SIAM J. NUMER. ANAL.
Vol. 59, No. 1, pp. 370–400

ARNOLDI ALGORITHMS WITH STRUCTURED ORTHOGONALIZATION*

PENGWEN CHEN†, CHUNG-KUAN CHENG‡, AND XINYUAN WANG§

Abstract. We study a stability preserved Arnoldi algorithm for matrix exponential in the time domain simulation of large-scale power delivery networks (PDNs), which are formulated as semi-explicit differential algebraic equations (DAEs). The solution can be decomposed to a sum of two projections, one in the range of the system operator and the other in its null space. The range projection can be computed with a shift-and-invert Krylov subspace method. The other projection can be computed with the algebraic equations. Differing from the ordinary Arnoldi method, the orthogonality in the Krylov subspace is replaced with the semi-inner product induced by the positive semidefinite system operator. With proper adjustment, numerical ranges of the Krylov operator lie in the right half-plane, and we obtain theoretical convergence analysis for the modified Arnoldi algorithm in computing phi-functions. Last, simulations on RLC networks are demonstrated to validate the effectiveness of the Arnoldi algorithm.

Key words. differential algebraic equations, Arnoldi algorithms, Krylov subspace methods, shift-and-invert Krylov subspace methods

AMS subject classifications. 68Q25, 68R10, 68U05

DOI. 10.1137/20M1336667

1. Introduction.

VLSI design verification relies heavily on the analysis of power delivery networks (PDNs) to estimate power supply noises [28, 43, 17, 32, 31, 67, 68, 58]. The performance of PDNs highly impacts on the quality of global, detailed, and mixed-size placement [33, 46, 34, 35], clock tree synthesis [63], and global and detailed routing [69], as well as timing [27] and power optimization. The challenges of power integrity analysis arise from the tighter noise margin with reducing power supply voltage, higher resistance on metal wires due to scaling, and strong coupling noise between the active devices.

Advanced technologies [74, 65], three-dimensional (3D) IC structures [70, 51], and increasing complexities of system designs all make VLSI PDNs extremely huge and the simulation tasks time-consuming and computationally challenging. Due to the enormous size of modern designs and long simulation runtime of many cycles, instead of general nonlinear circuit simulation [40, 41, 5], a PDN is often modeled as a large-scale linear circuit with voltage supplies and time-varying current sources [42, 30]. Those linear matrices are obtained by parasitic extraction process [73, 71, 66, 64, 32]. After those processes, we need time-domain large-scale linear circuit simulation to obtain the transient behavior of PDN with above inputs. The widely accepted backward Euler (BE) and trapezoidal (TR) usually serve as the baseline in traditional

*Received by the editors May 7, 2020; accepted for publication (in revised form) November 30, 2020; published electronically February 9, 2021.
https://doi.org/10.1137/20M1336667
Funding: The research of the first author was supported by grant 107-2115-M-005-006-MY3 from the Ministry of Science and Technology, Taiwan. The research of the second and third authors was supported by National Science Foundation grant CCF-1564302.
†Department of Applied Mathematics, National Chung Hsing University, Taichung City 402, Taiwan (pengwen@nchu.edu.tw).
‡Department of Computer Science and Engineering, and Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA 92093 USA (ckcheng@ucsd.edu).
§Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA 92093 USA (xiw193@eng.ucsd.edu).

linear multistep integration methods since they are proved to be *A-stable* [5, 40, 41]. However, solving a linear system is required at each time step and the performance of the implicit integration methods is impacted by the local truncation error (LTE).

A matrix exponential based integration method for PDN transient simulation is considered [4, 72]. Compared to the traditional linear multistep methods, the matrix exponential based method is not bounded by the Dahlquist stability barrier, and thus the step size breaks the limitation of the LTE [60, 72]. It has been explored with the efficient evaluation of matrix exponential and vector product (MEVP) via a Krylov subspace method, which is considered as a high order polynomial approximation [38, 50]. The stability of a matrix exponential based method when applied to ODEs has been well established in previous work [61, 72]. For general circuit simulation with differential algebraic equations (DAEs), the stability remains an interesting topic [11, 24, 62, 55]. Numerical stability issues are reported in [4, 59] and reveal the limitation of MEVP computations with Krylov subspace. A similar problem occurs in the eigenvalue problems [45] and model order reduction for interconnect simulation [48], where Krylov subspace methods are widely used.

Let $x(t) \in \mathbb{R}^N$ be the solution to a system of stiff differential equations [4]

$$\frac{dq(t)}{dt} + f(x(t)) = u(t), \quad x(0) = x_0, \tag{1.1}$$

where $u(t)$ is the input signal to the circuit system, $x(t) \in \mathbb{R}^N$ of large dimension N denotes nodal voltages, and branch currents at time t and q, $f \in \mathbb{R}^N$ are the charge (or flux) and current (or voltage) terms, respectively. The system is governed by Kirchhoff's current law and voltage law. With linearization, we have

$$C\frac{dx}{dt} + Gx = u(t), \quad x(0) = x_0, \tag{1.2}$$

where C and G both are real $N \times N$ matrices, which are the Jacobian matrices of q and f with respect to x, respectively.

In the study, we assume that C, G are constant matrices and

$$\begin{cases} G \text{ is positive definite but not necessarily symmetric;} \\ C \text{ is positive semidefinite and symmetric, } C \neq 0. \end{cases} \tag{1.3}$$

That is, $x^\top Gx > 0$ holds for nonzero vectors $x \in \mathbb{R}^N$. For PDNs, capacitances, inductances, and conductances have positive values. Through modified nodal analysis, matrix C is symmetric and positive semidefinite and matrix G is positive semidefinite. See [41, 42]. Every node is supposed to connect to power or ground via a path of resistors, which makes G nonsingular. For a stiff system, the solution can be of multiple timescales; i.e., the attractive solution is surrounded with fast-changing nearby solutions.

When C is nonsingular, the solution can be formulated as exponentials of the matrix $A := C^{-1}G$. There are various ways to compute the matrix exponentials (see, e.g., [37, 38]), depending on the state companion matrix A. When A is a matrix of small size, the most effective algorithm is a scaling-and-squaring method based on Padé approximation [56]. When A is sparse and large, one general and well-established technique is approximating the action of the matrix exponentials in the class of Krylov subspaces. One essential ingredient is the evaluation or approximation of the product of the exponential of the Jacobian A with a vector v. The application

of Krylov subspace techniques has been actively investigated in the literatures; see, e.g., [12, 50, 38, 22, 44, 25]. In general, the nonlinear form in (1.1) can be numerically handled by various exponential Runge–Kutta schemes with the aid of exponential integrators; see, e.g., [21] and references therein.

It is well known that Krylov subspace methods for matrix functions exhibit superlinear convergence behavior under sufficient large Krylov dimension [50, 20, 21]. Analogous to the inverse power method in computing eigenvectors corresponding to eigenvalues near a shift, the rational Krylov subspace method has the superiority over the standard Krylov subspace methods when the spectrum of the operator lies in the half-plane, e.g., the Laplacian operators in PDEs [8, 16, 57]. The convergence of computing exponential integrators of evolution equations in the resolvent Krylov subspace is independent of the operator norm of A from one numerical discretization when A in $\exp(-A)$ has numerical range (or called field of values) in the right half-plane [14, 15].

In this paper, we shall examine the modified shift-and-invert Arnoldi algorithm from the perspective of numerical ranges, which provides the theoretical foundation for the Arnoldi algorithm described in [4, 59]. Let x^* be the conjugate transpose of a vector x. Since the matrix C could be singular in PDN transient simulation, we introduce C semi-inner product as well as its induced norm,

$$\langle x, y\rangle_C := x^*Cy, \quad \|x\|_C := \sqrt{x^*Cx}, \tag{1.4}$$

to derive the error analysis, instead of the ordinary inner product $\langle x, y\rangle := x^*y$. Likewise, the C-norm $\|x\|_C := \sqrt{x^*Cx}$ is used to define the so-called C-numerical ranges in (2.1). The advantage of C semi-inner product introduced in the modified Arnoldi algorithm is two-fold: the null-space component is removed in the Arnoldi iterations, and the C-numerical range of the operator in the matrix exponentials lies in the right half-plane. The numerical range of the upper Hessenberg matrix is properly restricted within a disk with center at $1/2$ and radius $1/2$. The C semi-inner product has been employed in the Arnoldi algorithms, e.g, solving generalized eigenvector problems [9, 36] and generating stable and passive Arnoldi based model order reduction [52].

The main contributions are listed as follows. With the aid of eigenvectors of C as a basis, solutions $x(t)$ to PDNs can be expressed as a sum of $x_{\mathcal{R}}(t)$ and $x_{\mathcal{N}}(t)$. The shift-and-invert Krylov method in [59] actually computes $x_{\mathcal{R}}(t)$, which captures the dominant transient dynamical behaviors. The positive definite matrix G guarantees the C-numerical range of $G^{-1}C$ lying the right half-plane. With the C semi-inner product, the orthonormal basis of Krylov subspace preserves the passivity property of the system, which yields stable transient simulations. The shift parameter γ in the shift-and-invert method provides the flexibility to confine the spectrum of ill-conditioned systems [10]. The error with φ_k-functions tends to 0 as the dimension increases. In this paper, we employ Theorem 5.7 of [13] and the reasoning in [20, 39] to estimate the error bounds and establish the convergence to $x_{\mathcal{R}}(t)$.

The rest of this paper is organized as follows. The DAE framework is introduced in section 1.1. The explicit formulations of solutions in the basis of eigenvectors of $G^{-1}C$ and in the basis of eigenvectors of C are given in sections 1.2 and 1.3, respectively. In the paper, we focus on the computation of the projected solution $x_{\mathcal{R}}(t)$ in (1.29). In section 1.4, we introduce Krylov space corresponding to the shift-and invert method to generate one approximation in (1.43). In section 2, we give error estimations based on C-numerical ranges and establish convergence results. In section 3, we provide simulations on RLC networks with G only positive semidefinite to validate the effectiveness of the modified shift-and-invert Arnoldi algorithm. Results indicate

that stable simulations can be obtained under proper numerical pruning of spurious eigenvalues.

1.1. Solutions of nonsingular systems. Suppose that C is nonsingular with $A = C^{-1}G$. The variation-of-constants formula yields the solution $x(t)$ described by

$$x(t) = \exp(-tA)x_0 + \int_0^t \exp(-(t-s)A)C^{-1}u(s)\,ds. \tag{1.5}$$

Introducing so-called phi-functions,

$$\varphi_0(z) := \exp(z), \quad \varphi_{k+1}(z) := (\varphi_k(z) - (k!)^{-1})/z \quad \text{for } k \geq 0\ , \tag{1.6}$$

we can approximate (1.5) under linearization on the source term $C^{-1}u(s) \approx b + b's$ as a sum of the φ_0, φ_1, and φ_2 terms:

$$x(t+h) \approx \varphi_0(-hA)x(t) + h\varphi_1(-hA)b + h^2\varphi_2(-hA)b', \tag{1.7}$$

where $\varphi_0(z) = \exp(z)$ and $\varphi_1(z) = z^{-1}(\exp(z) - 1)$. One can employ the shift-and-invert Arnoldi transform to solve a nonsingular differential system as in [3]. Briefly, let $A = C^{-1}G$ and construct the Krylov subspace with respect to $(I + \gamma A)^{-1}$ with a parameter $\gamma > 0$, i.e.,

$$(I + \gamma A^{-1})^{-1}V_m = V_{m+1}\widetilde{H}_m,$$

where an orthogonal basis matrix $V_m \in \mathbb{R}^{N\times m}$ and an upper-Hessenburg matrix $\widetilde{H}_m \in \mathbb{R}^{(m+1)\times m}$ are generated. Let H_m be the submatrix of $\widetilde{H}_m$ without the last row. Then the terms φ_0, φ_1 in (1.7) can be approximated by the exponential function of H_m, e.g.,

$$\exp(-tA)x_0 \approx \|x_0\| \exp(-\gamma t(H_m^{-1} - I_m))e_1. \tag{1.8}$$

1.2. Solutions of singular systems. A nonsingular matrix C cannot always be achieved in general PDNs. For instance, the nodes without nodal capacitance or inductance would contribute to the algebraic equations and the corresponding matrix C would not be invertible. One major impact from the singularity is that the system in (1.2) is in fact one combination of differential equations and algebraic equations; i.e., $x(t)$ must satisfy the following range condition:

$$x(t) - G^{-1}u(t) \text{ in the range of } G^{-1}C. \tag{1.9}$$

In addition, since the projection H_m is constructed from an initial vector, without careful and proper handling, the matrix H_m could become a nearly degenerate matrix, and (1.8) boils down to be an erroneous approximation. Hence, it is natural to perform some proper decomposition on $x(t+h)$ based on nonzero and zero eigenvalues, so that H_m is not contaminated by null vectors and the solution $x(t)$ can be computed accurately.

We discuss two decompositions to express the solutions. Start with the standard approach in differential equations. (This approach is listed as Method 16 in [38].) Let $G^{-1}C = V\Lambda V^{-1}$ be the Jordan canonical form decomposition of $G^{-1}C$, where

$$\Lambda = \begin{pmatrix} J_{\mathcal{R}} & 0 \\ 0 & J_{\mathcal{Z}} \end{pmatrix} \in \mathbb{C}^{N\times N}$$

is in Jordan normal form. The submatrix $J_{\mathcal{R}} \in \mathbb{C}^{r\times r}$ consists of a few Jordan blocks corresponding to nonzero eigenvalues of $G^{-1}C$, and $J_{\mathcal{Z}} \in \mathbb{R}^{(N-r)\times(N-r)}$ is a nilpotent

matrix corresponding to eigenvalue zero of $G^{-1}C$. Since the null space of $G^{-1}C$ has dimension $N-n$, the algebraic multiplicity of the eigenvalue zero is not less than $N-n$. Write $V=[V_{\mathcal{R}},V_{\mathcal{Z}}]$, $V_{\mathcal{Z}}:=[V_{\mathcal{G}},V_{\mathcal{N}}]$, where columns of $V_{\mathcal{R}}$ and $V_{\mathcal{Z}}$ are the (generalized) eigenvectors of nonzero eigenvalues, respectively. Columns of $V_{\mathcal{G}}$ and $V_{\mathcal{N}}$ are the generalized eigenvectors and the eigenvectors of eigenvalue 0. That is, columns of $V_{\mathcal{N}}$ are the null vectors of $G^{-1}C$. Let $U:=(V^{-1})^*=[U_{\mathcal{R}},U_{\mathcal{Z}}]$, where A^* is the Hermitian transpose of a matrix A. Consider the solution decomposition

$$x(t)=x_{\mathcal{R}}(t)+x_{\mathcal{Z}}(t)=V_{\mathcal{R}}x_1(t)+V_{\mathcal{Z}}x_2(t) \tag{1.10}$$

with some vector functions $x_1(t),x_2(t)$. Let

$$U^*G^{-1}CV=\begin{pmatrix} J_{\mathcal{R}} & 0\\ 0 & J_{\mathcal{Z}}\end{pmatrix}. \tag{1.11}$$

Multiplying with U^*G^{-1} on (1.2) yields one differential equation for x_1:

$$J_{\mathcal{R}}\frac{dx_1}{dt}+x_1=U_{\mathcal{R}}^*G^{-1}u(t) \tag{1.12}$$

and

$$J_{\mathcal{Z}}\frac{dx_2}{dt}+x_2=U_{\mathcal{Z}}^*G^{-1}u(t). \tag{1.13}$$

Focus on (1.13) first. For simplicity, assume that $G^{-1}u(t)$ is a linear function in t, i.e., for some constant vectors w_0,w_1,

$$U_{\mathcal{Z}}^*G^{-1}u(t)=w_0+w_1t.$$

The solution $x_2(t)$ is also linear and can be expressed as

$$x_{\mathcal{Z}}(t)=V_{\mathcal{Z}}x_2(t)=V_{\mathcal{Z}}(w_1t+w_0-J_{\mathcal{Z}}w_1)=V_{\mathcal{Z}}\left(U_{\mathcal{Z}}^*G^{-1}u(t)-J_{\mathcal{Z}}U_{\mathcal{Z}}^*G^{-1}\frac{du(t)}{dt}\right).$$

Return to (1.12). Let $\widetilde{u}(t)=J_{\mathcal{R}}^{-1}U_{\mathcal{R}}^*G^{-1}u(t)$. The solution $x_1(t)$ in (1.12) can be expressed as

$$x_{\mathcal{R}}(t):=V_{\mathcal{R}}x_1(t)=V_{\mathcal{R}}\left\{\exp(-tJ_{\mathcal{R}}^{-1})U_{\mathcal{R}}^*x(0)+\exp(-tJ_{\mathcal{R}}^{-1})\int_0^t\exp(sJ_{\mathcal{R}}^{-1})\widetilde{u}(s)\,ds\right\}. \tag{1.14}$$

1.3. Solutions under eigenvectors of C. The matrices $V_{\mathcal{R}},U_{\mathcal{R}},J_{\mathcal{R}}$ in (1.14) are generally complex-valued, which makes the computation for large PDN systems very challenging. Next, we introduce one set of *real* basis vectors to express the solution in (1.2), the eigenvectors of C. Let $C=V_CC_1V_C^\top$ be the eigenvector decomposition of C, where columns of $V_C\in\mathbb{R}^{N\times n}$ are unit eigenvectors and $C_1\in\mathbb{R}^{n\times n}$ is diagonal and nonsingular. Let $P_C=V_CV_C^\top$ be the orthogonal projection matrix on the range of C. Also introduce orthogonal subspaces $\mathcal{R}$ and $\mathcal{N}$,

$$\mathcal{R}:=\{P_Cx:x\in\mathbb{R}^N\}, \tag{1.15}$$

$$\mathcal{N}:=\{x\in\mathbb{R}^N:P_Cx=0\}. \tag{1.16}$$

We employ

$$V:=[V_{\mathcal{R}},V_{\mathcal{N}}],\quad V_{\mathcal{R}}=V_C,U:=[U_{\mathcal{R}},U_{\mathcal{N}}]=(V^{-1})^\top, \tag{1.17}$$

to decouple the system in (1.2), where columns of $V_{\mathcal{R}} \in \mathbb{R}^{N\times n}, V_{\mathcal{N}} \in \mathbb{R}^{N\times(N-n)}$ are basis vectors in $\mathcal{R}$ and $\mathcal{N}$, respectively. Hence, $U = V$. Since $C \neq 0$, then $n \geq 1$. Write G, C in block forms,

$$V^\top G V = \begin{pmatrix} G_1 & G_2 \\ G_3 & G_4 \end{pmatrix}, \quad V^\top C V = \begin{pmatrix} C_1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{1.18}$$

where $C_1 \in \mathbb{R}^{n\times n}$ is a nonsingular, positive definite, and symmetric submatrix. Consider the following solution decomposition:

$$x(t) = x_{\mathcal{R}}(t) + x_{\mathcal{N}}(t) = V_C x_1(t) + V_{\mathcal{N}} x_2(t) \tag{1.19}$$

with some vector functions $x_1(t), x_2(t)$. Applying G^{-1} on (1.2) yields *one range consistency constraint on* $x(t)$, that $x(t) - G^{-1}u(t)$ must lie in the range of $G^{-1}C$, including the initial vector $x(0)$. Actually, from (1.18), the system in (1.2) is a combination of one differential system and one algebraic system, i.e.,

$$C_1 \frac{dx_1}{dt} = -G_1 x_1 - G_2 x_2 + u_1, \tag{1.20}$$

$$G_3 x_1 + G_4 x_2 = u_2. \tag{1.21}$$

Suppose that G_4 is invertible. With (1.21), we can eliminate x_2 in (1.20) and reach one *nonsingular* differential system of x_1, i.e.,

$$C_1 \frac{dx_1}{dt} = -(G_1 - G_2 G_4^{-1} G_3)x_1 - G_2 G_4^{-1} u_2 + u_1. \tag{1.22}$$

Equivalently, using the inverse formula of a 2×2 block matrix, we have

$$(G^{-1})_{1,1} C_1 \frac{dx_1}{dt} + x_1 = -(G^{-1})_{1,1}(G_2 G_4^{-1} u_2 + u_1). \tag{1.23}$$

Such a system of DAEs can also occur in the simulation of mechanical multibody systems; see, e.g., [53]. Finally, we can determine $x_{\mathcal{N}}$, i.e., $x_2(t)$ from (1.21), if G_4 is invertible. Hereafter we shall focus on the computation of $x_1(t)$. Keep in mind that the block form in (1.18) is only of theoretical interest, since the explicit formulation requires the information of eigenvectors of C. In practical applications of large dimension, $G_1, \ldots, G_4$ and C_1 are unlikely to be known in advance.

REMARK 1.1. *As pointed out by the referees, the seminorm* $\|\cdot\|_C$ *actually introduces a new Hilbert space* $\mathcal{V}$ *for the solution to* (1.2) *when* C *is singular. In abstract terms, this Hilbert space* $\mathcal{V}$ *can be regarded as one quotient space, i.e.,* $\mathbb{R}^N$ *(or* $\mathbb{C}^N$*) modulo the kernel space of the operator* C *and* $x_{\mathcal{R}}(t)$ *can be identified as one equivalence class of* x*. Indeed, multiplying* (1.2) *with* $P_C G^{-1}$ *yields one differential equation for an equivalent class,*

$$P_C G^{-1} C P_C \frac{dx}{dt} + P_C x = P_C G^{-1} u(t).$$

Once $x_{\mathcal{R}}(t) = P_C x(t)$ *is determined, the solution* x *in the equivalent class can be determined by the condition in* (1.9)*, i.e., the algebraic equation in* (1.21)*.*

In section 1.4, we shall derive a Krylov subspace approximation of $x_1(t)$ based on the invertibility of $V_C^\top G^{-1} C V_C$. Here we introduce some results regarding the eigenvalues of $B_{1,1} := V_C^\top G^{-1} C V_C$. Importantly, we provide one sufficient condition, the *positive definite* property on G, to ensure the invertibility of $B_{1,1} = V_C^\top G^{-1} C V_C$.

PROPOSITION 1.2. *Assume that C, G satisfy* (1.3). *Let columns of V_C be eigenvectors corresponding to nonzero eigenvalues of C. Let*

$$B = G^{-1}C,\ B_{1,1} = V_C^\top B V_C. \tag{1.24}$$

- *The matrix $B_{1,1}$ is invertible. Let $u_1 \in \mathbb{C}^n$ be an eigenvector of $B_{1,1}$ corresponding to a nonzero eigenvalue λ. Let*

$$u_2 = \lambda^{-1}(V_{\mathcal{N}}^\top G^{-1} C V_C u_1) \in \mathbb{C}^{N-n}.$$

  Then $V_C u_1 + V_{\mathcal{N}} u_2$ is an eigenvector of B corresponding to λ.
- *Let v be an eigenvector of B corresponding to eigenvalue λ. Then $\lambda = 0$ if and only if v is a null vector of C, i.e., $\|V_C^\top v\| = 0$. When $\lambda \neq 0$, then $\Re(\lambda) > 0$ and $V_C^\top v$ is an eigenvector of $B_{1,1}$ corresponding to the eigenvalue λ.*

Proof. We show the invertibility of G_4 first. Let v_2 be a null vector of G_4. Take $v = [0, v_2^\top]^\top \in \mathbb{R}^N$. The positive definite property of G yields $v^\top G v = v_2^\top G_4 v_2 = 0 \geq \epsilon \|v_2\|^2$ for some $\epsilon > 0$. Hence, we have $v_2 = 0$, i.e., the invertibility of G_4. Multiplying with $V_C^\top G^{-1}$ on (1.2) yields one differential equation for x_1:

$$B_{1,1}\frac{dx_1}{dt} + x_1 = V_C^\top G^{-1} V_C C_1 x_1 + x_1 = V_C^\top G^{-1} u. \tag{1.25}$$

Since $B_{1,1} = V_C^\top B V_C = (V_C^\top G^{-1} V_C) C_1$, according to the inverse formula of a 2×2 block matrix, the invertibility of G_4 implies the invertibility of $B_{1,1}$ and

$$B_{1,1}^{-1} = C_1^{-1}(V_C^\top G^{-1} V_C)^{-1} = C_1^{-1}(G_1 - G_2 G_4^{-1} G_3).$$

The remaining statements in the first part are obvious.

For the second part, the first statement is straightforward. For the second statement, since $v \in \mathbb{C}^N$ is one eigenvector of B corresponding to eigenvalue $\lambda \in \mathbb{C}$, i.e., $\lambda v = G^{-1} C v$, then

$$\lambda V_C^\top v = V_C^\top G^{-1} C v = V_C^\top G^{-1} V_C C_1 V_C^\top v = B_{1,1} V_C^\top v. \tag{1.26}$$

When v is not a null vector of C, then $V_C^\top v$ is an eigenvector of $B_{1,1}$. In addition, from (1.26),

$$\lambda v^* C v = \lambda (V_C^\top v)^* C_1 V_C^\top v = (V_C^\top v)^* C_1 V_C^\top G^{-1} C v = (Cv)^* G^{-1} C v.$$

The real part of both sides gives

$$\Re(\lambda) v^* C v = \Re(\lambda v^* C v) = \Re((Cv)^* G^{-1} C v) = \frac{1}{2}(\Re(G^{-1} C v)^\top (G^\top + G) \Re(G^{-1} C v)).$$

Hence, $\Re(\lambda) \geq 0$. ∎

PROPOSITION 1.3. *Assume that C, G satisfy* (1.3). *Let $V := [V_C, V_{\mathcal{N}}]$ in* (1.17). *Let $B_{1,1} := V_C^\top G^{-1} C V_C$. Let $\widetilde{u} := (B_{1,1})^{-1} V_C^\top G^{-1} u$. Then the projected solution $x_{\mathcal{R}}(t)$ is given by*

$$x_{\mathcal{R}}(t) := V_C x_1(t) = V_C \left\{ \exp(-t B_{1,1}^{-1}) V_C^\top x(0) + \exp(-t B_{1,1}^{-1}) \int_0^t \exp(s B_{1,1}^{-1}) \widetilde{u}(s)\, ds \right\}. \tag{1.27}$$

In addition, the projected solution $x_{\mathcal{N}}(t)$ is given by

$$x_{\mathcal{N}}(t) := V_{\mathcal{N}} x_2(t) = V_{\mathcal{N}}(U_{\mathcal{N}}^\top G V_{\mathcal{N}})^{-1} U_{\mathcal{N}}^\top (u(t) - G V_C x_1(t)). \tag{1.28}$$

Suppose $u(s)$ is linear, i.e., $u(s) = u(0) + su'(0)$ for some vectors $u(0), u'(0) = \frac{du}{ds}(0)$. Then

$$\begin{aligned} x_{\mathcal{R}}(t) = V_C\{&\exp(-tB_{1,1}^{-1})V_C^\top x(0) + tB_{1,1}^{-1}\varphi_1(-tB_{1,1}^{-1})V_C^\top G^{-1}u(0) \\ &+ t^2 B_{1,1}^{-1}\varphi_2(-tB_{1,1}^{-1})V_C^\top G^{-1}u'(0)\}. \end{aligned} \tag{1.29}$$

Proof. It is clear that the solution to (1.25) is given by (1.27). The second term in (1.27) can be further simplified, i.e.,

$$\begin{aligned} &\exp(-tB_{1,1}^{-1}) \int_0^t \exp(sB_{1,1}^{-1})\widetilde{u}(s)\, ds \\ &= (-B_{1,1})\{-I + \exp(-tB_{1,1}^{-1})\}\widetilde{u}(0) + B_{1,1}^2(-I + B_{1,1}^{-1}t + \exp(-tB_{1,1}^{-1}))\widetilde{u}'(0) \\ &= t\varphi_1(-tB_{1,1}^{-1})\widetilde{u}(0) + t^2\varphi_2(-tB_{1,1}^{-1})\widetilde{u}'(0). \end{aligned}$$

Thus, the projected solution $V_C V_C^\top x(t)$ is given by (1.29). ∎

REMARK 1.4. *What happens if G_4 is not invertible? This is one fundamental limitation of the decomposition described in section* 1.3: *when G_4 is not invertible, then $B_{1,1}$ has rank less than n and B can have generalized eigenvectors (in addition to null vectors) corresponding to the eigenvalue* 0. *Noninvertibility of G_4 leads to rank$(P_C G^{-1} C) <$ rank(C), i.e., $\mathcal{R} + \mathcal{N} \neq \mathbb{R}^N$. Actually, when G_4 is not invertible, i.e., $G_4 y = 0$ for some nonzero vector y, the algebraic multiplicity of the eigenvalue zero of $G^{-1}C$ is greater than its geometric multiplicity. The size of the Jordan normal form of $B = G^{-1}C$ corresponding to eigenvalue* 0 *could be* 2. *More discussions on this issue can be found in* [36, Theorem 1] *and* [9, Theorem 2.7]. *Further analysis on this issue is beyond the scope of the current paper.*

1.4. Krylov subspace approximation. Since $G^{-1}C$ is well-defined, it is intuitive to apply the shift-and-invert Arnoldi iterations to compute the requisite matrix exponentials in solving (1.2) with *singular* C. To compute $x_{\mathcal{R}}(t)$ from (1.27) or (1.29) for a large singular system in (1.2), we shall design one m-dimensional Arnoldi algorithm to construct a low-dimensional rational Krylov subspace approximation of the matrix exponential of $B_{1,1} := V_C^\top G^{-1} C V_C$.

Eigenvalues and eigenvectors of large matrices can be computed effectively by rational Krylov subspace methods [49]. One advantage lies in the fact that unlike polynomial approximants, rational best approximants of $\exp(-x)$ can converge geometrically in the domain $[0, \infty)$ [6]. It is known that matrix exponentials $\phi_k(-tA)$ acting on a vector v can be computed effectively by rational Krylov subspace methods when the numerical range of A is located somewhere in the right complex half-plane; see, e.g., [14].

For the circuit simulations, the numerical range of the matrix $B_{1,1}$ does not completely lie in the right half-plane. To overcome this difficulty, we introduce a new Arnoldi scheme with structured orthogonalization to generate one stable Krylov subspace and to compute matrix exponentials [59]. The orthogonality is based on the positive semidefinite matrix C, which plays a fundamental role in enforcing the numerical range of the operator in the right half-plane under the assumption in (1.3). Algorithm 1.1 can be regarded as one resolvent Krylov subspace method with C semi-inner product in [14].

1.4.1. Shift-and-invert methods.

REMARK 1.5. *Fix some parameter $\gamma > 0$. When all eigenvalues of A lie in the right complex half-plane, then $-1/\gamma$ lies in the resolvent set of A. We can employ the shift-and-invert method to approximate $\phi_k(-tA)v$ in the resolvent Krylov subspace,*

$$span\{v, (I+\gamma A)^{-1}v, \ldots, (I+\gamma A)^{-(m-1)}v\}.$$

As one reference, we list the result for the nonsingular case. Let $A = C^{-1}G$. From Proposition 1.2*, eigenvalues of A lie in the right complex half-plane. Use the standard Arnoldi iterations to construct (V_m, H_m) from*

$$(C+\gamma G)^{-1}CV_m = V_m H_m + h_{m+1,m}v_{m+1}e_m^\top,$$

where columns of V_m are a set of orthogonal vectors of m-dimensional Krylov subspace induced by $(C+\gamma G)^{-1}C$ and H_m satisfies

$$H_m = V_m^\top (C+\gamma G)^{-1}CV_m.$$

When $h_{m+1,m} = 0$, $(C+\gamma G)^{-1}C = V_m H_m V_m^\top$ motivates the approximation of the matrix exponential,

$$\exp(-tA)v \approx \|v\|V_k \exp(t(I - H_k^{-1})/\gamma)e_1 \quad \text{for } k < m.$$

DEFINITION 1.1. *To estimate the eigenstructure of $G^{-1}C$ subject to $\mathcal{R}$, we introduce a few matrices $S, S_{1,1}$ associated to B,*

$$S := P_C(C+\gamma G)^{-1}C, \quad \widetilde{S} := (C+\gamma G)^{-1}C, \tag{1.30}$$

$$S_{1,1} := V_C^\top S V_C = V_C^\top \widetilde{S} V_C, \quad \gamma > 0. \tag{1.31}$$

Let $\{w_1, w_2, \ldots, w_m\} \in \mathbb{R}^N$ be a set of C-orthogonal vectors spanning one Krylov subspace from S,

$$\mathcal{K}(S, w_1) := span\{w_1, Sw_1, S^2w_1, \ldots, S^{m-1}w_1\} = span\{w_1, w_2, \ldots, w_m\}.$$

Proposition 1.6 indicates that $W_m := [w_1, w_2, \ldots, w_m]$ is one low-dimensional subspace in the range of $P_C G^{-1}C$. Let H_m be one upper Hessenburg matrix corresponding to the projection of S on W_m. The algorithm to generate (W_m, H_m) is stated in Algorithm 1.1. Empirically we use the Arnoldi iterations in (1.33) to compute $\widetilde{W}_m$ and $\widetilde{H}_m$ instead. Proposition 1.6 suggests the computation of the approximate $x_a(t)$ in (1.43), where one single operation P_C is involved. Since W_m is the projection of $\widetilde{W}_m$ under P_C, the upper Hessenburg matrix H_m is identical to $\widetilde{H}_m$. Then the matrix exponentials can be approximated by (1.43), where only one P_C projection is applied. Observe that when $h_{m+1,m} = 0$ in (1.32), we have $S = W_m H_m W_m^\top C$, which suggests the approximation $W_m H_m W_m^\top C$ of S. The proof is straightforward and thus omitted.

PROPOSITION 1.6. *Consider the following two C-orthogonal Arnoldi iterations to generate (W_m, H_m) and $(\widetilde{W}_m, \widetilde{H}_m)$ from S and $\widetilde{S}$, respectively:*

$$SW_m = W_m H_m + h_{m+1,m}w_{m+1}e_m^\top, \tag{1.32}$$

$$\widetilde{S}\widetilde{W}_m = \widetilde{W}_m \widetilde{H}_m + \widetilde{h}_{m+1,m}\widetilde{w}_{m+1}e_m^\top, \tag{1.33}$$

where columns of W_m and $\widetilde{W}_m$ both form two sets of C-orthonormal vectors:

$$W_m = [w_1, w_2, \ldots, w_m], \ \widetilde{W}_m = [\widetilde{w}_1, \widetilde{w}_2, \ldots, \widetilde{w}_m], \ \ W_m^\top C W_m = \widetilde{W}_m^\top C \widetilde{W}_m = I.$$

- *Suppose the first column of W_m lies in the range of $P_C G^{-1} C$. Then all columns of W_m lie in the range of $P_C G^{-1} C$.*
- *Suppose the first column of $\widetilde{W}_m$ lies in the range of $G^{-1}C$. Then all columns of $\widetilde{W}_m$ lie in the range of $G^{-1}C$.*
- *Suppose $(\widetilde{W}_m, \widetilde{H}_m)$ satisfies* (1.33)*. Let $W_m = P_C \widetilde{W}_m$ and $H_m = \widetilde{H}_m$. Then (W_m, H_m) satisfies* (1.32)*.*

The C-orthogonality together with the positive definite assumption of G indicates the passivity and the invertibility of H_m. This is also known as the stability condition [52].

Algorithm 1.1 An Arnoldi algorithm with explicit structured orthogonalization and implicit regularization [59]

Input: C, G, k, γ, w, m
Output: H_m, W_m
Set $w = P_C w$, $w_1 = \frac{w}{\|w\|_C}$, where $\|w\|_C = \sqrt{w^\top C w}$ and $w_1^\top C w_1 = 1$
for $j = 1 : m$ **do**
  Solve $(\gamma G + C)w = C w_j$ and obtain w. Set $w = P_C w$
  for $i = 1 : j$ **do**
    $h_{i,j} = w^\top C w_i$, $w = w - h_{i,j} w_i$
  end
  $h_{j+1,j} = \|w\|_C$, $w_{j+1} = \frac{w}{h_{j+1,j}}$
  if *residual* $<$ *tolerance* **then**
    Results converge at dimension m
  end
end

Remark 1.7 (passivity property). *Assume G, C given in* (1.3)*. The advantage of the C-orthogonal iterations in* (1.32) *lies in the preservation of the passivity property of H_m, i.e., all eigenvalues of H_m have nonnegative real components. In particular, with G positive definite, we have the invertibility of H_m, which is crucial to the algorithm as well as the error analysis. Indeed, observe that* (1.32) *implies*

$$W_m^\top C S W_m = W_m^\top C P_C (C + \gamma G)^{-1} C W_m = W_m^\top C (C + \gamma G)^{-1} C W_m = H_m. \tag{1.34}$$

Then for each nonzero vector $x \in \mathbb{R}^m$, with $y := (C + \gamma G)^{-1}(C W_m x) \in \mathbb{R}^N$, we have

$$\langle x, H_m x \rangle = (C W_m x)^\top (C + \gamma G)^{-1} (C W_m x) = y^\top (C + \gamma G) y \geq 0.$$

The following shows the relation between $B_{1,1}$ and $S_{1,1}$.

Proposition 1.8. *Suppose C, G satisfy* (1.3)*. Let $\gamma > 0$, and introduce the function $g : \mathbb{C} \to \mathbb{C}$ and its inverse g_1,*

$$\lambda = g(\mu) = (1 + \gamma \mu^{-1})^{-1}, \quad \mu = g_1(\lambda) := g^{-1}(\lambda) = ((\lambda^{-1} - 1)/\gamma)^{-1}.$$

Then

$$B_{1,1} = g^{-1}(S_{1,1}),\ S_{1,1} = g(B_{1,1}). \tag{1.35}$$

Proof. By Proposition 1.2, $B_{1,1}$ is invertible. Let $C_1 = V_C^\top C V_C$. Using $V := [V_C, V_{\mathcal{N}}]$ in (1.17), we have

$$C V_{\mathcal{N}} = V_C C_1 V_C^\top V_{\mathcal{N}}^\top = 0 \quad \text{and} \quad B_{1,1} = V_C^\top G^{-1} C V_C. \tag{1.36}$$

Hence, $V^\top V = I$ gives

$$S_{1,1} = V_C^\top (G^{-1}(C+\gamma G))^{-1} G^{-1} C V_C \tag{1.37}$$

$$= V_C^\top (G^{-1}C + \gamma I)^{-1} V^\top V G^{-1} C V_C \tag{1.38}$$

$$= (B_{1,1} + \gamma I)^{-1} B_{1,1} = g(B_{1,1}), \tag{1.39}$$

where we used

$$V^\top G^{-1} C V = \begin{pmatrix} B_{1,1} & 0 \\ * & 0 \end{pmatrix},$$

$$(V^\top (G^{-1}C + \gamma I)V)^{-1} V^\top G^{-1} C V = \begin{pmatrix} (B_{1,1}+\gamma I)^{-1} B_{1,1} & 0 \\ * & 0 \end{pmatrix}. \qquad \square$$

Introduce a few notations. Let g, g_1 be given in Proposition 1.8 and

$$f(\lambda) := \varphi_0(-t(g^{-1}(\lambda))^{-1}) = \varphi_0(-t g_1(\lambda)^{-1}), \tag{1.40}$$

and let

$$f_k(\lambda) := g_1(\lambda)^{-1} \varphi_k(-t g_1(\lambda)^{-1}) \quad \text{for } k = 1, 2. \tag{1.41}$$

Now we are ready to state one approximation $x_a(t;m)$ for $x_{\mathcal{R}}(t)$ in (1.29). The error analysis will be given in the next section. For simplicity, we shall write $x_a(t) = x_a(t;m)$ if there is no confusion.

THEOREM 1.9. *Let $(\widetilde{W}_m, H_m)$ and (W_m, H_m) be generated from Arnoldi iterations with respect to $\widetilde{S}$ and S in Proposition* 1.6. *Let*

$$\begin{aligned} x_a(t;m) &:= W_m \{ f(H_m) W_m^\top C x(0) \\ &\qquad + t f_1(H_m) W_m^\top C u(0) + t^2 f_2(H_m) W_m^\top C u'(0) \} \end{aligned} \tag{1.42}$$

$$\begin{aligned} &= P_C \widetilde{W}_m \{ f(H_m) \widetilde{W}_m^\top C x(0) \\ &\qquad + t f_1(H_m) \widetilde{W}_m^\top C u(0) + t^2 f_2(H_m) \widetilde{W}_m^\top C u'(0) \}. \end{aligned} \tag{1.43}$$

Suppose $x(0)$, $u(0)$, and $u'(0)$ all lie in the range of W_m and $h_{m+1,m} = 0$. Then $x_a(t;m) = x_{\mathcal{R}}(t)$.

Proof. Write $x_{\mathcal{R}}(t)$ in (1.29) as follows:

$$x_{\mathcal{R}}(t) := z_1(t) + z_2(t) + z_3(t).$$

The first term in (1.29) gives

$$z_1(t) = V_C \exp(-t B_{1,1}^{-1}) V_C^\top x(0) \tag{1.44}$$

$$= V_C \exp(-t\{g^{-1}(S_{1,1})\}^{-1}) V_C^\top x(0) = V_C f(S_{1,1}) V_C^\top x(0). \tag{1.45}$$

The approximation of (1.45) is computed as follows. From

$$S_{1,1} \approx V_C^\top W_m H_m W_m^\top C V_C$$

and $V_C V_C^\top W_m = W_m$, we have

$$(S_{1,1})^k \approx V_C^\top W_m H_m^k W_m^\top C V_C. \tag{1.46}$$

Since columns of W_m lie in V_C, then with C-orthogonality, (1.45) yields

$$z_1(t) \approx W_m f(H_m) W_m^\top C V_C V_C^\top x(0) \approx W_m f(H_m) W_m^\top C x(0). \tag{1.47}$$

In the case of $h_{m+1,m} = 0$, the equalities in (1.46) hold and thus the equalities in (1.47) hold. For the remaining terms $z_2(t), z_3(t)$ of (1.29), we have

$$V_C \varphi_0(-tB_{1,1}^{-1}) V_C^\top \widetilde{u}(0) = V_C \exp(-tB_{1,1}^{-1}) V_C^\top \widetilde{u}(0) \approx W_m f(H_m) W_m^\top C \widetilde{u}(0).$$

Likewise, since $(B_{1,1})^{-1} = (g^{-1}(S_{1,1}))^{-1} = g_1(S_{1,1})$, then

$$\begin{aligned} &V_C B_{1,1}^{-1} \varphi_k(-tB_{1,1}^{-1}) V_C^\top \widetilde{u}'(0) = V_C g_1(S_{1,1})^{-1} \varphi_k(-t g_1(S_{1,1})^{-1}) V_C^\top \widetilde{u}'(0) \\ &\approx W_m f_k(H_m) W_m^\top C \widetilde{u}'(0). \end{aligned}$$

In summary, we have (1.43) and (1.43) by Proposition 1.6. ∎

REMARK 1.10 (solutions $x(t)$). *With* (1.43)*, we can compute one approximation for* $x_{\mathcal{R}}(t) + x_{\mathcal{N}}(t)$*; i.e., from* (1.2)*, we have*

$$x(t) = x_{\mathcal{R}}(t) + x_{\mathcal{N}}(t) = G^{-1}u(t) - G^{-1}C\frac{dx_{\mathcal{R}}(t)}{dt}. \tag{1.48}$$

REMARK 1.11. *Using* (1.43) *to compute the solution* $x_{\mathcal{R}}$ *in* (1.29) *can be regarded as the exponential Rosenbrock–Euler method (for nonautonomous systems)* [21]*. Suppose* $x(0)$*,* $u(0)$*, and* $u'(0)$ *are lying in* $\mathcal{R}$*. In the Krylov subspace methods, it is typical to choose them as the initial vector of the corresponding Arnoldi iterations with a proper* C*-normalization, i.e., the first column of* $\widetilde{W}_m$ *is the normalized vector* $w/\langle w, Cw\rangle^{1/2}$*. Note that when* $(\widetilde{W}_m^{(0)}, H_m^{(0)})$ *is generated from the* C*-orthogonal Arnoldi iterations with the initial vector* x_0 *in* $\mathcal{R}$*, the first term of* $x_a(t)$ *in* (1.43) *becomes* $\beta_0 P_C \widetilde{W}_m^{(0)} f(H_m^{(0)}) e_1$*, where* $\beta_0 = \|x_0\|_C$*. Since the computational cost of matrix exponentials is expensive, empirically one can collect all the exponential terms as one matrix-exponential-and-vector product (either* φ_0*,* φ_1*, or* φ_2*) and construct only one pair of* (W, H) *to conduct the computation. Numerical studies on circuit simulation applications can be found in* [59]*.*

2. Error analysis.

2.1. C-numerical range.

The numerical range (or called field of values) [29, 26, 54, 7, 1], which is the range of Rayleigh quotient, is one fundamental quantity in the stability analysis and the error analysis of matrix exponential computation.

To establish the convergence, for a square matrix $A \in \mathbb{C}^{N\times N}$ of the form $A = KC$ with some matrix $K \in \mathbb{R}^{N\times N}$, we introduce the C-numerical range

$$\mathcal{F}_C(A) = \{x^* CAx : x \in \mathbb{C}^N, \|x\|_C := \sqrt{x^* Cx} = 1\}, \tag{2.1}$$

which is one generalization of the standard numerical range

$$\mathcal{F}(A) = \{x^* Ax : x \in \mathbb{C}^N, \|x\| = 1\}.$$

Here A could be the matrix B in (1.24) or S in (1.30). Clearly, the set $\mathcal{F}_C(A)$ in (2.1) only depends on those vectors x in the range C.

When G is asymmetric, numerical range $\mathcal{F}_C(B)$ is not a line-segment on the real axis in general. The smallest disk covering $\mathcal{F}_C(B)$ is introduced to quantize the spectrum of $B = G^{-1}C$. For G, C in (1.3), let $C = V_C C_1 V_C^\top$ be the eigenvector

decomposition. Note that eigenvalues of $B_{1,1}$ lie in the right half-plane from Proposition 1.2, which does not imply that $\mathcal{F}(B_{1,1})$ lies in the right half-plane. By contrast, the C-numerical range $\mathcal{F}_C(B)$ always lies in the right half-plane according to the following proposition.

PROPOSITION 2.1. *Suppose that C satisfies* (1.3). *Let A be in the form of $A = KC$ for some matrix $K \in \mathbb{R}^{N\times N}$. Then both $\mathcal{F}(A)$ and $\mathcal{F}_C(A)$ contain all nonzero eigenvalues of A. In addition, if K is positive semidefinite, then $\mathcal{F}_C(A)$ lies in the right half-plane.*

Proof. Let x be a nonzero eigenvector of A corresponding to nonzero eigenvalue λ. Suppose that $x^*Cx > 0$. Then $Ax = \lambda x$ and the first statement is given by

$$\lambda = \frac{x^*Ax}{x^*x} = \frac{x^*CAx}{x^*Cx}.$$

Then λ lies in both $\mathcal{F}(A)$ and $\mathcal{F}_C(A)$. Suppose that $x^*Cx = 0$, i.e., the eigenvector x lies in the null space of C. Then

$$\frac{x^*Ax}{x^*x} = \frac{x^*KCx}{x^*x} = 0.$$

Thus, $\lambda = 0$ in this case.

In addition, when K is positive semidefinite,

$$\frac{x^*CAx}{x^*Cx} = \frac{x^*CKCx}{x^*Cx} = \frac{x^*C(K+K^\top)Cx}{2x^*Cx} \tag{2.2}$$

have a nonnegative real component for all $x \in \mathbb{C}^N$ with $x^*Cx > 0$. ∎

To proceed, we introduce a few notations. First, the set of a disk with center $c_1 \in \mathbb{C}$ and radius $\rho_1 > 0$ is denoted by $\mathcal{D}(c_1, \rho_1) \subset \mathbb{C}$. Second, since C is real, symmetric, and positive semidefinite, then there exists a unique positive semidefinite Hermitian matrix, a matrix square root denoted by $C^{1/2}$, such that $(C^{1/2})^2 = C$ (Theorem 7.2.6 in [23]). Indeed, with a unitary diagonalization $C = V_C C_1 V_C^\top$ in (1.18), the matrix square root $C^{1/2}$ is given by $C^{1/2} = V_C C_1^{1/2} V_C^\top$, where C_1 is diagonal, $C_1 = \operatorname{diag}((C_1)_{1,1}, \ldots, (C_1)_{n,n})$, and its matrix square root is given by $C_1^{1/2} := \operatorname{diag}((C_1)_{1,1}, \ldots, (C_1)_{n,n})$. Here are a few properties of $\mathcal{F}_C(B)$ if G is positive definite.

PROPOSITION 2.2. *Suppose that* (1.3) *holds for G, C. Let $H = G^{-1}$. Then, for some real parameters $\xi_1, \xi_2, \xi_3, \xi_4$, and ξ_5, we have the following:*

- *$(H + H^\top)/2$ is positive definite with eigenvalues in $[\xi_1, \xi_2]$ with $\xi_1 > 0$;*
- *$(H - H^\top)/2$ has eigenvalues in $[-i\xi_3, i\xi_3]$, $\xi_3 \geq 0$;*
- *C is positive semidefinite with eigenvalues in $\{0\} \cup [\xi_4, \xi_5]$, $\xi_4 > 0$.*

In addition, $\mathcal{F}_C(B)$ lies in $\mathcal{D}(c_1, \rho_1)$ with $c_1 > \rho_1$. Here c_1, ρ_1 only depend on these parameters $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$ of C, G.

Proof. Note that the C-numerical range $\mathcal{F}_C(B)$ of B can be expressed by

$$\left\{ \frac{x^*CG^{-1}Cx}{x^*Cx} : x \in \mathbb{C}^N, Cx \neq 0 \right\} = \{ z^* C_1^{1/2} V_C^\top G^{-1} V_C C_1^{1/2} z : \|z\| = 1, z \in \mathbb{C}^n \}. \tag{2.3}$$

From $G^{-1} = (H + H^\top)/2 + (H - H^\top)/2$, then $\mathcal{F}_C(B)$ lies within a box region in the right half-plane,

$$0 < \xi_1\xi_4 \leq \Re(\mathcal{F}_C(B)) \leq \xi_2\xi_5, \quad -\xi_3\xi_5 \leq \Im(\mathcal{F}_C(B)) \leq \xi_3\xi_5,$$

where equalities can hold only if z is a pure real vector or a pure imaginary vector. Thus, we can find some positive reals, $c_1 > 0$, $\rho_1 > 0$ with $c_1 - \rho_1 > 0$, such that $\mathcal{F}_C(B) \subset \mathcal{D}(c_1, \rho_1)$. Indeed, choose

$$\rho_1 := \sqrt{(\max(c_1 - \xi_1\xi_4, \xi_2\xi_5 - c_1))^2 + (\xi_3\xi_5)^2},$$

such that the aforementioned box region is covered by $\mathcal{D}(c_1, \rho_1)$. Note that $c_1^2 \geq \rho_1^2$ holds if and only if

$$c_1 \geq \max\{(2\xi_1\xi_4)^{-1}\{\xi_1^2\xi_4^2 + \xi_3^2\xi_5^2\}, (2\xi_2\xi_5)^{-1}\{\xi_2^2\xi_5^2 + \xi_3^2\xi_5^2\}\}. \tag{2.4}$$

Hence, (2.4) indicates that the disk $\mathcal{D}(c_1, \rho_1)$ containing $\mathcal{F}_C(B)$ lies in the right half-plane. ∎

In general, B is not normal. The following proposition and remark exhibit the dependence of $\mathcal{F}_C(S)$ and $\mathcal{F}(H_m)$ on $\mathcal{F}_C(B)$. As long as $\mathcal{F}_C(B)$ lies in the right half-plane, $\mathcal{F}(H_m)$ does as well. The following function g, which is one Möbius transformation, maps generalized circles to generalized circles, which actually lie within $\mathcal{D}(1/2, 1/2)$. Here since $\mu_1 \geq 0, \mu_2 \geq 0$, then $g(\mu_2) \leq 1$ and $g(\mu_1) \geq 0$. Thus, $\mathcal{F}_C(S) \subset \mathcal{D}(1/2, 1/2)$.

PROPOSITION 2.3. *Fix $\gamma > 0$. Suppose* (1.3). *Let $g : \mathbb{C} \to \mathbb{C}$ be the function, $g(\mu) = (1 + \gamma\mu^{-1})^{-1}$, as given in Proposition* 1.8. *Then $g(\mu)$ lies in $\mathcal{F}_C(S)$ for each $\mu \in \mathcal{F}_C(B)$. More precisely, let $\mathcal{F}_C(B)$ lie in the right half-plane,*

$$\mathcal{F}_C(B) \subset \mathcal{D}(c_1, \rho_1) \qquad \text{with some } c_1, \rho_1 \in \mathbb{R}. \tag{2.5}$$

Let $\mu_1 := c_1 - \rho_1 > 0$ and $\mu_2 := c_1 + \rho_1$. Then $\mathcal{F}_C(S) \subset \mathcal{D}(c_0, \rho_0)$, where $c_0 = (g(\mu_1) + g(\mu_2))/2$, $\rho_0 = (g(\mu_2) - g(\mu_1))/2$.

Proof. Let

$$T = (B - c_1 I)/\rho_1, \tag{2.6}$$

i.e., $B = c_1 I + \rho_1 T$, where $c_1 = (\mu_1 + \mu_2)/2$, $\rho_1 = (\mu_2 - \mu_1)/2$. By (2.5), we have $|\mathcal{F}_C(T)| \leq 1$. Choose one analytic function f on $z \in \mathcal{D}(0, 1) \to \mathcal{D}(0, 1)$,

$$f(z) = \frac{g(\rho_1 z + c_1) - c_0}{\rho_0}.$$

Since g is a function mapping a circle with center at the real axis to another circle with center at the real axis, by definition of c_0, c_1, ρ_1, $|f(z)| \leq 1$ for all $|z| \leq 1$. Clearly, $f(z)$ is analytic in $|z| < 1$ and continuous on the boundary. By Theorem 4 in [2], $\mathcal{F}_C(f(B))$ also lies in $\mathcal{D}(0, 1)$. Thus, with (2.6),

$$\mathcal{F}_C(S) = \mathcal{F}_C(g(B)) = c_0 + \rho_0 \mathcal{F}_C\left(\frac{g(\rho_1 T + c_1 I) - c_0 I}{\rho_0}\right)$$

lies in the disk $\mathcal{D}(c_0, \rho_0)$, i.e., with center c_0 and radius ρ_0. ∎

REMARK 2.4. *The passivity property of the system indicates $\mathcal{F}(H_m)$ in $\mathcal{D}(c_0, \rho_0)$. Indeed, from* (1.34) *and $W_m^\top C W_m = I$, the numerical range of H_m lies inside the C-numerical range,*

$$\mathcal{F}(H_m) \subset \mathcal{F}_C(S), \tag{2.7}$$

according to the definition of $\mathcal{F}$ and $\mathcal{F}_C$.

To establish the convergence, we need the following results, which relate the spectral norm to the radius of its numerical range.

PROPOSITION 2.5. *Let C satisfy* (1.3). *Let $A \in \mathbb{R}^{N\times N}$ be in the form of $A = KC$ with some $K \in \mathbb{R}^{N\times N}$. Suppose that $\mathcal{F}_C(A)$ lies in $\mathcal{D}(0,\rho)$ for some $\rho > 0$. Then*

$$\|A\|_C := \sup_v \{\|Av\|_C/\|v\|_C\} \le 2\rho.$$

Proof. For any $v \in \mathbb{C}^N$ with $\|v\|_C = 1$, we have

$$\frac{\|(K+K^\top)Cv\|_C}{\|v\|_C} = \frac{\|C^{1/2}(K+K^\top)C^{1/2}C^{1/2}v\|}{\|C^{1/2}v\|} \le \|C^{1/2}(K+K^\top)C^{1/2}\|$$
$$= \max_{x\neq 0} \Re\left(\frac{x^*(CKC+CK^\top C)x}{x^*Cx}\right) \le 2\max_{x\neq 0}\Re\left(\frac{x^*CKCx}{x^*Cx}\right) \le 2\rho.$$

Likewise,

(2.8)
$$\frac{\|(K-K^\top)Cv\|_C}{\|v\|_C} \le \max_{x\neq 0} \Im\left(\frac{x^*(CKC-CK^\top C)x}{x^*Cx}\right) \le 2\max_{x\neq 0}\Im\left(\frac{x^*CKCx}{x^*Cx}\right) \le 2\rho.$$

The sum of the above two inequalities gives $\|KCv\|_C/\|v\|_C \le 2\rho$. ∎

The following inequality, which is modified from one standard result (Theorem 4.1 in [54]), introduces numerical range $\mathcal{F}_C(A)$ in estimating error bounds for (2.22).

PROPOSITION 2.6. *Let Γ be a set in $\mathbb{C}$ and $d(\Gamma, \mathcal{F}_C(A))$ be the shortest distance between Γ and $\mathcal{F}_C(A)$. Then*

$$\max_{\lambda\in\Gamma} \|(\lambda I - A)^{-1}\|_C \le d(\Gamma, \mathcal{F}_C(A))^{-1}.$$

Proof. Let $u = (\lambda I - A)^{-1}v \in \mathbb{C}^n$. Then, for each $\lambda \in \Gamma$,

$$d(\Gamma, \mathcal{F}_C(A)) \le \frac{|\langle u, C(\lambda I - A)u\rangle|}{\|u\|_C^2} = \|u\|_C^{-2}|\langle u, v\rangle_C| \le \|u\|_C^{-1}\cdot\|v\|_C.$$

Hence, for each vector v, we have

$$\frac{\|(\lambda I - A)^{-1}v\|_C}{\|v\|_C} = \frac{\|u\|_C}{\|v\|_C} \le d(\Gamma, \mathcal{F}_C(A))^{-1},$$

which completes the proof after the maximization over Γ. ∎

2.2. Error bound inequality. In the following, we shall establish one upper bound depending on time span t, dimension m, and γ to show the convergence in computing the matrix exponentials. References [50, 20] show that the error of m-dimensional approximations of matrix exponentials could decay at least linearly (superlinearly) as the Krylov dimension increases. The following error arguments are roughly based on the Crouzeix inequality [7] to bound the rational Arnoldi approximation by its maximum norm on the numerical range of the operator, as stated in Corollary 3.4 in [18]. We believe that the error bounds here are not tight from the perspective of Theorem 5 in [20]. However, we do not pursue this direction. Instead our aim is simply to demonstrate one convergence analysis for the C-orthogonality Arnoldi iterations in the rational Krylov subspace method, i.e., an error bound for

the difference between $x_{\mathcal{R}}(t)$ of (1.2) and x_a in (1.43). Clearly, from (1.35), (1.29), and (1.43), we have the following inequality:

$$\begin{aligned}(2.9)\qquad \|x_{\mathcal{R}}(t)-x_a(t)\|_C &\le \|\{V_C f(S_{1,1})V_C^\top - W_m^{(0)} f(H_m^{(0)}){W_m^{(0)}}^\top C\}x(0)\|_C\\ &+\|\{V_C f_1(S_{1,1})V_C^\top - W_m^{(1)} f_1(H_m^{(1)}){W_m^{(1)}}^\top C\}u(0)\|_C\\ &+\|\{V_C f_2(S_{1,1})V_C^\top - W_m^{(2)} f_2(H_m^{(2)}){W_m^{(2)}}^\top C\}u'(0)\|_C.\end{aligned}$$

2.2.1. Convergence. Theorem 2.7 below is one error bound for φ_l functions for $l \ge 1$ (from Theorem 5.9 in [13]). Since the analysis cannot be used in the φ_0-case, we consider φ_1 for the $x(0)$ term; i.e., with $\varphi_0(-x) = (-x)\varphi_1(-x) + 1$, we have

$$f(S_{1,1}) = (-S_{1,1})(f_1(S_{1,1})) + I,$$

which gives the φ_1-computation for $f(S_{1,1})$,

$$(2.10)\qquad V_C f(S_{1,1})V_C^\top x(0) = (-V_C S_{1,1} V_C^\top)V_C f_1(S_{1,1})V_C^\top x(0) + x(0)$$

$$(2.11)\qquad \approx (-V_C S_{1,1} V_C^\top)(W_m f_1(H_m) W_m^\top C)x(0) + x(0).$$

Hence,

$$\begin{aligned}(2.12)\qquad &\|x_{\mathcal{R}}(t)-x_a(t)\|_C\\ &\le \|\{(-S)\{V_C f_1(S_{1,1})V_C^\top - W_m^{(0)} f_1(H_m^{(0)}){W_m^{(0)}}^\top C\}x(0)\|_C\\ &\quad+\|\{V_C f_1(S_{1,1})V_C^\top - W_m^{(1)} f_1(H_m^{(1)}){W_m^{(1)}}^\top C\}u(0)\|_C\\ &\quad+\|\{V_C f_2(S_{1,1})V_C^\top - W_m^{(2)} f_2(H_m^{(2)}){W_m^{(2)}}^\top C\}u'(0)\|_C.\end{aligned}$$

THEOREM 2.7. *Consider a matrix A whose $\mathcal{F}(A)$ lies in the left complex half-plane. Let $P_m = V_m V_m^\top$ be the orthogonal projection onto the shift-and-invert Krylov subspace $Q_m(A, v)$. For the restriction $A_m = P_m A P_m$ of A to $Q_m(A, v)$, we have the error bound*

$$\|\varphi_l(A)v - \varphi_l(A_m)v\| \le \frac{\mathcal{C}(l,\gamma)}{m^{l/2}}\|v\|, \quad l \ge 1.$$

Applying this theorem to (2.12) gives Theorem 2.8, which describes the convergence of $x_a(t; m)$ in (1.43) to $x_{\mathcal{R}}(t)$, at least sublinear in m.

THEOREM 2.8. *Let $x_a(t)$ be computed from* (1.43)*, where H_m is replaced with*

$$(2.13)\qquad \widehat{H}_m := H_m(I + \gamma h_{m+1,m} V_m^\top A v_{m+1} e_m^\top)^{-1} = H_m\left(I - \frac{V_m^\top A v_{m+1} e_m^\top}{(\gamma h_{m+1,m})^{-1} + e_m^\top V_m^\top A v_{m+1}}\right).$$

Suppose that C, G satisfy (1.3)*. Let $x_{\mathcal{R}}(t)$ be given in* (1.29)*. Then*

$$(2.14)\qquad \|x_{\mathcal{R}}(t) - x_a(t)\|_C \le \frac{\mathcal{C}(1,\gamma)}{m^{1/2}t}\|S\|_C\|x(0)\|_C + \frac{\mathcal{C}(2,\gamma)}{m^{2/2}t}\|u(0)\|_C + \frac{\mathcal{C}(3,\gamma)}{m^{3/2}t}\|u'(0)\|_C,$$

where $\mathcal{C}(l,\gamma)$ is a constant depending on l, γ, but independent of m or A.

Proof. We shall verify the conditions stated in Theorem 2.7. Let

$$(2.15)\qquad A := -C_1^{1/2} B_{1,1}^{-1} C_1^{-1/2}.$$

Since $V_C^\top G^{-1} V_C$ is positive semidefinite, the positive definite condition on G together with the calculation

$$-A = C_1^{1/2} B_{1,1}^{-1} C_1^{-1/2} = C_1^{1/2} (V_C^\top G^{-1} C V_C)^{-1} C_1^{-1/2} = C_1^{-1/2} (V_C^\top G^{-1} V_C)^{-1} C_1^{-1/2}$$

implies that the numerical range $\mathcal{F}(-A)$ lies in the right complex half-plane. Let $Q_m(A, v)$ be the shift-and-invert Krylov subspace

$$Q_m(A, v) = span\{v, (I - \gamma A)^{-1} v, \ldots, (I - \gamma A)^{-(m-1)} v\}.$$

Note that the definition of $S_{1,1}$ gives

$$S_{1,1} = V_C^\top (C + \gamma G)^{-1} C V_C = (I + \gamma B_{1,1}^{-1})^{-1}, \tag{2.16}$$

and

$$A = -\gamma^{-1} C_1^{1/2} (S_{1,1}^{-1} - I) C_1^{-1/2}. \tag{2.17}$$

From (2.15), we have

$$(I - \gamma A)^{-1} = C_1^{1/2} (I + \gamma B_{1,1}^{-1})^{-1} C_1^{-1/2} = C_1^{1/2} S_{1,1} C_1^{-1/2}.$$

Thus, the subspace $Q_m(A, v)$ is actually the Krylov subspace $K_m(C_1^{1/2} S_{1,1} C_1^{-1/2}, v)$, i.e.,

$$Q_m(A, v) = span\{v, C_1^{1/2} S_{1,1} C_1^{-1/2} v, \ldots, C_1^{1/2} S_{1,1}^{m-1} C_1^{-1/2} v\}.$$

Let V_m consist of orthogonal basis vectors in $K_m(C_1^{1/2} S_{1,1} C_1^{-1/2}, v)$. Then we have Arnoldi decomposition under the Gram–Schmidt process for some upper Hessenberg matrices

$$(I - \gamma A)^{-1} V_m = C_1^{1/2} S_{1,1} C_1^{-1/2} V_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^\top. \tag{2.18}$$

The orthogonality $V_m^\top V_m = I$ gives

$$H_m = V_m^\top C_1^{1/2} S_{1,1} C_1^{-1/2} V_m.$$

Simplifying (2.18) yields

$$V_m^\top A V_m = \gamma^{-1} \left(I - H_m^{-1} - h_{m+1,m} V_m^\top A v_{m+1} e_m^\top H_m^{-1} \right) = \gamma^{-1} (I - \widehat{H}_m^{-1}),$$

where we use the Sherman–Morrison formula and introduce $\widehat{H}_m \in \mathbb{R}^{m \times m}$ in (2.13). Let $P_m := V_m V_m^\top$ be the orthogonal projection onto $Q_m(A, v)$, and let A_m be the restriction of A on $Q_m(A, v)$. Note that when $h_{m+1,m} = 0$, then

$$A_m = P_m A P_m = V_m \gamma^{-1} (I - H_m^{-1}) V_m^\top, \quad H_m = \widehat{H}_m.$$

Note that from (1.6) and (1.41), we have for $l \geq 1$,

$$tf_l(S_{1,1}) = tg_1(S_{1,1})^{-1} \left(\frac{\varphi_{l+1}(-tg_1(S_{1,1})^{-1}) - I}{-tg_1(S_{1,1})^{-1}} \right) = I - \varphi_{l+1}(-tg_1(S_{1,1})^{-1}). \tag{2.19}$$

Let $v = C^{1/2}u(0) = V_C C_1^{1/2} V_C^\top u(0)$ and $V_m = C^{1/2}W_m$. The construction of W_m ensures that its columns lie in the range of V_C. From (2.17), (2.18), and (2.19), Theorem 2.7 indicates

$$
\begin{aligned}
(2.20)\ &\|\{V_C f_l(S_{1,1})V_C^\top - W_m f_l(\widehat{H}_m)W_m^\top C\}u(0)\|_C \\
&= \|C^{1/2}\{V_C f_l(S_{1,1})C_1^{-1/2}V_C - W_m f_l(\widehat{H}_m)W_m^\top C^{1/2}\}v\| \\
&= \|V_C V_C^\top C^{1/2} V_C f_l(S_{1,1})C_1^{-1/2}V_C - V_C V_C^\top V_m f_l(\widehat{H}_m)W_m^\top C^{1/2}\}v\| \\
&= t^{-1}\left\|V_C C_1^{1/2}\varphi_{l+1}(t\gamma^{-1}(I - S_{1,1}^{-1}))C_1^{-1/2}V_C^\top v - V_m\varphi_{l+1}(t\gamma^{-1}(I-\widehat{H}_m))V_m^\top v\right\| \\
&= t^{-1}\|\varphi_{l+1}(At)v - \varphi_{l+1}(A_m t)v\| \le \frac{\mathcal{C}(l+1,\gamma)}{m^{(l+1)/2}t}\|u(0)\|_C.
\end{aligned}
$$

Take $l = 1$ for the $u(0)$-term, and we obtain the second term of the right-hand side in (2.14). Similar arguments with $l = 2$ apply to the $u'(0)$-term, and we obtain the third term. Last, similar arguments with $l = 1$ apply to the $x(0)$-term, which completes the proof of (2.14). ∎

2.2.2. Linear convergence. Suppose that the size of $\mathcal{F}_C(B)$ can be estimated. We can derive (2.24) under the framework in [20, 39]. We can estimate the error $\{V_C f(S_{1,1}) - W_m^{(0)} f(H_m)^{(0)} {W_m^{(0)}}^\top C V_C\}v$ in (A.18) for any nonzero vector $w = V_C v$ as follows. Since f in (1.40) is an analytic function on $\mathbb{C} - \{0\}$, then $f(S_{1,1})v$ and its Krylov space approximation have the Cauchy integral expression (Definition 1.11 in [19]):

$$
(2.21)\qquad V_C f(S_{1,1})v = \frac{1}{2\pi i}\int_\Gamma f(\lambda)V_C(\lambda I - V_C^\top S V_C)^{-1}v\ d\lambda
$$

$$
(2.22)\qquad = \frac{1}{2\pi i}\int_\Gamma f(\lambda)(\lambda I - S)^{-1}w\ d\lambda,
$$

$$
(2.23)\qquad W_m f(H_m)W_m^\top C V_C v = \frac{1}{2\pi i}\int_\Gamma f(\lambda)W_m(\lambda I - H_m)^{-1}W_m^\top w\ d\lambda,
$$

where Γ can be a closed contour enclosing all the eigenvalues of $S_{1,1} := V_C^\top S V_C$ but not enclosing 0. The following result shows the effectiveness of C-orthogonality Arnoldi algorithms in solving $x_{\mathcal{R}}(t)$ of (1.2) under (1.3). Since $\rho_0/r < 1$, the error tends to 0 as $m \to \infty$. The proof is listed in the appendix.

THEOREM 2.1. *Suppose C, G satisfy* (1.3). *Then $\mathcal{F}_C(B)$ is bounded by $\mathcal{D}(c_1, \rho_1)$ with a real number $c_1 > \rho_1$, i.e., 0 not inside $\mathcal{F}_C(B)$, and thus Proposition* 2.3 *indicates that $\mathcal{F}_C(S)$ is bounded by a disk $\mathcal{D}(c_0, \rho_0)$ with $c_0 > \rho_0$. Take Γ as one circle with center c_0 and radius $r \in (\rho_0, c_0)$. Then*

(2.24)

$$
\|x_{\mathcal{R}}(t) - x_a(t;m)\|_C \le 4\frac{\max_{\lambda\in\Gamma}(|f(\lambda)|\|x(0)\|_C + |f_1(\lambda)|\|u(0)\|_C + |f_2(\lambda)|\|u'(0)\|_C)}{(1-(\rho_0/r))(\frac{\rho_0}{r})^{-m}}.
$$

2.3. Surrogate bounds $E(\gamma)$ with h/γ fixed. In this section, we shall examine the dependence of γ in the right-hand side of (2.24). From Proposition 2.2, $\mathcal{F}_C(B)$ lies in the right half-plane with $c_1 > \rho_1 > 0$,

$$
\mathcal{F}_C(B) \subset \mathcal{D}(c_1, \rho_1).
$$

Let $\mu_1 := c_1 - \rho_1$, $\mu_2 := c_1 + \rho_1$ be lower and upper bounds for $\Re(\mathcal{F}_C(B))$, respectively. Since Möbius transformations map generalized circles to generalized circles,

the function g stated in Proposition 1.8 maps $\mathcal{D}(c_1, \rho_1)$ in the μ-plane to $\mathcal{D}(c_0, \rho)$ in the λ-plane, where c_0, ρ are functions of γ,

(2.25)
$$c_0 = \frac{1}{2}\left((1+\gamma/\mu_2)^{-1} + (1+\gamma/\mu_1)^{-1}\right), \quad \rho = \frac{1}{2}\left((1+\gamma/\mu_2)^{-1} - (1+\gamma/\mu_1)^{-1}\right).$$

Consider the first term of the right-hand side, the φ_0 case, where f is defined in (1.40) with $t = h$,

$$f(\lambda) = \exp(-(h/\gamma)(\lambda^{-1} - 1)).$$

One upper bound is given by

$$|f(c_0 + r)| \cdot \frac{4}{(1-\rho/r)} \cdot \left(\frac{\rho}{r}\right)^m \|x(0)\|_C. \tag{2.26}$$

To simplify the computation, choose Γ to be one circle tangent to the imaginary axis at 0, sharing the same center with $\mathcal{D}(c_0, \rho)$; i.e., $r = c_0$ is chosen. Here we are interested in asymptotic results, i.e., $m \to \infty$, and thus for the sake of simplicity, we omit the absolute constant $4\|x(0)\|_C$ in (2.26) and examine the following "surrogate" error bound $E(\gamma)$ to illustrate the dependence on γ:

$$E(\gamma) := \exp((h/\gamma)(1 - (2c_0)^{-1})) \left(\frac{\rho}{c_0}\right)^m \frac{1}{(1-\rho/c_0)}. \tag{2.27}$$

2.3.1. Surrogate bounds for φ_0 functions. Suppose the eigenvalue information on $B_{1,1}$ is not available. It is natural to choose γ proportional to h, as in [59]. The following computation gives qualitative analysis on E with respect to γ. Here we focus on the φ_0 case. Arguments can be applied to other φ_k functions after some proper modifications. The proofs are tedious and appear in the appendix. Introduce ρ_*, γ_* as follows, where $c_0(\gamma_*) = 1/2$:

$$\gamma_* = \sqrt{\mu_1\mu_2},\ \rho(\gamma_*) = \rho_* := \frac{1}{2}\frac{\sqrt{\mu_2} - \sqrt{\mu_1}}{\sqrt{\mu_2} + \sqrt{\mu_1}}.$$

At $\gamma = \gamma_*$,

$$\frac{\rho}{c_0} = \frac{\sqrt{\mu_2} - \sqrt{\mu_1}}{\sqrt{\mu_2} + \sqrt{\mu_1}}.$$

PROPOSITION 2.9. *As γ increases in $[0, \infty)$, the radius ratio*

$$\frac{\rho}{c_0} = \frac{(\mu_2 - \mu_1)\gamma}{\mu_1(\mu_2 + \gamma) + \mu_2(\mu_1 + \gamma)}$$

increases.

Proposition 2.10 indicates that when $\delta = h/\gamma$ is kept fixed, the slope of $E(\gamma)$ decreases as γ increases from 0 to ∞. The graph of $E(\gamma)$ looks like a $\cap$-shaped curve.

PROPOSITION 2.10. *Let $\delta = h/\gamma$ be fixed. Let $\omega := \mu_1/\mu_2$ and*

$$\epsilon := \delta - \frac{2m\omega}{1+3\omega}(1+\sqrt{\omega})^2 - \frac{(1+\sqrt{\omega})^2}{1+\omega}.$$

When ω gets sufficiently close to 0 with $\epsilon > 0$, we have

$$-\frac{d}{d\gamma}\log E(\gamma) \geq (\sqrt{\mu_1} + \sqrt{\mu_2})^{-2}\epsilon,$$

which implies the exponential decay of $E(\gamma)$ for $\gamma > \mu_2$,

$$E(\gamma) = E(\mu_2)\exp(-\epsilon(\gamma - \mu_2)(\sqrt{\mu_1} + \sqrt{\mu_2})^{-2}).$$

2.3.2. Surrogate bounds for higher order functions φ_k. The phi-functions

$$\varphi_0(z) = \exp(z), \quad \varphi_k(z) = z^{-1}(\varphi_{k-1}(z) - 1/((k-1)!))$$

are initially proposed to serve as error bounds for the matrix exponential function; see, e.g., Theorem 5.1 in [50]. In applications, one can use any function φ_k, $k > 0$, to compute $\exp(-B_{1,1}^{-1}h)v$. The authors of [59] observe dissimilar error behaviors, even though two equivalent phi-functions are computed based on Krylov subspace approximations,

$$\varphi_0(-hB_{1,1}^{-1})B_{1,1}v, \tag{2.28}$$

$$-h\varphi_1(-hB_{1,1}^{-1})v + B_{1,1}v. \tag{2.29}$$

Here we focus on the computation framework in (2.29). With small Krylov dimensions, the error mainly originates from the Krylov approximation error of $h\varphi_1(-hB_{1,1}^{-1})$. To estimate the error, we can choose f in (2.24) to be

$$f(\lambda) := h\varphi_1((h/\gamma)(1-\lambda^{-1})) = h\{(h/\gamma)(1-\lambda^{-1})\}^{-1}\{\exp((h/\gamma)(1-\lambda^{-1}))-1\}. \tag{2.30}$$

For general $k \geq 1$, choose

$$f(\lambda) = f(g(\mu)) = h^k\varphi_k((h/\gamma)(1-\lambda^{-1})) = h^k\varphi_k(-h/\mu) \tag{2.31}$$

in estimating the error of the φ_k case,

$$\exp(-hB_{1,1}^{-1})u = u + \sum_{j=1}^{k-1}(-hB_{1,1}^{-1})^j u + (-h)^k\varphi_k(-hB_{1,1}^{-1}) \cdot (B_{1,1}^{-1})^k u. \tag{2.32}$$

Proposition 2.11 shows that $1/k!$ is one upper bound for each φ_k for $k \geq 1$, and thus f has an upper bound $h^k/k!$. This new upper bound mainly brings two adjustments to the original $\cap$-shaped error bound. First, the exponential fast dropping corresponding to large h disappears, since the upper bound for this function f is lifted to an increasing function $h^k/k!$. Second, polynomial decaying under small γ can be obtained, in contrast to the original stagnation in the φ_0-case.

PROPOSITION 2.11. *Consider integers $k > 0$. Let f be given as in* (2.31). *Then, with $g(\mu) = (1 + \mu^{-1}\gamma)^{-1}$, $|f(g(\mu))|$ can be bounded by $h^k/(k!)$.*

Proof. Let $\lambda = g(\mu)$. Claim: for each positive integer k, we have

$$|\varphi_k(-h\mu^{-1})| \leq (k!)^{-1}.$$

By Taylor's expansion theorem, if $z < 0$, then with ξ between 0 and z,

$$\varphi_k(z) = z^{-k}\left(\exp(z) - 1 - \sum_{j=1}^{k-1}\frac{z^j}{j!}\right) = \frac{\exp(\xi)z^k/k!}{z^k} = \frac{\exp(\xi)}{k!}.$$

Since $\xi \in [-h\mu^{-1}, 0]$, then

$$|f(g(\mu))| = |h^k\varphi_k(h\mu^{-1})| \leq (k!)^{-1}\max_{\xi}|\exp(\xi)| = (k!)^{-1}h^k. \tag{2.33}$$

∎

PROPOSITION 2.12. *Consider h being proportional to γ, $\delta = h/\gamma$. Error bounds corresponding to the φ_k case can be described by*

$$E(\gamma) := h^k \Big(\frac{\rho}{c_0}\Big)^m \frac{1}{1-\rho/c_0}. \tag{2.34}$$

Then

$$\frac{d\log E}{d\log \gamma} > k \quad \forall \gamma > 0.$$

Proof. From (2.34), we have

$$\log E = k\log(\delta\gamma) + m\log\frac{\rho}{c_0} - \log\Big(1-\frac{\rho}{c_0}\Big).$$

To explore the dependence on γ, taking derivative with respect to γ yields

$$\frac{d}{d\gamma}\log E(\gamma) = \frac{d}{d\gamma}\Big\{k\log\gamma + m\log\frac{\rho}{c_0} - \log\Big(1-\frac{\rho}{c_0}\Big)\Big\} \tag{2.35}$$

$$= \frac{k}{\gamma} + 2m\bigg(\Big(\frac{1}{\mu_1}+\frac{1}{\mu_2}\Big)\gamma^2 + 2\gamma\bigg)^{-1} + \frac{\mu_2(\mu_2-\mu_1)}{(\mu_2+\gamma)(\gamma(\mu_1+\mu_2)+2\mu_1\mu_2)}. \tag{2.36}$$

Hence, for all $\gamma > 0$, we have

$$\frac{d\log E}{d\log\gamma} = k + 2m\bigg(\Big(\frac{1}{\mu_1}+\frac{1}{\mu_2}\Big)\gamma + 2\bigg)^{-1} + \frac{\mu_2(\mu_2-\mu_1)\gamma}{(\mu_2+\gamma)(\gamma(\mu_1+\mu_2)+2\mu_1\mu_2)} > k. \quad \square$$

3. Simulations. Previous work in [4, 59] is recalled to illustrate the stability issue in solving semi-explicit DAEs by the ordinary Arnoldi method.

3.1. Stability problems of DAEs. We start from a one tank lumped RLC model, as shown in Figure 1. A step input current source I_S with rise time TR $= 1ps$ is applied. The DAEs $C\dot{x} + Gx = u$ of the one tank RLC follow the semi-explicit structure as expressed in (3.1). The node voltages and branch currents in the state vector are marked in Figure 1.

$$\begin{pmatrix} 0 & & & \\ & 0 & & \\ & & C1 & \\ & & & L1 \end{pmatrix}\begin{pmatrix} \dot{v_1} \\ \dot{v_2} \\ \dot{v_3} \\ \dot{i_L} \end{pmatrix} + \begin{pmatrix} \frac{1}{R_1}+\frac{1}{R_2} & -\frac{1}{R_1} & & \\ -\frac{1}{R_1} & \frac{1}{R_1} & & 1 \\ & & 0 & -1 \\ & & -1 & 1 \quad 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ i_L \end{pmatrix} = \begin{pmatrix} I_{bias} \\ 0 \\ -I_S \\ 0 \end{pmatrix}. \tag{3.1}$$

First, rational Krylov subspace is constructed through Arnoldi iterations with the ordinary inner product in the simulation to compute the matrix exponential with lower order φ_0 functions, i.e., (1.8). We set $h = TR$ for the input transition and use fixed step size for the stable stage. Since C is singular, we do observe the failure of the application of the Arnoldi algorithm. Figure 2 depicts the node voltages and the solution residual in the simulation, showing that the residual terms on algebraic variables v_1 and v_2 start to increase slightly at early stage, but soon these variables converge to wrong values. TR method results with fixed step size $100ps$ are plotted as comparison, which show a deviation from exact solution as well.

From the observations on ill-conditioned systems from DAEs, the numerical error occurs in the calculation of algebraic variables and could result in stability issues in later simulation stages. The Arnoldi algorithm with the C semi-inner product actually

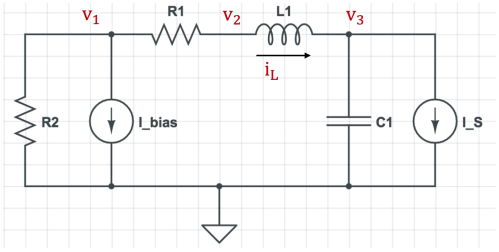

FIG. 1. *One tank RLC with* $R1 = 100\mu\Omega$, $L1 = 0.5nH$, $C1 = 0.5nF$, *and* $R2 \ll R1$.

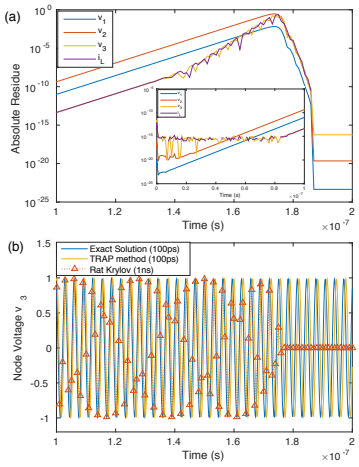

FIG. 2. *Simulation results of the one tank RLC (Figure* 1*).* (a) *absolute value of residual* $= C\dot{x}(t) + Gx(t) - u(t)$ *for each variable in* $x(t)$*;* (b) *simulation results on* v_3 *with rational Krylov subspace method as well as TR method. Exact solution is included as comparison.*

eliminates the error in the null space $\mathcal{N}(G^{-1}C) = \mathcal{N}(C)$; i.e., the algebraic variables are set to zero. The technique was called implicit regularization [4].

$$v = \begin{pmatrix} v_R \\ v_N \end{pmatrix} \Rightarrow P_C v = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_R \\ v_N \end{pmatrix} = \begin{pmatrix} v_R \\ 0 \end{pmatrix}. \tag{3.2}$$

Since C is diagonal, the matrix P_C only contains an identity matrix for the differential variables and zeros for the algebraic variables. The approach forces the computations in the range of C.

Simulation results of one tank RLC with implicit regularization are shown in Figure 3, which fit the exact solution. Residuals of v_3 and i_L remain at a low level ($\approx 10^{-15}$) when the input current is stable. The other variable could be solved algebraically, and the system no longer suffers from the singularity problem. More discussions on stability can be found in [59].

This simple example illustrates whether the numerical range of B is located in the right half-plane or does not affect the sensitivity of numerical integration methods. Indeed, since the matrix $P_C G^{-1} C$ is

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 5 \times 10^{-14} & 5 \times 10^{-10} \\ 0 & 0 & -5 \times 10^{-10} & 0 \end{pmatrix},$$

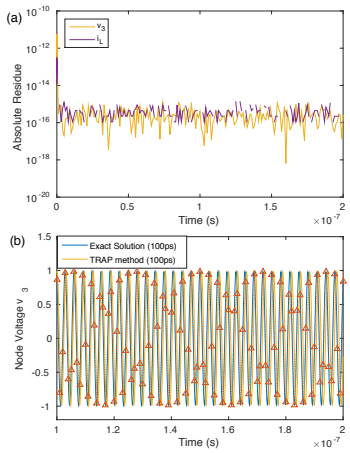

FIG. 3. *Simulation results of the one tank RLC (Figure* 1*) with implicit regularization.* (a) *The absolute residual no longer increase.* (b) *Simulation results well fit the exact solution. Node voltages* v_1, v_2 *are solved algebraically.*

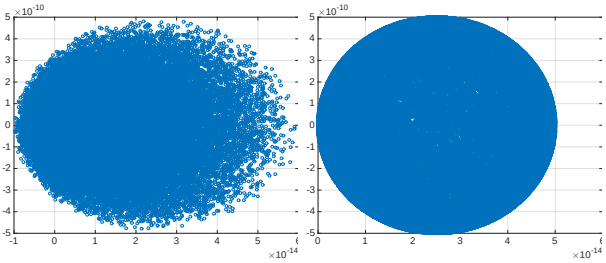

FIG. 4. *Illustration of* $\mathcal{F}(G^{-1}C)$ *(left) and* $\mathcal{F}_C(G^{-1}C)$ *(right) under* 5×10^5 *Rayleigh quotient realizations from* $\mathbb{C}^4$.

$\mathcal{F}_C(B)$ is the ellipse with center $(2.5 \times 10^{-14}, 0)$ and semimajor axis 5×10^{-14} and semiminor axis 5×10^{-10}, as shown in Figure 4. By (2.7) and Proposition 2.3, the Rayleigh quotient of the matrix H_m always lies in the image of the ellipse under the function g. Thus, $\mathcal{F}_C(H_m)$ lies in the disk $\mathcal{D}(1/2, 1/2)$. In contrast, the ordinary Arnoldi iterations generate upper Hessenberg matrix H_m, whose numerical range $\mathcal{F}(H_m)$ does not necessarily lies in $\mathcal{D}(1/2, 1/2)$, since part of $\mathcal{F}(B)$ even lies in the left half-plane.

3.2. RLC networks. To illustrate the performance of the proposed Arnoldi algorithm on the case with G only positive semidefinite, we use one PDN, consisting of 260 resistors, 160 capacitors, and 160 inductors.

The system matrix C is positive semidefinite and symmetric (actually diagonal). The matrix G is positive semidefinite but not symmetric. The eigenvalues of $B_{1,1} = V_C^\top G^{-1} C V_C$ are in the range of $[10^{-17}, 10^{-8}]$. The distribution of the eigenvalues is plotted in Figure 5. The transient response of the RLC mesh circuit is calculated with a single step integration. Assume the slope of input current source is unchanged within the current step. Starting from zero initial state $x(0)$, the response $x(h)$ of circuit

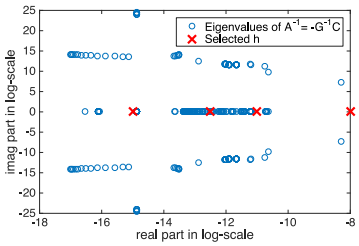

FIG. 5. *RLC network: eigenvalues of* $B = G^{-1}C$ *in log-scale.*

at time h is derived. The exact solution is computed by directly solving differential equations and algebraic equations in (1.20) and (1.21).

The shift parameter γ is set as $h/2$ empirically. The matrix exponentials in the solution are evaluated at different time step sizes h with increasing dimension m of Krylov subspace. For simplicity, we consider $x(0) = 0 = u(0)$ and the solution is given by $x(h) = h^2 V_C \varphi_2(-hB_{1,1}^{-1}) C_1^{-1} V_C^\top u'(0)$. Since

$$\varphi_0(t)v = t^2\varphi_2(t)v + v + tv = t\varphi_1(t)v + v,$$

the matrix exponential $\varphi_2(-hB_{1,1}^{-1})v$ that appeared in the solution can be computed with a Krylov subspace approximation of either φ_0, φ_1, or φ_2 functions. Consider the following three approaches to compute the Krylov subspace approximation:

(a) the original Arnoldi method with implicit regularization,
(b) the original Arnoldi method with implicit regularization + numerical pruning of spurious eigenvalues, and
(c) the Arnoldi method with structured orthogonality + numerical pruning of spurious eigenvalues.

From the left column to the right column in Figure 6, we include the distribution of absolute error after applying approach (a), (b), and (c), respectively. Here the absolute errors are focused on matrix exponentials, and thus subfigures from the top row to the bottom row show the absolute errors of the following matrix exponentials:

(i) φ_0 function: $V_C^\top \varphi_0(hB_{1,1}^{-1}) V_C G^{-1} V_C^\top C_1 V_C G^{-1} u'(0)$,
(ii) φ_1 function: $hV_C^\top \varphi_1(hB_{1,1}^{-1}) V_C G^{-1} u'(0)$, and
(iii) φ_2 function: $h^2 V_C^\top \varphi_1(hB_{1,1}^{-1}) C_1^{-1} V_C G^{-1} u'(0)$.

Experiments in Figure 6 show that the upper Hessenberg matrix can consist of many spurious eigenvalues. From (2.7) and (2.25), $\mathcal{F}_C(S) \subseteq \mathcal{D}(1/2, 1/2)$ and thus $\mathcal{F}(H_m) \subseteq \mathcal{D}(1/2, 1/2)$. The region with spurious eigenvalues is plotted in red (color is available online only). When the original Arnoldi iterations are used, the upper Hessenberg matrix could lose the positive definite property and the absolute error could grow extremely high. Clearly, the issue is resolved with (iii); see the right column. Notice that for γ close to 0, the set $\mathcal{F}(H_m)$ is very close to 1 from (2.25), and rounding errors could easily contaminate the computations of H_m, such that $\mathcal{F}(H_m)$ fails to lie in $\mathcal{D}(1/2, 1/2)$. Hence, proper numerical pruning is required. Observe that the error reduces quickly with all φ functions by increasing the dimension of rational Krylov subspace, which is consistent with Theorem 2.8. When h is larger than μ_2 (the upper bound for real components of eigenvalues of $B_{1,1}$), the calculation with the ϕ_0 function gives the best accuracy. On the other hand, if h is smaller than the spectrum, the errors (in the log-scale) with φ_1 and φ_2 exhibit a decrease proportional to γ in the log-scale, which alleviates the error stagnation in the solution with the φ_0 function. These results are consistent with empirical studies reported in [59].

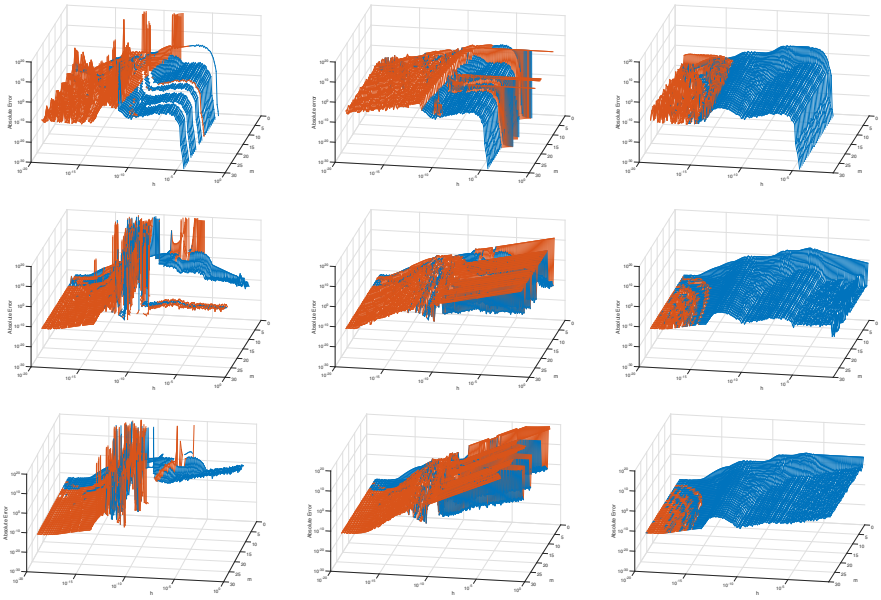

FIG. 6. *RLC network with $N = 507$: Left to right columns show the absolute error versus h and m with* (a) *original Arnoldi process,* (b) *original Arnoldi process + numerical pruning, and* (c) *Arnoldi process with explicit structured orthogonalization + numerical pruning.*

Appendix A. Proofs.

A.1. Proof of Theorem 2.1. Introduce the operator

$$\Delta_m := (\lambda I - S)^{-1} - W_m(\lambda I - H_m)^{-1}W_m^\top C.$$

Then the difference between (2.22) and (2.23) can be bounded by the operator on $w = V_C v$,

$$\{f(S) - W_m f(H_m)W_m^\top C\}V_C v \tag{A.1}$$

$$= \frac{1}{2\pi i}\int_\Gamma f(\lambda)\ \{(\lambda I - S)^{-1} - W_m(\lambda I - H_m)^{-1}W_m^\top C\}w\, d\lambda \tag{A.2}$$

$$= \frac{1}{2\pi i}\int_\Gamma f(\lambda)\,\Delta_m w\, d\lambda. \tag{A.3}$$

By computations, (1.34) gives

$$\Delta_m(\lambda I - S)W_m = \{W_m - W_m(\lambda I - H_m)^{-1}W_m^\top C(\lambda I - S)W_m\} \tag{A.4}$$

$$= \{W_m - W_m(\lambda I - H_m)^{-1}(\lambda I - H_m)\} = 0, \tag{A.5}$$

and thus

$$\Delta_m(w - (\lambda I - S)W_m y) = \Delta_m w \tag{A.6}$$

holds for any vector $y \in \mathbb{C}^m$. Note that columns of W_m lie in the subspace consisting of vectors

$$\{S^k w : k = 0, \ldots, m-1\}.$$

Hence, for each $y \in \mathbb{C}^m$, $w - (\lambda I - S)W_m y$ can be expressed as $p_m(S;\lambda)w$ for some polynomial $p_m(z;\lambda)$ of z with degree m. Note that $p_m(\lambda,\lambda) = 1$. Conversely, for

any (degree $\leq m$) polynomial $p_m(z;\lambda)$ with $p_m(\lambda;\lambda)=1$, there exists some vector $y\in\mathbb{C}^m$, such that

$$w-(\lambda I-S)W_m y = p_m(S;\lambda). \tag{A.7}$$

All together, from (A.3), (A.6), and (A.7), we have

$$V_C f(S_{1,1})v - W_m f(H_m)W_m^\top C V_C v = \frac{1}{2\pi i}\int_\Gamma f(\lambda)\Delta_m p_m(S;\lambda)w\ d\lambda. \tag{A.8}$$

Choose Γ to be the circle with center c_0 and radius r, and choose

$$p_m(z;\lambda)=\Big(\frac{z-c_0}{r}\Big)^m.$$

From (A.8), we have

$$V_C f(S_{1,1})v - V_C V_C^\top W_m f(H_m)W_m^\top C V_C v \tag{A.9}$$

$$= \frac{1}{2\pi i}\int_\Gamma f(\lambda)\Delta_m\left(\frac{S-c_0 I}{r}\right)^m w\,d\lambda \tag{A.10}$$

$$= \frac{(\rho_0/r)^m}{2\pi i}\int_\Gamma f(\lambda)\Delta_m p_m(S;\lambda)w\,d\lambda. \tag{A.11}$$

By Proposition 2.3, $\mathcal{F}_C(S)$ is bounded by a disk $\mathcal{D}(c_0,\rho_0)$. Then Proposition 2.5 and the power inequality (in [47]) indicate

$$\left|\mathcal{F}_C\left(\frac{S-c_0 I}{\rho_0}\right)\right|\leq 1,\ \left|\mathcal{F}_C\left(\left(\frac{S-c_0 I}{\rho_0}\right)^m\right)\right|\leq 1,$$

and thus Proposition 2.5 implies

$$\|(\rho_0^{-1}(S-c_0 I))^m u\|_C \leq 2\|u\|_C.$$

Hence, with the aid of Proposition 2.6 and (2.7), we have $\|(\lambda I-S)^{-1}\|_C \leq d(\Gamma,\mathcal{F}_C(S))^{-1}$ and

$$\|W_m(\lambda I-H_m)^{-1}W_m^\top C\|_C \leq \|(\lambda I-H_m)^{-1}\| \leq d(\Gamma,\mathcal{F}_C(S))^{-1},$$

which implies

$$\|\Delta_m w\|_C \leq 2d(\Gamma,\mathcal{F}_C(S))^{-1}\|w\|_C, \tag{A.12}$$

where $d(\Gamma,\mathcal{F}_C(S))$ is the shortest distance between Γ and $\mathcal{F}_C(S)$. From (A.6) and (A.12),

$$\|\Delta_m p_m(S;\lambda)w\|_C = \|\Delta_m w\|_C \leq \frac{2}{r-\rho_0}\cdot 2\|w\|_C. \tag{A.13}$$

From (A.11) and (A.13), we have for each unit vector v,

$$\|V_C f(S_{1,1})v - W_m f(H_m)W_m^\top C V_C v\|_C \tag{A.14}$$

$$\leq \frac{1}{2\pi}(\max_{\lambda\in\Gamma}|f(\lambda)|)\cdot 2\pi r\cdot\|\Delta_m p_m(S;\lambda)w\|_C \tag{A.15}$$

$$\leq (\max_{\lambda\in\Gamma}|f(\lambda)|)\cdot 2r\,d(\Gamma,\mathcal{F}_C(S))^{-1}\cdot\|p_m(S;\lambda)w\|_C \tag{A.16}$$

$$\leq (\max_{\lambda\in\Gamma}|f(\lambda)|)\cdot\frac{4}{1-(\rho_0/r)}\Big(\frac{\rho_0}{r}\Big)^m\|w\|_C. \tag{A.17}$$

From (1.35), (1.29), and (1.43), the quality of $x_a(t)$ in (1.43) can be analyzed in the following inequality:

$$\|x_{\mathcal{R}}(t) - x_a(t)\|_C \le \|\{V_C f(S_{1,1}) V_C^\top - W_m^{(0)} f(H_m^{(0)}) {W_m^{(0)}}^\top C\} x(0)\|_C \tag{A.18}$$

$$+ \|\{V_C f_1(S_{1,1}) V_C^\top - W_m^{(1)} f_1(H_m^{(1)}) {W_m^{(1)}}^\top C\} u(0)\|_C \tag{A.19}$$

$$+ \|\{V_C f_2(S_{1,1}) V_C^\top - W_m^{(2)} f_2(H_m^{(2)}) {W_m^{(2)}}^\top C\} u'(0)\|_C. \tag{A.20}$$

Consider $w = V_C v = x(0)$, $u(0)$, and $u'(0)$ in (A.17), respectively. We complete the proof.

A.2. Proof of Proposition 2.9.

Proof. Derivatives of ρ, c_0 with respect to γ are

$$\frac{dc_0}{d\gamma} = -\frac{1}{2}\left(\frac{\mu_2}{(\gamma+\mu_2)^2} + \frac{\mu_1}{(\gamma+\mu_1)^2}\right) < 0 \tag{A.21}$$

and

$$\frac{d\rho}{d\gamma} = \frac{1}{2}\left\{ - \frac{\mu_2}{(\mu_2+\gamma)^2} + \frac{\mu_1}{(\mu_1+\gamma)^2}\right\}. \tag{A.22}$$

Then

$$\frac{d}{d\gamma}(\log\rho - \log c_0) = \frac{1}{\rho}\frac{d\rho}{d\gamma} - \frac{1}{c_0}\frac{dc_0}{d\gamma} = 2((\mu_1^{-1} + \mu_2^{-1})\gamma^2 + 2\gamma)^{-1} > 0. \qquad \square$$

A.3. Proof of Proposition 2.10.

Proof. By computations,

$$\frac{d}{d\gamma}\log E(\gamma) = \frac{d}{d\gamma}\left\{\delta\left(1 - \frac{1}{2c_0}\right) + m\log\frac{\rho}{c_0} - \log\left(1 - \frac{\rho}{c_0}\right)\right\} \tag{A.23}$$

$$= \xi^{-1}\left\{ - \delta + \frac{m}{\gamma}\left(\frac{2\mu_1\mu_2}{(\mu_1+\mu_2)\gamma + 2\mu_1\mu_2}\right)\xi + \frac{\mu_2(\mu_2-\mu_1)}{(\mu_2+\gamma)(\gamma(\mu_1+\mu_2)+2\mu_1\mu_2)}\xi\right\}. \tag{A.24}$$

Here the function $\xi(\gamma)$ introduced has an upper bound decreasing with respect to γ,

$$\xi(\gamma) := \frac{\mu_1^2(\mu_2+\gamma)^2}{\mu_1(\mu_2+\gamma)^2}\frac{(1 + \frac{\mu_2(\mu_1+\gamma)}{\mu_1(\mu_2+\gamma)})^2}{1 + \frac{\mu_2(\mu_1+\gamma)^2}{\mu_1(\mu_2+\gamma)^2}} = \mu_1\frac{1 + 2\frac{\mu_2(\mu_1+\gamma)}{\mu_1(\mu_2+\gamma)} + (\frac{\mu_2(\mu_1+\gamma)}{\mu_1(\mu_2+\gamma)})^2}{1 + \frac{\mu_2(\mu_1+\gamma)^2}{\mu_1(\mu_2+\gamma)^2}} \tag{A.25}$$

$$\le \mu_1\left(1 + 2\left(\frac{\mu_2+\gamma}{\mu_1+\gamma}\right) + \frac{\mu_2}{\mu_1}\right) = \mu_1 + \mu_2 + 2\frac{\mu_1}{\mu_1+\gamma}(1 + \mu_2 - \mu_1). \tag{A.26}$$

Using the AM-GM inequality on the denominator for the second term of (A.25), we have one upper bound for ξ,

$$\xi(\gamma) \le \mu_1(1 + \sqrt{2\mu_2/\mu_1} + \mu_2/\mu_1) = (\sqrt{\mu_1} + \sqrt{\mu_2})^2.$$

Hence, for $\gamma \ge \mu_2$, (A.24) gives

$$-\frac{d}{d\gamma}\log E(\gamma) \ge (\sqrt{\mu_1} + \sqrt{\mu_2})^{-2}\left\{\delta - \left\{\frac{2m\mu_1 + \mu_2 - \mu_1}{\mu_2(3\mu_1+\mu_2)}\right\}(\sqrt{\mu_1} + \sqrt{\mu_2})^2\right\}. \qquad \square$$

Acknowledgment. We thank the anonymous referees for suggestions and corrections that have improved the presentation.

REFERENCES

[1] B. Beckermann and L. Reichel, *Error estimates and evaluation of matrix functions via the Faber transform*, SIAM J. Numer. Anal., 47 (2009), pp. 3849–3883, https://doi.org/10.1137/080741744.

[2] C. A. Berger and J. G. Stampfli, *Mapping theorems for the numerical range*, Amer. J. Math., 89 (1967), pp. 1047–1055.

[3] M. A. Botchev, V. Grimm, and M. Hochbruck, *Residual, restarting, and Richardson iteration for the matrix exponential*, SIAM J. Sci. Comput., 35 (2013), pp. A1376–A1397, https://doi.org/10.1137/110820191.

[4] P. Chen, C. K. Cheng, D. Park, and X. Wang, *Transient circuit simulation for differential algebraic systems using matrix exponential*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2018, 99.

[5] L. O. Chua and P.-M. Lin, *Computer Aided Analysis of Electric Circuits: Algorithms and Computational Techniques*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[6] W. Cody, G. Meinardus, and R. Varga, *Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems*, J. Approximation Theory, 2 (1969), pp. 50–65, https://doi.org/10.1016/0021-9045(69)90030-6.

[7] M. Crouzeix, *Numerical range and functional calculus in Hilbert space*, J. Funct. Anal., 244 (2007), pp. 668–690, https://doi.org/10.1016/j.jfa.2006.10.013.

[8] V. Druskin and L. Knizhnerman, *Extended Krylov subspaces: Approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771, https://doi.org/10.1137/S0895479895292400.

[9] T. Ericsson, *A generalised eigenvalue problem and the Lanczos algorithm*, in Large Scale Eigenvalue Problems, North-Holland Math. Stud. 127, J. Cullum and R. A. Willoughby, eds., North-Holland, Amsterdam, 1986, pp. 95–119, https://doi.org/10.1016/S0304-0208(08)72642-2.

[10] T. Ericsson and A. Ruhe, *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp., 35 (1980), pp. 1251–1268, https://doi.org/10.2307/2006390.

[11] R. W. Freund, *Krylov-subspace methods for reduced-order modeling in circuit simulation*, J. Comput. Appl. Math., 123 (2000), pp. 395–421.

[12] R. A. Friesner, L. S. Tuckerman, B. C. Dornblaser, and T. V. Russo, *A method for exponential propagation of large systems of stiff nonlinear differential equations*, J. Sci. Comput., 4 (1989), pp. 327–354, https://doi.org/10.1007/BF01060992.

[13] T. Göckler, *Rational Krylov Subspace Methods for Phi-Functions in Exponential Integrators*, Ph.D. thesis, Karlsruher Institut für Technologie, Karlsruhe, Germany, 2014, https://doi.org/10.5445/IR/1000043647.

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

[15] V. Grimm and T. Göckler, *Automatic smoothness detection of the resolvent Krylov subspace method for the approximation of C_0-semigroups*, SIAM J. Numer. Anal., 55 (2017), pp. 1483–1504, https://doi.org/10.1137/15M104880X.

[16] V. Grimm and M. Hochbruck, *Rational approximation to trigonometric operators*, BIT, 48 (2008), pp. 215–229, https://doi.org/10.1007/s10543-008-0185-9.

[17] M. S. Gupta, J. L. Oatley, R. Joseph, G.-Y. Wei, and D. M. Brooks, *Understanding voltage variations in chip multiprocessors using a distributed power-delivery network*, in Proceedings of IEEE Design, Automation, and Test in Europe Conference & Exhibition, 2007, pp. 1–6.

[18] S. Güttel, *Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection*, GAMM-Mitt., 36 (2013), pp. 8–31, https://doi.org/10.1002/gamm.201310002.

[19] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008, https://doi.org/10.1137/1.9780898717778.

[20] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925, https://doi.org/10.1137/S0036142995280572.

[21] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numer., 19 (2010), pp. 209–

286, https://doi.org/10.1017/S0962492910000048.
[22] M. Hochbruck, A. Ostermann, and J. Schweitzer, *Exponential Rosenbrock-type methods*, SIAM J. Numer. Anal., 47 (2009), pp. 786–803, https://doi.org/10.1137/080717717.
[23] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, UK, 2013, https://doi.org/10.1017/9781139020411.
[24] A. Ilchmann and T. Reis, eds., *Surveys in Differential-Algebraic Equations* II, Springer, Cham, 2015.
[25] J. Jimenez, H. de la Cruz, and P. D. Maio, *Efficient computation of phi-functions in exponential integrators*, J. Comput. Appl. Math., 374 (2020), 112758, https://doi.org/10.1016/j.cam.2020.112758.
[26] C. R. Johnson, *Numerical determination of the field of values of a general complex matrix*, SIAM J. Numer. Anal., 15 (1978), pp. 595–602, https://doi.org/10.1137/0715039.
[27] A. B. Kahng, S. Kang, H. Lee, I. L. Markov, and P. Thapar, *High-performance gate sizing with a signoff timer*, in Proceedings of IEEE/ACM International Conference on Computer-Aided Design, 2013, pp. 450–457.
[28] D. Kouroussis and F. N. Najm, *A static pattern-independent technique for power grid voltage integrity verification*, in Proceedings of the IEEE/ACM Design Automation Conference, 2003, pp. 99–104.
[29] P. D. Lax and B. Wendroff, *Difference schemes for hyperbolic equations with high order of accuracy*, Comm. Pure Appl. Math., 17 (1964), pp. 381–398, https://doi.org/10.1002/cpa.3160170311.
[30] Z. Li, R. Balasubramanian, F. Liu, and S. Nassif, 2012 *tau power grid simulation contest: Benchmark suite and results*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2012, pp. 643–646.
[31] S. Lin and N. Chang, *Challenges in power-ground integrity*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2001, pp. 651–654.
[32] S. Lin, M. Nagata, K. Shimazake, K. Satoh, M. Sumita, H. Tsujikawa, and A. T. Yang, *Full-chip vectorless dynamic power integrity analysis and verification against* $100uV/100ps$*-resolution measurement*, in Proceedings of the IEEE Custom Integrated Circuits Conference, 2004, pp. 509–512.
[33] J. Lu, P. Chen, C.-C. Chang, L. Sha, D. Huang, C.-C. Teng, and C.-K. Cheng, *ePlace: Electrostatics based placement using Nesterov's method*, in Proceedings of the IEEE/ACM Design Automation Conference, 2014, pp. 1–6.
[34] J. Lu, H. Zhuang, P. Chen, H. Chang, C.-C. Chang, Y.-C. Wong, L. Sha, D. Huang, Y. Luo, C.-C. Teng, and C. K. Cheng, *ePlace-MS: Electrostatics based placement for mixed-size circuits*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 34 (2015), pp. 685–698.
[35] J. Lu, H. Zhuang, I. Kang, P. Chen, and C.-K. Cheng, *ePlace-3D: Electrostatics based placement for 3D-ICs*, in Proceedings of the ACM International Symposium on Physical Design, 2016, pp. 11–18.
[36] K. Meerbergen and A. Spence, *Implicitly restarted Arnoldi with purification for the shift-invert transformation*, Math. Comp., 66 (1997), pp. 667–689.
[37] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836, https://doi.org/10.1137/1020098.
[38] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49, https://doi.org/10.1137/S00361445024180.
[39] I. Moret and P. Novati, *RD-rational approximations of the matrix exponential*, BIT, 44 (2004), pp. 595–615.
[40] L. Nagel, *SPICE*2: *A Computer Program to Simulate Semiconductor Circuits*, Ph.D. dissertation, University of California, Berkeley, CA, 1975.
[41] F. N. Najm, *Circuit Simulation*, Wiley, New York, 2010.
[42] S. R. Nassif, *Power grid analysis benchmarks*, in Proceedings of the Asia and South Pacific Design Automation Conference, 2008, pp. 376–381.
[43] S. R. Nassif and J. N. Kozhaya, *Fast power grid simulation*, in Proceedings of the IEEE/ACM Design Automation Conference, 2000, pp. 156–161.
[44] J. Nissen and W. M. Wright, *A Krylov subspace algorithm for evaluating the phi function appearing in exponential integrators*, ACM Trans. Math. Software, 38 (2012), pp. 1–21.
[45] B. Nour-Omid, B. N. Parlett, T. Ericsson, and P. S. Jensen, *How to implement the spectral transformation*, Math. Comp., 48 (1987), pp. 663–673.
[46] M. Pan, N. Viswanathan, and C. Chu, *An efficient and effective detailed placement algorithm*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design,

2005, pp. 48–55.
[47] C. Pearcy, *An elementary proof of the power inequality for the numerical radius.*, Michigan Math. J., 13 (1966), pp. 289–291, https://doi.org/10.1307/mmj/1031732779.
[48] J. Rommes and N. Martins, *Exploiting structure in large-scale electrical circuit and power system problems*, Linear Algebra Appl., 431 (2009), pp. 318–333.
[49] A. Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, Linear Algebra Appl., 58 (1984), pp. 391–405, https://doi.org/10.1016/0024-3795(84)90221-0.
[50] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228, https://doi.org/10.1137/0729014.
[51] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, *Full chip impact study of power delivery network designs in monolithic 3D ICs*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2014, pp. 565–572.
[52] L. M. Silveira, M. Kamon, I. Elfadel, and J. White, *A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 377–389.
[53] B. Simeon, C. Führer, and P. Rentrop, *The Drazin inverse in multibody system dynamics*, Numer. Math., 64 (1993), pp. 521–539.
[54] M. Spijker, *Numerical ranges and stability estimates*, Appl. Numer. Math., 13 (1993), pp. 241–249, https://doi.org/10.1016/0168-9274(93)90146-I.
[55] M. Takamatsu and S. Iwata, *Index characterization of differential–algebraic equations in hybrid analysis for circuit simulation*, Int. J. Circuit Theory Appl., 38 (2010), pp. 419–440.
[56] L. N. Trefethen, *Approximation Theory and Approximation Practice*, SIAM, Philadelphia, 2012.
[57] J. van den Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., 27 (2006), pp. 1438–1457, https://doi.org/10.1137/040605461.
[58] K. Wang, B. H. Meyer, R. Zhang, K. Skadron, and M. R. Stan, *Walking pads: Fast power-supply pad-placement optimization*, in Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference, 2014, pp. 537–543.
[59] X. Wang, P. Chen, and C. K. Cheng, *Stability and convergency exploration of matrix exponential integration on power delivery network transient simulation*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 39 (2020), pp. 2735–2748, https://doi.org/10.1109/TCAD.2019.2954473.
[60] G. Wanner, *Dahlquist's classical papers on stability theory*, BIT, 46 (2006), pp. 671–683.
[61] S.-H. Weng, Q. Chen, and C. K. Cheng, *Time-domain analysis of large-scale circuits by matrix exponential method with adaptive control*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 31 (2012), pp. 1180–1193.
[62] R. Winkler, *Stochastic differential algebraic equations of index 1 and applications in circuit simulation*, J. Comput. Appl. Math., 157 (2003), pp. 477–505.
[63] L. Xiao, Z. Xiao, Z. Qian, Y. Jiang, T. Huang, H. Tian, and E. F. Y. Young, *Local clock skew minimization using blockage-aware mixed tree-mesh clock network*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2010, pp. 458–462.
[64] W. Yu, H. Zhuang, C. Zhang, G. Hu, and Z. Liu, *RWCap: A floating random walk solver for 3-D capacitance extraction of very-large-scale integration interconnects*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 32 (2013), pp. 353–366.
[65] Z. Zeng, X. Ye, Z. Feng, and P. Li, *Tradeoff analysis and optimization of power delivery networks with on-chip voltage regulation*, in Proceedings of the IEEE/ACM Design Automation Conference, 2010, pp. 831–836.
[66] C. Zhang and W. Yu, *Efficient space management techniques for large-scale interconnect capacitance extraction with floating random walks*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 32 (2013), pp. 1633–1637.
[67] R. Zhang, B. H. Meyer, W. Huang, K. Skadron, and M. R. Stan, *Some limits of power delivery in the multicore era*, in Proceedings of the 4th Workshop on Energy-Efficient Design, 2012.
[68] R. Zhang, K. Wang, B. H. Meyer, M. R. Stan, and K. Skadron, *Architecture implications of pads as a scarce resource*, in Proceedings of the International Symposium on Computer Architecture, 2014, pp. 373–384.
[69] Y. Zhang and C. Chu, *GDRouter: Interleaved global routing and detailed routing for ultimate routability*, in Proceedings of the IEEE/ACM Design Automation Conference, 2012, pp. 597–602.
[70] H. Zhuang, J. Lu, K. Samadi, Y. Du, and C. K. Cheng, *Performance-driven placement for*

design of rotation and right arithmetic shifters in monolithic 3*D ICs*, in Proceedings of the IEEE International Conference on Communications, Circuits and Systems, Vol. 2, 2013, pp. 509–513.

[71] H. Zhuang, W. Yu, G. Hu, Z. Liu, and Z. Ye, *Fast floating random walk algorithm for multi-dielectric capacitance extraction with numerical characterization of Green's functions*, in Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference, 2012, pp. 377–382.

[72] H. Zhuang, W. Yu, S.-H. Weng, I. Kang, J.-H. Lin, X. Zhang, R. Coutts, J. Lu, and C. K. Cheng, *Simulation algorithms with exponential integration for time-domain analysis of large-scale power delivery networks*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 35 (2016), pp. 1681–1694.

[73] C. Zhuo, H. Gan, and W.-K. Shih, *Early-stage power grid design: Extraction, modeling and optimization*, in Proceedings of the IEEE/ACM Design Automation Conference, 2014, pp. 1–6.

[74] C. Zhuo, G. Wilke, R. Chakraborty, A. Aydiner, S. Chakravarty, and W.-K. Shih, *A silicon-validated methodology for power delivery modeling and simulation*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2012, pp. 255–262.