

# Backward Error Analysis for Perturbation Methods \*

Corless, Robert M.  
Western University  
rcorless@uwo.ca

Fillion, Nicolas  
Simon Fraser University  
nfillion@sfu.ca

September 7, 2016

## Abstract

We demonstrate via several examples how the backward error viewpoint can be used in the analysis of solutions obtained by perturbation methods. We show that this viewpoint is quite general and offers several important advantages. Perhaps the most important is that backward error analysis can be used to demonstrate the validity of the solution, however obtained and by whichever method. This includes a nontrivial safeguard against slips, blunders, or bugs in the original computation. We also demonstrate its utility in deciding when to truncate an asymptotic series, improving on the well-known rule of thumb indicating truncation just prior to the smallest term. We also give an example of elimination of *spurious* secular terms even when genuine secularity is present in the equation. We give short expositions of several well-known perturbation methods together with computer implementations (as scripts that can be modified). We also give a generic backward error based method that is equivalent to iteration (but we believe useful as an organizational viewpoint) for regular perturbation.

## 1 Introduction

As the title suggests, the main idea of this paper is to use backward error analysis (BEA) to assess and interpret solutions obtained by perturbation methods. The idea will seem natural, perhaps even obvious, to those who are familiar with the way in which backward error analysis has seen its scope increase dramatically since the pioneering work of Wilkinson in the 60s, e.g., [35, 36]. From early results in numerical linear algebraic problems and computer arithmetic, it has become a general method fruitfully applied to problems involving root finding, interpolation, numerical differentiation, quadrature, and the numerical solutions of ODEs, BVPs, DDEs, and PDEs, see, e.g., [9, 12, 19]. This is hardly a surprise when one considers that BEA offers several interesting advantages over a purely forward-error approach.

BEA is often used in conjunction with perturbation methods. Not only is it the case that many algorithms' backward error analyses rely on perturbation methods, but the backward error is related to the forward error by a coefficient of sensitivity known as the condition number, which is itself a kind of sensitivity to perturbation. In this paper, we examine an apparently new idea, namely, that perturbation methods themselves can also be interpreted within the backward error analysis framework. Our examples will have a classical feel, but the analysis and interpretation is what differs, and we will make general remarks about the benefits of this mode of analysis and interpretation.

However, due to the breadth of the literature in perturbation theory, we cannot determine with certainty the extent to which applying backward error analysis to perturbation methods is new. Still, none of the works we know, apart from [5], [7], and [39], even mention the possibility of using

---

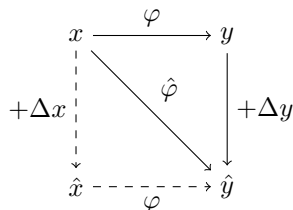
\*We would like to thank Pei Yu, Robert Moir and Julia Jankowski for their various contributions to this paper. We are also indebted to NSERC, Western University, as well as Galima Hassan for the key logistic support they provided.

of using BEA to explain or measure the success of a perturbation computation. Among the books we have consulted, only [5, p. 251 & p. 289] mentions the residual by name, but does not use it systematically. At the very least, therefore, the idea of using BEA in relation to perturbation methods might benefit from a wider discussion.

## 2 The basic method from the BEA point of view

The basic idea of BEA is increasingly well-known in the context of numerical methods. The slogan *a good numerical method gives the exact solution to a nearby problem* very nearly sums up the whole perspective. Any number of more formal definitions and discussions exist—we like the one given in [9, chap. 1], as one might suppose is natural, but one could hardly do better than go straight to the source and consult, e.g., [35, 36, 37, 38]. More recently [17] has offered a good historical perspective. In what follows we give a brief formal presentation and then give detailed analyses by examples in subsequent sections.

Problems can generally be represented as maps from an input space  $\mathcal{I}$  to an output space  $\mathcal{O}$ . If we have a problem  $\varphi : \mathcal{I} \rightarrow \mathcal{O}$  and wish to find  $y = \varphi(x)$  for some putative input  $x \in \mathcal{I}$ , lack of tractability might instead lead you to engineer a simpler problem  $\hat{\varphi}$  from which you would compute  $\hat{y} = \hat{\varphi}(x)$ . Then  $\hat{y} - y$  is the *forward error* and, provided it is small enough for your application, you can treat  $\hat{y}$  as an approximation in the sense that  $\hat{y} \approx \varphi(x)$ . In BEA, instead of focusing on the forward error, we try to find an  $\hat{x}$  such that  $\hat{y} = \varphi(\hat{x})$  by considering the *backward error*  $\Delta x = \hat{x} - x$ , i.e., we try to find for which set of data our approximation method  $\hat{\varphi}$  has exactly solved our reference problem  $\varphi$ . The general picture can be represented by the following commutative diagram:



We can see that, whenever  $x$  itself has many components, different backward error analyses will be possible since we will have the option of reflecting the forward error back into different selections of the components.

It is often the case that the map  $\varphi$  can be defined as the solution to  $\phi(x, y) = 0$  for some operator  $\phi$ , i.e., as having the form

$$x \xrightarrow{\varphi} \{y \mid \phi(x, y) = 0\} . \quad (1)$$

In this case, there will in particular be a simple and useful backward error resulting from computing the residual  $r = \phi(x, \hat{y})$ . Trivially  $\hat{y}$  then exactly solves the reverse-engineered problem  $\hat{\varphi}$  given by  $\hat{\varphi}(x, y) = \phi(x, y) - r = 0$ . Thus, when the residual can be used as a backward error, this directly computes a reverse-engineered problem that our method has solved exactly. We are then in the fortunate position of having both a problem and its solution, and the challenge then consists in determining how similar the reference problem  $\varphi$  and the modified problems  $\hat{\varphi}$  are, *and whether or not the modified problem is a good model for the phenomenon being studied*.

**Regular perturbation BEA-style** Now let us introduce a *general framework for perturbation methods* that relies on the general framework for BEA introduced above. Perturbation methods are so numerous and varied, and the problems tackled are from so many areas, that it seems a general scheme of solution would necessarily be so abstract as to be difficult to use in any particular

case. Actually, the following framework covers many methods. For simplicity of exposition, we will introduce it using the simple gauge functions  $1, \varepsilon, \varepsilon^2, \dots$ , but note that extension to other gauges is usually straightforward (such as Puiseux,  $\varepsilon^n \ln^m \varepsilon$ , etc), as we will show in the examples. To begin with, let

$$F(x, u; \varepsilon) = 0 \tag{2}$$

be the operator equation we are attempting to solve for the unknown  $u$ . The dependence of  $F$  on the scalar parameter  $\varepsilon$  and on any data  $x$  is assumed but henceforth not written explicitly. In the case of a simple power series perturbation, we will take the  $m$ th order approximation to  $u$  to be given by the *finite* sum

$$z_m = \sum_{k=0}^m \varepsilon^k u_k . \tag{3}$$

The operator  $F$  is assumed to be Fréchet differentiable. For convenience we assume slightly more, namely, that for any  $u$  and  $v$  in a suitable region, there exists a linear invertible operator  $F_1(v)$  such that

$$F(u) = F(v) + F_1(v)(u - v) + O(\|u - v\|^2) . \tag{4}$$

Here,  $\|\cdot\|$  denotes any convenient norm. We denote the *residual* of  $z_m$  by

$$\Delta_m := F(z_m) , \tag{5}$$

*i.e.*,  $\Delta_m$  results from evaluating  $F$  at  $z_m$  instead of evaluating it at the reference solution  $u$  as in equation (2). If  $\|\Delta_m\|$  is small, we say we have solved a “nearby” problem, namely, the reverse-engineered problem for the unknown  $u$  defined by

$$F(u) - F(z_m) = 0 , \tag{6}$$

which is exactly solved by  $u = z_m$ . Of course this is trivial. It is *not* trivial in consequences if  $\|\Delta_m\|$  is small compared to data errors or modelling errors in the operator  $F$ . We will exemplify this point more concretely later.

We now suppose that we have somehow found  $z_0 = u_0$ , a solution with a residual whose size is such that

$$\|\Delta_0\| = \|F(u_0)\| = O(\varepsilon) \quad \text{as} \quad \varepsilon \rightarrow 0 . \tag{7}$$

Finding this  $u_0$  is part of the art of perturbation; much of the rest is mechanical. Suppose now inductively that we have found  $z_n$  with residual of size

$$\|\Delta_n\| = O(\varepsilon^{n+1}) \quad \text{as} \quad \varepsilon \rightarrow 0 .$$

Consider  $F(z_{n+1})$  which, by definition, is just  $F(z_n + \varepsilon^{n+1} u_{n+1})$ . We wish to choose the term  $u_{n+1}$  in such a way that  $z_{n+1}$  has residual of size  $\|\Delta_{n+1}\| = O(\varepsilon^{n+2})$  as  $\varepsilon \rightarrow 0$ . Using the Fréchet derivative of the residual of  $z_{n+1}$  at  $z_n$ , we see that

$$\Delta_{n+1} = F(z_n + \varepsilon^{n+1} u_{n+1}) = F(z_n) + F_1(z_n) \varepsilon^{n+1} u_{n+1} + O(\varepsilon^{2n+2}) . \tag{8}$$

By linearity of the Fréchet derivative, we also obtain  $F_1(z_n) = F_1(z_0) + O(\varepsilon) = [\varepsilon^0]F_1(z_0) + O(\varepsilon)$ . Here,  $[\varepsilon^k]G$  refers to the coefficient of  $\varepsilon^k$  in the expansion of  $G$ . Let

$$A = [\varepsilon^0]F_1(z_0) , \tag{9}$$

that is, the zeroth order term in  $F_1(z_0)$ . Thus, we reach the following expansion of  $\Delta_{n+1}$ :

$$\Delta_{n+1} = F(z_n) + A\varepsilon^{n+1}u_{n+1} + O(\varepsilon^{n+2}) . \quad (10)$$

Note that, in equation (8), one could keep  $F_1(z_n)$ , not simplifying to  $A$  and compute not just  $u_{n+1}$  but, just as in Newton's method, double the number of correct terms. However, this in practice is often too expensive [16, chap. 6], and so we will in general use this simplification. As noted, we only need  $F_1(z_0)$  accurate to  $O(\varepsilon)$ , so in place of  $F_1(z_0)$  in equation (10) we use  $A$ .

As a result of the above expansion of  $\Delta_{n+1}$ , we now see that to make  $\Delta_{n+1} = O(\varepsilon^{n+2})$ , we must have  $F(z_n) + A\varepsilon^{n+1}u_{n+1} = O(\varepsilon^{n+2})$ , in which case

$$Au_{n+1} + \frac{F(z_n)}{\varepsilon^{n+1}} = Au_{n+1} + \frac{\Delta_n}{\varepsilon^{n+1}} = O(\varepsilon) . \quad (11)$$

Since by hypothesis  $\Delta_n = F(z_n) = O(\varepsilon^{n+1})$ , we know that  $\Delta_n/\varepsilon^{n+1} = O(1)$ . In other words, to find  $u_{n+1}$  we solve the linear operator equation

$$Au_{n+1} = -[\varepsilon^{n+1}]\Delta_n ,$$

where, again,  $[\varepsilon^{n+1}]$  is the coefficient of the  $(n+1)$ th power of  $\varepsilon$  in the series expansion of  $\Delta$ . Note that by the inductive hypothesis the right hand side has norm  $O(1)$  as  $\varepsilon \rightarrow 0$ . Then  $\|\Delta_{n+1}\| = O(\varepsilon^{n+2})$  as desired, so  $u_{n+1}$  is indeed the coefficient we were seeking. We thus need  $A = [\varepsilon^0]F(z_0)$  to be invertible. If not, the problem is singular, and essentially requires reformulation.<sup>1</sup> We shall see examples. If  $A$  is invertible, the problem is regular.

This general scheme can be compared to that of, say, [2]. Essential similarities can be seen. In Bellman's treatment, however, the residual is used implicitly, but not named or noted, and instead the equation defining  $u_{n+1}$  is derived by postulating an infinite expansion

$$u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \dots . \quad (12)$$

By taking the coefficient of  $\varepsilon^{n+1}$  in the expansion of  $\Delta_n$  we are implicitly doing the same work, but we will see advantages of this point of view. Also, note that in the frequent case of more general asymptotic sequences, namely Puiseux series or generalized approximations containing logarithmic terms, we can make the appropriate changes in a straightforward manner, as we will show below.

### 3 Algebraic equations

We begin by applying the regular method from section 2 to algebraic equations. We begin with a simple scalar equation and gradually increase the difficulty, thereby demonstrating the flexibility of the backward error point of view.

#### 3.1 Regular perturbation

In this section, after applying the method from section 2 to a scalar equation, we use the same method to solve a  $2 \times 2$  system; higher dimensional systems can be solved similarly. We give some computer algebra implementations (scripts that the reader may modify) of the basic method. Finally, in this section, we give an alternative method based on the Davidenko equation that is simpler to use in Maple.

---

<sup>1</sup>We remark that it is a sufficient but not necessary condition for regular expansion to be able to find our initial point  $u_0$  and to have invertible  $A = F_1(u_0; 0)$ . A regular perturbation problem can be defined in many ways, not just in the way we have done, with invertible  $A$ . For example, [3, Sec 7.2] essentially uses continuity in  $\varepsilon$  as  $\varepsilon \rightarrow 0$  to characterize it. Another characterization is that for regular perturbation problems infinite perturbation series are convergent for some non-zero radius of convergence.

### 3.1.1 Scalar equations

Let us consider a simple example similar to many used in textbooks for classical perturbation analysis. Suppose we wish to find a real root of

$$x^5 - x - 1 = 0 \quad (13)$$

and, since the Abel-Ruffini theorem—which says that in general there are no solutions in radicals to equations of degree 5 or more—suggests it is unlikely that we can find an elementary expression for the solution of this *particular* equation of degree 5, we introduce a parameter which we call  $\varepsilon$ , and moreover which we suppose to be small. That is, we embed our problem in a parametrized family of similar problems. If we decide to introduce  $\varepsilon$  in the degree-1 term, so that

$$u^5 - \varepsilon u - 1 = 0, \quad (14)$$

we will see that we have a so-called regular perturbation problem.

To begin with, we wish to find a  $z_0$  such that  $\Delta_0 = F(z_0) = z_0^5 - \varepsilon z_0 - 1 = O(\varepsilon)$ . Quite clearly, this can happen only if  $z_0^5 - 1 = 0$ . Ignoring the complex roots in this example, we take  $z_0 = 1$ . To continue the solution process, we now suppose that we have found

$$z_n = \sum_{k=0}^n u_k \varepsilon^k \quad (15)$$

such that  $\Delta_n = F(z_n) = z_n^5 - \varepsilon z_n - 1 = O(\varepsilon^{n+1})$  and we wish to use our iterative procedure. We need the Fréchet derivative of  $F$ , which in this case is just

$$F_1(u) = 5u^4 - \varepsilon, \quad (16)$$

because

$$F(u) = u^5 - \varepsilon u - 1 = v^5 - \varepsilon v - 1 + F'(v)(u - v) + O(u - v)^2. \quad (17)$$

Hence,  $A = 5z_0^4 = 5$ , which is invertible. As a result our iteration is  $\Delta_n = F(z_n)$ , i.e.,

$$5u_{n+1} = -[\varepsilon^{n+1}]\Delta_n. \quad (18)$$

Carrying out a few steps we have

$$\Delta_0 = F(z_0) = F(1) = 1 - \varepsilon - 1 = -\varepsilon \quad (19)$$

so

$$5 \cdot u_1 = -[\varepsilon]\Delta_0 = -[\varepsilon](-\varepsilon) = \varepsilon. \quad (20)$$

Thus,  $u_1 = \varepsilon/5$ . Therefore,  $z_1 = 1 + \varepsilon/5$  and

$$\Delta_1 = \left(1 + \frac{\varepsilon}{5}\right)^5 - \varepsilon \left(1 + \frac{\varepsilon}{5}\right) - 1 \quad (21)$$

$$= \left(1 + 5\frac{\varepsilon}{5} + 10\frac{\varepsilon^2}{25} + O(\varepsilon^3)\right) - \varepsilon - \frac{\varepsilon^2}{5} - 1 \quad (22)$$

$$= \left(\frac{2}{5} - \frac{1}{5}\right)\varepsilon^2 + O(\varepsilon^3) = \frac{1}{5}\varepsilon^2 + O(\varepsilon^3). \quad (23)$$

Then we find that  $Au_1 = -1/5$  and thus  $u_1 = -1/25$ . So,  $u = 1 + \varepsilon/5 - \varepsilon^2/25 + O(\varepsilon^3)$ . Finding more terms by this method is clearly possible although tedium might be expected at higher orders. Luckily

nowadays computers and programs are widely available that can solve such problems without much human effort, but before we demonstrate that, let's compute the residual of our computed solution so far:

$$z_2 = 1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2.$$

Then  $\Delta_2 = z_2^5 - \varepsilon z_2 - 1$  is

$$\begin{aligned} \Delta_2 &= \left(1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2\right)^5 - \varepsilon \left(1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2\right) - 1 \\ &= -\frac{1}{25}\varepsilon^3 - \frac{3}{125}\varepsilon^4 + \frac{11}{3125}\varepsilon^5 + \frac{3}{125}\varepsilon^6 - \frac{2}{15625}\varepsilon^7 \\ &\quad - \frac{1}{78125}\varepsilon^8 + \frac{1}{390625}\varepsilon^9 - \frac{1}{9765625}\varepsilon^{10}. \end{aligned} \tag{24}$$

We note the following. First,  $z_2$  exactly solves the modified equation

$$x^5 - \varepsilon x - 1 + \frac{1}{25}\varepsilon^3 + \frac{3}{25}\varepsilon^4 - \dots + \frac{1}{9765625}\varepsilon^{10} = 0 \tag{25}$$

which is  $O(\varepsilon^3)$  different to the original. Second, the complete residual was computed rationally: there is no error in saying that  $z_2 = 1 + \varepsilon/5 - \varepsilon^2/25$  solves equation (25) exactly. Third, if  $\varepsilon = 1$  then  $z_2 = 1 + 1/5 - 1/25 = 1.16$  exactly (or  $14/25$  if you prefer), and the residual is then  $(29/25)^5 - 29/25 - 1 \doteq -0.059658$ , showing that 1.16 is the exact root of an equation about 6% different to the original.

Something simple but importantly different to the usual treatment of perturbation methods has happened here. We have assessed the quality of the solution in an explicit fashion without concern for convergence issues or for the exact solution to  $x^5 - x - 1 = 0$ , which we term the reference problem. We use this term because its solution will be the reference solution. We can't call it the "exact" solution because  $z_2$  is *also* an "exact" solution, namely to equation (25).

Every numerical analyst and applied mathematician knows that this isn't the whole story—we need some evaluation or estimate of the effects of such perturbations of the problem. One effect is the difference between  $z_2$  and  $x$ , the reference solution, and this is what people focus on. We believe this focus is sometimes excessive. There are other possible views. For instance, if the backward error is physically reasonable. As an example, if  $\varepsilon = 1$  and  $z_2 = 1.16$  then  $z_2$  exactly solves  $y^5 - y - a = 0$  where  $a \neq 1$  but rather  $a \doteq 0.9403$ . If the original equation was really  $u^5 - u - \alpha = 0$  where  $\alpha = 1 \pm 5\%$  we might be inclined to accept  $z_2 = 1.16$  because, for all we know, we might have the true solution (even though we're outside the  $\pm 5\%$  range, we're only just outside; and how confident are we in the  $\pm 5\%$ , after all?).

### 3.1.2 Simple computer algebra solution

The following Maple script can be used to solve this or similar problems  $f(u; \varepsilon) = 0$ . Other computer algebra systems can also be used.

```
# Perturbation solution of F(u;epsilon) = 0
restart; #top down execution should have a clean state to begin.
macro(e = epsilon); #saves typing "epsilon" every time.
F := z -> z^5 - e*z - 1;
# Zeroth order solution, by inspection, is
z := 1; #solve(eval(F(z), e=0), z);
A := coeff(series((D(F))(z), e, 1), e, 0); #A must not be 0 for regularity
N := 3; #number of terms
Delta := F(z); #initial residual, must be O(e)
# Now, the iteration:
```

```

for k to N do
  u := -coeff(series(Delta, e, k+1), e, k);
  z := z + u*e^k/A;
  Delta := F(z);
end do;
z;
series(Delta, e, N+3);

```

That code is a straightforward implementation of the general scheme presented in subsection 2. Its results, translated into L<sup>A</sup>T<sub>E</sub>X and cleaned up a bit, are that

$$z = 1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2 + \frac{1}{125}\varepsilon^3 \quad (26)$$

and that the residual of this solution is

$$\Delta = \frac{21}{3125}\varepsilon^5 + O(\varepsilon^6) . \quad (27)$$

With  $N = 3$ , we get an extra order of accuracy as the next term in the series is zero, but this result is serendipitous.

### 3.1.3 Systems of algebraic equations

Regular perturbation for systems of equations using the framework from section 2 is straightforward. We include an example to show some computer algebra and for completeness. Consider the following two equations in two unknowns:

$$f_1(v_1, v_2) = v_1^2 + v_2^2 - 1 - \varepsilon v_1 v_2 = 0 \quad (28)$$

$$f_2(v_1, v_2) = 25v_1 v_2 - 12 + 2\varepsilon v_1 = 0 \quad (29)$$

When  $\varepsilon = 0$  these equations determine the intersections of a hyperbola with the unit circle. There are four such intersections:  $(3/5, 4/5)$ ,  $(4/5, 3/5)$ ,  $(-3/5, -4/5)$  and  $(-4/5, -3/5)$ . The Jacobian matrix (which gives us the Fréchet derivative in the case of algebraic equations) is

$$F_1(v) = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} \end{bmatrix} = \begin{bmatrix} 2v_1 & 2v_2 \\ 25v_2 & 25v_1 \end{bmatrix} + O(\varepsilon) . \quad (30)$$

Taking for instance  $u_0 = [3/5, 4/5]^T$  we have

$$A = F_1(u_0) = \begin{bmatrix} 6/5 & 8/5 \\ 20 & 15 \end{bmatrix} . \quad (31)$$

Since  $\det A = -14 \neq 0$ ,  $A$  is invertible and indeed

$$A^{-1} = \begin{bmatrix} -15/14 & 4/25 \\ 10/7 & -3/35 \end{bmatrix} . \quad (32)$$

The residual of the zeroth order solution is

$$\Delta_0 = F \left( \frac{3}{5}, \frac{4}{5} \right) = \begin{bmatrix} -12/25 \\ 6/5 \end{bmatrix} , \quad (33)$$

so  $-\varepsilon \Delta_0 = [12/25, -6/5]^T$ . Therefore

$$u_1 = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = A^{-1} \begin{bmatrix} 12/25 \\ -6/25 \end{bmatrix} = \begin{bmatrix} -114/175 \\ 138/175 \end{bmatrix} \quad (34)$$

and  $z_1 = u_0 + \varepsilon u_1$  is our improved solution:

$$z_1 = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix} + \varepsilon \begin{bmatrix} -114/175 \\ 138/175 \end{bmatrix}. \quad (35)$$

To guard against slips, blunders, and bugs (some of those calculations were done by hand, and some were done in Sage on an Android phone) we compute

$$\Delta_1 = F(z_1) = \varepsilon^2 \begin{bmatrix} 6702/6125 \\ -17328/1225 \end{bmatrix} + O(\varepsilon^3). \quad (36)$$

That computation was done in Maple, completely independently. Initially it came out  $O(\varepsilon)$  indicating that something was not right; tracking the error down we found a typo in the Maple data entry (183 was entered instead of 138). Correcting that typo we find  $\Delta_1 = O(\varepsilon^2)$  as it should be. Here is the corrected Maple code:

```
# Residual computation for a system of two equations
restart; #top down execution should have a clean state to begin.
macro(e = epsilon); #saves typing "epsilon" every time.
f1 := (v1,v2) -> v1^2 + v2^2 - 1 - e*v1*v2;
f2 := (v1,v2) -> 25*v1*v2 - 12 + 2*e*v1;
z11 := 3/5 + e*(-114/175);
z12 := 4/5 + e*138/175;
Delta11 := series( f1(z11,z12), e, 3);
Delta12 := series( f2(z11,z12), e, 3);
```

Just as for the scalar case, this process can be systematized and we give one way to do so in Maple, below. The code is not as pretty as the scalar case is, and one has to explicitly “map” the series function and the extraction of coefficients onto matrices and vectors, but this demonstrates feasibility.

```
# Residual computation for a system of two equations
restart; #top down execution should have a clean state to begin.
macro(e = epsilon); #saves typing "epsilon" every time.
z := Vector(2,[3/5,4/5]); # z_0 = u_0
F := u -> Vector(2,
    [ u[1]^2 + u[2]^2 - 1 - e*u[1]*u[2],
      25*u[1]*u[2] - 12 + 2*e*u[1] ] );
A := VectorCalculus[Jacobian](
    [ F([x,y])[1], F([x,y])[2] ], [x,y] );
A := eval( A, [x=z[1], y=z[2], e=0] );
N := 3;
Delta := F(z);
for k to N do
    u := map(t -> -coeff( t, e, k ),
             map( series, Delta, e, k+1 )
            );
    z := z + LinearAlgebra[LinearSolve]( A, u )*e^k;
    Delta := F( z );
end do;
z;
map( series, Delta, e, N+2 );
```

This code computes  $z_3$  correctly and gives a residual of  $O(\varepsilon^4)$ . From the backward error point of view, this code finds the intersection of curves that differ from the specified ones by terms of  $O(\varepsilon^4)$ .



In the next section, we show a way to use a built-in feature of Maple to do the same thing with less human labour.

### 3.1.4 The Davidenko equation

Maple has a built-in facility for solving differential equations in series that (at the time of writing) is superior to its built-in facility for solving algebraic equations in series, because the latter can only handle scalar equations. This may change in the future, but it may not because there is the following simple workaround. To solve

$$F(u; \varepsilon) = 0 \tag{37}$$

for a function  $u(\varepsilon)$  expressed as a series, simply differentiate to get

$$D_1(F)(u, \varepsilon) \frac{du}{d\varepsilon} + D_2(F)(u, \varepsilon) = 0. \tag{38}$$

Boyd [5] calls this the Davidenko equation. If we solve this in Taylor series with the initial condition  $u(0) = u_0$ , we have our perturbation series. Notice that what we were calling  $A = [\varepsilon^0]F_1(u_0)$  occurs here as  $D_1(F)(u_0, 0)$  and this needs to be nonsingular to be solved as an ordinary differential equation; if  $\text{rank}(D_1(F)(u_0, 0)) < n$  then this is in fact a nontrivial differential algebraic equation that Maple may still be able to solve using advanced techniques (see, e.g., [1]). Let us just show a simple case here:

```
# Residual computation for a system of two equations
restart; #top down execution should have a clean state to begin.
macro(e = epsilon); #saves typing "epsilon" every time.
Order := 4;
z := Vector(2, [3/5, 4/5]); # z_0 = u_0
F := u -> Vector(2,
    [ u[1]^2 + u[2]^2 - 1 - e*u[1]*u[2],
      25*u[1]*u[2] - 12 + 2*e*u[1] ] );
Zer := F( [ x(e), y(e) ] ); #This asks for F evaluated at functions x(e)
# and y(e) that are yet unspecified.
diffeqs := { diff( Zer[1], e ), diff( Zer[2], e ) }; #This creates a set
# of two differential equations, one from each component of F.
# Each equation will contain both dx/de and dy/de.
iniconds := { x(0) = z[1] , y(0) = z[2] };
sol := dsolve( diffeqs union iniconds , {x(e), y(e)}, type=series );
Delta := eval( F( [x(e), y(e)] ), map(convert, sol, polynom) );
map( series, Delta, e, Order+2 );
```

This generates (to the specified value of the order, namely, **Order=4**) the solution

$$x(\varepsilon) = \frac{3}{5} - \frac{114}{175}\varepsilon + \frac{119577}{42875}\varepsilon^2 - \frac{43543632}{2100875}\varepsilon^3 \tag{39}$$

$$y(\varepsilon) = \frac{4}{5} + \frac{138}{175}\varepsilon - \frac{119004}{42875}\varepsilon^2 + \frac{43245168}{2100875}\varepsilon^3, \tag{40}$$

whose residual is  $O(\varepsilon^4)$ . Internally, Maple uses its own algorithms, which occasionally get improved as algorithmic knowledge advances.

## 3.2 Puiseux series

Puiseux series are simply Taylor series or Laurent series with fractional powers. A standard example is

$$\sin \sqrt{x} = x^{1/2} - \frac{1}{3!}x^{3/2} + \frac{1}{5!}x^{5/2} + \dots \quad (41)$$

A simple change of variable (e.g.  $t = \sqrt{x}$  so  $x = t^2$ ) is enough to convert to Taylor series. Once the appropriate power  $n$  is known for  $\varepsilon = \mu^n$ , perturbation by Puiseux expansion reduces to computations similar to those we've seen already. For instance, had we chosen to embed  $u^5 - u - 1$  in the family  $u^5 - \varepsilon(u + 1)$  (which is somehow conjugate to the family of the last section), then because the equation becomes  $u^5 = 0$  when  $\varepsilon = 0$  we see that we have a five-fold root to perturb, and we thus suspect we will need Puiseux series.

For scalar equations, there are built-in facilities in Maple for Puiseux series, which gives yet another way in Maple to solve scalar algebraic equations perturbatively. One can use the `RootOf` construct to do so as follows:

```
restart;
macro(e = epsilon);
Order := 2;
alias(alpha = RootOf(z^5-1, z));
f := u -> u^5 - e*(u+1);
z := convert(series(RootOf(f(u), u), e), polynom);
Delta := series(f(z), e, Order+2);
map(simplify, Delta);
```

This yields

$$z = \alpha \varepsilon^{1/5} + \frac{1}{5} \alpha^2 \varepsilon^{2/5} - \frac{1}{25} \alpha^3 \varepsilon^{3/5} + \frac{1}{125} \alpha^4 \varepsilon^{4/5} - \frac{21}{15626} \alpha \varepsilon^{6/5}. \quad (42)$$

This series describes all paths, accurately for small  $\varepsilon$ . Note that the command

```
alias(alpha = RootOf(u^5-1, u))
```

is a way to tell Maple that  $\alpha$  represents a fixed fifth root of unity. Exactly which fixed root can be deferred till later. Working instead with the default value for the environment variable `Order`, namely `Order := 6`, gets us a longer series for  $z$  containing terms up to  $\varepsilon^{29/5}$  but not  $\varepsilon^{30/5} = \varepsilon^6$ . Putting the resulting  $z_6$  back into  $f(u)$  we get a residual

$$\Delta_6 = f(z_6) = \frac{23927804441356816}{14551915228366851806640625} \varepsilon^7 + O(\varepsilon^8) \quad (43)$$

Thus we expect that for small  $\varepsilon$  the residual will be quite small. For instance, with  $\varepsilon = 1$  the exact residual is, for  $\alpha = 1$ ,  $\Delta_6 = 1.2 \cdot 10^{-9}$ . This tells us that this approximation ought to get us quite accurate roots, and indeed we do.

We conclude this discussion with two remarks. The first is that by a discriminant analysis as we describe in section 3.3, we find that the nearest singularity is at  $\varepsilon = 3125/256$ , and so we expect this series to actually converge for  $\varepsilon = 1$ . Again, this fact was not used in our analysis above. Secondly, we could have used the `series/RootOf` technique to do both the regular perturbation in subsection 3.1 or the singular one we will do in subsection 3.3. The Maple commands are quite similar:

```
series(RootOf(u^5-e*u-1, u), e);
```

and

```
series(RootOf(e*u^5-u-1, u), e);
```

However, in both cases only the real root is expanded. Some “Maple art” (that one of us more readily characterizes as black magic) can be used to complete the computation, but the previous code (both the loop and the Davidenko equation) are easier to generalize. Making the `dsolve/series` code for the Davidenko equation work in the case of Puiseux series requires a preliminary scaling.

### 3.3 Singular perturbation

Suppose that instead of embedding  $u^5 - u - 1 = 0$  in the regular family we used in the previous section, we had used  $\varepsilon u^5 - u - 1 = 0$ . If we run our previous Maple programs, we find that the zeroth order solution is unique, and  $z_0 = -1$ . The Fréchet derivative is  $-1$  to  $O(\varepsilon)$ , and so  $u_{n+1} = [\varepsilon^{n+1}] \Delta_n$  for all  $n \geq 0$ . We find, for instance,

$$z_7 = -1 - \varepsilon - 5\varepsilon^2 - 35\varepsilon^3 - 285\varepsilon^4 - 2530\varepsilon^5 - 23751\varepsilon^6 - 231880\varepsilon^7 \quad (44)$$

which has residual  $\Delta_7 = O(\varepsilon^8)$  but with a larger integer as the constant hidden in that  $O$  symbol. For  $\varepsilon = 0.2$ , the value of  $z_7$  becomes

$$z_7 \doteq -7.4337280 \quad (45)$$

while  $\Delta_7 = -4533.64404$ , which is not small at all. Thus we have no evidence this perturbation solution is any good: we have the exact solution to  $u^5 - 0.2u - 1 = -4533.64404$  or  $u^5 - 0.2u + 4532.64404 = 0$ , probably not what was intended (and if it was, it would be a colossal fluke). Note that we do not need to know a reference value of a root of  $u^5 - 0.2u - 1$  to determine this. Trying a smaller  $\varepsilon$ , we find that if  $\varepsilon = 0.05$  we have  $z_7 \doteq -1.07$  and  $\Delta_7 \doteq -1.2 \cdot 10^{-4}$ . This means  $z_7$  is an exact root of  $u^5 - 0.05u - 1.00012$ ; which may very well be what we want.

The following remark is not really germane to the method but it’s interesting. Taking the discriminant with respect to  $u$ , i.e., the resultant of  $f$  and  $\partial f/\partial u$ , we find  $\text{discrim}(f) = \varepsilon^3(3125\varepsilon - 256)$ . Thus  $f$  will have multiple roots if  $\varepsilon = 0$  (there are 4 multiple roots at infinity) or if  $\varepsilon = 256/3125 = 0.08192$ . Thus our perturbation expansion can be expected to diverge<sup>2</sup> for  $\varepsilon \geq 0.08192$ . What happens to  $z_7$  if  $\varepsilon = 256/3125$ ?  $z_7 \doteq -1.1698$  and  $\Delta_7 = -9.65 \cdot 10^{-3}$ , so we have an exact solution for  $u^5 - 256/3125u - 1.00965$ ; this is not bad. The reference double root is  $-1.25$ , about 0.1 away, although this fact was not used in the previous discussion.

But this computation, valid as it is, only found one root out of five, and then only for sufficiently small  $\varepsilon$ . We now turn to the roots that go to infinity as  $\varepsilon \rightarrow 0$ . Preliminary investigation from similar to that of subsection 3.2 shows that it is convenient to replace  $\varepsilon$  by  $\mu^4$ . Many singular perturbation problems including this one can be turned into regular ones by rescaling. Putting  $u = y/\mu$ , we get

$$\mu^4 \left( \frac{y}{\mu} \right)^5 - \frac{y}{\mu} - 1 = 0, \quad (46)$$

which reduces to

$$y^5 - y - \mu = 0. \quad (47)$$

This is now regular in  $\mu$ . The zeroth order the equation is  $y(y^4 - 1) = 0$  and the root  $y = 0$  just recovers the regular series previously attained; so we let  $\alpha$  be a root of  $y^4 - 1$ , i.e.,  $\alpha \in \{1, -1, i, -i\}$ . A very similar Maple program (to either of the previous two) gives

$$y_5 = \alpha + \frac{1}{4}\mu - \frac{5}{32}\alpha^3\mu^2 + \frac{5}{32}\alpha^2\mu^3 - \frac{385}{2048}\alpha\mu^4 + \frac{1}{4}\mu^5 \quad (48)$$

---

<sup>2</sup>A separate analysis leads to the identification of  $u_k = \frac{1}{5k+1} \binom{5k+1}{k}$  (via [27]). The ratio test confirms that the series converges for  $|\varepsilon| < 256/3125$ , and diverges if  $\varepsilon = 256/3125$ .

so our approximate solution is  $y_5/\mu$  or

$$z_5 = \frac{\alpha}{\mu} + \frac{1}{4} - \frac{5}{32}\alpha^3\mu^2 - \frac{385}{2048}\alpha\mu^3 + \frac{1}{4}\mu^4 \quad (49)$$

which has residual *in the original equation*

$$\Delta_5 = \mu^4 z_5^5 - z - 1 = \frac{23205}{16384}\alpha^3\mu^5 - \frac{21255}{65536}\alpha^2\mu^6 + O(\mu^7). \quad (50)$$

That is,  $z_5$  exactly solves  $\mu^4 u^5 - u - 1 - \frac{23205}{16384}\alpha^2\mu^5 = O(\mu^6)$  instead of the one we had wanted to solve. This differs from the original by  $O(|\varepsilon|^{5/4})$ , and for small enough  $\varepsilon$  this may suffice.

**Optimal backward error** Interestingly enough, we can do better. The residual is only one kind of backward error. Taking the lead from the Oettli-Prager theorem [9, chap. 6], we look for equations of the form

$$\left( \mu^4 + \sum_{j=10}^{15} a_j \mu^j \right) u^5 - u - 1 \quad (51)$$

for which  $z_5$  is a better solution yet. Simply equating coefficients of the residual

$$\tilde{\Delta}_5 = \left( \mu^4 + \sum_{j=10}^{15} a_j \mu^j \right) z_5^5 - z_5 - 1 \quad (52)$$

to zero, we find

$$\left( \mu^4 - \frac{23205}{16384}\alpha^2\mu^{10} + \frac{2145}{1024}\alpha\mu^{11} \right) z_5^5 - z_5 - 1 = \frac{12165535425}{1073741824}\alpha\mu^{11} + O(\mu^{12}) \quad (53)$$

and thus  $z_5$  solves an equation that is  $O(\mu^{10/4}) = O(\varepsilon^{5/2})$  close to the original, not just an equation (50) that is  $O(\mu^6) = O(|\varepsilon|^{5/4})$ . This is a superior explanation of the quality of  $z_5$ . This was obtained with the following Maple code:

```
# Perturbation solution of F(u;epsilon) = 0
restart;
e := mu^4;
Forig := z -> e*z^5 - z - 1;
F := y -> y^5 - y - mu;
# Zeroth order solution, by inspection:
alias(alpha = RootOf(Z^4-1, Z));
y := alpha;
A := coeff( series( (D(F))(y), mu, 1), mu, 0);
A := simplify(A);
N := 5;
Delta := simplify( F(y) );
for k to N do
    u := -coeff( series(Delta, mu, k+1), mu, k);
    y := y+u*mu^k/A;
    Delta := simplify( F(y) );
end do;
y;
series(Delta, mu, N+3);
```

```

M := 5+2*N;
modified := u -> (mu^4 + add(a[j]*mu^j, j = 5+N..M))*u^5 - u - 1;
z := series(y/mu, mu, N+1);
zer := series(modified(z), mu, M+1);
eqs := [seq(simplify(coeff(zer, mu, k)), k = N .. M-5)];
sol := solve(eqs, [seq(a[j], j = 5+N .. M)]);
perreq := eval(modified(U), sol[1]);
newresid := eval(perreq, U = z);
map(simplify, series(newresid, mu, M+2));

```

Computing to higher orders (see the worksheet) gives e.g. that  $z_8$  is the exact solution to an equation that differs by  $O(\mu^{13})$  from the original, or better than  $O(\varepsilon^3)$ . This in spite of the fact that the basic residual  $\Delta_8 = O(\varepsilon^{9/4})$ , only slightly better than  $O(\varepsilon^2)$ .

We will see other examples of improved backward error over residual for singularly-perturbed problems. In retrospect it's not so surprising, or shouldn't have been: singular problems are sensitive to changes in the leading term, and so it takes less effort to match a given solution.

### 3.4 Perturbing all roots at once

The preceding analysis found a nearby equation for each root independently; this might suffice, but there are circumstances in which it might not. Perhaps we want a “nearby” equation satisfied by all roots at once. Sadly this is more difficult, and in general may not be possible. But it is possible for the example we've considered and we demonstrate how the backward error is used in such a case. Let

$$\zeta_1 = z_5(1) = \frac{1}{\mu} + \frac{1}{4} - \frac{5}{32}\mu - \frac{385}{2048}\mu^3 + \frac{1}{4}\mu^4 \quad (54)$$

$$\zeta_2 = z_5(-1) = -\frac{1}{\mu} + \frac{1}{4} - \frac{5}{32}\mu + \frac{385}{2048}\mu^3 + \frac{1}{4}\mu^4 \quad (55)$$

$$\zeta_3 = z_5(i) = \frac{i}{\mu} + \frac{1}{4} + \frac{5}{32}\mu - \frac{385i}{2048}\mu^3 + \frac{1}{4}\mu^4 \quad (56)$$

$$\zeta_4 = z_5(-i) = -\frac{i}{\mu} + \frac{1}{4} + \frac{5}{32}\mu + \frac{385}{2048}\mu^3 + \frac{1}{4}\mu^4 \quad (57)$$

$$\zeta_5 = z_5 = -1 - \mu^4 - 5\mu^8, \quad (58)$$

$\zeta_5$  is the regular root we have found first in the previous subsection. Now put

$$\tilde{p}(x) = \mu^4(x - \zeta_1)(x - \zeta_2)(x - \zeta_3)(x - \zeta_4)(x - \zeta_5) \quad (59)$$

and expand it. The result, by Maple, is

$$\begin{aligned} \mu^4 x^5 - 5\mu^{12} x^4 + \left( \frac{23205}{16384} \mu^8 + \frac{45}{8} \mu^{12} \right) x^3 - \left( \frac{5435}{32768} \mu^8 + \frac{195697915}{33554432} \mu^{12} \right) x^2 \\ + \left( \frac{2575665}{2097152} \mu^8 + \frac{5696429035}{1073741824} \mu^{12} - 1 \right) x + \frac{8453745}{2097152} \mu^8 - \frac{5355037365}{1073741824} \mu^{12} - 1 \end{aligned} \quad (60)$$

which equals

$$\varepsilon x^5 - x - 1 - 5\varepsilon^3 x^4 + \left( \frac{23205}{16384} \varepsilon^2 + \frac{45}{8} \varepsilon^3 \right) x^3 - \left( \frac{5435}{32768} \varepsilon^2 + \dots \right) x^2 + O(\varepsilon^2) \quad (61)$$

As we see, this equation is remarkably close to the original, although we see changes in all the coefficients. The backward error is  $O(\mu^8)$ , i.e.,  $O(\varepsilon^2)$ . Thus for algebraic equations it's possible to talk about simultaneous backward error.

### 3.5 A hyperasymptotic example

In [5, sect. 15.3, pp. 285-288], Boyd takes up the perturbation series expansion of the root near  $-1$  of

$$f(x, \varepsilon) = 1 + x + \varepsilon \operatorname{sech}\left(\frac{x}{\varepsilon}\right) = 0, \quad (62)$$

a problem he took from [20, p. 22]. After computing the desired expansion using a two-variable technique, Boyd then sketches an alternative approach suggested by one of us (based on [10]), namely to use the Lambert  $W$  function. Unfortunately, there are a number of sign errors in Boyd's equation (15.28). We take the opportunity here to offer a correction, together with a residual-based analysis that confirms the validity of the correction. First, the erroneous formula: Boyd has

$$z_0 = \frac{W(-2e^{1/\varepsilon})\varepsilon - 1}{\varepsilon} \quad (63)$$

and  $x_0 = -\varepsilon z_0$ , so allegedly  $x_0 = 1 - \varepsilon W(-2\varepsilon^{1/\varepsilon})$ . This can't be right: as  $\varepsilon \rightarrow 0^+$ ,  $e^{1/\varepsilon} \rightarrow \infty$  and the argument to  $W$  is negative and large; but  $W$  is real only if its argument is between  $-e^{-1}$  and 0, if it's negative at all. We claim that the correct formula is

$$x_0 = -1 - \varepsilon W(2e^{-1/\varepsilon}) \quad (64)$$

which shows that the errors in Boyd's equation (15.28) are explainable as trivial. Indeed, Boyd's derivation is correct up to the last step; rather than fill in the algebraic details of the derivation of formula (64), we here verify that it works by computing the residual:

$$\Delta_0 = 1 + x_0 + \varepsilon \operatorname{sech}\left(\frac{x_0}{\varepsilon}\right). \quad (65)$$

For notational simplicity, we will omit the argument to the Lambert  $W$  function and just write  $W$  for  $W(2e^{-1/\varepsilon})$ . Then, note that  $\operatorname{sech}(x_0/\varepsilon) = \operatorname{sech}(1 + \varepsilon W/\varepsilon)$  since each  $\operatorname{sech}$  is even, and that

$$\operatorname{sech}\left(\frac{x_0}{\varepsilon}\right) = \frac{2}{e^{x_0/\varepsilon} + e^{-x_0/\varepsilon}} = \frac{1}{e^{(1/\varepsilon)+W} + e^{-1/\varepsilon-W}}. \quad (66)$$

Now, by definition,

$$We^W = 2e^{-1/\varepsilon} \quad (67)$$

and thus we obtain

$$e^W = \frac{2e^{-1/\varepsilon}}{W} \quad \text{and} \quad e^{-W} = \frac{We^{1/\varepsilon}}{2}. \quad (68)$$

It follows that

$$\operatorname{sech}\left(\frac{x_0}{\varepsilon}\right) = \frac{2}{2/W + W/2} = \frac{W}{1 + W^2/4}, \quad (69)$$

and hence the residual is

$$\begin{aligned} \Delta_0 &= 1 + (-1 - \varepsilon W) + \varepsilon \frac{W}{1 + W^2/4} = \frac{-\varepsilon W(1 + W^2/4) + \varepsilon W}{1 + W^2/4} \\ &= \frac{-\varepsilon W^3/4}{1 + W^2/4} = \frac{-\varepsilon W^3}{4 + W^2}. \end{aligned} \quad (70)$$

Now  $W = W(2e^{-1/\varepsilon})$  and as  $\varepsilon \rightarrow 0^+$ ,  $2e^{-1/\varepsilon} \rightarrow 0$  rapidly; since the Taylor series for  $W(z)$  starts as  $W(z) = z - z^2 + \frac{3}{2}z^3 + \dots$ , we have that  $W(2e^{-1/\varepsilon}) \sim 2e^{-1/\varepsilon}$  and therefore

$$\Delta_0 = -\varepsilon 2e^{-3/\varepsilon} + O(e^{-5/\varepsilon}). \quad (71)$$

We see that this residual is very small indeed. But we can say even more. Boyd leaves us the exercise of computing higher order terms; here is our solution to the exercise. A Newton correction would give us

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \quad (72)$$

and we have already computed  $f(x_0) = \Delta_0$ . What is  $f'(x_0)$ ? Since  $f(x) = 1 + x + \varepsilon \operatorname{sech}(x/\varepsilon)$ , this derivative is

$$f'(x) = 1 - \operatorname{sech}\left(\frac{x}{\varepsilon}\right) \tanh\left(\frac{x}{\varepsilon}\right). \quad (73)$$

Simplifying similarly to equation (69), we obtain

$$\tanh\left(\frac{x_0}{\varepsilon}\right) = \frac{e^{1/\varepsilon+W} - e^{-1/\varepsilon-W}}{e^{1/\varepsilon+W} + e^{-1/\varepsilon+W}} = \frac{\frac{2}{W} - \frac{W}{2}}{\frac{2}{W} + \frac{W}{2}} = \frac{4 - W^2}{4 + W^2}. \quad (74)$$

Thus

$$f'(x_0) = 1 - \operatorname{sech}\left(\frac{x_0}{\varepsilon}\right) \tanh\left(\frac{x_0}{\varepsilon}\right) = 1 - \frac{W(1 - W^2/4)}{(1 + W^2/4)^2}. \quad (75)$$

It follows that

$$x_1 = x_0 - \frac{\Delta_0}{f'(x_0)} = -1 - \varepsilon W + \frac{\varepsilon W^3/4 + W^2}{1 - \frac{W(1 - W^2/4)}{(1 + W^2/4)^2}} \quad (76)$$

$$= -1 - \varepsilon W + \frac{\varepsilon W^3(4 + W^2)}{16 - 16W + 8W^2 + 4W^3 + W^4} \quad (77)$$

$$= -1 - \varepsilon W + \frac{\varepsilon}{4}W^3 + \frac{\varepsilon}{4}W^4 + \frac{3}{16}\varepsilon W^5 - \frac{11}{64}\varepsilon W^6 + O(W^7) \quad (78)$$

Finally, the residual of  $x_1$  is

$$\Delta_1 = 4\varepsilon e^{7/\varepsilon} + O(\varepsilon e^{-8/\varepsilon}). \quad (79)$$

We thus see an example of the use of  $f'(x_0)$  instead of just  $A$ , as discussed in section 2, to approximately double the number of correct terms in the approximation.

This analysis can be implemented in Maple as follows:

```
restart;
with(MultiSeries);
macro(e = epsilon);
alias(W = LambertW);
f := x -> 1 + x + e*sech(x/e);
df := D(f);
x[0] := -1 - e*W(2*exp(-1/e));
Delta[0] := f(x[0]);
series(Delta[0], e, 3);
```

```

x[1] := x[0] - Delta[0]/df(x[0]);
Delta[1] := f(x[1]);
s := multiseries(x[1], e=0);
scale := SeriesInfo[Scale](s);
multiseries(x[1], scale, 3);
multiseries(Delta[1], scale, 5);
# In what follows we have substituted expressions in W for sech and tanh
# since Maple couldn't simplify the expression well.
restart;
macro(e = epsilon);
x[1] := -1-e*W+e*W^3/((4+W^2)*(1-W*(1-(1/4)*W^2)/(1+(1/4)*W^2)^2));
change := factor(x[1]+1+e*W);
series(change, W=0, 8);

```

Note that we had to use the MultiSeries package [31] to expand the series in equation (79), for understanding how accurate  $z_2$  was.  $z_2$  is slightly more lacunary than the two-variable expansion in [5], because we have a zero coefficient for  $W^2$ .

## 4 Divergent Asymptotic Series

Before we begin, a note about the section title: some authors give the impression that the word “asymptotic” is used *only* for divergent series, and so the title might seem redundant. But the proper definition of an asymptotic series can include convergent series (see, e.g., [11]), as it means that the relevant limit is not as the number of terms  $N$  goes to infinity, but rather as the variable in question (be it  $\varepsilon$ , or  $x$ , or whatever) approaches a distinguished point (be it 0, or infinity, or whatever). In this sense, an asymptotic series might diverge as  $N$  goes to infinity, or it might converge, but typically we don't care. We concentrate in this section on divergent asymptotic series.

Beginning students are often confused when they learn the usual “rule of thumb” for optimal accuracy when using divergent asymptotic series, namely to truncate the series *before* adding in the smallest (magnitude) term. This rule is usually motivated by an analogy with *convergent* alternating series, where the error is less than the magnitude of the first term neglected. But why should this work (if it does) for divergent series?

The answer we present in this section isn't as clear-cut as we would like, but nonetheless we find it explanatory. Perhaps you and your students will, too. The basis for the answer is that one can measure the residual  $\Delta$  that arises on truncating the series at, say,  $M$  terms, and choose  $M$  to minimize the residual. Since the forward error is bounded by the condition number times the size of the residual, by minimizing  $\|\Delta\|$  one minimizes a bound on the forward error. It often turns out that this method gives the same  $M$  as the rule of thumb, though not always.

An example may clarify this. We use the large- $x$  asymptotics of  $J_0(x)$ , the zeroth-order Bessel function of the first kind. In [13, section 10.17(i)], we find the following asymptotic series, which is attributed to Hankel:

$$J_0(x) = \left(\frac{2}{\pi x}\right)^{1/2} \left(A(x) \cos\left(x - \frac{\pi}{4}\right) - B(x) \sin\left(x - \frac{\pi}{4}\right)\right) \quad (80)$$

where

$$A(x) = \sum_{k \geq 0} \frac{a_{2k}}{x^{2k}} \quad \text{and} \quad B(x) = \sum_{k \geq 0} \frac{a_{2k+1}}{x^{2k+1}} \quad (81)$$



and where

$$a_0 = 1$$

$$a_k = \frac{(-1)^k}{k!8^k} \prod_{j=1}^k (2j-1)^2. \quad (82)$$

For the first few  $a_k$ s, we get

$$a_0 = 1, a_1 = -\frac{1}{8}, a_2 = -\frac{9}{128}, a_3 = \frac{75}{1024}, \quad (83)$$

and so on. The ratio test immediately shows the two series (81) diverge for all finite  $x$ .

Luckily, we always have to truncate anyway, and if we do, the forward errors get arbitrarily small so long as we take  $x$  arbitrarily large. Because the Bessel functions are so well-studied, we have alternative methods for computation, for instance

$$J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin \theta) d\theta \quad (84)$$

which, given  $x$ , can be evaluated numerically (although it's ill-conditioned in a relative sense near any zero of  $J_0(x)$ ). So we can directly compute the forward error. But let's pretend that we can't. We have the asymptotic series, and not much more. Or course we have to have a defining equation—Bessel's differential equation

$$x^2 y'' + xy' + x^2 y = 0 \quad (85)$$

with the appropriate normalizations at  $\infty$ . We look at

$$y_{N,M} = \left(\frac{2}{\pi x}\right)^{1/2} A_N(x) \cos\left(x - \frac{\pi}{4}\right) - \frac{2}{\pi x} B_M(x) \cos\left(x - \frac{\pi}{4}\right) \quad (86)$$

where

$$A_N(x) = \sum_{k=0}^N \frac{a_{2k}}{x^{2k}} \quad \text{and} \quad B_M(x) = \sum_{k=0}^M \frac{a_{2k+1}}{x^{2k+1}}. \quad (87)$$

Inspection shows that there are only two cases that matter: when we end on an even term  $a_{2k}$  or on an odd term  $a_{2k+1}$ . The first terms omitted will be odd and even. A little work shows that the residual

$$\Delta = x^2 y''_{N,M} + xy'_{N,M} + x^2 y_{N,M} \quad (88)$$

is just

$$\frac{(k+1/2)^2 a_k}{x^{k+1/2}} \cdot \left\{ \begin{array}{l} \cos(x - \pi/4) \\ \sin(x - \pi/4) \end{array} \right\} \quad (89)$$

if the final term kept, odd or even, is  $a_k$ . If even, then multiply by  $\cos(x - \pi/4)$ ; if odd, then  $\sin(x - \pi/4)$ .

Let's pause a moment. The algebra to show this is a bit finicky but not hard (the equation is, after all, linear). This end result is an extremely simple (and exact!) formula for  $\Delta$ . The finite series  $y_{N,M}$  is then the exact solution to

$$x^2 y'' + xy' + xy = \Delta \quad (90)$$

$$= \frac{(k+1/2)^2 a_k}{x^{k+1/2}} \cdot \left\{ \begin{array}{l} \cos(x - \frac{\pi}{4}) \\ \sin(x - \frac{\pi}{4}) \end{array} \right\} \quad (91)$$

and, provided  $x$  is large enough, this is only a small perturbation of Bessel's equation. In many modelling situations, such a small perturbation may be of direct physical significance, and we'd be done. Here, though, Bessel's equation typically arises as an intermediate step, after separation of variables, say. Hence one might be interested in the forward error. By the theory of Green's functions, we may express this as

$$J_0(x) - y_{N,M}(x) = \int_x^\infty K(x, \xi) \Delta(\xi) d\xi \quad (92)$$

for a suitable kernel  $K(x, \xi)$ . The obvious conclusion is that if  $\Delta$  is small then so will  $J_0(x) - y_{N,M}(x)$ ; but  $K(x, \xi)$  will have some effect, possibly amplifying the effects of  $\Delta$ , or perhaps even damping its effects. Hence, the connection is indirect.

To have an error in  $\Delta$  of at most  $\varepsilon$ , we must have

$$\left(k + \frac{1}{2}\right)^2 \frac{|a_k|}{x^{k+1/2}} \leq \varepsilon \quad (93)$$

(remember,  $x > 0$ ). This will happen only if

$$x \geq \left(\left(k + \frac{1}{2}\right)^2 \frac{|a_k|}{\varepsilon}\right)^{2/(2k+1)} \quad (94)$$

and this, for fixed  $k$ , goes to  $\infty$  as  $\varepsilon \rightarrow 0$ . Alternatively, we may ask which  $k$ , for a fixed  $x$ , minimizes

$$\left(k + \frac{1}{2}\right)^2 \frac{|a_k|}{x^{k+1/2}} \quad (95)$$

and this answers the truncation question in a rational way. In this particular case, minimizing  $\|\Delta\|$  doesn't necessarily minimize the forward error (although, it's close). For  $x = 2.3$ , for instance, the sequence  $(k + 1/2)^2 |a_k| x^{-k-1/2}$  is (no  $\sqrt{2/\pi}$ )

$k$	0	1	2	3	4	5	
$A_k$	0.165	0.081	0.055	0.049	0.054	0.070	(96)

The clear winner seems to be  $k = 3$ . This suggests that for  $x = 2.3$ , the best series to take is

$$y_3 = \left(\frac{2}{\pi x}\right)^{1/2} \left( \left(1 - \frac{9}{128x^2}\right) \cos\left(x - \frac{\pi}{4}\right) + \left(\frac{1}{8x} - \frac{75}{1024x^3}\right) \sin\left(x - \frac{\pi}{4}\right) \right). \quad (97)$$

This gives  $5.454 \cdot 10^{-2}$  for  $x = 2.3$ . But the cosine versus sine plays a role, here:  $\cos(2.3 - \pi/4) \doteq 0.056$  while  $\sin(2.3 - \pi/4) \doteq 0.998$ , so we should have included this. When we do, the estimates for  $\Delta_0, \Delta_2$  and  $\Delta_4$  are all significantly reduced—and this changes our selection, and makes  $k = 4$  the right choice;  $\Delta_6 > \Delta_4$  as well (either way). But the influence of the integral is mollifying. Comparing to a better answer (computers via the integral formula) 0.0555398, we see that the error is about  $8.8 \cdot 10^{-4}$  whereas  $((4 + 1/2)^2 a_4 / 2.3^{4+1/2}) \cos(2.3 - \pi/4)$  is  $3.06 \cdot 10^{-3}$ ; hence the residual overestimates the error slightly.

How does the rule of thumb do? The first term that is neglected here is  $(1/x)^{1/2} a_5 x^{-5} \sin(x - \pi/4)$  which is  $\sim 2.3 \cdot 10^{-3}$  apart from the  $(2/\pi)^{1/2} = 0.797$  factor, so about  $1.86 \cdot 10^{-3}$ . The next term is, however,  $(2/\pi x)^{1/2} a_6 x^{-6} \cos(x - \pi/4) \doteq -1.14 \cdot 10^{-4}$  which is smaller yet, suggesting that we should keep the  $a_5$  term. But we shouldn't. Stopping with  $a_4$  gives a better answer, just as the residual suggests that it should.

We emphasize that this is only a slightly more rational rule of thumb, because minimizing  $\|\Delta\|$  only minimizes a bound on the forward error, not the forward error itself. Still, we have not seen

this discussed in the literature before. A final comment is that the defining equation and its scale, define also the scale for what's a "small" residual.

So, a justification for the "rule of thumb" would be as follows. In our general scheme,

$$Au_{n+1} = -[\varepsilon^{n+1}]\Delta_n \quad (98)$$

and thus, loosely speaking,

$$u_{n+1} \sim -A^{-1}\Delta_n + O(\varepsilon^{n+1}). \quad (99)$$

Thus, if we stop when  $u_{n+1}$  is smallest, this would tend to happen at the same integer  $n$  that  $\Delta_n$  was smallest.

This isn't going to be always true. For instance, if  $A$  is a matrix with largest singular value  $\sigma_1$  and smallest  $\sigma_N > 0$ , with associated vectors  $\hat{u}_k$  and  $\hat{v}_k$ , so that

$$A\hat{v}_k = \sigma_k\hat{u}_k. \quad (100)$$

Then, if  $u_{n+1}$  is like  $\hat{v}_1$  then  $\Delta_n$  will be like  $\sigma_1\hat{u}_1$ , which can be substantially larger; contrariwise, if  $u_{n+1}$  is like  $\hat{v}_N$  then  $A\hat{v}_N = \sigma_N\hat{u}_N$  and  $\Delta_n$  can be substantially smaller. The point is that directions of  $\Delta_n$  can change between steps in the perturbation expansion; we thus expect correlation but not identity.

## 5 Initial-Value problems

BEA has successfully been applied to the *numerical* solution of differential equations for a long time, now. Examples include the works of Enright since the 1980s, e.g., [14, 15], and indeed the Lanczos  $\tau$ -method is yet older [23]. It was pointed out in [8] and [7] that BEA could be used for perturbation and other series solutions of differential equations, also. We here display several examples illustrating this fact. We use regular expansion, matched asymptotic expansions, the renormalization group method, and the method of multiple scales.

### 5.1 Duffing's Equation

This proposed way of interpreting solutions obtained by perturbation methods has interesting advantages for the analysis of series solutions to differential equations. Consider for example an unforced weakly nonlinear Duffing oscillator, which we take from [3]:

$$y'' + y + \varepsilon y^3 = 0 \quad (101)$$

with initial conditions  $y(0) = 1$  and  $y'(0) = 0$ . As usual, we assume that  $0 < \varepsilon \ll 1$ . Our discussion of this example does not provide a new method of solving this problem, but instead it improves the interpretation of the quality of solutions obtained by various methods.

#### 5.1.1 Regular expansion

The classical perturbation analysis supposes that the solution to this equation can be written as the power series

$$y(t) = y_0(t) + y_1(t)\varepsilon + y_2(t)\varepsilon^2 + y_3(t)\varepsilon^3 + \dots \quad (102)$$

Substituting this series in equation (101) and solving the equations obtained by equating to zero the coefficients of powers of  $\varepsilon$  in the residual, we find  $y_0(t)$  and  $y_1(t)$  and we thus have the solution

$$z_1(t) = \cos(t) + \varepsilon \left( \frac{1}{32} \cos(3t) - \frac{1}{32} \cos(t) - \frac{3}{8} t \sin(t) \right). \quad (103)$$

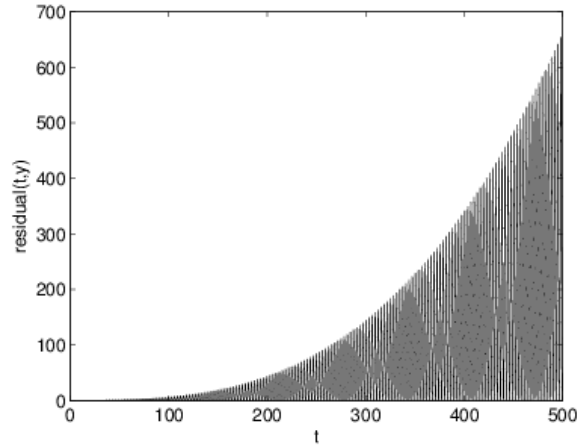


Figure 1: Absolute Residual for the first-order classical perturbative solution of the unforced weakly damped Duffing equation with  $\varepsilon = 0.1$ .

The difficulty with this solution is typically characterized in one of two ways. Physically, the secular term  $t \sin t$  shows that our simple perturbative method has failed since the energy conservation prohibits unbounded solutions. Mathematically, the secular term  $t \sin t$  shows that our method has failed since the periodicity of the solution contradicts the existence of secular terms.

Both these characterizations are correct, but require foreknowledge of what is physically meaningful or of whether the solutions are bounded. In contrast, interpreting (103) from the backward error viewpoint is much simpler. To compute the residual, we simply substitute  $z_2$  in equation (101), that is, the residual is defined by

$$\Delta_1(t) = z_1'' + z_1 + \varepsilon z_1^3. \quad (104)$$

For the first-order solution of equation (103), the residual is

$$\Delta_1(t) = \left( -\frac{3}{64} \cos(t) + \frac{3}{128} \cos(5t) + \frac{3}{128} \cos(3t) - \frac{9}{32} t \sin(t) - \frac{9}{32} t \sin(3t) \right) \varepsilon^2 + O(\varepsilon^3). \quad (105)$$

$\Delta_1(t)$  is exactly computable. We don't print it all here because it's too ugly, but in figure 1, we see that the complete residual grows rapidly. This is due to the secular term  $-\frac{9}{32}t(\sin(t) - \sin(3t))$  of equation (105). Thus we come to the conclusion that the secular term contained in the first-order solution obtained in equation (103) invalidate it, but this time we do not need to know in advance what to physically expect or to prove that the solution is bounded. This is a slight but sometimes useful gain in simplicity.<sup>3</sup>

A simple Maple code makes it possible to easily obtain higher-order solutions:

```
#Regular Expansion for Duffing's Equation
#We choose initial conditions y(0)=1 and y'(0)=0 so y(t)=cos(t) to O(e).
restart;
macro(e=epsilon);
```

<sup>3</sup>In addition, this method makes it easy to find mistakes of various kinds. For instance, a typo in the 1978 edition of [3] was uncovered by computing the residual. That typo does not seem to be in the later editions, so it's likely that the authors found and fixed it themselves.

```

N := 3;
Order := N+1;
z := add(y[k](t)*epsilon^k, k = 0 .. N);
DE := y -> diff(y, t, t)+y+e*y^3;
des := series( DE(z), e);
dos := dsolve({coeff(des, e, 0), y[0](0) = 1, (D(y[0]))(0) = 0}, y[0](t));
assign(dos);
for k to N do
    tmp:=dsolve({coeff(des, e, k), y[k](0)=0, (D(y[k]))(0)=0}, y[k](t));
    assign(tmp);
end do;
Delta := DE(z);
ResidualSeries := map(combine, series(Delta, e, Order+3), trig);

```

Experiments with this code suggests the conjecture that  $\Delta_n = O(t^n \varepsilon^{n+1})$ . For this to be small, we must have  $\varepsilon t = o(1)$  or  $t < O(1/\varepsilon)$ .

### 5.1.2 Lindstedt's method

The failure to obtain an accurate solution on unbounded time intervals by means of the classical perturbation method suggests that another method that eliminates the secular terms will be preferable. A natural choice is Lindstedt's method, which rescales the time variable  $t$  in order to cancel the secular terms. The idea is that if we use a rescaling  $\tau = \omega t$  of the time variable and chose  $\omega$  wisely the secular terms from the classical perturbation method will cancel each other out.<sup>4</sup> Applying this transformation, equation (101) becomes

$$\omega^2 y''(\tau) + y(\tau) + \varepsilon y^3(\tau) \quad y(0) = 1, \quad y'(0) = 0. \quad (106)$$

In addition to writing the solution as a truncated series

$$z_1(\tau) = y_0(\tau) + y_1(\tau)\varepsilon \quad (107)$$

we expand the scaling factor as a truncated power series in  $\varepsilon$ :

$$\omega = 1 + \omega_1 \varepsilon. \quad (108)$$

Substituting (107) and (108) back in equation (106) to obtain the residual and setting the terms of the residual to zero in sequence, we find the equations

$$y_0'' + y_0 = 0, \quad (109)$$

so that  $y_0 = \cos(\tau)$ , and

$$y_1'' + y_1 = -y_0^3 - 2\omega_1 y_0'' \quad (110)$$

subject to the same initial conditions,  $y_0(0) = 1, y_0'(0) = 0, y_1(0) = 0$ , and  $y_1'(0) = 0$ . By solving this last equation, we find

$$y_1(\tau) = \frac{31}{32} \cos(\tau) + \frac{1}{32} \cos(3\tau) - \frac{3}{8} \tau \sin(\tau) + \omega_1 \tau \sin(\tau). \quad (111)$$

So, we only need to choose  $\omega_1 = 3/8$  to cancel out the secular terms containing  $\tau \sin(\tau)$ . Finally, we simply write the solution  $y(t)$  by taking the first two terms of  $y(\tau)$  and plug in  $\tau = (1 + 3\varepsilon/8)t$ :

$$z_1(t) = \cos \tau + \varepsilon \left( \frac{31}{32} \cos \tau + \frac{1}{32} \cos 3\tau \right) \quad (112)$$

---

<sup>4</sup>Interpret this as: we choose  $\omega$  to keep the residual small over as long a time-interval as possible.

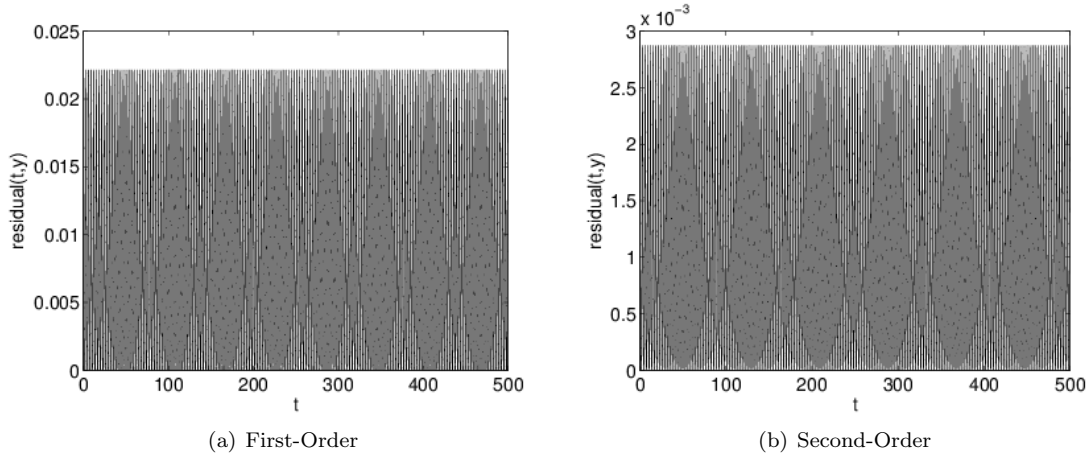


Figure 2: Absolute Residual for the Lindstedt solutions of the unforced weakly damped Duffing equation with  $\varepsilon = 0.1$ .

This truncated power series can be substituted back in the left-hand side of equation (101) to obtain an expression for the residual:

$$\Delta_1(t) = \left( \frac{171}{128} \cos(t) + \frac{3}{128} \cos(5t) + \frac{9}{16} \cos(3t) \right) \varepsilon^2 + O(\varepsilon^3) \quad (113)$$

See figure 2(a). We then do the same with the second term  $\omega_2$ . The following Maple code has been tested up to order 12:

```
# Elimination of Secular Terms in the Solution of the Duffing Equation
# with the Poincare-Lindstedt method.
restart;
macro(e=epsilon);
N := 12;
Order := N+1;
z := add(y[k](tau)*e^k, k = 0..N);
omega := 1+add(a[k]*e^k, k = 1..N);
DE := y -> omega^2*(diff(y, tau, tau))+y+e*y^3;
des := series(DE(z), e);
dos := dsolve({coeff(des, e, 0), y[0](0)=1, (D(y[0]))(0)=0}, y[0](tau));
assign(dos);
for k to N do
    tmp := convert(combine(coeff(des, e, k), trig), exp);
    UZ := eval(tmp, [exp(I*tau) = Z, exp(-I*tau) = 1/Z]);
    ah := coeff(UZ, Z, 1);
    antiseccular := solve(ah = 0, a[k]);
    if {antiseccular} <> {} then
        a[k] := antiseccular;
    end if;
    tmp := dsolve({evalc(tmp), y[k](0)=0, (D(y[k]))(0)=0}, y[k](tau));
    assign(tmp);
end do;
Delta := DE(z);
```

Sdelta := map(simplify, series(Delta, epsilon, Order+4));  
map(combine, Sdelta, trig);

The significance of this is as follows: The normal presentation of the method first requires a proof (an independent proof) that the reference solution is bounded and therefore the secular term  $\varepsilon t \sin t$  in the classical solution is spurious. *But* the residual analysis needs no such proof. It says directly that the classical solution solves not

$$f(t, y, y', y'') = 0 \quad (114)$$

nor  $f + \Delta f = 0$  for uniformly small  $\Delta$  but rather that the residual *departs* from 0 and is *not* uniformly small whereas the residual for the Lindstedt solution *is* uniformly small.

## 5.2 Morrison's counterexample

In [28, pp. 192-193], we find a discussion of the equation

$$y'' + y + \varepsilon(y')^3 + 3\varepsilon^2(y') = 0. \quad (115)$$

O'Malley attributed the equation to [25]. The equation is one that is supposed to illustrate a difficulty with the (very popular and effective) method of multiple scales. We give a relatively full treatment here because a residual-based approach shows that the method of multiple scales, applied somewhat artfully, can be quite successful and moreover we can demonstrate *a posteriori* that the method was successful. The solution sketched in [28] uses the complex exponential format, which one of us used to good effect in his PhD, but in this case the real trigonometric form leads to slightly simpler formulæ. We are very much indebted to our colleague, Professor Pei Yu at Western, for his careful solution, which we follow and analyze here.<sup>5</sup>

The first thing to note is that we will use three time scales,  $T_0 = t$ ,  $T_1 = \varepsilon t$ , and  $T_2 = \varepsilon^2 t$  because the DE contains an  $\varepsilon^2$  term, which will prove to be important. Then the multiple scales formalism gives

$$\frac{d}{dt} = \frac{\partial}{\partial T_0} + \varepsilon \frac{\partial}{\partial T_1} + \varepsilon^2 \frac{\partial}{\partial T_2} \quad (116)$$

This formalism gives most students some pause, at first: replace an ordinary derivative by a sum of partial derivatives using the chain rule? What could this mean? But soon the student, emboldened by success on simple problems, gets used to the idea and eventually the conceptual headaches are forgotten.<sup>6</sup> But sometimes they return, as with this example.

To proceed, we take

$$y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + O(\varepsilon^3) \quad (117)$$

and equate to zero like powers of  $\varepsilon$  in the residual. The expansion of  $d^2 y/dt^2$  is straightforward:

$$\left( \frac{\partial}{\partial T_0} + \varepsilon \frac{\partial}{\partial T_1} + \varepsilon^2 \frac{\partial}{\partial T_2} \right)^2 (y_0 + \varepsilon y_1 + \varepsilon^2 y_2) = \frac{\partial^2 y_0}{\partial T_0^2} + \varepsilon \left( \frac{\partial^2 y_1}{\partial T_0^2} + 2 \frac{\partial^2 y_0}{\partial T_0 \partial T_1} \right) + \varepsilon^2 \left( \frac{\partial^2 y_2}{\partial T_0^2} + 2 \frac{\partial^2 y_1}{\partial T_0 \partial T_1} + \frac{\partial^2 y_0}{\partial T_1^2} + 2 \frac{\partial^2 y_0}{\partial T_0 \partial T_1} \right) \quad (118)$$

<sup>5</sup>We had asked him to solve this problem using one of his many computer algebra programs; instead, he presented us with an elegant handwritten solution.

<sup>6</sup>This can be made to make sense, after the fact. We imagine  $F(T_1, T_2, T_3)$  describing the problem, and  $d/dt = \partial F/\partial T_1 \partial T_1/\partial t + \partial F/\partial T_2 \partial T_2/\partial t + \partial F/\partial T_3 \partial T_3/\partial t$  which gives  $d/dt = \partial F/\partial T_1 + \varepsilon \partial F/\partial T_2 + \varepsilon^2 \partial F/\partial T_3$  if  $T_1 = t$ ,  $T_2 = \varepsilon t$  and  $T_3 = \varepsilon^2 t$ .

For completeness we include the other necessary terms, even though this construction may be familiar to the reader. We have

$$\begin{aligned} \varepsilon \left( \frac{dy}{dt} \right)^3 &= \varepsilon \left( \left( \frac{\partial}{\partial T_0} + \varepsilon \frac{\partial}{\partial T_1} \right) (y_0 + \varepsilon y_1) \right)^3 \\ &= \varepsilon \left( \frac{\partial y_0}{\partial T_0} \right)^3 + 3\varepsilon^2 \left( \frac{\partial y_0}{\partial T_0} \right)^2 \left( \frac{\partial y_0}{\partial T_1} + \frac{\partial y_1}{\partial T_0} \right) + \dots, \end{aligned} \quad (119)$$

and  $y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2$  is straightforward, and also

$$3\varepsilon^2 \left( \left( \frac{\partial}{\partial T_0} + \dots \right) (y_0 + \dots) \right) = 3\varepsilon^2 \frac{\partial y_0}{\partial T_0} + \dots \quad (120)$$

is at this order likewise straightforward. At  $O(\varepsilon^0)$  the residual is

$$\frac{\partial^2 y_0}{\partial T_0^2} + y_0 = 0 \quad (121)$$

and without loss of generality we take as solution

$$y_0 = a(T_1, T_2) \cos(T_0 + \varphi(T_1, T_2)) \quad (122)$$

by shifting the origin to a local maximum when  $T_0 = 0$ . For notational simplicity put  $\theta = T_0 + \varphi(T_1, T_2)$ . At  $O(\varepsilon^1)$  the equation is

$$\frac{\partial^2 y_1}{\partial T_0^2} + y_1 = - \left( \frac{\partial y_0}{\partial T_0} \right)^3 - 2 \frac{\partial^2 y_0}{\partial T_0 \partial T_1} \quad (123)$$

where the first term on the right comes from the  $\varepsilon \dot{y}^3$  term whilst the second comes from the multiple scales formalism. Using  $\sin^3 \theta = 3/4 \sin \theta - 1/4 \sin 3\theta$ , this gives

$$\frac{\partial^2 y_1}{\partial T_0^2} + y_1 = \left( 2 \frac{\partial a}{\partial T_1} + \frac{3}{4} a^3 \right) \sin \theta + 2a \frac{\partial \varphi}{\partial T_1} \cos \theta - \frac{a^3}{4} \sin 3\theta \quad (124)$$

and to suppress the resonance that would generate secular terms we put

$$\frac{\partial a}{\partial T_1} = -\frac{3}{8} a^3 \quad \text{and} \quad \frac{\partial \varphi}{\partial T_1} = 0. \quad (125)$$

Then  $y_1 = \frac{a^3}{32} \sin 3\theta$  solves this equation and has  $y_1(0) = 0$ , which does not disturb the initial condition  $y_0(0) = a_0$ , although since  $dy_1/dT_0 = 3a^2/32 \cos 3\theta$  the derivative of  $y_0 + \varepsilon y_1$  will differ by  $O(\varepsilon)$  from zero at  $T_0 = 0$ . This does not matter and we may adjust this by choice of initial conditions for  $\varphi$ , later.

The  $O(\varepsilon^2)$  term is somewhat finicky, being

$$\begin{aligned} \frac{\partial^2 y_2}{\partial T_0^2} + y_2 &= -2 \frac{\partial^2 y_0}{\partial T_0 \partial T_2} - 2 \frac{\partial^2 y_1}{\partial T_0 \partial T_1} \\ &\quad - 3 \left( \frac{\partial y_0}{\partial T_0} \right)^2 \left( \frac{\partial y_0}{\partial T_1} + \frac{\partial y_1}{\partial T_0} \right) - \frac{\partial^2 y_0}{\partial T_1^2} - 3 \frac{\partial y_0}{\partial T_0} \end{aligned} \quad (126)$$

where the last term came from  $3(\dot{y})\varepsilon^2$ . Proceeding as before, and using  $\partial\varphi/\partial T_1 = 0$  and  $\partial a/\partial T_1 = -3/8 a^3$  as well as some other trigonometric identities, we find the right-hand side can be written as

$$\left( 2 \frac{\partial a}{\partial T_2} + 3a \right) \sin \theta + \left( 2a \frac{\partial \varphi}{\partial T_2} - \frac{9}{128} a^5 \right) \cos \theta - \frac{27}{1024} a^5 \cos 3\theta + \frac{9}{128} a^5 \cos 5\theta. \quad (127)$$



Again setting the coefficients of  $\sin \theta$  and  $\cos \theta$  to zero to prevent resonance we have

$$\frac{\partial a}{\partial T_2} = -\frac{3}{2}a \quad (128)$$

and

$$\frac{\partial \varphi}{\partial T_2} = \frac{9}{256}a^4 \quad (a \neq 0). \quad (129)$$

This leaves

$$y_2 = \frac{27}{1024}a^5 \cos 3\theta - \frac{3a^5}{1024} \cos 5\theta \quad (130)$$

again setting the homogeneous part to zero.

Now comes a bit of multiple scales magic: instead of solving equations (125) and (128) in sequence, as would be usual, we write

$$\begin{aligned} \frac{da}{dt} &= \frac{\partial a}{\partial T_0} + \varepsilon \frac{\partial a}{\partial T_1} + \varepsilon^2 \frac{\partial a}{\partial T_2} = 0 + \varepsilon \left( -\frac{3}{8}a^3 \right) + \varepsilon^2 \left( -\frac{3}{2}a \right) \\ &= -\frac{3}{8}\varepsilon a(a^2 + 4\varepsilon). \end{aligned} \quad (131)$$

Using  $a = 2R$  this is equation (6.50) in [28]. Similarly

$$\frac{d\varphi}{dt} = \varepsilon \frac{\partial \varphi}{\partial T_1} + \varepsilon^2 \frac{\partial \varphi}{\partial T_2} = 0 + \varepsilon^2 \frac{9}{256}a^4 \quad (132)$$

and once  $a$  has been identified,  $\varphi$  can be found by quadrature. Solving (131) and (132) by Maple,

$$a = \frac{\sqrt{\varepsilon}a_0}{\sqrt{\varepsilon e^{3\varepsilon^2 t} + \frac{a_0^2}{4}(e^{3\varepsilon^2 t} - 1)}} = 2 \frac{\sqrt{\varepsilon}a_0}{\sqrt{u}} \quad (133)$$

and

$$\varphi = -\frac{3}{16}\varepsilon^2 \ln u + \frac{9}{16}\varepsilon^4 t - \frac{3}{16} \frac{\varepsilon^2 a_0^2}{u} \quad (134)$$

where  $u = 4\varepsilon e^{3\varepsilon^2 t} + a_0^2(e^{3\varepsilon^2 t} - 1)$ . The residual is (again by Maple)

$$\varepsilon^3 \left( \frac{9}{16}a_0^3 \cos 3t + a_0^7 \left( -\frac{351}{4096} \sin t - \frac{9}{512} \sin 7t + \frac{333}{4096} \sin 3t + \frac{459}{4096} \sin 5t \right) \right) + O(\varepsilon^4) \quad (135)$$

and there is no secularity visible in this term.

It is important to note that the construction of the equation (131) for  $a(t)$  required both  $\partial a/\partial T_1$  and  $\partial a/\partial T_2$ . Either one alone gives misleading or inconsistent answers. While it may be obvious to an expert that both terms must be used at once, the situation is somewhat unusual and a novice or casual user of perturbation methods may well wish reassurance. (We did!) Computing (and plotting) the residual  $\Delta = \ddot{z} + z + \varepsilon(\dot{z})^3 + 3\varepsilon^2 \dot{z}$  does just that (see figure 3). It is simple to verify that, say, for  $\varepsilon = 1/100$ ,  $|\Delta| < \varepsilon^3 a$  on  $0 < t < 10^5 \pi$ . Notice that  $a \sim O(e^{-3/2\varepsilon^2 t})$  and  $e^{-3/2 \cdot 10^{-4} \cdot 10^5 \cdot \pi} = e^{-15\pi} \doteq 10^{-15}$  by the end of this range. The method of multiple scales has thus produced  $z$ , the exact solution of an equation uniformly and relatively near to the original equation. In trigonometric form,

$$\begin{aligned} z &= a \cos(t + \varphi) + \varepsilon \frac{a^3}{32} \cos(3(t + \varphi)) \\ &\quad + \varepsilon^2 \left( \frac{27}{1024}a^5 \cos(3(t + \varphi)) - \frac{3}{1024}a^5 \cos^5((5(t + \varphi))) \right) \end{aligned} \quad (136)$$

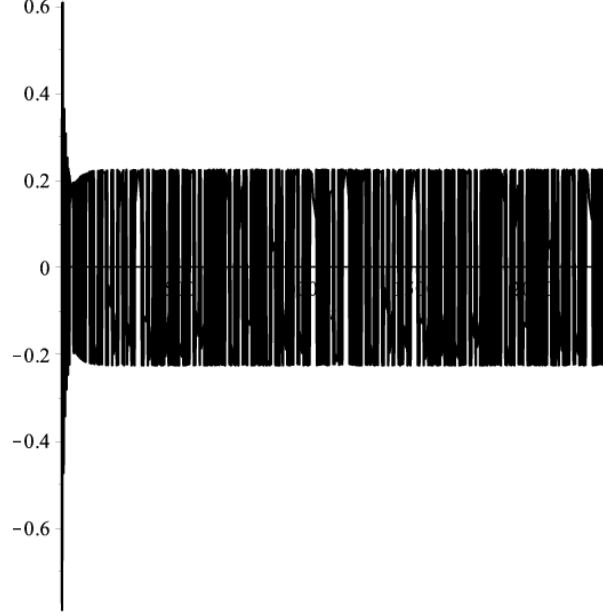


Figure 3: The residual  $|\Delta_3|$  divided by  $\varepsilon^3 a$ , with  $\varepsilon = 0.1$ , where  $a = O(e^{-3/2 \varepsilon^2 t})$ , on  $0 \leq t \leq 10 \ln(10)/\varepsilon^2$  (at which point  $a = 10^{-15}$ ). We see that  $|\Delta_3/\varepsilon^3 a| < 1$  on this entire interval.

and  $a$  and  $\varphi$  are as in equations (131) and (132). Note that  $\varphi$  asymptotically approaches zero. Note that the trigonometric solution we have demonstrated here to be correct, which was derived for us by our colleague Pei Yu, appears to differ from that given in [28], which is

$$y = Ae^{it} + \varepsilon B e^{3it} + \varepsilon^2 C e^{5it} + \dots \quad (137)$$

where (with  $\tau = \varepsilon t$ )

$$C \sim \frac{3}{64} A^5 + \dots \quad \text{and} \quad B \sim -\frac{A^3}{8} \left( i + \frac{45}{8} \varepsilon |A|^2 + \dots \right) \quad (138)$$

and, if  $A = R e^{i\varphi}$ ,

$$\frac{dR}{d\tau} = -\frac{3}{2} (R^3 + \varepsilon R + \dots) \quad \text{and} \quad \frac{d\varphi}{d\tau} = -\frac{3}{2} R^2 \left( 1 + \frac{3\varepsilon}{8} R^2 + \dots \right) \quad (139)$$

Of course with the trigonometric form  $y = a \cos(t + \varphi)$ , the equivalent complex form is

$$y = a \left( \frac{e^{it+i\varphi} + e^{-it-i\varphi}}{2} \right) = \frac{a}{2} e^{i\varphi} e^{it} + c.c. \quad (140)$$

and so  $R = a/2$ . As expected, equation (6.50) in [28] becomes

$$\frac{da}{d\tau} \left( \frac{a}{2} \right) = -\frac{3}{2} \frac{a}{2} \left( \frac{a^2}{4} + \varepsilon \right) \quad (141)$$

or, alternatively,

$$\frac{da}{d\tau} = -\frac{3}{8} \varepsilon a (a^2 + 4\varepsilon) \quad (142)$$

which agrees with that computed for us by Pei Yu. However, O'Malley's equation (6.48) gives

$$C \cdot e^{i \cdot 5t} = \frac{3}{64} A^5 e^{i5t} = \frac{3}{64} R^5 e^{i5\theta} = \frac{3}{2048} a^5 e^{i5\theta}, \quad (143)$$

so that

$$C e^{i5t} + c.c. = \frac{3}{1024} a^5 \cos 5\theta, \quad (144)$$

whereas Pei Yu has  $-3/1024$ . As demonstrated by the residual in figure 3, Pei Yu is correct. Well, sign errors are trivial enough.

More differences occur for  $B$ , however. The  $-A^3/s i e^{3it}$  term becomes  $a^3/32 \cos 3\theta$ , as expected, but  $-45/64 A^3 \cdot |A|^2 e^{3it} + c.c.$  becomes  $-45/32 a^5/32 \cos 3\theta = -45/1024 a^5 \cos 3\theta$ , not  $27/1024 a^5 \cos 3\theta$ . Thus we believe there has been an arithmetic error in [28]. This is also present in [29]. Similarly, we believe the  $d\varphi/dt$  equation there is wrong.

Arithmetic errors in perturbation solutions are, obviously, a constant hazard even for experts. We do not point out this error (or the other errors highlighted in this paper) in a spirit of glee—goodness knows we've made our own share. No, the reason we do so is to emphasize the value of a separate, independent check using the residual. Because we have done so here, we are certain that equation (136) is correct: it produces a residual that is uniformly  $O(\varepsilon^3)$  for bounded time, and which is  $O(\varepsilon^{9/2} e^{-3/2 \varepsilon^2 t})$  as  $t \rightarrow \infty$ . (We do not know why there is extra accuracy for large times).

Finally, we remark that the difficulty this example presents for the method of multiple scales is that equation (131) cannot be solved itself by perturbation methods (or, at least, we couldn't do it). One has to use all three terms at once; the fact that this works is amply demonstrated afterwards. Indeed the whole multiple scales procedure based on equation (116) is really very strange when you think about it, but it can be justified afterwards. It really doesn't matter how we find equation (136). Once we have done so, verifying that it is the exact solution of a small perturbation of the original equation is quite straightforward. The implementation is described in the following Maple code:

```
restart;
r := epsilon;
de := u -> (diff(u, t, t)+u+r*(diff(u, t))^3+3*r^2*(diff(u, t)));
U := -a0^2+4*exp(3*epsilon^2*t)*epsilon+exp(3*epsilon^2*t)*a0^2;
a := 2*sqrt(r)*a0/sqrt(U);
phi := -(3/16)*epsilon^2*ln(U)+(9/16)*epsilon^4*t-(3/16)*epsilon^2*a0^2/U;
z := a*cos(t+phi)+(1/32)*r*a^3*sin(3*t+3*phi)
+r^2*((27/1024)*a^5*cos(3*(t+phi))-(3/1024)*a^5*cos(5*(t+phi)));
resid := de(z);
zer := MultiSeries[series](resid, r, 4);
map(combine, zer, trig);
eps := 1/10;
plot(eval(resid/(a*r^3), [a0 = 1.0, r = eps]), t = 0 .. 10*ln(10)/eps^2,
colour = BLACK);
```

### 5.3 The lengthening pendulum

As an interesting example with a genuine secular term, [4] discuss the lengthening pendulum. There, Boas solves the linearized equation exactly in terms of Bessel functions. We use the model here as an example of a perturbation solution in a physical context. The original Lagrangian leads to

$$\frac{d}{dt} \left( m\ell^2 \frac{d\theta}{dt} \right) + mg\ell \sin \theta = 0 \quad (145)$$

(having already neglected any system damping). The length of the pendulum at time  $t$  is modelled as  $\ell = \ell_0 + vt$ , and implicitly  $v$  is small compared to the oscillatory speed  $d\theta/dt$  (else why would it be a pendulum at all?). The presence of  $\sin \theta$  makes this a nonlinear problem; when  $v = 0$  there is an analytic solution using elliptic functions [24, chap. 4].

We *could* do a perturbation solution about that analytic solution; indeed there is computer algebra code to do so automatically [30]. For the purpose of this illustration, however, we make the same small-amplitude linearization that Boas did and replace  $\sin \theta$  by  $\theta$ . Dividing the resulting equation by  $\ell_0$ , putting  $\varepsilon = v/\ell_0\omega$  with  $\omega = \sqrt{g/\ell_0}$  and rescaling time to  $\tau = \omega t$ , we get

$$(1 + \varepsilon\tau) \frac{d^2\theta}{d\tau^2} + 2\varepsilon \frac{d\theta}{d\tau} + \theta = 0. \quad (146)$$

This supposes, of course, that the pin holding the base of the pendulum is held perfectly still (and is frictionless besides).

Computing a regular perturbation approximation

$$z_{\text{reg}} = \sum_{k=0}^N \theta_k(\tau) \varepsilon^k \quad (147)$$

is straightforward, for any reasonable  $N$ , by using computer algebra. For instance, with  $N = 1$  we have

$$z_{\text{reg}} = \cos \tau + \varepsilon \left( \frac{3}{4} \sin \tau + \frac{\tau^2}{4} \sin \tau - \frac{3}{4} \tau \cos \tau \right). \quad (148)$$

This has residual

$$\Delta_{\text{reg}} = (1 + \varepsilon\tau) z_{\text{reg}}'' + 2\varepsilon z_{\text{reg}}' + z_{\text{reg}} \quad (149)$$

$$= -\frac{\varepsilon^2}{4} (\tau^3 \sin \tau - 9\tau^2 \cos \tau - 15\tau \sin \tau) \quad (150)$$

also computed straightforwardly with computer algebra. By experiment with various  $N$  we find that the residuals are always of  $O(\varepsilon^{N+1})$  but contain powers of  $\tau$ , as high as  $\tau^{2N+1}$ . This naturally raises the question of just when this can be considered “small.” We thus have the *exact* solution of

$$(1 + \varepsilon\tau) \frac{d^2\theta}{d\tau^2} + 2\varepsilon \frac{d\theta}{d\tau} + \theta = \Delta_{\text{reg}}(\tau) = P(\varepsilon^{N+1} \tau^{2N+1}) \quad (151)$$

and it seems clear that if  $\varepsilon^{N+1} \tau^{2N+1}$  is to be considered small it should at least be smaller than  $\varepsilon\tau$ , which appear on the left hand side of the equation. [ $d^2/d\tau^2$  is  $-\cos \tau$  to leading order, so this is periodically  $O(1)$ .] This means  $\varepsilon^N \tau^{2N}$  should be smaller than 1, which forces  $\tau \leq T$  where  $T = O(\varepsilon^{-q})$  with  $q < \frac{1}{2}$ . That is, this regular perturbation solution is valid only on a limited range of  $\tau$ , namely,  $\tau = O(\varepsilon^{-1/2})$ .

Of course, the original equation contains a term  $\varepsilon\tau$ , and this itself is small only if  $\tau \leq T_{\text{max}}$  with  $T_{\text{max}} = O(\varepsilon^{-1+\delta})$  for  $\delta > 0$ . Notice that we have discovered this limitation of the regular perturbation solution without reference to the ‘exact’ Bessel function solution of this linearized equation. Notice also that  $\Delta_{\text{reg}}$  can be interpreted as a small forcing term; a vibration of the pin holding the pendulum, say. Knowing that, say, such physical vibrations, perhaps caused by trucks driving past the laboratory holding the pendulum, are bounded in size by a certain amount, can help to decide what  $N$  to take, and over which  $\tau$ -interval the resulting solution is valid.

Of course, one might be interested in the forward error  $\theta - z_{\text{reg}}$ ; but then one should be interested in the forward errors caused by neglecting physical vibrations (e.g. of trucks passing by) and the same theory—what a numerical analyst calls a condition number—can be used for both.

But before we pursue that farther, let us first try to improve the perturbation solution. The method of multiple scales, or equivalent but easier in this case the renormalization group method [22] which consists for a linear problem of taking the regular perturbation solution and replacing  $\cos \tau$  by  $(e^{i\tau} + e^{-i\tau})/2$  and  $\sin \tau$  by  $(e^{i\tau} - e^{-i\tau})/2i$ , gathering up the result and writing it as  $1/2 A(\tau; \varepsilon)e^{i\tau} + 1/2 \bar{A}(\tau; \varepsilon)e^{-i\tau}$ . One then writes  $A(\tau; \varepsilon) = e^{L(\tau; \varepsilon)} + O(\varepsilon^{N+1})$  (that is, taking the logarithm of the  $\varepsilon$ -series for  $A(\tau; \varepsilon) = A_0(\tau) + \varepsilon A_1(\tau) + \dots + \varepsilon^N A_N(\tau) + O(\varepsilon^{N+1})$ , a straightforward exercise (especially in a computer algebra system) and then (if one likes) rewriting  $1/2 e^{L(\tau; \varepsilon) + i\tau} + \text{c.c.}$  in real trigonometric form again, gives an excellent result. If  $N = 1$ , we get

$$\tilde{z}_{\text{renorm}} = e^{-3/4 \varepsilon \tau} \cos \left( \frac{3}{4} \varepsilon + \tau - \varepsilon \frac{\tau^2}{4} \right) \quad (152)$$

which contains an irrelevant phase change  $\frac{3}{4}\varepsilon$  which we remove here as a distraction to get

$$z_{\text{renorm}} = e^{-3/4 \varepsilon \tau} \cos \left( \tau - \varepsilon \frac{\tau^2}{4} \right). \quad (153)$$

This has residual:

$$\begin{aligned} \Delta_{\text{renorm}} &= (1 + \varepsilon \tau) \frac{d^2 z_{\text{renorm}}}{d\tau^2} + 2\varepsilon \frac{dz_{\text{renorm}}}{d\tau} + z_{\text{renorm}} \\ &= \varepsilon^2 e^{-\frac{3}{4}\varepsilon\tau} \left( \left( \frac{3}{4}\tau^2 - \frac{15}{16} \right) \cos\left(\tau - \varepsilon \frac{\tau^2}{4}\right) - \frac{9}{4}\tau \sin\left(\tau - \varepsilon \frac{\tau^2}{4}\right) \right) + O(\varepsilon^3 \tau^3 e^{-\frac{3}{4}\varepsilon\tau}). \end{aligned} \quad (154)$$

By inspection, we see that this is superior in several ways to the residual from the regular perturbation method. First, it contains the damping term  $e^{-3/4 \varepsilon \tau}$  just as the computed solution does; this residual will be small compared even to the decaying solution. Second, at order  $N$  it contains only  $\tau^{N+1}$  as its highest power of  $\varepsilon$ , not  $\tau^{2N+1}$ . This will be small compared to  $\varepsilon \tau$  for times  $\tau < T$  with  $T = O(\varepsilon^{-1+\delta})$  for *any*  $\delta > 0$ ; that is, this perturbation solution will provide a good solution so long as its fundamental assumption, that the  $\varepsilon \tau$  term in the original equation, can be considered ‘small’, is good.

Note that again the quality of this perturbation solution has been judged without reference to the exact solution, and quite independently of whatever assumptions are usually made to argue for multiple scales solutions (such as boundedness of  $\theta$ ) or the renormalization group method. Thus, we conclude that the renormalization group method gives a superior solution in this case, and this judgement was made possible by computing the residual. We have used the following Maple implementation:

```
restart;
macro(e = epsilon);
de := y -> (1+e*t)*(diff(y, t, t))+2*e*(diff(y, t))+y;
z := cos(t);
N := 1;
Order := N+1;
for i to N do
    zt := z+e^i*y[i](t);
    res := series(de(zt), e, i+1);
    eqs := coeff(res, e, i);
    yi := dsolve({eqs, y[i](0) = 0, (D(y[i]))(0) = 0}, y[i](t));
    z := eval(zt, yi);
end do;
res := de(z);
expform := convert(z, exp);
```

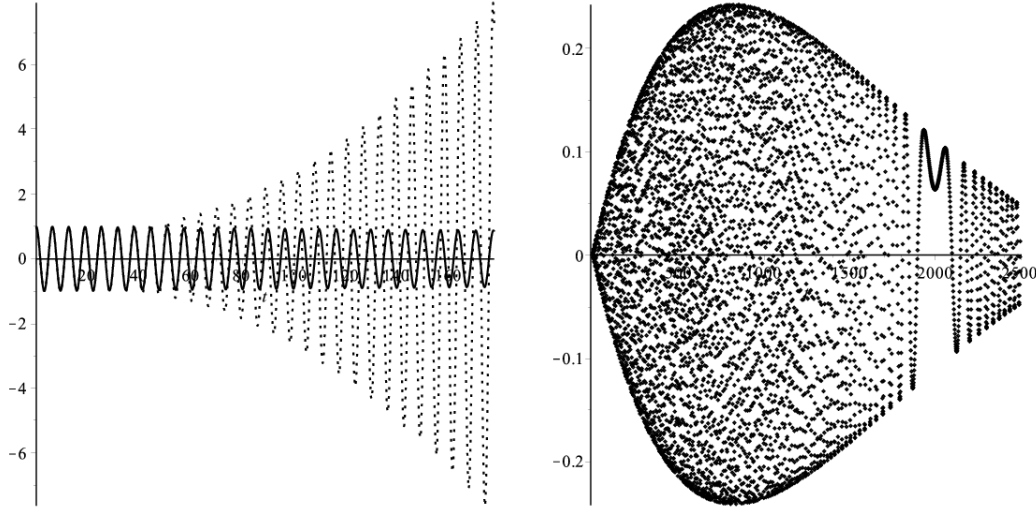


Figure 4: On the left, solutions to the lengthening pendulum equation (the renormalized solution is the solid line). On the right, residual of the renormalized solution, which is orders of magnitudes smaller than that of the regular expansion.

```

expform := collect(expform, [exp(I*t), exp(-I*t)], factor);
zp := coeff(expform, exp(I*t));
lg := convert(series(ln(series(zp+O(e^Order), e)), e), polynomial);
lg := collect(lg, e, factor);
zrg := exp(lg)*exp(I*t);
zrg := zrg+evalc(conjugate(zrg));
zrg := combine(evalc(zrg), trig);
zrg := simplify(zrg);
zrg := exp(-(3/4)*epsilon*t)*cos(t-(1/4)*epsilon*t^2);
resrg := collect(de(zrg), e, t -> combine(simplify(t), trig));
tiny := 1/1000;
plot(eval([z, zrg], e = tiny), t = 0 .. 1/tiny^(3/4), colour = BLACK, linestyle = [2, 1]);
plot(eval(res, e = tiny), t = 1 .. 2500, colour = BLACK, linestyle = 2);
plot(eval(resrg/(tiny*t), e = tiny), t = 1 .. 2500, colour = BLACK,
style = POINT, numpoints=2016, symbolsize=1);

```

See figure 4.

Note that this renormalized residual contains terms of the form  $(\varepsilon\tau)^k e^{-3/4\varepsilon\tau} i$ . No matter what order we compute to, these have maxima  $O(1)$  when  $\tau = O(1/\varepsilon)$ , but as noted previously the fundamental assumption of perturbation has been violated by that large a  $\tau$ .

**Optimal backward error again** Now, one further refinement is possible. We may look for an  $O(\varepsilon^2)$  perturbation of the lengthening of the pendulum, that explains part of this computed residual! That is, we look for  $p(t)$ , say, so that

$$\Delta_2 := (1 + \varepsilon\tau + \varepsilon p(\tau))z''_{\text{renorm}} + 2(\varepsilon + \varepsilon^2 p'(\tau))z'_{\text{renorm}} + z_{\text{renorm}} \quad (155)$$

has only *smaller* terms in it than  $\Delta_{\text{renorm}}$ . Note the correlated changes,  $\varepsilon^2 p(\tau)$  and  $\varepsilon^2 p'(\tau)$ .

At this point, we don't know if this is possible or useful, but it's a good thing to try. In numerical analysis terms, we are trying to find a structured backward error for this computed solution.

The procedure for identifying  $p(\tau)$  in equation (155) is straightforward. We put  $p(\tau) = a_0 + a_1\tau + a_2\tau^2$  with unknown coefficients, compute  $\Delta_2$ , and try to choose  $a_0$ ,  $a_1$ , and  $a_2$  in order to make as many coefficients of powers of  $\varepsilon$  in  $\Delta_2$  to be zero as we can. When we do this, we find that

$$p = -\frac{15}{16} + \frac{3}{4}\tau^2 \quad (156)$$

makes

$$\Delta_{\text{mod}} = \left(1 + \varepsilon\tau + \varepsilon^2 \left(\frac{3}{4}\tau^2 - \frac{15}{16}\right)\right) z''_{\text{renorm}} + 2 \left(\varepsilon + \varepsilon^2 \left(\frac{3}{2}\tau\right)\right) z'_{\text{renorm}} + z_{\text{renorm}} \quad (157)$$

$$= \varepsilon^2 e^{-3/4 \varepsilon\tau} \left(-\frac{3}{4}\tau \sin(\tau - 1/4 \varepsilon\tau^2)\right) + O(\varepsilon^3 \tau^3 e^{-3\varepsilon\tau/4}). \quad (158)$$

This is  $O(\varepsilon^2 \tau e^{-3\varepsilon\tau/4})$  instead of  $O(\varepsilon^2 \tau^2 e^{-3\varepsilon\tau/4})$ , and therefore smaller. This *interprets* the largest term of the original residual, the  $O(\varepsilon^2 \tau^2)$  term, as a perturbation in the lengthening of the pendulum. The gain is one of interpretation; the solution is the same, but the equation it solves exactly is slightly different. For  $O(\varepsilon^N \tau^N)$  solutions the modifications will probably be similar. Now, if  $z \doteq \cos \tau$  then  $z' \doteq -\sin \tau$ ; so if we include a damping term

$$\left(+\varepsilon^2 \cdot \frac{3}{8} \cdot \tau \theta'\right) \quad (159)$$

in the model, we have

$$\begin{aligned} \left(1 + \varepsilon\tau + \varepsilon^2 \left(\frac{3}{4}\tau^2 - \frac{15}{16}\right)\right) z''_{\text{renorm}} + 2 \left(\varepsilon - \varepsilon^2 \left(\frac{3}{2}\tau\right) + \varepsilon^2 \frac{3}{8}\tau\right) z'_{\text{renorm}} + z_{\text{renorm}} \\ = O\left(\varepsilon^3 \tau^3 e^{-3/4 \varepsilon\tau}\right) \end{aligned} \quad (160)$$

and *all* of the leading terms of the residual have been “explained” in the physical context. If the damping term had been negative, we might have rejected it; having it increase with time also isn’t very physical (although one might imagine heating effects or some such).

## 5.4 Vanishing lag delay DE

For another example we consider an expansion that “everybody knows” can be problematic. We take the DDE

$$\dot{y}(t) + ay(t - \varepsilon) + by(t) = 0 \quad (161)$$

from [2, p. 52] as a simple instance. Expanding  $y(t - \varepsilon) = y(t) - \dot{y}(t)\varepsilon + O(\varepsilon^2)$  we get

$$(1 - a\varepsilon)\dot{y}(t) + (b + a)y(t) = 0 \quad (162)$$

by ignoring  $O(\varepsilon^2)$  terms, with solution

$$z(t) = \exp\left(-\frac{b+a}{1-a\varepsilon}t\right)u_0 \quad (163)$$

if a simple initial condition  $u(0) = u_0$  is given. Direct computation of the residual shows

$$\Delta = \dot{z} + az(t - \varepsilon) + bz(t) \quad (164)$$

$$= O(\varepsilon^2)z(t) \quad (165)$$

uniformly for all  $t$ ; in other words, our computed solution  $z(t)$  exactly solves

$$\dot{y} + ay(t - \varepsilon) + (b + O(\varepsilon^2))y(t) = 0 \quad (166)$$

which is an equation of the same type as the original, with only  $O(\varepsilon^2)$  perturbed coefficients. The initial history for the DDE should be prescribed on  $-\varepsilon \leq t < 0$  as well as the initial condition, and that's an issue, but often that history is an issue anyway. So, in this case, contrary to the usual vague folklore that Taylor series expansion in the vanishing lag “can lead to difficulties”, we have a successful solution and we know that it's successful.

We now need to assess the sensitivity of the problem to small changes in  $b$ , but we all know that has to be done anyway, even if we often ignore it.

Another example of Bellman's on the same page,  $\ddot{y}(t) + ay(t - \varepsilon) = 0$ , can be treated in the same manner. Bellman cautions there that seemingly similar approaches can lead to singular perturbation problems, which can indeed lead to difficulties, but even there a residual/backward error analysis can help to navigate those difficulties.

## 5.5 Artificial viscosity in a nonlinear wave equation

Suppose we are trying to understand a particular numerical solution, by the method of lines, of

$$u_t + uu_x = 0 \quad (167)$$

with initial condition  $u(0, x) = e^{i\pi x}$  on  $-1 \leq x \leq 1$  and periodic boundary conditions. Suppose that we use the method of modified equations (see, for example, [18], [33], or [9, chap 12]) to find a perturbed equation that the numerical solution more nearly solves. Suppose also that we analyze the same numerical method applied to the divergence form

$$u_t + \frac{1}{2}(u^2)_x = 0. \quad (168)$$

Finally, suppose that the method in question uses backward differences  $f'(x) = (f(x) - f(x - 2\varepsilon))/2\varepsilon$  (the factor 2 is for convenience) on an equally-spaced  $x$ -grid, so  $\Delta x = -2\varepsilon$ . The method of modified equations gives

$$u_t + uu_x - \varepsilon(uu_{xx}) + O(\varepsilon^2) = 0 \quad (169)$$

for equation (167) and

$$u_t + uu_x - \varepsilon(u_x^2 + uu_{xx}) + O(\varepsilon^2) = 0 \quad (170)$$

for equation (168).

The outer solution to each of these equations is just the reference solution to both equations (167) and (168), namely,

$$u = \frac{1}{i\pi t} W(i\pi t e^{i\pi x}) \quad (171)$$

where  $W(z)$  is the principal branch of the Lambert  $W$  function, which satisfies  $W(z)e^{W(z)} = z$ . See [10] for more on the Lambert  $W$  function. That  $u$  is the solution for this initial condition was first noticed by [34]. The residuals of these outer solutions are just  $-\varepsilon uu_{xx}$  and  $-\varepsilon(u_x^2 + uu_{xx})$  respectively. Simplifying, and again suppressing the argument of  $W$  for tidiness, we find that

$$-\varepsilon uu_{xx} = -\frac{\varepsilon W^2}{t^2(1 + W^3)} \quad (172)$$



and

$$-\varepsilon(u_x^2 + uu_{xx}) = -\frac{\varepsilon W^2(2+W)}{t^2(1+W^3)} \quad (173)$$

where  $W$  is short for  $W(i\pi t e^{i\pi x})$ . We see that if  $x = 1/2$  and  $t = 1/(\pi e)$ , both of these are singular:

$$-\varepsilon uu_{xx} \sim -\varepsilon \left( \frac{i\pi^2 e^2 \sqrt{2}}{4(et\pi - 1)^{3/2}} + O\left(\frac{1}{et\pi - 1}\right) \right) \quad (174)$$

and

$$-\varepsilon(u_x^2 + uu_{xx}) \sim -\varepsilon \left( \frac{i\pi^2 e^2 \sqrt{2}}{4(et\pi - 1)^{3/2}} + O\left(\frac{1}{\sqrt{et\pi - 1}}\right) \right). \quad (175)$$

We see that the outer solution makes the residual very large near  $x = 1/2$  as  $t \rightarrow 1/(\pi e)^-$  suggesting that the solution of the modified equation—and thus the numerical solution—will depart from the outer solution. Both the original form and the divergence form are predicted to have similar behaviour, and this is confirmed by numerical experiments.

We remark that using forward differences instead just changes the sign of  $\varepsilon$ , and given the similarity of  $\varepsilon uu_{xx}$  to  $\varepsilon u_{xx}$ , we intuit that this will blow up rather quickly, like the backward heat equation, because the exact solution to Burger’s equation  $u_t + uu_x = \varepsilon u_{xx}$  involves a change in variable to the heat equation [21, pp. 352-353]. We also remark also that this use of residual is a bit perverse: we here substitute the reference solution into an approximate (reverse-engineered) equation. Some authors do use ‘residual’ or even ‘defect’ in this sense., e.g., [6]. It only fits our usage because the reference solution to the original equation is just the outer solution of the perturbation problem of interest here.

Finally, we can interpolate the numerical solution using a trigonometric interpolant in  $x$  tensor producted with the interpolant in  $t$  provided by the numerical solver (e.g., `ode15s` in Matlab). We can then compute the residual  $\Delta(t, x) = z_t + zz_x$  in the original equation and we find that, away from the singularity, it is  $O(\varepsilon)$ . If we compute the residual in the modified equation

$$\Delta_1(t, x) = z_t + zz_x - \varepsilon zz_{xx} \quad (176)$$

we find that, away from the singularity, it is  $O(\varepsilon^2)$ . This is a more traditional use of residual in a numerical computation, and is done without knowledge of any reference solution. The analogous use we are making for perturbation methods can be understood from this numerical perspective.

## 6 Concluding Remarks

Decades ago, van Dyke had already made the point that, in perturbation theory, “[t]he possibilities are too diverse to be subject to rules” [32, p. 31]. Van Dyke was talking about the useful freedom to choose expansion variables artfully, but the same might be said for perturbation methods generally. This paper has attempted (in the face of that observation) to lift a known technique, namely the residual as a backward error, out of numerical analysis and apply it to perturbation theory. The approach is surprisingly useful and clarifies several issues, namely

- BEA allows one to directly use approximations taken from divergent series in an optimal fashion without appealing to “rules of thumb” such as stopping before including the smallest term.
- BEA allows the justification of removing spurious secular terms, even when true secular terms are present.

- Not least, residual computation and *a posteriori* BEA makes detection of slips, blunders, and bugs all but certain, as illustrated in our examples.
- Finally BEA interprets the computed solution  $z$  as the exact solution to just as good a model.

In this paper we have used BEA to demonstrate the validity of solutions obtained by the iterative method, by Lindstedt’s method, by the method of multiple scales, by the renormalization group method, and by matched asymptotic expansions. We have also successfully used the residual and BEA in many problems not shown here: eigenvalue problems from [26]; an example from [32] using the method of strained coordinates; and many more.

The examples here have largely been for algebraic equations and for ODEs, but the method was used to good effect in [39] for a PDE system describing heat transfer between concentric cylinders, with a high-order perturbation series in Rayleigh number. Aside from the amount of computational work required, there is no theoretical obstacle to using the technique for other PDE; indeed the residual of a computed solution  $z$  (perturbation solution, in this paper) to an operator equation  $\varphi(y; x) = 0$  is usually computable:  $\Delta = \varphi(z; x)$  and its size (in our case, leading term in the expansion in the gauge functions) easily assessed.

It’s remarkable to us that the notion, while present here and there in the literature, is not used more to justify the validity of the perturbation series.

We end with a caution. Of course, BEA is not a panacea. There are problems for which it is not possible. For instance, there may be hidden constraints, something like solvability conditions, that play a crucial role and where the residual tells you nothing. A residual can even be zero and if there are multiple solutions, one needs a way to get the right one. There are things that can go wrong with this backward error approach. First, the final residual computation might not be independent enough from the computation of  $z$ , and repeat the same error. An example is if one correctly solves

$$\ddot{y} + y + \varepsilon \dot{y}^3 + 3\varepsilon^2 \dot{y} = 0 \tag{177}$$

and verifies that the residual is small, while *intending* to solve

$$\ddot{y} + y + \varepsilon \dot{y}^3 - 3\varepsilon^2 \dot{y} = 0, \tag{178}$$

i.e., getting the wrong sign on the  $\dot{y}$  term, both times. Another thing that can go wrong is to have an error in your independent check but not your solution. This happened to us with 183 instead of 138 in subsection 3.1.3; the discrepancy alerted us that there was a problem, so this at least was noticeable. A third thing that can go wrong is that you verify the residual is small but forget to check the boundary conditions. A fourth thing that can go wrong is that the residual may be small in an absolute sense but still larger than important terms in the equation—the residual may need to be smaller than you expect, in order to get good qualitative results. A fifth thing is that the residual may be small but of the ‘wrong character’, i.e., be unphysical. Perhaps the method has introduced the equivalent of negative damping, for instance. This point can be very subtle.

A final point is that a good solution needs not just a small backward error, but also information about the sensitivity (or robustness) of the model to physical perturbations. We have not discussed computation of sensitivity, but we emphasize that even if  $\Delta \equiv 0$ , you still have to do it, because real situations have real perturbations. Nonetheless, we hope that we have convinced you that BEA can be helpful.

## References

- [1] Konstantin E. Avrachenkov, Jerzy A. Filar, and Phil G. Howlett. *Analytic perturbation theory and its applications*. SIAM, 2013.

- [2] Richard E. Bellman. *Perturbation techniques in mathematics, physics, and engineering*. Dover Publications, 1972.
- [3] C.M. Bender and S.A. Orszag. *Advanced mathematical methods for scientists and engineers: Asymptotic methods and perturbation theory*, volume 1. Springer Verlag, 1978.
- [4] Mary L. Boas. *Mathematical Methods in the Physical Sciences*. John Wiley, New York, 1966.
- [5] John P. Boyd. *Solving Transcendental Equations*. SIAM, 2014.
- [6] Hayato Chiba. Extension and unification of singular perturbation methods for odes based on the renormalization group method. *SIAM Journal on Applied Dynamical Systems*, 8(3):1066–1115, 2009.
- [7] Robert M. Corless. What is a solution of an ODE? *ACM SIGSAM Bulletin*, 27(4):15–19, 1993.
- [8] Robert M. Corless and George F. Corliss. Rationale for guaranteed ODE defect control. In L. Atanassova and J. Herzberger, editors, *Computer Arithmetic and Enclosure Methods*, pages 3–12. North-Holland, 1992.
- [9] Robert M. Corless and N. Fillion. *A Graduate Introduction to Numerical Methods, From the Viewpoint of Backward Error Analysis*. Springer, New York, 2013. 868pp.
- [10] Robert M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and Donald E. Knuth. On the Lambert  $W$  function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.
- [11] Nicolaas Govert De Bruijn. *Asymptotic methods in analysis*, volume 4. Dover, 1970.
- [12] P. Deuffhard and A. Hohmann. *Numerical analysis in modern scientific computing: an introduction*, volume 43. Springer Verlag, 2003.
- [13] NIST Digital Library of Mathematical Functions. <http://dlmf.nist.gov/>, Release 1.0.10 of 2015-08-07.
- [14] Wayne H. Enright. Analysis of error control strategies for continuous runge-kutta methods. *SIAM Journal on Numerical Analysis*, 26(3):588–599, 1989.
- [15] Wayne H. Enright. A new error-control for initial value solvers. *Applied Mathematics and Computation*, 31:288–301, 1989.
- [16] K. O. Geddes, S. R. Czapor, and G. Labahn. *Algorithms for computer algebra*. Kluwer Academic, Boston, 1992.
- [17] J.F. Grcar. John von Neumann’s analysis of Gaussian elimination and the origins of modern numerical analysis. *SIAM review*, 53(4):607–682, 2011.
- [18] D.F. Griffiths and J.-M. Sanz-Serna. On the scope of the method of modified equations. *SIAM Journal on Scientific and Statistical Computing*, 7(3):994–1008, 1986.
- [19] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 2nd edition, 2002.
- [20] M.H. Holmes. *Introduction to perturbation methods*. Springer, 1995.
- [21] Jirair Kevorkian and Julian D Cole. *Perturbation methods in applied mathematics*. Springer, 2013.
- [22] Eleftherios Kirkinis. The renormalization group: A perturbation method for the graduate curriculum. *SIAM Review*, 54(2):374–388, 2012.

- [23] Cornelius Lanczos. *Applied analysis*. Dover Pubns, 1988.
- [24] Derek F. Lawden. *Elliptic functions and applications*, volume 80. Springer Science & Business Media, 2013.
- [25] J.A. Morrison. Comparison of the modified method of averaging and the two variable expansion procedure. *SIAM Review*, 8(1):66–85, 1966.
- [26] Ali H Nayfeh. *Introduction to perturbation techniques*. John Wiley & Sons, 2011.
- [27] The On-Line Encyclopedia of Integer Sequences. <https://oeis.org/>.
- [28] Robert E. O’Malley. *Historical Developments in Singular Perturbations*. Springer, 2014.
- [29] Robert E. O’Malley and Eleftherios Kirkinis. A combined renormalization group-multiple scale method for singularly perturbed problems. *Studies in Applied Mathematics*, 124(4):383–410, 2010.
- [30] Richard Rand and Dieter Armbruster. *Perturbation methods, bifurcation theory and computer algebra*, volume 65. Springer Science & Business Media, 2012.
- [31] Bruno Salvy and John Shackell. Measured limits and multiserries. *Journal of the London Mathematical Society*, 82(3):747–762, 2010.
- [32] Milton Van Dyke. *Perturbation methods in fluid mechanics*. Academic Press, 1964.
- [33] R.F. Warming and B.J. Hyett. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of computational physics*, 14(2):159–179, 1974.
- [34] J.A.C. Weideman. Computing the dynamics of complex singularities of nonlinear PDEs. *SIAM J. Appl. Dyn. Syst*, 2(2):171–186, 2003.
- [35] James H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, 1963.
- [36] James H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, New York, 1965.
- [37] James H. Wilkinson. Modern error analysis. *SIAM Review*, 13(4):548–568, 1971.
- [38] James H. Wilkinson. The perfidious polynomial. In Gene H. Golub, editor, *Studies in Numerical Analysis*, volume 24, pages 1–28. Mathematical Assosication of America, 1984.
- [39] Yiming Zhang and Robert M. Corless. High-accuracy series solution for two-dimensional convection in a horizontal concentric cylinder. *SIAM Journal on Applied Mathematics*, 74(3):599–619, 2014.