1983

Some Asymptotic Estimates for Higher-Order Averaging and a Comparison with Iterated Averaging

James Murdock
Iowa State University, jmurdock@iastate.edu

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

0036-1410/83/1403-0001 $01.25/0

SOME ASYMPTOTIC ESTIMATES FOR HIGHER ORDER AVERAGING AND A COMPARISON WITH ITERATED AVERAGING*

JAMES A. MURDOCK†

Abstract. Asymptotic estimates for classical higher order averaging are obtained on intervals of length greater than $O(1/\varepsilon)$ when some of the averages vanish. These results are compared with results of Persek using iterated averaging, and the classical methods are found to be more powerful.

1. Introduction. We shall be concerned with the n-dimensional system of differential equations

$$\dot{x}=\varepsilon f(x,t,\varepsilon)=\varepsilon f_1(x,t)+\varepsilon^2 f_2(x,t)+\cdots+\varepsilon^n f_n(x,t)+\varepsilon^{n+1} r_{n+1}(x,t,\varepsilon) \tag{1}$$

where f is 2π-periodic in t and ε is a small positive real number. For such systems the traditional nth order averaging method, as described for instance in Perko [2], yields approximate solutions which retain accuracy $O(\varepsilon^n)$ on time intervals of length $O(1/\varepsilon)$. It is of interest whether one can "trade off" some of this accuracy for validity on a longer time interval; that is to say, we may ask whether the same approximate solution retains the decreased accuracy $O(\varepsilon^{n-j})$ on the expanded time interval of length $O(1/\varepsilon^{1+j})$ for certain integers j. Our first theorem (§2) asserts that this is true for $j=0,1,\cdots,l-1$ if the nth order averaged system corresponding to (1) begins with the term of order ε^l, in other words takes the form

$$\dot{z}=\varepsilon^l g_l(z)+\cdots+\varepsilon^n g_n(z). \tag{2}$$

The proof involves no new methods.

Persek [3] has defined a method which he calls "iterated averaging" which under certain conditions approximates system (1) by a system of the form

$$\dot{z}=\varepsilon^l h_l(z). \tag{3}$$

He then proves that solutions of (3) approximate those of (1) to order $O(\varepsilon)$ on an interval of length $O(1/\varepsilon^l)$. It is evident that (3) has the same form as (2) in the case $n=l$, although it is not clear a priori whether the function h_l constructed by Persek coincides with the classical g_l in this case. Nevertheless Persek's estimate of the accuracy of (3) coincides with our estimate for (2) if $n=l$ and if j is taken to be $l-1$. This prompts a comparison between g_l and h_l, which we carry out in §3 for the case $l=2$. Briefly the result is that g_2 and h_2 coincide when both are defined, but that the defineability of h_2 depends upon a hypothesis which is unnecessary under the traditional method. Thus Persek's result is (at least for the case $l=2$) a special case of ours (in §2). For $l>2$ it would be tedious to compare h_l and g_l but it is again apparent that g_l is always defined (and thus equation (2) exists provided that g_1 through g_{l-1} vanish), whereas h_l is defined only if h_1 through h_{l-1} vanish and in addition a further condition is satisfied, which does not correspond to any condition in the traditional theory.

*Received by the editors November 13, 1981.
†Department of Mathematics, Iowa State University, Ames, Iowa 50011.

2. Asymptotic estimates. It is shown in the classical theory of averaging that there exists a coordinate transformation

$$x=u(y,t,\varepsilon)=y+\varepsilon u_1(y,t)+\cdots+\varepsilon^n u_n(y,t) \tag{4}$$

2π-periodic in t and carrying (1) to

$$\dot{y}=\varepsilon g_1(y)+\cdots+\varepsilon^n g_n(y)+\varepsilon^{n+1}R_{n+1}(y,t,\varepsilon). \tag{5}$$

The transformation (4) is not unique and is usually normalized either by requiring that the average of each u_k vanish, or by requiring that each u_k vanish for $t=0$. The latter is called the stroboscopic method and has the advantage that (4) reduces to $x=y$ at $t=0$ and at all stroboscopic times $t=2\pi, 4\pi, \cdots$. Associated with (5) is the truncated system

$$\dot{z}=\varepsilon g_1(z)+\cdots+\varepsilon^n g_n(z) \tag{6}$$

which is wholly autonomous. Let $x(t,\varepsilon)$, $y(t,\varepsilon)$, $z(t,\varepsilon)$ denote solutions of (1), (5), and (6) defined in an interval $0\le\varepsilon\le\varepsilon_0$ whose initial conditions are related by $x(0,\varepsilon)=u(y(0,\varepsilon),0,\varepsilon)$ and $y(0,\varepsilon)=z(0,\varepsilon)$; note that in the stroboscopic case this reduces to $x(0,\varepsilon)=y(0,\varepsilon)=z(0,\varepsilon)$. The classical method of averaging proposes to construct from $z(t)$ an approximation to $x(t)$. Since $z(t)$ ought to be close to $y(t)$, and since $y(t)$ is related to $x(t)$ by (4), in view of the fact that (4) is assumed to hold at $t=0$, $x(t)$ should be well approximated by $u(z(t,\varepsilon),t,\varepsilon)$, an expression which is called the *improved nth approximation* to $x(t)$. It turns out, however, that there is no loss in asymptotic accuracy if the term of order n in (4) is omitted in forming the approximation (it must not be omitted, of course, in transforming (1) into (5)). Therefore the *nth approximation* to $x(t)$ is defined by

$$X(t,\varepsilon)=\hat{u}(z(t,\varepsilon),t,\varepsilon), \tag{7}$$

where

$$\hat{u}(z,t,\varepsilon)=z+\varepsilon u_1(z,t)+\cdots+\varepsilon^{n-1}u_{n-1}(z,t).$$

THEOREM. *Under the above hypotheses there exist positive constants c and c_0 such that $|x(t,\varepsilon)-X(t,\varepsilon)|\le c_0\varepsilon^n$ for $0\le t\le c/\varepsilon$ and $0\le\varepsilon\le\varepsilon_0$. Under the additional hypothesis that g_1 through g_{l-1} vanish, so that* (6) *takes the form* (2), *there exist for each $j=0,\cdots,l-1$ positive constants c_j such that $|x(t,\varepsilon)-X(t,\varepsilon)|\le c_j\varepsilon^{n-j}$ for $0\le t\le c/\varepsilon^{1+j}$.*

Proof. We prove the second assertion. The first, which is classical, is included by taking $l=1, j=0$.

Let K be a closed ball centered at $x(0,0)$ of radius R sufficiently large that $x(0,\varepsilon)$ and $y(0,\varepsilon)=z(0,\varepsilon)$ are contained in the concentric ball of radius $R/2$ for $0\le\varepsilon\le\varepsilon_0$. Since K is compact there exists in view of (1), (2) and (5) a constant M such that $d|y|/dt$ and $d|z|/dt$ are less than $M\varepsilon^l$ for $0\le\varepsilon\le\varepsilon_0$ as long as y and z remain in K. Since $|y|$ and $|z|$ must drift by at least $R/2$ from their initial positions in order to leave K, and since in time t they can drift at most $M\varepsilon^l t$ while in K, it follows that there exists a constant $c>0$ such that y and z remain in K for $0\le t\le c/\varepsilon^l$. Hence on this interval (the longest considered in the theorem) one may use, for all functions of y and z, their upper bounds and Lipschitz constants on K.

Letting $\rho=|y(t,\varepsilon)-z(t,\varepsilon)|$ one immediately finds from (2) and (5) Lipschitz constants $L_l,\cdots,L_n$ and a bound B such that $d\rho/dt\le(\varepsilon^l L_l+\cdots+\varepsilon^n L_n)\rho+\varepsilon^{n+1}B$. Solving this linear differential inequality with initial condition $\rho=0$ yields $\rho\le\varepsilon^{n+1}B\delta^{-1}(e^{\delta t}-1)$ where $\delta=\varepsilon^l L_l+\cdots+\varepsilon^n L_n$. Estimating the factor $e^{\delta t}-1$ requires some care. The tempting answer $e^{\delta t}-1=O(\delta t)=O(\varepsilon^l t)$ is correct for the time intervals with which we

are concerned, but this requires proof. Namely $e^x-1=x(1+x/2!+\cdots)=x\phi(x)$ and hence for x in any bounded interval there is a constant k such that $e^x-1\le kx$; if $x=\delta t$ and $0\le t\le c/\varepsilon^l$ and $0<\varepsilon\le\varepsilon_0$ then x is bounded and $e^{\delta t}-1\le k\delta t$. (On longer intervals $e^{\delta t}-1$ can approach infinity faster than δt.) Since $\delta^{-1}=O(1/\varepsilon^l)$ we have $\rho=O(\varepsilon^{n+1}t)$ as long as y and z remain in K, that is, at least on the interval $0\le t\le c/\varepsilon^l$. It follows that on any interval $0\le t\le c/\varepsilon^{1+j}$, $j=0,\cdots,l-1$, one has $|y(t,\varepsilon)-z(t,\varepsilon)|=\rho=O(\varepsilon^{n-j})$; that is, this quantity is bounded by a constant times ε^{n-j} for $0\le t\le c/\varepsilon^{1+j}$ and $0\le\varepsilon\le\varepsilon_0$. Now using the Lipschitz constant for u on K one finds $|x(t,\varepsilon)-u(z(t,\varepsilon),t,\varepsilon)|=|u(y(t,\varepsilon),t,\varepsilon)-u(z(t,\varepsilon),t,\varepsilon)|=O(\varepsilon^{n-j})$ on the same interval. But

$$|u(z(t,\varepsilon),t,\varepsilon)-X(t,\varepsilon)|=|u(z(t,\varepsilon),t,\varepsilon)-\hat{u}(z(t,\varepsilon),t,\varepsilon)=O(\varepsilon^n)$$

for all time; adding the last two estimates proves the theorem. Q.E.D.

In the proof it is seen that the final term in the transformation u is unnecessary in constructing $X(t)$ because the error committed by leaving it out is of the same order as the error already present. By the same reasoning we see that for $j>0$, where the possible accuracy is at most $O(\varepsilon^{n-j})$, we may omit j additional terms from u in forming X. In particular, in the case $n=l$, $j=l-1$, it is not necessary to use u at all and we obtain

COROLLARY. *When* (2) *reduces to* $\dot{z}=\varepsilon^l g_l(z)$, *there exist positive constants* c *and* c_0 *such that* $|x(t,\varepsilon)-z(t,\varepsilon)|<c_0\varepsilon$ *for* $0\le t\le c/\varepsilon^l$.

This form of the theorem is the one most directly comparable to the work of Persek. The comparison is carried out in the next section.

3. Comparison of two averaging methods. In order to calculate $g_2(z)$ it is necessary to recall how (4) and (5) are constructed. It is clear a priori that any transformation of the form (4) carries (1) into a system of the form

$$\dot{y}=\varepsilon g_1(y,t)+\cdots+\varepsilon^n g_n(y,t)+\varepsilon^{n+1}R_{n+1}(y,t,\varepsilon). \tag{8}$$

From (1), (4), and (8) one calculates that the f's, u's, and g's are related by

$$\begin{aligned}\frac{\partial u_1}{\partial t}(y,t)&=f_1(y,t)-g_1(y,t),\\ \frac{\partial u_2}{\partial t}(y,t)&=\left\{f_2+\frac{\partial f_1}{\partial y}u_1-\frac{\partial u_1}{\partial y}g_1\right\}_{(y,t)}+g_2(y,t)\end{aligned} \tag{9}$$

with similar equations for the higher u_n's. (Briefly, to obtain (9) differentiate (4), insert (5) and compare this with the result of inserting (4) into (1).) It is clear that (9) admits solutions for u_1 and u_2 which are periodic in t if and only if the right-hand sides have zero mean value. Now to achieve (5) the g_i must be independent of t; thus (9) dictates that $g_1(y)$ must be the average of $f_1(y,t)$ and $g_2(y)$ must be the average of the expression in braces.

We wish to consider the case in which the averaged system takes the form (2) with $l=2$. Thus we now assume $g_1(y)=0$, which is to say that the average of $f_1(y,t)$ vanishes. In this case $u_1(y,t)=\int_a^t f_1(y,s)\,ds$ with a arbitrary ($a=0$ gives the stroboscopic method). Inserting this into the expression in braces and averaging gives the following formula for g_2, in which the dependence upon the choice of a is made explicit:

$$g_2(y,a)=\frac{1}{2\pi}\int_0^{2\pi}\left\{f_2(y,t)+\frac{\partial f_1}{\partial y}(y,t)\int_a^t f_1(y,s)\,ds\right\}dt. \tag{10}$$

Persek's function $h_2(z)$ is defined as follows in our notation (compare his $\bar{E}^{(2)}$, [3, p. 416]). First assume the average of $f_1(y,t)$ vanishes, as we have done above. Next define

$$H_2(z,\tau)=\frac{1}{2\pi}\int_{\tau}^{\tau+2\pi}\left\{f_2(z,t)+\frac{\partial f_1}{\partial y}(y,t)\int_{\tau}^{t}f_1(y,s)\,ds\right\}dt. \tag{11}$$

If the latter expression is independent of τ, it is defined to be $h_2(z)$; if it is not independent of τ, $h_2(z)$ is not defined.

Now it is clear that the bracketed expression in (10) is periodic in t, since we have assumed that f_1 has zero mean value. Therefore $\int_0^{2\pi}$ in (10) may be replaced by $\int_{\tau}^{\tau+2\pi}$ for any τ. The only remaining difference between (10) and (11) is that a in (10) is replaced by τ in (11). Thus we see that

$$H_2(z,\tau)=g_2(z,\tau). \tag{12}$$

Thus the sole difference between Persek's average (for $n=l=2$) and ours is that Persek must assume (11) independent of τ, whereas the corresponding quantity a in (10) enters as an arbitrary constant and requires no additional assumptions. It is clear that the assumption that (11) is independent of τ is a very strong assumption and one which gains no advantage.

With regard to higher order terms the following situation obtains. Our g_k is always defineable, and is not unique but rather depends upon the choices of integration constants in solving for u. On the other hand h_k is only defined if two conditions are met: $h_1,\cdots,h_{k-1}$ must be defined and vanish, and a certain function $H_k(z,\tau)$ (which Persek calls $\bar{E}^{(k)}$) must be independent of τ. The presence of the latter restriction indicates that there are likely to be many cases in which (2) takes the form $\dot{z}=\epsilon^l g_l(z)$ and yet (3) cannot be formulated. It seems reasonable to conjecture (based on the case $l=2$) that h_l exists precisely when g_l is unique (i.e., independent of the choices made in u) and that in this case $h_l=g_l$. Proof of such a theorem, if true, would involve notational difficulties but might be attempted (if it were considered important) by the use of the operators constructed by Musen [1] for use in averaging methods, based upon a formula of St. Faa de Bruno.

REFERENCES

[1] P. Musen, *On the high order effects in the methods of Krylov–Bogoliubov and Poincaré*, J. Astronaut. Sci. 12 (1965), pp. 129–134.

[2] L. M. Perko, *Higher order averaging and related methods for perturbed periodic and quasi-periodic systems*, SIAM J. Appl. Math., 17 (1968), pp. 698–723.

[3] S. C. Persek, *Hierarchies of iterated averages for systems of ordinary differential equations with a small parameter*, SIAM J. Math. Anal., 12 (1981), pp. 413–420.