# Numerical Stability and Catalan Numbers

Arash Ghasemi[*], Kidambi Sreenivas[†] and Lafayette K. Taylor[‡]
SimCenter: National Center for Computational Engineering
701 E. M.L. King Blvd. Chattanooga, TN 37403, UTC

**Abstract**

To predict allowable time-step size for the fully discretized nonlinear differential equations, a stability theory is developed using exact determination of an infinite perturbation series. Mathematical induction is used to determine the coefficients of the series. It is discovered that the closed-form equation for the nonlinear shift of generic polynomial non-linearity can be written as a series expansion where the coefficients are the Pfaff-Fuss-Catalan numbers in Combinatorics. This reveals criteria which can be used to analytically determine the allowable time step. It is shown that stability region decreases when the nonlinearity of the differential equation increases. Therefore, the maximum allowable time step is severely limited by the nonlinearity even if an unconditionally stable scheme (in a linear sense) is used. The theory is applied to the case of general system of time-dependent nonlinear Partial Differential Equations.

**Keywords** Numerical Stability, Nonlinear Differential Equations, Catalan Numbers, Time Marching

## 1 Introduction

The spatial discretization of the system of partial differential equations

$$\frac{\partial v_k}{\partial t} = G\left(v_k, \frac{\partial^i v_k}{\partial x^i}, \frac{\partial^j v_k}{\partial y^j}, \ldots\right) \tag{1}$$

for $\vec{v} = (v_1, v_2, \ldots) = v_k$ and some range of $i, j, \ldots$ leads to the system of semi-discrete form $d\mathbf{v}/dt = R(\mathbf{v}(t))$ or

$$\mathbf{v} = \mathbf{v}_0 + \int_{t_0}^{t} R\left(\mathbf{v}(\xi)\right) d\xi, \tag{2}$$

where $\mathbf{v} = \mathbf{v}(t)$ is the spatially distributed nodal/modal solution vector at time $t$ and $\mathbf{v}_0 = \mathbf{v}(t_0)$ designates the initial condition of the system and $R$ is the residual of the spatial discretization. The integral in (2) can be arbitrarily discretized to obtain the following space-time discretization

$$\mathbf{u} = \mathbf{u}_0 + \Delta t \mathbb{S} \otimes R\left(\mathbf{u}\right) \tag{3}$$

where $\mathbf{u} = \left[\mathbf{v}(t_1)^T, \mathbf{v}(t_2)^T, \ldots, \mathbf{v}(t_s)^T\right]^T$ is the space-time vector containing solution $\mathbf{v}(t_i)$ at temporal collocation point $t_0 < t_i \leq t$, $\mathbf{u}_0 = \left[\mathbf{v}_0^T, \mathbf{v}_0^T, \ldots, \mathbf{v}_0^T\right]^T$ is the space-time initial condition, $\Delta t = (t - t_0)/s$ is the time-step, and $\mathbb{S}$ is the integration operator. For the first-order truncated Riemannian integration, $\mathbb{S}$ is a lower-diagonal unity matrix while it can be a full matrix when orthogonal polynomials such as Chebyshev polynomials are used [8].

Equation (3) represents a nonlinear system of equations which requires an iterative method to be solved in practice. A *sequential iterative* solution of (3) constitutes the Discrete Picard Iteration (DPI) which is the exact numerical counterpart of original Picard iterations to find the fixed-point of nonlinear system (3). Thus,

$$\mathbf{u}_{n+1} = \mathbf{u}_0 + \Delta t \mathbb{S} \otimes R\left(\mathbf{u}_n\right) \tag{4}$$

[*] Doctorate Candidate
[†] Research Professor
[‡] Professor

is considered here as the basic target system in which the stability of the iterative procedure is sought. However, a *simultaneous* update to (3), i.e.

$$\mathbf{u}_{n+1} = \mathbf{u}_0 + \Delta t \mathbb{S} \otimes R\left(\bar{a}\,\mathbf{u}_{n+1} + \bar{b}\,\mathbf{u}_n\right), \qquad \bar{a} + \bar{b} = 1, \tag{5}$$

yields an implicit form of the Discrete Picard Iteration which is also studied here. In particular, one is interested to know:

1. For the case where $R\left(\mathbf{u}_n\right)$ has polynomial nonlinearity, under what conditions are the iterative forms (4) and (5) stable?
2. If $\mathbb{S}$ is chosen such that (5) is linearly unconditional stable and assuming that the Jacobian of linearization is computed exactly, then does this imply nonlinear stability for arbitrary $\Delta t$?

Answers to the above questions may improve the understanding of nonlinear instability of numerical methods which is important for researchers in the field of Computational Sciences. In general, the nonlinear systems (4) and (5) can be written as $F_h(u_h) = 0$ as an approximation to the original nonlinear system obtained without the discretization of the differential operators, i.e. $F(u) = 0$. Keller [3] used this notation to obtain the stability criteria based on Lipschitz continuous linearization. Later Lopez-Marcos et. al. [4] worked on the same approach to interpret nonlinear stability based on local linear stability near the exact solution of the nonlinear system. As pointed out by Pirovino [5], these linearization approaches have a disadvantage that the Lipschitz constant of the derivative $F_h(u_h) = 0$ must be known which is not possible in practice. To overcome this, Pirovino used the linearization approach in a neighborhood of $u_h$ to determine nonlinear stability. The approaches just mentioned here use norm-based inequalities to investigate the contraction of the nonlinear operator and corresponding stability. These inequality relations estimate upperbound for the solution behavior but not the *exact* nonlinear mechanism which induces instability. Therefore the exact nonlinear shift ("thresholds" according to [4]) in the stability region and in the solution remains unanswered. The exact mechanism of generation of the nonlinear shift is important. Such knowledge might stimulate the design of faster algorithms with less stringent stability limits.

The structure of the paper is summarized as follows. In § 2 a loosely coupled form of (4) is considered where the perturbation parameter $\epsilon$ is introduced. This is a special case of the general theory presented at the end of the paper in § 6. Then the perturbation analysis is performed in § 3 and the exact nonlinear shift is obtained. In § 4, the results of § 3 are generalized to arbitrary polynomial nonlinearity. For implicit discretization (5), the perturbation analysis is performed in § 5. The main result of the paper is presented in § 6 where the stability of the general time-dependent PDE (1) is related to the concepts developed in (§ 2- § 5).

# 2 The perturbation parameter

Toward the stability analysis of (4) and (5) it is insightful to assume that the system is lossely coupled meaning that the k$^{th}$ state variable at the n$^{th}$ Picard iteration, i.e. $u_{k,n}(t)$ is almost independent of other variables $u_{l,n}(t), l \neq k$ in the solution vector $\mathbf{u}_n$. This assumption is exact when the residual arises from the discretization of an ordinary differential equations. However, this still remains as an approximation for the case of lower-order spatial discretization of PDEs where the system is very similar to a lossely coupled one. [1]. In this case (4) can be written as $u_{k,n+1} = u_{k,0} + \Delta t \mathbb{S} \otimes R\left(u_{k,n}\right)$ or in compact form

$$u_{n+1} = u_0 + \Delta t \mathbb{S} \otimes R\left(u_n\right) \tag{6}$$

where $u_{n+1}$ is a scalar. The second assumption in this section is that the number of temporal collocation points is limited to one. Consequently the integration operator $\mathbb{S} = \mathbb{S}_{1\times 1} = \lambda$ reduces to a scalar value which yields (6) to reduce to the following scalar equation

$$u_{n+1} = u_0 + \Delta t \lambda R\left(u_n\right). \tag{7}$$

[1] In § 6, it will be shown that the results can be consistently extended to the more general cases (4) and (5) without such assumption

Assuming that the residual $R$ is analytic over the time span, (7) can be expanded as

$$R(u_n) = R(u_0) + \left.\frac{\partial R}{\partial u}\right|_{u_0} \Delta u_n + \frac{1}{2} \left.\frac{\partial^2 R}{\partial u^2}\right|_{u_0} \Delta u_n^2 + \ldots \tag{8}$$

where $\Delta u_n = u_n - u_0$. Substituting (8) into (7) results in

$$\Delta u_{n+1} = \Delta t\, \lambda \left( R_0 + \left.\frac{\partial R}{\partial u}\right|_{u_0} \Delta u_n + \frac{1}{2} \left.\frac{\partial^2 R}{\partial u^2}\right|_{u_0} \Delta u_n^2 + \ldots \right). \tag{9}$$

The perturbation parameter $\epsilon$ is introduced here as a relation between first derivative (Jacobian) and higher-order derivative (Hessian)[2].

$$\left.\frac{\partial^2 R}{\partial u^2}\right|_{u_0} = 2\,\epsilon \left.\frac{\partial R}{\partial u}\right|_{u_0} \tag{10}$$

The intuition for selecting $\epsilon$ as the perturbation parameter is described as follows. One can propose that the value of $\epsilon$ should be small for a weakly nonlinear residual where the second derivative is small compared to the first derivative. To validate this proposition, consider the scenario where the nonlinear residual converges to the linear functional $R \to c\,u$ where $c$ is a constant. Then $\partial R/\partial u|_{u_0} \to c$ and $\partial^2 R/\partial u^2\big|_{u_0} \to 0$ which means that $\epsilon \to 0$ must hold in eq.(10) as $R \to c\,u$. However it should be noted that the analysis presented in the following sections is valid for arbitrarily large $\epsilon$ since the perturbation series is not truncated. Substituting (10) into (9) yields

$$\Delta u_{n+1} = \Delta t\, \lambda \left( R_0 + \left.\frac{\partial R}{\partial u}\right|_{u_0} \Delta u_n + \epsilon \left.\frac{\partial R}{\partial u}\right|_{u_0} \Delta u_n^2 + \ldots \right) \tag{11}$$

Defining linear stability number as

$$r = \Delta t\, \lambda \left.\frac{\partial R}{\partial u}\right|_{u_0} \tag{12}$$

Equation (11) can be written as

$$\Delta u_{n+1} = \Delta t\, \lambda\, R_0 + r\Delta u_n + \epsilon r \Delta u_n^2 + \ldots \tag{13}$$

To simplify notation define $U_0 = \Delta t\, \lambda\, R_0$ and $U = \Delta u$. Hence (13) yields

$$U_{n+1} = U_0 + r(1 + \epsilon U_n)U_n + \ldots \tag{14}$$

This is the final form which will be analyzed using the formal perturbation technique. Note that in this case, the nonlinear residual is

$$R_n = cU_n + c\epsilon U_n^2, \tag{15}$$

where $c = \frac{\partial R}{\partial u}\big|_{u_0}$ is the Jacobian of the linearization.

## 3 Perturbation Analysis

The solution to eq. (14) is expanded in the term of $\epsilon$ and the i$^{th}$ perturbation amplitudes at the n$^{th}$ Picard iteration, i.e., $u_{i,n}$ such that

$$U_n = \sum_{i=0}^{\infty} u_{i,n} \epsilon^i, \tag{16}$$

subject to initial condition

$$U_{n=0} = u_0 + 0\epsilon + 0\epsilon^2 + \ldots \tag{17}$$

[2] The generalization of (10) is presented in (75).

Substituting (16) into (14) and matching the coefficients of $\epsilon^i$, a cascade of linear equations is obtained which are recursively solved to find perturbation amplitudes. It is shown in Appendix (A) that the i$^{th}$ perturbation amplitude converges to

$$\frac{u_{i,\infty}}{u_0^{i+1}} = C(i)\, \frac{r^i}{(1-r)^{2i+1}} \quad (i = 0, 1, 2, \ldots, \quad |r| \leq 1)\,, \tag{18}$$

where $C(i) = \{1, 1, 2, 5, 14, 42, 132, 429\ldots\}$ is the well-known Catalan sequence [1] given explicitly as

$$C(i) = \frac{(2i)!}{i! \times (i+1)!} = \frac{\text{binomial}\,(2i, i)}{i+1} \tag{19}$$

According to [1], this sequence has many different interpretations in Combinatorics but nothing about nonlinear stability of time-stepping methods has been reported so far. Substituting the perturbation amplitudes (125) into (16), the final nonlinear solution to DPI (14) is obtained as follows.

$$\frac{U}{u_0} = \sum_{i=0}^{\infty} C(i)\, \frac{r^i}{(1-r)^{2i+1}} (\epsilon u_0)^i$$

or

$$\frac{U}{u_0} = \underbrace{\frac{1}{1-r}}_{\text{Linear}} + \underbrace{\sum_{i=1}^{\infty} C(i)\, \frac{r^i}{(1-r)^{2i+1}} (\hat{\epsilon})^i}_{\text{Nonlinear Shift}}, \qquad \hat{\epsilon} = \epsilon u_0 \tag{20}$$

where $\hat{\epsilon}$ is introduced as the *combined perturbation amplitude*. Although all perturbation amplitudes converge in the linear (original) stability region $|r| < 1$ [3], their partial sum identified as *nonlinear shift* in (20) may or may not converge in this region. Therefore one can conclude that the linear stability region is affected as a consequence of the existence of the nonlinear shift.

In fact (14) is stable for some stability number $r$, if the nonlinear shift in (20) remains finite for the given perturbation amplitude $\epsilon$ and initial condition $u_0$. In order to derive an exact analytical relation for the stability region, the nonlinear shift in (20) is rearranged as follows.

$$\text{Nonlinear Shift} = \sum_{i=1}^{\infty} C(i)\, \frac{r^i}{(1-r)^{2i+1}} (\hat{\epsilon})^i = \frac{1}{1-r} \sum_{i=1}^{\infty} C(i)\, \frac{r^i}{(1-r)^{2i}} (\hat{\epsilon})^i \tag{21}$$

Substituting the Catalan sequence from (126) into (21) yields

$$\text{Nonlinear Shift} = \frac{1}{1-r} \sum_{i=1}^{\infty} \frac{(2i)!}{i! \times (i+1)!} \left( \frac{r\hat{\epsilon}}{(1-r)^2} \right)^i \tag{22}$$

Therefore in order to find criteria for convergence, it is only required to find the convergence of (22). To achieve more compact notation define

$$\theta = \frac{r\hat{\epsilon}}{(1-r)^2}. \tag{23}$$

where $\theta$ is named here as *Nonlinear Stability Number*[4]. Therefore

$$\text{Nonlinear Shift} = \frac{1}{1-r} \sum_{i=1}^{\infty} \frac{(2i)!}{i! \times (i+1)!} (\theta)^i \tag{24}$$

Thus the primary goal is to find the conditions for which the above series converges. Using the generalized hypergeometric function, it can be shown that

[3] as shown in (117), (119), (121), (122), (123), (124) and (125)

[4] According to analysis in § 5

$$\sum_{i=1}^{k} \frac{(2i)!}{i! \times (i+1)!} (\theta)^i = \frac{4\,\theta}{\left(1+\sqrt{1-4\,\theta}\right)^2} - \frac{{}_2F_1(1, k+\frac{3}{2};\, k+3;\, 4\,\theta)\,\theta^{k+1}\,(2\,k+2)!}{(k+1)!\,(k+2)!} \tag{25}$$

where the standard hypergeometric function [6, 7] is expanded in terms of Gamma functions as follows

$$_2F_1(1, k+\frac{3}{2};\, k+3;\, 4\,\theta) = \sum_{j=0}^{\infty} \frac{(4\theta)^j \times \frac{\Gamma(1+j)}{\Gamma(1)} \times \frac{\Gamma(k+3/2+j)}{\Gamma(k+3/2)}}{j! \times \frac{\Gamma(k+3+j)}{\Gamma(k+3)}} = \sum_{j=0}^{\infty} \frac{(4\theta)^j \times\, (1)^j \times (k+3/2)^j}{j! \times\, (k+3)^j} \tag{26}$$

Each Gamma ratio is a Pochhammer symbol. Since $k \to \infty$ then

$$_2F_1(1, k+\frac{3}{2};\, k+3;\, 4\,\theta) = \sum_{j=0}^{\infty} \frac{(4\theta)^j \times\, (1)^j \times (k+3/2)^j}{j! \times\, (k+3)^j} = \sum_{j=0}^{\infty} \frac{(4\theta)^j}{j!} \times \left(\frac{k+3/2}{k+3}\right)^j = \sum_{j=0}^{\infty} \frac{(4\theta)^j}{j!}, \tag{27}$$

which converges to

$$_2F_1(1, k+\frac{3}{2};\, k+3;\, 4\,\theta) = \exp\left(4\,\theta\right). \tag{28}$$

Substituting (28) in (25) yields

$$\sum_{i=1}^{k} \frac{(2i)!}{i! \times (i+1)!} (\theta)^i = \frac{4\,\theta}{\left(1+\sqrt{1-4\,\theta}\right)^2} - \frac{\exp(4\;\theta)\,\theta^{k+1}\,(2\,k+2)!}{(k+1)!\,(k+2)!} \tag{29}$$

For $\theta \leq 1/4$, (29) yields real values. Hence for $k \to \infty$, (29) reduces to

$$\sum_{i=1}^{k} \frac{(2i)!}{i! \times (i+1)!} (\theta)^i = \frac{4\,\theta}{\left(1+\sqrt{1-4\,\theta}\right)^2}, \quad \theta \leq \frac{1}{4}, \tag{30}$$

Substituting (30) into (24) the exact nonlinear shift can be written as

$$\text{Nonlinear Shift} = \frac{4\theta}{(1-r)\left(1+\sqrt{1-4\,\theta}\right)^2} \tag{31}$$

Also the exact converged nonlinear solution is obtained by substituting (31) into (20). The final result is

$$\frac{U}{u_0} = \frac{1}{(1-r)}\left(1 + \frac{4\theta}{\left(1+\sqrt{1-4\,\theta}\right)^2}\right). \tag{32}$$

For the fully linear case $\epsilon = 0$ hence $\hat{\epsilon} = \epsilon u_0 = 0$ for any initial condition and therefore $\theta = 0$ which according to (32), solution for the linear case is retrieved as follows.

$$\frac{U}{u_0} = \frac{1}{(1-r)} \tag{33}$$

However for the fully nonlinear case $\epsilon = 1$ hence $\hat{\epsilon} = u_0$ and therefore

$$\frac{U}{u_0} = \underbrace{\left(\frac{2}{1+\sqrt{\frac{1-2\,r+r^2-4\,r\,u_0}{(-1+r)^2}}}\right)}_{\text{Correction Factor}} \frac{1}{1-r} \tag{34}$$

Note that the correction factor converges to unity as $r \to 0$ which is consistent with the fact that for the small values of the stability number (intuitively small $\Delta t$), the problem is essentially linear. Using the definition of $\theta$ in (23) one can find the stability borders as follows. Solving (23) for $r$ yields

$$r_{1,2} = 1 + \frac{\hat{\epsilon}}{2\theta} \pm \sqrt{\frac{\hat{\epsilon}}{\theta} + \left(\frac{\hat{\epsilon}}{2\theta}\right)^2} \tag{35}$$

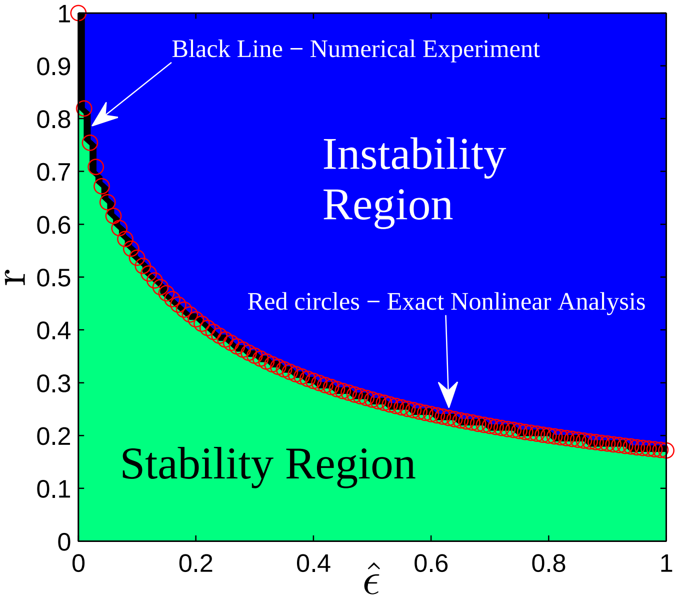


Figure 1: The stability region of nonlinear explicit DPI (14). This green area corresponds to (36).

For positive perturbations $\hat{\epsilon}$ and $\theta$, the first root $r_1$ corresponding to the plus sign in (35) violates $|r| < 1$, i.e., the convergence interval of (117, 119, 121, 122, 123, 124, 125). Therefore only the second root is acceptable. Hence

$$r \leq 1 + \frac{\hat{\epsilon}}{2\theta} - \sqrt{\frac{\hat{\epsilon}}{\theta} + \left(\frac{\hat{\epsilon}}{2\theta}\right)^2} \tag{36}$$

The above equation determines the stability region which is plotted in green in fig.(1). For small perturbation amplitude $\epsilon \to 0$ and/or small initial condition $u_0 \to 0$, the combined perturbation amplitude $\hat{\epsilon} = \epsilon\, u_0 \to 0$ and therefore the linear stability condition $r \leq 1$ is retrieved by vertical axis $(\hat{\epsilon} = 0, r)$ according to fig.(1). The border between stability and instability regions is obtained by substituting $\theta = 1/4$ into (36) which yields

$$r = 1 + 2\,\hat{\epsilon} - 2\,\sqrt{\hat{\epsilon} + \hat{\epsilon}^2} \tag{37}$$

The result is a parabola and is plotted in fig.(1) using red circles. This is in exact agreement with the values (black line) obtained from the numerical solution to (14) using a brute-force method for parameters $0 \leq r, \hat{\epsilon} \leq 1$.

## 4 Generalization to polynomial nonlinearity

The stability analysis of nonlinear explicit DPI presented in the previous section can be consistently extended to the more general case where the residual is assumed to be a polynomial function of the dependent variable. The result is presented as follows.

**Conjecture.** *The exact solution to the following explicit Discrete Picard Iteration*

$$U_{n+1} = U_0 + r\left(1 + \epsilon U_n^Z\right) U_n, \quad Z = 1, 2, 3, \ldots \tag{38}$$

*is*

$$\frac{U}{u_0} = \left(1 + \sum_{i=1}^{\infty} C(i, Z)\,\theta^i\right) \frac{1}{1 - r}, \tag{39}$$

*where $\hat{\epsilon} = \epsilon u_0^Z$ is the combined perturbation amplitude and*

$$\theta = \frac{r\hat{\epsilon}}{(1-r)^{Z+1}} \tag{40}$$

*is the Nonlinear Stability Number and $C(i,Z)$ is a generalized form of Catalan sequence given as*

$$C(i,Z) = \frac{binomial((Z+1)\times i, i)}{Z\times i+1} = \frac{((Z+1)\times i)!}{i!\times(Z\times i+1)!}. \tag{41}$$

*In addition, the stability border is the solution to*

$$\tilde{r}^{Z+1} + b\tilde{r} = b, \tag{42}$$

*where $\tilde{r} = 1-r$ and $b = \frac{(Z+1)^{Z+1}}{Z^Z}\hat{\epsilon}$.*

The above conjecture is validated for $Z = 1, 2, 3$ using symbolic processing [9]. The generalized Catalan sequence given in (41) is known in Combinatorics as the Pfaff-Fuss-Catalan or k-Raney sequence [2]. It is used in Graph Theory to enumerate (Z-ary) trees (rooted, ordered, incomplete) with $Z$ vertices including the root [1].

This conjecture shed light on the mechanisms of nonlinear numerical instability. Obviously, the stability is governed by the convergence of $\sum_{i=1}^{\infty} C(i,Z)\,\theta^i$ in (39). To understand this, it is better to find the converged value of the series for $Z = 2, 3, 4, 5, \ldots$ by method of mathematical induction. Case $Z = 1$ was studied before. For $Z = 2$, one can write

$$\sum_{i=1}^{\infty} C(i,2)\,\theta^i = \theta \times {}_3\mathrm{F}_2(1, \frac{4}{3}, \frac{5}{3}; 2, \frac{5}{2}; \frac{27}{4}\theta) \tag{43}$$

where the above hypergeometric series ${}_3\mathrm{F}_2$ is convergent if $\frac{27}{4}\theta \le 1$. Therefore the stability border is obtained as

$$\theta_{max}(2) = \frac{4}{27} = \frac{Z^Z}{(Z+1)^{Z+1}}. \tag{44}$$

Similarly $Z = 3$ yields

$$\sum_{i=1}^{\infty} C(i,3)\,\theta^i = \theta \times {}_4\mathrm{F}_3(1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4}; \frac{5}{3}, 2, \frac{7}{3}; \frac{256}{27}\theta) \tag{45}$$

which is convergent for

$$\theta_{max}(3) = \frac{27}{256} = \frac{Z^Z}{(Z+1)^{Z+1}}. \tag{46}$$

For $Z = 4$ the partial sum reduces to

$$\sum_{i=1}^{\infty} C(i,4)\,\theta^i = \theta \times {}_5\mathrm{F}_4(1, \frac{6}{5}, \frac{7}{5}, \frac{8}{5}, \frac{9}{5}; \frac{3}{2}, \frac{7}{4}, 2, \frac{9}{4}; \frac{3125}{256}\theta) \tag{47}$$

which yields

$$\theta_{max}(4) = \frac{256}{3125} = \frac{Z^Z}{(Z+1)^{Z+1}}. \tag{48}$$

Similarly for $Z = 5$ one obtains

$$\sum_{i=1}^{\infty} C(i,5)\,\theta^i = \theta \times {}_6\mathrm{F}_5(1, \frac{7}{6}, \frac{4}{3}, \frac{3}{2}, \frac{5}{3}, \frac{11}{6}; \frac{7}{5}, \frac{8}{5}, \frac{9}{5}, 2, \frac{11}{5}; \frac{46656}{3125}\theta) \tag{49}$$

which is convergent for

$$\theta_{max}(5) = \frac{3125}{46656} = \frac{Z^Z}{(Z+1)^{Z+1}}, \tag{50}$$

Therefore it is concluded that for arbitrary $Z$

$$\theta_{max}(Z) = \frac{Z^Z}{(Z+1)^{Z+1}}. \tag{51}$$

To understand the effect of *increasing nonlinearity*, i.e. $Z$ on the stability region, a geometrical interpretation of (40) is possible. At stability border $\theta = \theta_{max}$ or

$$\theta_{max} = \frac{r\hat{\epsilon}}{(1-r)^{Z+1}} \tag{52}$$

where $\theta_{max}$ is given in (51). Once the above equation is solved the maximum allowable stability number $r_{max}$ can be precisely determined. Unfortunately (52) is $Z+1$ degree polynomial equation and can't be solved analytically. However geometrical interpretative tools can be used. Here (52) is rearranged to define function $y$ as the below

$$y = (1-r)^{Z+1} = \left(\frac{\hat{\epsilon}}{\theta_{max}}\right) r \tag{53}$$

The set of curves $y = (1-r)^{Z+1}$ and $y = (\hat{\epsilon}/\theta_{max})r$ intersect at some point $0 \leq r_{max} < 1$ which is a solution to the original unsolvable nonlinear equation (52). This is schematically shown in fig.(2) where the red curves represent the lhs and rhs of (53).

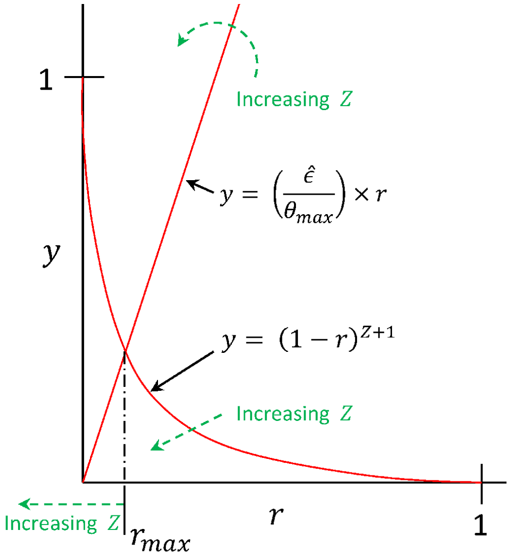


Figure 2: The effect of increasing the degree of nonlinearity '$Z$' on the maximum allowable stability number.

According to fig.(2), with increasing $Z$ the value of '$\theta_{max}$' defined in (51) decreases, hence the slope of the straight line increases. On the other hand, increasing $Z$ forces the curve to "bow" closer to the origin. As the total result, the point of intersection of these two curves moves closer to the origin and this proves that the maximum stability number decreases.

It should be noted that the stability border is always a *canonical curve*. This can be easily shown by writing

$$\frac{r\hat{\epsilon}}{(1-r)^{Z+1}} = \theta_{max} = \frac{Z^Z}{(Z+1)^{Z+1}} \tag{54}$$

or

$$\left(\frac{1-r}{1+Z}\right)^{1+Z} - \left(\frac{\hat{\epsilon}}{Z^Z}\right) r = 0, \tag{55}$$

where the stability region is specified by

$$r \leq \text{RootsOf}\left[\left(\frac{1-r}{1+Z}\right)^{1+Z} = \left(\frac{\hat{\epsilon}}{Z^Z}\right) r\right]. \tag{56}$$

Equation (55) can be written in the standard canonical form by changing variable $\tilde{r} = 1 - r$

$$\tilde{r}^{Z+1} + b\tilde{r} = b, \tag{57}$$

where $b$ is a constant given as

$$b = \frac{(Z+1)^{Z+1}}{Z^Z}\hat{\epsilon}. \tag{58}$$

# 5 Nonlinear Stability Analysis of Implicit DPI

The perturbation analysis of § 3 can be applied to the case where DPI is performed implicitly. In this case the residual vector in (5) can be written as the weighted average between two iterative steps

$$\mathbf{u}_{n+1} = \mathbf{u}_0 + \Delta t \mathbb{S} \otimes \left( a\, R(\mathbf{u}_{n+1}) + b\, R(\mathbf{u}_n) \right). \tag{59}$$

where weights satisfy $a + b = 1$ and $0 \leq a, b \leq 1$. For $a = 0$ and $b = 1$ the explicit DPI (4) is retrieved. For $a = 1$ and $b = 0$, (59) yields

$$\mathbf{u}_{n+1} = \mathbf{u}_0 + \Delta t \mathbb{S} \otimes R(\mathbf{u}_{n+1}). \tag{60}$$

Substituting the linearization of residual, i.e. $R(\mathbf{u}_{n+1}) \simeq R(\mathbf{u}_n) + \partial R(\mathbf{u}_n)/\partial \mathbf{u}_n\, (\mathbf{u}_{n+1} - \mathbf{u}_n)$ in (60) yields

$$\left( \mathbb{I} - \Delta t \mathbb{S} \otimes \left. \frac{\partial R}{\partial \mathbf{u}_n} \right|_{\mathbf{u}_n} \right) \mathbf{u}_{n+1} = \mathbf{u}_0 + \Delta t \mathbb{S} \otimes \left( R(\mathbf{u}_n) - \left. \frac{\partial R}{\partial \mathbf{u}_n} \right|_{\mathbf{u}_n} \mathbf{u}_n \right) \tag{61}$$

Following th assumptions made in § 2, (61) reduces to

$$\left( 1 - \Delta t\, \lambda \frac{\partial R_n}{\partial U_n} \right) U_{n+1} = u_0 + \Delta t\, \lambda \left( R_n - \frac{\partial R_n}{\partial U_n} U_n \right) \tag{62}$$

where the Jacobian is the derivative of the residual defined in eq.(15)

$$\frac{\partial R_n}{\partial U_n} = c + 2\, c\, \epsilon\, U_n = c \left( 1 + 2\, \epsilon\, U_n \right), \tag{63}$$

Substituting (63) and (15) into (62) results in

$$\left( 1 - \Delta t\, \lambda\, c \left( 1 + 2\, \epsilon\, U_n \right) \right) U_{n+1} = u_0 + \Delta t\, \lambda\, \left( cU_n + c\epsilon U_n^2 - c \left( 1 + 2\, \epsilon\, U_n \right) U_n \right) \tag{64}$$

Using the definition of the stability number $r = c\lambda\Delta t$ (64) simplifies to

$$\begin{aligned} \left( 1 - r \left( 1 + 2\, \epsilon\, U_n \right) \right) U_{n+1} &= u_0 - r\epsilon U_n^2 \\ U_{n=0} &= u_0, \end{aligned} \tag{65}$$

This is the sequence that is analyzed here. The perturbation series (16) is then substituted into (65) which generates expressions for perturbation amplitudes. It is shown in Appendix (B) that the i$^{th}$ perturbation amplitude of implicit DPI is written as

$$\frac{u_{i,n}}{u_0^{i+1}} = C(i) \frac{r^i}{(1-r)^{2i+1}} \quad i = 0, 1, 2, \ldots, \tag{66}$$

Where $C(i)$ are the Catalan numbers. Comparing the above with (125) it is clear that both results are equal except $r$ is not constrained in (139) since its convergence is independent of $n$. Substituting (139) into the perturbation series (16) yields

$$\frac{U}{u_0} = \frac{1}{1-r} + \frac{1}{1-r} \sum_{i=1}^{\infty} \frac{(2i)!}{i! \times (i+1)!}\, \theta^i, \quad \hat{\epsilon} = \epsilon u_0, \quad \theta = \frac{r\hat{\epsilon}}{(1-r)^2} \tag{67}$$

This is consistent with explicit DPI for $|r| \leq 1$ (see (22)). In fact the stability and convergence of implicit DPI is only governed by the convergence of the nonlinear shift in (67) not the original stability number $r$. Therefore the nonlinear stability number $\theta$ acts like a stability number governing the nonlinear nature of the residual and this is the reasoning behind its name.

According to (30), $\theta$ must be less than or equal to 1/4 so that the nonlinear shift converges. This implies that

$$\left| \frac{r\hat{\epsilon}}{(1-r)^2} \right| \leq \frac{1}{4} \tag{68}$$

The stability regions of (68) is shown in figure (3-Left). The area under $r = 1$ is exactly equal to the nonlinear stability theory of the explicit DPI described by (36) and presented in fig.(1).

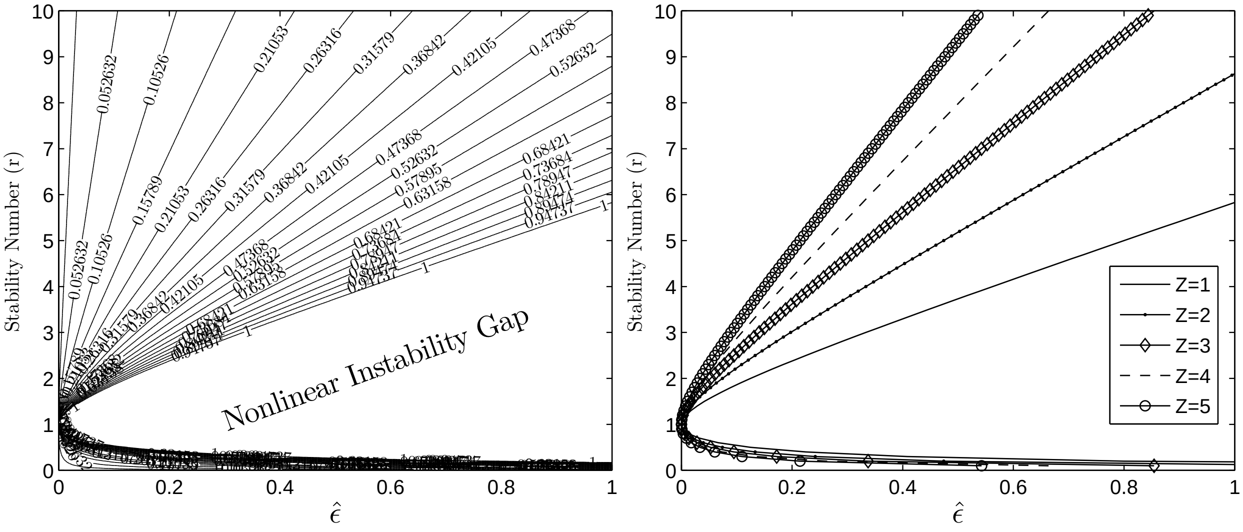


Figure 3: Left) A contour plot of $\frac{4\, r\hat{\epsilon}}{(1-r)^2}$ versus $\hat{\epsilon}$ and $r$. Right) Variation of maximum stability number for increasing nonlinearity. The nonlinear instability gap increases as nonlinearity $Z$ increases.

For the case $r > 1$ the nonlinear implicit DPI still remains stable. However there is a parabolic nonlinear instability gap which must be avoided in practice. This important results reveals a weakness of the implicit DPI. While it is *linearly* unconditionally stable for $a = 1$ and $b = 0$[5], it has a instability gap due to nonlinearity of the residuals. This analysis can be extended to general nonlinear polynomial residual where $R_n = c\left(1 + \epsilon U_n^Z\right) U_n,\ Z = 1, 2, 3, \ldots$. According to eq.(51), in this case the bound for the modified nonlinear stability number $\theta$ is

$$\theta_{max} = \frac{r_{max}\hat{\epsilon}}{(1 - r_{max})^2} = \frac{Z^Z}{(Z+1)^{Z+1}}, \tag{69}$$

where $Z = 1$ is the second-order nonlinearity (see (68). A plot of the stability number versus $\hat{\epsilon}$ is shown in fig.(3-Right). As nonlinearity increases, i.e. $Z$ increases the instability gap widens rapidly.

It can be conclude that in practice the governing equations should be slightly nonlinear or with small initial conditions. In this case either $Z$, $\epsilon$ or $u_0$ are small thus $\hat{\epsilon}$ is small enough to neglect the instability gap according to fig.(3-Right).

# 6 Generalization to the stability of system of PDEs

The assumptions made in § 2 transformed the general nonlinear systems (4) and (5) into scalar equations (14) and (62) which in this case an exact analysis was possible. However this analysis still can be utilized when (4) and (5) are considered to be system of arbitrary size. For a moment, lets assume

[5] See fig.(3-Left) for linear case ($\hat{\epsilon} = 0, r = \text{arbit.}$) is always in the stability region.

that the PDE (1) is not discretized. In this case, consider the corresponding integral form of (1) i.e. $v_k = v_{0k} + \int_{t_0}^{t} G\left(v_k(\xi), \frac{\partial^i v_k(\xi)}{\partial x^i}, \frac{\partial^j v_k(\xi)}{\partial y^j}, \dots\right) d\xi$ and introduce the space-time analytical operator $\tilde{G} = \int_{t_0}^{t} G$. Then one can write

$$v_k = v_{0k} + \tilde{G} \tag{70}$$

The analytical residual $\tilde{G}$ is now expanded $\tilde{G} = \tilde{G}\Big|_{v_{0k}} + \mathbf{J}\Delta v_k + \frac{1}{2}\Delta {v_k}^T\, \mathbf{H}\, \Delta v_k + H.O.T$ where $\mathbf{J} = \frac{\partial \tilde{G}}{\partial v_k}\Big|_{v_{0k}}$ is the Jacobian and $\mathbf{H} = J\left(\nabla \tilde{G}\right)(v_{0k})$ is the Hessian matrix evaluated at $v_{0k}$ and $\Delta v_k = v_k - v_{0k}$. This is analogous to the procedure in § 2 for the scalar case. Doing so (70) yields

$$v_k = v_{0k} + \tilde{G}\Big|_{v_{0k}} + \mathbf{J}\Delta v_k + \frac{1}{2}\Delta {v_k}^T\, \mathbf{H}\, \Delta v_k + H.O.T, \tag{71}$$

or

$$\|\Delta v_k\|_p \leq \|\, \tilde{G}\Big|_{v_{0k}} \|_p + \|\mathbf{J}\Delta v_k\|_p + \frac{1}{2}\|\Delta {v_k}^T\, \mathbf{H}\, \Delta v_k + H.O.T.\|_p \tag{72}$$

On the other hand consider the following linear Sturm-Liouville problem

$$\mathbf{J}\, \Delta v_k = \frac{\partial \tilde{G}}{\partial v_k}\Bigg|_{v_{0k}} \Delta v_k = \lambda_k \Delta v_k, \tag{73}$$

which can be solved analytically for fair broad range of PDEs with prescribed boundary conditions since it is a linear equation. The supremum of the eigenvalue spectrum of (73) is denoted by $r = \max\{|\lambda_k|\}$. Therefore using (73), (72) can be bounded by

$$\|\Delta v_k\|_p \leq \|\, \tilde{G}\Big|_{v_{0k}} \|_p + \|r\Delta v_k\|_p + \frac{1}{2}\|\Delta {v_k}^T\, \mathbf{H}\, \Delta v_k + H.O.T.\|_p \tag{74}$$

According to the discussion in § 2 the second derivative can be related to the first derivative using a perturbation parameter. As a generalization to (10), one can write

$$\frac{1}{2}\, \|\Delta v_k^T\, \mathbf{H}\, \Delta v_k + H.O.T\|_p = \epsilon\, \|\mathbf{J}\, \Delta v_k\|_p\, \|\Delta v_k\|_p \tag{75}$$

for some arbitrary $\epsilon$ in the *entire* space-time. The parameter $\epsilon$ is small when $v_k$ is close to a linear functional according to § 2. However, as mentioned before, there is no restriction on the size of $\epsilon$ since perturbation series is not truncated. Thus (75) should always hold. Substituting (75) into (74) yields

$$\|\Delta v_k\|_p \leq \|\, \tilde{G}\Big|_{v_{0k}} \|_p + r\|\Delta v_k\|_p + \epsilon\, \|\mathbf{J}\, \Delta v_k\|_p\, \|\Delta v_k\|_p \tag{76}$$

Substituting (73) in (76) yields

$$\|\Delta v_k\|_p \leq \|\, \tilde{G}\Big|_{v_{0k}} \|_p + r\|\Delta v_k\|_p + \epsilon\, \|r\Delta v_k\|_p\, \|\Delta v_k\|_p \tag{77}$$

Introducing $V = \|\Delta v_k\|_p$ and $V_0 = \|\, \tilde{G}\Big|_{v_{0k}} \|_p$, (77) can be written as

$$V \leq V_0 + rV + \epsilon r V^2 \tag{78}$$

or

$$V \leq V_0 + (1 + \epsilon V)\, rV \tag{79}$$

which is analogous to (14). The iterative class $V_{n+1} \leq V_0 + (1 + \epsilon V_n)\, rV_n$ is the explicit Analytical Picard Iteration (API) for the norm of the solution satisfying the general PDE (1). Since $V_n \geq 0$ and $r \geq 0$ the perturbation method of § 3 can be consistently used here. Perturbing $V_n$ in the terms of $\epsilon$ similar to (16) and the solving the corresponding recursive sequences one will obtain

$$\frac{V}{V_0} \leq \frac{1}{1-r} + \sum_{i=1}^{\infty} C(i)\, \frac{r^i}{(1-r)^{2i+1}} \hat{\epsilon}^i, \qquad \hat{\epsilon} = \epsilon V_0 \tag{80}$$

which remains bounded if $\left|\theta = r\hat{\epsilon}/(1-r)^2\right| \leq 1/4$ according to discussion in § 3. It can be shown that the same condition applies when the API is performed implicitly.

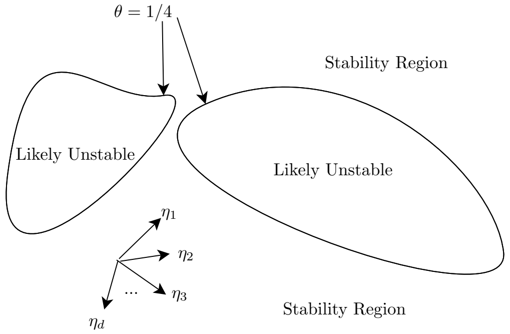


Figure 4: The stability region of general nonlinear system of PDEs (1) in the Fourier space granted by (86).

There is an interesting discussion regarding the linear Sturm-Liouville problem (73). Since $\mathbf{J}$ is always a *linear operator* it can be represented via

$$\mathbf{J} = \sum_{l=1}^{d} \sum_{m} a_{lm}\left(v_{0k}\right) \frac{\partial^m}{\partial x_l^m} \tag{81}$$

in the d-dimensional space $\mathbb{R}^d$. On the other hand, the Fourier transform

$$\hat{\Delta v}_k = \int_{\mathbb{R}^d} \Delta v_k \, e^{-\mathbf{i} x_j \, \eta_j} dx_1 \, dx_2 \, \ldots \, dx_d, \quad j = 1 \ldots d, \tag{82}$$

maps $\Delta v_k$ defined in the physical domain $x_j$ to $\hat{\Delta v}_k$ in the frequency domain $\eta_j$. Taking Fourier transform of (73) and using (82) yields

$$\int_{\mathbb{R}^d} \mathbf{J} \, \Delta v_k \, e^{-\mathbf{i} x_j \, \eta_j} dx_1 \, dx_2 \, \ldots \, dx_d = \lambda_k \int_{\mathbb{R}^d} \Delta v_k \, e^{-\mathbf{i} x_j \, \eta_j} dx_1 \, dx_2 \, \ldots \, dx_d, \tag{83}$$

or

$$\left( \sum_{l=1}^{d} \sum_{m} a_{lm}\left(v_{0k}\right) \left(\mathbf{i} \, \eta_l\right)^m \right) \hat{\Delta v}_k = \lambda_k \, \hat{\Delta v}_k. \tag{84}$$

Therefore the k-th eigenvalue of the Sturm-Liouville problem is obtained as

$$\lambda_k = \sum_{l=1}^{d} \sum_{m} a_{lm}\left(v_{0k}\right) \left(\mathbf{i} \, \eta_l\right)^m, \tag{85}$$

and hence the stability number is $r = \max |\lambda_k|$. Therefore $\theta \leq 1/4$ implies that a solution to the system of nonlinear PDEs (1) remains stable and finite in space-time $\mathbb{R}^d$ if

$$|\theta| = \left| \frac{\max \left| \sum\limits_{l=1}^{d} \sum\limits_{m} a_{lm}\left(v_{0k}\right) \left(\mathbf{i} \, \eta_l\right)^m \right| \hat{\epsilon}}{\left(1 - \max \left| \sum\limits_{l=1}^{d} \sum\limits_{m} a_{lm}\left(v_{0k}\right) \left(\mathbf{i} \, \eta_l\right)^m \right| \right)^2} \right| \leq \frac{1}{4} \tag{86}$$

The iso-level contours of (86) for $\theta = 1/4$ (which are not necessarily closed curves) define the stability borders as depicted in fig.(4) where stability is guaranteed by (86) outside of these regions. Inside these regions, however, the solution may or may not be stable because (86) yields a least upperbound.

Also any numerical solution to (1) is nonlinearly stable if (86) is valid when the frequency $\eta_l$ is replaced with the frequency modified by the numerical method. Such a modified frequency can be easily obtained using Discrete Fourier Transform.

**Example:** For nonlinear Poisson equation on defined on $x \in I = [-1, 1]$

$$\frac{\partial v}{\partial t} = \frac{\partial^2 R}{\partial x^2}, \quad R = v + v^2, \tag{87}$$

with IBVs

$$v\left(x = -1, t\right) = v\left(x = 1, t\right) = 0, \quad v\left(x, t = 0\right) = \left(1 - x\right)\left(1 + x\right), \tag{88}$$

the explicit DPI (4) leads to

$$u_{n+1} = u_0 + \Delta t \, R\left(u_n\right). \tag{89}$$

Therefore (70) yields

$$\tilde{G} = \Delta t \, R\left(u_n\right) = \Delta t \, \frac{\partial^2}{\partial x^2} \left(u_n + u_n^2\right) \tag{90}$$

The Jacobian is

$$J = \left. \frac{\partial \tilde{G}}{\partial u} \right|_{u_0} = \Delta t \, \left( \frac{\partial^2}{\partial x^2} \Box + 2 \frac{\partial^2}{\partial x^2} u_0 \Box \right), \tag{91}$$

and hence the general Sturm-Liouville problem (73) reduces to the following

$$\Delta t\,\frac{\partial^2}{\partial x^2}\left(\left(1+2u_0\right)\Delta u_k\right)=\lambda_k\,\Delta u_k \tag{92}$$

with the stability criteria given by

$$|\theta|=\left|\frac{r\hat{\epsilon}}{\left(1-r\right)^2}\right|\leq\frac{1}{4}, \tag{93}$$

where $r=\max|\lambda_k|$. If (92) is solved *analytically* for *infinite* eigenvalues $\lambda_{k=1\ldots\infty}$ then (93) leads to semi-discrete stability regions. In the semi-discrete approach, an infinite dimensional banded matrix is indeed considered for the Jacobian operator (91) and the equations are only discretized in time. This is while, in a fully discrete numerical solution, a finite-dimensional matrix (not necessarily banded) is employed. In this case, numerical stability regions can be investigated by finding whether (93) is satisfied for finite-dimensional eigen-spectrum $\lambda_{k=1\ldots M}$. These eigenvalues uniquely correspond to the numerical method used for discretization and also the type of boundary conditions used. This incorporates all details of a numerical solution in the current stability theory in a unified and consistent way. Therefore for different discretization method and/or BCs types, the discretized form of Jacobian matrix given in (91) changes and thus the eigen spectrum (92) changes and as a result, the stability regions obtained from (93) changes accordingly.

Focusing on the numerical stability, consider a symmetric second-order discretization of Laplacian $\frac{\partial^2}{\partial x^2}\approx\frac{1}{\Delta x^2}\,\mathrm{diag}\left(1,-2,1\right)_{M\times M}$ where $M+2$ collocation points (including the boundaries) are used on interval $I$. In this case (92) can be written as

$$\left[\beta\,.\,\mathrm{diag}\left(1,-2,1\right)\,.\,\mathrm{diag}\left(1+2u_0(x_i)\right)\right]\Delta u_k=\lambda_k\,\Delta u_k \tag{94}$$

where $\beta=\frac{\Delta t}{\Delta x^2}$ is the CFL number. According to [10], the eigenvalues of the tridiagonal matrix diag $(1,-2,1)$ can be obtained as

$$\bar{\lambda}_k=2\left(\cos\left(\frac{k\,\pi}{M+1}\right)-1\right)\quad,k=1\ldots M. \tag{95}$$

Substituting (95) in (94) yields

$$\lambda_k=2\beta\left(\cos\left(\frac{k\,\pi}{M+1}\right)-1\right)\left(1+2(1-x_k)(1+x_k)\right) \tag{96}$$

where $x_k=-1+2\,k/(M+1)$. For convenience, define a new variable

$$\gamma=\frac{k}{M+1} \tag{97}$$

Substituting $x_k=-1+2\gamma$ and (97) in (96) yields

$$\lambda_k=2\beta\left(\cos(\gamma\pi)-1\right)\left(1+8\gamma\left(1-\gamma\right)\right) \tag{98}$$

Hence $r$ can be obtained by finding the maximum value of $\lambda_k$ over $I$. The extremum happens at the root of the derivative of (98) which is a nonlinear equation. Therefore an exact solution is not possible and hence it is estimated as follows.

$$r=\max|\lambda_k|\approx 8.562\beta \tag{99}$$

At this moment, the value of $\hat{\epsilon}$ is required according to (93) to complete the analysis. Since $\hat{\epsilon}=\epsilon V_0$, it is easier to compute $V_0$ and $\epsilon$ separately. The value of $V_0$ is obtained as follows.

$$V_0=\|\left.\tilde{G}\right|_{u_0}\|_p=\|\Delta t\frac{\partial^2}{\partial x^2}\left(u_0+u_0^2\right)\|_p \tag{100}$$

For the second-order central numerical discretization used here, the Laplacian operator in (100) should be replaced with the corresponding discretized form as below.

$$V_0=\|\frac{\Delta t}{\Delta x^2}\mathrm{diag}(1,-2,1)\,\mathrm{diag}(u_0+u_0^2)\|_p=\frac{\Delta t}{\Delta x^2}\,\max\left(\mathrm{eig}(\mathrm{diag}(1,-2,1)\,\mathrm{diag}(u_0+u_0^2))\right). \tag{101}$$

Substituting corresponding eigenvalues, (101) can be written as follows.

$$V_0 = 2\beta \left|\max\left\{(\cos(\gamma\pi) - 1)\ (u_0(\gamma) + u_0^2(\gamma))\right\}\right| \tag{102}$$

Since $x_k = -1 + 2\gamma$ then $u_0(x_k) = u_0(\gamma) = 4\gamma(1-\gamma)$ hence (102) leads to

$$V_0 = 2\beta \left|\max\left\{(\cos(\gamma\pi) - 1)\ (4\gamma + 12\gamma^2 - 32\gamma^3 + 16\gamma^4)\right\}\right| \tag{103}$$

which can be approximated as

$$V_0 \approx 5.054\,\beta \tag{104}$$

The value of the Hessian in (75) can be obtained by taking the deivative of (91) which yields

$$H = \left.\frac{\partial J}{\partial u}\right|_{u_0} = 2\,\Delta t\,\frac{\partial^2}{\partial x^2}\Box \tag{105}$$

Also note that $H.O.T = 0$ in (75) since higher derivatives of Hessian are identically zero. Substituting (105) and (91) in (75) yields

$$\left\|\frac{\partial^2}{\partial x^2}\Box\right\|_p = \epsilon\left\|\frac{\partial^2}{\partial x^2}(\Box + 2\,u_0\Box)\right\|_p = \epsilon\frac{r}{\Delta t}, \tag{106}$$

or equivalently

$$\epsilon = \frac{\max\left|\mathrm{eig}\left(\frac{\partial^2}{\partial x^2}\Box\right)\right|}{\max\left|\mathrm{eig}\left(\frac{\partial^2}{\partial x^2}\left(\Box + 2\,u_0\Box\right)\right)\right|} = \frac{\Delta t\max\left|\mathrm{eig}\left(\frac{\partial^2}{\partial x^2}\Box\right)\right|}{r} = \frac{\Delta t\max\left|\mathrm{eig}\left(\frac{\mathrm{diag}(1,-2,1)}{\Delta x^2}\right)\right|}{r}. \tag{107}$$

Hence

$$\epsilon = \frac{\beta\max\left|\mathrm{eig}\left(\mathrm{diag}(1,-2,1)\right)\right|}{r} \tag{108}$$

Substituting (95) in (108) yields

$$\epsilon = \frac{2\,\beta\left(1 - \cos(\frac{M\pi}{M+1})\right)}{r} \approx \frac{4\beta}{r} \tag{109}$$

Therefore from (109) and (104) it is concluded that

$$\hat{\epsilon} = \epsilon\,V_0 = \frac{20.216\beta^2}{r} \tag{110}$$

Substituting (110) and (99) in (93) yields

$$\left|\frac{20.216\,\beta^2}{(1 - 8.562\beta)^2}\right| \leq \frac{1}{4} \tag{111}$$

Solving (111) it can be easily verified that the stability region is $0 \leq \beta \leq 0.0570$. This is a great reduction in the allowable CFL number $\beta$ compared to the linear Poisson equation where $0 \leq \beta \leq 0.5$. This spectacular result can not be justified using linear stability theories.

To validate the analytical stability region $0 \leq \beta \leq 0.0570$, a computer program [9] is written which solves nonlinear Poisson equation (87) with the given initial and boundary conditions using second-order spatial discretization. The value of CFL number is experimentally modified to find the stability region. It is found that $0 \leq \beta \leq 0.0885$ which is consistent with the analytical result since the current theory gives a least upperbound.

# 7 Conclusions

The analysis presented in this paper determines the stability region of nonlinear system of PDE (1) when the corresponding space-time integral (2) is discretized in explicit form (4) and implicit form (5). Important conclusions are summarized as follows.

1. The analysis presented in this paper determines the shift that occurs in the linear stability criteria due to the existence of nonlinear terms in residual. This shift was shown to be exact when (4) and (61) are scalar and can be regarded as a least upperbound when (4) and (61) are general system of equations. This answers the first question in the introduction.
2. For both explicit and implicit discretization, there is a canonical instability gap in the $r-\hat{\epsilon}$ plane for polynomial nonlinearity (see fig.(3)-left). Outside of this region, the solution remains stable while inside of this gap, the scalar version of (4) and (61) are guaranteed to be unstable. However the general form (4) and (61) may or may not be unstable in this region according to fig.(4) and discussions in § 6. This result implies that even if linearization is done perfectly, and the Jacobian of linearization is computed analytically, and a linearly unconditional stable is applied for the discretization of (1), then still the resulting numerical method is nonlinearly unstable inside the instability gap. This address question (2) in the introduction.
3. The area of the instability gap increases when the degree of the nonlinearity of the residual increases (see fig.(3)-right). In this case, the space-time discretization of (1) is strongly limited by nonlinear instability. However, from a practical point of view, application of a different discretization of the original Cauchy problem such as multi-step Runge-Kutta methods may or may not reduce the nonlinear instability gap. This prompts further investigation of the nonlinear instability of RK methods which may or may not be canonical.

AG acknowledge help and support form SimCenter University of Tennessee at Chattanooga.

# A Derivation of Perturbation Amplitudes For Explicit DPI

The details of derivation of perturbation amplitudes for explicit DPI is presented as follows. For $i = 0$, the corresponding equation would be

$$\begin{aligned} u_{0,n+1} &= r\, u_{0,n} + u_0 \\ u_{0,n=0} &= u_0. \end{aligned} \tag{112}$$

which is the only perturbation amplitude when the residual is linear, i.e. $R = cu$. Matching the coefficient of $\epsilon^1$ yields

$$\begin{aligned} u_{1,n+1} &= r u_{1,n} + r u_{0,n}^2 \\ u_{1,n=0} &= 0 \end{aligned} \tag{113}$$

Similarly for the coefficient of $\epsilon^2$ one obtains

$$\begin{aligned} u_{2,n+1} &= r u_{2,n} + 2\, r\, u_{0,n}\, u_{1,n} \\ u_{2,n=0} &= 0. \end{aligned} \tag{114}$$

The coefficients of $\epsilon^3$ and $\epsilon^4$ generates the following sequences.

$$\begin{aligned} u_{3,n+1} &= r\, u_{3,n} + 2\, r\, u_{0,n}\, u_{2,n} + r\, {u_{1,n}}^2 \\ u_{4,n+1} &= r\, u_{4,n} + 2\, r\, \left(u_{2,n}\, u_{1,n} + u_{0,n}\, u_{3,n}\right) \\ u_{3,n=0} &= u_{4,n=0} = 0, \end{aligned} \tag{115}$$

It should be noted that the sequences generated in this way always consist of a linear core in the form of $r\, u_{i,n}$ plus a nonlinear source term which *only* depends on the previous Picard iterations. This is the desired property of the perturbation method which makes it possible to analytically obtain the i$^{th}$ nonlinear amplitude using recursive solution of linear sequences. Similar expressions can be derived for higher order terms; however, the resulting expressions are very long to be included here. A symbolic was written to derive and solve the equation for the i$^{th}$ perturbation amplitude[9].

At this point, the perturbation amplitudes need to be solved recursively. First (112) is solved yielding

$$\frac{u_{0,n}}{u_0} = \frac{r^{n+1}-1}{r-1} \tag{116}$$

which is the partial sum of the first $n$ terms of geometric series obtained by recursively expanding (112). Equation (116) converges at arbitrarily large iterations if and only if $|r| < 1$. In this case the converged solution is

$$\frac{u_{0,\infty}}{u_0} = -\frac{1}{r-1} \tag{117}$$

Since for the linear residual, $u_{0,n}$ is the only available perturbation amplitude it can be concluded that the sufficient linear stability requirement is $|r| < 1$. The second perturbation amplitude is obtained by substituting (116) into (113) and finding the partial sum. The final result is

$$\frac{u_{1,n}}{u_0^2} = \frac{\left(-2\,r^{2+n}+2\,r^{n+1}\right)n}{\left(-1+r\right)^3} + \frac{-r+r^{2n+2}+r^{n+1}-r^{2+n}}{\left(-1+r\right)^3}, \tag{118}$$

which converges to

$$\frac{u_{1,\infty}}{u_0^2} = -\frac{r}{\left(r-1\right)^3} \quad \text{iff } |r| < 1 \tag{119}$$

Substituting (118) and (116) into (114) the second perturbation amplitude can be found. The final result can be written as follows.

$$\begin{aligned}\frac{u_{2,n}}{u_0^3} = &-2\,\frac{\left(-r^{4+n}+r^{3+n}+r^{2+n}-r^{n+1}\right)n^2}{\left(-1+r\right)^4\left(r^2-1\right)} - 2\,\frac{\left(2\,r^{4+2n}-2\,r^{2+2n}-r^{2+n}+r^{4+n}\right)n}{\left(-1+r\right)^4\left(r^2-1\right)}\\ &-2\,\frac{-r^{2+n}-r^{4+2n}+r^{4+n}+r^3-r^{3+3n}+r^2-r^{3+2n}+r^{3+n}}{\left(-1+r\right)^4\left(r^2-1\right)}\end{aligned} \tag{120}$$

which converges to

$$\frac{u_{2,\infty}}{u_0^3} = -2\,\frac{r^2}{\left(r-1\right)^5} \quad \text{iff } |r| < 1 \tag{121}$$

The partial sum of the third and higher amplitudes are exceedingly lengthy. The converged solutions are provided here. The third amplitude yields

$$\frac{u_{3,\infty}}{u_0^4} = -5\,\frac{r^3}{\left(r-1\right)^7} \quad \text{iff } |r| < 1 \tag{122}$$

The full partial sum of the fourth amplitude converges to

$$\frac{u_{4,\infty}}{u_0^5} = -14\,\frac{r^4}{\left(r-1\right)^9} \quad \text{iff } |r| < 1 \tag{123}$$

and the fifth amplitude converges to

$$\frac{u_{5,\infty}}{u_0^6} = -42\,\frac{r^5}{\left(r-1\right)^{11}} \quad \text{iff } |r| < 1 \tag{124}$$

A symbolic processor was used to derive full partial sums and finding limits where it is determined (using mathematical induction) that the i$^{th}$ perturbation amplitude converges to

$$\frac{u_{i,\infty}}{u_0^{i+1}} = C(i)\,\frac{r^i}{\left(1-r\right)^{2i+1}} \quad \left(i = 0,1,2,\ldots,\quad |r| \le 1\right), \tag{125}$$

where $C(i) = \{1,1,2,5,14,42,132,429\ldots\}$ is the well-known Catalan sequence [1] given explicitly as

$$C(i) = \frac{(2i)!}{i!\times(i+1)!} = \frac{\text{binomial}\,(2i,i)}{i+1}. \tag{126}$$

# B Perturbation Amplitudes of Implicit DPI

The zeroth perturbation amplitude yields

$$(1-r)u_{0,n+1} = u_0 \tag{127}$$

which is easily solved for $u_{0,n+1}$ as

$$\frac{u_{0,n+1}}{u_0} = -\frac{1}{r-1}. \tag{128}$$

Evidently $u_{0,n+1}$ for implicit DPI converges in the first iteration thus it is independent of $n$. This shows that *unlike* the zeroth perturbation amplitude of *explicit DPI* given in (116), the zeroth perturbation amplitude of implicit DPI is always stable independent of the iteration number $n$. The first amplitude is obtained as

$$u_{1,n+1} = \frac{r\,u_{0,n}\left(-2\,u_{0,n+1} + u_{0,n}\right)}{r-1} \tag{129}$$

Substituting (128) into (129) yields

$$\frac{u_{1,n+1}}{u_0^2} = -\frac{r}{\left(r-1\right)^3} \tag{130}$$

Again, by comparing (130) with (118) one realizes that the first perturbation amplitude of the implicit DPI scheme is independent of the Picard iterations. In fact the first perturbation amplitude of implicit DPI for arbitrary $r$ and $n$ is *exactly equal* to the first perturbation amplitude of explicit DPI when it converges (compare to (119)). Similarly the second perturbation amplitude is read as

$$u_{2,n+1} = 2\,\frac{r\left(-u_{0,n}\,u_{1,n+1} - u_{1,n}\,u_{0,n+1} + u_{0,n}u_{1,n}\right)}{r-1} \tag{131}$$

Substituting (128, 130) into (131) yields

$$\frac{u_{2,n+1}}{u_0^3} = -2\,\frac{r^2}{\left(r-1\right)^5} \tag{132}$$

The same conclusion again holds here whereas (132) and (121) are equal. Similarly for the third, fourth and the fifth perturbation amplitudes are obtained as follows.

$$\begin{aligned}\left(\frac{r-1}{2r}\right)u_{3,n+1} &= u_{0,n}\,u_{2,n} - u_{0,n}\,u_{2,n+1} - u_{1,n}\,u_{1,n+1} \\ &- u_{2,n}\,u_{0,n+1} + \frac{1}{2}\,u_{1,n}^2, \end{aligned} \tag{133}$$

$$\begin{aligned}\left(\frac{r-1}{2r}\right)u_{4,n+1} &= -u_{3,n}\,u_{0,n+1} - u_{0,n}\,u_{3,n+1} - u_{1,n}\,u_{2,n+1} \\ &- u_{2,n}\,u_{1,n+1} + u_{0,n}\,u_{3,n} + u_{1,n}\,u_{2,n}. \end{aligned} \tag{134}$$

$$\begin{aligned}\left(\frac{r-1}{2r}\right)u_{5,n+1} &= -u_{3,n}\,u_{1,n+1} - u_{0,n}\,u_{4,n+1} - u_{1,n}\,u_{3,n+1} \\ &- u_{2,n}\,u_{2,n+1} - u_{4,n}\,u_{0,n+1} + u_{0,n}\,u_{4,n} \\ &+ u_{1,n}\,u_{3,n} + \frac{1}{2}\,u_{2,n}^2. \end{aligned} \tag{135}$$

By substituting previous perturbation amplitudes into (133), (134) and (135) one obtains

$$\frac{u_{3,n+1}}{u_0^4} = -5\,\frac{r^3}{\left(r-1\right)^7} \tag{136}$$

$$\frac{u_{4,n+1}}{u_0^5} = -14\,\frac{r^4}{\left(r-1\right)^9} \tag{137}$$

$$\frac{u_{5,n+1}}{u_0^6} = -42\,\frac{r^5}{(r-1)^{11}} \tag{138}$$

As mentioned before, the above relations are independent of iteration number and they are in fact the exact converged value of corresponding explicit DPI relations given in (122), (123) and (124). In general the i$^{th}$ perturbation amplitude of implicit DPI is written as

$$\frac{u_{i,n}}{u_0^{i+1}} = C(i)\,\frac{r^i}{(1-r)^{2i+1}} \quad i = 0, 1, 2, \ldots . \tag{139}$$

# References

[1] N. J. A. Sloane, "The On-Line Encyclopedia of Integer Sequences", *Catalan Numbers*, `http://oeis.org/A000108`, retrieved 2011.

[2] W. Lang, "Combinatorial Interpretation of Generalized Stirling Numbers", *Journal of Integer Sequences*, **12**, 1-24, 2009.

[3] H. B. Keller, "Approximation Methods for Nonlinear Problems with Application to Two-Point Boundary Value Problems", *Mathematics of Computation*, **29**, 464-474, 1975.

[4] J.C. Lopez-Marcos and J.M. Sanz-Serna., "Stability and convergence in numerical analysis III: Linear investigation of nonlinear stability", *IMA J. Numer. Anal.*, **8**, 71-84, 1988.

[5] Magnus Pirovino, "On the Denition of Nonlinear Stability for Numerical Methods", Seminar fur Angewandte Mathematik, Eidgenossische Technische Hochschule, Technical Report 1991.

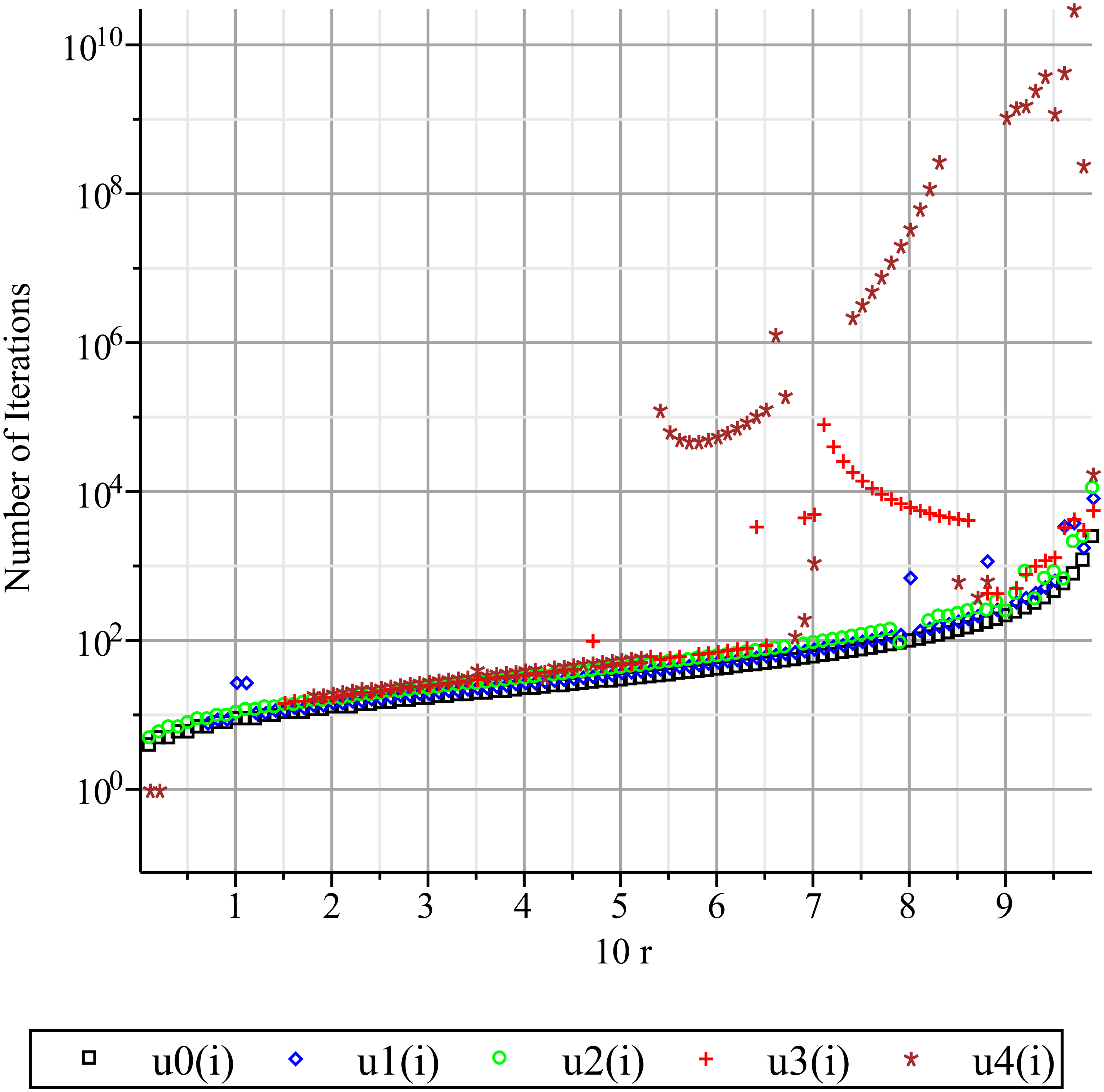
[6] M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions", *Dover Publications*, 1965.

[7] W. W. Bell, "Special Functions for Scientists and Engineers", *Dover Publications*, 2004.

[8] L. N. Trefethen, "Spectral Methods in MATLAB", SIAM, 2001.

[9] A. Ghasemi and K. Sreenivas and L. K. Taylor, "Anlysis of nonlinear stability of discrete Picard iteration using symbolic processing", SimCenter Archive, UTC, Chattanooga, 2012.

[10] W. Yueh, "Eigenvalues of Several Tridiagonal Matrices", *Applied Mathematics E-Notes*, **5**, 66-74, 2005.


Number of Iterations
$10^{10}$
$10^8$
$10^6$
$10^4$
$10^2$
$10^0$
1
2
3
4
5
6
7
8
9
10 r
□ u0(i)
◇ u1(i)
○ u2(i)
+ u3(i)
* u4(i)