Modulation instability: The beginning

V.E. Zakharov\textsuperscript{a,b}, L.A. Ostrovsky\textsuperscript{c,d,∗}

\textsuperscript{a} Department of Mathematics, University of Arizona, Tucson, AZ, USA
\textsuperscript{b} Lebedev Institute of Physics, Russian Acad. Sci., Russia
\textsuperscript{c} Zel Technologies/NOAA Earth Science Research Laboratory, USA
\textsuperscript{d} Institute of Applied Physics, Russian Acad. Sci., Russia

\textbf{A B S T R A C T}

We discuss the early history of an important field of “sturm and drang” in modern theory of nonlinear waves. It is demonstrated how scientific demand resulted in independent and almost simultaneous publications by many different authors on modulation instability, a phenomenon resulting in a variety of nonlinear processes such as envelope solitons, envelope shocks, freak waves, etc. Examples from water wave hydrodynamics, electrodynamics, nonlinear optics, and convection theory are given.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

To give the reader an idea of how widespread is the notion of modulation instability (MI), we can recommend to do a simple Internet search. There are between one and two million entries on “Modulation instability” and even more for “Self-modulation” in, e.g., Yahoo. Even if these references are not all equally relevant, the numbers are still impressive. We believe that most of the researchers in the area of nonlinear waves would agree that the MI is one of the most ubiquitous types of instabilities in nature. Thus, it seems useful to briefly outline the beginnings of the research in this area: it is remarkable that different groups of physicists in different countries have started research in this area almost simultaneously, albeit independently, an indicator that the idea was emerging when the time was indeed ripe.

In its simplistic version, the effect of modulation instability is the result of interaction between a strong carrier harmonic wave at a frequency $\omega$, and small sidebands $\omega \pm \Omega$. This is the particular case of four-wave interaction (two quanta at $\omega$ create the quanta at $\omega + \Omega$ and $\omega - \Omega$). Growth of the sidebands can be treated in terms of amplification of weak modulation imposed on a harmonic wave (Fig. 1).

At the same time, in modern nonlinear physics, MI (or self-modulation) is considered as a basic process that classifies the qualitative behavior of modulated waves (“envelope waves”) and may initialize the formation of stable entities such as envelope solitons. It was observed in numerous physical situations including water waves, plasma waves, laser beams, and electromagnetic transmission lines. In theoretical models, the phenomenon was considered for even broader range of phenomena, from biological molecules to galactics.

As mentioned, the development of the theory of MI started almost simultaneously and occurred in parallel in hydrodynamics and electrodynamics/nonlinear optics. Thus, it should be stressed from the very beginning that our goal is not to set priorities, but on the contrary, to show how the same or similar ideas may arise independently when they are in demand.
The mathematical models used in all works considered below are largely similar and universal. Most of the authors understood well this universality and often explicitly stated it or even used the universal approach from the very beginning. Still, our narrative here will follow, whenever possible, this parallel development that seems instructive and characteristic of the early progress. We limit ourselves by relatively few works, mostly from the 1960s and early 1970s, and only briefly mention the later, intensive development. Whenever possible we preserve notation and logics of the original papers.

It might be interesting to note that the research in this area had been started by the Western and Soviet scientists in the 1960s almost independently, and often implied different physical applications. Most of the early Western work has been related to classical hydrodynamics: water waves, convection etc. On the other hand, Soviet works on MI of about the same period were largely based on the then recent progress in electromagnetics, including nonlinear optics (lasers, self-focusing, nonlinear radio waves etc.), and plasma physics (there were exceptions, however, even at that time; e.g., the Russian paper [11] cited below is concerned with water waves). Both authors of this paper began their scientific careers in the Soviet Union, with rather limited international contacts. We subsequently acquired an access to the papers by Whitham, then by Lighthill, and somewhat later by Benjamin and Feir. A publication in Russian of the materials of a discussion on nonlinear dispersive waves organized by Lighthill and published in English in 1967 (see Ref. [2]), especially enhanced our knowledge of the Western contributions to the area.

Naturally, here we have cited only published works and may have missed some interesting historical details, in particular relating to priorities. Indeed, A. Newell, in a private communication, has told us that when his advisor, D. Benney, first asked him, in spring of 1965, to look into the possibility of such an instability, it was because Benney had heard from Benjamin that he had had trouble reproducing the Stokes wave experimentally and believed it was unstable. Moreover, Newell also recalls Whitham saying at a much later date that he was initially puzzled by the fact that his modulation equations could be both hyperbolic (expected) and elliptic (unexpected) and it was only after he heard of Benjamin’s result that “the penny dropped”. It therefore may well have been that Benjamin was the first to derive, in the context of water waves, the criterion for the modulational instability. However, because the approach is quite general, and because it was from the Lighthill paper of 1965 [3] and from the volume [2] edited by Lighthill that we first learned of the result, we have decided to tell the story using Lighthill’s version. In Section 3, we carry out the calculation for water waves, the path that Benjamin and Feir and, later, one of us followed.

2. Benjamin–Feir–Lighthill criterion

In 1965, Whitham [4] suggested the averaged variational principle for quasi-periodic waves based on a period-averaged Lagrangian, \( \mathcal{L}(\omega, k, a) \) which depends on the wave phase \( \theta \) (actually on its derivatives, local frequency \( \omega = -\partial \theta / \partial t \) and wave number \( k = \partial \theta / \partial x \)), amplitude \( a \), and possibly other slowly varying parameters. Using \( \theta \) and \( a \) as canonical variables, one obtains equations describing slowly varying wave characteristics having in a 1-D case the form

\[
\frac{\partial \mathcal{L}}{\partial a} = 0, \quad \frac{\partial \mathcal{L}}{\partial \omega} - \frac{\partial \mathcal{L}}{\partial k} = 0, \quad \frac{\partial k}{\partial t} + \frac{\partial \omega}{\partial x} = 0. \tag{1}
\]

Lighthill [3] further developed Whitham’s theory, considering a specific case of small nonlinearity where the averaged Lagrangian can be reduced to

\[
\mathcal{L} = G(\omega, k)a^2 + B(\omega, k)a^4. \tag{2}
\]

Variation of this over \( a \) gives

\[
G(\omega, k) + 2B(\omega, k)a^2 = 0, \tag{3}
\]

or, after resolving with respect to \( \omega \),

\[
\omega = \omega_0(k) + \omega_1(k)a^2. \tag{4}
\]

The latter expression can be considered as a nonlinear dispersion equation in which \( \omega_0(k) \) follows from \( G(\omega, k) = 0 \) corresponding to the linear approximation, and \( \omega_1 \) is due to nonlinearity. Note that in the linear case when \( B = 0 \), from Eq. (3) we have \( \mathcal{L} = 0 \). This relation has a simple mechanical interpretation: average values of the kinetic and potential energy densities are equal in a linear traveling wave.

The rest of the two equations (1) gives

\[
\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x} (v_g a^2) = 0, \tag{5}
\]

and

\[
\frac{\partial k}{\partial t} + v_g \frac{\partial k}{\partial x} + \frac{\partial}{\partial x} (\omega_1 a^2) = 0, \tag{6}
\]

where \( v_g = -G_\omega/G_\omega = \omega_0/\omega_1 \) is the linear group velocity. Characteristic velocities for this system are

\[
C_k = v_g(k) \pm \sqrt{v_g/\omega_1 a^2 + O(a^2)}. \tag{7}
\]

Note that nonlinearity in Eq. (5) leading to the terms of order \( a^2 \) in Eq. (7) is neglected; these terms become important if dispersion

---

Fig. 1. Top: evolution of a nonlinear wave train in the course of MI. Bottom: the corresponding evolution of wave spectrum.
is small (see Section 3.3 below). Hence, the above equations are hyperbolic \((C \text{ is real})\)

\[
\beta = \omega_1 \frac{d\nu}{dk} > 0 \tag{8}
\]

and elliptic \((C_{\pm} \text{ are complex})\) if \(\beta < 0\). We shall refer to this condition as Benjamin–Feir–Lighthill (BFL) criterion. As it is easy to deduce from these expressions, if Eqs. (5) and (6) are linearized around the harmonic wave with constant \(a, \omega_c, k_c,\) and perturbations are sought in the form of \(\exp(i(kx - \Omega t))\), the result is

\[
\Omega = C_{\pm} K. \tag{9}
\]

Hence, in the hyperbolic case the harmonic wave is stable, and, according to (7), can propagate with two slightly different “group velocities”, whereas in the elliptic case it is unstable with respect to small modulation.

Lighthill then considered a weakly nonlinear Stokes wave on deep water when the nonlinear dispersion equation (4) reads

\[
\omega = \omega_0(k) \left(1 + \frac{1}{2} k^2 a^2\right), \quad \omega_0 = \sqrt{gk}. \tag{10}
\]

In this case, from (8) it follows that

\[
\beta = -\frac{\omega_0^2}{8} < 0. \tag{11}
\]

Hence, this is an elliptic case. Although Lighthill did not explicitly discuss wave stability in that paper, it is clear that, according to (7) and (9), a weakly nonlinear Stokes wave is unstable with an increment

\[
\nu = \text{Im}(C) K = \frac{\omega_0(k)}{2} \frac{d\nu}{dk} (k) \Omega. \tag{12}
\]

In this approximation, the increment increases monotonically with modulation frequency. As described below, a limitation of this result was established shortly thereafter in both water wave theory and in electrodynamics.

Modulation instability can also be explained as follows. Suppose that at some moment a local “bump” of intensity in a propagating wave appears. If, for example, \(\omega_1 > 0\) in (4), the derivative \(\omega_1\) is positive before the bump and negative after it. According to the wave phase conservation expressed in the last equation (1), that means that \(k_1 < 0\) before maximum and \(k_1 < 0\) after it. Suppose now that \(\omega_0(k) = \frac{d\nu}{dk} < 0\). Thus, the group velocity (more exactly, its linear part) tends to increase behind the peak and decrease in front of it. This means that the wave groups neighboring the amplitude maximum tend to compress the bump; due to the energy conservation, the amplitude increases cumulatively. The same reasoning shows that an initial trough in a harmonic wave would deepen. This is the case of modulation instability. In case of \(\omega_0(k) > 0\), the effect is the opposite: the initial bump tends to be smeared; this is the case of neutral stability.

From the spectral viewpoint, a simple interpretation of the BFL criterion is as follows: small sidebands interact with the strong carrier wave; for their effective interaction, the simultaneous fulfillment of resonance (synchronism) conditions is needed:

\[
\begin{align*}
\omega_1 + \omega_2 &= 2\omega_c, \quad k_1 + k_2 = 2k_c, \quad \text{or} \\
\omega_{1,2} &= \omega_c \pm \Omega, \quad k_{1,2} = k_c \pm K
\end{align*} \tag{13}
\]

(the latter equalities are for slow modulation when \(\Omega\) and \(K\) are small). Here subscripts 1, 2 correspond to the carrier wave and the sideband waves, respectively. In the linear case these conditions are not met because of dispersion. In the nonlinear case, the velocities of these waves differ due to two factors: dispersion (due to their frequency differences) and nonlinearity (they propagate on the background of the nonlinear carrier wave). In cases when these detunings are of the same sign, no synchronism occurs and the sidebands do not increase (hyperbolic case). If, however, these detunings differ in signs, they can compensate each other, and the waves interact in a synchronous, resonant manner, which results in their amplification. It is reflected in the BFL criterion; indeed, \(\beta\) is a product of the parameters of dispersion and nonlinearity.

### 3. Higher-order dispersion. Nonlinear Schrödinger equation

#### 3.1. Water waves. Benjamin–Feir instability

Whitham’s equations (1) can be considered non-dissipative with respect to the “complex envelope” \(a(x, t)\). As a result, the BFL criterion (8) does not depend on modulation frequency (provided it is small as compared to the carrier frequency). However, a more detailed account of the effects of dispersion imposes an additional limitation on modulation instability.

For water waves it was demonstrated by Benjamin and Feir [5,6], who discovered modulation instability for nonlinear Stokes waves on the water surface. Such a discovery came as a surprise. Indeed, for decades the existence of stationary nonlinear (Stokes) waves on deep water was the subject of an involved mathematical proof (e.g., [7,8]). Suddenly it was determined that although such solutions do exist mathematically, they are unstable! Benjamin and Feir [5] experimentally demonstrated and theoretically explained this fact. Feir’s experiments were performed in a water channel with a wave maker producing a wave with a length of 2.2 m; some results are shown in Figs. 2 and 3. Their theoretical model for such instability used a spectral approach, starting from the equations and boundary conditions for the one-dimensional potential, \(\psi(x, z, t)\), and the surface displacement, \(z = \eta(x, z, t)\), in the form (for deep water)

\[
\begin{align*}
\psi_{xx} + \psi_{zz} &= 0, \\
\eta_t + \eta_x \psi_{x=\nu} - |\psi_{z=\nu}| &= 0, \\
g \eta + \left[\psi_1 + \frac{1}{2} \left(\psi_+^2 + \psi_-^2\right)_{z=\nu}\right] &= 0,
\end{align*}
\]

where \(g\) is gravity acceleration; \(z = 0\) corresponds to a non-perturbed surface. A known solution of these equations is a progressive (Stokes) wave water, in which only the basic (first) and the second harmonics are retained:

\[
\begin{align*}
\eta &= H \approx a \left(\cos \zeta + \frac{1}{2} ka \cos 2\zeta\right), \\
\psi &= \Phi \approx ak^{-1} ae^{i2 \sin \zeta}, \\
\omega^2 &= gk(1 + k^2 a^2).
\end{align*}
\]

Here \(\zeta = kx - \omega t, \) and \(\alpha\) is its amplitude. Then small perturbations are added to this solution, each being represented as a sum of spectral components at frequencies \(\omega \pm \Omega\), where \(\Omega\) is a modulation frequency and \(\Omega \ll \omega\). In other words, the wave is now represented in the form \(\eta = H + \eta_1 + \eta_2, \psi = \Phi + \theta_1 + \theta_2\). The sideband waves \(\eta_1\) and \(\eta_2\) are supposed to have amplitudes \(\epsilon_{1,2}\) and phases

\[
\zeta_{1,2} = k(1 \pm \delta) x - \omega(1 \pm \delta) t - \gamma_{1,2}, \tag{16}
\]

where \(k\) and \(\delta = \Omega / \omega\) are small fractions satisfying the relation \(\delta_0 = c_0 k,\) and \(c_0 = g/(2\omega_0)\) is the linear group velocity at the main frequency. The parameters \(\gamma_{1,2}\) are corrections that arise due to dispersion (a difference of group velocities at the main wave and the side components) and to nonlinearity. If \(\theta = \gamma_1 + \gamma_2\), the four-wave resonance mentioned in the Introduction occurs when \(2\zeta = \zeta_1 + \zeta_2 + \text{const}\). As a result, the perturbations can increase, which is equivalent to the MI.
Fig. 2. Evolution of a wave train with the main frequency of 0.85 Hz in a water tank. Upper record is taken at a distance of 60 m from the wave maker; lower record is for 120 m. Time marks are at each 0.1 s. Vertical bar is in inches. From [6].

Fig. 3. Photographs of progressive wave trains illustrating the wave breaking due to the instability. Upper photograph is made near the wave maker; lower at 60 m from it. The main wave length is 2.2 m. From [6].

After substitution of the perturbed $\eta$ and $\varphi$ with slowly varying $\varepsilon_1(t)$ and $\theta(t)$ into Eq. (14) and keeping only resonance terms, the following equations follow after some transformations:

$$\frac{d\varepsilon_\pm}{dt} = \frac{1}{2} \left( \omega k^2 a^2 \sin \theta \right) \varepsilon_\mp, \quad \frac{d\theta}{dt} = \omega k^2 a^2 \left( 1 + \frac{\varepsilon_1^2 + \varepsilon_2^2}{2k} \cos \theta \right) - \Omega^2 \omega / \Omega. \quad (17)$$

For a Stokes wave this yields instability with growth rate

$$\gamma = \frac{1}{2} \delta \left( 2k^2 a^2 - \delta^2 \right)^{1/2}. \quad (18)$$

From here it is evident that the instability exists in a limited range of modulation frequencies,

$$\Omega < \Omega_s = \omega k a \sqrt{2}. \quad (19)$$

The maximum growth rate (increment) is achieved at $\Omega = \Omega_s / \sqrt{2} = \omega k a$ (Fig. 4).

At small $\Omega$ formula (18) reduces to (12) following from Lighthill’s consideration, but in general it represents a more specific condition for instability.

The works by Whitham, Lighthill, Benjamin, and Feir stimulated a lively discussion organized by M. J. Lighthill [2]. Within that event, the contributions by the above authors have expanded their previous studies. For example, Benjamin and Feir have shown that for water of finite depth $h$, the instability takes place at $kh > 1.363$, i.e., as expected, waves on shallow water are modulationally stable. Lighthill suggested a phenomenological averaged Lagrangian for Stokes waves. Whitham showed an equivalence between the spectral method used by Benjamin and Feir and his own modulational approach as regards to the MI. The range of relevant problems has been broadened by other authors. In particular, Phillips, Hasselmann, and Longuet-Higgins and Gill have studied resonance interactions of waves beyond the limits of the MI.

3.2. Hamiltonian approach for water waves

Zakharov [1] has shown that the equations of type (14) for weakly nonlinear waves on the surface of deep fluid can be reduced to a Hamiltonian form

$$\frac{\partial \eta}{\partial t} = \frac{\delta E}{\delta \varphi_s}, \quad \frac{\partial \varphi_s}{\partial t} = -\frac{\delta E}{\delta \eta}, \quad (20)$$

where $\varphi_s$ is the potential at the surface, $z = \eta$, and $E$ is energy (Hamiltonian). Then the dynamic equations are expressed in terms of Fourier components $a(k)$ that can be considered as new complex canonical variables:

$$\eta(k) = \sqrt{\frac{|k|}{2\omega(k)}} \left[ a(k) + a^*(-k) \right], \quad (21)$$

$$\varphi_s(k) = -i \sqrt{\frac{\omega(k)}{2|k|}} \left[ a(k) - a^*(-k) \right].$$

The resulting Hamiltonian equation is

$$\frac{\partial a(k)}{\partial t} = -i \frac{\delta E}{\delta a^*(k)}. \quad (22)$$

The energy $E$ is then represented as a series in powers of $a(k)$ and $a^*(k)$ up to the quartic terms, integrated over all ranges of wave vectors. For weakly nonlinear waves, the complex amplitudes can be presented in the form $a(k) \approx \tilde{A}(k, \tau) \exp \left[ -i \omega(k) \tau \right]$, where $\tilde{A}$ is a slowly varying function.

For a wave packet with a narrow spectrum, the nonlinear Schrödinger equation (NSE) follows from here for the wave envelope; in one-dimensional case it has the form

$$\frac{\partial \psi_s}{\partial t} - \frac{i\lambda}{2} \frac{\partial^2 \psi_s}{\partial x^2} = -w |\psi_s|^2 \psi_s, \quad (23)$$

where $\xi = x - v_p \tau$, $\omega_p = d\omega/dk$, $\lambda = d^2 \omega/dk^2$. This equation has an obvious solution in the form of a constant-amplitude harmonic wave, the phase velocity of which depends on amplitude. Namely, at a given $k = k_0$, the frequency is $\omega = w_0 k_0^2$, where $b_0$ is proportional to the wave amplitude. Adding a perturbation so that
\[ \psi_t = e^{-i|\omega|t} \left( b_0 + \alpha e^{-i2\Omega t + i\xi} + \alpha^* e^{i2\Omega t - i\xi} \right) \]  and linearizing Eq. (23), one obtains

\[ \Omega^2 = w \lambda k^2 |b_0|^2 + \lambda^2 k^2 / 4. \]  (24)

It is seen from here that the instability is possible if \( w\lambda < 0 \), which corresponds exactly to the BFL criterion (8). Besides, it exists at relatively long modulation spatial scales when \( k^2 < 4w|b_0|^2 / \lambda \). For water waves, this is equivalent to the above condition (19). It should be noted that in all cases considered above, at slow modulation the increment, \( \text{Im}(\Omega) \), is proportional to the wave amplitude rather than its square.

Later, Zakharov and Kharitonov [9] generalized this result to a finite-depth fluid and two-dimensional case and obtained results similar to those of Benjamin mentioned above, but have also shown that this type of instability exists only in a narrow range of angles between the direction of the main wave and perturbation.

### 3.3. Modulation instability of electromagnetic waves

Analogous research in electromagnetics was occurring simultaneously with the hydrodynamic ones and was stimulated, in particular, by the rapid progress in studies of lasers and the related problems of nonlinear optics.

A study of self-modulation of nonlinear electromagnetic waves was performed by Ostrovsky [10] in application to waves in a nonlinear dielectric described by one-dimensional coupled equations for an electric field \( E \) and polarization \( P \):

\[ \begin{align*}
 c^2 E_{xx} &= E_{tt} + 4\pi P_t, \\
 P_{tt} + \alpha^2 P - \alpha^2 P &= (\omega^2 / 4\pi)E, 
\end{align*} \]  (25)

where \( \omega_0, \alpha_0 \), and \( \alpha \) are constant parameters (resonance frequency of a dipole, plasma frequency, and nonlinearity parameter, respectively), and \( c \) is light speed in a vacuum. Seeking a solution with slowly varying amplitude and phase,

\[ E = A(x, t) \cos \left[ (\omega t - kx + \psi(x, t)) \right] \]  (26)

and similarly for \( P \), one obtains the equations for “envelope waves”:

\[ A_t + v_g (1 + \kappa \delta + q_1 A^2) A_x = \frac{K}{2} \delta t, \]  (27)

\[ \delta t + v_g (1 + \kappa \delta - q_2 A^2) \delta_x + g(A^2) \delta_{xx} = -\frac{K}{2\omega_0^2} \left( \frac{A_t}{A} \right) \]  (28)

Here \( v_g(\omega) \) is linear group velocity, \( \kappa \) is proportional to \( d^2 k / d\omega^2 \), and \( q_1 \) and \( g \) are constant parameters given in the cited paper. The variable \( \delta = \psi_t / \omega \ll 1 \) is a relative frequency perturbation. If the terms with \( q_1,2 \) are neglected as in Lighthill's equation (5), the system (27) is completely equivalent to the nonlinear Schrödinger equation (23) after representing the complex amplitude in terms of real amplitude and phase. After also neglecting the term in the r.h.s. of the second equation, one would obtain a hydrodynamic-type system similar to that considered by Lighthill; this system can be either hyperbolic (stable) or elliptic (unstable). The terms with \( q_{1,2} \) are responsible for asymmetric distortion of the envelope at symmetric initial conditions; later, similar terms were included by Dysthe in water wave equations; these terms are important in the case of relatively weak dispersion when \( \kappa \) is small, i.e., \( v_g \) is weakly dependent on \( \omega \).

System (27) has a straightforward particular solution in the form of a harmonic wave with \( A = A_0 = \text{const} \), \( \delta = \delta_0 = \text{const} \); here one can let \( \delta = 0 \) without losing generality. After perturbing it by \( A' = A - A_0 \simeq \exp[i(\kappa x - \Omega t)] \), \( \delta \sim \exp[i(\kappa x - \Omega t)] \), one obtains a dispersion equation similar to (18) and (24):

\[ \kappa = \Omega \left[ 1 \pm |s| \left( -g A_0^2 + (\Omega / \omega)^2 \right)^{1/2} \right], \quad g' \sim g/s. \]  (28)

### 3.4. Electromagnetic experiments

Although in the above papers some estimates for nonlinear optics were made, the first experimental observation of modulation instability (i.e., amplification of small initial modulation) for electromagnetic waves was made in a radio wave range by Zagryadskaya and Ostrovsky [13] in a line, with ferrites as nonlinear elements. They used an exciting signal at frequency \( \nu \) which ranged from 200 to 500 kHz; modulation frequency varied between 0.02\( \nu \) and 0.15\( \nu \); and initial modulation depth from 7% to 10.5%. In the course of propagation, the modulation depth increased by 25% on average, depending on the wave amplitude. Note that in these experiments, initial amplitude modulation produced phase (frequency) modulation, as follows from the above theory. Later, Ostrovsky and Soustov [14] observed a stronger effect in a semi-conductor line (Fig. 5).

Numerous later experiments in nonlinear optics typically involved media with possible modulation instability, but in most cases the observations dealt with nonlinear stages such as envelope solitons ("bright solitons") which, as mentioned, can exist only in unstable situations and can be a result of evolving modulation instability, and with "dark solitons" corresponding to the stable case. In their paper of 1973, Hasegawa and Tappert [15] suggested to use bright solitons for transmission of light in fibers, (with reference to [10]); later this idea has turned to be practically important.

### 4. Three-dimensional, space-time instabilities

#### 4.1. Electromagnetic waves

The relevant (and very important) topic was self-focusing of laser beams predicted as a general idea by Askaryan in 1962 [16] and elaborated in many works beginning in 1964–65 (e.g., [17,18]); a review of early results in this area can be...
found in [19]. Actually, self-focusing can be understood as “spatial self-modulation”: a small initial “ripple” on the front of a plane wave can increase in the course of propagation. In the 1960s, self-focusing was one of the “hottest” topics in nonlinear optics. Note also that propagation of an intensive optical pulse in a medium with cubic nonlinearity causes the “phase self-modulation” (frequency modulation due to the amplitude variation) which can result in a significant frequency shift and broadening of the wave spectrum. This effect was described by Ostrovsky [20] and observed by Brewer [21]). This is, however, not an instability yet.

Modulation instability and spatial instability have many common features. These two effects were combined by Litvak and Talanov [22], who considered electromagnetic waves in a dispersive medium with cubic nonlinearity. They first obtained a linear parabolic equation for complex amplitude that combines a stationary linear parabolic equation for a small-angle diffracting beam (obtained earlier by Leontovich and Fock for electromagnetic waves), and then by Maluzhinets for acoustic waves), and the linearized Schrödinger equation for one-dimensional electric field in the form of $E = A(x, t) \exp(i\omega(t-k_0x))$. Then the derivation was extended to the nonlinear case to yield the nonlinear parabolic equation

$$\Delta_{\perp}A + k_0(\partial v_{gr}/\partial\omega)A_{\perp} - 2ik_0A_\perp + k_0'^2e^2|A|^2A = 0,$$  

where $\xi = x - v_{gr}t$, and $e^2$ is proportional to the nonlinear permeability of the medium at the carrier frequency.

Then the authors considered modulation instability by imposing small perturbations on a plane harmonic wave. After some transformations, seeking a perturbation proportional to $\exp(\imath \xi - k_{\perp}x_{\perp} - \imath\omega t)$, a dispersion equation follows:

$$4\hbar^2 = (\kappa^2 + sk^2)(\kappa^2 + sk^2 - 2e^2|A|^2),$$  

where $s = k_0(\partial v_{gr}/\partial\omega)$. This equation includes the particular cases of $\hbar = 0$ that is spatial growth for $\kappa^2 < 2e^2|A|^2$, and the 1-D case when $k_{\perp} = 0$. This case is similar to those considered above, except instead of perturbation increasing in time, it increases along $x$ (at $\xi's > 0$ and $k < 2e^2|A|^2$/s); the spatial and temporal increments differ only by a factor of $v_{gr}$. In a general case, the region of instability in the $(\kappa^2, k^2)$ plane lies between the straight lines $\kappa^2 = -sk^2$ and $\kappa^2 = -sk^2 + 2e^2|A|^2$ with the maximum increment lying on the line $\kappa^2 = -sk^2 + e^2|A|^2$.

Note that the consideration in [22] was based on a general form of electric dispersion, when the relation between electric induction $D$ and electric field $E$ is given in an integral form. In general, the frequency dependence of nonlinear permeability was also considered (the latter is absent in Eq. (30)).

The authors calculated a nonlinear stage of evolution of a 1-D wave initially modulated sinusoidally, and also of a Gaussian pulse. In addition to the evolution of the wave amplitude, they calculated its phase and have shown that at the point of maximal self-compression, the phase is close to constant along a pulse (a known fact for the focal area of a steady linear beam). Also, the form of a steady envelope soliton was explicitly presented via hyperbolic cosine.

A space-time equation similar to (29) for an optical beam has also been written by Zakharov [23] who used it to analyze stability of nonlinear beams and pulses.

The effects combining self-focusing and self-modulation have also been analyzed for plasma. In particular, Litvak [24] has written a parabolic equation similar to (29), for a magnetoeactive plasma. In general, plasma is subject to various types of instabilities including the modulation instability according to [25] (where, however, the three-wave interactions were considered).
Now suppose that $T(k_0, k_0, k_0, k_0)$ is a continuous function if all $k_i \to k_0$. Then $\Delta \Omega(k_0) = T(k_0, k_0, k_0, k_0) = T$. In the limit $k_i \to k_0$ the expression (39) for $\nu$ can be simplified. Denoting $k_1 = k_0 + p$, $k_2 = k_0 - p$, one has
\[
\nu = \left( \frac{1}{4} \Phi^2 - \Phi T^2 \right)^{1/2}; \quad \Phi = \frac{\partial^2 \omega}{\partial k_a \partial k_b} \rho_a \rho_b.
\] (40)

In the 1-D case, this expression coincides with the above results, cf. Eqs. (24) and (28).

When $T$ is a continuous function, equations (38) for a narrow-band wave can be reduced to the NSE. Note that for gravity waves on water of a finite depth, $T$ is not a continuous function, and NSE is replaced by the Davey–Stewartson equation.

4.3. Modulation approach

Benney and Newell [27] started from multi-wave equations for complex Fourier amplitudes, in the form
\[
\frac{dA_i}{dt} = i \sum_{m,n} \alpha_{imm} A_m^* A_n + i \varepsilon \left\{ \sum_q \beta_q A_q A_0^* + \sum_{q,r} \gamma_{qr} A_q A_r^* \right\}.
\] (41)

with $\alpha_{0l} = \omega(k_0) = -\omega_1$, and $A_l = \alpha(k_0, t) = -A_0^*$. One of the particular cases considered, based on this system, excludes the explicit resonance triads and quartets but allows slow modulation of a given wave, $A_l = a_l(k, X, T) \exp(-\eta T)$ with large (slow) scales $T$ and $X$, so that the terms with $\eta / \partial T$ and $\partial / \partial X$ are of order $\varepsilon$. As a result, the authors obtained a single-mode equation for wave amplitude:
\[
\frac{\partial a}{\partial T} = i \sum_{r,s} \delta_{rs} \frac{\partial^2 a}{\partial X_r \partial X_s} + \beta a^2 a^*.
\] (42)

where $\delta_{rs}$ is (1/2)$\partial^2 \omega / \partial k_r \partial k_s$ is dispersion parameter; $T' = \varepsilon T$, $X' = X - c_0 T$, and $c_0 = \nabla k_0 \omega$ is the group velocity vector. NSE immediately follows from here for the 1-D case. The authors considered modulational stability of a harmonic wave and found a dispersion equation for the harmonic perturbation at a frequency $\Omega'$ (this notation is ours) and wave vector $K'$:
\[
\Omega'^2 = \sum_{r,s} \delta_{rs} K'_r K'_s \left( \sum_{r,s} \delta_{rs} K'_r K'_s - 2 \beta |a_0|^2 \right).
\] (43)

Here again, modulation stability/stability depends on the sign of the real quadratic form $\beta \delta_{rs} K'_r K'_s$.

4.4. Phenomenological approach

Yet another option is to use a semi-phenomenological approach based on the nonlinear dispersion equation and energy conservation. This method was used by Karpman and Krushkal [28]. They used the nonlinear dispersion equation
\[
\omega \approx \omega_0(k) + \left( \partial \omega / \partial a^2 \right) a^2.
\] (44)

where $a$ is the wave amplitude. Representing the wave in a form similar to (26) and expanding (44) in powers of perturbation of the wave number $k = k_0 + \nu \psi(r, t)$, one obtains an equation for the phase $\psi$ as
\[
\psi_t + v_p \psi_x + \frac{1}{2} \left( \frac{\partial v_p}{\partial k} \right)_0 \psi_x^2 + \frac{v_p}{2k_0} (\nabla \psi)^2
+ \left( \frac{\partial \omega}{\partial a^2} \right)_0 |a_0|^2 = 0.
\] (45)

Another equation follows from the energy conservation equation with a linear group velocity: $(a^2)_{,x} + (\nabla \times v_p) a^2 = 0$. From this and the above form of the wave vector, it follows that
\[
(a^2)_{,x} + (v_p / k_0) \nabla \cdot (\nabla \psi a^2) = 0.
\] (46)

In the 1-D case, equations (45) and (46) are analogous to the system (27) with $a_{1,2} = 0$. The same results regarding stability criterion and limitation on modulation frequency of growing modulation are applicable here. As in Litvak and Talanov [22], these equations take into account transverse modulation and were represented in a complex form of the parabolic equation similar to (29).

5. Active systems

Finally, we briefly outline a more general case when the basic system is active, i.e., it has some energy pumping. When the pumping is balanced by some kind of nonlinear losses, an equilibrium is possible. In case of a progressive wave, it is often called an autowave. An important difference from the conservative systems considered above is that the equilibrium wave amplitude is not arbitrary but has only one or several discrete values prescribed by the balance mentioned above. Such processes are typical of convection, supercritical hydrodynamic flows, reaction-diffusion systems in chemistry and biology, lasers, and electric lines with active elements such as transistors. The entire problem is far beyond the framework of the present consideration; here we only mention possible modulation instabilities, in which two small side components satisfying the resonance conditions (13) can be amplified, as above.

Early results in this area have been obtained by Eckhaus as early as 1963 (published in English in [29]), then, in 1969, by Segel [30] and (in a form which is closer to the approaches considered above) by Newell and Whitehead [31], all in application to the convection theory. The latter authors introduced the complex Ginzburg–Landau (GL) equation which describes a complex envelope $W$ of a quasi-harmonic, almost plane wave,
\[
W_t = g W + (a + ib) \nabla^2 W - (d + ic) |W|^2 W.
\] (47)

Evidently, at $g = a = d = 0$, in the 1-D case this equation reduces to the NSE considered above, and in a 3-D case, to the space-time parabolic equation of type (29). If $b = c = 0$ and real $g > 0$, (the case of GL with real coefficients considered by Eckhaus), there exist harmonic envelope waves, $W_0 = A_0 \exp(i k x + \theta_0)$ with $A_0 = \sqrt{g - K^2}$. In the case of convection (for which most of the early results were obtained), these solutions describe a stationary roll pattern with a wave number $k = k_0 + K$. Eckhaus has shown that a small modulation leads to instability if $g > 3K^2$.

(48)

In later works, various other perturbation modes and their possible growth in specific hydrodynamic structures have been studied; still, the criterion (48) seems to remain most simple and universal.

The general, complex version of GL is rich in variants. Newell and Whitehead performed a detailed analysis of space-time instabilities of two-dimensional structures (patterns) in this equation by applying the spectral (“finite bandwidth”) approach. In the later work [33], Newell has shown that the steady harmonic solution of (47),
\[
W = \sqrt{\frac{g}{d}} \exp \left( -\frac{i g}{d} x \right)
\] (49)

is modulationaly unstable when $ad + bc < 0$.

(50)
which is the direct analog of the modulational instability considered above.

In the papers [31–34] the steady autowave solutions of (47) with real coefficients (patterns) and their stability have also been studied. At a small diffusion coefficient, these solutions can be of a meandering type, i.e., they consist of (almost) constants, \( W = \pm \sqrt{g/\alpha} \), connected by fronts (kinks), and have a solution in the form of a periodic plane wave which, for a relatively small diffusion coefficient \( \sigma \), is of a meandering type (i.e., it consists of (almost) constants, connected by fronts (kinks)).

Different variations and applications of this problem have been thoroughly studied in many other papers, e.g., in the work by Kogelman and DiPrima of 1970 [35]. It is really interesting that many hydrodynamic structures and patterns can be interpreted in terms of complex envelopes. For an outline of a similar (and somewhat more general) approach to pattern formation, see the later review paper by Newell et al. [36] and references therein.

Among the other early works, we note that that Zakharov et al. [37] derived the complex GLE equation for traveling and standing spin waves, and Ostrovsky [38] considered an example of modulation instability in an electromagnetic wave with amplification and losses (this effect is relevant to the active medium in laser resonators). In a broader sense, auto-synchronization of modes in lasers (leading to so-called giant pulses) is just a more complex case of self-modulation when many sideband frequencies (resonator modes) become phase correlated due to nonlinearity.

6. Conclusions

Here we mentioned only the very early works from which the theory of nonlinear self-action leading to modulation instability has begun. Our aim was to show a broader picture of how the research was being unfolded at that time, both in the Western countries and in the Soviet Union. As the reader can see, various methods have been used for treating MI which may formally differ but essentially leads to the same mathematical and physical results. Such a variety of works that have appeared within the span of a few years is an indicator of the importance of the phenomenon in many areas of physics. Many (albeit not all) of these studies were based on the nonlinear Schrödinger equation (NSE) which was derived, either in real or in complex variables, in many papers of the 1960s, including many of those cited above. Actually, however, a similar equation had been first derived as early as 1961 for the Bose–Einstein condensates; it is known as the Gross–Pitaevskii equation (see, e.g., in the book [39]). The corresponding topic is evidently very important at present but it was different from the MI in nonlinear waves considered here. In 1971 Zakharov and Shbat [40] showed that NSE belongs to a class of equations that are completely integrable by the inverse scattering method. This equation served as a basis for a broad spectrum of nonlinear effects including envelope solitons which, as already mentioned, can be formed at the nonlinear stage of modulation instability, and envelope shocks and “dark solitons” existing in the stable case. Also MI of water waves is a possible cause of the occurrence of “freak waves”, sporadic bursts on the water surface strongly exceeding the average level. We already mentioned the role of solitons in fiber optics. Finally, many effects of hydrodynamic instability, convective patterns etc., have an origin in the same class of processes. However, outlining, even briefly, all important effects related to MI and the resulting “envelope waves” would draw us into a boundless area. Thus, we limit ourselves to referring the reader to the books [41–44] and encyclopedic articles [45,46] in which different aspects of the problem are discussed.

References

[28] T.J. Stuart, R.C. DiPerna, Bose–Einstein condensates; it is known as the Gross–Pitaevskii equation (see, e.g., in the book [39]). The corresponding topic is evidently very important at present but it was different from the MI in nonlinear waves considered here. In 1971 Zakharov and Shbat [40] showed that NSE belongs to a class of equations that are completely integrable by the inverse scattering method. This equation served as a basis for a broad spectrum of nonlinear effects including envelope solitons which, as already mentioned, can be formed at the nonlinear stage of modulation instability, and envelope shocks and “dark solitons” existing in the stable case. Also MI of water waves is a possible cause of the occurrence of “freak waves”, sporadic bursts on the water surface strongly exceeding the average level. We already mentioned the role of solitons in fiber optics. Finally, many effects of hydrodynamic instability, convective patterns etc., have an origin in the same class of processes. However, outlining, even briefly, all important effects related to MI and the resulting “envelope waves” would draw us into a boundless area. Thus, we limit ourselves to referring the reader to the books [41–44] and encyclopedic articles [45,46] in which different aspects of the problem are discussed.

References