Computing 37, 195–218 (1986)

Computing

On the Numerical Integration of Second-Order Initial Value Problems with a Periodic Forcing Function

P. J. van der Houwen and **B. P. Sommeijer**, Amsterdam
K. Strehmel and **R. Weiner**, Halle

Received February 25, 1986

Abstract — Zusammenfassung

On the Numerical Integration of Second-Order Initial Value Problems with a Periodic Forcing Function. Runge-Kutta-Nyström type methods and special predictor-corrector methods are constructed for the accurate solution of second-order differential equations of which the solution is dominated by the forced oscillation originating from an external, periodic forcing term. For a family of second-order explicit and linearly implicit Runge-Kutta-Nyström methods it is shown that the forced oscillation is represented with zero phase lag. For a family of predictor-corrector methods of fourth-order, it is shown that both the phase lag order and the dissipation of the forced oscillation can be made arbitrarily high. Numerical examples illustrate the effectiveness of our reduced phase lag methods.

1980 Mathematical Subject Classification: 65L05.

1982 CR Categories: G.1.7, G.1.8.

Key words: Numerical analysis, ordinary differential equations, Runge-Kutta methods, predictor-corrector methods, periodic solutions.

Zur numerischen Integration von Anfangswertaufgaben für Differentialgleichungen zweiter Ordnung mit einer periodischen Kraftfunktion. Für die numerische Behandlung von Differentialgleichungen zweiter Ordnung, bei denen die Lösung im wesentlichen durch die von einer äußeren periodischen Kraft erzwungenen Schwingung bestimmt wird, werden Diskretisierungsmethoden vom Runge-Kutta-Nyström Typ und spezielle Prädiktor-Korrektor Methoden konstruiert. Für eine Klasse expliziter und linear-impliziter Runge-Kutta-Nyström Methoden der Ordung zwei zeigen wir, daß die erzwungene Schwingung keinen Phasenfehler aufweist. Für eine Klasse von Prädiktor-Korrektor Methoden vierter Ordnung wird nachgewiesen, daß die Phasen- und Dissipationsfehlerordnung beliebig groß gemacht werden kann. Numerische Beispiele bestätigen die Wirksamkeit unserer Methoden mit reduziertem Phasenfehler.

1. Introduction

We shall be concerned with the special second-order differential equation

$$y'' = f(t, y) \tag{1.1}$$

with initial conditions $y(0) = y_0$ and $y'(0) = y'_0$. In particular, we will consider problems, where it is known in advance that the solution $y(t)$ is periodic due to some external forcing term. To be more precise, we aim at problems of the form

$$y''(t) = M(t, y)\, y + g(t, y), \tag{1.2}$$

where M is a matrix with a negative spectrum and $g(t, y(t))$ is a periodic function of t; furthermore, M and g are slowly varying with (t, y) and y, respectively. The solution component representing the forced oscillation, introduced by g, will be called the "inhomogeneous" solution component. In our analysis it will be assumed that the inhomogeneous solution component dominates the solution and forces $y(t)$ to be periodic with frequency ω.

The methods to be analysed in this paper are explicit Runge-Kutta-Nyström methods, the adaptive Runge-Kutta-Nyström methods proposed in Strehmel/Weiner [11], and the special predictor-corrector methods proposed in van der Houwen/Sommeijer [7]. Conditions will be derived for tuning these families of methods to the given problem, and concrete methods will be constructed that satisfy these conditions. The resulting methods are characterized by the property that the phase error of the inhomogeneous solution component is significantly smaller than the phase errors produced by conventional methods. Our main results are, relative to the usual test equation (cf. (2.1)), (i) families of second-order Runge-Kutta-Nyström methods of explicit type and of linearly implicit type (adaptive methods) with zero phase lag in the inhomogeneous solution component, and (ii) a family of fourth-order predictor-corrector methods of arbitrarily high phase lag order and dissipation order.

In a number of earlier papers (cf. e.g. Brusa/Nigro [1], Gladwell/Thomas [5], Thomas [12] and Strehmel/Weiner [11]), the reduction, or even the elimination, of the *inhomogeneous phase error* has already been studied. The present paper extends this work, firstly, by treating the important classes of Runge-Kutta-Nyström type methods and predictor-corrector methods in a systematic way, and secondly, by a simultaneous reduction of the *inhomogeneous phase lag and dissipation error*.

Finally, we remark that the phase lag analysis of the *homogeneous* solution component (using a homogeneous test equation) has been studied in Chawla/Rao [2], Twizell [13], van der Houwen/Sommeijer [8, 9], and Chawla/Rao/Neta [3].

2. Preliminaries

In this section we shall derive recursions for the approximate solution of the test equation

$$y''(t) = -\delta^2 y(t) + c e^{i\omega t};\ \delta^2 > 0;\ \omega^2 \neq \delta^2;\quad c, \omega \in \mathbb{R}\backslash\{0\} \tag{2.1}$$

when integrated by a numerical method. Here, δ and ω respectively correspond to the dominating frequencies in homogeneous and inhomogeneous solution components of the given equation (1.1); ω will be assumed to be given and δ represents an eigenvalue of the matrix M in (1.2). Three classes of numerical methods will be considered, viz. Runge-Kutta-Nyström methods, adaptive Runge-Kutta-Nyström methods, and predictor-corrector type methods.

Throughout this paper we use the notation

$$z_0 := -\tau^2 \delta^2,\ v_0 := \tau \omega,$$

where τ denotes the integration step of the numerical method.

2.1 Runge-Kutta-Nyström Methods

Consider the explicit, m-stage Runge-Kutta-Nyström (RKN) method

$$\begin{aligned} y_{n+1}^{(0)} &= y_n, \\ y_{n+1}^{(j)} &= y_n + \mu_j \tau \dot{y}_n + \tau^2 \sum_{l=0}^{j-1} \lambda_{jl} f(t_n + \mu_l \tau, y_{n+1}^{(l)}), \quad j = 1, \ldots, m, \ \mu_0 = 0, \mu_m = 1, \\ y_{n+1} &:= y_{n+1}^{(m)}, \ \dot{y}_{n+1} := \dot{y}_n + \tau \sum_{l=0}^{m-1} \lambda_l^* f(t_n + \mu_l \tau, y_{n+1}^{(l)}), \end{aligned} \tag{2.2}$$

where $\mu_j, \lambda_{jl}, \lambda_l^*$ satisfy certain order conditions. For a survey of order conditions for RKN methods we refer to [6].

Suppose that not all parameters μ_j, λ_{jl} and λ_l^* are used for satisfying the order conditions. Then, one may try to use the remaining degrees of freedom for increasing the accuracy when integrating the special test equation (2.1).

Theorem 2.1: *Let the polynomials $A_m(z)$, $B_m(z)$, $C_m(z)$ be defined by*

$$\begin{aligned} A_0(z) &:= 1, \ A_j(z) := 1 + z \sum_{l=0}^{j-1} \lambda_{jl} A_l(z), \quad j = 1, \ldots, m, \\ B_0(z) &:= 0, \ B_j(z) := \mu_j + z \sum_{l=0}^{j-1} \lambda_{jl} B_l(z), \quad j = 1, \ldots, m, \\ C_0(z) &:= 0, \ C_j(z) := \sum_{l=0}^{j-1} \lambda_{jl} [z C_l(z) + e^{i\mu_l \nu_0}], \quad j = 1, \ldots, m \end{aligned} \tag{2.3}$$

and let

$$\begin{aligned} A_m^*(z) &:= z \sum_{l=0}^{m-1} \lambda_l^* A_l(z), \ B_m^*(z) := 1 + z \sum_{l=0}^{m-1} \lambda_l^* B_l(z), \\ C_m^*(z) &:= \sum_{l=0}^{m-1} \lambda_l^* [z C_l(z) + e^{i\mu_l \nu_0}], \\ \Phi_2(z, \zeta) &:= \zeta^2 - (A_m(z) + B_m^*(z))\zeta + A_m(z) B_m^*(z) - A_m^*(z) B_m(z). \end{aligned} \tag{2.4}$$

Then the numerical solution of (2.1) *satisfies the recursion*

$$\Phi_2(z_0, E) y_n = c \tau^2 [C_m(z_0) e^{i\nu_0} + B_m(z_0) C_m^*(z_0) - B_m^*(z_0) C_m(z_0)] e^{in\nu_0}, \tag{2.5}$$

where E is the forward shift operator.

Proof: Application of the RKN method to (2.1) yields

$$\begin{aligned} y_{n+1} &= A_m y_n + \tau B_m \dot{y}_n + \tau^2 C_m c e^{in\nu_0}, \\ \tau \dot{y}_{n+1} &= A_m^* y_n + \tau B_m^* \dot{y}_n + \tau^2 C_m^* c e^{in\nu_0}, \end{aligned} \tag{2.6}$$

where the polynomials are evaluated at z_0. Elimination of $\dot{y}_{n+1}, \dot{y}_n, \ldots$ from the recursion (2.6) leads to the recursion (2.5). □

It should be remarked that the polynomials C_m and C_m^* depend on the value of ν_0. In fact, as we shall see later, the coefficients of the polynomials A_m, A_m^*, B_m and B_m^* will also depend on ν_0 and, in a few cases, on z_0.

2.2 Adaptive Runge-Kutta-Nyström Methods

We consider the m-stage adaptive Runge-Kutta-Nyström method (ARKN method)

$$y_{n+1}^{(0)} = y_n + \mu_0 \tau \dot{y}_n,$$

$$y_{n+1}^{(j)} = V_0((\mu_j\tau)^2 T) y_n + \mu_j \tau V_1((\mu_j\tau)^2 T)\dot{y}_n + \tau^2 \sum_{l=0}^{j-1} A_{jl} g_l, \quad j=1,\ldots,m, \tag{2.7a}$$

$$y_{n+1} := y_{n+1}^{(m)}, \quad \dot{y}_{n+1} = V_0(\tau^2 T)\dot{y}_n + \tau\left[TV_1(\tau^2 T) y_n + \sum_{l=0}^{m-1} B_l g_l\right],$$

where the functions $V_l(z)$ are defined by

$$V_0(z) := \frac{1}{2}[R_0(\sqrt{z}) + R_0(-\sqrt{z})], \quad V_1(z) := \frac{1}{2}\frac{R_0(\sqrt{z}) - R_0(-\sqrt{z})}{\sqrt{z}},$$
$$V_{l+1}(z) := \frac{1}{2(l-1)!}\frac{R_l(\sqrt{z}) - R_l(-\sqrt{z})}{\sqrt{z}}, \quad l=1,2,\ldots. \tag{2.7b}$$

The rational function $R_0(z)$ is an approximation to $\exp(z)$. The rational functions $R_l(z), l=1,2,\ldots$ are recursively given by

$$R_1(z) := \frac{R_0(z)-1}{z}, \quad R_{l+1}(z) := \frac{lR_l(z)-1}{z}, \quad l=1,2,\ldots \tag{2.7c}$$

and the functions A_{jl} and B_l are defined by

$$A_{jl}(z) := \sum_{s=0}^{\rho_j} \alpha_{sl}^{(j)} V_{l+2}(z), \quad B_l := \sum_{s=0}^{\rho} \gamma_{sl} V_{l+1}(z), \tag{2.7d}$$

where $\alpha_{sl}^{(j)}, \gamma_{sl}$ are parameters which determine the method, T is a constant matrix on $[t_n, t_n+\tau]$, usually an approximation to the Jacobian matrix of the system. The values g_l are given by

$$g_l := f(t_n + \mu_l\tau, y_{n+1}^{(l)}) - Ty_{n+1}^{(l)}.$$

For $T=0$ we obtain a classical explicit RKN method. A detailed description of ARKN methods (with $\mu_0=0$) can be found in [11].

Theorem 2.2: *Let the rational functions $C_m(z)$, $C_m^*(z)$ be defined by*

$$C_m(z) := \sum_{l=0}^{m-1} A_{ml} e^{i\mu_l\nu_0},$$

$$C_m^*(z) := \sum_{l=0}^{m-1} B_l e^{i\mu_l\nu_0},$$

and let

$$\Phi_2(z,\zeta) := \zeta^2 - 2V_0(z)\zeta + V_0^2(z) - zV_1^2(z).$$

Then the numerical solution, when integrating the test equation (2.1), *satisfies the recursion*

$$\Phi_2(z_0, E) y_n = c\tau^2 [C_m(z_0) e^{i\nu_0} + V_1(z_0) C_m^*(z_0) - V_0(z_0) C_m(z_0)] e^{in\nu_0}. \tag{2.8}$$

Proof: Application of the adaptive RKN method to (2.1) yields

$$y_{n+1} = V_0\, y_n + \tau V_1\, \dot{y}_n + \tau^2\, c\, C_m\, e^{in\nu_0},$$
$$\tau\, \dot{y}_{n+1} = z V_1\, y_n + \tau V_0\, \dot{y}_n + \tau^2\, c\, C_m^*\, e^{in\nu_0}.$$

Elimination of $\dot{y}_{n+1}, \dot{y}_n, \ldots$ from these recursions leads to the recursion (2.8). □

2.3 Predictor-Corrector Methods

The predictor-corrector method, as defined in [7, 9], is an iteration scheme for approximating the solution of the implicit linear k^*-step method

$$\rho^*(E)\, y_n = \tau^2\, \sigma^*(E) f_n,\ f_n := f(t_n, y_n). \tag{2.9}$$

The initial approximation to y_{n+k^*} is denoted by $y^{(0)}_{n+k^*}$ and is assumed to be provided by an explicit linear $\tilde{k}$-step method with characteristic polynomials $\{\tilde{\rho}, \tilde{\sigma}\}$. It will be assumed that the coefficients a_0^* and b_0^* of ζ^{k^*} in ρ^* and σ^* are respectively 1 and $\neq 0$. If the corrector equation (2.9) is written in the form

$$y_{n+1} - b_0^*\, \tau^2 f_{n+1} = \Sigma_n, \tag{2.9′}$$

where Σ_n represents the back values in (2.9), then the successive iterates in the iteration scheme are defined by

$$y^{(j)}_{n+1} = \sum_{l=1}^{j} [\mu_{jl}\, y^{(l-1)}_{n+1} + \bar{\mu}_{jl}\, \tau^2 f^{(l-1)}_{n+1}] + \lambda_j\, \Sigma_n, \quad j = 1, \ldots, m, \tag{2.10a}$$

where the parameters μ_{jl}, $\bar{\mu}_{jl}$ and λ_j are assumed to satisfy the compatibility conditions

$$\sum_{l=1}^{j} \mu_{jl} = 1 - \lambda_j,\ \sum_{l=1}^{j} \bar{\mu}_{jl} = b_0^*\, \lambda_j, \quad j = 1, \ldots, m. \tag{2.10b}$$

The iterate $y^{(m)}_{n+1}$ will be adopted as the final approximation to the solution of (2.9) and is therefore denoted by y_{n+1}.

In the analysis of the method (2.10) the iteration polynomial $P_m(z)$, recursively defined by

$$P_0(z) = 1,\ P_j(z) = \sum_{l=1}^{j} [\mu_{jl} + \bar{\mu}_{jl}\, z]\, P_{l-1}(z), \quad j = 1, \ldots, m, \tag{2.11}$$

plays a central role. It governs both the accuracy and the stability of the method. In the present paper, the parameters μ_{jl} and $\bar{\mu}_{jl}$, and therefore the coefficients of $P_m(z)$, are allowed to depend on τ. Notice that (2.10b) implies that $P_m(1/b_0^*) = 1$.

Theorem 2.3: *Let the predictor $\{\tilde{\rho}, \tilde{\sigma}\}$ and the corrector $\{\rho^*, \sigma^*\}$ be of order $\tilde{p}$ and p^*, respectively, and let $P_m(\tau^2) = 0(\tau^s)$ as $\tau \to 0$. Then the predictor-corrector method (2.10) is of order $p = \min\{p^*, \tilde{p} + s, 4 + 2\tilde{p}\}$.*

Proof: Cf. [7]. □

We see that, by specifying the LM methods $\{\tilde{\rho},\tilde{\sigma}\}$ and $\{\rho^*,\sigma^*\}$, and the iteration polynomial, the order of the method is completely determined. In the choice of the LM methods and $P_m(z)$ we shall be led by our wish to minimize the (global) error that arises when the test equation (2.1) is integrated.

Theorem 2.4: *Let $\{\rho^*,\sigma^*\}$ and $\{\tilde{\rho},\tilde{\sigma}\}$ be normalized in the sense that $a_0^*=\tilde{a}_0=1$, and let*

$$R(z,\zeta):=\rho^*(\zeta)-\frac{(1-b_0^* z)P_m(z)}{P_m(z)-1}\tilde{\rho}(\zeta)\zeta^{k^*-\tilde{k}},$$

$$S(z,\zeta):=\sigma^*(\zeta)-\frac{(1-b_0^* z)P_m(z)}{P_m(z)-1}\tilde{\sigma}(\zeta)\zeta^{k^*-\tilde{k}}, \tag{2.12}$$

$$\Phi_k(z,\zeta):=R(z,\zeta)-zS(z,\zeta).$$

Then the solution, when integrating the test equation (2.1), *satisfies the recursion*

$$\Phi_k(z_0,E)\,y_n=c\tau^2 S(z_0,e^{i\nu_0})\,e^{in\nu_0}. \tag{2.13}$$

Proof: Application of the predictor-corrector method (2.10) to the test equation (2.1) yields

$$y_{n+1}^{(j)}=P_j(z_0)\,y_{n+1}^{(0)}+Q_j(z_0)\,\Sigma_n+Q_j^*(z_0)\,c\,e^{i(n+1)\nu_0}, \tag{2.14}$$

where P_j is defined in (2.11), and Q_j and Q_j^* respectively satisfy the recursion

$$Q_0(z)=0,\ Q_j(z)=\lambda_j+\sum_{l=1}^{j}[\mu_{jl}+\bar{\mu}_{jl}z]\,Q_{l-1}(z),$$

$$Q_0^*(z)=0,\ Q_j^*(z)=b_0^*\lambda_j\tau^2+\sum_{l=1}^{j}[\mu_{jl}+\bar{\mu}_{jl}z]\,Q_{l-1}^*(z).$$

By virtue of (2.10b) it can be verified that Q_j and Q_j^* are related to P_j according to

$$Q_j^*(z)=b_0^*\tau^2 Q_j(z)=b_0^*\tau^2\frac{1-P_j(z)}{1-b_0^*}.$$

On substituting into (2.14) and observing that

$$y_{n+1}^{(0)}=[E^{\tilde{k}}-\tilde{\rho}(E)]\,y_{n+1-\tilde{k}}+\tau^2\tilde{\sigma}(E)f_{n+1-\tilde{k}},$$

$$\Sigma_n=[E^{k^*}-\rho^*(E)]\,y_{n+1-k^*}-\tau^2[b_0^*E^{k^*}-\sigma^*(E)]f_{n+1-k^*},\ f_n=-\delta^2 y_n+c\,e^{in\nu_0}$$

we arrive at the recursion (2.13). □

2.4 The Numerical Solution of the Test Equation

We shall derive an explicit expression for the numerical solution of the test equation determined by the recursion (2.5), (2.8) and (2.13). These recursions are of the form

$$\Phi(z_0,E)\,y_n=c\tau^2 F(z_0,v_0)\,e^{in\nu_0}, \tag{2.15}$$

where $\Phi(z,\zeta)$ is a polynomial in ζ with coefficients depending on z but not on n, and $F(z,v)$ is a given function of z and v, again not depending on n. By virtue of this special form, the following theorem holds:

Theorem 2.5: *The general solution of* (2.15) *assumes the form*

$$y_n = c\tau^2 \frac{F(z_0, v_0)}{\Phi(z_0, e^{iv_0})} e^{inv_0} + \sum_{j=1}^{r} \zeta_j^n \tilde{c}_{m_j-1}(n), \tag{2.16}$$

where ζ_j, $j=1,\ldots,r$, *are zeros of* $\Phi(z_0,\zeta)$ *of multiplicity* m_j *and* $\{\tilde{c}_{m_j-1}(n)\}$ *are polynomials in* n *of degree* m_j-1.

Proof: By writing

$$y_n = u_n e^{inv_0} \tag{2.17}$$

and on substitution into (2.15), we obtain the recursion

$$\Phi(z_0, e^{iv_0} E) u_n = c\tau^2 F(z_0, v_0).$$

The general solution of this recursion is given by (cf. e.g. Lambert [10, p. 8])

$$\begin{aligned} u_n = c\tau^2 \frac{F(z_0, v_0)}{\Phi(z_0, e^{iv_0})} + \sum_{j=1}^{r} (e^{-iv_0}\zeta_j)^n [c_{j1} + c_{j2}\, n + c_{j3}\, n(n-1) + \ldots \\ + c_{jm_j} n(n-1)\ldots(n-m_j+2)], \end{aligned} \tag{2.18}$$

where m_j is the multiplicity of the characteristic root ζ_j of Φ and where the constants c_{jl} are arbitrary. From (2.17) and (2.18) the assertion of the theorem readily follows. □

The numerical solutions provided by the RKN type and predictor-corrector methods can now be obtained explicitly by substituting the corresponding quantities Φ, F and ζ_j into (2.16); these quantities are given in the Theorems 2.1, 2.2 and 2.4, respectively.

3. Reduction of Phase Errors and Dissipation Errors

3.1 Possible Strategies

Having derived the numerical solution to our test equation (2.1) we are in a position to compare its error with respect to the exact solution of (2.1); the exact solution can be represented by

$$y(t_n) = \frac{-c\tau^2}{z_0 + v_0^2} e^{inv_0} + c_+ e^{in\sqrt{-z_0}} + c_- e^{-in\sqrt{-z_0}}, \tag{3.1}$$

where c_+ and c_- are constants determined by the initial conditions. In the expressions (2.16), for the numerical solution, and (3.1), for the exact solution, the first term is called the *inhomogeneous solution component* and the subsequent terms are called the *homogeneous solution components*. The inhomogeneous solution component is one source of possible phase errors caused by a different argument of the expressions in front of $\exp(inv_0)$. The homogeneous solution components give

rise to two sources of phase errors: firstly, the principal characteristic roots ζ_1 and ζ_2 in (2.16) may differ in phase with $\exp(\pm i\sqrt{-z_0})$, and secondly, the corresponding coefficients $\tilde{c}_{m_1-1}$ and $\tilde{c}_{m_2-1}$ may differ in phase with the coefficients $c_\pm$. The phase errors caused by the phase lag of the characteristic roots, increase linearly in time and was termed *propagated dispersion* in [8]. The two other forms of phase errors do not depend on t and may be considered as *initial dispersion* (or *initial phase lag*) of *inhomogeneous* and *homogeneous* type, respectively.

It should be observed that a vanishing inhomogeneous component ($c=0$) in the exact solution implies a vanishing inhomogeneous component in the numerical solution, and vice versa. This is not true for the homogeneous components. Thus, when integrating an equation whose exact solution does not contain homogeneous components ($c_+ = c_- = 0$), the numerical solution will generally have homogeneous components. In such cases, the effect of the homogeneous components on the total phase error is not clear, because we cannot compare the arguments of corresponding components.

A second observation concerns the weight factor in front of the forced oscillation $\exp(in\tau\omega)$. Let c be fixed in (2.1) and let ω increase. Then, it follows from (3.1) that the inhomogeneous component in the exact solution is decreasing in magnitude. Suppose that we have no homogeneous components in the exact solution; then it may happen that the inhomogeneous component in the numerical solution is dominating for small values of the forced frequency ω, but is becoming insignificant (with respect to the numerical homogeneous components) if ω increases. Thus, when the method is deviced in order to represent the forced oscillation accurately, then it will only be effective if the inhomogeneous component is dominating in the numerical solution. Such methods lose their effectiveness if ω increases. This phenomenon was observed experimentally by Thomas [12] and by Strehmel/Weiner [11].

In this paper we shall concentrate on the reduction of the *inhomogeneous phase error*. Following Brusa/Nigro [1], we estimate the magnitude of phase errors relative to the phase of the corresponding exact solution component. For the inhomogeneous phase error this leads to the following definition:

Definition 3.1: The *inhomogeneous phase error* introduced by the numerical scheme (2.15) is defined by

$$P_{inh} := \left| \frac{z_0 + v_0^2}{c\tau^2} \right| \left| \arg\left[\frac{F(z_0, v_0)}{\Phi(z_0, e^{iv_0})} \right] - \arg\left[\frac{-1}{z_0 + v_0^2} \right] \right|.$$

If $P_{inh} = 0(\tau^q)$ as $\tau \to 0$, then the method is said to have *inhomogeneous phase lag of order* q. □

Similarly, we define the inhomogeneous dissipation error:

Definition 3.2: The *dissipation error* of the inhomogeneous component of the numerical solution determined by (2.15), is defined by

$$D_{inh} := \left| (z_0 + v_0^2) \frac{F(z_0, v_0)}{\Phi(z, e^{iv_0})} \right| - 1.$$

If $D_{inh}=0(\tau^r)$, then the method is said to have *dissipation of order r* with respect to the inhomogeneous solution component. □

In reducing the magnitude of the phase errors, there are several avenues open to us; for instance:

(i) If the frequencies δ and ω (and therefore z_0 and v_0) are both precisely known, then one could try to determine the free parameters such that P_{inh} vanishes for the given values of z_0 and v_0.

(ii) If the frequency δ is not precisely known, but instead, the corresponding value of z_0 is known to be small, then one could try to maximize the phase lag order q.

(iii) If the frequency δ is known to lie in an interval $[\underline{\delta}, \bar{\delta}]$, then one could try to minimize P_{inh} over the corresponding interval on the z-axis.

In a similar way, one can reduce the dissipation error D_{inh}, or if desired, one can reduce P_{inh} and D_{inh} simultaneously. In fact, this approach is to be preferred. Firstly, because the reduction of the phase error alone leads to complicated formulas defining the parameters of the methods, resulting in difficult computer implementations, and secondly, because the dissipation error may decrease the accuracy of the method to such an extent that the advantage of a small phase lag is completely lost.

In the subsequent Sections 3.2 – 3.4 we discuss the explicit RKN methods, the ARKN methods and the predictor-corrector methods.

3.2 Runge-Kutta-Nyström Methods

For the RKN method (2.2) we obtain

$$\frac{F(z,v)}{\Phi(z,e^{iv})}=\frac{B_m C_m^*-B_m^* C_m+C_m e^{iv}}{\Phi_2(z,e^{iv})}, \tag{3.2}$$

where Φ_2 is defined in (2.4). In order to simplify the method and, at the same time, to guarantee that the method has a nonempty interval of periodicity, we will require that

$$A_m B_m^*-A_m^* B_m\equiv 1 \tag{3.3}$$

for all values of z. Inserting (3.3) into (3.2) yields

$$\frac{F(z,v)}{\Phi(z,e^{iv})}=\frac{e^{-iv}[B_m C_m^*-B_m^* C_m]+C_m}{2\cos(v)-(A_m+B_m^*)}. \tag{3.2'}$$

Theorem 3.1: *Let*

$$\varepsilon(z,v):=\frac{e^{-iv}[B_m C_m^*-B_m^* C_m]+C_m}{2\cos(v)-(A_m+B_m^*)}+\frac{1}{z+v^2}. \tag{3.4}$$

If $\varepsilon=0(\tau^\gamma)$ *then the phase lag order and dissipation order are both greater than or equal to* $\gamma+2$.

Proof: From (3.4) and the definition of P_{inh} and D_{inh} it follows that

$$P_{inh}=\left|\frac{(z_0+v_0^2)}{c\tau^2}\right|\cdot\left|\arctan\frac{(z_0+v_0^2)\,\mathrm{Im}\,(\varepsilon(z_0,v_0))}{1-(z_0+v_0^2)\,\mathrm{Re}\,(\varepsilon(z_0,v_0))}\right|$$

$$=\left|\frac{(z_0+v_0^2)^2}{c\tau^2}\right|\cdot\left|\frac{\mathrm{Im}\,(\varepsilon(z_0,v_0))}{1-(z_0+v_0^2)\,\mathrm{Re}\,(\varepsilon(z_0,v_0))}\right|+0(\tau^{4+3\gamma}),$$

$$D_{inh}=|1-\varepsilon(z_0,v_0)(z_0+v_0^2)|-1.$$

The assertion of the theorem is now immediate. □

We shall restrict our discussion to the RKN methods considered in [8]. These methods are generated by the array

$\mu_1=\frac{1}{2}$	0			
$\mu_2=\frac{1}{2}$	0	λ_{21}		
.	.	.	.	
.	.	.	.	
.	.	.	.	
$\mu_{m-1}=\frac{1}{2}$	0	...	0	$\lambda_{m-1,m-2}$
$\mu_m=1$	0	...	0	$\frac{1}{2}=\lambda_{m,m-1}$
	0	...	0	$1=\lambda^*_{m-1}$

and are second-order accurate for all values of $\lambda_{j,j-1}$, $j=2,\ldots,m-1$. In addition, they satisfy condition (3.3) so that, for z_0 lying in the periodicity interval, the *homogeneous* solution component is presented by the method *without dissipation error*.

When we calculate ε, as defined by (3.4), then it turns out that ε is the real-valued function

$$\varepsilon(z,v)=\frac{\cos(\frac{1}{2}v)\,(S_m(z)-2)/z}{2\cos(v)-S_m(z)}+\frac{1}{z+v^2},\tag{3.4'}$$

where $S_m(z)$ is the polynomial

$$S_m(z)=A_m(z)+B_m^*(z)=2+z+\sigma_2 z^2+\ldots+\sigma_{m-1}z^{m-1},$$

$$\lambda_{j,j-1}=\frac{\sigma_{m-j+1}}{\sigma_{m-j}},\quad j=2,\ldots,m-1,\ \sigma_1=1.$$

Since ε is real, it follows that $P_{inh}=0$ for all z_0 and v_0. The (inhomogeneous) dissipation error is given by

$$D_{inh}=-1-\frac{z_0+v_0^2}{z_0}\,\frac{\cos\left(\frac{1}{2}v_0\right)[S_m(z_0)-2]}{2\cos(v_0)-S_m(z_0)}=\frac{1}{24}(\tau\omega)^2\,\frac{1+3(8\sigma_2-1)\left(\frac{\delta}{\omega}\right)^2}{1-\left(\frac{\delta}{\omega}\right)^2}+0(\tau^4)$$

provided that z_0 and v_0 are sufficiently small. This expression reveals that the free parameters in the polynomial $S_m(z)$ can only be exploited for the reduction of D_{inh} if we know both δ and ω. For instance, if we choose

$$\sigma_2 = \frac{1}{8}\left(1 - \frac{\omega^2}{3\delta^2}\right),$$

then $D_{inh} = 0(\tau^4)$ (of course, we can do better by setting $D_{inh} = 0$ for the given values of $z_0 = -\tau^2\delta^2$ and $v_0 = \omega\tau$, see Example 3.1). If δ is not known, or if several frequencies δ are involved, then it is not clear how we can choose $S_m(z)$ such that the dissipation of the inhomogeneous solution component is reduced so that the dissipation order exceeds $r=2$. In such cases, we propose to use the polynomial $S_m(z)$ to reduce the phase error of the *homogeneous* solution component. This has been investigated in [8] where it was pointed out that choosing

$$\sigma_j = \frac{2}{(2j)!},\ j = 1, \ldots, m-1 \tag{3.5}$$

leads to the maximal attainable phase lag order $2m-2$ for the homogeneous solution.

Example 3.1: Consider the method

$$\begin{array}{c|ccc} \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \sigma_2 & \\ \hline & 0 & 0 & \frac{1}{2} \\ & 0 & 0 & 1 \end{array} \qquad p=2,\ q=\infty,\ r\geq 2 \tag{3.6a}$$

with (inhomogeneous) dissipation error

$$D_{inh} = -1 - (z_0 + v_0^2)\frac{\cos(\frac{1}{2}v_0)(1+\sigma_2 z_0)}{2(\cos(v_0)-1) - z_0 - \sigma_2 z_0^2}.$$

Setting $D_{inh} = 0$ yields

$$\sigma_2 = \frac{1}{z_0}\frac{\left(1-\cos\left(\frac{v_0}{2}\right)\right)z_0 - \cos\left(\frac{v_0}{2}\right)v_0^2 - 2(\cos(v_0)-1)}{\cos\left(\frac{v_0}{2}\right)v_0^2 - \left(1-\cos\left(\frac{v_0}{2}\right)\right)z_0} \approx \frac{1}{8}\left(1 - \frac{\omega^2}{3\delta^2}\right). \tag{3.6b}$$

The resulting method has zero phase lag and zero dissipation. □

3.3 Adaptive Runge-Kutta Methods

For the ARKN methods (2.7) we obtain

$$\frac{F(z,v)}{\Phi(z,e^{iv})} = \frac{V_1 C_m^* - V_0 C_m + C_m e^{iv}}{\Phi_2(z,e^{iv})},$$

where $\Phi_2(z, e^{iv})$ is defined in (2.8). In order to simplify these methods we require that the rational function $R_0(z)$ satisfies the condition

$$|R_0(ix)|=1 \quad \text{for all } x\in\mathbb{R}. \tag{3.7}$$

This condition guarantees that the ARKN-method possesses an infinite interval of periodicity, i.e. the method is P-stable, and that the dissipation error of the homogeneous solution components is zero. The propagated phase lag order is equal to the approximation order of $R_0(z)$, so that a reduction of the propagated phase error is always possible without difficulty. From the definition of the rational functions $V_0(z)$ and $V_1(z)$ and with condition (3.7) we obtain the relation

$$V_0^2(z)-z\,V_1^2(z)=1 \quad \text{for all } z\in\mathbb{C}.$$

Thus we get

$$\frac{F(z,v)}{\Phi(z,e^{iv})}=\frac{e^{-iv}[V_1\,C_m^*-V_0\,C_m]+C_m}{2(\cos(v_0)-V_0)}.$$

In analogy to Theorem 3.1 we obtain

Theorem 3.2: *Let*

$$\varepsilon(z_0,v_0):=\frac{e^{-iv_0}[V_1\,C_m^*-V_0\,C_m]+C_m}{2(\cos(v_0)-V_0)}+\frac{1}{z_0+v_0^2}. \tag{3.8}$$

If $\varepsilon=0(\tau^\gamma)$ then the phase lag order and the dissipation order are greater than or equal to $\gamma+2$. □

Example 3.2: We consider the one-stage ARKN method

$$\begin{aligned}
y_{n+1}^{(0)}&=y_n+\tfrac{1}{2}\tau\,\dot y_n,\\
y_{n+1}&=V_0(\tau^2T)\,y_n+\tau\,V_1(\tau^2T)\,\dot y_n+\tau^2\,V_2(\tau^2T)\,g(t_n+\tfrac{1}{2}\tau,y_{n+1}^{(0)}),\\
\dot y_{n+1}&=V_0(\tau^2T)\,\dot y_n+\tau[TV_1(\tau^2T)\,y_n+V_1(\tau^2T)\,g(t_n+\tfrac{1}{2}\tau,y_{n+1}^{(0)})].
\end{aligned} \tag{3.9}$$

This method possesses the (algebraic) order 2 if the approximation order of $R_0(z)$ is greater than or equal to 2. For ε we obtain

$$\varepsilon(z_0,v_0)=\frac{[V_1^2(z_0)-V_0(z_0)\,V_2(z_0)]\,e^{-iv_0/2}+V_2(z_0)\,e^{iv_0/2}}{2(\cos(v_0)-V_0(z_0))}+\frac{1}{z_0+v_0^2}.$$

From (2.7b) it follows

$$V_1^2(z)=V_2(z)(1+V_0(z)) \quad \text{for all } z\in\mathbb{C}. \tag{3.10}$$

Hence

$$\varepsilon(z_0,v_0)=\frac{V_2(z_0)}{\cos(v_0)-V_0(z_0)}\cos(\tfrac{1}{2}v_0)+\frac{1}{z_0+v_0^2}.$$

Since ε is real we obtain $P_{inh}=0$ for all z_0 and v_0. The dissipation error is given by

$$D_{inh}=-1-(z_0+v_0^2)\frac{V_2(z_0)}{\cos(v_0)-V_0(z_0)}\cos(\tfrac{1}{2}v_0).$$

Let us assume that

$$R_0(z)-\exp(z)=0(z^k) \quad \text{as } z\to 0 \text{ with } k\geq 2,$$

then we obtain for $V_0(z)$ and $V_2(z)$ the asymptotic expansions

$$V_0(z_0)=1+\tfrac{1}{2}z_0+d_0 z_0^2+0(z_0^4),$$
$$V_2(z_0)=\tfrac{1}{2}+d_2 z_0+0(z_0^2)$$

with $d_0, d_2 \in \mathbb{R}$; therefore we get

$$D_{inh}=-\tau^2\frac{2(d_0-d_2)\delta^4+\frac{1}{24}\omega^4-(\frac{1}{8}-2d_2)\omega^2\delta^2}{\omega^2-\delta^2}+0(\tau^4).$$

Finally, we show that an m-stage ARKN method with a phase lag order $q=\infty$ possesses the maximal (algebraic) order 2. The following holds:

Theorem 3.3: *Let $c_k, k=0,1,\ldots,M-1$, be the distinct nodes $\mu_l, l=0,1,\ldots,m-1$ with $m \geq M$, of an m-stage ARKN method. Then the ARKN method*

$$\begin{array}{c|cccc} \mu_1 & a_{10}V_2 & & & \\ \mu_2 & a_{20}V_2 & a_{21}V_2 & & \\ \vdots & \vdots & & & \\ \mu_{m-1} & a_{m-1,0}V_2 & \cdots & a_{m-1,m-2}V_2 & \\ \hline & a_0 V_2 & \cdots & a_{m-1}V_2 & \\ & b_0 V_1 & \cdots & b_{m-1}V_1 & \end{array} \quad \text{with } \mu_0 \in [0,1] \tag{3.11}$$

possesses the phase lag order $q=\infty$ if and only if

$$\sum_{\mu_j=c_k} b_j = \sum_{\mu_j=1-c_k} a_j; \quad \sum_{\mu_j=c_k} b_j = \sum_{\mu_j=c_k} a_j \tag{3.12}$$

for all $k=0,1,\ldots,M-1$.

Proof: From (3.8) it follows that

$$\varepsilon(z_0,v_0)=\frac{\sum_{l=0}^{m-1}[V_1^2 b_l - V_0 V_2 a_l]e^{iv_0(\mu_l-1)}+\sum_{l=0}^{m-1}V_2 a_l e^{iv_0\mu_l}}{2(\cos(v_0)-V_0)}+\frac{1}{z_0+v_0^2}.$$

With (3.10) we obtain

$$\varepsilon(z_0,v_0)=\frac{V_2}{2(\cos(v_0)-V_0)}\sum_{l=0}^{m-1}[b_l e^{iv_0(\mu_l-1)}+a_l e^{iv_0\mu_l}+V_0(b_l-a_l)e^{iv_0(\mu_l-1)}]+\frac{1}{z_0+v_0^2}.$$

$\varepsilon(z_0,v_0)$ is real if and only if the conditions (3.12) are fulfilled which is what we wanted to show. □

Theorem 3.4: *An m-stage ARKN method* (3.11) *with $q=\infty$ possesses the maximal (algebraic) order 2.*

Proof: By application of Theorem 3.3, it follows from (3.12) that

$$\sum_{l=0}^{m-1} a_l \mu_l = \sum_{k=1}^{M-1} c_k \sum_{\mu_l=c_k} a_l = \sum_{k=1}^{M-1}\sum_{\mu_l=c_k} b_l = \sum_{l=0}^{m-1} b_l \mu_l.$$

The condition

$$\sum_{l=0}^{m-1} b_l \mu_l = \tfrac{1}{2}$$

for algebraic order two yields

$$\sum_{l=0}^{m-1} a_l \mu_l = \tfrac{1}{2}. \tag{3.13}$$

For an ARKN method of order $p \geq 3$ we have the consistency condition

$$\sum_{l=0}^{m-1} V_2(0)\, a_l \mu_l = \tfrac{1}{6}.$$

Because of $V_2(0) = \frac{1}{2}$, it follows that

$$\sum_{l=0}^{m-1} a_l \mu_l = \tfrac{1}{3}.$$

Thus we have a contradiction to (3.13) and the theorem is proved. □

3.4 Predictor-Corrector Methods

For predictor-corrector methods of type (2.10) we find

$$\frac{F(z,v)}{\Phi(z,e^{iv})} = \frac{-1}{z - \dfrac{R}{S}(z,e^{iv})}, \tag{3.14}$$

where R and S are defined in (2.12). The analogue of the Theorems 3.1 and 3.2 reads:

Theorem 3.5: *Let*

$$\phi^*(v) := \rho^*(e^{iv}) + v^2 \sigma^*(e^{iv}),\ \tilde{\phi}(v) := \tilde{\rho}(e^{iv}) + v^2 \tilde{\sigma}(e^{iv})$$

and define

$$\varepsilon(z,v) := P_m(z) - \frac{\phi^*(v)}{\phi^*(v) - (1 - b_0^* z)\,\tilde{\phi}(v)\, e^{i(k^* - \tilde{k})v}}. \tag{3.15}$$

If $\varepsilon = 0(\tau^\gamma)$, $P_m(0) \neq 1$ *and* $S(0,1) \neq 0$, *then the phase lag order and the dissipation order are both greater than or equal to* $\gamma + \min\{p^*, \tilde{p}\}$, *where* p^* *and* $\tilde{p}$ *are the orders of the corrector and the predictor, respectively.*

Proof: We first express the function R/S in terms of the functions $P_m(z)$, $\varepsilon(z,v)$, $\phi^*(v)$ and $\tilde{\phi}(v)$. From (2.12) and (3.15) we derive the relation

$$\frac{R}{S}(z,e^{iv}) = -v^2 + \varepsilon(z,v)\frac{\phi^*(v) - (1 - b_0^* z)\,\tilde{\phi}(v)\, e^{i(k^* - \tilde{k})v}}{(P_m(z) - 1)\, S(z,e^{iv})}. \tag{3.16}$$

Since $\phi^* = 0(\tau^{p^*+2})$ and $\tilde{\phi}(v) = 0(\tau^{\tilde{p}+2})$ it follows from the assumptions of the theorem that

$$\frac{R}{S}(z,e^{iv}) = -v^2 + 0(\tau^\gamma[\tau^{p^*+2} + \tau^{\tilde{p}+2}]). \tag{3.16'}$$

The phase lag P_{inh} is now given by (cf. Definition 3.1):

$$\begin{aligned} P_{inh} &= \left|\frac{z_0+v_0^2}{c\tau^2}\right| \left|\arg\left(\frac{-1}{z_0-R/S}\right)-\arg\left(\frac{-1}{z_0+v_0^2}\right)\right| \\ &= \left|\frac{z_0+v_0^2}{c\tau^2}\right| |\arg(z_0+v_0^2)-\arg(z_0-R/S)| \\ &= \left|\frac{z_0+v_0^2}{c\tau^2}\right| \left|\arctan\frac{\operatorname{Im}(R/S)}{z_0-\operatorname{Re}(R/S)}\right|, \end{aligned} \tag{3.17}$$

where R and S are evaluated at (z_0, e^{iv_0}). Similarly, we find

$$D_{inh} = \left|\frac{z_0+v_0^2}{z_0-R/S}\right| - 1. \tag{3.18}$$

Substitution of (3.16′) leads to

$$P_{inh} = \left|\frac{-\delta^2+\omega^2}{c}\right| \arctan\left(\frac{0(\tau^{p^*+\gamma+2}+\tau^{\bar{p}+\gamma+2})}{0(\tau^2)}\right) = 0(\tau^{p^*+\gamma}+\tau^{\bar{p}+\gamma})$$

and

$$D_{inh} = \left|\frac{z_0+v_0^2}{z_0+v_0^2+0(\tau^{p^*+\gamma+2}+\tau^{\bar{p}+\gamma+2})}\right| - 1 = 0(\tau^{p^*+\gamma}+\tau^{\bar{p}+\gamma}),$$

proving the theorem. □

Before discussing the maximization of the order γ of ε as $\tau\to 0$, we consider the case where (z, v) assumes a fixed value (z_0, v_0). Then, by choosing $P_m(z)$ such that (3.15) is satisfied for $(z, v, \varepsilon)=(z_0, v_0, 0)$, it follows from (3.16) that

$$\frac{R}{S}(z_0, e^{iv_0}) = -v_0^2,$$

and from (3.17) and (3.18) we obtain $P_{inh}=D_{inh}=0$. Thus, for given (z_0, v_0) there is no phase lag and no dissipation.

Example 3.3: Consider the Störmer predictor

$$(\tilde{\rho}, \tilde{\sigma}) = ((\zeta-1)^2, \zeta) \tag{3.19a}$$

and the Numerov corrector

$$(\rho^*, \sigma^*) = ((\zeta-1)^2, \tfrac{1}{12}(\zeta^2+10\zeta+1)). \tag{3.19b}$$

Then

$$\tilde{\phi} = 2e^{iv}[\cos(v)-1+\tfrac{1}{2}v^2], \quad \phi^* = 2e^{iv}[(1+\tfrac{1}{12}v^2)\cos(v)-1+\tfrac{5}{12}v^2].$$

Substitution into (3.15) with $(z, v, \varepsilon)=(z_0, v_0, 0)$ leads to the condition $P_m(z_0)=c_0$, where

$$c_0 := \frac{(12+v_0^2)\cos(v_0)-12+5v_0^2}{(v_0^2+z_0)\cos(v_0)-v_0^2-z_0+\frac{1}{2}v_0^2 z_0}. \tag{3.19c}$$

For instance, for $m=1$ this condition reads

$$P_1(z_0) = \beta_0+\beta_1 z_0 = c_0.$$

Since we should also satisfy the compatibility condition (2.10b), we have to require $P_m(1/b_0^*)=1$, i.e.

$$P_1(12)=\beta_0+12\beta_1=1.$$

Thus, the iteration polynomial assumes the form

$$P_1(z)=\frac{12c_0-z_0+(1-c_0)z}{12-z_0}. \tag{3.19d}$$

The one-stage predictor-corrector method generated by (3.19a), (3.19b), (3.19c) and (3.19d) has zero phase lag and zero dissipation as far as the inhomogeneous solution component is concerned. Its (algebraic) order p can be derived from Theorem 2.3. Since $p^*=4$, $\tilde{p}=2$ and, because $c_0\approx-\frac{1}{20}v_0^2$ as $\tau\to 0$, $s=2$, it follows that $p=4$. □

Finally, we consider the maximization of γ, that is, the maximization of the phase lag and dissipation order.

Theorem 3.6: *The iteration polynomial*

$$P_m(z)=\sum_{j=0}^{m}\beta_j z^j;\quad \beta_0:=\frac{\phi^*(v_0)}{\phi^*(v_0)-\bar{\phi}(v_0)e^{i(k^*-\bar{k})v_0}};$$
$$\beta_j:=-b_0^*(\beta_0-1)\beta_{j-1},\ j=1,\ldots,m-1;\quad \beta_m:=(b_0^*)^m\left(1-\sum_{j=0}^{m-1}\beta_j(b_0^*)^{-j}\right) \tag{3.20}$$

satisfies the compatibility condition (2.10b) *and yields the maximal attainable phase lag and dissipation order* $q=r=2m+\min(p^*,\tilde{p})$, *while the maximal algebraic order* $p=\min\{p^*,4+2\tilde{p}\}$.

Proof: First of all we impose the condition that $P_m(z)$ satisfies the compatibility condition $P_m(1/b_0^*)=1$. This is achieved if β_m is defined as in (3.20). The coefficients $\beta_0,\ldots,\beta_{m-1}$ are now free for maximizing the order γ of ε as $\tau\to 0$. It follows from (3.15) that γ is maximized if $P_m(z)$ is a Taylor approximation of the function

$$T(z):=\frac{\phi^*(v_0)}{\phi^*(v_0)-(1-b_0^*z)\bar{\phi}(v_0)e^{i(k^*-\bar{k})v_0}}$$

of highest possible order. Thus,

$$\beta_j=\frac{1}{j!}\frac{d^jP_m}{dz^j}(0)=\frac{1}{j!}\frac{d^jT}{dz^j}(0),\quad j=0,\ldots,m-1.$$

An elementary calculation leads to the coefficients given in (3.20). The order of ε is evidently $\gamma=2m$.

Since $P_m(z)$ satisfies the condition

$$P_m(\tau^2)=0(\beta_0)=0(\tau^s),\ s=\max(p^*-\tilde{p},0)$$

it follows from Theorem 2.3 that the maximal possible algebraic order is given as in the theorem. □

Example 3.4: We again consider a method based on the Störmer predictor (3.19a) and the Numerov corrector (3.19b). Choosing $m=2$, the iteration polynomial defined by Theorem 3.6 assumes the form

$$P_2(z)=\beta_0-\tfrac{1}{12}\beta_0(\beta_0-1)z+\tfrac{1}{144}(\beta_0-1)^2 z^2,$$

where

$$\beta_0=\frac{(12+v_0^2)\cos(v_0)-12+5v_0^2}{v_0^2(\cos(v_0)-1)}. \tag{3.21}$$

The resulting 2-stage method has algebraic order $p=4$, phase lag order $q=6$ and dissipation order $r=6$. Notice that the value of z_0 does not enter in the iteration polynomial. □

4. Numerical Examples

4.1 Testing Strategy

In this section we will test the RKN type method and the predictor-corrector (PC) methods as described in the preceding sections.

Because the major aim of this paper is an accurate treatment of the inhomogeneous solution component, we will apply the methods to test examples in which the forced oscillation strongly dominates the homogeneous solution components.

In the examples we will concentrate on the *phase errors* in the numerical solution. To measure the *total* phase lag at the endpoint $t=T$, we define

$$cd(T):=-^{10}\log(\|(y_N-y(T))/y'(T)\|_\infty),\ N=T/\tau, \tag{4.1}$$

where N denotes the number of steps performed and T is a zero of the exact solution. If the numerical solution y_N is small at $t_N=T$, then, by taking the slope $y'(T)$ into account, this cd-value is an adequate measure for the phase lag.

Because the number of f-evaluations per step is not the same for all methods, we adjusted the step sizes in such a way as to obtain an equal amount of computational effort (in terms of f-evaluations) over the whole range of integration. This strategy is only valid for comparing the efficiency of the explicit schemes. The computational effort of the ARKN method is determined not only by function evaluations, but also by the solution of linear systems of equations (including the evaluation of the Jacobian). If a constant stepsize is used, a substantial reduction of this effort is possible, in case the Jacobian is constant or if the matrix $T\approx f_y$ is kept constant for some steps (the algebraic order is independent of T). A new LU-decomposition is required only after a change of T.

4.2 Specification of the Methods

Now, we will discuss the methods which will be actually tested. First, we briefly mention the schemes and at the end of this subsection we summarize their characteristics.

RKN Methods

To start with, from the family of Runge-Kutta-Nyström methods we will test the two schemes as given by (3.6a), (3.5) and by (3.6a), (3.6b).

ARKN Methods

To keep the computational effort small we implemented the one-stage method (3.9). For $R_0(z)$ we used the Padé-approximations P_{11} (ARKN1) and P_{22} (ARKN2). The choice of P_{22} requires more work (matrix-by-matrix multiplication) but yields the propagated phase lag order 4.

PC Methods

Next, we implemented the one- and two-stage PC schemes based on the Störmer predictor and Numerov corrector. The iteration polynomials, defining these schemes, are given in the Examples 3.3 and 3.4, respectively. From these polynomials, the actual PC schemes are straightforwardly constructed (see also [7]). For both schemes we start with

$$\Sigma_n = 2y_n - y_{n-1} + \tfrac{1}{12}\tau^2(10f_n + f_{n-1}),\quad y_{n+1}^{(0)} = 2y_n - y_{n-1} + \tau^2 f_n. \tag{4.2a}$$

Now, for $m=1$, the final result y_{n+1} is obtained by

$$y_{n+1} = [(12c_0 - z_0)y_{n+1}^{(0)} + (12 - 12c_0)\Sigma_n + (1-c_0)\tau^2 f_{n+1}^{(0)}]/(12 - z_0), \tag{4.2b}$$

where c_0 is given by (3.19c). This fourth-order scheme should only be applied when exact values for z_0 and v_0 are available. In that case the inhomogeneous solution component is integrated without any error.

The two-stage scheme, which is to be used in case of small z_0-values, also starts with (4.2a) but proceeds with

$$\begin{aligned} y_{n+1}^{(1)} &= \beta_0 y_{n+1}^{(0)} + (1-\beta_0)\Sigma_n + \tfrac{1}{12}(1-\beta_0)\tau^2 f_{n+1}^{(0)}, \\ y_{n+1} &= \beta_0 y_{n+1}^{(0)} + (1-\beta_0)\Sigma_n + \tfrac{1}{12}(1-\beta_0)\tau^2 f_{n+1}^{(1)}, \end{aligned} \tag{4.2c}$$

where β_0 is given in (3.21).

In implementing this two-step method, we need the starting value y_1. This value was provided by the classical Nyström method, using an extremely small time step. Hence, this value can be considered as the exact starting value $y(t_1)$.

Conventional Methods

Finally, for reasons of comparison, we also applied two commonly-used explicit methods: the well-known, second-order Störmer method (cf. (3.19a))

$$y_{n+1} = 2y_n - y_{n-1} + \tau^2 f(t_n, y_n) \tag{4.3}$$

will be used to compare the second-order RKN type methods with, whereas the classical fourth-order Nyström scheme, given by

$$\begin{array}{c|ccc} \frac{1}{2} & \frac{1}{8} & & \\ 1 & 0 & \frac{1}{2} & \\ \hline & \frac{1}{6} & \frac{1}{3} & 0 \\ & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \tag{4.4}$$

will serve as a reference for the fourth-order PC methods.

In the following table we summarize the characteristics of the above methods (an infinite order of dissipation refers to methods for which the stepsize satisfies the periodicity condition):

definition of the method	(4.3)	(3.6)	(3.6a),(3.5)	(3.9)	(3.9)	(4.4)	(4.2a+b)	(4.2a+c)
abbreviation to be used	ST	RKN1	RKN2	ARKN1	ARKN2	NYS	PC1	PC2
algebraic order	2	2	2	2	2	4	4	4
phase lag order: propagated	2	2	4	2	4	4	4	4
phase lag order: inhom.	2	∞	∞	∞	∞	4	∞	6
dissipation order: propagated	∞	∞	∞	∞	∞	4	∞	∞
dissipation order: inhom.	2	∞	2	2	2	4	∞	6
number of f-evaluations per step	1	2	2	1	1	3	2	3
periodicity interval*	$(0,2^2)$	$(0,(2.85)^2)$	$(0,(3.46)^2)$	$(0,\infty)$	$(0,\infty)$	$(0,(2.58)^2)$	$(0,(2.44)^2)$	$(0,(2.30)^2)$

* In the cases (3.6) and (4.2) this periodicity interval varies slowly with v_0.

4.3 A Model Problem

As a first example we consider the model equation

$$y''+\delta^2 y=c\sin(\omega t),\ 0\le t\le T,\ y(0)=0,\ y'(0)=\theta\delta-\frac{\omega c}{-\delta^2+\omega^2}. \tag{4.5}$$

Obviously, the solution is given by

$$y(t)=\theta\sin(\delta t)-\frac{c}{-\delta^2+\omega^2}\sin(\omega t). \tag{4.6}$$

In our experiments, we selected the parameter values $\delta=2$, $\omega=1$ and $c=1$. By means of the parameter θ we can adjust the influence of the homogeneous solution component. Let us start with $\theta=1$. As the endpoint of the integration interval we choose $T=100\pi$; additionally, we measured the dispersion after the first 5 periods. The results of the various schemes can be found in Table 4.1. These results clearly demonstrate that the propagated phase lag in the homogeneous solution component is the major source of phase errors, as is to be expected. The accuracies, listed in this table are in good agreement with the propagated phase lag orders as tabulated in the previous subsection, and it is clear that the PC and RKN type methods hardly benefit from their special features with respect to the inhomogeneous solution component.

Table 4.1. *cd(T)-values for problem (4.5) with $\theta=1$ for several values of T*

method	τ	$T=2\pi$	$T=4\pi$	$T=6\pi$	$T=8\pi$	$T=10\pi$	$T=100\pi$
ST	$\pi/30$	2.0	1.7	1.5	1.4	1.3	0.4
RKN1	$\pi/15$	1.8	1.5	1.4	1.2	1.2	0.4
RKN2	$\pi/15$	3.6	3.3	3.2	3.0	2.9	1.9
ARKN1	$\pi/30$	1.7	1.4	1.2	1.1	1.0	0.5
ARKN2	$\pi/30$	4.8	4.5	4.4	4.2	4.1	3.1
NYS	$\pi/10$	2.6	2.3	2.1	2.0	1.9	1.0
PC1	$\pi/15$	3.6	3.3	3.1	3.0	2.9	1.9
PC2	$\pi/10$	2.8	2.5	2.4	2.2	2.1	1.1

Next, we suppress the homogeneous solution component in the *analytical* solution by setting $\theta=0$. As pointed out in Section 3.1, this does not imply the absence of this component in the *numerical* solution; as a matter of fact, this component is introduced by all schemes and its phase error is propagated as the integration proceeds. Let us write $y(t)=A\sin(\omega t)$ and let the numerical solution be represented by $y_n=\tilde{A}\sin(\omega t_n+\varepsilon)+\tilde{\theta}\sin(\tilde{\delta} t_n)$, where ε and $\tilde{\theta}$ are small. Then

$$y(t_n)-y_n\approx(A-\tilde{A})\sin(\omega t_n)-\tilde{\theta}\sin(\delta t_n)-\varepsilon\omega\tilde{A}\cos(\omega t_n)-\tilde{\theta}(\tilde{\delta}-\delta)t_n\cos(\delta t_n).$$

The behaviour of the error heavily depends on the values of the frequencies ω and δ, and on the points $t_n=T$ where the phase shift is estimated. For instance, in the present experiment (see Table 4.2), $\delta=2\omega$ and T is a multiple of π. Hence,

$$y(t_n)-y_n\approx-\varepsilon\omega\tilde{A}\cos(\omega t_n)-\tilde{\theta}(\tilde{\delta}-\delta)t_n\cos(\delta t_n).$$

Table 4.2. *cd(T)-values for problem (4.5) with $\theta=0$ for several values of T*

method	τ	$T=2\pi$	$T=4\pi$	$T=6\pi$	$T=8\pi$	$T=10\pi$	$T=100\pi$
ST	$\pi/30$	5.5	5.2	5.0	4.9	4.8	3.9
RKN1	$\pi/15$	4.2	3.9	3.7	3.6	3.5	2.7
RKN2	$\pi/15$	6.3	6.0	5.8	5.7	5.6	4.6
ARKN1	$\pi/30$	4.2	3.9	3.7	3.6	3.5	3.0
ARKN2	$\pi/30$	7.3	7.0	6.8	6.7	6.6	5.6
NYS	$\pi/10$	6.0	5.7	5.5	5.4	5.3	4.4
PC1	$\pi/15$	14.0	13.3	13.0	13.0	13.1	11.5
PC2	$\pi/10$	8.3	8.0	7.8	7.7	7.6	6.6

If now the inhomogeneous phase error ε is small with respect to $\tilde{\theta}(\tilde{\delta}-\delta)t_n$, we will observe a linearly increasing phase error at the points T. Generally, however, when δ is not a multiple of ω, we will have an oscillating phase error.

Apart from the parameters ω and δ, the values of $\tilde{A}$, ε and $\tilde{\theta}$ do also determine the error behaviour. For example, both the PC1 and RKN1 methods do not possess an initial phase error of the inhomogeneous type, because they were provided with the exact z_0-value. However, the PC1 method, having a larger propagated phase lag order, behaves much more accurately. For the same reason, the PC2 and RKN2 methods

are superior to the RKN1 scheme. As a consequence, these methods behave significantly more efficiently than the corresponding classical method of the same order. The Störmer method behaves slightly more accurately than ARKN1. This is due to the fact, that the coefficient of the main error term of the propagated phase error is 1/24 for ST and 1/12 for ARKN1. Finally, we conclude from this example, that it is of great importance to use a method by which the numerical homogeneous solution components are also treated adequately, even in cases where the analytical solution only contains inhomogeneous components.

4.4 A Non-Linear Example

As a second example, we consider Duffing's equation, forced by a harmonic function (van Dooren [4])

$$y''(t)+y(t)+y^3(t)=c\cos(\omega t),\ 0\leq t\leq T, \tag{4.7}$$

with the parameter values $c=2_{10}-3$ and $\omega=1.01$. The initial conditions read

$$y(0)=A,\ y'(0)=0, \tag{4.8}$$

where A is obtained from the Galerkin approximation y_G, evaluated at $t=0$:

$$y_G(t)=\sum_{i=0}^{\infty} a_{2i+1}\cos((2i+1)\omega t). \tag{4.9a}$$

Van Dooren calculated an approximation of order 9, having the same frequency as the forcing term; with an absolute precision of 10^{-12}, the coefficients are given by

$$\begin{aligned} &a_1=.200179477536,\ a_3=.246946143_{10}-3,\\ &a_5=.304014_{10}-6,\ a_7=.374_{10}-9,\ a_9=0. \end{aligned} \tag{4.9b}$$

The exact solution (4.9) has its zeros at $t=l\cdot\dfrac{\pi}{2\omega}$, l odd. Table 4.3 shows the phase errors produced by the various schemes at $T=\{1,11,101\}\cdot\pi/2\omega$. The methods PC1 and RKN1, which need a δ-value, were given $\delta=1.0$.

Table 4.3. *cd(T)-values for problem* (4.7), (4.8) *for several values of T*

method	$\tau\cdot\dfrac{2\omega}{\pi}$	$T=\dfrac{\pi}{2\omega}$	$T=11\dfrac{\pi}{2\omega}$	$T=101\dfrac{\pi}{2\omega}$
ST	1/30	3.8	2.7	2.1
RKN1	1/15	4.5	3.5	2.9
RKN2	1/15	4.6	3.6	3.0
ARKN1	1/30	3.4/3.5	2.4/2.4	1.8/1.8
ARKN2	1/30	5.3/4.7	4.3/3.6	3.5/3.0
NYS	1/10	5.5	4.5	3.7
PC1	1/15	7.2	6.2	5.7
PC2	1/10	6.8	6.8	7.4

The *cd*-values for the ARKN methods correspond to $T=f_y(t_n, y_n)/T=$constant part of f_y, respectively. The conclusions for this example are similar to those of the previous one. Again, the accumulated (homogeneous) phase errors mainly determine the accuracy behaviour of the method with the exception of PC2 which behaves different for this example. With $\delta=1$, $\omega=1.01$ we obtain $\sigma_2\approx 1/12$ for RKN1, so that RKN1 and RKN2 are nearly equal for this example. This explains the good performance of RKN1 for this problem.

4.5 A Hyperbolic Equation

As a last example, we test the wave equation (see also [8])

$$\frac{\partial^2 u}{\partial t^2}=g\,d(x)\frac{\partial^2 u}{\partial x^2}+\frac{1}{4}\lambda^2(x,u)\,u+s(t,x;\omega),\ 0\le x\le b,\ 0\le t\le T, \tag{4.10a}$$

where the source term s is prescribed by

$$s(t,x;\omega)=A\cos\left(\frac{\pi x}{b}\right)\sin(\omega t). \tag{4.10b}$$

Here, $d(x)$ is the depth function given by $d=d_0[2+\cos(2\pi x/b)]$, g denotes the acceleration of gravity, and $\lambda(x,u)$ is the coefficient of bottom friction defined by $\lambda=g|u|/C^2 d$ with Chezy coefficient C, where ω is a parameter. The boundary conditions are of the type

$$\frac{\partial u}{\partial x}(t,0)=\frac{\partial u}{\partial x}(t,b)=0 \tag{4.10c}$$

and the initial conditions are given by

$$u(0,x)=0,\ u_t(0,x)=A\,\omega\cos\left(\frac{\pi x}{b}\right). \tag{4.10d}$$

By choosing the initial and boundary conditions consistent with the forced oscillation (4.10b), we expect a solution which is dominated by the inhomogeneous solution component, possessing the same frequency ω. For not too large ω-values, this turned out to be the case.

In the numerical tests, we selected the parameter values

$$g=9.81,\ b=100,\ A=0.1,\ C=50,\ d_0=10. \tag{4.10e}$$

We semi-discretized (4.10) on an equidistant space grid with $\Delta x=b/10$, using second-order symmetric differences. The resulting system of ODEs will be integrated over two periods in time. Results are given for the ninth component of this ODE, i.e. the one which approximates $u(t,x)$ at $x=8\,\Delta x$. As we have no analytical solution available, we determined numerically (using an extremely small time step) the point T where this component has its fourth zero and additionally, we calculated $y'(T)$ (cf. (4.1)).

This test was performed for two values of the parameter ω, viz. $\omega=0.5$ and $\omega=0.1$. For these ω-values we found $T\approx 28.818867$, $y'(T)\approx 1.18$ and $T\approx 125.75714$, $y'(T)\approx 0.067$, respectively. The results for several step sizes are given in Table 4.4;

Table 4.4. *cd(T)-values for problem* (4.10) *with* $\omega=0.5$ *and* $\omega=0.1$, *for several step sizes* τ; $\tau=T/N$

method	$T=28.818867$; $\omega=0.5$				$T=125.75714$; $\omega=0.1$					
	N	cd	N	cd	N	cd	N	cd	N	cd
ST	120	2.1	240	2.7	300	1.0	600	1.6	1200	2.2
RKN1	60	1.7	120	2.9	150	*	300	1.3	600	3.1
RKN2	60	3.0	120	3.6	150	3.0	300	4.4	600	4.5
ARKN1	120	1.8/1.8	240	2.4/2.4	300	0.7/0.7	600	1.3/1.3	1200	1.9/1.9
ARKN2	120	3.5/3.5	240	4.1/4.1	300	4.1/4.1	600	4.4/4.4	1200	4.9/4.9
NYS	40	3.3	80	4.5	100	*	200	2.9	400	4.1
PC1	60	1.7	120	2.6	150	*	300	1.3	600	2.2
PC2	40	3.4	80	5.7	100	*	200	3.0	400	4.2

an "*" denotes an unstable behaviour. The presentation for the ARKN methods is analogous to Table 4.3. The methods PC1 and RKN1, which need a δ-value, were provided with $\delta=10$. However, because this example deals with a *system* of (coupled) ODEs, it is not clear in advance what δ-value should be chosen for these "fitted" methods; and indeed, Table 4.4 shows a rather poor behaviour for these schemes. Moreover, their performance is quite sensitive to the value of δ, as is clear from the following table, where we repeated the experiment for $\omega=0.5$ and $N=60$:

δ-value	1	5	10	15	20
cd-value for RKN1	2.0	2.3	1.7	1.6	1.5
cd-value for PC1	3.4	2.0	1.7	1.6	1.6

4.6 Conclusions

These numerical tests show, that the efficiency of our reduced phase lag methods essentially depends on the propagated phase lag order. It is therefore reasonable to exploit the free parameters of the methods in order to reduce the propagated phase error ($\sigma_2=1/12$ for RKN2, $R_0=P_{22}$ for ARKN2). We want to remark, that the specific advantages of the ARKN methods (unbounded interval of periodicity) appear only at problems possessing quickly oscillating solution components, which however have no or only small influence to the solution (see [11]). If stability requirements are not very strong, then with respect to accuracy and computational effort RKN2 is the best of the methods of algebraic order 2, whereas the PC methods are best of the fourth-order methods.

References

[1] Brusa, L., Nigro, L.: A one-step method for direct integration of structural dynamic equations. Int. J. Num. Meth. in Engng. *15*, 685–699 (1980).

[2] Chawla, M. M., Rao, P. S.: A Noumerov-type method with minimal phase-lag for the integration of second order periodic initial-value problems II. Explicit method. J. Comput. Appl. Math. *15*, 329–337 (1985).

[3] Chawla, M. M., Rao, P. S., Neta, B.: Two-step fourth order P-stable methods with phase-lag of order six for $y'' = f(t, y)$. To appear in J. Comput. Appl. Math. (1986).
[4] Dooren, R. van: Stabilization of Cowell's classical finite difference method for numerical integration. J. Comp. Physics *16*, 186–192 (1974).
[5] Gladwell, I., Thomas, R. M.: Damping and phase analysis for some methods for solving second-order ordinary differential equations. Int. J. Num. Meth. Engng. *19*, 495–503 (1983).
[6] Hairer, E., Wanner, G.: A theory for Nyström methods. Num. Math. *25*, 383–400 (1976).
[7] Houwen, P. J. van der, Sommeijer, B. P.: Predictor-corrector methods with improved absolute stability regions. IMA J. Num. Anal. *3*, 417–437 (1983).
[8] Houwen, P. J. van der, Sommeijer, B. P.: Explicit Runge-Kutta (-Nyström) methods with reduced phase errors for computing oscillating solutions. Report NM-R 8504, Centre of Mathematics and Computer Science, Amsterdam (1985). To appear in SIAM Numer. Anal.
[9] Houwen, P. J. van der, Sommeijer, B. P.: Predictor-corrector methods for periodic second-order initial value problems. Report NM-R 8509, Centre of Mathematics and Computer Science, Amsterdam (1985). To appear in IMA J. Num. Anal. 1987.
[10] Lambert, J. D.: Computational methods in ordinary differential equations. John Wiley & Sons, London (1973).
[11] Strehmel, K., Weiner, R.: Nichtlineare Stabilität und Phasenuntersuchung adaptiver Nyström-Runge-Kutta Methoden. Computing *35*, 325–344 (1985).
[12] Thomas, R. M.: Phase properties of high order, almost P-stable formulae. BIT *24*, 225–238 (1984).
[13] Twizell, E. H.: Phase-lag analysis for a family of multiderivative methods for second order periodic initial value problems. Submitted for publication.

P. J. van der Houwen
B. P. Sommeijer
Centre for Mathematics and
Computer Science
Kruislaan 413
NL-1098 SJ Amsterdam
The Netherlands

K. Strehmel
R. Weiner
Sektion Mathematik
Martin-Luther-Universität
Halle-Wittenberg
Weinbergweg 17
DDR-4020 Halle
German Democratic Republic