The double cnoidal wave of the Korteweg–de Vries equation: An overview

John P. Boyd
Department of Atmospheric and Oceanic Science, University of Michigan, Ann Arbor, Michigan 48109

(Received 22 March 1984; accepted for publication 22 June 1984)

Earlier work of the author on the spatially periodic solutions of the Korteweg–de Vries equation is here extended via an in-depth treatment of a special case. The double cnoidal wave is the simplest generalization of the ordinary cnoidal wave discovered by Korteweg and de Vries in 1895. In the limit of small amplitude, the double cnoidal wave is the sum of two noninteracting linear sine waves. In the oppositie limit of large amplitude, it is the sum of solitary waves of two different heights repeated periodically over all space. Although special, the double cnoidal wave is important because it is but the particular case $N = 2$ of a broad family of solutions known variously as "N-polycnoidal waves," "finite gap," "finite zone" solutions, "waves on a circle," or "N-phase wave trains." It has been shown by others that the set of N-polycnoidal waves gives the general initial value solution to the Korteweg–de Vries equation. This present work is the core of a three-part treatment of the double cnoidal wave. This part, the overview, presents graphic examples in all the important parameter regimes, explains how collision phase shifts alter the average speed of the two wave phases from the "free" velocities of the two solitary waves, describes the different branches or modes of the double cnoidal wave (it is possible to have many solitary waves on each spatial period provided they are of only two distinct sizes), and contrasts the results of this work with the very limited numerical calculations of previous authors. The second part describes how the problem of numerically calculating the double cnoidal wave can be reduced down to solving four algebraic equations by perturbation theory. The third part explains how the so-called "modular transformation" of the Riemann theta functions is important in interpreting N-polycnoidal waves.

PACS numbers: 02.30.Jr, 02.60.Lj

I. INTRODUCTION

The "Hill's spectrum method," developed in the mid-1970's by Lax, Novikov, McKean, and others, has been a powerful theoretical tool for understanding the spatially periodic solutions of the Korteweg–de Vries and other soliton equations. In particular, it showed that there existed solutions which generalize the simple cnoidal waves found by Korteweg and de Vries themselves in 1895. These generalizations were dubbed "polycnoidal waves" in Ref. 1 but they are known alternatively as "finite band" or "finite gap" solutions in the Russian literature and sometimes as "N-phase wave trains" in the American journals. The N-polycnoidal wave is a function of N "phase" variables of the form

$$\zeta_i = k_i(x - c_i t) + \phi_i, \tag{1.1}$$

where the k_i are wavenumbers, the c_i are phase speeds, and the ϕ_i are constant phase factors. The most compact expression for $u(x,t)$ is in terms of a N-dimensional Riemann theta function whose arguments are the N "phase" variables defined in (1.1). Although the polycnoidal waves, like the ordinary cnoidal wave which is the special case $N = 1$, are thus special solutions, it has been shown that the class of polycnoidal waves is dense on the set of solutions of the Korteweg–de Vries (KdV) equation which are spatially periodic. To put it another way, the solution to the KdV equation for an arbitrary initial condition can be approximated for an arbitrary finite time interval to an arbitrary degree of accuracy by an N-polycnoidal wave of appropriate parameters and sufficiently large N. Thus, to understand these special solutions is to understand the general solution, too, at least for finite time.

Unfortunately, like its counterpart, the inverse scattering method for a spatially unbounded domain, the Hill's spectrum method is very complicated and a poor tool for actual numerical calculations. To quote Ferguson *et al.*,[2] "the exact formulas seem to be of little practical use." An alternative approach was discovered by Hirota[3,4] and subsequently generalized to the spatially periodic problem independently by Nakamura[5] and Boyd.[1] The reason for the alternative's effectiveness is that the theta functions satisfy not the KdV equation itself, but rather Hirota's transformed version, which will be called the "Hirota–Korteweg–de Vries" or "HKdV" equation; the solution of the KdV equation is obtained by taking the second derivative with respect to x of the logarithm of the theta function. Because the theta function depends on only a finite number of parameters, it is possible to reduce the problem down to that of solving a finite set of algebraic equations to determine these theta function parameters.

The aim of this paper, which is a sequel to Ref. 1, is to exploit this Hirota-theta function approach to deepen our understanding of the spatially periodic solutions of the Korteweg–de Vries equation, paying particular attention to $N = 2$, the double cnoidal wave. This article and its two companion papers,[6,7] are a single connected work. The other two papers discuss a perturbative (and numerical) solution of the implicit dispersion relation for the theta function parameters and the role of the "special" modular transformation of the theta functions in physically interpreting the polycnoidal wave solutions. This paper will strive to provide a general overview of the physics and mathematics of polycnoidal waves, leaving the technical details to the other two articles wherever possible.

Before giving an outline of this work, it is useful to compare and contrast its aims with those of three other schools of polycnoidal wave studies. A. Nakamura and his collaborators R. Hirota, M. Ito, and Y. Matsuno[5,8–10] have developed the direct theta function method by showing, via a mixture of clever theorems and occasional numerical calculations, that it can be used in principle to reduce a large number of different soliton-admitting partial differential equations to a finite set of algebraic equations for the theta function parameters. Equations whose Hirota-transformed equivalent is a set of coupled bilinear equations or a complex equation are discussed as well as the simpler case of those which, like the Korteweg–de Vries equation, transform into a single bilinear equation with real coefficients. They emphasize that a number of as yet unresolved technical difficulties exist for these other classes of equations, which is why this present article is focused specifically on the KdV equation. The limitations of their work are a lack of explicit calculations (except for ordinary cnoidal waves and some numerical computations described in Sec. VII), omission of perturbation theory such as is given in Ref. 6, and restriction to theta Fourier series only. The alternative Gaussian series for the theta function, introduced in Ref. 1, is a better way to explore the near-solitary wave regime.

Forest, McLaughlin, Flaschka, and Ferguson[2,11] have, like the author, attempted to explore polycnoidal waves in the spirit of applied mathematics rather than pure mathematics by taking a "concrete viewpoint," to borrow a phrase from the title of Ferguson *et al.*[2] Though the philosophy thus is similar, the line of attack is very different: this work and Refs. 1, 6, and 7 scrupulously avoid any explicit use of the Hill's spectrum method while Ferguson *et al.*[2] have "Spectral theory" as the first words of their title. Their whole approach is oriented toward understanding polycnoidal waves via calculation of the spectrum of Hill's equation and they avoid all mention of Hirota's transformed bilinear equations, perturbation theory, the special modular transformation, and most of the other topics we will discuss. Thus, their work is complementary to what will be presented here.

The Polish school of Zagrodziński and Jaworski[12] has written an interesting series of papers on the sine-Gordon equation. Their approach is inverse to that used here in that they completey specify the theta matrix and then solve for the wavenumbers k_j. This simplifies much of the analysis at the expense of obtaining generally nonintegral k_j so that their solutions are "almost periodic" rather than periodic in space.

II. AN OVERVIEW OF THE DOUBLE CNOIDAL WAVE

The Hill's spectrum method has shown that the N-polycnoidal wave is most easily expressed in terms of an N-dimensional Riemann theta function via

$$u(x,t) = 12\frac{d^2}{dx^2}\ln[\theta(x,t)], \tag{2.1}$$

where $\theta(x,t)$ is the N-dimensional Riemann theta function and where $u(x,t)$ is the actual solution of the Korteweg–de Vries equation

$$u_t + uu_x + u_{xxx} = 0. \tag{2.2}$$

For the special case $N=2$, which will be henceforth called the "double cnoidal wave," the theta function is defined by

$$\theta = \sum_{n_1=-\infty}^{\infty}\sum_{n_2=-\infty}^{\infty}\exp(-\{T_{11}n_1{}^2 + 2T_{12}n_1n_2 + T_{22}n_2{}^2\}) \times\exp[2\pi i(n_1X + n_2Y)], \tag{2.3}$$

where the T_{ij} are the elements of a 2×2 positive definite symmetric matrix known as the "theta matrix" and where X and Y are the "phase variables" defined, as in (1.1), by

$$X = k_1(x - c_1t) + \phi_1, \tag{2.4}$$

$$Y = k_2(x - c_2t) + \phi_2. \tag{2.5}$$

Mathematicians normally define the theta function in terms of an imaginary theta matrix as explained in Appendix A, but the real-valued T_{ij} employed in (2.3) are more convenient for calculations. The independent parameters are the wavenumbers k_1 and k_2, and the diagonal theta matrix elements T_{11} and T_{22}. The dependent parameters are the phase speeds c_1 and c_2, plus the diagonal theta matrix element T_{12}. [There is a fourth dependent parameter, the constant of integration A in the "Hirota–Korteweg–de Vries equation" described in Ref. 6, but this is only a calculational tool and does not appear in the final answer (2.1).]

The wavenumbers k_1 and k_2 can be arbitrary; Novikov[13] has emphasized from his earliest papers that if the wavenumbers are incommensurable, i.e., if k_1/k_2 is an irrational number, then the double cnoidal will be "almost periodic" in space rather than strictly periodic, but this is mathematically legitimate. Although some applications of "spatial almost periodicity" can be envisaged,[14] it is sufficient for most physical problems to take $k_1 = 1$ and $k_2 = 2$. The reasons are that (i) in most Fourier series, the second harmonic ($k=2$) is the largest component after the fundamental ($k=1$), and (ii) one can change the spatial period from unity [as in (2.3) with $k_1 = 1$] to an arbitrary period through a trivial rescaling of the coordinates. The spatial period is equal to one in all the figures and cases described in the rest of this paper.

The diagonal theta matrix elements are thus the more important parameters because they specify the amplitude of the two waves that make up the double cnoidal wave. Figure 1 indicates the different wave regimes of the $T_{11} - T_{22}$ plane. When T_{11} and T_{22} are both large, the double cnoidal wave is approximately equal to the sum of two linear, noninteracting sine waves of different wavenumbers and phase speeds, i.e.,

$$u(x,t) = -48\pi^2[k_1^2e^{-T_{11}}\cos(2\pi X) + k_2^2e^{-T_{22}}\cos(2\pi Y)]. \tag{2.6}$$

When both T_{11} and T_{22} are small, the double cnoidal wave is approximately given by the usual Korteweg–de Vries double solitary wave with one tall soliton and one short soliton on each unit interval in x. The Fourier series (2.3) converges very slowly for small T_{11} and T_{22}. The central theme of the author's previous paper[1] is that one should substitute instead the series

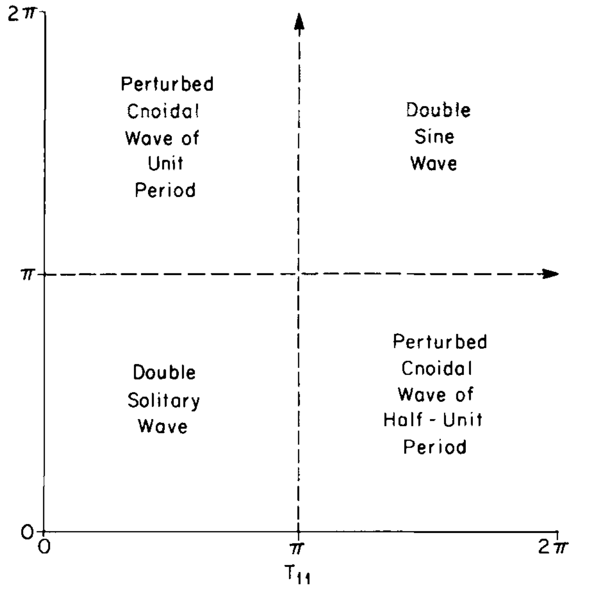
2π
π
0
0
π
2π
T11
Perturbed Cnoidal Wave of Unit Period
Double Sine Wave
Double Solitary Wave
Perturbed Cnoidal Wave of Half-Unit Period

FIG. 1. Schematic diagram showing the four main regimes of the double cnoidal wave in $T_{11} - T_{22}$ plane, where T_{11} and T_{22} are the diagonal theta matrix elements, which are always positive.

$$\theta = \sum_{\substack{n_1 = -\infty \\ [\text{half-integers}]}}^{\infty} \sum_{n_2 = -\infty}^{\infty} \exp\left(- \left\{ \left(\frac{R_{11}}{2}\right)(X + n_1)^2 + R_{12}(X + n_1)(Y + n_2) + \left(\frac{R_{22}}{2}\right)(Y + n_2)^2 \right\} \right), \tag{2.7}$$

where the sums are over the half-integers, $\pm \frac{1}{2}, \pm \frac{3}{2}, \pm \frac{5}{2}, \ldots$, and where the R_{ij} are proportional to the elements of the inverse of theta matrix formed by the T_{ij}. For obvious reasons, (2.7) will be referred to as the "Gaussian" series of the theta function since each term is a Gaussian function of X and Y; this series is the Poisson sum of the Fourier series. As explained in Appendix B of Ref. 6, the usual double solitary wave can be obtained from (2.7) by truncating it to four terms and taking the second logarithmic derivative as in (2.1), but the result is too messy to repeat here.

The strength of using two alternative series representations, (2.3) and (2.7), is that the Fourier series converges rapidly in the double sine wave regime where (2.7) converges slowly, while the Gaussian series converges rapidly in the double soliton regime where the Fourier series is almost useless. Consequently, in this paper and its two companions, we shall move from Fourier series to Gaussian series and back again with great freedom. As explained in Ref. 6, the mechanics of calculating the unknown phase speeds and diagonal theta matrix element (either T_{12} or R_{12}) are such that the Fourier-based computation is merely a special case of that for the Gaussian series.

Unfortunately, neither series is rapidly convergent along the T_{11} and T_{22} axes where one diagonal theta matrix element is large in comparison to the other, but this is not of vital importance because these near-axis regimes represent a single solitary wave perturbed by a very small amplitude sine wave. As such, these regimes are much less interesting than those in which the two waves are of equal amplitude since theories for the single soliton subject to an arbitrary perturbation have been developed by R. Grimshaw[15] and others he references. In practice, there is actually a high degree of overlap between the Fourier and Gaussian series both with each other and with the perturbed one-soliton regimes, so the need for special methods for these near-axis double cnoidal waves is usually academic.

The double solitary wave regime is the most interesting case of all. In Sec. IV, the geometry of the X-Y plane is deduced from the Gaussian series. To some extent, this will merely repeat the construction given in Ref. 1 for the single cnoidal wave, but it will also bring out several features such as phase shifts and the special modular transformation which are unique to polycnoidal waves with $N \geqslant 2$, and have no counterpart for the ordinary $N = 1$ cnoidal wave. First, however, some sample graphs are presented to give the reader a feeling for each of the four regimes of the double cnoidal wave.

III. SAMPLE DOUBLE CNOIDAL WAVES

Figures 2–5 illustrate $u(x,t)$ for each of the wave regimes indicated schematically in Fig. 1. The graphs were computed in a frame of reference moving at the phase velocity c_1 so that the tallest peak is approximately stationary; in this frame of reference, the double cnoidal wave is simply periodic in time, so it suffices to show half of one temporal period. Strictly speaking, the double cnoidal wave solution has a mean value of 0, i.e., the integral of $u(x,t)$ over a period is 0, but for visual clarity, a constant[16] has been added to the graphs.

The first case is that of a classic double solitary wave: The tall soliton overtakes the short soliton and only a single peak is visible at the time of maximum interaction. In time, however, the two separate and emerge unchanged by their interaction except for a shift of phase. In other words, the tall peak is briefly accelerated and the short peak briefly deaccelerated by their encounter so that the tall soliton is farther to the right than it would have been in the absence of the collision. In a spatially unbounded domain, where there are just the two solitons on the whole interval $x \in [-\infty, \infty]$, this col-

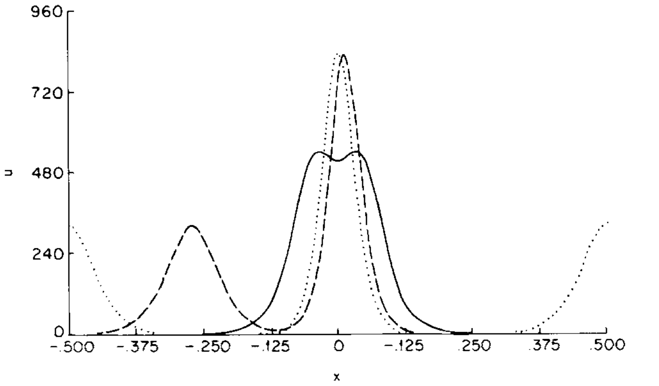

FIG. 2. A Korteweg–de Vries double cnoidal wave in the double soliton regime. The mode in this and the next three figures is $[1,2]^P$ or equivalently, $\{1,1\}^P$, in the notation defined in Sec. V. The angle variable X, defined by (2.4) was set equal to x, the spatial coordinate, for all curves so that we are looking at the wave in a frame of reference moving with the phase speed, c_1. The double cnoidal wave is simply periodic in time in this reference frame with a period $P = 1/c_2$. Solid curve ($t = 0$), dashed curve ($t = P/4$), and dotted curve ($t = P/2$) show one half of a time period. $T_{11} = 0.397$, $T_{12} = 0.359$, and $T_{22} = 0.892$ (with $k_1 = 1$ and $k_2 = 2$, here and in the next three figures).

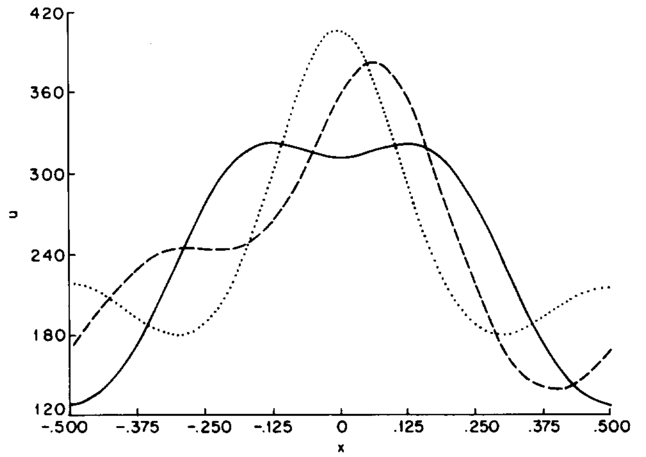

FIG. 3. Same as Fig. 2 except that the polycnoidal wave is in that intermediate parameter range where it can be regarded (and accurately approximated) as either a pair of linear sine waves or a pair of solitary waves. Solid curve ($t = 0$), dashed curve ($t = P/4$), and dotted curve ($t = P/2$), where P is the time period.

lision is a once-in-a-lifetime event, and therefore does not affect the average speed of the solitons. On the periodic domain, the collision is repeated endlessly, so the repeated phase shifting does alter the average phase speed of the solitons. The implications of this are discussed in the next section and more particularly in Sec. V.

Figure 3 shows the double cnoidal wave when both peaks are much smaller and wider. The parameter values are such that the polycnoidal wave lies in that intermediate regime where it can be equally well considered to be a solitary wave or a pair of sine waves: both lowest-order approximations agree with the exact solution to within a few percent of accuracy. The qualitative behavior is very similar to that of the extreme double soliton case shown in Fig. 2, and can likewise be interpreted as colliding solitary waves. The alternative sine wave interpretation is equally straightforward.[17] At $t = 0$ (solid curve, Fig. 3), a trough of the second harmonic is 180 degrees out of phase with the wavenumber one component at $X = 0$. The result is a dimple at $X = 0$, where the peak of the fundamental is partially cancelled by a trough of the second harmonic, two peaks on either side of the origin near nodes of the second harmonic, and very deep troughs at $X = \pm\frac{1}{2}$, where both the fundamental and harmonic have negative maxima. When the second harmonic has moved a quarter unit in X (dotted curve), there is a single tall, narrow peak at $X = 0$ where the fundamental and second harmonic are in phase, and smaller secondary peaks at $X = \pm\frac{1}{2}$ where the narrow crests of the second harmonic rise from the flatter troughs of the fundamental.

Figure 4 illustrates the rather boring case of a single soliton modified by a small superharmonic (wavenumber two) perturbation ($T_{22} \gg T_{11}$, where $k_2 = 2\,k_1$). Lax has shown[18] that when the two solitons are sufficiently unequal in size, the tall soliton becomes shorter and broader during the collision (i.e., while out of phase with the crest of the perturbation) but the dimple at or near $X = 0$ (so that Figs. 2 and 3 always have two local maxima) does not occur so that there is only a single local maximum for part of each period in time.

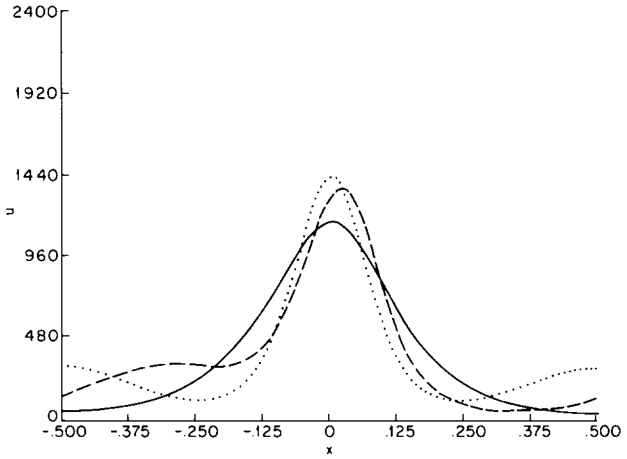

FIG. 4. Same as Figs. 2 and 3 except that the polycnoidal wave is actually a weakly perturbed ordinary cnoidal wave. Solid curve ($t = 0$), dashed curve ($t = P/4$), and dotted curve ($t = P/2$), where $T = 1/c_2$ is the time period. $T_{11} = 1.00$, $T_{12} = 0.759$, and $T_{22} = 3.00$.

Figure 5 shows the other perturbed soliton regime ($T_{11} \gg T_{22}$). This is a cnoidal wave of half-unit spatial period weakly affected by a subharmonic perturbation of unit period. For clarity, a slightly different convention was used than with the preceding three figures: Instead of keeping the phase of X fixed while advancing that of Y by a half unit, the phase of X was decreased by 0.25 while that of Y was increased by 0.25 to trace out half a time period so that the peaks are quasistationary in the graphical frame of reference.

The twin crests of the cnoidal wave do not merge under the influence of the perturbation, but instead execute a small oscillation about their mean positions. This is perfectly consistent with interpreting this case as the collision of two solitons that differ slightly in amplitude. Lax[18] has shown that, in the words of Fornberg and Whitham,[19] "there are always two maxima; the wave approach each other and exchange roles, but then shear away and do not pass through each other." Another way to look at this to examine the dimple at $x = 0$ at the time of the maximum soliton overlap in Fig. 2. As the ratio of the amplitude of the two solitons becomes closer and closer to 1.0, this local minimum at $x = 0$ be-

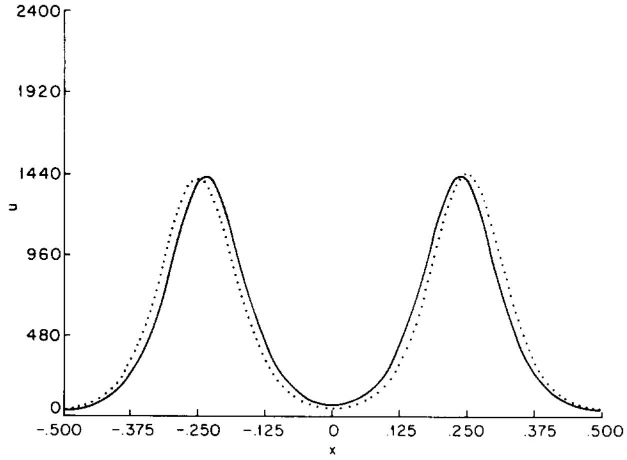

FIG. 5. Same as Figs. 2–4 except that the polycnoidal wave is a simple cnoidal wave of half-unit period subject to a weak perturbation of unit spatial period. For clarity, a different frame of reference was used such that the phase of the angle variable X was decreased by 0.125 between graphs while that of Y was increased by the same amount. Solid curve ($\phi_1 = 0$, $\phi_2 = 0$), dashed curve ($\phi_1 = -0.125$, $\phi_2 = 0.125$), and dotted curve ($\phi_1 = -0.25$, $\phi_2 = 0.25$). $T_{11} = 3.00$, $T_{12} = 0.851$, and $T_{22} = 1.811$.

comes deeper and deeper until the two solitons are separated by a wide, deep trough even at the time of closest approach.

One can also interpret Fig. 5 in terms of constructive and destructive interference between two periodic waves of different phase speeds. Although not obvious on the graph, the right peak in Fig. 5(b) is in fact slightly taller than the left peak as a result of constructive interference at $x = 0.25$ with the crest of the $k_1 = 1$ component while the left soliton is shrunk a bit because it rests on the trough of the perturbation at $x = -0.25$. As the perturbation continues to move relative to the tall peaks, it will reinforce and weaken each large crest in turn. Thus, one has two alternative interpretations of this case that lead to the same conclusions: (i) two colliding solitary waves of almost identical amplitude on each periodicity interval, or (ii) a simple cnoidal wave of half-unit period whose crests swell and accelerate or shorten and slow down as the crests and troughs of the sine wave perturbation move through them.

IV. THE GEOMETRY OF THE X-Y PLANE

Although the samples of the preceding section illustrate the general characteristics of double cnoidal waves, there are some important, but subtle, aspects of polycnoidal waves which can be explained only by examining $\theta(X,Y)$ and its relation to $u(x,t)$. As noted in Ref. 1, a heuristic way of constructing a polycnoidal wave is to simply repeat the usual multiple soliton solution over the whole x-axis. The resulting approximation is obviously periodic, but generally is not an exact[20] solution of the Korteweg–de Vries equation.

Boyd[1] shows, however, that Hirota's transformed single solitary wave solution,

$$F = 1 + \exp(2sX)\,, \tag{4.1}$$

which gives the usual hyperbolic secant squared soliton upon taking the second logarithmic derivative, can be generalized to a "bi-Gaussian"

$$\begin{aligned}\Theta(x,t) = {}& \exp[-s(X-\pi/2)^2/\pi] \\ & + \exp[-s(X+\pi/2)^2/\pi]\,. \end{aligned} \tag{4.2}$$

If one repeats (4.2) over the whole interval, one obtains the Gaussian series of the one-dimensional theta function, which is an exact solution of the Hirota–Korteweg–de Vries equation, and therefore generates an exact solution of the KdV equation upon taking the second logarithmic derivative. Figure 6, which is borrowed from Boyd,[6] illustrates the procedure. The shape of the polycnoidal wave is determined by the theta function; the only remaining unknown (for the ordinary cnoidal wave) is to solve a pair of algebraic equations to determine the nonlinear phase speed c_1 in the "angle" variable X.

The same concept applies for higher polycnoidal waves. In particular, a "tetra-Gaussian" consisting of four Gaussian functions of identical shape but with peaks located at the four corners of a unit square ($X = \pm 0.5$, $Y = \pm 0.5$) gives the usual double soliton of the KdV equation on an infinite domain in x. (A proof is given in Appendix B of Ref. 6.) When this tetra-Gaussian is repeated with unit spacing over the whole of the X-Y plane, it generates the Gaussian series of the theta function.

In the near-double soliton regime (small T_{11}, T_{22} or equivalently, large R_{11} and R_{22}), the Gaussians are sharply peaked so that the full infinite series can be approximated on the unit square by the sum of the four Gaussians whose peaks are at its corners. The reason that it is not possible to approximate the series by a single Gaussian is that $u(x,t)$ is obtained by taking the second logarithmic derivative, which for a single Gaussian would be $u(x,t) = \text{const}$. The solitons actually lie in the valleys between the peaks of the Gaussians, and the center of the square where the two valleys meet is also where the solitons collide.

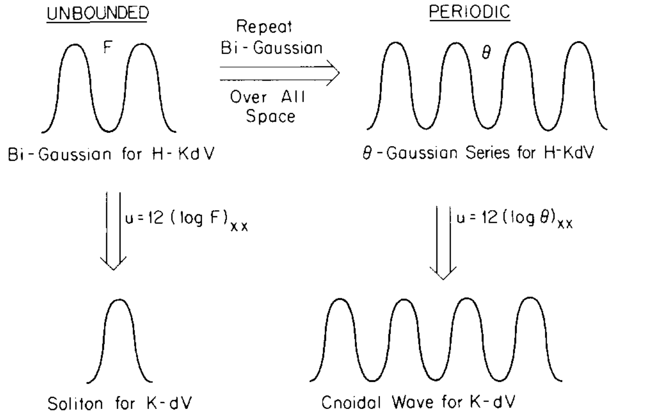

FIG. 6. Schematic diagram showing the relationship between the bi-Gaussian and theta function solutions to Hirota's transformed version of the KdV equation. The left side shows the situation when the domain is unbounded: The solution to the transformed KdV equation has just two peaks on all of $X\in[-\infty,\infty]$, and the second logarithmic derivative of this gives a single crest (corresponding to the valley betwen the two peaks of the bi-Gaussian) which is the usual solitary wave. When the bi-Gaussian pattern is repeated with even spacing over all X, it generates the Gaussian series of the theta function. This, as shown on the right, is a spatially periodic solution of the transformed KdV equation and its second logarithmic derivative is the simple ($N = 1$) cnoidal wave. [Taken from Boyd[1].] For the double cnoidal wave, the basic unit is a tetra-Gaussian with peaks at the four corners of a unit square in the X-Y plane which generates the double solitary wave when the domain is unbounded. The idea is the same, however, repeating this basic unit over all of X-Y space with even spacing gives a periodic solution to the transformed KdV equation whose second logarithmic derivative with respect to x is the double KdV cnoidal wave.

Figure 7 shows the graph of the theta function in the X-Y plane with the contours of the function

$$\begin{aligned} U(X,Y) = 12\{ & k_1^2(\log\theta)_{XX} + 2k_1k_2(\log\theta)_{XY} \\ & + k_2^2(\log\theta)_{YY} + \alpha\} \end{aligned} \tag{4.3}$$

also plotted. (The constant α has been added so that the solitons asymptote to 0, as in Figs. 2–5.) The function $u(x,t)$ which actually solves the KdV equation is obtained from $U(X,Y)$ by drawing a line of slope k_2/k_1 through the origin ($X = 0$, $Y = 0$). The values of $U(X,Y)$ along this line then give the values of $u(x,t=0)$. The function $u(x,t)$ is obtained at later times by moving the line with the velocity $-c_1$ in X and $-c_2$ in Y consistent with the definitions (for $k_1 = k_2 = 1$)

$$X = x - c_1t, \quad Y = x - c_2t. \tag{4.4}$$

[The reason for the minus signs is so that $u(x=0,t) = U(-c_1t, -c_2t)$ and similarly for other x to agree with (4.4).]

If the solitary waves collided without a shift of phase, then (i) the theta matrix and inverse theta matrix would be diagonal, i.e., $T_{12} = R_{12} = 0$; (ii) the ridges of $U(X,Y)$ would be parallel to the X and Y axes. In reality, however, there is a

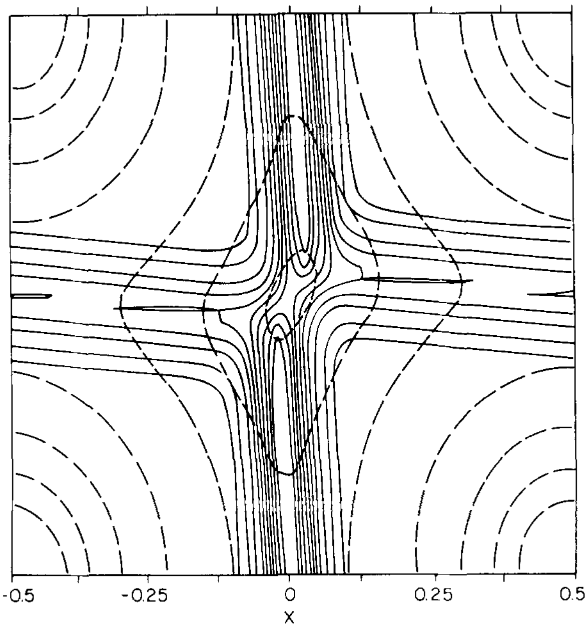

FIG. 7. Contours of the two-dimensional theta function (dashed lines) and of $U(X,Y)$ in the unit square whose corners are $X = \pm 0.5$ and $Y = \pm 0.5$. This is in the double solitary wave regime; the Gaussian series was used with $k_1 = k_2 = 1$, and $R_{11} = 50$, $R_{12} = 2.913$, and $R_{22} = 30$. The function $u(x,t)$ for this case is shown in Fig. 2.

phase shift of both solitary waves after the collision—the taller soliton is temporarily accelerated while the shorter one is deaccelerated during their encounter—so the ridges of $U(X,Y)$ are tilted with respect to the axes. The magnitude of the slope is given in Appendix C of Ref. 6 along with other formulas describing the contours of $U(X,Y)$ and so on, but the mere fact of the slope is enough to show one rather startling fact: The phase velocities c_1 and c_2 are not the speeds at which the solitons travel when outside the collision region.

In the next section, the reason will be discussed in detail. In brief, one concludes that c_1 and c_2 represent the average velocities of the two solitary waves, and these averages are changed from the usual noncolliding soliton speeds because of the phase shifts that occur during the collision. When the spatial domain is unbounded and there are but two solitons, the collision occurs but once. With spatial periodicity, the collisions recur endlessly and the average speed of the solitons is altered. Before turning to this, however, we must first explore the role of wavenumbers.

Figure 7, which shows a unit square in the X-Y plane, implicitly assumes $k_1 = k_2 = 1$. When $k_2 = 2$, however, Y varies by 2 when x varies by 1. Thus, for $k_1 = 1$ but $k_2 = 2$, the whole of the rectangle shown in Figure 8 projects on a unit interval in x. The line which takes $U(X,Y)$ to $u(x,t)$ now has a slope of 2, and the reader can see (by laying a ruler between the lower left and upper right corner) that for part of each temporal period, there are three solitons on each unit interval in x: one tall solitary wave and two short solitary waves. Figure 9 shows $u(x,t)$ for the same wave as in Fig. 8. Thus, the wavenumbers are extremely important in determining the qualitative nature of the flow, and Sec. VI will examine that role in detail.

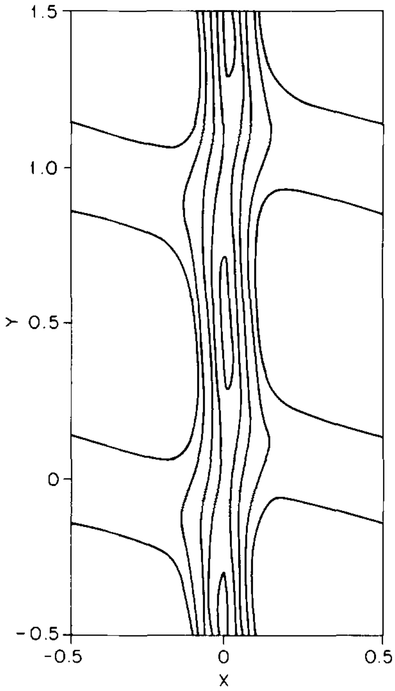

FIG. 8. Contours of $U(X,Y)$ in the rectangle whose corners are $X = \pm 0.5$ and $Y = 0.5, + 1.5$ for $k_1 = 1$ but $k_2 = 2$. When converted from X and Y to the actual spatial coordinate X, all of this rectangle projects onto a unit interval in x. $R_{11} = 32$, $R_{12} = 2.20$, and $R_{22} = 8$. The corresponding $u(x,t)$ is shown in Fig. 9.

V. PHASE SPEEDS AND SOLITON VELOCITIES

As shown in Ref. 6, the overlap of the solitons on one unit periodicity interval in x with those of another creates corrections to c_1 and c_2 which can be calculated as a double perturbation series in the parameters $\exp(-R_{11})$ and $\exp(-R_{22})$. Since the solitons decay exponentially with x [as $\exp(-R_{11}|x|)$ and $\exp(-R_{22}|x|)$], it follows that these "overlap" corrections decrease exponentially with the half-widths of the solitary waves. The differences between c_1 and c_2 and the velocities of the solitons, however, decrease only linearly with the widths of the solitons, and are therefore

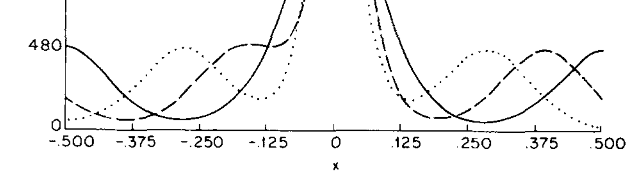

FIG. 9. The KdV solution $u(x,t)$ for the wave whose theta function is plotted in Fig. 8. As with Figs. 2–4, the phase of X is kept fixed so that we view $u(x,t)$ in a frame of reference moving with the phase velocity c_1. In this reference frame, the wave is periodic in time with a period $P = 1/c_2$. Solid curve ($t = 0$), dashed curve ($t = P/4$), and dotted curve ($t = P/2$).

something quite different in nature.

One proof of this comes from the observation that the slopes of the ridges of $U(X,Y)$, which are responsible for making the phase and soliton velocities differ, are given by $-R_{11}/R_{12}$ and $-R_{12}/R_{22}$, respectively, as shown in Appendix C of Ref. 6. Since R_{12} remains $O(1)$ when R_{11} and R_{22} become large, it follows that the slopes of the soliton ridges in the X-Y plane become increasingly parallel to the Y and X axis, respectively. The angles between the solitons and the axes, however, are linear functions of $1/R_{11}$ and $1/R_{22}$ while the "overlap" corrections, i.e., the higher-order terms in the perturbation series of Ref. 6, are decreasing exponentially in these same variables.

A more direct way is to simply calculate these quantities to zeroth order in perturbation theory, which is equivalent to truncating the infinite theta function series to the minimum of four Gaussian functions needed to generate the double solitary wave. It is shown in Ref. 6 that the phase velocities c_1 and c_2 that appear in the "angle" variables X and Y are obtained from the "pseudofrequencies" ϵ_1 and ϵ_2 by solving the pair of linear equations

$$\begin{vmatrix} (-R_{11}k_1) & (-R_{12}k_2) \\ (-R_{12}k_1) & (-R_{22}k_2) \end{vmatrix} \begin{vmatrix} c_1 \\ c_2 \end{vmatrix} \begin{matrix} = \\ = \end{matrix} \begin{vmatrix} \epsilon_1 \\ \epsilon_2 \end{vmatrix} . \tag{5.1}$$

To lowest order

$$\epsilon_i = -c_i^{\rm sol}\,\delta_i, \quad i = 1,2, \tag{5.2}$$

where

$$\delta_i \equiv R_{ii}\,k_i + R_{12}\,k_j, \quad i = 1,2, \; j \neq i \tag{5.3}$$

gives the width of each soliton and where $c_i^{\rm sol}$ is the "free" velocity of a soliton, i.e., the speed at which the soliton travels when not in collision with another. When "free," $u(x,t) \simeq 3\delta_i^2 \,\mathrm{sech}^2[\delta_i(x - c_i^{\rm sol}\,t)]$ in the neighborhood of the ith soliton. If we add a constant[16] to $u(x,t)$ and the phase speeds so that the solitons asymptote to 0 for large x—the result is still a polycnoidal wave solution of the KdV equation—then

$$c_i^{\rm sol} = \delta_i^2, \quad i = 1,2 \tag{5.4}$$

which is the usual formula as given in Whitham,[21] for example, although he uses κ in place of our δ.

Through elementary algebra, one can show from (5.1) through (5.4) that

$$c_1 = c_1^{\rm sol} + \frac{(c_1^{\rm sol} - c_2^{\rm sol})\,R_{12}\delta_2}{k_1(R_{11}\,R_{22} - R_{12}\,R_{12})} . \tag{5.5}$$

In the extreme soliton regime (R_{11}, $R_{22} \gg 1$), R_{11}, $R_{22} \gg R_{12}$, which permits (5.5) to be simplified to

$$c_1 = c_1^{\rm sol} + (c_1^{\rm sol} - c_2^{\rm sol})\,k_2\{R_{12}/\delta_1\}. \tag{5.6}$$

Now it can be shown (Whitham[21] and Appendix C of Ref. 6) that the phase shift experienced by a soliton of amplitude determined by R_{11} (which we shall call "type 1" for short) after collision with a soliton of the other size is (R_{12}/δ_1), so (5.7) implies, reasonably enough, that the difference between the "free" speed of the soliton and the corresponding phase velocity in X is proportional to this phase shift—which argues strongly that it is the phase shift that is the cause of this difference. If this explanation is correct, however, then (5.6) should also depend upon the frequency with which a soliton of type 1 collides with a soliton of type 2. Since k_2 determines the number of solitons of type 2 per unit interval in x, it follows that $k_2(c_1^{\rm sol} - c_2^{\rm sol})$ is the frequency with which a soliton of type 1 will collide with a soliton of the other size per unit time. The wavenumber k_1, which determines the density of type 1 solitons per unit interval of x, is conspicuously missing from (5.6); it has no bearing on the number of collisions between a particular soliton of type 1 and all the solitons of the other height because a type 1 soliton collides only with the solitary waves of the other amplitude. Thus, (5.6) can be rewritten schematically as

$$c_1 = c_1^{\rm sol} + \{\text{number of collisions/unit time}\}\{\text{phase shift/collision}\} \tag{5.7}$$

and similarly for c_2.

Thus, as mentioned earlier, c_1 and c_2 may be properly interpreted as the average speeds of the solitary waves while their instantaneous speeds (outside collision zones) are given by the different quantities $c_1^{\rm sol}$ and $c_2^{\rm sol}$.

VI. WAVENUMBERS AND THE SPECIAL MODULAR TRANSFORMATION

The wavenumbers k_1 and k_2 have different roles in the double-sine wave and double-soliton regime. In the near-linear regime, k_1 and k_2 are the actual wavenumbers of the two sinusoidal, noninteracting waves that approximate the polycnoidal wave. In the double-soliton regime, the widths of the solitary waves are given by the "pseudowavenumbers" defined by (5.3) above, and k_1 and k_2 instead give the number of solitons on each interval. This was shown explicitly by Figs. 8 and 9 in Sec. IV, where a double cnoidal wave with three solitons on each unit interval was displayed. Since $R_{11} > R_{22}$ for this case and k_2 was the wavenumber equal to two, the pair of identical solitons was shorter than the third, but one could mix two tall solitons with a single shorter one on each unit interval by either choosing $k_1 = 2$ instead or taking R_{22} larger than R_{11}. More exotic combinations are possible and it will be argued in the next section that Hyman[22] computed a double cnoidal wave with four solitary waves on each spatial period, three tall and one short.

This all seems rather straightforward, but in reality the issue of wavenumbers is so complicated as to demand an entire separate article unto itself (Ref. 7). The Serpent in Eden is that the different roles assigned to the wavenumbers for solitons and sine waves are contradictory. Figures 7–9 show clearly that the usual situation of two solitons of unequal size per unit interval in x demands $k_1 = k_2 = 1$, but in the sine wave regime, this is absurd because the linear dispersion relation demands that two infinitesimal amplitude waves of the same wavenumber must also have the same phase speed, and the double cnoidal wave collapses into the ordinary single cnoidal wave. The simplest possibility that preserves two distinct phase speeds and "phase" variables and is a true double cnoidal wave is to take $k_2 = 2k_1$, i.e., one wave is the second harmonic of the other.

The resolution of this difficulty lies in a remarkable fact that at first seems only to put us into more trouble: Each theta function of two or more dimensions can be written in a denumerable infinity of ways via the so-called "special modular transformation" which is the central theme of Ref. 7. The theta matrices and wavenumbers are transformed by matrices whose elements are integers so that the equivalent representations of a theta function with integral wavenumbers are restricted to those for which the new wavenumbers are integers also.

Physically, of course, there is no ambiguity at least in the limits of very large or very small wave amplitudes: In the double-soliton regime, there is only one representation for which the wavenumbers give the actual density of solitary waves on the unit interval and the phase speeds of the phase variables are the average velocities of the solitons, and in the double-sine wave regime, there is again only one way of writing the theta function in which the wavenumbers and phase speed of its arguments X and Y are the actual wavenumbers and phase speeds of the two sine waves. The special modular transformation is thus a way of providing the theta function with a mathematical disguise which alters the arguments and parameters of the theta function without altering the Korteweg–de Vries solution which it generates. It would be quite foolish, however, to dismiss the modular transformation as a mere mathematical curiosity.

In the first place, it implies that the nonlinear implicit dispersion relation given in Ref. 6, which must be solved to determine c_1, c_2, and the diagonal theta matrix element, has nonunique solutions. (In fact, an infinite number of them.) Some care is needed to insure that one computes in the "physical" representation so that the phase speeds computed are those of the actual components of the polycnoidal wave being sought, and not merely mathematical disguises for something quite different.

In the second place, the special modular transformation resolves the dilemma of needing different wavenumbers to make sense of the simplest double-soliton and double-sine wave regimes. If one solves the residual equations by varying the diagonal theta matrix elements in small steps, the so-called "continuation" method, one finds upon graphing $u(x,t)$ that the mode which is the sum of one sine wave with $k_1 = 1$ plus another with $k_2 = 2$ does indeed smoothly continue into a pair of solitary waves, one tall and one short, on each unit interval. The phase speeds so computed, however, are not those of the actual solitons, but can be made into them by taking that modular transformation which reduces the wavenumber from $k_2 = 2$ to $k_1 = 1$. In a similar way, if one begins with the double soliton for $k_1 = k_2 = 1$ and marches in the opposite direction of decreasing amplitude, the phase speeds computed from the residual equation will not be those of the sine wave and its second harmonic that dominate $u(x,t)$ when the amplitude is small, but can be changed into the physical wave speeds through the modular transformation that sends k_2 from 1 to 2. The whole business is discussed thoroughly with numerical tables in Ref. 7.

The modular transformation makes it necessary to introduce some notation. A pair of numbers written in square brackets, for example, [1,2], is used to denote the wavenumbers of the Fourier representation with k_1 written first. A superscript "P" can be added to denote that the "physical" representation is meant and not one of the infinite number of disguises allowed by the mathematics. (When there is no danger of confusion, the superscript P will be omitted; when this notation is used elsewhere in this series of papers, the "physical" representation will always be meant unless expressly stated otherwise.) In a similar way, curly brackets, i.e., $\{1,1\}$ will be used to denote the wavenumbers of the Gaussian series of the theta function. The author apologizes for burdening physics with more notation, but it is unavoidable. It is necessary to introduce separate notation for the Fourier and Gaussian series because

$$[1,2]^P = \{1,1\}^P. \tag{6.1}$$

In words, the mode which is the sum of a wave and its second harmonic for small amplitude is the sum of one tall and one short solitary wave for large amplitude.

Reference 7 goes on to describe in some detail the identifying characteristics of the "physical" representation. First, it is that for which the off-diagonal theta matrix element is small in comparison to the diagonal theta matrix elements. Second, it is the representation employed by the perturbation series of Ref. 6—the perturbation series always give answers in the "right" representation, in other words. The perturbation series suggest T_{12} and R_{12} are always positive, so a representation in which either of these off-diagonal elements is negative is almost certainly not the physical representation.

Finally, one can give a graphical definition. Figure 10 compares $U(X,Y)$ for two different $\{1,2\}$ modes. The left panel is simply a repeat of Fig. 8; the corresponding $u(x,t)$ is given by Fig. 9 and truly has three solitary waves on each unit interval in x. The right panel, however, is in an unphysical representation. Notice that the repeated soliton ridges have a steep positive slope rather than a shallow negative slope as in the left panel. The reason is that R_{12} is large and negative instead of being small and positive as it should be.

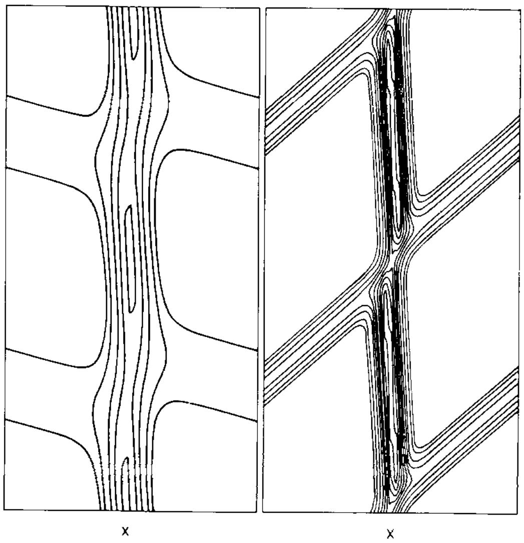

FIG. 10. Contours of $U(X,Y)$ for two theta functions with $k_1 = 1$ and $k_2 = 2$. (a) [left panel] This is identical with that shown in Fig. 8; this choice of wave numbers is the physical representation of this wave, so this mode is denoted $\{1,2\}^P$. (b) [right panel] This is actually a $\{1,1\}^P$ mode in disguise with $R_{11} = 74.17$, $R_{12} = -27.09$, and $R_{22} = 30$. Although (b) looks quite different from Fig. 7, they are plots of the same theta function in different representations; when the function shown in (b) is converted back into (x,t) coordinates, the resulting $u(x,t)$ is that shown in Fig. 2.

By laying a ruler across the figure at a slope of k_2/k_1, i.e., 60 degrees, one can convince oneself that even though one wavenumber is 2, there are in fact no more than two solitons present at any time. The actual $u(x,t)$ for Fig. 10(b) is in fact that graphed in Fig. 2.

VII. PREVIOUS CALCULATIONS OF DOUBLE CNOIDAL WAVES

Although there have been a huge number of abstract, theoretical papers on polycnoidal waves, there have been only two explicit attempts to calculate and graph KdV polycnoidal waves before this present work. Both have limitations which illustrate the usefulness of the ideas developed in the two companion papers (Refs. 6 and 7).

Hyman[22] used a variational principle of Lax' to numerically calculate a number of case studies of double cnoidal waves, although only one is described in detail in his paper. By carefully computing the trajectory of the maxima, he showed "the peaks move with two distinct speeds. In any spatial period three of the peaks are traveling with one speed while the fourth is traveling faster." This inspired the remark by other researchers,[23] "The general shape [of $u(x,t)$] is still obscure, though a large body of numerical information has been obtained by J. M. Hyman; for example, he finds that for $N = 2$, the number of peaks and valleys is usually 4 and on occasion 5." The case illustrated in Hyman's own paper has 4 peaks and 4 valleys.

In light of what has been presented earlier here, it is difficult to escape the conclusion that Hyman actually computed only double cnoidal waves with the physical representation $\{1,3\}^P$, i.e., four solitons on each unit interval with three of one size and a fourth of another, and missed the $\{1,2\}^P$ or $\{1,1\}^P$ modes. Figures 2 through 5 show clearly that the conclusion that the "number of peaks and valleys is usually four" is nonsense; the $\{1,1\}^P - [1,2]^P$ mode has only two peaks and two valleys, sometimes less. The conclusion would seem to be that Lax' variational principle combined with numerical nonlinear optimization is a poor way to investigate polycnoidal waves.

Hyman's paper is still of interest, however, because he superimposed random perturbations upon his double cnoidal waves and found them to be remarkably stable. It seems probable that this is true of all polycnoidal waves, but a proof is lacking, and Hyman's paper is at present the only evidence in favor of this hypothesis.

Hirota and Ito[8] have computed a double cnoidal wave by numerically solving the implicit dispersion relation. Table I gives their results in their original notation, translates their results into the notation used here, and then compares the results with the Fourier and Gaussian perturbation series derived in Ref. 6. The result is a rather resounding triumph for perturbation theory: The second-order Fourier series gives all three physically significant unknowns to within 4% relative error while the zeroth order Gaussian series, i.e., the tetra-Gaussian double soliton, gives these same three quantities to within 4% error also. The conclusion is that number crunching is not really necessary: for most purposes, the perturbation series of Ref. 6 are more than adequate.

TABLE I. A comparison of the numerical calculations of a double cnoidal wave from Hirota and Ito[8] with Fourier and Gaussian perturbation theory. The first line of the table gives the numerical results of Hirota and Ito in their own notation. The second line gives the same exact solution in terms of the notation and conventions employed here. (Their theta matrix elements must be multiplied by π, their constant of integration λ divided by -2 to give my A, and their frequencies converted into phase speeds by multiplying by $-1/k_1$. Because I normalize k_1 to 1, it is also necessary to multiply the phase speeds by 6.25^2 and A by 6.25^4 to increase the wavenumbers by a factor of $6.25 = 1/0.16$.) The third part of the table gives the results of Fourier perturbation theory; because of the smallness of the nome $q_2 \sim q_1^2$, the terms in q_2^2 were neglected in computing the first-order solution and q_2^4 in the second-order solution. Relative errors are given in square brackets. The fourth part of the table gives the results of Gaussian perturbation theory for $R_{11} = 14.38$, $R_{22} = 6.478$, which correspond to the T_{11} and T_{22} values employed in the rest of the table. Normally, it would be necessary to determine these R_{11} from the corresponding T_{11} through some kind of iterative procedure as explained in the text.

Hirota–Ito Notation

k_1	k_2	τ_{11}	τ_{22}	λ	ω_1	ω_2	τ_{12}
0.16	0.32	0.464	1.16	-2.01	-0.086	1.23	0.297

Boyd Notation

k_1	k_2	T_{11}	T_{22}	A	c_1	c_2	T_{12}
1.0	2.0	1.458	3.64	1 533	21.00	-150.2	0.933

Fourier Perturbation Theory

	A	c_1	c_2	T_{12}
0th order	0 [100%]	-39.5 [300%]	-157.9 [5.2%]	1.099 [17.8%]
1st order	1 013 [34.5%]	11.8 [43.5]	-157.9 [5.2%]	0.936 [0.27%]
2nd order	1 547 [0.89%]	20.2 [3.8%]	-150.4 [0.15%]	0.933 [<0.1%]

Gaussian Perturbation Theory

	A	c_1	c_2	R_{12}
0th order	1 443 [5.8%]	20.3 [3.6%]	-149.0 [0.79%]	2.335 [0.56%]

Their paper, however, is of further interest because it also computes a triple cnoidal wave. This has only seven unknowns but there are *eight* residual equations. Knowing from the "Hill's spectrum method" that theta function solutions should exist, they boldly chose seven of the eight equations and solved them as a closed system, and then verified after the fact that the extra equation was also satisfied to within machine precision. It would be extremely interesting to have an analytical proof of the redundancy of the residual equations for $N = 3$ and higher, as opposed to their numerical proof, but none is yet known.

Thus, although the analysis of Refs. 6 and 7 makes it possible to improve on these early, limited calculations by Hyman and by Hirota and Ito, both papers are still valuable for their intelligent use of numerical solutions to suggest as yet unproven theorems for the future.

VIII. THE DOUBLE CNOIDAL WAVE IN PERSPECTIVE

The methods employed here and in Refs. 1, 6, and 7 can be extended, with a few additional tricks, to most or all of the "exactly integrable," soliton-admitting equations which are now known to be solvable via theta functions through the "Hill's spectrum" method. The Korteweg–de Vries equation is one of several whose Hirota-transformed equivalent is a single bilinear differential equation: applying the new algorithms to the Boussinesq equation,

$$u_{tt} - u_{xx} - u_{xxxx} - [u^2]_{xx} = 0, \tag{8.1}$$

for example, is merely a matter of altering the function $\zeta(p,q)$ which is defined in Ref. 6. Other soliton equations like the sine-Gordon equation and cubic Schrödinger equation have Hirota equivalents which are systems of bilinear equations rather than a single equation. For these, there are still some holes even in the Hill's spectrum method, so the class of "coupled bilinear" equations requires further work. Still, there seems little doubt that most of the concepts developed here (using the Gaussian series for large amplitude and the Fourier series for small, reducing the partial differential equation to the algebraic residual equations, computing explicit perturbation series, and applying the modular transformation) will be important for these other types of soliton equations, too.

A much harder question is to relate the KdV polycnoidal waves to the nonlinear solutions of similar differential equations that are not "exactly integrable" via the inverse scattering or Hill's spectrum algorithms. The Gaussian series, which converges most rapidly when the wave amplitude is large, is a specific property of theta functions and does not carry over to waves that cannot be described in terms of theta functions.

Reference 1 (Appendix B) has shown, however, that it is possible to compute Fourier series representations for polycnoidal waves by using Stokes' expansions, which is a particular case of the singular perturbation technique known as the "method of multiple scales," without employing theta functions in any sense at all. The Stokes' expansion strongly suggests that double and triple and N-polycnoidal waves exist for almost any species of neutral, nondissipative waves whether the governing equation is "exactly integrable" or not.

This hypothesis must be qualified in several obvious ways. First, a perturbation series for a wave is not quite the same thing as an existence proof for the wave. For the Korteweg–de Vries equation, the Hill's spectrum method shows that the theta series converges for all values of the wave amplitude; the corresponding Fourier series for a nonintegrable equation may have only a finite radius of convergence, or perhaps be an asymptotic series with no radius of convergence at all.

Second, numerical experiments with nonintegrable differential equations have shown that their solitons collide inelastically with often the creation of a new soliton or the permanent destruction of an old one; such solutions cannot be classified as (limiting cases of) polycnoidal waves. However, this does not contradict the hypothesis that polycnoidal waves exist for nonintegrable equations, too. What makes polycnoidal waves so important for the Korteweg–de Vries equation is that they are complete, that is, the general initial value solution can be approximated to an arbitrary degree of accuracy by an N-polycnoidal wave of sufficiently large N. It seems probable that polycnoidal waves exist for at least some nonintegrable partial differential equations, but lack this property of initial value completeness. In other words, for nonintegrable equations, there are solutions which cannot be approximated to arbitrary accuracy by polycnoidal waves.

It is known, however, that for some nonintegrable equations which are closely related to integrable equations, the degree of inelasticity seems to be small. (This notion of "nearly integrable" equations is well developed with many examples in the review by Makhankov.[24]) Perhaps with better understanding of polycnoidal waves, it will be possible to put a bound on the nonpolycnoidal part of the solution and still apply the concept of a polycnoidal wave, at least qualitatively, to such nearly integrable equations.

IX. SUMMARY AND CONCLUSIONS

This article and its two companions (Boyd[6,7]) have tried to show that much can be learned about the generalized cnoidal waves of the Korteweg–de Vries equations and related equations by using rather elementary methods. The perturbation series of Boyd[6] provide an accurate means of calculating both phase speeds and $u(x,t)$ itself in all the interesting parameter regimes. The Gaussian series is especially useful because it converges rapidly in precisely that domain—large amplitude—where all normal perturbation theories fail. The special modular transformation, which involves nothing more esoteric than multiplying the theta matrix by another matrix whose elements are explicitly given integers, is essential in correctly interpreting the various modes of the double cnoidal wave. The most important mode is shown to be the sum of two solitary waves on each unit interval in x for large amplitude and to be the superposition of two linear sine waves, with one being the second harmonic of the other, for small amplitude.

The directions of future research are fairly clear. One is to simply apply the formalism developed here to other soliton equations like the Boussinesq equation (8.1) and turn the crank.

A second, more interesting direction is to explore the connection between polycnoidal waves and the general initial value problem with spatial periodicity. The Hill's spectrum method provides one complicated and indirect means of calculating that polycnoidal wave which approximates a given, arbitrary initial condition. It is known, however, that one can obtain a simpler answer by employing the method of multiple scales (a Stokes' expansion-with-a-twist, if you will) for small amplitude, and it appears possible to extend this into an effective numerical algorithm for any amplitude.

A third line of attack is to explore those other soliton equations whose Hirota bilinear form is a pair of equations rather than just one. The sine-Gordon equation and the cubic Schrödinger equation are examples. There are still some gaps even in the Hill's spectrum theory for these equations, so the extension of the ideas presented here to the coupled-bilinear class of systems is far from trivial. Nonetheless, one expects that perturbation theory, Gaussian series, the algebraic residual equations, and the modular transformation will all play a role.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation through grants OCE 8108530 and OCE 8305648.

APPENDIX A: THETA FUNCTION NOTATION

Mathematicians normally define the theta function via

$$\theta\begin{bmatrix}\boldsymbol{\epsilon}\\ \boldsymbol{\epsilon}'\end{bmatrix}(\boldsymbol{\zeta},\mathbf{T})=\sum_{\mathbf{n}}\exp\left\{\pi i\left[\sum_{i=1}^{N}\sum_{j=1}^{N}t_{ij}\left(n_i+\frac{\epsilon_i}{2}\right)\right.\right.$$
$$\left.\left.\times\left(n_j+\frac{\epsilon_j}{2}\right)+2\sum_{i=1}^{N}\left(n_i+\frac{\epsilon_i}{2}\right)\left(\zeta_i+\frac{\epsilon_i'}{2}\right)\right]\right\}. \tag{A1}$$

$\boldsymbol{\zeta}$ is the N-dimensional vector of dependent variables; in the theory of polycnoidal waves, $\zeta_1=k_i(x-c_it)+\phi_i$, $i=1,\ldots,N$ as in (1.1). The quantity $[{}^{\boldsymbol{\epsilon}}_{\boldsymbol{\epsilon}'}]$, the "characteristic" of the theta function, consists of two N-dimensional row vectors written one above the other with each element restricted to be either 0 or 1. The vector $\mathbf{n}=(n_1,n_2,\ldots,n_N)$, and the summation is taken over all possible positive and negative integers (including 0) for each of $n_1,n_2,\ldots,n_N$.

In applications to KdV polycnoidal waves, one can pick the characteristic at will. The usual choice, as in Nakamura[5] and Boyd,[1] is to use $\theta\,[{}^0_0]\,(\boldsymbol{\zeta},\mathbf{T})$. For the Gaussian series (soliton regime calculations), the formulas are a little simpler if one employs

$$\theta\begin{bmatrix}0\\1\end{bmatrix}(\boldsymbol{\zeta},\mathbf{T})=\theta\begin{bmatrix}0\\0\end{bmatrix}\left(\boldsymbol{\zeta}+\frac{1}{2},\mathbf{T}\right). \tag{A2}$$

Note that the two differ only in choice of the phase of $\boldsymbol{\zeta}$, but like all wave phases, these are arbitrary anyway. The choice of theta characteristic is physically irrelevant.

Although Ref. 7 uses the theta matrix in the mathematician's form (A1) [for convenience in discussing the derivation of the "special" modular transformation from the general transformation given by Rauch and Farkas[25]], it is easier in most applications to eliminate the factor of πi by defining the real theta matrix elements

$$T_{ij}\equiv-\pi it_{ij}. \tag{A3}$$

For the ordinary cnoidal wave $T_{11}\equiv\pi/s$, where s is the parameter used in Ref. 1.

For the Gaussian series, it is similarly convenient to define the elements R_{ij} of a square matrix $\mathbf{R}$, where

$$\mathbf{R}=2\pi^2\mathbf{T}^{-1}, \tag{A4}$$

$$\mathbf{T}=2\pi^2\mathbf{R}^{-1}, \tag{A5}$$

where $\mathbf{T}$ in (A4) and (A5) is the matrix whose elements are T_{ij}. The factors of π in (A4) arise from the factor of π in (A1) and (A3) and also from a similar factor of π when the Gaussian series of the theta function is expressed in terms of the inverse of the matrix whose elements are t_{ij}. The factor of 2 is inserted into (A4) to eliminate a huge number of 2's that would otherwise appear in the formulas of the Gaussian series perturbation theory.

APPENDIX B: CORRECTIONS AND CLARIFICATIONS FOR BOYD[1]

This earlier paper contains a number of typographical errors. A comma should be inserted between n' and c on the left-hand side of (6.6). The letter $\boldsymbol{\delta}$ in the argument of θ on the left-hand side of (7.1) should be replaced by $\boldsymbol{\zeta}$. In Eq. (5.3), 12 sech^2 $[sX]$ should be $12s^2$ sech$[sX]$. In (7.9), a Gaussian factor was omitted from the right-hand side of (7.9); the correct transformation is given by (2.10) of Ref. 7.

The author's earlier article makes the remark (p. 384) that "it is conventional to define the multidimensional theta function so that it is periodic with period 2." This is technically true for the general theta function, but it is somewhat misleading since the special cases $\theta\,[{}^0_0]$ and $\theta\,[{}^0_1]$—the only ones needed for polycnoidal theory—are periodic with period 1, as true of all the solutions discussed in this present, later article and its companions (Refs. 6 and 7).

Finally, as noted in Ref. 20, Toda showed the ordinary cnoidal wave has the exact series representation

$$u(x,t)=\frac{-24s}{\pi}+12s^2\sum_{\substack{n=-\infty\\ \text{[integers]}}}^{\infty}\operatorname{sech}^2[s(X-n\pi)]. \tag{B1}$$

The remark in Ref. 1 that repeating solitary waves with even spacing over $X\in[-\infty,\infty]$ as in (B1) could give only an approximate solution to the KdV equation is incorrect. Toda's proof was based on the infinite product of the theta function. Reference 26 shows that a more general method of proof is to apply Poisson summation—the same transformation that also generates the Gaussian series of the theta function—directly to the Fourier series of $u(x,t)$ given by (A9) of Ref. 1, and gives similar hyperbolic series for the elliptic functions dn, cn, and sn. The handbook of Gradshteyn and Ryzhik[27] lists some 21 other known Fourier series for various ratios and combinations of elliptic functions, and all can presumably be Poisson summed in the same way.

Unfortunately, the Fourier coefficients for the hyperelliptic functions, i.e., $u(x,t)$ for $N>1$, are not known although the theta function coefficients are known for all N. As a result, the Poisson summation method can only be applied to the theta function except for the special case of the ordinary cnoidal wave. Consequently, the author's earlier comment that the theta functions provide the only efficient way of generalizing solitary waves to spatially periodic functions remains true for $N > 1$.

[1] J. P. Boyd, J. Math. Phys. **23**, 375 (1982).

[2] W. E. Ferguson, Jr., H. Flaschka, and D. W. McLaughlin, J. Comput. Phys. **45**, 157 (1982).

[3] R. Hirota, Phys. Rev. Lett. **27**, 1192 (1971).

[4] R. Hirota and J. Satsuma, Prog. Theor. Phys. Suppl. **59**, 64 (1976).

[5] A. Nakamura, J. Phys. Soc. Jpn. **47**, 1701 (1949).

[6] J. P. Boyd, J. Math. Phys. **25**, 3402 (1984).

[7] J. P. Boyd, J. Math. Phys. **25**, 3415 (1984).

[8] R. Hirota and M. Ito, J. Phys. Soc. Jpn. **50**, 338 (1981).

[9] A. Nakamura and Y. Matsuno, J. Phys. Soc. Jpn. **48**, 653 (1980).

[10] A. Nakamura, J. Phys. Soc. Jpn. **48**, 1365 (1980).

[11] M. G. Forest and D. W. McLaughlin, J. Math. Phys. **23**, 1248 (1982); H. Flaschka, M. G. Forest, and D. W. McLaughlin, Commun. Pure. Appl. Math. **33**, 739 (1980).

[12] J. Zagrodziński, Lett. Nuovo Cimento **30**, 266 (1981); J. Zagrodziński and M. Jaworski, Phys. Lett A **92**, 427 (1982); J. Zagrodziński and M. Jaworski, Z. Phys. B **49**, 75 (1982); J. Zagrodziński, J. Math. Phys. **24**, 46–52 (1983).

[13] S. P. Novikov, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer-Verlag, New York, 1980), p. 325.

[14] Forest and McLaughlin[11] have stressed the usefulness of polycnoidal wave theory in studying a "high density of solitons," which does not necessarily imply periodic boundary conditions. The theory of baroclinic instability in the atmosphere, which has been studied via the sine-Gordon equation [J. D. Gibbon, I. N. James, and I. M. Moroz, Proc. R. Soc. London. Ser. A **367**, 219 (1979)], most emphatically does involve periodic boundary conditions, but the unstable waves are wavenumbers 4, 5, and 6 for typical values of the parameters, so that this is a problem where the components of a polycnoidal wave have ratios k_2/k_1 and so on which are fractions rather than integers.

[15] R. Grimshaw, Proc. R. Soc. London. Ser. A **368**, 359 (1979).

[16] The constant is 12α where $\alpha \equiv R_{11}\, k_1{}^2 + 2R_{12}\, k_1\, k_2 + R_{22}\, k_2{}^2$.

[17] Some care is necessary in interpreting Fig. 3 correctly in terms of sine waves. First, the theta function $\theta\left[{0 \atop 1}\right](X,Y;\mathbf{T}) \equiv \theta\left[{0 \atop 0}\right](X+\frac{1}{2}, Y+\frac{1}{2};\mathbf{T})$ is used to generate the graphs; as explained in Appendix A, $\theta\left[{0 \atop 1}\right]$ is more convenient in the solitary wave regime, but it differs by phase factors from the $\theta\left[{0 \atop 0}\right]$ which is used everywhere when discussing theta Fourier series. Second, expanding the logarithm of the theta function and then taking the second derivative multiplies all Fourier components by a minus sign. Thus, the "fundamental" referred to in the text is $-\cos[2\pi(X+\frac{1}{2})] = \cos(2\pi X)$. Third, the graphs were made in the [1,1] representation, i.e., $k_1 = k_2 = 1$, which is unnatural for Fourier series as explained in Sec. VI of this work, in Appendix A of Ref. 6, and Sec. VI of Ref. 7. The second harmonic is proportional to $-\cos[2\pi(X+Y+\frac{1}{2}+\frac{1}{2})] = -\cos[2\pi(X+Y)]$. Thus, the fundamental and second harmonic are out of phase at $t=0$ at $x=0$ and the fundamental has a peak there, even though naive use of (2.3) would seem to imply both should be negative for $x=t=0$. I ask the reader's indulgence for this long-winded explanation, but as is the theme of Ref. 7, the need to use different sets of wavenumbers to interpret the double cnoidal wave as sine waves or as solitons sometimes even left the author confused!

[18] P. D. Lax, Commun. Pure. Appl. Math. **21**, 467 (1968). He showed that if δ_1 and δ_2 are as defined in Sec. V of Ref. 6 and one defines $r \equiv \delta_1{}^2/\delta_2{}^2$, then the nonoverlapping collision (Fig. 5) occurs if $r > 2.618$. When $r > 3.0$, the large peak simply decreases to a certain lower bound and then begins to increase again, but no local minimum appears at $x = 0$ (Fig. 4). For intermediate r, there is an interval in time when there is but a single maximum (as true also for larger r), but there are two peaks with a shallow minimum between them when the phase factors ϕ_1 and ϕ_2 both $= 0$, as true for smaller r.

[19] B. Fornberg and G. B. Whitham, Philos. Trans. R. Soc. London. Ser. A **289**, 373 (1978).

[20] Although the author was not aware of it at the time Boyd[1] was written, M. Toda, Phys. Rep. **18**, 1 (1975), has shown that the series of displaced single solitons, (3.8) of Ref. 1, in fact is an exact solution of the KdV equation! The implied statement in Ref. 1 that it is only an approximate solution is therefore incorrect. See Appendix B for further discussion.

[21] G. B. Whitham, *Nonlinear Waves* (Wiley, New York, 1974), p. 584.

[22] J. M. Hyman, Rocky Mount. J. Math. **8**, 95 (1978).

[23] H. P. McKean and P. van Moerbeke, Invent. Math. **30**, 217 (1975).

[24] V. Makhankov, Comput. Phys. Comm. **21**, 1 (1980).

[25] H. E. Rauch and H. M. Farkas, *Theta Functions with Applications to Riemann Surfaces* (Williams and Wilkins, Baltimore, 1974), p. 229.

[26] J. P. Boyd, SIAM J. Appl. Math. (in press).

[27] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, 4th ed. (Academic, New York, 1965), pp. 911–912.