Available online at http://scik.org
J. Math. Comput. Sci. 5 (2015), No. 4, 454-461
ISSN: 1927-5307

FALSE CONVERGENCE IN THE NONLINEAR SHOOTING METHOD

J.S.C. PRENTICE

Department of Pure and Applied Mathematics,

University of Johannesburg, Johannesburg, South Africa

Abstract. We study the concept of false convergence in the nonlinear shooting method for boundary-value problems. We show that false convergence is due to the global error that exists in the Runge-Kutta solution to the associated initial-value problem. We show how the situation can be remedied through the device of global error control in the Runge-Kutta method. We suggest the use of the RKQ algorithm for this purpose.

Keywords: False convergence; Runge-Kutta; Boundary Value Problem; Global error.

2010 AMS Subject Classification: 65L05, 65L06.

1. Introduction

A well-known algorithm for finding a numerical solution to the nonlinear boundary-value problem

$$y'' = f\left(x,y,y'\right) \tag{1}$$
$$x \in [a,b] \qquad y(a)=\alpha \qquad y(b)=\beta$$

Received January 6, 2015

is the iterative nonlinear *shooting* method [1 – 5]. This method, to be discussed in more detail in the next section, requires the solution of an initial-value problem of the form

$$\begin{aligned} y' &= z, \\ z' &= f(x,y,z), \\ x &\in [a,b] \qquad y(a)=\alpha \quad z(a)=y'(a)=\theta. \end{aligned} \tag{2}$$

This initial-value problem is usually solved numerically using a Runge-Kutta (RK) method. Of course, an approximation error exists in the RK solution. The effect that this error - the RK global error - has on the convergence of the shooting algorithm is the subject of this paper. We will see that the RK error can lead to erroneous convergence, and we will propose a remedy such that, even if convergence is erroneous, the shooting algorithm will still yield an acceptably accurate result.

In the next few sections we describe notation, terminology and concepts relevant to our work; we present a theoretical discussion, followed by a numerical example; and we make a few relevant comments.

2. Notation, terminology and relevant concepts

The shooting method for (1) is summarized by the iterative procedure

$$\theta_{k+1} = \theta_k - (w_k(b) - \beta)\left(\frac{\theta_k - \theta_{k-1}}{w_k(b) - w_{k-1}(b)}\right). \tag{3}$$

In this expression k indicates the iteration count, θ_k is the approximate value of $y'(a)$ after iteration k, and $w_k(b)$ is the approximation to β obtained using an RK method applied to (2), with $z(a) = \theta_k$. Note that the term in parentheses on the RHS is a finite-difference approximation to the reciprocal of the derivative $dy/d\theta$. This procedure requires that two initial guesses for the slope, θ_1 and θ_2, are specified. The iteration then proceeds until the *residual* $|w_{k+1}(b) - \beta|$ is less than a user-specified tolerance ε. The resulting θ_{k+1} is taken as the slope $y'(a)$, and the RK solution of (2), with $z(a) = \theta_{k+1}$, is taken as the solution to (1).

We define the errors

$$\begin{aligned} \delta_k &\equiv y_k(b) - \beta, \\ \Delta_k &\equiv w_k(b) - y_k(b), \end{aligned}$$

where $y_k(b)$ indicates the *exact* value of (2) when $z(a) = \theta_k$. Of course, $y_k(b)$ is not known and is approximated by $w_k(b)$. Using these definitions we find, for the residual,

$$|w_{k+1}(b) - \beta| = |\delta_{k+1} + \Delta_{k+1}|. \tag{4}$$

The quantity Δ_k is known as the *global error* in the RK solution at b and, in general, we must assume that $\Delta_k \neq 0$. It is clear that $\delta_k = 0$ when $\theta_k = \theta$.

3. Analysis

Our analysis centers on the residual (4). Clearly, the convergence condition

$$|w_{k+1}(b) - \beta| = |\delta_{k+1} + \Delta_{k+1}| \leqslant \varepsilon \tag{5}$$

could be satisfied if

$$\Delta_{k+1} \simeq -\delta_{k+1},$$

and this could be true even if the magnitudes of the errors are larger than ε. In such a case, we would find convergence of (3), even though $|\delta_{k+1}| > \varepsilon$. We refer to this state of affairs as *false convergence.* Also, even if $\theta_k = \theta$, we could find that the residual is larger than the tolerance, because of a large RK global error. In other words, the exact value θ might not yield convergence! It is clear, then, that the RK global error has the capacity to corrupt the shooting algorithm.

We propose the following remedy: assume that the RK global error can be controlled, such that

$$|\Delta_{k+1}| \leqslant \frac{\varepsilon}{2}. \tag{6}$$

Then, if both (5) and (6) are true, we must necessarily have

$$|\delta_{k+1}| \leqslant \frac{3\varepsilon}{2}. \tag{7}$$

In this case, even if false convergence occurs, the magnitude of δ_{k+1} is still bounded by a known value, which can, via ε, be made acceptably small.

4. Numerical example

The false convergence described above, and its remedy, may be demonstrated by means of the test problem

$$\begin{aligned} y'' &= \frac{32+2x^3+yy'}{8} \\ x &\in [1,3] \qquad y(1)=17 \qquad y(3)=\frac{43}{3}, \end{aligned}$$

which has solution $y(x)=x^2+16/x$, so that $\theta=-14$. In applying the shooting method, the initial-value problem that must be solved, for each θ_k, is

$$\begin{aligned} y' &= z, \\ z' &= \frac{32+2x^3+yz}{8}, \\ x &\in [1,3] \qquad y(1)=17 \qquad z(1)=y'(1)=\theta_k. \end{aligned} \tag{8}$$

Of course, when $\theta_k=\theta=-14$, the solution to this initial-value problem is the solution to the test problem.

We use two tolerances - $\varepsilon=10^{-6}$ and $\varepsilon=10^{-10}$ - and, for each, we consider the cases where the RK global error at $x=3$ is controlled and uncontrolled. When controlled, it is limited by $\varepsilon/2$. We use a third-order RK method [6], denoted RK3, to solve (8). We always choose $\theta_1=-6$ and $\theta_2=-7$, although it is understood that this choice is arbitrary. The 'true' solution, for each θ_k, is obtained using an eighth-order RK method (RK8) [7] with a suitably small stepsize (maximal error in the RK8 solution was estimated to be $\sim 10^{-13}$ or less). This RK8 solution enables us to compute δ_k. Obviously, we then compute Δ_k from $\Delta_k=w_{k+1}(3)-43/3-\delta_k$. We also consider the controlled case where $\varepsilon=10^{-4}$, but $|\Delta_k|\leqslant 5\times 10^{-11}$. Results are shown in Table 1. In this table, we have dropped the subscript k; the δ and Δ shown here are the values at the point of convergence.

Table 1. Results for the test problem.

$\varepsilon = 10^{-6}$	**Uncontrolled**	**Controlled**
Δ	0.322513×10^{-4}	0.4983×10^{-6}
δ	-0.322512×10^{-4}	-0.4982×10^{-6}
$\lvert\delta+\Delta\rvert$	0.91×10^{-10}	0.9×10^{-10}
$\varepsilon = 10^{-10}$	**Uncontrolled**	**Controlled**
Δ	0.461549×10^{-6}	0.499×10^{-10}
δ	-0.461456×10^{-6}	0.429×10^{-10}
$\lvert\delta+\Delta\rvert$	0.93×10^{-10}	0.928×10^{-10}
$\varepsilon = 10^{-4}$		**Controlled**
Δ		0.5×10^{-10}
δ		-0.25×10^{-5}
$\lvert\delta+\Delta\rvert$		0.25×10^{-5}

In all cases, $|\delta+\Delta| \leqslant \varepsilon$ but, clearly, $|\delta| \gg \varepsilon$ for the uncontrolled cases. This is the false convergence discussed previously. For the controlled cases, we have $|\delta| < 3\varepsilon/2$, as expected. For the last case, where the tolerance on the shooting iteration is fairly loose, and the RK tolerance is tight, we see that the RK global error does not contaminate the convergence condition, since here $|\Delta| \ll |\delta|$. We believe, however, that tight tolerances on the shooting algorithm are preferable. The resulting solutions for the various cases differed from $\theta = -14$ by $\sim 10^{-5}$ (for the first case, uncontrolled) to $\sim 10^{-11}$ (second case, controlled). Furthermore, the second case, controlled, yields an RK3 solution to (8), with $z(1) = -13.99999999992703$, that differs from the exact solution to the test problem by no more than 1.2×10^{-10} anywhere on $[1,3]$.

5. Comments

(1) In the above example, for the sake of a clear demonstration, we have controlled the RK error only at the endpoint of the interval, by simply choosing a suitably small stepsize.

We have not attempted to control the error anywhere else on the interval (although the small stepsize happens to result in a globally accurate RK solution). In practice, we would want to control the RK error everywhere on the interval, not only at the endpoint. This can be achieved using the RKQ algorithm [8,9], developed recently by us, which facilitates step-by-step control of the global error in the RK solution. We did not use RKQ in the example, since the stepsize adjustments arising in RKQ resulted in a global error at the endpoint that was much smaller than the tolerance, and so the effect of false convergence was not demonstrated clearly. Rather, such a small error at the endpoint $\left(\sim 10^{-12} \text{ when } \varepsilon = 10^{-10}\right)$ actually corresponds to the third case in Table 1, where the RK error is so small that it does not contaminate the shooting algorithm.

(2) In principle, we should write

$$w_k(b) - y_k(b) = \Delta_k + \mu_k,$$

where μ_k denotes a roundoff error that may be present in the numerical quantity $w_k(b)$. This roundoff component, unlike Δ_k, is not proportional to some power of the RK stepsize. However, in our work here we have assumed that it is small enough, compared to Δ_k, to be ignored. This is a good assumption provided that the RK stepsize is not very small, compared to the length of the interval $[a,b]$. This condition is likely to be met if the the RK method used is of reasonably high order, and that the RK tolerance ε is not so small that it is similar to machine precision.

(3) The shooting algorithm used here is based on the root-finding method of Linear Interpolation. A more sophisticated version exists, based on Newton's Method [1], wherein a second initial-value problem must be solved in conjunction with (2). However, even in this version of the method, the condition (5) must still be satisfied, so that false convergence is likely to occur, for the same reasons discussed above.

(4) We note that RK error must be present in the denominator on the RHS of (3). Such error could affect the performance of the iteration process, but we do not think that it would contribute to false convergence, per se. We do not study this effect here.

(5) The tolerance ε considered above is an *absolute* tolerance. If we were to impose a *relative* tolerance on the problem, as in

$$\begin{aligned} \left|\frac{w_{k+1}(b)-\beta}{\beta}\right| &= \left|\frac{\delta_{k+1}+\Delta_{k+1}}{\beta}\right| \leqslant \varepsilon \\ \Rightarrow |\delta_{k+1}+\Delta_{k+1}| &\leqslant \varepsilon|\beta|, \end{aligned}$$

then we simply replace ε with $\varepsilon|\beta|$ in (6) and (7), and the same analysis holds. We make this point because relative error control would usually be preferred when $|\beta|>1$, particularly when $|\beta|$ is very large.

6. Conclusion

We have studied false convergence in the nonlinear shooting method, showing how it arises as a consequence of Runge-Kutta global error. We have suggested that control of the Runge-Kutta global error will result in control of the convergence condition, leading to a satisfactory solution. We have considered an example wherein the Runge-Kutta error has been controlled at the endpoint of the interval of integration (for the sake of demonstration), although we have proposed that control of the Runge-Kutta error over the entire interval might be preferable, and for this purpose the RKQ algorithm may be well-suited.

Conflict of Interests

The author declares that there is no conflict of interests.

REFERENCES

[1] R.L. Burden, J.D. Faires, Numerical Analysis, 9th ed., Brooks/Cole, Pacific Grove (2011).

[2] E. Isaacson, H.B. Keller, Analyisis of Numerical Methods, Dover, New York (1994).

[3] A. Granas, R. B. Guenther, J. W. Lee, The shooting method for the numerical solution of a class of nonlinear boundary value problems, SIAM J. Numer. Anal. 16 (1979), 828–836.

[4] S. N. Ha, A nonlinear shooting method for two-point boundary value problems, Comput. Math. Appl. 42 (2001), 1411–1420.

[5] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, vol. 12 of Texts in Applied Mathematics 2nd ed., Springer, New York (1993).

[6] D. Kincaid, W. Cheney, Numerical Analysis: Mathematics of Scientific Computing 3rd ed., Brooks/Cole, Pacific Grove (2002).

[7] J.C. Butcher, Numerical Methods for Ordinary Differential Equations, Wiley, Chichester (2003).

[8] J. S. C. Prentice, Stepwise global error control in an explicit Runge-Kutta method using local extrapolation with high-order selective quenching, J. Math. Res. 3 (2011), 126-136.

[9] J. S. C. Prentice, Relative global error control in the RKQ algorithm for systems of ordinary differential equations, J. Math. Res. 3 (2011), 59-66.