Perturbation Methods using Backward Error

Robert M. Corless

CUNEF University, Madrid, Oct 23 2024

Western University, Canada

Slides available at `rcorless.github.io`; please download them

Joint work with Nic Fillion

Maple Transactions

Announcing Maple Transactions

a "Diamond" class open access journal with no page charges

Now listed by DBLP

mapletransactions.org

An exemplary paper:

Some Instructive Mathematical Errors by Richard P. Brent,

(we will remark on this paper later)

There must be “a few books” already on Perturbation Methods

Figure 1: RMC and Don Quixote, in Alcalá de Henares, 2017

Why, on Earth, write another?

Fools rush in where angels fear to tread.
—Alexander Pope, *An essay on criticism*, written 1709

- There are only two other books that use backward error [5, 6]
- We claim backward error is *very* useful for perturbation methods*
- We think computer algebra is still under-utilized nowadays, although there are some works that use it systematically
- Even though scientific computing has progressed *far* beyond perturbation methods, there is still a need for them.

* This fact may seem obvious in retrospect. We contend that the obstacle of hand labour has discouraged full use of backward error in practice till now. We will see its advantages!

The goal: short lucid formulae

Numerical solution and graphs (and animations) are truly valuable, but sometimes a short lucid formula can tell you just as much as an hour with a simulator and visualization tools can.

This *depends* on the scientist (or student!) understanding the terms in the formula, of course!

Backward error is not purely mathematical

Although backward analysis is a perfectly straightforward concept there is strong evidence that **a training in classical mathematics leaves one unprepared to adopt it**. . . . *I have even detected a note of moral disapproval in the attitude of many to its use and there is a tendency to seek a forward error analysis even when a backward error analysis has been spectacularly successful.*

—J. H. Wilkinson, in [**Wilkinson1985**]

What is "Backward Error?"

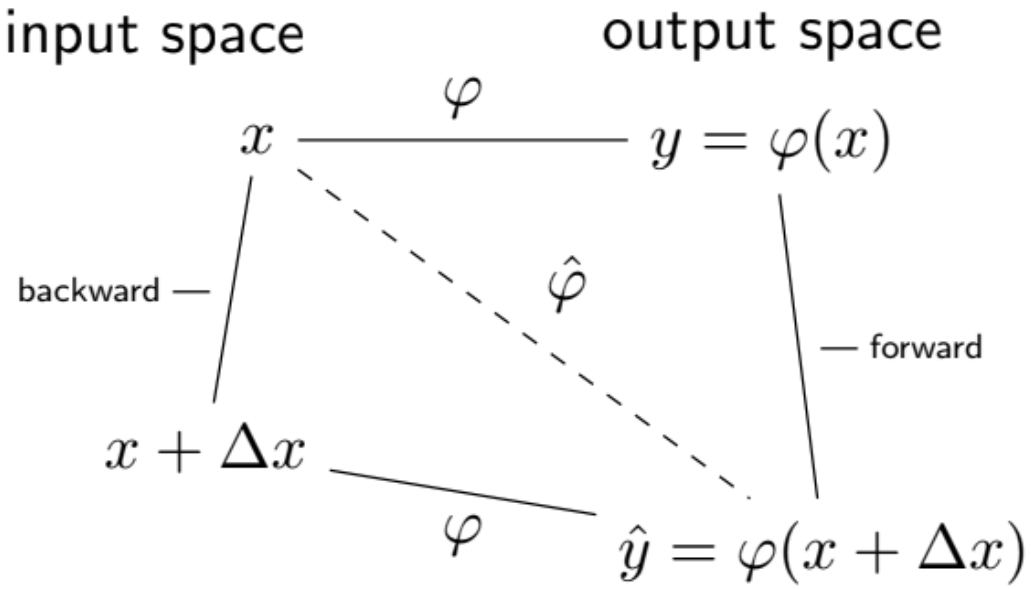

Figure 2: We want to compute $\varphi(x)$ but we cannot, for some reason. We *can* compute $\widehat{y} = \widehat{\varphi}(x)$. This has forward error $y - \widehat{y}$. But perhaps $\widehat{y} = \varphi(x + \Delta x)$ exactly; Δx is a "backward error" (this need not be unique). Or perhaps $\widehat{y} = (\varphi + \Delta\varphi)(x)$; then $\Delta\varphi$ is another kind of backward error.

That's not mathematics

Changes in the input data x, to $x + \Delta x$, are usual in science (engineering, economics, psychology, anything). Changes in the mathematical model φ are also usual: one normally neglects terms and effects that are considered to be "small" or "unimportant."

If we can put our errors-in-solution in the same context as these kinds of data or modelling errors, then we can *reuse* the tools that we have to use for such (e.g. the "sensitivity" or "conditioning" of the problem).

A numerical example

I used *linearized Newton's method* to solve $we^w = 3.0$, that is, to evaluate $W(3.0)$ where W is the Lambert W function. I did this by hand and I can show you the computations, complete with mistakes.

Initial estimate, $w_0 = 1.0$ because $1.0e^{1.0} \approx 2.7182$ so $f(w_0) = w_0 e^{w_0} - 3 \approx -0.2818$. Notice $W(2.7182\ldots) = w_0$ exactly.

The iteration is $w_{k+1} = w_k - f(w_k)/f'(w_0)$. Note we never recompute the derivative $f'(w) = (1 + w)e^w$. At $w = w_0$, $f'(w_0) = 2e \approx 5.4364$. Then $w_1 \approx 1.05018$.

The residual $f(w_1)$ is $w_1 e^{w_1} - 3 \approx 2.99985 - 3 \approx -0.00015$.

Equivalently, $w_1 = W(2.99985)$. We have found (nearly) the exact value of W evaluated at a *nearby* point.

This is a kind of backward error analysis.

Relation to forward error

From Taylor series, $W(x+\Delta x) \approx W(x) + W'(x)\Delta x$ (indeed, by the Mean Value Theorem $W(x+\Delta x) = W(x) + W'(x+\theta\Delta x)\Delta x$ for some $\theta \in (0,1)$). So we need to compute $W'(x)$: $W(x)\exp(W(x)) - x = 0$ so $W'(x)\exp(W(x)) + W(x)W'(x)\exp(W(x)) - 1 = 0$ or

$$W'(x) = \frac{1}{(1+W(x))e^{W(x)}} = \frac{W(x)}{x(1+W(x))}\,. \tag{1}$$

Therefore

$$W'(2.99985) \approx \frac{1.05}{2.99985(1+1.05)} \approx \frac{1}{6}\,, \tag{2}$$

so we see that the forward error (in this case) is about $1/6$ the backward error. This problem is not sensitive to changes in x near this x. We say the problem is *well-conditioned*.

That's more important than just forward error

We have learned more than just that we have a good, accurate answer. We have also learned that other errors in the data (and possibly the model) will not be amplified. We have learned that this equation is (near $x = 3$) insensitive to changes.

Computing Symbolically

Let's use the same method to solve $f(w) = we^w - x = 0$ symbolically, for small x. Choose an initial estimate $w_0 = 0$ so that $f'(w_0) = 1$. Then

$$w_1 = w_0 - \frac{w_0 e^{w_0} - x}{1} = 0 - \frac{-x}{1} = x \tag{3}$$

$$w_2 = x - \frac{xe^x - x}{1} \approx x - (x(1 + x + \cdots) - x) = x - x^2 \tag{4}$$

$$w_3 = w_2 - \frac{w_2 e^{w_2} - x}{1} \approx x - x^2 + \frac{3}{2}x^3 \tag{5}$$

getting one more term in the series correct with each iteration. Notice that at each stage we have exactly solved $f(w) - f(w_k) = 0$, a nearby equation if the residual $f(w_k)$ is small. This is trivial but profound.

For the record

$$f(w_3) = \left(x - x^2 + \frac{3}{2}x^3\right) \mathrm{e}^{x - x^2 + \frac{3}{2}x^3} - x \tag{6}$$

$$= \frac{8}{3}x^4 + \frac{1}{8}x^5 + \mathrm{O}\left(x^6\right) \tag{7}$$

which is very small if x is small.

In other words, w_3 is exactly $W(x + 8x^4/3 + x^5/8 + \cdots)$, the Lambert W function evaluated not at x but nearby.

Small x

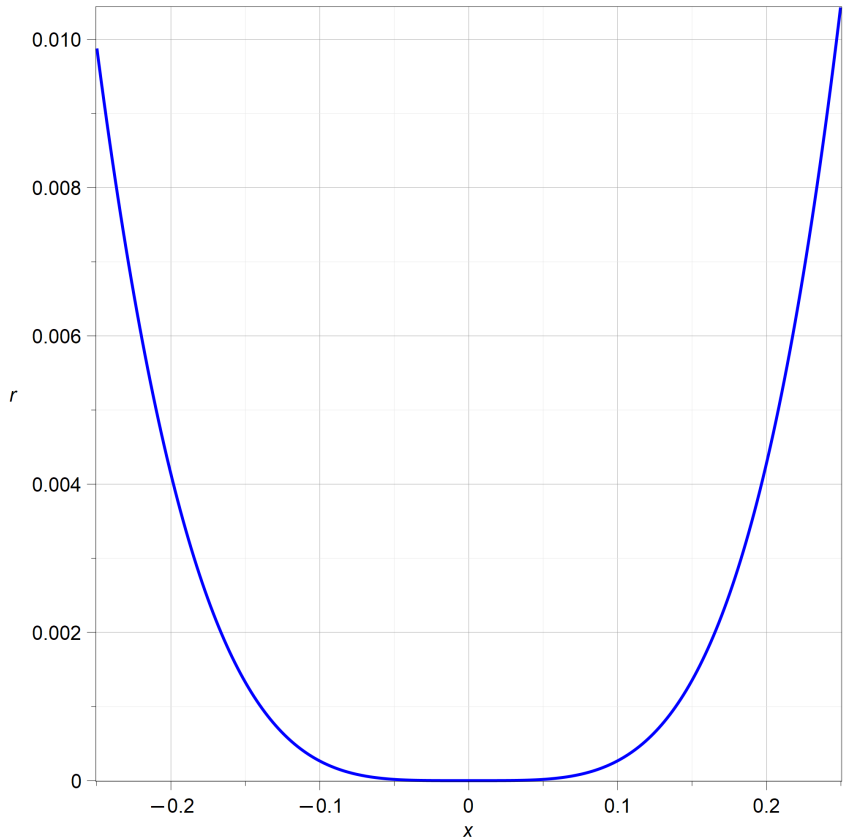

Figure 3: $w = x - x^2 + 3x^3/2$ is the exact value of $W(x + r)$, where r is pictured here. On $-0.25 \leq x \leq 0.25$ the difference is less than one percent.

Not limited to small x

If our initial estimate is $w_0 = \ln x - \ln\ln x$, then the residual is

$$f(w_0) = (\ln(x) - \ln(\ln(x)))\, \mathrm{e}^{\ln(x)-\ln(\ln(x))} - x \tag{8}$$

$$= -x\frac{\ln(\ln(x))}{\ln(x)} \tag{9}$$

which doesn't look small; but it is, compared to x. Already this w_0 is the *exact* value of

$$W\left(x\left(1+\frac{\ln(\ln(x))}{\ln(x)}\right)\right) \tag{10}$$

which, as $x \to \infty$, is closer and closer to x.

(*Tediously* slowly: $\ln\ln x/\ln x$ is still more than 10% at $x = 10^{15}$. The relative condition number $xW'/W = 1/(1+W)$ goes to zero as $x \to \infty$ but also slowly, being about $1/32$ at $x = 10^{15}$.)

An important philosophical point

We have *not* ever assumed the existence or the convergence of any infinite series or process.

Everything in this procedure is finite. At the end of every stage we can decide if our answer is good enough, or not.

As a practical matter, if our answers do not get *better* each iteration, the method will fail. So the method **can fail**. Generally speaking, the procedure will succeed for "small enough" ["large enough"] values of the parameter (and we will be able to tell if the values are "small enough" ["large enough"]).

Aging spring, Lengthened pendulum

Some physical problems have natural "secular" (slowly-varying) terms in them. For instance, consider the "aging spring" [1]:

$$\ddot{y} + e^{-\varepsilon t} y = 0\,. \tag{11}$$

Cheng and Wu claimed to have used the "two-scale" method to get the solution $\exp(\varepsilon t/4)\sin(2(1-\exp(-\varepsilon t/2))/\varepsilon)$. The "WKB method" gets this solution directly. Its residual is

$$\frac{1}{16}\varepsilon^2 \mathrm{e}^{\varepsilon t/4} \sin\left(\frac{2(1-\mathrm{e}^{-\varepsilon t/2})}{\varepsilon}\right)\,. \tag{12}$$

Is that a "small" residual? It's a bit hard to tell.

Better backward error

But! Notice that the residual in equation (12) is just $\varepsilon^2 Y(t)/16$ where $Y(t)$ is the computed solution. This means that $Y(t)$ is the *exact* solution to

$$y'' + \left(e^{-\varepsilon t} - \frac{\varepsilon^2}{16}\right) y(t) = 0\,. \tag{13}$$

This is an equation that we can *directly* interpret in terms of the original model.

Notice that the spring constant becomes *zero* when $\exp(-\varepsilon t) = \varepsilon^2/16$, or $t = -2\ln(\varepsilon/4)/\varepsilon$. We thus learn that the approximate solution is likely not valid for t larger than this, *in a way that is consonant with the mathematical modelling*. [Cheng and Wu say that this equation is used in some kind of quantum application.]

The aging spring is sensitive to some changes

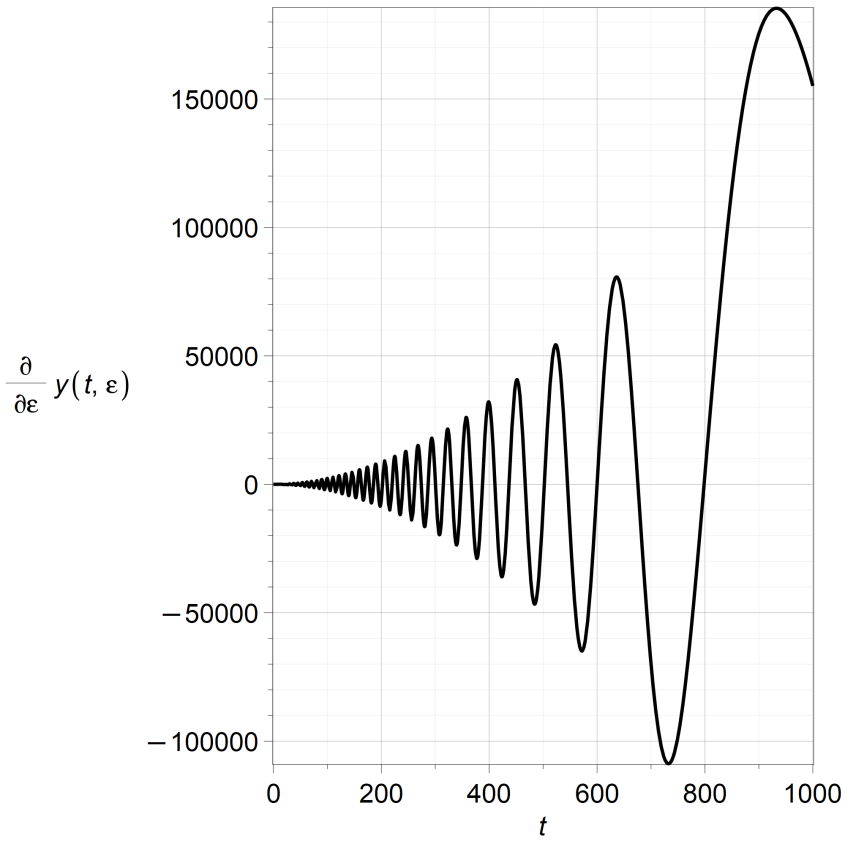

Figure 4: Taking the derivative with respect to ε shows that the solution is sensitive to changes in ε. $\varepsilon = 1/100$ here.

The context matters

Both of those details matter. Changing $e^{-\varepsilon t}$ to $e^{-\varepsilon t} - \varepsilon^2/16$ introduces a *spurious turning point* into the equation. This is likely "not physical" and demonstrates that for t large enough the Cheng–Wu solution will not be valid.

The fact that the solution varies strongly when tiny ε is changed by even a tinier amount is also a *kind* of ill-conditioning (but somehow it's "under control" in the model because we can see its consequences directly).

For both cases, to draw conclusions we need to know the physical context.

Perturbation vs Exact Solution

The analysis of the aging spring just performed—exhibiting an approximate solution that is the exact solution of a nearby problem of similar type, together with a residual and a condition number—tells us at least as much information as the exact solution in terms of Bessel functions would have.

We have identified an important issue, namely the sensitivity of the solution to changes in the problem, that will still be important for the exact (reference) solution.

WKB and backward error

The WKB (Wentzel–Kramers–Brillouin) method (or WKBJ method where the J is for Jeffreys, or LG method for Liouville–Green, or the "phase integral" method, even) gives the "solution of physical optics" of $\varepsilon^2 y'' = Q(x)y$ as

$$y_{WKB} = c_1 Q(x)^{-1/4} e^{S(x)/\varepsilon} + c_2 Q(x)^{-1/4} e^{-S(x)/\varepsilon} \tag{14}$$

where $S(x) = \int_{x_0}^{x} \sqrt{Q(\xi)}\, d\xi$. It's amazingly simple (once you get used to it); it's inspired by the integrating factor for $\varepsilon y' = P(x)y$ which is $I(x) = \int^x P(\xi)\, d\xi/\varepsilon$.

How good is the solution?

Backward error for WKB

y_{WKB} gives the *exact solution* to $\varepsilon^2 y'' = \widehat{Q}(x)y$ where

$$\widehat{Q}(x) = Q(x) + \varepsilon^2 \left(\frac{Q''}{4Q} - 5 \left(\frac{Q'}{4Q} \right)^2 \right) . \tag{15}$$

There is no further approximation there. That's a finite formula for the exact backward error $r(x) = \varepsilon^2 Q_2(x)$. The WKB method gives an exact solution to a nearby equation (provided $Q(x) \neq 0$—places where $Q(x) = 0$ are called *turning points*).

We have not seen this fact mentioned in any other textbook.

Green's functions and forward error

The forward error is then

$$\int_{x_0}^{x} G(x,\xi) r(\xi) y_{WKB}(\xi)\, d\xi \tag{16}$$

where $G(x,\xi)$ is the Green's function. We can compute it (pretty easily) for the WKB solution; it is $O(1/\varepsilon)$ in size, so the forward error will be $O(\varepsilon)$ as $\varepsilon \to 0$.

The Green's function *also* (and more importantly) measures the sensitivity to changes in the equation or model, such as added noise.

“Small” vs “Small enough”

Since Backward Error Analysis requires the *context* of the original problem to be taken into account, this explicitly allows us to consider whether the computed residual (or other backward error form) is actually small compared to other neglected effects.

This is *not* mathematics! Mathematics abstracts, as far as possible, with the goal of making its results and predictions independent of context.

This is the crux of the matter.

Once this is settled, we can consider conditioning: are such small effects amplified to the point where we lose all predictability or control, or is the solution useful?

A classical ODE

The "Renormalization Group Method" [2] makes short work of many nonlinear oscillator problems. Consider the Rayleigh equation

$$\ddot{y} - \varepsilon\dot{y}\left(1 - \frac{4}{3}\dot{y}^2\right) + y = 0\,. \tag{17}$$

"This RG method works, although it is somewhat inefficient since it first obtains the naive expansion. . ."

— Robert E. O'Malley [3, p. 187]

The Renormalization Group method in a nutshell

Take the terms that cause trouble, gather them up, and replace the troublesome series T by the exponential of the logarithm of T.

Explicitly:

- Choose N and compute the regular solution to $O(\varepsilon^{N+1})$ with secular terms in it, starting with $y_0(t) = 2A\cos(t) = A\exp(it)+\text{c.c.}$
- Gather up the term $Ay_A(t)e^{it}$ (we want the coefficient $y_A(t) = 1 + O(\varepsilon)$ of the resonant term)
- Compute the logarithm $L_A(t)$ of the series $y_A(t) = 1 + \cdots + O(\varepsilon^{N+1})$ and write $y_A(t) = \exp(L_A(t))$

Continued

- Replace the initial amplitude A by the time-dependent amplitude $R(t)$ determined by solving the equation $f(A,R,\varepsilon) = R^2 - A^2(\Re(y_A)^2 + \Im(y_A)^2) = 0$ perturbatively (regular perturbation works!)
- Determine the differential equations for $R(t)$ and $\theta(t)$ via

$$\frac{R'(t)}{R(t)} + i\theta'(t) = \frac{y_A'(t)}{y_A(t)} \tag{18}$$

  Set $t = 0$ after differentiation in the right-hand side, and replace every A by the R that you found; cf Lie group–Lie algebra exponential.
- Redo the computation with initial approximation $y(t) = 2R(t)\cos(t + \theta(t))$, using the differential equations for R and θ as you go along.

Results for the Rayleigh oscillator

It turns out $\theta'(t) = 0$. Then if $y = 2R(t)\cos t + \varepsilon R^3(t)\cos 3t/3$ where

$$\frac{d}{dt}R(t) = \frac{\varepsilon}{2}R(t)\left(1 - 4R^2(t)\right) \tag{19}$$

(an equation we may expect students to be able to solve by hand; note the stable limit cycle at $R = 1/2$) then $y(t)$ exactly solves

$$\ddot{y} - \varepsilon\dot{y}\left(1 - \frac{4}{3}\dot{y}^2\right) + y = r(t) = \varepsilon^2 v(t) \tag{20}$$

where $v(t)$ is precisely known (we have a formula for it, which we can inspect explicitly, but it's a little long to present here). Moreover, $v(t)$ is bounded for all time t.

Among other things, this shows that our computation was free of blunders. We did this up to order $N = 16$ (that is, $O(\varepsilon^{17})$), by the way, essentially for fun (the code took 35 minutes to do that case).

The residual $r(t) = \varepsilon^2 v(t)$

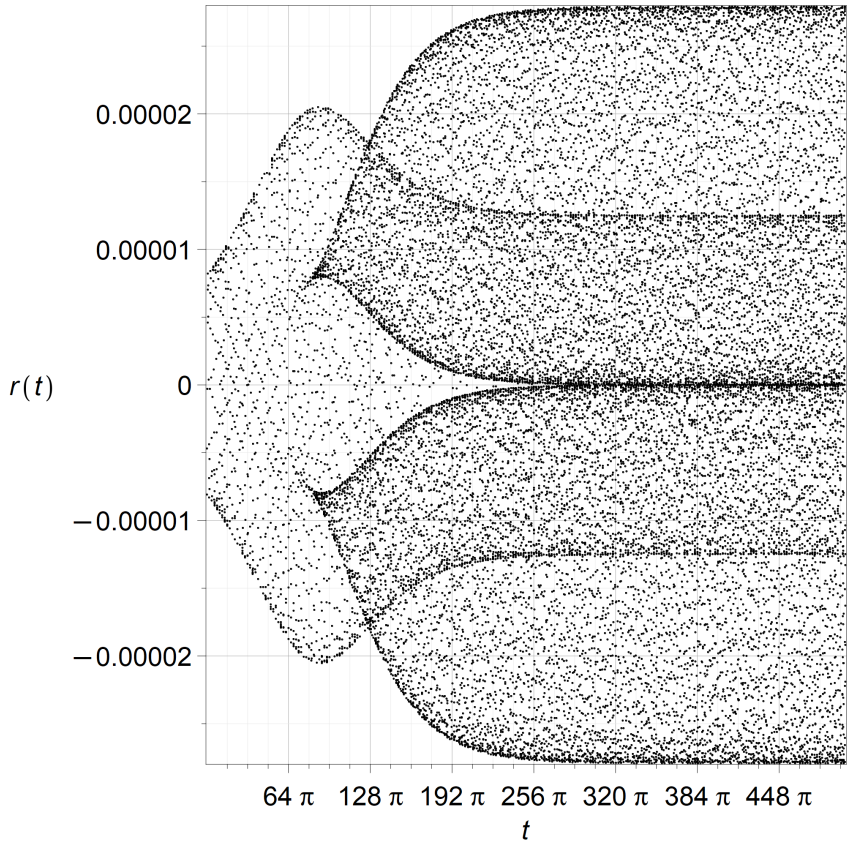

Figure 5: $R_0 = 0.15$ and $\varepsilon = 0.01$. Here $N = 1$ so the backward error is $O(\varepsilon^2)$.

The effects of a small residual

The Rayleigh equation is (near its limit cycle) extremely well-conditioned, although phase error can accumulate algebraically. The residual $r(t)$, which is small for small $\varepsilon > 0$, has very little effect on the solution—which is as it should be, because real physical oscillators are frequently subject to small forcing oscillations (shaky ground, shaky table, and the like).

Our highest-order computation, $O(\varepsilon^{17})$, had uniformly very small residual, which at the limit cycle was well below rounding error size when $\varepsilon = 0.2$. With $N = 13$ and $\varepsilon = 0.4$, the residual was about 10^{-8}.

Resonant terms

Curiously enough, that residual contains resonant terms, in that the coefficients of $\sin t$ and $\cos t$ are not zero! This might be surprising, but one can compute a *structured* backward error, at the limit cycle, that looks like this:

$$\ddot{y} - \varepsilon\dot{y}\left(1 - (1+\varepsilon^2 A)\frac{4}{3}\dot{y}^2\right) + (1+\varepsilon^2 B)y = r_1(t) = \varepsilon^2 v_1(t) \qquad (21)$$

for some bounded quantities A and B and where now the new residual $r_1(t)$ contains no resonant terms. That is, we might interpret the residual as containing a slight change in frequency and in damping, as well as a small forcing.

In fact we find $A = -1/32 + O(\varepsilon^2)$ and $B = 1/8 + O(\varepsilon^2)$ at the limit cycle.

This is a *structured backward error*.

Absence of secularity

The normal treatment of secularity assumes that one knows ahead of time that the reference solution to the model problem is bounded. Sometimes that's obvious physically, but sometimes one has to prove boundedness (e.g. with the Duffing equation one finds a conserved integral).

But here if we can find a solution (as here) with a uniformly bounded residual, this provides its own proof of validity.

The argument looks a bit circular, but it's not. We used the RG method to find a solution that had no secular terms in it, and hence which had a bounded residual, which proved the solution valid (in the context of modelling errors).

Morrison's counterexample

In [3, pp. 192–193], we find a discussion of the equation

$$y'' + y + \varepsilon(y')^3 + 3\varepsilon^2(y') = 0\,. \tag{22}$$

O'Malley's solution, there and in [4], is incorrect. We claim that had he computed a residual, he would have identified the blunder*.

* By "blunder" we mean arithmetic error, or algebra error, no more. It's just that the word "error" is a bit overused in this field already. Also, I feel some worry* in pointing out this blunder: O'Malley was a giant of perturbation methods. But we are certain that our solution is correct.

* inquietud, intranquilo, desasogado, ...

RG Solution

All we need do is to change the differential equation in our Jupyter notebook script, and alter the interrogations of the solution afterward. At $N = 2$, we get

$$\begin{aligned} z(t) =& 2R(t)\cos(t+\theta(t)) + \frac{\varepsilon R(t)^3 \sin(3t+3\theta(t))}{4} \\ &+ \varepsilon^2 \left(\frac{27R(t)^5 \cos(3t+3\theta(t))}{32} - \frac{3R(t)^5 \cos(5t+5\theta(t))}{32} \right) \end{aligned} \tag{23}$$

with

$$\dot{R}(t) = -\frac{3\varepsilon}{2} R(t) \left(R(t)^2 + \varepsilon \right) \tag{24}$$

and

$$\dot{\theta}(t) = \frac{9}{16} R^4(t) \varepsilon^2 \,. \tag{25}$$

With this, we get a uniformly small residual, which is small even compared to the decaying amplitude.

An important tangent: Published blunders

Off the top of my head, blunders in published perturbation computations have been exhibited by

- John P. Boyd (a hyperasymptotic expansion)
- Robert E. O'Malley (Morrison's counterexample)
- Émile Mathieu (in his 1868 paper which defined what are now called Mathieu functions)
- Bender & Orszag (a plain multiple scales computation, fixed in later editions)

at least. I claim that had they computed a final residual, they would have detected their blunders. Given that *all* of the above are/were experts, and therefore it's clear that the rest of us make blunders at least as frequently, I claim that residuals are even more necessary for us.

To be fair

Richard Brent was fair enough to include some of his own errors in the paper "Some instructive mathematical errors" I mentioned previously and so I should say explicitly that I make blunders, too. In my paper "A Sequence of Series for the Lambert W function " I claimed a certain series had infinite radius of convergence. Richard Crandall pointed out that I was wrong and the series had radius of convergence $\sqrt{2\pi}$.

So I am guilty, too!

Perturbation is just taking derivatives

The simplicity of a perturbation computation hides its importance. We are investigating what happens if a small part of the model changes.

This is itself a fundamental question of science. It's not surprising that the old techniques are still valuable; maybe it's a surprise just how valuable they can be.

That said, nowadays one can do a heck of a lot with a simulation window and a slider bar.

Thank you for listening.

This work supported by NSERC, and by the Spanish MICINN. I also thank CUNEF University for the opportunity to give this talk.

I am also happy to announce that SIAM has offered Nic and me a contract for this book, and we are to deliver it to them by December. Your feedback today will help to improve the book, and we will acknowledge you all.

Book text available at `https://github.com/rcorless/rcorless.github.io/blob/main/PerturbationBEABook.pdf`. Please download it and read it and send me (or Nic) your comments, by Nov 10 if possible.

Let's open the topic for discussion.

References

[1] Hung Cheng and Tai Tsun Wu. **"An aging spring".** In: *Studies in applied Mathematics* 49.2 (1970), pp. 183–185.

[2] Eleftherios Kirkinis. **"The Renormalization Group: A perturbation method for the graduate curriculum".** In: *SIAM Review* 54.2 (2012), pp. 374–388.

[3] Robert E. O'Malley. **Historical Developments in Singular Pertubations.** Springer, 2014.

[4] Robert E. O'Malley and Eleftherios Kirkinis. **"A Combined Renormalization Group-Multiple Scale Method for Singularly Perturbed Problems".** In: *Studies in Applied Mathematics* 124.4 (2010), pp. 383–410.

References ii

[5] Anthony John Roberts. **Model emergent dynamics in complex systems.** SIAM, 2014.

[6] Donald R. Smith. **Singular-perturbation Theory.** Cambridge University Press, 1985.

Using computation to illustrate formulae

Let's try to understand which is bigger, the term $\exp(-1/\varepsilon)$ or any algebraic term ε^j. L'Hopital's rule shows that as $\varepsilon \to 0^+$ the exponential is transcendentally smaller than any ε^j. But what happens if we ask when the two are equal?

$$e^{-1/\varepsilon} = \varepsilon^j \tag{26}$$

exactly when $\varepsilon_{-1} = e^{W_{-1}(-1/j)}$ (on the left) and when $\varepsilon_0 = e^{W_0(-1/j)}$. Here W_{-1} and W_0 are the two real branches of the Lambert W function. [Short, lucid formulae, just what we want*.]

So ε^j is *smaller* than $\exp(-1/\varepsilon)$ if $\varepsilon_{-1} < \varepsilon < \varepsilon_0$. Paradoxically, this is most of the interval, for large j!

* Heh. $\varepsilon_{-1} \sim 1/(j \ln j)$ and $\varepsilon_0 \sim 1 - 1/j$ might be easier to understand!

When is that?

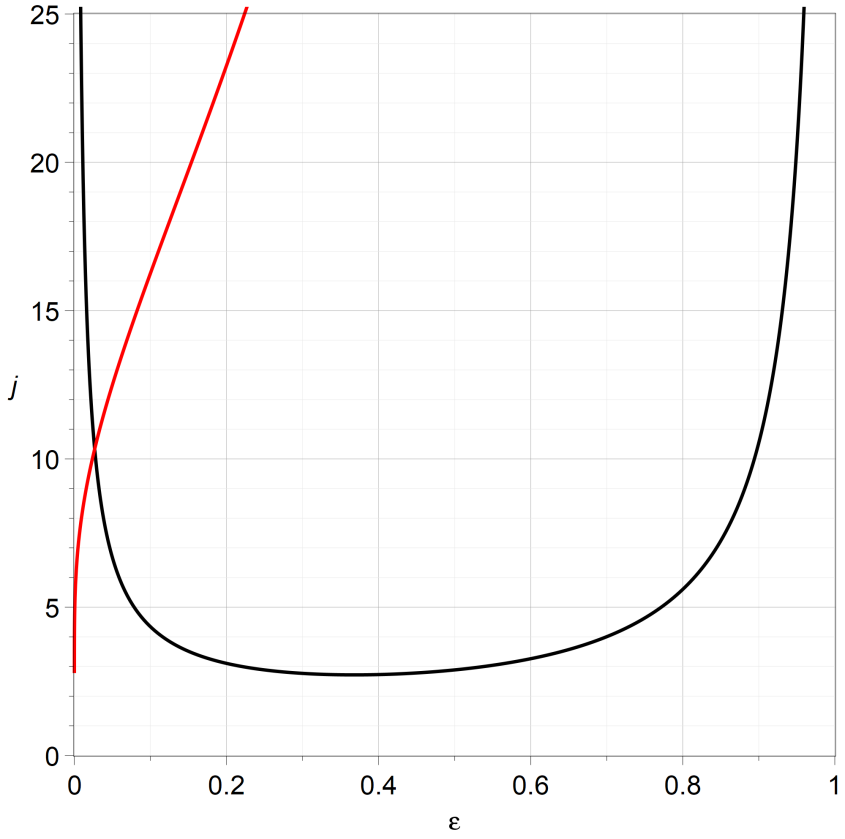

Figure 6: For values of j above this curve, $\varepsilon^j < \exp(-1/\varepsilon)$. That is, the "exponentially small" term is more important! Left of the red line is lost to rounding error in double precision. Note $\exp(-1/\varepsilon) = 2^{-54}$ already when $\varepsilon = \ln(2)/54 \approx 0.0267$.