UNCONDITIONAL STABILITY OF A CRANK-NICOLSON/ADAMS-BASHFORTH 2 IMPLICIT/EXPLICIT METHOD FOR ORDINARY DIFFERENTIAL EQUATIONS

by

Andrew D. Jorgenson

B.S., Gonzaga University, 2009

M.A., University of Pittsburgh, 2010

Submitted to the Graduate Faculty of
the Department of Mathematics in partial fulfillment
of the requirements for the degree of

Master of Science

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

MATHEMATICS DEPARTMENT

This thesis was presented

by

Andrew D. Jorgenson

It was defended on

April 17th, 2012

and approved by

Dr. Catalin Trenchea, University of Pittsburgh, Mathematics

Dr. Myron Sussman, University of Pittsburgh, Mathematics

Dr. William Layton, University of Pittsburgh, Mathematics

Thesis Advisor: Dr. Catalin Trenchea, University of Pittsburgh, Mathematics

UNCONDITIONAL STABILITY OF A CRANK-NICOLSON/ADAMS-BASHFORTH 2 IMPLICIT/EXPLICIT METHOD FOR ORDINARY DIFFERENTIAL EQUATIONS

Andrew D. Jorgenson, M.S.

University of Pittsburgh, 2012

Systems of non-linear partial differential equations modeling turbulent fluid flow and other processes present special challanges in numerical analysis. A time-stepping Crank-Nicolson/Adams-Bashforth 2 implicit-explicit method for solving spatially-discretized systems of this type is proposed and proven to be unconditionally stable and second-order convergent.

TABLE OF CONTENTS

LIST OF FIGURES

1.0 INTRODUCTION

The motivation of this work is to consider the stability of numerical methods when applied to ordinary differential equations (henceforth "ODEs") of the form

$$u'(t) + Au(t) - Cu(t) + B(u)u(t) = f(t), \tag{1.1}$$

in which $A, B(u)$ and C are $d \times d$ matrices, $u(t)$ and $f(t)$ are d-vectors, and

$$A = A^T \succ 0, B(u) = -B(u)^T, C = C^T \succcurlyeq 0 \text{ and } \quad A - C \succ 0 \quad . \tag{1.2}$$

Here $\succ$ and $\succcurlyeq$ denote the positive definite and positive semidefinite ordering, respectively.

Models of the behavior of turbulent fluid flow using convection diffusion partial differential equations discretized in the spatial variable give rise to a system of ODEs, such as

$$\dot{u}_{ij}(t) + b \cdot \nabla^h u_{ij} - (\epsilon_0(h) + \nu)\Delta^h u_{ij} + \epsilon_0(h) P_H(\Delta^h P_H(u_{ij})) = f_{ij}, \tag{1.3}$$

where Δ^h is the discrete Laplacian, ∇^h is the discrete gradient, $\epsilon(h)$ is the artificial viscosity parameters, and P_H denotes a projection onto a coarser mesh [1]. System (1.3) is of the form (1.1), (1.2) where

$$A = -(\epsilon_0(h) + \nu)\Delta^h, \quad C = \epsilon_0(h) P_h \Delta^h P_h, \quad B(u) = b \cdot \nabla^h.$$

In this case the matrix $B(\cdot)$ is constant, but in general it may depend on u, and thus the system is allowed to have a nonlinear part. A linear multistep method for the numerical integration of the system $u'(t) = F(t, u)$, such as (1.1), is

$$\sum_{j=-1}^{k} \alpha_j u_{n-j} = \Delta t \sum_{j=-1}^{k} \beta_j F_{n-j}, \tag{1.4}$$

where t is defined on $\mathcal{I} = [t_0, t_0 + T] \subset \mathbb{R}$, $u_{n-j} \in \mathbb{R}^d$, $F_{n-j} = F(t_{n-j}, u_{n-j})$.

In [1], Anitescu et al. show that the first-order implicit-explicit (IMEX) method

$$\frac{u_{n+1} - u_n}{\Delta t} + Au_{n+1} - Cu_n + B(u_n)u_{n+1} = f_{n+1} \tag{1.5}$$

is unconditionally stable (its stability properties are independent of the choice of step-size Δt). The aim of this work is to prove unconditional stability and second-order convergence for a proposed Crank-Nicolson/Adams-Bashforth 2 IMEX numerical method,

$$
\begin{aligned}
&\frac{u_{n+1}-u_n}{\Delta t} + (A-C)^{\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\tfrac{1}{2}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big) \\
&\quad + B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big) \\
&= f_{n+\frac{1}{2}},
\end{aligned}
\tag{1.6}
$$

where $\mathcal{E}_{n+\frac{1}{2}} = \frac{3}{2}u_n + \frac{1}{2}u_{n-1}$, an explicit approximation of $u(t_{n+\frac{1}{2}})$. As will be shown in Theorem 4, this method is a second-order convergent numerical scheme of the form (1.4), where $k = 2$ and

$$
\begin{aligned}
&\alpha_{-1} = I, \qquad \alpha_0 = -I, \qquad \alpha_1 = 0 \\
&\beta_{-1} = -(A-C)^{\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}} - B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}} \\
&\beta_0 = -(A-C)^{\frac{1}{2}}\big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}} - B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}} \\
&\beta_1 = -(A-C)^{\frac{1}{2}}\tfrac{1}{2}C(A-C)^{-\frac{1}{2}} - B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}.
\end{aligned}
$$

Applying this method for solving the system (1.1) will require solving for the vector u_{n+1} in terms of u_n, u_{n-1} (given two initial conditions u_0, u_1), that is

$$
u_{n+1} = \big[I - h\beta_{-1}\big]^{-1}\Big[\big[I + h\beta_0\big]u_n + h\beta_1 u_{n-1}\Big]. \tag{1.7}
$$

The method requires the inversion

$$
\big[I - h\beta_{-1}\big]^{-1} = \big[I + h(A-C)^{\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}} + hB(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}\big]^{-1}.
$$

In practice (1.7) will not be solved by computing the inverse since this would be overly costly and introduce large round-off error, which would have adverse effects on the method at each step n. Also of note is that in general, A, B, and C do not commute, and thus the calculation in (1.7) appears to be somewhat more costly than in the case of (1.5) due to the additional $(A-C)^{-\frac{1}{2}}$ and $(A-C)^{\frac{1}{2}}$ terms. As will be seen, the fact that these matrices do not commute plays a critical role in the stability analysis developed in Chapter 3.

2.0 STABILITY CONCEPTS FOR CAUCHY PROBLEMS

2.1 STANDARD CAUCHY PROBLEM

Before considering the stability properties of a non-linear system such as (1.1), let us first analyze the stability of a more well-behaved system of first-order ODE and initial conditions

$$y'(t) = F(t, y(t)), \qquad y(t_0) = y_0 \tag{2.1}$$

defined on F: $\mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$, $y : \mathbb{R} \to \mathbb{R}^d$, $t \in I \subset \mathbb{R}$, that satisfy the Lipschitz condition

$$\|F(t, y) - F(t, z)\| \leq L\|y - z\|, \tag{2.2}$$

for some positive constant L, where $\| \cdot \|$ is an appropriate norm.

Often finding analytical solutions for systems of the form of (2.1) is difficult or impossible, so it is worth exploring suitable numerical schemes that give good approximate solutions under the broadest possible conditions. Of particular concern are the numerical method's order of convergence, consistency, and stability properties, the last of which is the main focus of this paper.

2.2 WELL-POSEDNESS OF THE CAUCHY PROBLEM

Before stability of the numerical method is analyzed it is first worth ensuring that the underlying problem (1.1) is itself stable, and thus the problem is "well-posed." "Indeed, it is not appropriate to pretend the numerical method can cure the pathologies of an intrinsically ill-posed problem"[3].

To examine the stability of the Cauchy problem (2.1), consider

$$z'(t) = F(t, z(t)) + \delta(t), \qquad z(t_0) = y_0 + \delta_0, \tag{2.3}$$

which is (2.1) but perturbed in both the initial condition and in F, where $\delta_0 \in \mathbb{R}^d$ and $\delta : \mathbb{R} \to \mathbb{R}^d$ is a continuous function.

Definition 1. *Liapunov Stability [3]. The Cauchy problem (2.1) is stable, or "Liapunov-stable," if for any perturbation ($\delta(t)$, δ_0),*

$$\|\delta(t)\|_\infty < \epsilon, \qquad \|\delta_0\|_\infty < \epsilon, \qquad \forall t \in \mathcal{I} \tag{2.4}$$

where $\|\cdot\|_\infty$ is the infinity vector norm, and $\epsilon > 0$ but small enough to ensure the solution exists, there $\exists$ a C which is independent of ϵ, such that

$$\|y(t) - z(t)\|_\infty < C\epsilon, \qquad \forall t \in \mathcal{I}. \tag{2.5}$$

Theorem 1. *If the Lipschitz condition (2.2) holds, the system of ODEs (2.1) is Liapunov-stable, and*

$$\|w(t)\|_\infty \leq (1 + |t - t_0|)\epsilon e^{L|t-t_0|}, \tag{2.6}$$

which implies $C = (1+K)e^{LK}$, where $K = max_{t\in\mathcal{I}}|t - t_0|$.

Proof. See Appendix (A.1.1). □

Thus since the largest difference in solutions between the perturbed and unperturbed Cauchy problems of any individual ODE in the system remains bounded for all $t \in I$, the underlying problem (2.1) is Liapunov-stable and is thus "well-posed." This is the same as saying that in a general sense, small changes in "data" (perturbation in the forcing term and initial conditions (2.3)) give small changes in the solution, or as stated in Theorem 1, bounded changes in the solution.

Once the stability of the underlying problem is ensured, one can then consider what kind of numerical methods applied to the problem are also stable. What follows is a review of some basic assymptotic stability concepts, which are applied to scalar examples that motivate a discussion of the stability of the proposed CN/AB2 method when applied to system (1.1).

2.3 A-STABILITY

Consider the Cauchy problem

$$\begin{aligned} &y'(t) = (\epsilon + \nu)\lambda y(t) - \epsilon\lambda y(t), \\ &y : \mathbb{R} \to \mathbb{R}, \quad y(0) = 1, \quad \lambda < 0, \quad 0 < \nu, \quad 0 < \epsilon. \end{aligned} \tag{2.7}$$

Note that this is the classic Dahlquist test-problem $y'(t) = \nu\lambda y(t)$, with exact solution $y(t) = e^{\nu\lambda t}$, broken into two parts.

Definition 2. *A-stability (Dahlquist 1963). The multistep method (1.4) applied to the Cauchy test problem (2.10) is A-stable if $A \supseteq \mathbb{C}^-$ (where A is entire region of stability for the method). This is equivalent to requiring the numerical solutions $|u_n| \to 0$ as $t_n \to +\infty$ [2].*

A method's A-stability region can be illustrated by plotting its root locus curve, that is, the values of $\Delta t\lambda\nu$ corresponding to the stability boundary roots $|\zeta(\Delta t\lambda\nu)| = 1$ of its generating polynomials (see Appendix A.1.2). Recall that for stability the roots of these polynomials must be lie within the unit circle ($\zeta_j(\Delta t\lambda\nu) \leq 1$ in modulus) [2].

The aim here is to explore methods which, when applied to the Cauchy test problem (2.7) as stated, display stable behavior. Let us consider methods which apply an implicit scheme to the first part, and an explicit scheme to the second part, which are thus called implicit/explicit (IMEX) methods. A-priori it is not obvious under which conditions such a mixed method will exhibit stable behavior (if at all), and if so, whether the stability properties of the implicit or explicit part will dominate.

2.3.1 Backward Euler/Forward Euler IMEX

Let us first investigate an IMEX method which is Backward Euler for the implicit part and Forward Euler for the explicit part;

$$\frac{u_{n+1}-u_n}{\Delta t} = (\epsilon+\nu)\lambda u_{n+1} - \epsilon\lambda u_n. \tag{2.8}$$

This method can be solved for u_n in terms of λ, ϵ, ν, and an initial condition u_0. Iterating backward n times gives

$$u_n = u_0\Big(\frac{1-\epsilon\lambda\Delta t}{1-(\epsilon+\nu)\lambda\Delta t}\Big)^n.$$

As $n \to +\infty$, $|u_n| \to 0$ if $\big|\frac{1-\epsilon\lambda\Delta t}{1-(\epsilon+\nu)\lambda\Delta t}\big| < 1$. $\lambda < 0$, $0 < \nu$, and $\epsilon < 0$ are sufficient for this to hold. None of these conditions is dependent on the choice of step-size Δt, so given our assumptions on λ, ϵ, and ν we can immediately conclude that this method is unconditionally stable.

Note that if ϵ is allowed to be zero, we recover the Backward Euler method, which has the solution

$$u_n = u_0\Big(\frac{1}{1-\nu\lambda\Delta t}\Big)^n.$$

Figure 2.1 shows the convergence of the energy (u_n^2) of the solutions of Backward Euler and Backward Euler/Forward Euler for initial condition $u_0 = 1$, $\lambda = -10000$, $\nu = .001$, $\Delta t = .01$, $\epsilon = .01$ (for the BE/FE scheme). Notice that Backward Euler converges faster than Backward Euler/Forward Euler mixed method. This illustrates that the advantages of using an IMEX method as described in Chapter 1 come at the cost of decreased speed of convergence of the method's solutions.

To see the stability region of the BE/FE method in terms of step-size and eigenvalues, we take $\zeta^n = u_n$, $\mu = \Delta t\lambda\nu$, and solving the method (2.8) for μ gives the root locus curve [2]

$$\mu = \nu\frac{\rho(\zeta)}{\sigma(\zeta)} = \nu\frac{\zeta-1}{(\alpha+\nu)\zeta-\epsilon}. \tag{2.9}$$

Since $|e^{i\theta}| = 1$ for all θ, taking $\zeta = e^{i\theta}$ in (2.9) and letting θ vary in $[0, 2\pi]$ produces the desired stability region (with ν=.001).

Figure 2.2 illustrates that the BE/FE IMEX method is stable for any choice of μ outside the solid blue line, which is to say the method (2.8) is A-stable since any choice of $\Delta t\lambda\nu$ in $\mathbb{C}^-$ will be stable and the solution u_n will converge to zero as n gets large. This plot is the same, except for the size of the stability region, for any choice of ϵ.

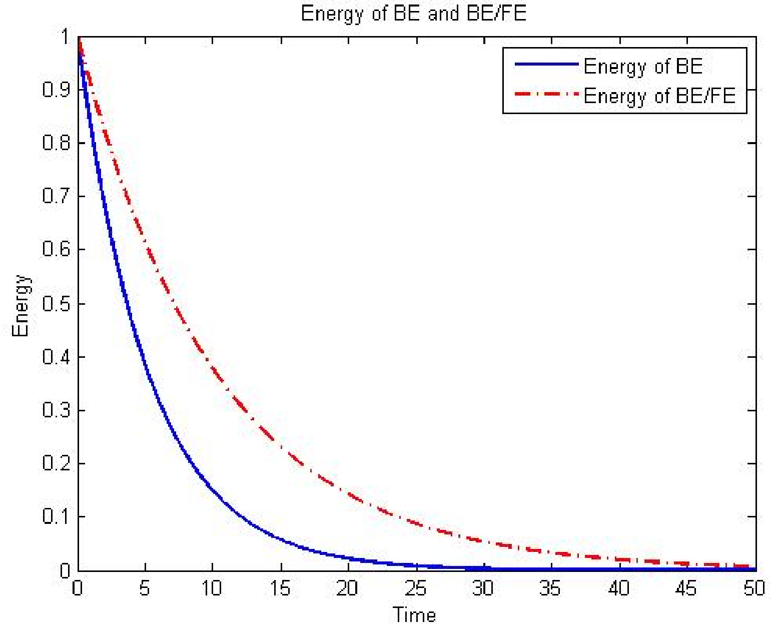

Figure 2.1: Energy decay of BE and BE/FE methods

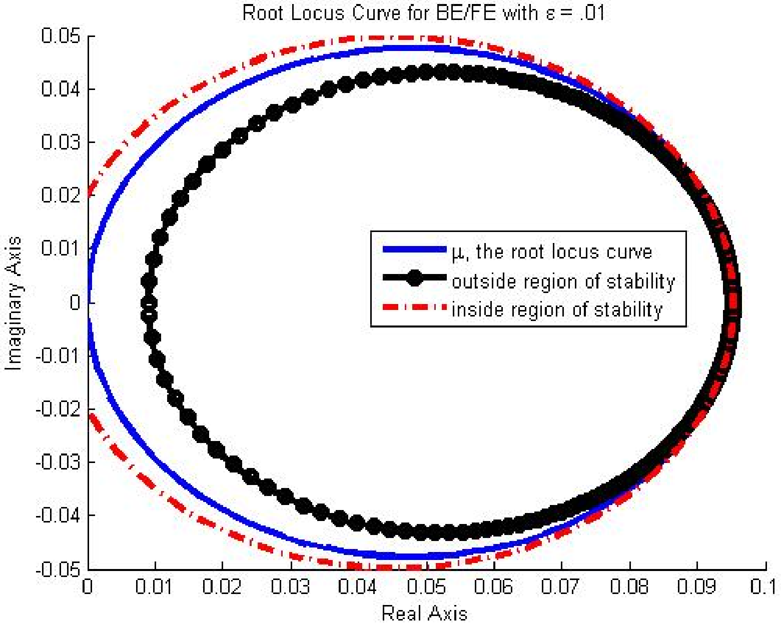

Figure 2.2: Root locus curve denoting the region of A-stability for Backward Euler/Forward Euler method with $\epsilon = .01$

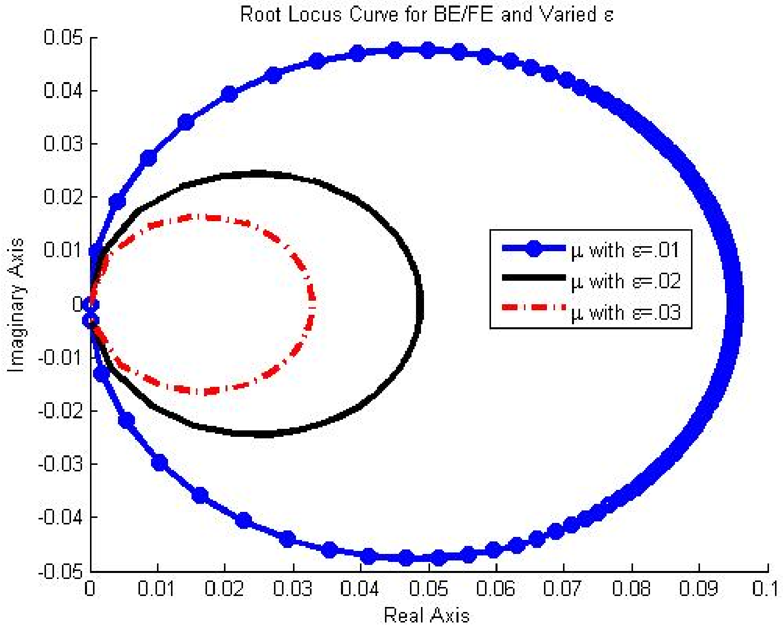

Figure 2.3: Root locus curve denoting the regions of A-stability for Backward Euler/Forward Euler method for different values of ϵ

Firgure 2.3 shows, somewhat counterintuitively, that the stability region of BE/FE IMEX method is growing with ϵ (since the stability region is outside the circles as shown in Figure 2.2).

2.3.2 Crank-Nicolson/Adams-Bashforth 2 IMEX

We are interested in finding a second-order convergent IMEX method that is also A-stable. We consider

$$\frac{u_{n+1} - u_n}{\Delta t} = (\epsilon + \nu)\lambda(\frac{u_{n+1} + u_n}{2}) - \epsilon\lambda(\tfrac{3}{2}u_n - \tfrac{1}{2}u_{n-1}), \tag{2.10}$$

which is a Crank-Nicolson second-order (implicit) method for the first part of the Cauchy problem (2.7), and Adams-Bashforth 2 second-order (explicit) for the second part. If ϵ is allowed to be zero we recover Crank-Nicolson:

$$u_n = \Big[\frac{1 + \frac{1}{2}\Delta t\nu\lambda}{1 - \frac{1}{2}\Delta t\nu\lambda}\Big]^n.$$

The characteristic polynomial of method (2.10) is

$$\Pi(r) = (1 - \tfrac{1}{2}\Delta t(\epsilon + \nu)\lambda)r^2 - (1 + \tfrac{1}{2}\Delta t(\epsilon + \nu)\lambda - \tfrac{3}{2}\Delta t\epsilon\lambda)r - \tfrac{1}{2}\epsilon\lambda\Delta t r^0 = 0.$$

This second-degree polynomial has two roots,

$$r_{1,2} = \frac{\big(1 - \Delta t\epsilon\lambda - \frac{1}{2}\Delta t\nu\big) \pm \frac{1}{2}\sqrt{4 + 4\Delta t\lambda\nu + \Delta t^2\nu\lambda^2(\nu - 8\epsilon)}}{2 - \Delta t\lambda\epsilon - \Delta t\lambda\nu}.$$

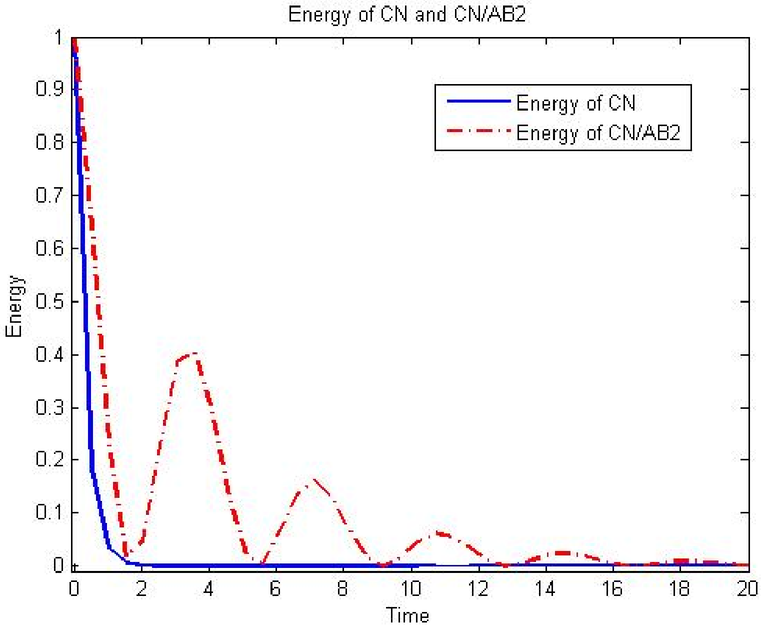

Figure 2.4: Energy decay of CN and CN/AB2 methods

This implies that the analytical solutions of the CN/AB2 scalar method (2.10) can be written as

$$u_n = \gamma_1 r_1^n + \gamma_2 r_2^n.$$

Using initial conditions u_0, u_1 to solve for γ_1, γ_2 gives

$$u_n = +\frac{u_1 + r_2 u_0}{r_1 + r_2} r_1^n + \frac{u_1 + r_1 u_0}{r_1 + r_2} r_2^n.$$

Figure 2.4 shows the convergence of the energy of the solutions of Crank-Nicolson and Crank-Nicolson/Adams Bashforth 2 for initial conditions $u_0 = 1$, $u_1 = .8$ (for CN/AB2), $\lambda = -10000$, $\nu = .001$, $\Delta t = .5$, $\epsilon = .01$ (for the CN/AB2 scheme). Like BE and BE/FE in Figure 2.1, pure Crank-Nicolson converges faster than the mixed method.

For method (2.10) the root-locus curve is

$$\Delta t \lambda \nu = \nu \frac{\rho(\zeta)}{\sigma(\zeta)} = \nu \frac{\zeta^2 - \zeta}{(\epsilon + \nu)(\frac{\zeta^2 + \zeta}{2}) - \epsilon(\frac{3}{2}\zeta - \frac{1}{2})}. \tag{2.11}$$

Figure 2.5 shows the region of stability for CN/AB2 IMEX is similar to that of BE/FE IMEX, and this method is also A-stable.

Figure 2.6 shows the root locus curves corresponding to different values of ϵ. As with BE/FE, the region of stability is growing with ϵ. This plot is similar, except for the size of the stability region, for any choice of $\epsilon \neq 0$.

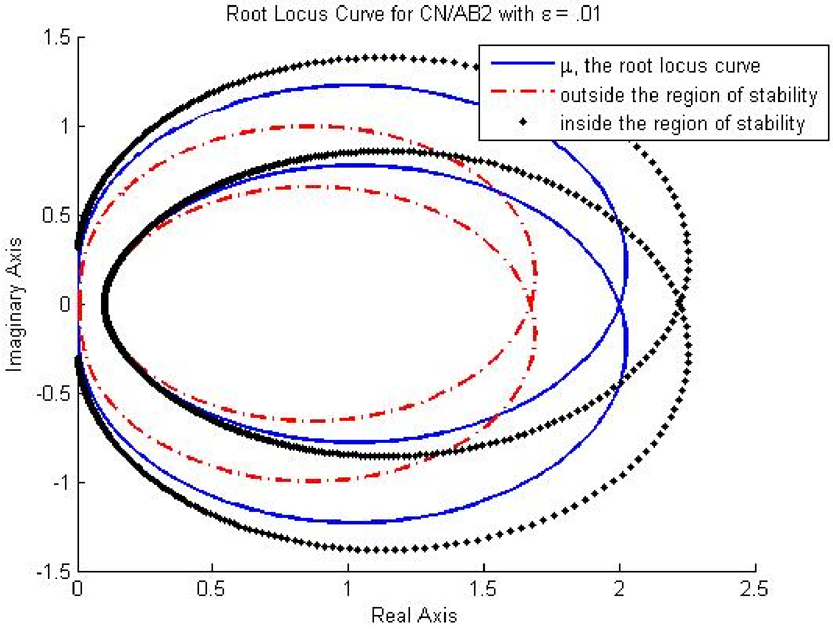

Figure 2.5: Root locus curve denoting the region of A-stability for Crank-Nicolson/Adams Bashforth 2 method with $\epsilon = .01$

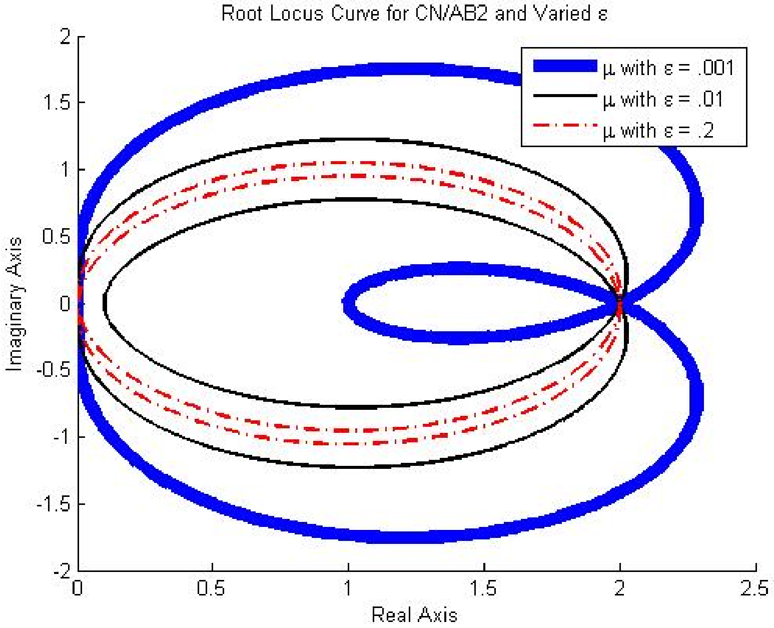

Figure 2.6: Root locus curve denoting the regions of A-stability for Crank-Nicolson/Adams Bashforth 2 method for different values of ϵ

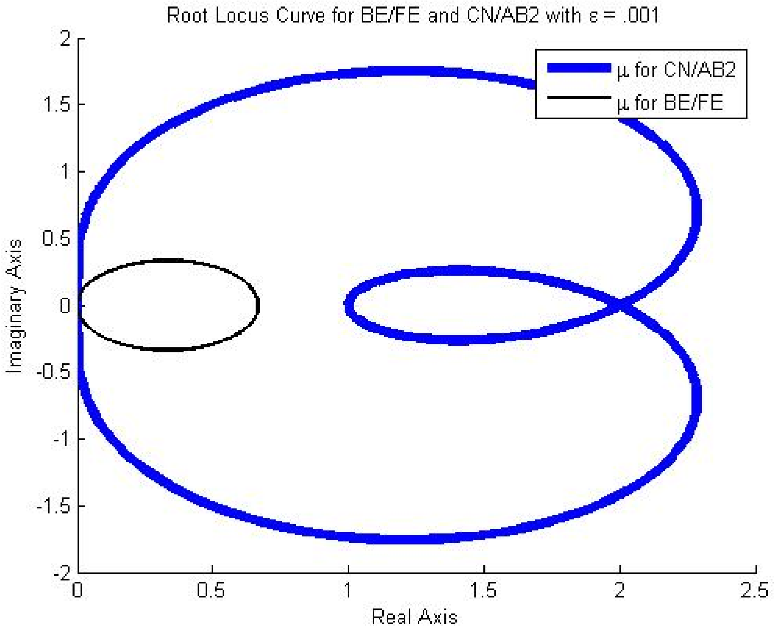

Figure 2.7: The root locus curves for BE/FE and CN/AB2 with $\epsilon = .001$

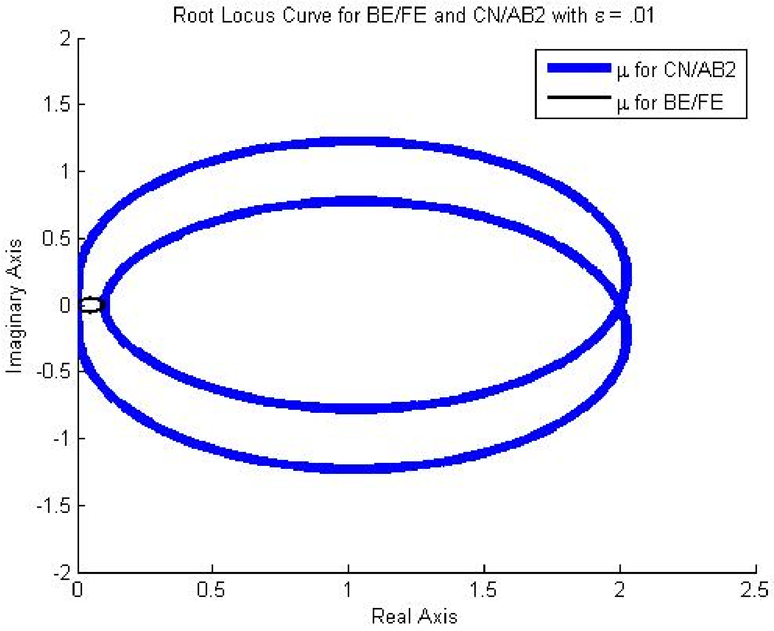

Figure 2.8: The root locus curves for BE/FE and CN/AB2 with $\epsilon = .01$

Figures 2.7 and 2.8 show that for for $\epsilon = .001$ and $\epsilon = .01$ the region of stability for BE/FE is relatively larger than that of CN/AB2, and this holds true for any ϵ. This reflects the fact that using a higher order method (CN/AB2, which is second-order) comes at the cost of a decreased region of stability.

2.4 G-STABILITY

Now let us study the stability of the two aforementioned methods under the lens of a stability definition that is both more complicated and in some cases more useful. Consider the Lipschitz condition

$$Re\langle F(t,y) - F(t,z), y - z\rangle \leq L\|y - z\|^2. \tag{2.12}$$

If the system (2.1) satisfies (2.12) with $L = 0$, then its solutions are contractive. In this case we want to know which linear multi-step methods applied to (2.1) also have contractive solutions, and are thus G-stable as defined in Definition 3 stated below. Let

$$Y_n = (y_{n+k-1}, y_{n+k-2}, ..., y_n)^T$$

be a sequence of numerical solutions to (1.4), and define the G-norm of Y_n to be $\|Y_n\|_G^2 = Y_n^T G Y_n$.

Definition 3. *G-stability (Dahlquist 1975)[2]. A multi-step numerical method is G-stable if the system of ODEs $y' = F(t,y)$ satisfy (2.12) with $L = 0$, and if there exists a symmetric positive-definite matrix (SPD) G, such that*

$$\|Y_{n+1} - \hat{Y}_{n+1}\|_G \leq \|Y_n - \hat{Y}_n\|_G, \tag{2.13}$$

for all steps n and step-sizes $\Delta t > 0$ where $\hat{Y}_n$ is a sequence of solutions for (1.4) that correspond to different initial conditions than Y_n.

Thus, we can use G-stability to test the behavior of a method where the underlying ODE is linear or non-linear, providing that it satisfies the Lipschitz condition with $L = 0$. In the non-linear case, we have G-stability when the difference of the solutions $Y_n - \hat{Y}_n$ are not growing in the G-norm.

2.4.1 G-stability of Scalar Crank-Nicolson/Adams-Bashforth 2

Showing that the method in question is G-stable involves checking that the conditions of the G-stability definition hold. Since in this case our underlying ODE (2.7) is linear, we can consider the Lipschitz and G-norm conditions

$$\langle f(t,y), y)\rangle \leq 0, \quad \|Y_{n+1}\|_G \leq \|Y_n\|_G, \tag{2.14}$$

respectively. It is easy to see that if $\lambda < 0$ and $\nu > 0$ then $\langle \nu\lambda y, y\rangle = \nu\lambda y^2 \leq 0$, and the Lipschitz condition is satisfied. Thus, the task is to see if we can construct a G that satisfies the G-stability definition.

The G-matrix corresponding to this method can be generated directly following along similar lines as the example in [2] or using the proof Dahlquist's equivalence theorem as is done in the preceding subsection (both of which are reproduced for the reader's convenience in Appendix Section A.1.3).

2.4.1.1 Direct Computation of G

First, consider the inner-product

$$\left\langle (\epsilon+\nu)\lambda\left(\frac{u_{n+1}+u_n}{2}\right) - \epsilon\lambda\left(\tfrac{3}{2}u_n - \tfrac{1}{2}u_{n-1}\right), \frac{u_{n+1}-u_n}{\Delta t}\right)\right\rangle \geq 0, \tag{2.15}$$

which holds because they are the RHS and LHS of method (2.10) under consideration. Multiplying by $-\frac{1}{\Delta t}$ and expanding gives

$$-c_1 u_{n+1}^2 - (c_2 - c_1)u_{n+1}u_n + c_2 u_n^2 - c_3 u_{n+1}u_{n-1} + c_3 u_n u_{n-1} \leq 0 \tag{2.16}$$

where

$$c_1 = \tfrac{1}{2}(\epsilon+\nu)\lambda, \qquad c_2 = \tfrac{1}{2}(\epsilon+\nu)\lambda - \tfrac{3}{2}\epsilon\lambda, \qquad c_3 = \tfrac{1}{2}\epsilon\lambda. \tag{2.17}$$

Now consider the equation

$$E = \|Y_{n+1}\|_G^2 - \|Y_n\|_G^2 + \|a_2 y_{n+1} + a_1 y_{n+1} + a_0 y_n\|^2, \qquad a_0, a_1, a_2 \in \mathbb{R}. \tag{2.18}$$

Imposing $E \leq 0$ implies $\|Y_{n+1}\|_G^2 \leq \|Y_n\|_G^2$, since $\|a_2 y_{n+1} + a_1 y_{n+1} + a_0 y_n\|^2 \geq 0$. Let

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \tag{2.19}$$

Thus, if the matrix G produced by matching the coefficients of (2.16) to those of (2.18) is SPD, method (2.10) is G-stable by Definition 3.

Following this approach and letting $g_{12} = g_{21}$ produces the following non-linear system of six equations in six unknowns:

$$\begin{array}{llll} y_{n+1}^2: & -c_1 = g_{11} + a_2^2, & y_{n+1}y_n: & c_1 - c_2 = 2g_{12} + 2a_2a_1 \\ y_n^2: & c_2 = g_{22} - g_{11} + a_1^2 & y_{n+1}y_n: & -c_3 = 2a_2a_0 \\ y_{n-1}^2: & 0 = -g_{22} + a_0^2 & y_ny_{n-1}: & c_3 = -2g_{12} + 2a_1a_0. \end{array} \tag{2.20}$$

Solving this system produces the G-matrix

$$G = \frac{\lambda}{4}\begin{pmatrix} -\epsilon - 2\nu & \epsilon \\ \epsilon & -\epsilon \end{pmatrix}. \tag{2.21}$$

This matrix is symmetric by construction, and it is easy to see that if $\lambda < 0$ all its principle minors have a positive determinant, and therefore this G is positive-definite by Sylvester's Criterion (see Appendix A.1.4). Thus, by Definition 3 this IMEX method is G-stable (as well as A-stable, as demonstrated in the previous section). This, as Dahlquist was finally able to prove in 1978, is not a coincidence.

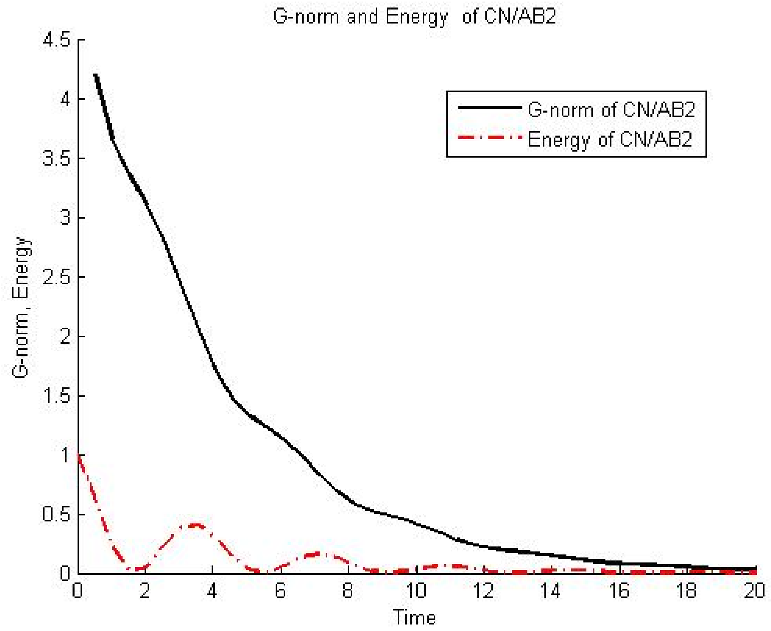

Figure 2.9: G-norm and Energy decay of CN/AB2 method

Figure 2.9 shows the convergence of the G-norm and energy of the solutions of the Crank-Nicolson/Adams Bashforth 2 schemes for initial conditions $u_0 = 1$, $u_1 = .8$, $\epsilon = .01$, $\lambda = -10000$, $\nu = .001$, and $\Delta t = .5$,). Notice that the G-norm is monotonically decreasing, as the G-stability definition requires.

Theorem 2. *(Dahlquist 1978): If a method's generating polynomials ρ, σ, have no common divisor, then the method is G-stable if and only it is A-stable.*

The proof of Theorem 2 is not included here, refer to [2], Chapter V.6 for its details.

2.4.1.2 Constructing G Using Generating Polynomials

A second, more universally applicable method of checking for G-stability is to use the proof of Theorem 2 to construct a G-stability matrix, then check its positive-definiteness. This is a useful technique since it does not involve solving a non-linear system as was required in computing G directly.

The generating polynomials (see Appendix A.1.2) for scalar CN/AB2 IMEX are

$$\rho(\zeta) = \zeta^2 - \zeta, \qquad \sigma(\zeta) = (\epsilon + \nu)(\tfrac{\zeta^2+\zeta}{2}) - \epsilon(\tfrac{3}{2}\zeta - \tfrac{1}{2}).$$

Define the function

$$E(\zeta) = \tfrac{1}{2}(\rho(\zeta)\sigma(\tfrac{1}{\zeta}) + \rho(\tfrac{1}{\zeta})\sigma(\zeta)),$$

which for CN/AB2 is

$$\begin{aligned}
E(\zeta) &= \tfrac{1}{2}((\zeta^2-\zeta)((\alpha+\nu)(\tfrac{\frac{1}{\zeta^2}+\frac{1}{\zeta}}{2}) - \epsilon(\tfrac{3}{2}\zeta-\tfrac{1}{2}))(\tfrac{1}{\zeta}) + (\tfrac{1}{\zeta^2}-\tfrac{1}{\zeta})((\epsilon+\nu)(\tfrac{\zeta^2+\zeta}{2}) - \epsilon(\tfrac{3}{2}\zeta-\tfrac{1}{2})))\\
&= \frac{\epsilon(\zeta-1)^2}{4\zeta^2}\\
&= \Big[\frac{\sqrt{\epsilon}}{2}[(\zeta-1)^2]\Big]\Big[\frac{\sqrt{\epsilon}}{2}][(\tfrac{1}{\zeta}-1)^2]\Big]\\
&= a(\zeta)a(\tfrac{1}{\zeta}).
\end{aligned}$$

Define the function $P(\zeta,\omega) = \frac{1}{2}(\rho(\zeta)\sigma(\omega)+\rho(\omega)\sigma(\zeta)) - a(\zeta)a(\omega)$, which with some simplification and factoring becomes

$$\begin{aligned}
P(\zeta,\omega) &= \tfrac{1}{4}[-\epsilon(\zeta-1)^2(\omega-1)^2 + \epsilon(\omega-1)\omega - \zeta(\alpha(2\nu-4\epsilon)\omega + 3\alpha\omega^2) + \zeta^2(\epsilon - 3\epsilon\omega + 2(\epsilon+\nu)\omega^2))]\\
&= (\zeta\omega-1)\Big(\frac{(\epsilon+2\nu)}{4}\zeta\omega - \frac{\epsilon}{4}\zeta - \frac{\epsilon}{4}\omega + \frac{\epsilon}{4}\Big)\\
&= (\zeta\omega-1)(g_{11}\zeta\omega - g_{12}\zeta - g_{21}\omega + g_{22}).
\end{aligned}$$

This yields the matrix

$$G = \frac{1}{4}\begin{pmatrix} \epsilon+2\nu & -\epsilon \\ -\epsilon & \epsilon \end{pmatrix}, \tag{2.22}$$

which is SPD. Multiplying (2.22) by the positive constant $-\lambda$ gives the same result as computing G directly as in the previous section.

Somewhat surprisingly, the method's stability properties are driven by its implicit part, and this is consistent with the idea that separation of linear parts A and C as done in (1.1) will lead to a more stable regime.

3.0 CRANK-NICOLSON/ADAMS BASHFORTH 2 IMEX METHOD: UNCONDITIONAL STABILITY, CONSISTENCY, AND CONVERGENCE

Let us now return to considering a system of ODEs (1.1) under assumptions (1.2).

The Crank-Nicolson/Adams Bashforth 2 method proposed in Section 1 is a member of a broader family of three level, second order time-stepping schemes [4]:

$$\begin{aligned}
&\frac{(\theta+\frac{1}{2})u_{n+1}-2\theta u_n+(\theta-\frac{1}{2})u_{n-1}}{\Delta t}\\
&+(A-C)^{\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\theta u_{n+1}+\big((1-\theta)A-(\theta+1)C\big)(A-C)^{-\frac{1}{2}}u_n+C(A-C)^{-\frac{1}{2}}\theta u_{n-1})\Big)\\
&\quad+B(\mathcal{E}_{n+\theta})(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\theta u_{n+1}+\big((1-\theta)A-(\theta+1)C\big)(A-C)^{-\frac{1}{2}}u_n+C(A-C)^{-\frac{1}{2}}\theta u_{n-1})\Big)\\
&=f_{n+\theta},
\end{aligned}\tag{3.1}$$

where $\mathcal{E}_{n+\theta}$ is an implicit or explicit second order approximation of u_{n+1} and $\theta\in[\frac{1}{2},1]$. This work will focus on the case of $\theta=\frac{1}{2}$, and take $\mathcal{E}_{n+\theta}=\frac{3}{2}u_n-\frac{1}{2}u_{n-1}$, which is

$$\begin{aligned}
&\frac{u_{n+1}-u_n}{\Delta t}+(A-C)^{\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\tfrac{1}{2}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\\
&\quad+B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\Big)\\
&=f_{n+\frac{1}{2}}.
\end{aligned}\tag{3.2}$$

Of particular interest are the numerical stability and convergence properties of (3.2).

3.1 STABILITY ANALYSIS

Here we will consider the stability properties of the method (3.2). However, unlike the examples in the previous chapter, the system under consideration will have be d-dimensional and be in terms of non-commmuting coefficient matrices A, B, C, where B is allowed to be nonlinear. This added complexity is worthwhile since many processes have highly non-linear behavior, but it comes at the cost of greatly complicating stability analysis.

3.1.1 Well-Posedness of the Problem

Unlike the Cauchy problems analyzed in Section 2.2, nonlinearity will prohibit the Lipschitz condition (2.2) from holding globally. This is not trivial, since this means none of the stability theory for the underlying problem developed in the Chapter 2 will necessarily hold. The system is, however, locally stable.

Theorem 3. *Local Stability of Nonlinear System (1.1). Under assumptions (1.2),*

$$u'(t) + Au(t) - Cu(t) + B(u)u(t) = f(t)$$

is stable for all $t \in \mathcal{I} = [0, T]$, for all finite T.

Proof. The proof, which is taken from Anitescu et al. [1], is included in Appendix A.1.5. □

By Theorem 3 problem (1.1) is well-behaved locally, which is to say that its solutions $u(t)$ do not blow up on $t \in \mathcal{I}$. This allows us to conclude that the problem is sufficiently well-posed in at least a local sense, and we can thus discuss stability of a numerical method for solving it.

3.1.2 Transformation of the Method

In [1] the numerical solution u_n provided by the BE/FE IMEX method (1.5) is shown to be nonincreasing,

$$\|u_{n+1}\|_E \leq \|u_n\|_E, \qquad E = I + \Delta t C,$$

in the energy norm E, and this condition is sufficient to conclude the method is unconditionally stable. The aim of this section will be analogous in nature. Borrowing heavily from the G-stability concepts developed in Section 2.4, given an appropriately chosen transformation of the method, it can be proved that the numerical solutions are decreasing at each time step in the G-norm, and thus the method is unconditionally stable on the interval of interest $\mathcal{I}$.

Since $B(u)$ is assumed to be skew-symmetric, multiplying method (3.2) from the left by the vector

$$\Big[(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big]^T \tag{3.3}$$

will cause the nonlinear term

$$B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)$$

to disappear, leaving

$$\begin{aligned}
&\Big\langle (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big),\\
&\frac{u_{n+1}-u_n}{\Delta t} + (A-C)^{\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\tfrac{1}{2}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle\\
&= \Big\langle (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big), f_{n+\frac{1}{2}}\Big\rangle.
\end{aligned}$$

By the properties of the inner-product and Euclidian norm, this can be rearranged as

$$\begin{aligned}&\frac{1}{\Delta t}\Big\langle u_{n+1}-u_n\,,\,(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\tfrac{1}{2}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle\\&+\Big\|\Big(A(A-C)^{-\frac{1}{2}}\tfrac{1}{2}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\|_2^2\\&=\Big\langle f_{n+\frac{1}{2}}\,,\,(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\tfrac{1}{2}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle.\end{aligned}\tag{3.4}$$

Focusing on the first line of (3.4), the goal will be to simplify the transformed method into positive pieces using the G-norm to group and compare terms, as was done in the G-stability examples in Section 2.4.1. The G-stability matrix is calculated using Dahlquist's equivalence theorem as demonstrated in Section 2.4.1.2. Method (3.2) and its corresponding characteristic polynomials yield matrix

$$G=\begin{pmatrix}(A-C)^{-\frac{1}{2}}(\tfrac{1}{2}A-\tfrac{1}{4}C)(A-C)^{-\frac{1}{2}} & -(A-C)^{-\frac{1}{2}}(\tfrac{1}{4}C)(A-C)^{-\frac{1}{2}}\\ -(A-C)^{-\frac{1}{2}}(\tfrac{1}{4}C)(A-C)^{-\frac{1}{2}} & (A-C)^{-\frac{1}{2}}(\tfrac{1}{4}C)(A-C)^{-\frac{1}{2}}\end{pmatrix}.\tag{3.5}$$

Referring to the G-stability examples in Section 2.4.1, taking $A=-(\alpha+\nu)\lambda$, $C=-\alpha\lambda$, and ignoring $(A-C)^{-\frac{1}{2}}$ terms, G matches matrix (2.21).

3.1.3 G-norm

We now wish to check that G is a symmetric positive-definite matrix so it can be used to finish putting the transformed method (3.4) into norms and positive terms.

3.1.3.1 Symmetry of G The G matrix defined in (3.5) is a 2×2 block-partitioned matrix with submatrices of size $d\times d$. Since the off-diagonal blocks are the same, if each of the four blocks is symmetric this is sufficient to conclude G is symmetric also.

Since A and C are both symmetric by assumptions (1.2), adding or subtracting positive-definite multiples of them also results in symmetric matrices. $(A-C)^{-\frac{1}{2}}$ is symmetric by Lemma 4. Thus

$$(A-C)^{-\frac{1}{2}}(\tfrac{1}{2}A-\tfrac{1}{4}C)(A-C)^{-\frac{1}{2}}=\big[(A-C)^{-\frac{1}{2}}(\tfrac{1}{2}A-\tfrac{1}{4}C)(A-C)^{-\frac{1}{2}}\big]^T.$$

Making the same argument, the other three blocks are symmetric as well, and so G is also.

3.1.3.2 Positive-Definiteness of G If G is PD the following will be strictly positive for any choice of $d\times 1$ vectors u, v not equal to zero:

$$\Big\|\begin{bmatrix}u\\v\end{bmatrix}\Big\|_G^2=\Big\langle\begin{bmatrix}u\\v\end{bmatrix},G\begin{bmatrix}u\\v\end{bmatrix}\Big\rangle.\tag{3.6}$$

Lemma 1. *G is a positive-definite matrix.*

Proof. Expanding equation (3.6) gives

$$\begin{aligned}\Big\langle\begin{bmatrix}u\\v\end{bmatrix},G\begin{bmatrix}u\\v\end{bmatrix}\Big\rangle=\tfrac{1}{4}\Big[&u^T[(A-C)^{-\frac{1}{2}}(2A-C)(A-C)^{-\frac{1}{2}}]u-u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]v\\&-v^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u+v^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]v\Big].\end{aligned}\tag{3.7}$$

Subtracting and adding $\frac{1}{4}\big[u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u\big]$ to (3.7) and using the fact that

$$u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]v = \big[u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]v\big]^T = v^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u$$

gives

$$\Big\langle \begin{bmatrix}u\\v\end{bmatrix}, G\begin{bmatrix}u\\v\end{bmatrix}\Big\rangle = \frac{1}{4}\Big[u^T[(A-C)^{-\frac{1}{2}}(2A-C)(A-C)^{-\frac{1}{2}}]u - u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u$$
$$+ u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u - 2u^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]v + v^T[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]v\Big].$$

The first two terms can be combined and simplified to be

$$\tfrac{1}{4}u^T[(A-C)^{-\frac{1}{2}}(2A-2C)(A-C)^{-\frac{1}{2}}]u = \tfrac{1}{2}u^Tu.$$

Take $F = (A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}$. $(A-C)^{-1}$ is SPD by Lemma 3 and by Lemma 4 so is $(A-C)^{-\frac{1}{2}}$. These results allow us to conclude that $F = F^T$, and the remaining terms can be factored as

$$\begin{aligned}
&\tfrac{1}{4}[u^T(F^{\frac{1}{2}})^TF^{\frac{1}{2}}u - 2u^T(F^{\frac{1}{2}})^TF^{\frac{1}{2}}v + v^T(F^{\frac{1}{2}})^TF^{\frac{1}{2}}v]\\
&= \tfrac{1}{4}\langle F^{\frac{1}{2}}u - F^{\frac{1}{2}}v, F^{\frac{1}{2}}u - F^{\frac{1}{2}}v\rangle\\
&= \tfrac{1}{4}\|F^{\frac{1}{2}}u - F^{\frac{1}{2}}v\|_2^2 \geq 0.
\end{aligned}$$

Thus we can conclude that $\Big\|\begin{bmatrix}u\\v\end{bmatrix}\Big\|_G^2 \geq \frac{1}{2}u^Tu > 0$ for all non-zero u, v, and G is positive-definite. □

3.1.4 Unconditional Stability Result

As proved above, the matrix G is symmetric and positive-definite, and therefore the expression defined in (3.6) is a G-norm.

Lemma 2. *Let u_n satisfy (3.2) for all $n \in \{2, \ldots, \frac{T}{\Delta t}\}$. Then*

$$\frac{1}{\Delta t}\Big\langle u_{n+1} - u_n\,,\,(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle$$
$$= \frac{1}{\Delta t}\Big\|\begin{bmatrix}u_{n+1}\\u_n\end{bmatrix}\Big\|_G^2 - \frac{1}{\Delta t}\Big\|\begin{bmatrix}u_n\\u_{n-1}\end{bmatrix}\Big\|_G^2 + \frac{1}{4\Delta t}\|\big(u_{n+1}-2u_n+u_{n-1}\big)\|_F^2. \tag{3.8}$$

Proof. Expanding the LHS of (3.8) gives

$$\begin{aligned}
&\frac{1}{\Delta t}\Big\langle u_{n+1} - u_n\,,\,(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle\\
&= \frac{1}{\Delta t}\Big[u_{n+1}^T(A-C)^{-\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + u_{n+1}^T(A-C)^{-\frac{1}{2}}(\tfrac{1}{2}A - \tfrac{3}{2}C)(A-C)^{-\frac{1}{2}}u_n\\
&\quad + u_{n+1}^T(A-C)^{-\frac{1}{2}}\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1} - u_n^T(A-C)^{-\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}\\
&\quad - u_n^T(A-C)^{-\frac{1}{2}}(\tfrac{1}{2}A - \tfrac{3}{2}C)(A-C)^{-\frac{1}{2}}u_n - u_n^T(A-C)^{-\frac{1}{2}}\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\Big]
\end{aligned}$$

Next, note that expanding each piece of the RHS of the previous equation and omittting $\frac{1}{\Delta t}$ gives

$$u_{n+1}^T(A-C)^{-\frac{1}{2}}\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}n_{n+1} = u_{n+1}^T\tfrac{1}{4}\Big[(A-C)^{-\frac{1}{2}}(2A-C)(A-C)^{-\frac{1}{2}}]u_{n+1}$$

$$+ u_{n+1}^T \tfrac{1}{4}\Big[(A-C)^{-\frac{1}{2}} C (A-C)^{-\frac{1}{2}}] u_{n+1}$$

$$\begin{aligned}
u_{n+1}^T[(A-C)^{-\frac{1}{2}}(-\tfrac{3}{2}C)(A-C)^{-\frac{1}{2}}]u_n &= u_{n+1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(-C)(A-C)^{-\frac{1}{2}}]u_n \\
&+ u_n^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(-C)(A-C)^{-\frac{1}{2}}]u_{n+1} + u_{n+1}^T[\tfrac{1}{4}(A-C)^{-\frac{1}{2}}(-2C)(A-C)^{-\frac{1}{2}}]u_n \\
&+ u_n^T[(A-C)^{-\frac{1}{2}}(-2C)(A-C)^{-\frac{1}{2}}]u_{n+1}
\end{aligned}$$

$$\begin{aligned}
u_{n+1}^T[(A-C)^{-\frac{1}{2}}\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}]u_{n-1} &= u_{n+1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_{n-1} \\
&+ u_{n-1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_{n+1}
\end{aligned}$$

$$\begin{aligned}
u_n^T[(A-C)^{-\frac{1}{2}}(\tfrac{3}{2}C-\tfrac{1}{2}A)(A-C)^{-\frac{1}{2}}]u_n &= u_n^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_n \\
&- u_n^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(2A-C)(A-C)^{-\frac{1}{2}}]u_n + u_n^T \tfrac{1}{4}[(-2)(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_n \\
&+ u_n^T \tfrac{1}{4}[(-2)(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_n
\end{aligned}$$

$$\begin{aligned}
u_n^T[(A-C)^{-\frac{1}{2}}(-\tfrac{1}{2}C)(A-C)^{-\frac{1}{2}}]u_{n-1} &= u_n^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_{n-1} \\
&+ u_{n-1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}]u_n + u_n^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(-2C)(A-C)^{-\frac{1}{2}}]u_{n-1} \\
&+ u_{n-1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(-2C)(A-C)^{-\frac{1}{2}}]u_n
\end{aligned}$$

$$\begin{aligned}
u_{n-1}^T(A-C)^{-\frac{1}{2}}(0)(A-C)^{-\frac{1}{2}}u_{n-1} &= u_{n-1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(-C)(A-C)^{-\frac{1}{2}}]u_{n-1} \\
&+ u_{n-1}^T \tfrac{1}{4}[(A-C)^{-\frac{1}{2}}(C)(A-C)^{-\frac{1}{2}}]u_{n-1}.
\end{aligned}$$

Summing these exapanded RHS terms and replacing $\frac{1}{\Delta t}$ gives

$$\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_{n+1}\\ u_n\end{bmatrix}\Big\|_G^2 - \frac{1}{\Delta t}\Big\|\begin{bmatrix}u_n\\ u_{n-1}\end{bmatrix}\Big\|_G^2 + \frac{1}{4\Delta t}\|\big(u_{n+1}-2u_n+u_{n-1}\big)\|^2_{(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}},$$

and since $F = (A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}$ we have the result. □

3.1.4.1 Energy Equality To see that the method is unconditionally stable consider the following energy equality, which holds for u_0, u_1 given inital conditions at all time steps $n = 1$ through $N-1$:

$$\begin{aligned}
&\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_N\\ u_{N-1}\end{bmatrix}\Big\|_G^2 + \frac{1}{4\Delta t}\sum_{n=1}^{N-1}\|u_{n+1}-2u_n+u_{n-1}\|_F^2 \\
&+ \sum_{n=1}^{N-1}\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\big)\Big\|^2 \\
= &\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_1\\ u_0\end{bmatrix}\Big\|_G^2
\end{aligned}$$

$$+\sum_{n=1}^{N-1}\Big\langle f_{n+\frac{1}{2}}, (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle. \tag{3.9}$$

Proof. The method (3.2) multiplied by the $d\times 1$ vector (3.3) at $n=1$ is

$$\begin{aligned}&\frac{1}{\Delta t}\Big\langle u_2-u_1\,,\,(A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_2+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_1+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_0)\Big)\Big\rangle\\&+\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_2+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_1+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_0\Big\|^2\\&=\Big\langle f_{\frac{3}{2}}, (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle.\end{aligned}$$

By Lemma 2 this becomes

$$\begin{aligned}&\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_2\\u_1\end{bmatrix}\Big\|_G^2-\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_1\\u_0\end{bmatrix}\Big\|_G^2+\frac{1}{4\Delta t}\big\|(u_2-2u_1+u_0)\big\|_F^2\\&\quad+\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_2+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_1+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_0\Big\|^2\\&=\Big\langle f_{\frac{3}{2}}, (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_2+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_1+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_0)\Big)\Big\rangle.\end{aligned}$$

By the same argument, at $n=2$ we have

$$\begin{aligned}&\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_3\\u_1\end{bmatrix}\Big\|_G^2-\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_2\\u_1\end{bmatrix}\Big\|_G^2+\frac{1}{4\Delta t}\big\|(u_3-2u_2+u_1)\big\|_F^2\\&\quad+\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_3+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_2+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_1\Big\|^2\\&=\Big\langle f_{\frac{3}{2}}, (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_3+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_2+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_1)\Big)\Big\rangle.\end{aligned}$$

Solving this for $\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_2\\u_1\end{bmatrix}\Big\|_G^2$ and plugging in to the previous equation gives

$$\begin{aligned}&\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_3\\u_2\end{bmatrix}\Big\|_G^2+\frac{1}{4\Delta t}\sum_{n=1}^{2}\big\|u_{n+1}-2u_n+u_{n-1}\big\|_F^2\\&\quad+\sum_{n=1}^{2}\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\Big\|^2-\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_1\\u_0\end{bmatrix}\Big\|_G^2\\&=\sum_{n=1}^{2}\Big\langle f_{n+\frac{1}{2}}, (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big)\Big\rangle.\end{aligned}$$

Repeating this calculation $N-2$ more times yields the result. □

Notice Lemma 2 and the Energy Equality immediately imply G-stability in the case of $f_{n+\frac{1}{2}}=0$. That is, if the the energy source (forcing function) is removed , stability of the method requires that the energy in the system decays to zero. To see that this is so, note that

$$\begin{aligned}&\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_{n+1}\\u_n\end{bmatrix}\Big\|_G^2-\frac{1}{\Delta t}\Big\|\begin{bmatrix}u_n\\u_{n-1}\end{bmatrix}\Big\|_G^2+\frac{1}{4\Delta t}\big\|(u_{n+1}-2u_n+u_{n-1})\big\|_F^2\\&+\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1}+\big(\tfrac{1}{2}A-\tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n+\tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big\|^2=0\end{aligned}$$

holds for all $n \in \{1, N-1\}$. Further,

$$\Big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1})\Big\|^2 \geq 0,$$

$$\frac{1}{4\Delta t}\big\|\big(u_{n+1} - 2u_n + u_{n-1}\big)\big\|_F^2 \geq 0,$$

since F is a positive-definite matrix. Thus we have

$$\Big\|\begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}\Big\|_G^2 \leq \Big\|\begin{bmatrix} u_n \\ u_{n-1} \end{bmatrix}\Big\|_G^2.$$

Since this result is independent of the the size of time-step Δt, we have unconditional stablility.

3.1.4.2 Energy Estimate In the case of $f(t) \neq 0$ for some $t \in \mathcal{I}$, the effect of $f_{n+\frac{1}{2}}$ on the energy equality (3.9) is ambiguous. For this case we can derive the following energy estimate to bound the effect of f on the energy in the system:

$$\begin{aligned} &\|u_N\|^2 + \tfrac{1}{2}\sum_{n=1}^{N-1}\big\|u_{n+1} - 2u_n + u_{n-1}\big\|_F^2 \\ &\quad + \Delta t\sum_{n=1}^{N-1}\big\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\big\|^2 \\ &\leq 2\Big\|\begin{bmatrix} u_1 \\ u_0 \end{bmatrix}\Big\|_G^2 + \Delta t\sum_{n=1}^{N-1}\big\|(A-C)^{-\frac{1}{2}}f_{n+\frac{1}{2}}\big\|^2. \end{aligned} \tag{3.10}$$

Proof. The energy estimate (3.9) is a consequence of Lemma 2 and the proof of Lemma 1. Using the Cauchy-Schwarz and Young inequalities, we have that the forcing term in (3.9) can be bounded as follows:

$$\begin{aligned} &\Big\langle f_{n+\frac{1}{2}}, (A-C)^{-\frac{1}{2}}\Big(\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\Big)\Big\rangle \\ &\leq \frac{1}{2}\|(A-C)^{-\frac{1}{2}}f_{n+\frac{1}{2}}\|^2 + \frac{1}{2}\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\|^2, \end{aligned}$$

which gives

$$\begin{aligned} &\Big\|\begin{bmatrix} u_N \\ u_{N-1} \end{bmatrix}\Big\|_G^2 + \frac{1}{4}\sum_{n=1}^{N-1}\|u_{n+1} - 2u_n + u_{n-1}\|_F^2 \\ &\quad + \frac{\Delta t}{2}\sum_{n=1}^{N-1}\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\|^2 \\ &\leq \Big\|\begin{bmatrix} u_1 \\ u_0 \end{bmatrix}\Big\|_G^2 + \frac{\Delta t}{2}\sum_{n=1}^{N-1}\|(A-C)^{-\frac{1}{2}}f_{n+\frac{1}{2}}\|^2. \end{aligned}$$

Using the conclusion of the proof of Lemma 1 we obtain

$$\begin{aligned} &\|u_N\|^2 + \frac{1}{2}\sum_{n=1}^{N-1}\|u_{n+1} - 2u_n + u_{n-1}\|_F^2 \\ &\quad + \Delta t\sum_{n=1}^{N-1}\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}u_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}u_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}u_{n-1}\|^2 \\ &\leq +2\Big\|\begin{bmatrix} u_1 \\ u_0 \end{bmatrix}\Big\|_G^2 + \Delta t\sum_{n=1}^{N-1}\|(A-C)^{-\frac{1}{2}}f_{n+\frac{1}{2}}\|^2, \end{aligned}$$

which by induction completes the proof. □

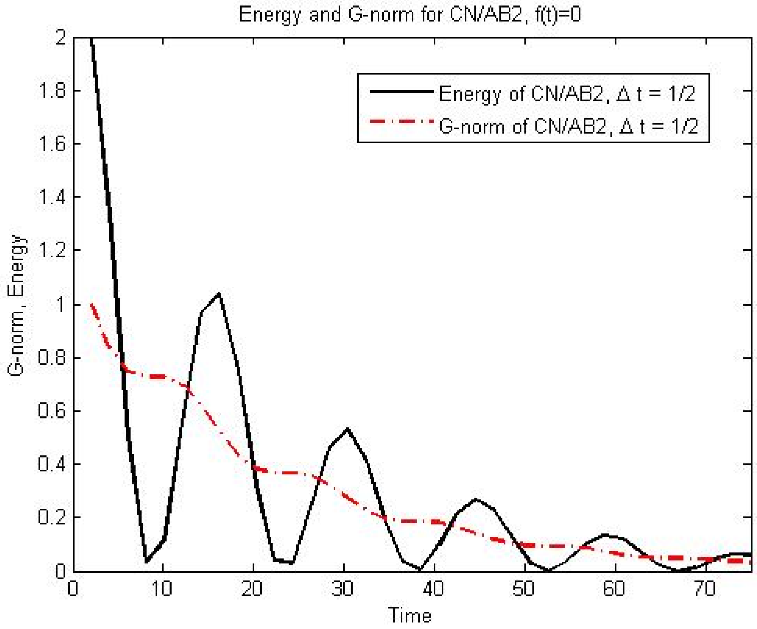

Figure 3.1: Energy and G-norm convergence of CNAB2, d=2, $f(t) = 0$

3.2 NUMERICAL EXPERIMENTS

To demonstrate that the proposed CN/AB2 method (3.2) is unconditionally stable consider the following numerical experiments.

3.2.1 Experiment 1

Take

$$A = (\epsilon + \nu)\begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}, \qquad C = \epsilon\begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}, \qquad B(u) = \sqrt{u_1^2 + u_2^2}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad f(t) = 0,$$

where u_1 and u_2 denote the first and second elements of the vector u (not the time step). Let $\nu = .001$, and initial conditions be $u_1 = u_2 = [1, 1]^T$.

Figure 3.1 shows the convergence of the energy and G-norms for CN/AB2 (3.2) with $d = 2$, and $\epsilon = .01$. Notice that as in the scalar example in Chapter 2, Figure 2.9, the G-norm decreases monotonically, even though the energy of the solution does not.

Figure 3.2 shows the convergence of the G-norm for CN/AB2 (3.2) with $d = 2$ and $f(t) = 0$ for various Δt. As expected from the theory, the smaller Δt is the faster the method converges. Nonethless, even when Δt is taken to be very large ($\Delta t = 5$), the method's solutions still converge in the G-norm, which illustrates the unconditional stability result derived in the previous section.

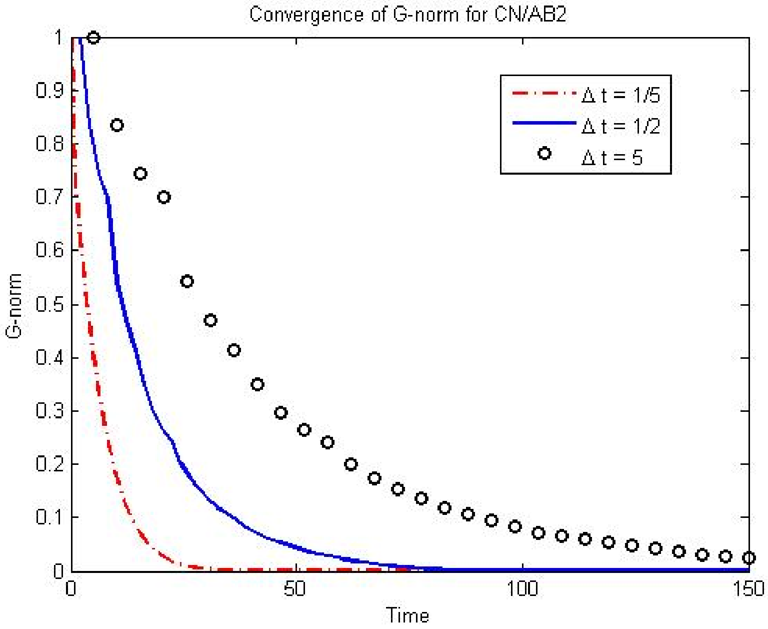

Figure 3.2: G-norm for for various Δt

3.2.2 Experiment 2

Now we relax the restrictions on C and $f(t)$ and consider the case where C is not diagonal, and $f(t) \neq 0$ for some t. Taking

$$C = \epsilon \begin{pmatrix} 100 & -10 \\ -10 & 100 \end{pmatrix}, \qquad f(t) = e^{-t},$$

implies that

$$A - C = \begin{pmatrix} 100\nu & 10\epsilon \\ 10\epsilon & 100\nu \end{pmatrix}.$$

Recalling that $A - C$ is required to be positive-definite, by Sylvester's Criterion (Appendix A.1.4) we see that $A-C$ is positive-definite when $10,000\nu^2 - 100\epsilon^2 > 0$, which is satisfied by all ν and ϵ such that $10\nu > \epsilon$. $\epsilon = .009$, satisfies the inequality for $\nu = .001$, and Figure 3.3 shows the expriment with the revised conditions for this choice of ϵ. Notice that this is nearly the same plot as the previous figures with $f(t) = 0$ and C diagonal except the amplitute of the initial osccilations are greater.

Figure 3.4 shows the energy and G-norms for various Δt under the new conditions on C and $f(t)$. Notice that when $\Delta t = \frac{1}{5}$ the G-norm is not monotically decreasing until $t \simeq 5$. This is due to the forcing function $f(t) = e^{-t}$, and is an illustration of the Energy Estimate (3.10), which says that the solutions in the G-norm are bounded by solutions at previous steps *and* a norm depending on the forcing function $f(t)$.

If we take $\epsilon = .015$, $A - C$ is no longer positive-definite when $\nu = .01$. Figure 3.5 shows the energy and G-norm of the CN/AB2 solutions for this value of ϵ. Given that $A - C$ is no longer positive-definite, the

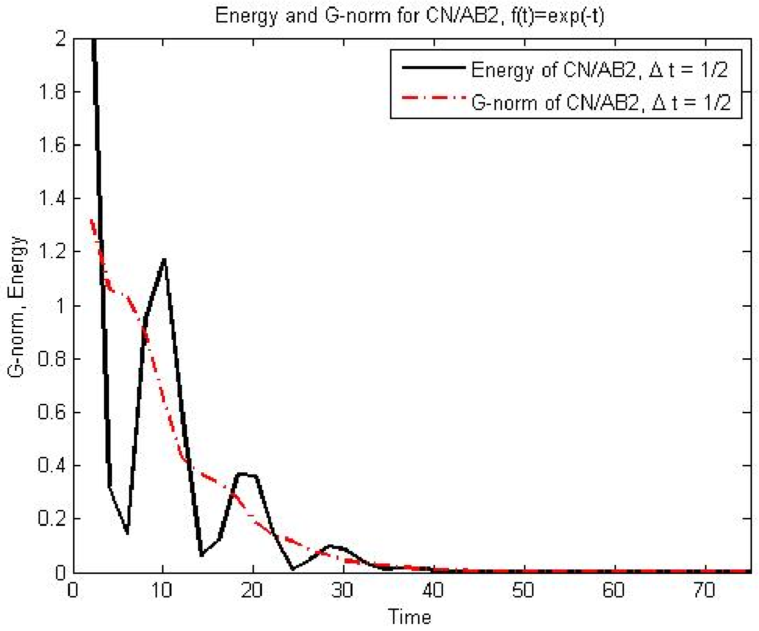

Figure 3.3: Energy and G-norm convergence of CN/AB2, d=2, C not diagonal, $f(t) = e^{-t}$

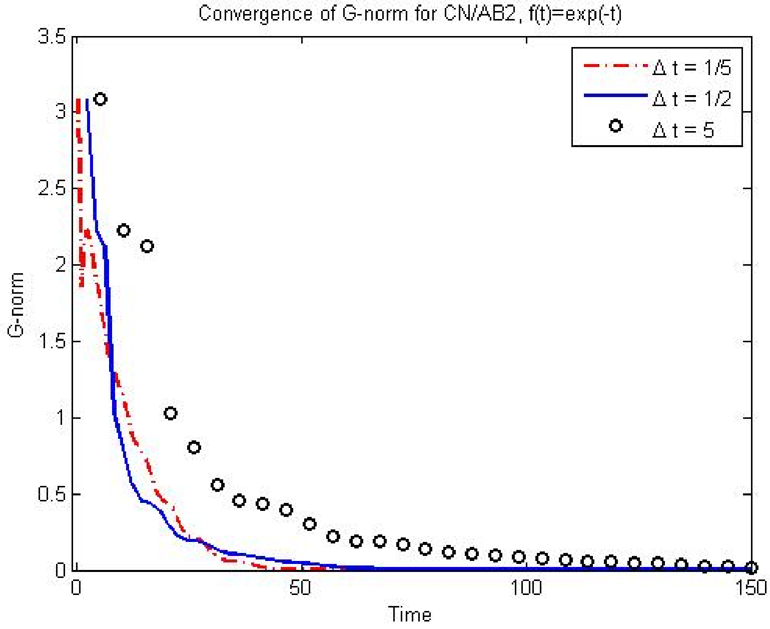

Figure 3.4: Energy and G-norm convergence of CN/AB2, d=2, C not diagonal, $f(t) = e^{-t}$, for various Δt

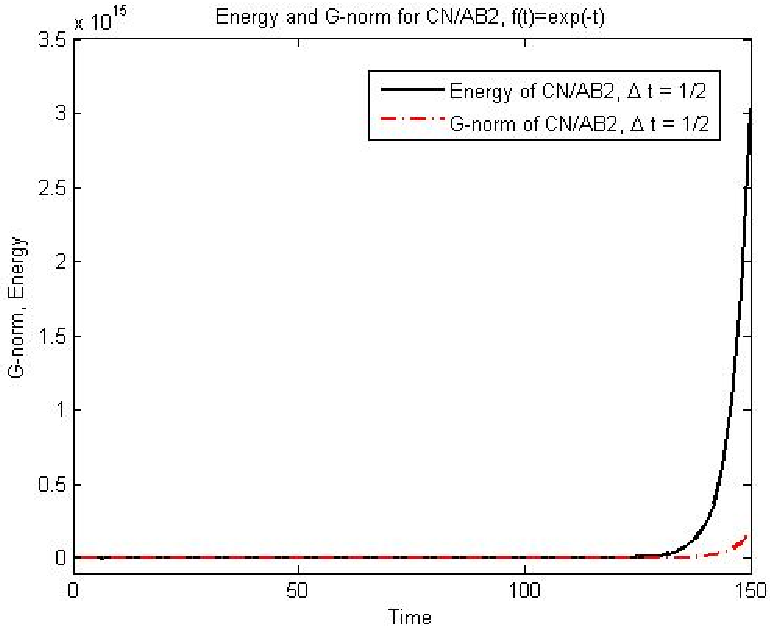

Figure 3.5: Energy and G-norm divergence of CN/AB2, d=2, C not diagonal, $f(t) = e^{-t}$

divergence at the end of the interval is not surprising.

3.3 CONSISTENCY AND CONVERGENCE

The local truncation error (LTE) of a method is the error that arises from substituting the exact solution (denoted $u(t_n)$ for exact solution at point t_n) into the method. With $\mathcal{E}^t_{n+\frac{1}{2}} = \frac{3}{2}u(t_n) - \frac{1}{2}u(t_{n-1})$, corresponding to $\mathcal{E}_{n+\frac{1}{2}}$, the explicit approximation of $u(t_{n+\frac{1}{2}})$, the local truncation error for method (3.2) is

$$\begin{aligned}\tau_{n+1}(\Delta t) = {}& \frac{u(t_{n+1}) - u(t_n)}{\Delta t} \\ & + (A-C)^{\frac{1}{2}} A\big(\tfrac{1}{2}(A-C)^{-\frac{1}{2}} u(t_{n+1}) + \tfrac{1}{2}(A-C)^{-\frac{1}{2}} u(t_n)\big) \\ & - (A-C)^{\frac{1}{2}} C\big(\tfrac{3}{2}(A-C)^{-\frac{1}{2}} u(t_n) - \tfrac{1}{2}(A-C)^{-\frac{1}{2}} u(t_{n-1})\big) \\ + B(\mathcal{E}^t_{n+\frac{1}{2}})(A-C)^{-\frac{1}{2}} & \Big(A(A-C)^{-\frac{1}{2}}(\tfrac{1}{2}(A-C)^{-\frac{1}{2}} u(t_{n+1}) + \tfrac{1}{2}(A-C)^{-\frac{1}{2}} u(t_n))\Big) - f(t_{n+\frac{1}{2}}).\end{aligned} \tag{3.11}$$

To remove the autonomous function $f(t_{n+\frac{1}{2}})$ take the ODE (1.1) at $(t_{n+\frac{1}{2}})$,

$$u'(t_{n+\frac{1}{2}}) + (A-C)u(t_{n+\frac{1}{2}}) + B(u(t_{n+\frac{1}{2}}))u(t_{n+\frac{1}{2}}) - f(t_{n+\frac{1}{2}}) = 0$$

and subtract it from (3.11). This, after some rearranging of terms, gives

$$
\begin{aligned}
\tau_{n+1}(\Delta t) = {} & \frac{u(t_{n+1}) - u(t_n)}{\Delta t} - u'(t_{n+\frac{1}{2}}) \\
& + (A{-}C)^{\frac{1}{2}} A\big(\tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n+1}) + \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_n) - (A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}})\big) \\
& - (A{-}C)^{\frac{1}{2}} C\big(\tfrac{3}{2}(A{-}C)^{-\frac{1}{2}} u_n - \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u_{n-1} - (A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}})\big) \\
& + \Big(B(\mathcal{E}^t_{n+\frac{1}{2}}) - B(u(t_{n+\frac{1}{2}}))\Big)(A{-}C)^{-\frac{1}{2}} \Big(A(A{-}C)^{-\frac{1}{2}}(\tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n+1}) {+} \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_n)) \\
& \qquad\qquad - C\big(\tfrac{3}{2}(A{-}C)^{-\frac{1}{2}} u(t_n) {-} \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n-1})\big)\Big) \\
& + B(u(t_{n+\frac{1}{2}}))\Big[(A{-}C)^{-\frac{1}{2}} \Big(A(\tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n+1}) {+} \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_n)) \\
& \qquad\qquad - C\big(\tfrac{3}{2}(A{-}C)^{-\frac{1}{2}} u(t_n) {-} \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n-1})\big)\Big) - (A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}})\Big].
\end{aligned}
$$

The consistency and convergence of the general class of methods (3.1) for $\theta \in [\frac{1}{2}, 1]$ is proven in [4]. This proof is presented here for $\theta = \frac{1}{2}$.

Theorem 4. *Assume that* (1.2) *holds and* $f \in C^1([0,T]), u \in C^2([0,T])$. *Then the local truncation error is* $\mathcal{O}(\Delta t^2)$, *the method* (3.2) *is convergent, and if* $e_0 = e_1 = 0$ *the global error satisfies*

$$
\|e_N\|^2 \leq 2\exp(4T\kappa\|(A-C)^{-\frac{1}{2}}\|^2)\|(A-C)^{-\frac{1}{2}}\|^2 U^2 \Delta t^4,
$$

where

$$
\begin{aligned}
U = {} & \max_{[t_{n-1},t_{n+1}]} \|u''(t)\|_2 \big(\frac{7}{6}\big) \\
& + \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{-\frac{1}{2}} A (A{-}C)^{-\frac{1}{2}} u'(t)\|_2 \big(\frac{1}{8}\big) \\
& + \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{-\frac{1}{2}} C (A{-}C)^{-\frac{1}{2}} u'(t)\|_2 \big(\frac{3}{4}\big) \\
& + \max_{[t_{n-1},t_{n+1}]} \Big\|\tfrac{d}{dt} B(u(\cdot))\Big\|_2 \max_{[t_{n-1},t_{n+1}]} \|u(\cdot)\|_2 \big(\frac{7}{8}\big) \\
& + \max_{[t_{n-1},t_{n+1}]} \Big\|\tfrac{d}{dt} B(u(\cdot))\Big\|_2 \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{-\frac{1}{2}} A (A{-}C)^{-\frac{1}{2}} u'(\cdot)\|_2 \big(\frac{1}{8}\big)\Delta t^2 \\
& + \max_{[t_{n-1},t_{n+1}]} \Big\|\tfrac{d}{dt} B(u(\cdot))\Big\|_2 \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{-\frac{1}{2}} C (A{-}C)^{-\frac{1}{2}} u'(\cdot)\|_2 \big(\frac{3}{4}\big)\Delta t^2 \\
& + \max_{[t_n,t_{n+1}]} \|B(u(\cdot))\|_2 \max_{[t_n,t_{n+1}]} \|(A{-}C)^{-\frac{1}{2}} A (A{-}C)^{-\frac{1}{2}} u'(\cdot)\|_2 \big(\frac{1}{8}\big) \\
& + \max_{[t_n,t_{n+1}]} \|B(u(\cdot))\|_2 \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{-\frac{1}{2}} C (A{-}C)^{-\frac{1}{2}} u'(\cdot)\|_2 \big(\frac{3}{4}\big).
\end{aligned}
$$

Proof. Using the Taylor expansion around $t_{n+\frac{1}{2}} := t_n + \frac{1}{2}\Delta t$ we obtain

$$\begin{aligned}
&\Big\| \frac{1}{\Delta t}\big(u(t_{n+1}) - u(t_n)\big) - u'(t_{n+\theta})\Big\|_2 \\
&= \frac{1}{\Delta t}\Big\| \int_{t_n+\frac{1}{2}\Delta t}^{t_n+\Delta t} \frac{u''(t)}{2!}(t_n+\Delta t - t)^2 dt - \int_{t_n+\frac{1}{2}\Delta t}^{t_n} \frac{u''(t)}{2!}(t_n - t)^2 dt \Big\|_2 \\
&\leq \frac{1}{2\Delta t} \max_{[t_{n-1},t_{n+1}]} \|u''(t)\|_2 \left(\Big| \int_{t_n+\frac{1}{2}\Delta t}^{t_n+\Delta t} (t_n+\Delta t - t)^2 dt \Big| \; + \; \Big| \int_{t_n+\frac{1}{2}\Delta t}^{t_n} (t_n - t)^2 dt \Big| \right) \\
&= \max_{[t_{n-1},t_{n+1}]} \|u''(t)\|_2 \frac{7}{6}\Delta t^2,
\end{aligned}$$

$$\begin{aligned}
&\|(A{-}C)^{\frac{1}{2}} A\big(\tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n+1}) + \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_n) - (A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}})\big)\|_2 \\
&= \Big\|(A{-}C)^{\frac{1}{2}} A\Big(\tfrac{1}{2}\int_{t_n+\frac{1}{2}\Delta t}^{t_{n+1}} (A{-}C)^{-\frac{1}{2}} u'(t)(t_n+\Delta t - t)dt + \tfrac{1}{2}\int_{t_n+\frac{1}{2}\Delta t}^{t_n} (A{-}C)^{-\frac{1}{2}} u'(t)(t_n+\Delta t - t)dt\Big)\Big\|_2 \\
&\leq \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{\frac{1}{2}} A(A{-}C)^{-\frac{1}{2}} u'(t)\|_2 \frac{1}{8}\Delta t^2
\end{aligned}$$

$$\begin{aligned}
&\| -(A{-}C)^{\frac{1}{2}} C\big(\tfrac{3}{2}(A{-}C)^{-\frac{1}{2}} u(t_n) - \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n-1}) - (A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}})\big)\|_2 \\
&\leq \Big\|(A{-}C)^{\frac{1}{2}} C\Big(\tfrac{3}{2}\int_{t_n+\frac{1}{2}\Delta t}^{t_{n+1}} (A{-}C)^{-\frac{1}{2}} u'(t)(t_n+\Delta t - t)dt - \tfrac{1}{2}\int_{t_n+\frac{1}{2}\Delta t}^{t_n-\Delta t} (A{-}C)^{-\frac{1}{2}} u'(t)(t_n-\Delta t - t)dt\Big)\Big\|_2 \\
&\leq \max_{[t_{n-1},t_{n+1}]} \|(A{-}C)^{\frac{1}{2}} C(A{-}C)^{-\frac{1}{2}} u'(t)\|_2 \frac{3}{4}\Delta t^2
\end{aligned}$$

$$\begin{aligned}
&\Big\|\Big(B(\mathcal{E}^t_{n+\frac{1}{2}}(u)) - B(u(t_{n+\frac{1}{2}}))\Big)(A{-}C)^{-\frac{1}{2}}\Big(A(\tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n+1}) + \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_n)\big) \\
&\qquad\qquad - C\big(\tfrac{3}{2}(A{-}C)^{-\frac{1}{2}} u(t_n) - \tfrac{1}{2}(A{-}C)^{-\frac{1}{2}} u(t_{n-1})\big)\Big)\Big\|_2 \\
&= \Big\|\Big(B(\mathcal{E}^t_{n+\frac{1}{2}}(u)) - B(u(t_{n+\frac{1}{2}}))\Big)(A{-}C)^{-\frac{1}{2}}\Big(A(A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}}) \\
&+ A(A{-}C)^{-\frac{1}{2}}\big(\tfrac{1}{2}\int_{t_{n+\frac{1}{2}\Delta t}}^{t_{n+1}} u'(t)(t_n+\Delta t - t)dt + \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_n} u'(t)(t_n - t)dt\big) \\
&\qquad\qquad - C(A{-}C)^{-\frac{1}{2}} u(t_{n+\frac{1}{2}}) - C(A{-}C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n) - \tfrac{1}{2}u(t_{n-1})\big)\Big)\Big\|_2 \\
&= \Big\|\Big(B(\mathcal{E}^t_{n+\frac{1}{2}}(u)) - B(u(t_{n+\frac{1}{2}}))\Big)\Big[u(t_{n+\frac{1}{2}}) + (A{-}C)^{-\frac{1}{2}}\Big(A(A{-}C)^{-\frac{1}{2}}\big(\tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_n+\Delta t - t)dt \\
&\qquad + \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_n} u'(t)(t_n - t)dt\big) - C(A{-}C)^{-\frac{1}{2}}\big(\tfrac{3}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_{n+1} - t)dt - \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n-1}} u'(t)(t_{n-1} - t)dt\big)\Big)\Big]\Big\|_2 \\
&\leq \Big\|\Big(B(\mathcal{E}^t_{n+\frac{1}{2}}(u)) - B(u(t_{n+\frac{1}{2}}))\Big)u(t_{n+\frac{1}{2}})\Big\|_2 \\
&+ \Big\|\Big(B(\mathcal{E}^t_{n+\frac{1}{2}}(u)) - B(u(t_{n+\frac{1}{2}}))\Big)(A{-}C)^{-\frac{1}{2}}\Big(A(A{-}C)^{-\frac{1}{2}}\big(\tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_n+\Delta t - t)dt \\
&\; + \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_n} u'(t)(t_n - t)dt\big) - C(A{-}C)^{-\frac{1}{2}}\big((\theta{+}1)\int_{t_{n+\theta}}^{t_{n+1}} u'(t)(t_{n+1} - t)dt - \theta\int_{t_{n+\theta}}^{t_{n-1}} u'(t)(t_{n-1} - t)dt\big)\Big)\Big]\Big\|_2
\end{aligned}$$

$$\begin{aligned}
&\leq \max_{[t_{n-1},t_{n+1}]} \Big\| \tfrac{d}{dt} B(u(\cdot)) \Big\|_2 \max_{[t_{n-1},t_{n+1}]} \|u(\cdot)\|_2 \frac{3}{4}\Delta t^2 \\
&+ \max_{[t_{n-1},t_{n+1}]} \Big\| \tfrac{d}{dt} B(u(\cdot)) \Big\|_2 \max_{[t_{n-1},t_{n+1}]} \|(A-C)^{-\frac{1}{2}} A (A-C)^{-\frac{1}{2}} u'(\cdot)\|_2 \frac{1}{8}\Delta t^4 \\
&+ \max_{[t_{n-1},t_{n+1}]} \Big\| \tfrac{d}{dt} B(u(\cdot)) \Big\|_2 \max_{[t_{n-1},t_{n+1}]} \|(A-C)^{-\frac{1}{2}} C (A-C)^{-\frac{1}{2}} u'(\cdot)\|_2 \frac{3}{4}\Delta t^4,
\end{aligned}$$

$$\begin{aligned}
&\Big\| B(u(t_{n+\frac{1}{2}}))\Big[(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1})+\tfrac{1}{2}u(t_n)\big) - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n)-\tfrac{1}{2}u(t_{n-1})\big)\Big) \\
&\qquad\qquad\qquad - (A-C)^{-\frac{1}{2}}u(t_{n+\frac{1}{2}})\Big]\Big\|_2 \\
&= \Big\| B(u(t_{n+\frac{1}{2}}))\Big[(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_n+\Delta t - t)dt + \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_n} u'(t)(t_n - t)dt\big) \\
&\qquad\qquad - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_{n+1}-t)dt - \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n-1}} u'(t)(t_{n-1}-t)dt\big)\Big)\Big]\Big\|_2 \\
&\leq \Big\| B(u(t_{n+\frac{1}{2}}))(A-C)^{-\frac{1}{2}}A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_n+\Delta t - t)dt + \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_n} u'(t)(t_n - t)dt\big)\Big\|_2 \\
&\quad + \Big\| B(u(t_{n+\frac{1}{2}}))(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_{n+1}-t)dt - \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n-1}} u'(t)(t_{n-1}-t)dt\big)\Big)\Big]\Big\|_2 \\
&\leq \max_{[t_n,t_{n+1}]} \|B(u(\cdot))\|_2 \Big\|(A-C)^{-\frac{1}{2}}A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_n+\Delta t - t)dt + \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_n} u'(t)(t_n - t)dt\big)\Big\|_2 \\
&\quad + \max_{[t_n,t_{n+1}]} \|B(u(\cdot))\|_2 \Big\|(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n+1}} u'(t)(t_{n+1}-t)dt - \tfrac{1}{2}\int_{t_{n+\frac{1}{2}}}^{t_{n-1}} u'(t)(t_{n-1}-t)dt\big)\Big)\Big\|_2 \\
&\leq \max_{[t_n,t_{n+1}]} \|B(u(\cdot))\|_2 \max_{[t_n,t_{n+1}]} \|(A-C)^{-\frac{1}{2}}A(A-C)^{-\frac{1}{2}}u'(\cdot)\|_2 \frac{1}{8}2\Delta t^2 \\
&\qquad + \max_{[t_n,t_{n+1}]} \|B(u(\cdot))\|_2 \max_{[t_{n-1},t_{n+1}]} \|(A-C)^{-\frac{1}{2}}C(A-C)^{-\frac{1}{2}}u'(\cdot)\|_2 \frac{3}{4}\Delta t^2.
\end{aligned}$$

Therefore

$$\|\tau_{n+1}(\Delta t)\|_2 \leq U\Delta t^2, \tag{3.12}$$

which proves the consistency of method (3.2). The error $e_n = u(t_n) - u_n$ satisfies

$$\begin{aligned}
&\frac{e_{n+1}-e_n}{\Delta t} + (A-C)^{\frac{1}{2}}A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}e_{n+1}+\tfrac{1}{2}e_n\big) - (A-C)^{\frac{1}{2}}C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}e_n - \tfrac{1}{2}e_{n-1}\big) \\
&+ B(\mathcal{E}_{n+\frac{1}{2}})(A-C)^{\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}e_{n+1}+\tfrac{1}{2}e_n\big) - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}e_n - \tfrac{1}{2}e_{n-1}\big)\Big) \\
= {} & \tau_{n+1}(\Delta t) \\
&- \big(B(\mathcal{E}^t_{n+\frac{1}{2}}) - B(\mathcal{E}_{n+\frac{1}{2}})\big)(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1})+\tfrac{1}{2}u(t_n)\big) \\
&\qquad\qquad - C(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_n) - \tfrac{1}{2}u(t_{n-1})\big)\Big).
\end{aligned} \tag{3.13}$$

From (3.10) we have

$$\|u_n\|_2 \leq \Lambda_1 := \Big(\|u_0\|^2 + 2\Big\|\begin{bmatrix}u_1\\u_0\end{bmatrix}\Big\|_G^2 + T\|(A-C)^{-\frac{1}{2}}\|^2 \max_{t\in[0,T]}\|f(t)\|^2\Big)^{\frac{1}{2}} \quad \forall n = 1,\ldots,N,$$

also from (2.1) we obtain

$$\|u(t)\|_2 \leq \Lambda_2 := \Big(\|u(0)\|^2 + \int_0^T \|(A-C)^{-1}f(s)\|^2 ds\Big)^{\frac{1}{2}},$$

and define

$$\Lambda_3 = \max_{n=1,\dots,N} \|(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1})+\tfrac{1}{2}u(t_n)\big) - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n)-\tfrac{1}{2}u(t_{n-1})\big)\Big)\|_2.$$

The last term in the RHS of (3.13) is

$$\begin{aligned}
&\big(B(\mathcal{E}^t_{n+\frac{1}{2}}) - B(\mathcal{E}_{n+\frac{1}{2}})\big)(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1}) + \tfrac{1}{2}u(t_n)\big) \\
&\qquad\qquad\qquad\qquad\qquad - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n) - \tfrac{1}{2}u(t_{n-1})\big)\Big) \\
&= \int_0^1 \frac{d}{ds}\big[B(s\mathcal{E}^t_{n+\frac{1}{2}} + (1-s)\mathcal{E}_{n+\frac{1}{2}})\big]ds(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1}) + \tfrac{1}{2}u(t_n)\big) \\
&\qquad\qquad\qquad\qquad\qquad - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n) - \tfrac{1}{2}u(t_{n-1})\big)\Big) \\
&= \int_0^1 \nabla_u\big[B(u)(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1}) + \tfrac{1}{2}u(t_n)\big) \\
&\qquad\qquad - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n) - \tfrac{1}{2}u(t_{n-1})\big)\Big)\big]|_{u=s\mathcal{E}^t_{n+\frac{1}{2}}+(1-s)\mathcal{E}_{n+\frac{1}{2}}} ds\,\big(\mathcal{E}^t_{n+\frac{1}{2}} - \mathcal{E}_{n+\frac{1}{2}}\big),
\end{aligned}$$

which since $B(\cdot)$ is C^1 implies

$$\begin{aligned}
&\|\big(B(\mathcal{E}^t_{n+\frac{1}{2}}) - B(\mathcal{E}_{n+\frac{1}{2}})\big)(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\tfrac{1}{2}u(t_{n+1}) + \tfrac{1}{2}u(t_n)\big) \\
&\qquad\qquad\qquad\qquad - C(A-C)^{-\frac{1}{2}}\big(\tfrac{3}{2}u(t_n) - \tfrac{1}{2}u(t_{n-1})\big)\Big)\|_2 \\
&\leq 2\kappa(\|e_n\|_2 + \|e_{n-1}\|), \forall n,
\end{aligned}$$

where

$$2\kappa = \max_{s\in[0,1], \|U_1\|_2\leq 2\Lambda_1, \|U_2\|_2\leq\Lambda_3, \|V_2\|_2\leq 2\Lambda_2} \|\nabla_u\big[B(sV_2 + (1-s)U_1)U_2\big]\|_2.$$

Multiplying (3.13) by $(A-C)^{-\frac{1}{2}}\Big(A(A-C)^{-\frac{1}{2}}\big(\frac{1}{2}e_{n+1} + \frac{1}{2}e_n\big) - C(A-C)^{-\frac{1}{2}}\big(\frac{3}{2}e_n - \frac{1}{2}e_{n-1}\big)\Big)$ and taking the sum from $n = 1$ to $N-1$ we obtain

$$\begin{aligned}
&\Big\|\begin{bmatrix} e_N \\ e_{N-1}\end{bmatrix}\Big\|_G^2 + \frac{1}{4}\sum_{n=1}^{N-1}\|e_{n+1} - 2e_n + e_{n-1}\|_F^2 \\
&+\frac{\Delta t}{2}\sum_{n=1}^{N-1}\|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}e_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}e_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}e_{n-1}\|^2 \\
&\leq \Big\|\begin{bmatrix} e_1 \\ e_0\end{bmatrix}\Big\|_G^2 + \Delta t\sum_{n=1}^{N-1}\|(A-C)^{-\frac{1}{2}}\tau_{n+1}(\Delta t)\|^2 + \kappa\|(A-C)^{-\frac{1}{2}}\|^2\Delta t\sum_{n=1}^{N-1}(\|e_n\|^2 + \|e_{n-1}\|^2).
\end{aligned} \tag{3.14}$$

Using again the proof of Lemma 1, after some calculation we get

$$
\begin{aligned}
&\|e_N\|^2 + \frac{1}{2}\sum_{n=1}^{N-1}\|e_{n+1}-2e_n+e_{n-1}\|_F^2 \\
&\quad + \Delta t \sum_{n=1}^{N-1} \|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}e_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}e_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}e_{n-1}\|^2 \\
&\le 2\Big(\Big\|\begin{bmatrix} e_1 \\ e_0 \end{bmatrix}\Big\|_G^2 + \Delta t \sum_{n=1}^{N-1}\|(A-C)^{-\frac{1}{2}}\tau_{n+1}(\Delta t)\|^2\Big) + 4\kappa\|(A-C)^{-\frac{1}{2}}\|^2 \Delta t \sum_{n=0}^{N-1}\|e_n\|^2.
\end{aligned}
$$

Therefore, from the discrete Grönwall lemma, we deduce the following error estimate

$$
\begin{aligned}
&\|e_N\|^2 + \tfrac{1}{2}\sum_{n=1}^{N-1}\|e_{n+1}-2e_n+e_{n-1}\|_F^2 \\
&\quad + \Delta t \sum_{n=1}^{N-1} \|\tfrac{1}{2}A(A-C)^{-\frac{1}{2}}e_{n+1} + \big(\tfrac{1}{2}A - \tfrac{3}{2}C\big)(A-C)^{-\frac{1}{2}}e_n + \tfrac{1}{2}C(A-C)^{-\frac{1}{2}}e_{n-1}\|^2 \\
&\le 2\exp(4T\kappa\|(A-C)^{-\frac{1}{2}}\|^2)\Big(\Big\|\begin{bmatrix} e_1 \\ e_0 \end{bmatrix}\Big\|_G^2 + \Delta t \sum_{n=1}^{N-1}\|(A-C)^{-\frac{1}{2}}\tau_{n+1}(\Delta t)\|^2\Big).
\end{aligned}
$$

Finally, the convergence result follows from the consistency bound (3.12). □

4.0 CONCLUSION

The Crank-Nicolson/Adams-Bashforth 2 second-order method analyzed herein offers an improvement over the first-order method proposed in [1] in terms of accuracy, though it does so at the expense of being considerably more computationally expensive. Nonetheless, the fact that it is unconditionally stable, and thus accomodates any choice of Δt makes it an attractive method in terms of its stability properties.

A.1 APPENDIX

A.1.1 Proof of Theorem 1

Proof. Consider the difference of solutions of the two Cauchy problems (2.1) and (2.3)

$$w(t) = z(t) - y(t), \qquad w : \mathbb{R} \to \mathbb{R}^d$$

and take the derivative with respect to t:

$$w'(t) = z'(t) - y'(t) = F(t, z(t)) - F(t, y(t)) + \delta(t).$$

Integrating both sides from t to t_0 and recalling intial conditions yields

$$w(t) - \delta_0 = \int_{t_0}^{t} [F(s, z(s)) - F(s, y(s))]ds + \int_{t_0}^{t} \delta(s)ds.$$

Taking the infinity norm of both sides and applying the Triangle Inequality gives

$$\|w(t)\|_\infty \leq \|\delta_0\|_\infty + \int_{t_0}^{t} \|F(s, z(s)) - F(s, y(s))\|_\infty ds + \int_{t_0}^{t} \|\delta(s)\|_\infty ds.$$

Using the Lipschitz condition and assumptions (2.4),

$$\|w(t)\|_\infty \leq \epsilon + \int_{t_0}^{t} L\|z(t) - y(s)\|_\infty ds + \int_{t_0}^{t} \epsilon \ ds.$$

Integrating further and recalling $w(t) = z(t) - y(t)$ gives

$$\|w(t)\|_\infty \leq \epsilon + |t - t_0|\epsilon + \int_{t_0}^{t} L\|w(s)\|_\infty ds.$$

Since L is a nonnegative integrable function on $\mathcal{I}$, $\epsilon + |t - t_0|\epsilon$ and $\|w(t)\|_\infty$ are continuous on $\mathcal{I}$, and $\epsilon + |t - t_0|\epsilon$ is nondecreasing, we can apply the Grönwall Lemma and conclude

$$\|w(t)\|_\infty \leq (1 + |t - t_0|)\epsilon e^{L|t-t_0|}.$$

Choosing $C = (1 + K)e^{LK}$, where $K = \max_{t \in \mathcal{I}} |t - t_0|$ gives the result, and Definition 1.1 is satisfied [3]. □

A.1.2 Generating Polynomials for a Linear Multistep Method

For the general multistep method (1.4), the generating polynomials are

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j. \tag{1}$$

For a general introduction to the theory of difference equations on which this concept is based see [2], Chapter 10.4.

A.1.3 G-stability for Scalar BDF 2

The general 2-step BDF (Backward Differentiation Formula) method is

$$\tfrac{3}{2}y_{n+2} - 2y_{n+1} + \tfrac{1}{2}y_n = hf(t_{n+2}, y_{n+2}). \tag{2}$$

To allow for the possibility of f being a nonlinear function, take a second numerical solution sequence $\{\hat{y}_n\}$ and denote its difference from $\{y_n\}$ as $\Delta y_n = y_n - \hat{y}_n$. Next, assume the method satisfies the Lipschitz condition (2.12) with $L = 0$. Substituting (2) into (2.12) and multiplying by $\frac{1}{\Delta t}$ gives

$$E = \mathrm{Re}\Big\langle \tfrac{3}{2}\Delta y_{n+2} - 2\Delta y_{n+1}\tfrac{3}{2}\Delta y_n, \Delta y_{n+2}\Big\rangle \leq 0, \tag{3}$$

Now consider the equation

$$E = \|\Delta Y_{n+1}\|_G^2 - \|\Delta Y_n\|_G^2 + \|a_2\Delta y_{n+1} + a_1\Delta y_{n+1} + a_0\Delta y_n\|^2, \qquad a_0, a_1, a_2 \in \mathbb{R}. \tag{4}$$

Using $E \leq 0$ from above implies $\|\Delta Y_{n+1}\|_G^2 \leq \|\Delta Y_n\|_G^2$, since $\|a_2\Delta y_{n+1} + a_1\Delta y_{n+1} + a_0\Delta y_n\|^2 \geq 0$. Let

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$$

If matching the coefficients of (3) to (4) produces an SPD G, method (2) is G-stable by Definition 3.

Imposing symmetry by letting $g_{12} = g_{21}$ produces the following nonlinear system of six equations in six unknowns:

$$\begin{array}{llll} \Delta y_{n+1}^2 : & \tfrac{3}{2} = g_{11} + a_2^2, & \Delta y_{n+1}y_n : & -2 = 2g_{12} + 2a_2a_1 \\ \Delta y_n^2 : & 0 = g_{22} - g_{11} + a_1^2 & \Delta y_{n+1}y_n : & \tfrac{1}{2} = 2a_2a_0 \\ \Delta y_{n-1}^2 : & 0 = -g_{22} + a_0^2 & \Delta y_n\Delta y_{n-1} : & 0 = -2g_{12} + 2a_1a_0. \end{array}$$

Solving this system produces the G-matrix

$$G = \frac{1}{4}\begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}. \tag{5}$$

By Sylvester's Criterion (see A.1.4) this is a positive-definite matrix, and thus the BDF 2 method (2) is G-stable.

Alternatively, one can use the method's generating polynomials (see A.1.2), which for scalar BDF 2 are

$$\rho(\zeta) = \tfrac{3}{2}\zeta^2 - 2\zeta + \tfrac{1}{2}, \qquad \sigma(\zeta) = \zeta^2.$$

Define the function

$$E(\zeta) = \tfrac{1}{2}(\rho(\zeta)\sigma(\tfrac{1}{\zeta}) + \rho(\tfrac{1}{\zeta})\sigma(\zeta)),$$

which for BDF 2 simplifies to

$$E(\zeta) = \tfrac{1}{4}(\zeta^2 + \frac{1}{\zeta}) - (\zeta + \frac{1}{\zeta}) + \tfrac{3}{2} = \tfrac{1}{2}(\zeta - 1)^2\tfrac{1}{2}(\frac{1}{\zeta} - 1)^2 = a(\zeta)a(\tfrac{1}{\zeta}).$$

Define a function $P(\zeta,\omega) = \frac{1}{2}(\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta)) - a(\zeta)a(\omega)$, which with some simplification and factoring becomes

$$\begin{aligned} P(\zeta,\omega) &= (\zeta\omega - 1)\Big(\frac{5}{4}\zeta\omega - \frac{1}{2}\zeta - \frac{1}{2}\omega + \frac{1}{4}\Big) \\ &= (\zeta\omega - 1)(g_{11}\zeta\omega - g_{12}\zeta - g_{21}\omega + g_{22}), \end{aligned}$$

which gives matrix

$$G = \frac{1}{4}\begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}. \tag{6}$$

which is the same as when calculated directly.

A.1.4 Sylvester's Criterion

Theorem 5. *If a matrix A and all its principal-minors have strictly positive determinants, then A is positive-definite.*

A.1.5 Proof of Theorem 3

Proof. Since B(u) is skew-symmetric, multiplying method (1.5) through by $u^T(t)$ gives

$$u^T(t)u'(t) + u^T(t)(A - C)u(t) = u^T(t)f(t),$$

and since $A - C \succcurlyeq 0$,

$$u^T(t)u'(t) \leq u^T(t)f(t) \iff 2u^T(t)u'(t) \leq 2u^T(t)f(t) \iff \frac{d}{dt}\|u(t)\|_2^2 \leq 2u^T(t)f(t).$$

Letting $F_T^2 = \max_{t\in[0,T]} \|f(t)\|_2^2$, using Young's Inequality gives

$$\frac{d}{dt}\|u(t)\|_2^2 \leq \|u(t)\|_2^2 + F_2^T.$$

Multiplying both sides by integrating factor e^{-t}, letting $v(t) = \|u(t)\|_2^2$, and rearranging terms gives

$$v'(t)e^{-t} - v(t)e^{-t} \leq F_T^2 e^{-t} \iff \frac{d}{dt}[v(t)e^{-t}] \leq F_T^2 e^{-t}.$$

Finally, integrating both sides from 0 to t gives

$$\int_0^t \frac{d}{ds}[v(s)e^{-s}]ds \leq \int_0^t F_T^2 e^{-s}ds \iff v(t)e^{-t} - v(0)e^0 \leq -F_T^2 e^{-t} + F_T^2 \iff v(t) \leq v(0)e^t + F_T^2(e^t - 1),$$

and taking $\|u(t)\|_2^2 = v(t)$ and $\|u(0)\|_2^2 = v(0)$ gives the result. □

A.1.6 Inverse of a Symmetric Positive-Definite Matrix

Lemma 3. *The inverse of a symmetric positive-definite matrix is symmetric positive-definite.*

Proof. Let D be an SPD matrix. Since D is symmetric, there exists a diagonalization of D

$$\Lambda = Q^{-1}DQ, \qquad \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n),$$

where $Q^{-1}Q = I$, the identity matrix, and $Q^{-1} = Q^T$. Then

$$D^{-1} = (Q\Lambda Q^{-1})^{-1} = (Q^{-1})^{-1}\Lambda^{-1}Q^{-1} = Q\Lambda^{-1}Q^{-1} = (Q^{-1})^T(\Lambda^{-1})^TQ^T = (Q\Lambda^{-1}Q^{-1})^T = (D^{-1})^T.$$

Thus D^{-1} is symmetric, and it is positive-definite since

$$\Lambda^{-1} = diag\Big(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_n}\Big), \qquad \lambda_i > 0 \ \ \forall i.$$

□

A.1.7 More Properties of Symmetric and Positive-Definite Matrices

Lemma 4. *If D is a symmetric positive-definite matrix it can be diagonalized as $Q\Lambda^{-1}Q^{-1}$ by Lemma 3. Define $D^{\frac{1}{2}}$ as a matrix such that $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$ (derived more specifically in the proof below). Then $D^{\frac{1}{2}}$ is also positive definite, and $(D^{\frac{1}{2}})^T = (D^T)^{\frac{1}{2}}$.*

Proof. Since D is symmetric positive-definite, there exists a diagonalization of D such that

$$\Lambda = Q^{-1}DQ, \qquad \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n), \quad \lambda_i > 0 \ \ \forall i,$$

where $Q^{-1}Q = I$, the identity matrix, and $Q^{-1} = Q^T$. Then

$$D^{\frac{1}{2}} = (Q\Lambda Q^{-1})^{\frac{1}{2}} = (Q\sqrt{\Lambda}\sqrt{\Lambda}Q^{-1})^{\frac{1}{2}} = (Q\sqrt{\Lambda}Q^{-1}Q\sqrt{\Lambda}Q^{-1})^{\frac{1}{2}} = ((Q\sqrt{\Lambda}Q^{-1})^2)^{\frac{1}{2}} = Q\sqrt{\Lambda}Q^{-1}.$$

Taking the transpose gives

$$(D^{\frac{1}{2}})^T = (Q\sqrt{\Lambda}Q^{-1})^T = (Q^{-1})^T(\sqrt{\Lambda})^TQ^T = Q\sqrt{\Lambda}Q^T = D^{\frac{1}{2}} = (D^T)^{\frac{1}{2}},$$

where the last equality holds since $D = D^T$ from the fact that it is symmetric by assumption. $D^{\frac{1}{2}}$ is thus positive-definite since

$$\Lambda = diag\big(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_n}\big),$$

and $\sqrt{\lambda_i} > 0$ for all i since $\lambda_i > 0$ for all i. □

BIBLIOGRAPHY

[1] M. Anitescu, F. Pahlevani, and W. J. Layton, *Implicit for local effects and explicit for nonlocal effects is unconditionally stable*, Electron. Trans. Numer. Anal., 18 (2004), pp. 174–187 (electronic).

[2] E. Hairer and G. Wanner, *Solving ordinary differential equations. II*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition.

[3] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, vol. 37 of Texts in Applied Mathematics, Springer-Verlag, Berlin, second ed., 2007.

[4] C. Trenchea, *Second order implicit for local effects and explicit for nonlocal effects is unconditionally stable*, submitted, (2012).