{aligned} \tag{15}$$

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, $f_m,\ f \in \mathcal{H}(\Sigma)$. From equation (9), $W(K_m) \subseteq W(X_m)$, and from Crouzeix [4], there is $1 \le C \le 11.08$ such that:

$$\begin{aligned}\|f_m(X_m) - r(X_m)\| &\le C\|f_m - r\|_\Sigma,\\ \|f_m(K_m) - r(K_m)\| &\le C\|f_m - r\|_\Sigma.\end{aligned} \tag{16}$$

$r \in \mathcal{P}_{m-1}/q_{m-1}$ is arbitrary, and ϕ_k is represented as $\phi_k(z) = \int_0^1 e^{sz}(1-s)^{k-1}/(k-1)!ds$, so it is deduced that:

$$\begin{aligned}&\|\phi_k(A)v - V_m f_m(K_m)V_m^* v\|\\ &\quad\le \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} [\|f_m(X_m) - r(X_m)\|\,\|v\| + \|f_m(K_m) - r(K_m)\|\,\|v\|] \qquad (\because (15))\\ &\quad\le 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|f_m - r\|_\Sigma \qquad (\because (16))\\ &\quad= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} |\phi_k(N-m-z^{-1}) - r(z)|\end{aligned}$$

$$
\begin{aligned}
&= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left| \int_0^1 e^{s(N-m-z^{-1})} \frac{(1-s)^{k-1}}{(k-1)!} - e^{s(N-m)} \left(e^{-s(N-m)} r(z)\right) ds \right| \\
&\leq 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left| e^{N-m} \left\{ \int_0^1 e^{-sz^{-1}} \frac{(1-s)^{k-1}}{(k-1)!} ds - \int_0^1 e^{-s(N-m)} r(z) ds \right\} \right| \\
&= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \sup_{z\in\Sigma} \left| e^{-m} \left\{ \int_0^1 e^{N-sz^{-1}} \frac{(1-s)^{k-1}}{(k-1)!} ds - \int_0^1 e^{N-s(N-m)} ds\, r(z) \right\} \right| \\
&= 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} e^{-m} \sup_{z\in\Sigma} \left| \int_0^1 e^{N-sz^{-1}} \frac{(1-s)^{k-1}}{(k-1)!} ds - r(z) \right|.
\end{aligned}
$$

□

Choosing γ_j as $N-j$ results in the space $\mathcal{P}_{m-1}/q_{m-1}$ expanding with each iteration, because q_m has the form $q_m(z) = (1+mz)\cdots(1+z)$. Therefore, $\min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|f-r\|_\Sigma$ in error bound (14) becomes smaller as m becomes larger. In addition, e^{-m} becomes smaller as m becomes larger. The term e^{-m} accelerates the convergence.

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 $j = 1 \ldots m$, 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.

$$
\begin{aligned}
&(\gamma_j I - A)^{-1} v_j - f_j^{\text{sys}} = \sum_{k=1}^{j+1} h_{k,j} v_k, \\
&v_j = \sum_{k=1}^{j+1} h_{k,j} (\gamma_j I - A) v_k + r_j^{\text{sys}}, \\
&V_m = V_m H_m D_m - A V_m H_m + h_{m+1,m} (\gamma_m I - A) v_{m+1} e_m^* + R_m^{\text{sys}}, \\
&(\gamma_j I - A) V_m = V_m K_m^{-1} - h_{m+1,m} (\gamma_m I - A) v_{m+1} e_m^* H_m^{-1} - R_m^{\text{sys}} H_m^{-1},
\end{aligned} \tag{17}
$$

where V_m is the $n \times m$ matrix with orthonormal columns, H_m is an $m \times m$ upper Hessenberg matrix, and $K_m = H_m(I - H_m D_m + \gamma_m H_m)^{-1}$. The matrices V_m, H_m and K_m in equation (17) are different matrices from equation (9). For the approximation, the same fomula used by the SIRK is employed:

$$
\phi_k(A)v \approx V_m f_m(K_m) V_m^* v. \tag{18}
$$

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

$$
E_m = \tilde{f}_m(\gamma_m I - A)v - V_m \tilde{f}_m(K_m^{-1}) V_m^* v
$$

$$= \frac{1}{2\pi i}\int_\Gamma \tilde{f}_m(\lambda)\left[(\lambda I-\gamma_m I+A)^{-1}v - V_m(\lambda I-K_m^{-1})^{-1}V_m^* v\right] d\,\lambda$$
$$= \frac{1}{2\pi i}\int_\Gamma \tilde{f}_m(\lambda) e_m^{\mathrm{lin}}\, d\lambda, \tag{19}$$

where Γ is a contour enclosing the eigenvalues of $\gamma_m I - A$ and K_m^{-1}, $e_m^{\mathrm{lin}} = [(\lambda I - \gamma_m I + A)^{-1}v - V_m(\lambda I - K_m^{-1})^{-1}V_m^* v]$. 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:

$$\begin{aligned}
&\|\hat{f}((\gamma_m I - A))v - V_m\hat{f}(K_m)V_m^* v\| \\
&\quad \le \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \left[\|\hat{f}(X_m) - r(X_m)\|\,\|v\| + \|\hat{f}(K_m) - r(K_m)\|\,\|v\|\right] \\
&\quad \le 2C\|v\| \min_{r\in\mathcal{P}_{m-1}/q_{m-1}} \|\hat{f} - r\|_\Sigma,
\end{aligned} \tag{20}$$

where Σ is the same set as Theorem 3.1. In this case, $\hat{f}$ does not depend on m, so the upper bound (20) decreases as m becomes large. Therefore, this approximation converges. Using equation (17), the residual r_m^{lin} of this approximation for the linear equation is represented as

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

Replacing e_m^{lin} with r_m^{lin} in equation (19), the generalized residual $r_{\phi,m}^{\mathrm{real}}$ [14] of the approximated $\phi_k(A)v$ is

$$\begin{aligned}
r_{\phi,m}^{\mathrm{real}} &= -h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1}\tilde{f}_m(K_m^{-1})V_m^* v - R_m^{\mathrm{sys}}H_m^{-1}\tilde{f}_m(K_m^{-1})V_m^* v \\
&= -h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^* H_m^{-1}\phi_k\left((H_m D_m - I)H_m^{-1}\right)V_m^* v \\
&\qquad\qquad - R_m^{\mathrm{sys}}H_m^{-1}\phi_k\left((H_m D_m - I)H_m^{-1}\right)V_m^* v \\
&= -\beta h_{m+1,m}(\gamma_m I - A)v_{m+1}e_m^*\phi_k\left(D_m - H_m^{-1}\right)H_m^{-1}e_1 \\
&\qquad\qquad - \beta R_m^{\mathrm{sys}}\phi_k\left(D_m - H_m^{-1}\right)H_m^{-1}e_1
\end{aligned} \tag{21}$$

In order to evaluate equation (21), the following lemma and propositions are used.

Lemma 4.1 (see [10, Proposition 2]) *Let f be the holomorphic function in $\mathbb{C}^+$ (resp. $\mathbb{C}^-$). If the sequence of the upper Hessenberg matrices $\{H_m \in \mathbb{R}^{m\times m}\}_{m=1}^n$ satisfies*

$$W(H_m) \subseteq \mathbb{C}^+ \ (\textit{resp. } \mathbb{C}^-) \quad (1 \le^\forall m \le n), \tag{22}$$

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

$$\left|[f(H_m)]_{i,j}\right| \leq \alpha\lambda^{i-j} \quad (i \geq j). \tag{23}$$

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 which satisfies*

$$|(K_m)_{i,j}| \leq \alpha\lambda^{i-j} \ (i \geq j), \tag{24}$$

where $\alpha > 0$ *and* $0 < \lambda < 1$ *which do not depend on* m*. Let* $\hat{\alpha} = \alpha + \alpha\sqrt{1/(1-\lambda^2)}$*. If*

$$\lambda \leq \sqrt{\frac{1}{2\hat{\alpha}^2+1}}, \tag{25}$$

$$\lambda < \frac{1}{\sqrt{2}}, \tag{26}$$

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 satisfies:

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

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 $(I_{m-j} + \tilde{Q}_{j+1})k_{j+1:m,j} = \eta_j e_1$. Therefore, 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 (24), 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, 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

$$\begin{cases} |(\hat{Q}_{k_r}\ldots\hat{Q}_{k_1}\hat{Q}_l)_{i,j}| \leq \dfrac{\check{\alpha}^{r+1}\lambda^{-2(l-r-1)}}{(1-\lambda^2)^r}\lambda^{i+j} =: \alpha''(l,r)\lambda^{i+j} \quad (i,j \leq k_r) \\ |(\hat{Q}_{k_r}\ldots\hat{Q}_{k_1}\hat{Q}_l)_{i,j}| = 0 \quad (i > k_r \text{ or } j > k_r) \end{cases} \tag{27}$$

Inequality (27) 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 (27) 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 (27). In inequality (27), if $\lambda \leq \sqrt{1/(1+\check{\alpha})}$, 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:

$$\begin{aligned}Q_m &= (I_m + \hat{Q}_{m-1})\dots(I_m+\hat{Q}_2) \\ &= I_m + \sum_{k=3}^{m-1}\sum_{l=2}^{k-1}\sum_{(a_1,a_2,\dots,a_{k-l-1})\in\{0,1\}^{k-l-1}}\hat{Q}_k\hat{Q}_{k-1}^{a_1}\hat{Q}_{k-1}^{a_2}\dots\hat{Q}_{l-1}^{a_{k-l-1}}\hat{Q}_l + \sum_{k=2}^{m-1}\hat{Q}_k. \qquad (28)\end{aligned}$$

As a result, for $2 \leq \min\{i,j\} \leq m-1$, under the assumption of (25), equation (28) and inequality (27) shows that:

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

Therefore, for $2 \leq \min\{i,j\} \leq m-1$ and $i \leq j$, under the assumptions of (25) and (26), it is deduced that:

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

$$
\begin{aligned}
&\leq 2\frac{\check{\alpha}^2\lambda^4(2\lambda^2)^{-1}}{(1-\lambda^2)(1-2\lambda^2)}\lambda^{i+j}\frac{\lambda^{-2}}{\lambda^{-2}-1}(\lambda^{-2})^{i-2}+\frac{\check{\alpha}^2\lambda^4}{(1-\lambda^2)^2}\lambda^{i+j}\lambda^{-2i}\\
&=\frac{\check{\alpha}^2\lambda^6}{(1-\lambda^2)^2(1-2\lambda^2)}\lambda^{j-i}+\frac{\check{\alpha}^2\lambda^4}{(1-\lambda^2)^2}\lambda^{j-i}\\
&=\frac{\check{\alpha}^2\lambda^4}{(1-\lambda^2)(1-2\lambda^2)}\lambda^{j-i}=:\alpha'\lambda^{j-i},
\end{aligned}
$$

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

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

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

$$
\begin{cases}
|(Q_m)_{1,1}|=1\\
|(Q_m)_{i,j}|=0\quad(i\neq1\text{ or }j\neq1)
\end{cases}
$$

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 $D_m-H_m^{-1}$ satisfies*

$$
W(D_m-H_m^{-1})\subset\mathbb{C}^-\quad(1\leq^{\forall}m\leq n),\tag{29}
$$

then there exist $\alpha>0$ and $0<\lambda<1$ which do not depend on m such that:

$$
\left|\left[\phi_k\left(D_m-H_m^{-1}\right)H_m^{-1}\right]_{i,1}\right|\leq\frac{1}{2}\alpha(i+1)i\lambda^{i-1}.\tag{30}
$$

Proof : It is based on the assumption of D_m that $W(D_m)\subseteq\mathbb{C}^+$ are derived. This fact and condition (29) imply that the numerical range of H_m^{-1} satisfies:

$$
W(H_m^{-1})\subseteq W(D_m)-W(D_m-H_m^{-1})\subseteq\mathbb{C}^+.
$$

Furthermore, for matrix $H\in\mathbb{R}^{m\times m}$, we have

$$
x^*Hx=x^*H^*H_m^{-*}Hx=(Hx)^*H^{-*}(Hx)=\|Hx\|^2\frac{(Hx)^*H^{-*}(Hx)}{\|Hx\|^2}\quad(^{\forall}x\in\mathbb{C}^m,\ \|x\|=1),
$$

$$
\Re(x^*H^{-*}x)=\Re(x^*H^{-1}x)\quad(^{\forall}x\in\mathbb{C}^m,\ \|x\|=1).
$$

Therefore, we have

$$
W(H_m)\subseteq\|H_m\|^2W(H^{-*})\subseteq\mathbb{C}^+.\tag{31}
$$

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:

$$
|[H_m^{-1}]_{i,j}|\leq\hat{\alpha}\hat{\lambda}^{i-j}\ (i\geq j).\tag{32}
$$

D_m is a diagonal matrix, so redefining $\hat{\alpha}$ as the sum of $\|D_m\| = N-1$ and the previous $\hat{\alpha}$ leads to:

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

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

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

The second equality is held, because from inequality (32), 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, from condition (29), $\tilde{H}_m$ satisfies

$$W(\tilde{H}_m) = W\left(Q_m(D_m - H_m^{-1})Q_m^*\right) = W(D_m - H_m^{-1}) \subset \mathbb{C}^-. \tag{34}$$

Therefore, setting $f = \phi_k$ and using Lemma 4.1 derive that there exist $\check{\alpha} > 0$ and $0 < \check{\lambda} < 1$ such that $|[\phi_k(\tilde{H}_m)]_{i,j}| \leq \check{\alpha}\check{\lambda}^{i-j}$ $(i \geq j)$. Since $|e^z| \leq 1$ is satisfied when $z \in \mathbb{C}^-$, $|\phi_k(z)|$ is bounded as

$$|\phi_k(z)| = \left|\int_0^1 e^{sz} \frac{(1-s)^{k-1}}{(k-1)!} ds\right| \leq |e^z| \left|\int_0^1 \frac{(1-s)^{k-1}}{(k-1)!} ds\right| \leq \frac{1}{k!} \quad (z \in \mathbb{C}^-). \tag{35}$$

Using the theorem by Crouzeix [4, Theorem 2], condition (34) and inequality (35), there exists $1 \leq C \leq 11.08$ such that

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

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

$$\left|\left[\phi_k(\tilde{H}_m)\right]_{i,j}\right| \leq \check{\alpha}\check{\lambda}^{i-j} \quad (i \geq j), \tag{36}$$

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

From the upper bounds (36) and (37), it is deduced that:

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

$$= i\bar{\alpha}\bar{\lambda}^{i-1}. \tag{38}$$

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 (38), it is deduced that:

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

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

□

If the residual of solving the linear equation satisfies $\|r_m^{\text{sys}}\| \leq \delta$ for some $\delta > 0$, then there exist $\alpha > 0$ and $0 < \lambda < 1$ such that the first term of equation (21) becomes:

$$\begin{aligned}
&\beta \left|h_{m+1,m} e_m^* \phi_k(D_m - H_m^{-1})H_m^{-1} e_1\right| \, \|(\gamma_m I - A)v_{m+1}\| \\
&\quad\leq \beta|h_{m+1,m}| \left|\left[\phi_k(D_m - H_m^{-1})H_m^{-1}\right]_{m,1}\right| \, \|\gamma_m I - A\| \, \|v_{m+1}\| \\
&\quad\leq \beta|h_{m+1,m}| \, \|\gamma_m I - A\| \frac{1}{2}\alpha m(m+1)\lambda^{m-1} \quad (\because (30)) \\
&\quad\leq \frac{\beta}{2}\|(\gamma_m I - A)^{-1}v_m - f_m^{\text{sys}} - h_{1,m}v_1 - \ldots - h_{m,m}v_m\| \\
&\qquad\qquad \|\gamma_m I - A\|\alpha m(m+1)\lambda^{m-1} \\
&\quad\leq \frac{\beta}{2}(\|(\gamma_m I - A)^{-1}v_m\| + \|f_m^{\text{sys}}\|)\|\gamma_m I - A\|\alpha m(m+1)\lambda^{m-1} \\
&\quad\leq \frac{\beta}{2}(1 + \|r_m^{\text{sys}}\|)\|(\gamma_m I - A)^{-1}\| \, \|\gamma_m I - A\|\alpha m(m+1)\lambda^{m-1} \\
&\quad\leq \frac{\beta}{2}(1+\delta)\kappa(\gamma_m I - A)\alpha m(m+1)\lambda^{m-1}.
\end{aligned} \tag{39}$$

Since $0 < \lambda < 1$, the upper bound (39) implies that under the assumptions of (25) and (26), if $\kappa(\gamma_m I - A)$ does not increase as m becomes larger, the first term of equation (21) decreases as m becomes larger.

Concerning the second term of equation (21), the following theorem is deduced:

Theorem 4.1 *Let* $[\phi_k(D_m - H_m^{-1})H_m^{-1}]_{i,j} =: g_{i,j}^m$. *Moreover, let* $\text{tol}_\phi > 0$ *be the tolerance for computing the* ϕ*-function and* $m^{\max}$ *be the maximum number of iterations. Under the assumptions of (25), (26) and (29), If,*

$$\|r_1^{\text{sys}}\| \leq \frac{\text{tol}_\phi}{2m^{\max}\beta\|\phi_k(D_m - H_m^{-1})H_m^{-1}e_1\|}, \tag{40}$$

$$\|r_j^{\rm sys}\| \leq \frac{|g_{1,1}^m|}{|g_{j-1,1}^m|}\|r_1^{\rm sys}\| \quad (2 \leq j \leq m), \tag{41}$$

then,

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

Proof : Based on the above assumptions (40), (41) and Proposition 4.2, the upper bound is derived:

$$\begin{aligned}
&\beta\|R_m^{\rm sys}\phi_k(D_m - H_m^{-1})H_m^{-1}e_1\| \\
&\quad\leq \beta(|g_{1,1}^m|\,\|r_1^{\rm sys}\| + |g_{2,1}^m|\,\|r_2^{\rm sys}\| + \ldots + |g_{m,1}^m|\,\|r_m^{\rm sys}\|) \\
&\quad\leq \beta\left(|g_{1,1}^m|\,\|r_1^{\rm sys}\| + |g_{2,1}^m|\frac{|g_{1,1}^m|}{|g_{1,1}^m|}\|r_1^{\rm sys}\| + |g_{3,1}^m|\frac{|g_{1,1}^m|}{|g_{2,1}^m|}\|r_1^{\rm sys}\| + \ldots + |g_{m,1}^m|\frac{|g_{1,1}^m|}{|g_{m-1,1}^m|}\|r_1^{\rm sys}\|\right) \\
&\qquad\qquad (\because (41)) \\
&\quad= \beta|g_{1,1}^m|\,\|r_1^{\rm sys}\|\left(1 + \frac{|g_{2,1}^m|}{|g_{1,1}^m|} + \frac{|g_{3,1}^m|}{|g_{2,1}^m|} + \ldots + \frac{|g_{m,1}^m|}{|g_{m-1,1}^m|}\right) \\
&\quad\lesssim \beta\|\phi_k(D_m - H_m^{-1})H_m^{-1}e_1\|\,\|r_1^{\rm sys}\|\left(1 + 3\lambda + 2\lambda\ldots + \frac{m(m+1)}{(m-1)m}\lambda\right) \quad (\because (30)) \\
&\quad\leq \beta\|\phi_k(D_m - H_m^{-1})H_m^{-1}e_1\|\,\|r_1^{\rm sys}\| \cdot 2m^{\max} \\
&\quad\leq {\rm tol}_\phi \quad (\because (40)).
\end{aligned}$$

□

The right-hand side of inequality (41) 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 $(\gamma_m I - A)x_m = v_m$ becomes more inexact, and the computational cost decreases compared to the SIRK. However, if the linear equations are solved, satisfying inequalities (40) and (41), then the second term of equation (21) is no longer an issue. In this scenario, the first term of equation (21), $r_{\phi,m}^{\rm comp}$, is used as the stopping criterion for the convergence of ISIRK.

Remark 4.1 *In practical computation, the values depending on m in inequalities (40) and (41) are unavailable in advance. Thus, for the exponential integrator at the $(i+1)$th step, $\phi_k(D_m - H_m^{-1})H_m^{-1}e_1$ is replaced with the ones in the largest Krylov subspace at the ith step. For the computation of equation (3):*

$$\begin{aligned}
V_m^*(\gamma_m I - A)V_m &\approx (I - H_m D_m + \gamma_m H_m)H_m^{-1} \\
H_m^{-1} &\approx H_m D_m H_m^{-1} - V_m^* A V_m,
\end{aligned}$$

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 $H_m^{-1}e_1 \approx (H_m^{-1})_{1,1}e_1$. Thus,

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

Algorithm 4.1 ISIRK method for ϕ-functions in the exponential integrator of the ith step

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

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

$$\begin{aligned}\|H_m^{-1}\phi_k((H_m D_m - I)H_m^{-1})e_1\| &\approx \|H_m^{-1}[\phi_k((H_m D_m - I)H_m^{-1})]_{1,1}e_1\| \\ &\approx \|V_m^*(\gamma_1 I - A)V_m \phi_k((H_m D_m - I)H_m^{-1})e_1\| \\ &\approx \|(\gamma_1 I - A)y(t)\| \\ &\approx \|(\gamma_1 I - A)y(0)\|. \end{aligned} \tag{42}$$

Approximation (42) is employed for inequality (40) in the computation of equation (3). Moreover, since α *and* λ *do not depend on* m, *the following approximation is used:*

$$|g_{1,1}^m| \approx |g_{1,1}^{j-1}|, \quad |g_{1,j-1}^m| \approx |g_{1,j-1}^{j-1}| \qquad (2 \le j \le m).$$

In summary, Algorithm 4.1 is proposed, where $(f_m)_j$ is the jth element of f_m. For the computation of equation (3), the second line is replaced by $tol_1^{\mathrm{sys}} = \mathrm{tol}_\phi / [m^{\max}\beta \|(\gamma_1 I - A)y(0)\|]$. The linear equation in the sixth 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 [25] or the GMRES [22], are viable options. $(H_m D_m - I)H_m^{-1}$ in the twelfth line is a small matrix, and it can be computed via a direct method inexpensively. After computing $(H_m D_m - I)H_m^{-1}$, $\phi_k((H_m D_m - I)H_m^{-1})$ is also computed using a direct method, such as the scaling and squaring method [11].

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 this computation.

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 [22] with an ILU(0) preconditioner were applied to solve the linear equation in the sixth 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, a wave equation was implemented in region $(-1.5, 1.5) \times (-1, 1) \subseteq \mathbb{R}^2$:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \Delta u = f(x,t) & \text{in } (0,T] \times \Omega, \\ u = e^{-10(x_1-0.5)^2-10(x_2-0.5)^2} & \text{on } \{0\} \times \Omega, \\ u = 0 & \text{on } (0,T] \times \partial\Omega_1, \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } (0,T] \times \partial\Omega_2, \end{cases} \tag{43}$$

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

$$\begin{cases} \tilde{M}\ddot{\tilde{y}}(t) = \tilde{L}\tilde{y}(t) + \tilde{b}(t), \\ \tilde{y}(0) = \tilde{v}. \end{cases} \tag{44}$$

Equations (44) 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 $n = 237378$. The 1-step exponential integrator [16] whose scheme was:

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

was used.

In order to treat $M^{-1}L_{i+1}$ instead of $\Delta t M^{-1} L_{i+1}$, $\gamma_j/\Delta t$ was used instead of γ_j. Table 1 shows the CPU time and the iteration numbers 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} with SIA, RK, and SIRK. Figure 3 shows the relative error of each algorithm. The shifts introduced by Göckler [7] for SIA and RK were used. For SIA it was shown that choosing $\gamma = m^\alpha$ where $\alpha = (r-2)/(r+2)$ for step m resulted in a convergence rate of $O(m^{-k(1+\alpha)/2})$. Thus, $r \to \infty$ provides a

Table 1: Example 1: Comparison of SIA, RK, and SIRK.

Algorithm	γ_j	CPU time(s)	Iterations
SIRK	$(50-j)/\Delta t$	7.1	25
RK	$r = 20/\Delta t,\ h = 1.5/\Delta t$	–	–
RK	$r = 1.0/\Delta t,\ h = 0.1/\Delta t$	–	–
SIA	$20^{2/3}/\Delta t$	85.6	67
SIA	$20^{12/13}/\Delta t$	19.0	34
SIA	$30^{2/3}/\Delta t$	54.2	52
SIA	$30^{12/13}/\Delta t$	10.6	29
SIA	$40^{2/3}/\Delta t$	35.3	44
SIA	$40^{12/13}/\Delta t$	7.6	26

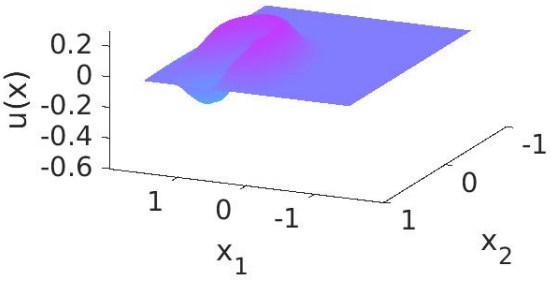

Figure 1: Example 1: Numerical solution of $t = 1$ with EI and SIRK.

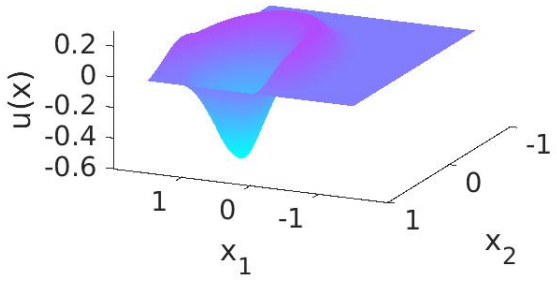

Figure 2: Example 1: Numerical solution of $t = 2$ with IE and SIRK.

convergence rate of $O(m^{-1})$. However, Göckler also showed that the constant in the rate $O(m^{-k(1+\alpha)/2})$ became larger as r grew larger. For this reason, both $r = 10$ and $r = 50$ were tested. Concerning m, $m = 20, 30, 40$ were tested. For RK, Göckler proposed setting $\gamma_j = r + h \cdot (-1)^{j-1} \lceil (j-1)/2 \rceil i$ at the jth step. $r = 20$, $h = 1.5$, $r = 1.0$ and $h = 0.1$, were tested. For SIRK, $m^{\max} = 50$ was set. In this example, SIA converged quickly for the large m and large r, but it converged slowly for the small r and small m. RK uses complex shifts even though the matrices and vector M, L and v are real. Thus, additional computational costs become necessary with complex values. Moreover, it does not converge in this case. On the other hand, SIRK uses real shifts, so its computation is faster, and the shifts in SIRK are determined automatically. Figure 1 and 2 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 2

The next problem is the convection diffusion equation in region $\Omega = ((-1.5, 1.5) \times$

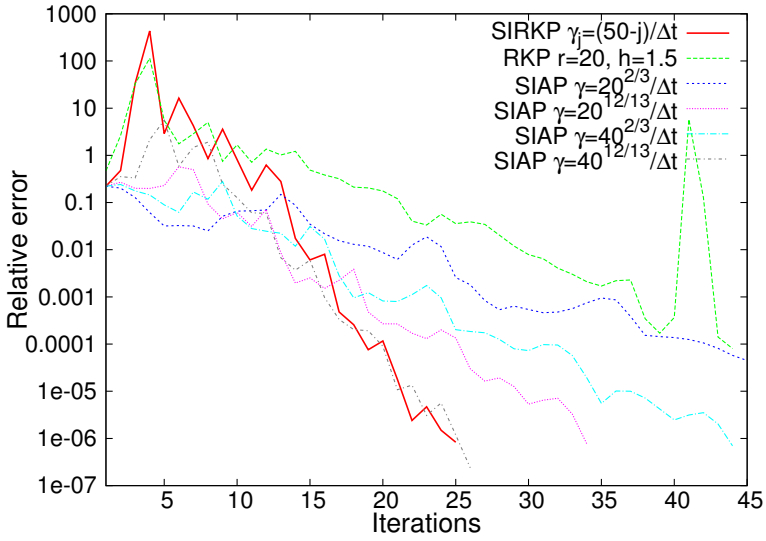

Figure 3: Example 1: The relative error of SIA, RK, and SIRK.

$(-1,1)) \setminus ([-0.5,0.5] \times [-0.25,0.25]) \subseteq \mathbb{R}^2$:

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

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 $\lambda = 0.025$. The ISIA, which were previously proposed by Hashimoto and Nodera [10], and ISIRK were compared. 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, ISIA, SIRK, and ISIRK. The CPU times and iteration numbers were compared. The detailed results are shown in Table 2. The residual tolerance for computing $\phi_0(tM^{-1}L)(v + L^{-1}c)$, tol_ϕ was 10^{-6}, and $t = 270$. $m^{\max} = 100$. In addition, $\delta = 0.01$ for the ISIA and ISIRK. Concerning the shift in the SIA and ISIA, $10/\Delta t$, $80/\Delta t$, and $160/\Delta t$. The results show that the ISIA and ISIRK are efficient. It should be noted that, ISIA does not converge or converges slowly depending on the choice of the shift. On the other hand, ISIRK does not require choosing shifts, and converges in a reasonable amount of time. Figure 4 shows the residual tolerance for solving linear equations at each Krylov step of the ISIRK. It was observed that the exactness needed to obtain a solution for the linear equation decreased as m became larger. The solutions computed with the ISIRK are tabulated in Figure 5. 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.

Example 3

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

Table 2: Example 2: Comparison of SIA, RK, and SIRK.

Algorithm	γ_j	CPU time(s)	Iterations
ISIRK	$(100-j)/t$	67.5	60
SIRK	$(100-j)/t$	159.8	60
ISIA	$640/t$	–	–
SIA	$640/t$	–	–
ISIA	$80/t$	58.0	59
SIA	$80/t$	138.2	61
ISIA	$10/t$	300.0	84
SIA	$10/t$	917.0	89

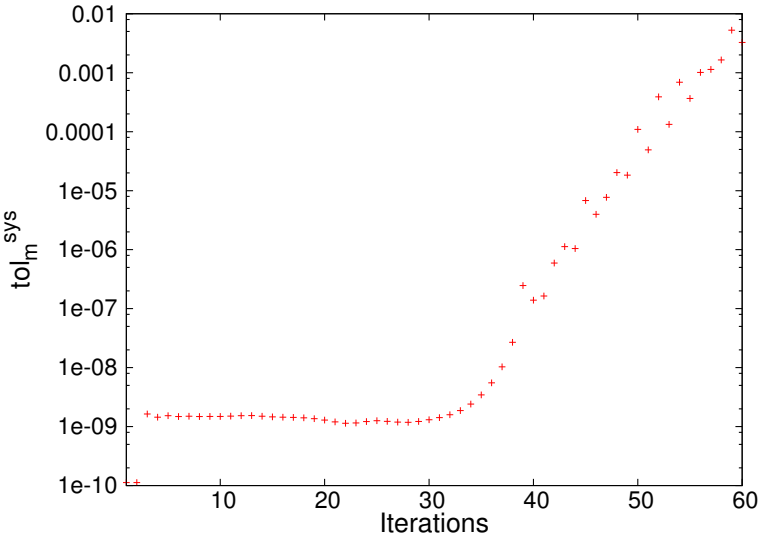

Figure 4: Example 2: Iterations versus $\mathrm{tol}_m^{\mathrm{sys}}$.

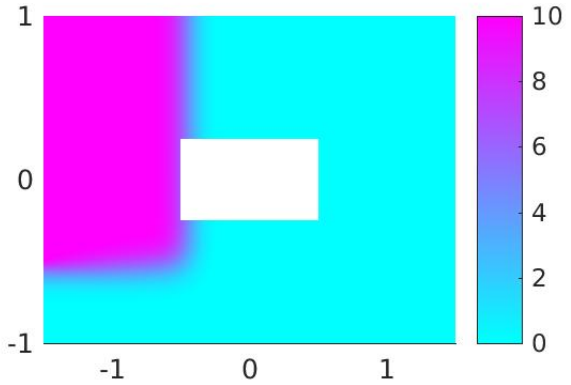

Figure 5: Example 2: Numerical solutions of ISIRK.

for confirming the effectiveness of ISIRK:

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

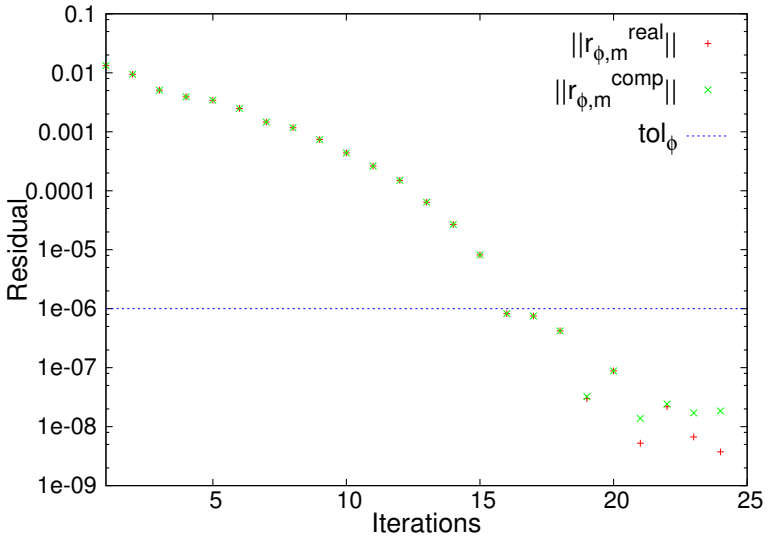

Figure 6: Example 3: Iterations versus $\|r_{\phi,m}^{\mathrm{real}}\|$ and $\|r_{\phi,m}^{\mathrm{comp}}\|$ with $\mathrm{tol}_\phi = 10^{-6}$.

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 [16]. The 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 $\phi_2(\Delta t M^{-1}L_{i+1})[n_i(y_i)-n_i(y_{i-1})]$ was 10^{-6} or 10^{-8} and $m^{\max} = 50$, $\delta = 0.01$. Figure 6 and Figure 7 show the relationship between the number of iterations and the residuals of ISIRK with $n = 118689$. The real residual $r_{\phi,m}^{\mathrm{real}}$ 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 $r_{\phi,m}^{\mathrm{comp}}$ decreased even after it had reached tol_ϕ. Moreover, the behavior of $r_{\phi,m}^{\mathrm{real}}$ and $r_{\phi,m}^{\mathrm{comp}}$ were the same before they reached tol_ϕ. Thus, $r_{\phi,m}^{\mathrm{comp}}$ was an appropriate stopping criterion for ISIRK. Figure 8 shows the residual tolerance for solving linear equations at each rational Krylov step with $\mathrm{tol}_\phi = 10^{-6}$ and $n = 118689$. It was observed that the exactness needed to obtain a solution for the linear equation decreased as m became larger. Table 3 shows the CPU time of ISIRK and SIRK with $\mathrm{tol}_\phi = 10^{-6}$. ISIRK was faster than the SIRK.

In view of Example 1, 2 and 3, the following observation were made: SIRK was more effective than RK and SIA. However, solving linear equations in SIRK inexactly with ISIRK was more effective and efficient.

6 Conclusion

The uses of SIRK and ISIRK were explored in this paper. The advantage of SIRK is that it determines the shifts of real values automatically. These shifts enable the faster

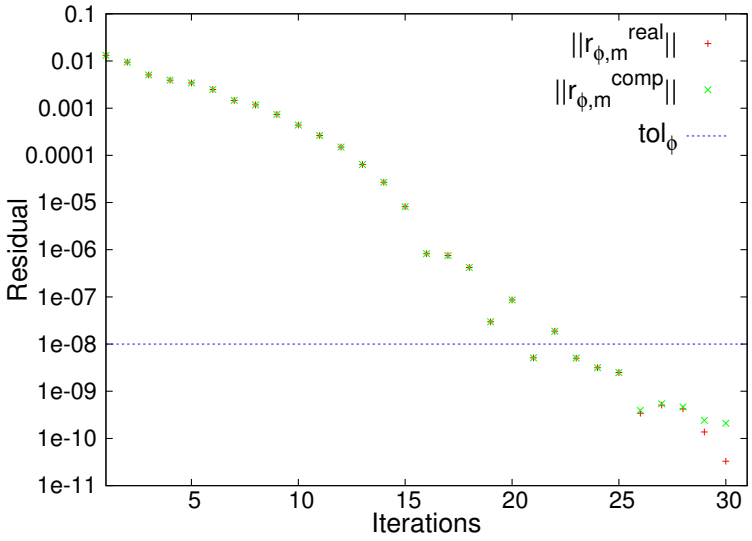

Figure 7: Example 3: Iterations versus $\|r_{\phi,m}^{\mathrm{real}}\|$ and $\|r_{\phi,m}^{\mathrm{comp}}\|$ with $\mathrm{tol}_\phi = 10^{-8}$.

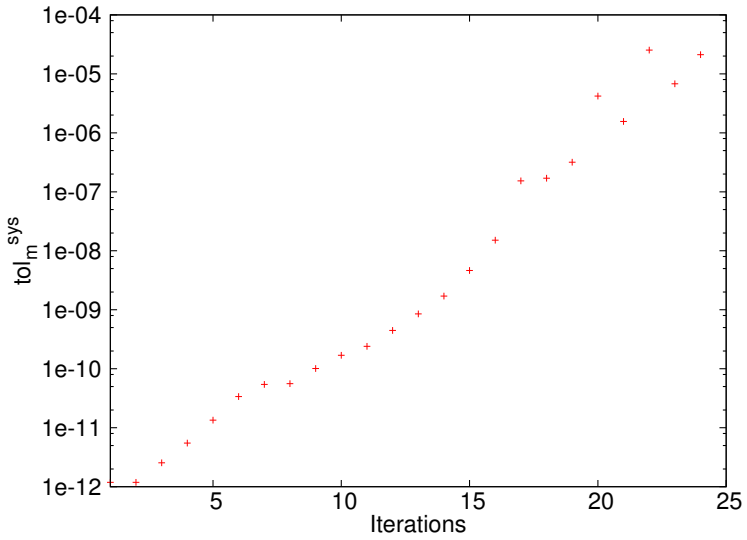

Figure 8: Example 3: Iterations versus $\mathrm{tol}_m^{\mathrm{sys}}$.

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

Algorithm	n	CPU time(s)	Iterations
ISIRK	29649	0.42	14
SIRK	29649	0.59	14
ISIRK	118689	2.4	16
SIRK	118689	3.6	16

convergence of SIRK. In addition, SIRK uses matrices appearing in every step of the Krylov process. This makes SIRK the effective method for computing ϕ-functions in the exponential integrator. Further to this, the computational 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 the convergence 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.

[2] Benzi, M. and Boito, P., Decay Properties for Functions of Matrices over C^*-algebras, *Linear Algebla and its Appricatations*, 456(1):174–198, 2014.

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

[4] Crouzeix, M., Numerical Range and Functional Calculus in Hilbert Space, *Journal of Functional Analysis*, 244:668–690, 2007.

[5] 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.

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

[7] Göckler, T., *Rational Krylov Subspace Methods for 'ϕ'-functions in Exponential Integrators*, PhD thesis, Karlsruher Instituts für Technologie, 2014.

[8] Güttel, S., *Rational Krylov Methods for Operator Functions*, PhD thesis, Technischen Universität Bergakademie Freiberg, 2010.

[9] Güttel, S., Rational Krylov Approximation of Matrix Functions: Numerical Methods and Optimal Pole Selection, *GAMM-Mitteilungen*, 38(1):8–31, 2013.

[10] Hashimoto, Y. and Nodera, T., Inexact Shift-invert Arnoldi Method for Evolution Equations, *ANZIAM Journal*, 58(E):E1–E27, 2016.

[11] 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.

[12] Hochbruck, M., A Short Course on Exponential Integrators, *Series in Contemporary Applied Mathematics*, 17:29–49, 2015.

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

[14] 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.

[15] Hochbruck, M. and Ostermann, A., Exponential Runge-Kutta Methods for Parabolic Problems, *Applied Numerical Mathematics*, 53(2–4):323–339, 2005.

[16] Hochbruck, M. and Ostermann, A., Exponential Integrators, *Acta Numerica*, 19:209–286, 2010.

[17] 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.

[18] Lee, S., Pang, H., and Sun, H., Shift-invert Arnoldi Approximation to the Toeplitz Matrix Exponential, *SIAM Journal on Scientific Computing*, 32(2):774–792, 2010.

[19] 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.

[20] Moret, I. and Novati, P., RD-Rational Approximations of the Matrix Exponential, *BIT Numerical Mathematics*, 44(3):595–615, 2004.

[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.

[22] 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.

[23] 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.

[24] Svoboda, Z., The Convective-diffusion Equation and Its Use in Building Physics, *International Journal on Architectural Science*, 1(2):68–79, 2000.

[25] Van der Vorst, 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.

KSTS/RR-17/001
January 27, 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