

## W

## Water Waves and the Korteweg–de Vries Equation

LOKENATH DEBNATH

Department of Mathematics,

University of Texas – Pan American, Edinburg, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

The Euler Equation of Motion in Rectangular Cartesian and Cylindrical Polar Coordinates

Basic Equations of Water Waves with Effects of Surface Tension

The Stokes Waves and Nonlinear Dispersion Relation

Surface Gravity Waves on a Running Stream in Water

History of Russell's Solitary Waves and Their Interactions

The Korteweg–de Vries and Boussinesq Equations

Solutions of the KdV Equation:

Solitons and Cnoidal Waves

Derivation of the KdV Equation from the Euler Equations

Two-Dimensional and Axisymmetric KdV Equations

The Nonlinear Schrödinger Equation and Solitary Waves

Whitham's Equations of Nonlinear Dispersive Waves

Whitham's Instability Analysis of Water Waves

The Benjamin–Feir Instability of the Stokes Water Waves

Future Directions

Bibliography

### Glossary

**Axisymmetric (concentric) KdV equation** This is a partial differential equation for the free surface  $\eta(R, t)$  in the form  $(2\eta_R + \frac{1}{R}\eta + 3\eta\eta_\xi) + \frac{1}{3}\eta\xi\xi\xi = 0$ .

**Benjamin–Feir instability of water waves** This describes the instability of nonlinear water waves.

**Bernoulli's equation** This is a partial differential equation which determines the pressure in terms of the ve-

locity potential in the form  $\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{P}{\rho} + gz = 0$ , where  $\phi$  is the velocity potential,  $P$  is the pressure,  $\rho$  is the density and  $g$  is the acceleration due to gravity.

**Boussinesq equation** This is a nonlinear partial differential equation in shallow water of depth  $h$  given by

$$u_{tt} - c^2 u_{xx} + \frac{1}{2}(u^2)_{xx} = \frac{1}{3}h^2 u_{xxtt},$$

where  $c^2 = \sqrt{gh}$ .

**Cnoidal waves** Waves are represented by the Jacobian elliptic function  $cn(z, m)$ .

**Continuity equation** This is an equation describing the conservation of mass of a fluid. More precisely, this equation for an incompressible fluid is  $div \mathbf{u} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$ , where  $\mathbf{u} = (u, v, w)$  is the velocity field, and  $\mathbf{x} = (x, y, z)$ .

**Continuum hypothesis** It requires that the velocity  $\mathbf{u} = (u, v, w)$ , pressure  $p$  and density  $\rho$  are continuous functions of position  $\mathbf{x} = (x, y, z)$  and time  $t$ .

**Crapper's nonlinear capillary waves** Pure progressive capillary waves of arbitrary amplitude.

**Dispersion relation** A mathematical relation between the wavenumber, frequency and/or the amplitude of a wave.

**Euler equations** This is a nonlinear partial differential equation for an inviscid incompressible fluid flow governed by the velocity field  $\mathbf{u} = (u, v, w)$  and pressure  $P(\mathbf{x}, t)$  under the external force  $\mathbf{F}$ . More precisely,  $\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \mathbf{F}$ , where  $\rho$  is the constant density of the fluid.

**Group velocity** The velocity defined by the derivative of the frequency with respect to the wavenumber ( $c_g = d\omega/dk$ ).

**Johnson's equation** This is a nearly concentric KdV equation for  $\eta(R, \xi, \theta)$  in cylindrical polar coordinates in the form

$$\left(2\eta_R + \frac{1}{R}\eta + 3\eta\eta_\xi + \frac{1}{3}\eta\xi\xi\xi + \frac{1}{R^2}\eta\theta\theta\right) = 0.$$

**Kadomtsev–Petviashvili (KP) equation** This is a two-dimensional KdV equation in the form

$$\left(2\eta_t + 3\eta\eta_\xi + \frac{1}{3}\eta_\xi\xi\xi\right)_\xi + \eta_{yy} = 0.$$

**KdV–Burgers equation** This is a nonlinear partial differential equation in the form

$$\eta_t + c_0\eta_x + d\eta\eta_x + \mu\eta_{xxx} - \nu\eta_{xx} = 0,$$

where  $\mu = \frac{1}{6}c_0h_0^2$ .

**Korteweg–de Vries (KdV) equation** This is a nonlinear partial differential equation for a solitary wave (or soliton). This equation in a shallow water of depth  $h$  is governed by the free surface elevation  $\eta(x, t)$  in the form  $\frac{\partial\eta}{\partial t} + c(1 + \frac{3}{2h}\eta)\eta_x + (\frac{ch^2}{6})\eta_{xxx} = 0$ , where  $c = \sqrt{gh}$  is the shallow water speed.

**Laplace equation** This is partial differential equation of the form  $\nabla^2\phi = \phi_{xx} + \phi_{yy} + \phi_{zz}$  where the  $\phi = \phi(x, y, z)$  is the potential.

**Linear dispersion relation** A mathematical relation between the wavenumber  $k$  and the frequency  $\omega$  of waves ( $\omega = \omega(k)$ ).

**Linear dispersive waves** Waves with the given dispersion relation between the wavenumber and the frequency.

**Linear Schrödinger equation** This can be written in the form  $i a_t + \frac{1}{2}\omega''(k)\frac{\partial^2 a}{\partial x^2} = 0$ , where  $a = a(x, t)$  is the amplitude and  $\omega = \omega(k)$ .

**Linear wave equation in Cartesian coordinates**

In three dimensions this can be written in the form  $u_{tt} = c^2\nabla^2u$ , where  $c$  is a constant and  $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$  is the three-dimensional Laplacian.

**Linear wave equation in cylindrical polar coordinates**

This can be written in the form  $u_{tt} = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}$ .

**Navier–Stokes equations** This is a nonlinear partial differential equation for an incompressible and viscous fluid flow governed by the velocity field  $\mathbf{u} = (u, v, w)$  and pressure  $P(\mathbf{x}, t)$  under the action of external force  $\mathbf{F}$ . More precisely,  $\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \nu^2\mathbf{u} + \mathbf{F}$ , where  $\rho$  is the density and  $\nu = (\mu/\rho)$  is the kinematic viscosity.

**Nonlinear dispersion relation** A mathematical relation between the wavenumber  $k$ , frequency  $\omega$  and the amplitude  $a$  that is  $D(\omega, k, a) = 0$ .

**Nonlinear dispersive waves** Waves with the given dispersion relation between the wavenumber, frequency and the amplitude.

**Non-linear Schrödinger (NLS) equation** This is a nonlinear partial differential equation for the nonlinear

modulation of a monochromatic wave. The amplitude,  $a(x, t)$  of the modulation satisfies the equation  $i(\frac{\partial a}{\partial t} + \omega'_0\frac{\partial a}{\partial x}) + \frac{1}{2}\omega''_0\frac{\partial^2 a}{\partial x^2} + \gamma|a|^2a = 0$ , where  $\omega_0 = \omega_0(k)$ , and  $\gamma$  is a constant.

**Ocean waves** Waves observed on the surface or inside the ocean.

**Phase velocity** The velocity defined by the ratio of the frequency  $\omega$  and the wavenumber  $k$ , ( $c_p = \frac{\omega}{k}$ ).

**Resonant or critical phenomenon** Waves with unbounded amplitude.

**Sinusoidal (or exponential) wave** A wave is of the form  $u(x, t) = a \operatorname{Re} \exp[i(kx - \omega t)] = a \cos(kx - \omega t)$ , where  $a$  is amplitude,  $k$  is the wavenumber ( $k = \frac{2\pi}{\lambda}$ ),  $\lambda$  is the wavelength and  $\omega$  is the frequency.

**Solitary waves (or soliton)** Waves describing a single hump of given height travel in a medium without change of shape.

**Stokes expansion** This is an expansion of the frequency in terms of the wavenumber  $k$  and the amplitude  $a$ , that is,  $\omega(k) = \omega_0(k) + \omega_2(k)a^2 + \dots$ .

**Stokes wave** Water waves with dispersion relation involving the wavenumber, frequency and amplitude.

**Surface-capillary gravity waves** Waves under the joint action of the gravitational field and surface tension.

**Surface gravity waves** Water waves under the action of the gravitational field.

**Variational principle** For three-dimensional water waves, it is of the form  $\delta I = \delta \iint_D L \, dx \, dt = 0$ , where  $L$  is called the Lagrangian.

**Velocity potential** A single valued function  $\phi = \phi(\mathbf{x}, t)$  defined by  $\mathbf{u} = \nabla\phi$ .

**Water waves** Waves observed on the surface or inside of a body of water.

**Waves on a running stream** Waves observed on the surface or inside of a body of fluid which is moving with a given velocity.

**Whitham averaged variational principle** This can be formulated in the form  $\delta \iint \mathcal{L} \, dx \, dt = 0$ , where  $\mathcal{L}$  is called the Whitham average Lagrangian over the phase of the integral of the Lagrangian  $L$  defined by  $\mathcal{L}(\omega, \mathbf{k}, a, \mathbf{x}, t) = \frac{1}{2\pi} \int_0^{2\pi} L \, d\theta$ , where  $L$  is the Lagrangian.

**Whitham's conservation equations** These are first order nonlinear partial differential equations in the form  $\frac{\partial k}{\partial t} + \frac{\partial \omega}{\partial x} = 0$ ,  $\frac{\partial}{\partial t} \{f(k)A^2\} + \frac{\partial}{\partial x} \{f(k)C(k)A^2\} = 0$ , where  $k = k(x, t)$  is the density of waves,  $\omega = \omega(x, t)$  is the flux of waves,  $A = A(x, t)$  is the amplitude and  $f(k)$  is an arbitrary function.

**Whitham's equation** This first order nonlinear partial differential equation represents the conservation of waves. Mathematically,  $(\partial k/\partial t) + (\partial \omega/\partial x) = 0$  where

$k = k(x, t)$  is the density of waves and  $\omega = \omega(x, t)$  is the flux of waves.

**Whitham’s equation for slowly varying wavetrain** This is written in the form  $\frac{\partial}{\partial t} \mathcal{L}\omega - \frac{\partial}{\partial x_i} \mathcal{L}k_i = 0$ , where  $\mathcal{L}$  is the Whitham averaged Lagrangian.

**Whitham’s nonlinear nonlocal equations** It is in the form  $u_t + duu_x + \int_{-\infty}^{\infty} K(x-s) u_s(s, t) ds = 0$ , where  $K(x) = \mathcal{F}^{-1} \{c(k) = \omega/k\}$  and  $\mathcal{F}^{-1}$  is the inverse Fourier transformation.

**Definition of the Subject**

A wave is usually defined as the propagation of a disturbance in a medium.

The simplest example is the exponential or sinusoidal wave which has the form

$$\begin{aligned} u(x, t) &= a \operatorname{Re} \exp [i(kx - \omega t)] \\ &= a \cos(kx - \omega t), \end{aligned} \tag{1}$$

where  $a$  is called the *amplitude*,  $\operatorname{Re}$  stands for the real part,  $k (= \frac{2\pi}{\lambda})$  is called the *wavenumber* and  $\lambda$  is called the *wavelength* of the wave, and  $\omega$  is called the *frequency* and it is a definite function of the wavenumber  $k$  and hence,  $\omega = \omega(k)$  is determined by the particular equation of the problem. The quantity  $\theta = kx - \omega t$  is called the *phase* of the wave so that a wave of a constant phase propagate with  $kx - \omega t = \text{constant}$ .

The mathematical relation

$$\begin{aligned} \omega &= \omega(k) \quad \text{or} \\ D(\omega, k) &= 0, \end{aligned} \tag{2}$$

is called the *dispersion relation*.

The *phase* (or *wave*) *velocity* is defined by

$$c(k) = \frac{\omega(k)}{k}. \tag{3}$$

This shows that the phase velocity, in general, depends on the wavenumber  $k$  (or wavelength  $\lambda$ ) so that waves with different wavelength propagate with different phase velocity. The waves are called *dispersive* if the phase velocity is not a constant, but depends on the wavenumber  $k$ . On the other hand, waves are called *nondispersive* if  $c(k)$  is constant, that is, independent of  $k$ .

In general, the dispersion relation (2) can be written in the complex form

$$\begin{aligned} \omega &= \omega(k) \\ &= \sigma(k) + i\nu(k), \end{aligned} \tag{4}$$

where  $\nu(k) < 0$ ,  $\nu(k) = 0$ , or  $\nu(k) > 0$ .

In this case, the sinusoidal wave takes the form

$$u(x, t) = a \exp [\nu(k)t] \exp [i(kx - \sigma t)]. \tag{5}$$

If  $\nu(k) < 0$ , the amplitude of the wave decays to zero as  $t \rightarrow \infty$  and the waves are called *dissipative*. When  $\nu(k) = 0$ , waves are dispersive. On the other hand, if  $\nu(k) > 0$  for some or all  $k$ , the solution grows exponentially as  $t \rightarrow \infty$ . This case corresponds to instability.

The group velocity of the wave (1) is defined by

$$C(k) = \frac{d\omega}{dk}. \tag{6}$$

So, in general,  $c(k) \neq C(k)$ . It is convenient to modify the definition slightly. Waves are called dispersive if  $\omega'(k)$  is not a constant, that is,  $\omega''(k) \neq 0$ .

In higher dimensions, all the above ideas can be generalized without any difficulty. The sinusoidal waves in three space dimensions are defined by

$$u(\mathbf{x}, t) = a \operatorname{Re} \exp [i(\boldsymbol{\kappa} \cdot \mathbf{x} - \omega t)], \tag{7}$$

where  $a$  is the amplitude,  $\mathbf{x} = (x, y, z)$  is the displacement vector,  $\boldsymbol{\kappa} = (k, l, m)$  is the wavenumber vector and  $\omega$  is the frequency which is related to  $\boldsymbol{\kappa}$  by the dispersion relation

$$\omega = \omega(\boldsymbol{\kappa}) \quad \text{or} \quad D(\omega, \boldsymbol{\kappa}) = 0. \tag{8}$$

This function  $D$  is also determined by the particular equation of the problem. In this case,  $\theta = \boldsymbol{\kappa} \cdot \mathbf{x} - \omega t$  is the phase function.

Similarly, the phase velocity of the waves are defined as follows:

$$c(k) = \frac{\omega(\boldsymbol{\kappa})}{\boldsymbol{\kappa} \cdot \hat{\boldsymbol{\kappa}}}, \tag{9}$$

where  $\hat{\boldsymbol{\kappa}} = (\hat{k}, \hat{l}, \hat{m})$  is the unit vector in the  $\boldsymbol{\kappa}$  direction and the group velocity is defined by

$$C(\boldsymbol{\kappa}) = \nabla_{\boldsymbol{\kappa}} \omega(\boldsymbol{\kappa}) = \hat{\mathbf{k}} \frac{\partial \omega}{\partial k} + \hat{\mathbf{l}} \frac{\partial \omega}{\partial l} + \hat{\mathbf{m}} \frac{\partial \omega}{\partial m}. \tag{10}$$

If the dispersion relation (2) depends on the amplitude  $a$  so that the dispersion relation becomes

$$\omega = \omega(k, a) \quad \text{or} \quad D(\omega, k, a) = 0. \tag{11}$$

In general, the *linear plane waves* are recognized by the existence of periodic wavetrains in the form

$$u(x, t) = f(\theta) = f(kx - \omega t), \tag{12}$$

where  $f$  is a periodic function of the phase  $\theta$ .

For linear problems, solutions more general than (1) can be obtained by the principle of superposition to form Fourier integrals as

$$u(x, t) = \int_{-\infty}^{\infty} U(k) \exp [i \{kx - \omega(k)t\}] dk, \quad (13)$$

where the arbitrary function  $U(k)$  may be chosen to fit arbitrary initial or boundary conditions, provided the data are reasonable enough to admit Fourier transform  $U(k) = \mathcal{F} \{u(x, 0)\}$ , and  $\omega = \omega(k)$  is the dispersion relation (2) appropriate to the problem.

On the other hand, the *nonlinear dispersive* waves are also recognized by the existence of periodic wavetrains (12) and the solution must include the amplitude parameter  $a$  and it also requires a nonlinear dispersion relation of the form (11).

In shallow water of constant depth  $h$ , the free surface elevation  $\eta(x, t)$  satisfies the Korteweg and de Vries (KdV) equation

$$\eta_t + c \left( 1 + \frac{3}{2h} \eta \right) \eta_x + \left( \frac{ch^2}{6} \right) \eta_{xxx} = 0, \quad (14)$$

where  $c = \sqrt{gh}$  is the shallow water velocity,  $g$  is the acceleration due to gravity, and the total depth  $H = h + \eta(x, t)$ . The first two terms ( $\eta_t + c \eta_x$ ) describe the wave evolution at speed  $c$ , the third term with the coefficient  $(3c/2h)$  represents the nonlinear wave steepening and the last term with the coefficient  $(ch^2/6)$  describes linear dispersion.

The KdV equation admits an exact solution in the form

$$\eta(x, t) = a \operatorname{sech}^2 \left[ \left( \frac{3a}{4h^3} \right)^{\frac{1}{2}} X \right], \quad (15)$$

where  $X = x - Ut$ , and  $U$  is the wave velocity given by

$$U = C \left( 1 + \frac{a}{2h} \right). \quad (16)$$

The solution (15) is called the *solitary wave* (or *soliton*) describing a single hump of height  $a$  above the undisturbed depth  $h$  and tending rapidly to zero away from  $X = 0$ . The solitary wave propagates to the right with velocity  $U (> c)$  which is directly proportional to the amplitude  $a$  and has width  $b^{-1} = (3a/4h^3)^{-\frac{1}{2}}$ , that is,  $b^{-1}$  is inversely proportional to the square root of the amplitude  $a$ . Another significant feature of the solitary wave is that it travels in the medium *without* change of shape.

In general, the solution for  $\eta(x, t)$  can be expressed in terms of the Jacobian elliptic function  $cn(X, m)$

$$\eta(X) = a cn^2 \left[ \left( \frac{3b}{4h^3} \right)^{\frac{1}{2}} X, m \right], \quad (17)$$

where  $m = (a/b)^{\frac{1}{2}}$  is the modulus of the  $cn$  function. Consequently, the solution (17) is called the *cnoidal wave*.

The *nonlinear Schrödinger (NLS) equation* can be written in the standard form

$$i \psi_t + \psi_{xx} + \gamma |\psi|^2 \psi = 0, \quad -\infty < x < \infty, \quad t > 0, \quad (18)$$

and  $\gamma$  is a constant.

With  $X = x - Ut$ , we seek the solution in the form

$$\psi = \exp [i(mX - nt)] f(X), \quad (19)$$

where  $f(X)$  can be expressed in terms of the Jacobian elliptic function in the form

$$f(X) = (\alpha_1/\alpha_2)^{\frac{1}{2}} sn(\sigma X, \kappa), \quad (20)$$

where  $\alpha_1, \alpha_2, \sigma = (\alpha_2\beta_2/\beta_1\alpha_2)$  and  $\kappa = (\alpha_1\beta_2)/(\beta_1\alpha_2)$  are constants.

The limiting case of the solitary wave is possible and has the form

$$f(X) = \left( \frac{2\alpha}{\gamma} \right)^{\frac{1}{2}} \operatorname{sech} [\sqrt{\alpha}(x - Ut)], \quad (21)$$

where  $\alpha$  and  $\gamma$  are positive constants. This solution represents a solitary wave solution which propagates without change of shape with constant velocity. However, unlike the KdV solitary waves, the amplitude and the velocity are independent parameters. It is important to note that the solution (21) is possible only in the unstable case  $\gamma > 0$ . This suggests that the end result of an unstable wavetrain subject to small modulation is a series of solitary waves.

### Introduction

Water waves are the most common observable phenomena in Nature. The subject of water waves is most fascinating and highly mathematical, and varied of all areas in the study of wave motions in the physical world. The mathematical as well as physical problems deal with water waves and their breaking on beaches, with flood waves in rivers, with ocean waves from storms, with ship waves on water, with free oscillations of enclosed waters such as lakes and harbors. The study of water waves and their various ramifications remain central to fluid dynamics in general, and to the dynamics of oceans in particular. The mathematical theory of water waves is quite interesting in its own merit and intrinsically beautiful. It has provided the solid background and impetus for the development of the theory of nonlinear dispersive waves. Indeed, most of the fundamental ideas and results for nonlinear dispersive waves and solitons originated in the investigation of water waves.

Historically, the problems of water waves in oceans originated from the classic work of Leonhard Euler (1707–1783), A.G. Cauchy (1789–1857), S.D. Poisson (1781–1840), Joseph Boussinesq (1842–1929), George Airy (1801–1892), Lord Kelvin (1824–1907), Lord Rayleigh (1842–1919), George Stokes (1819–1903), Scott Russell (1808–1882) and many others. Indeed, Euler formulated the boundary value problem to understand water wave phenomena and provided successful mathematical and physical description of inviscid and incompressible fluid motion in general. Based on Newton's second law of motion, Euler first formulated his celebration equation of motion of an inviscid fluid about 250 years ago. Euler's equation is still considered the basis of all inviscid fluid flows. In 1821, Claude Navier (1785–1836) included the effect of viscosity to the Euler equation, and first developed the equations of motion of viscous fluid. In 1845, Sir George Stokes provided a sound mathematical foundation of viscous fluid flows and rederived the equations of motion for an incompressible viscous fluid that is universally known as the *Navier–Stokes equations* in the form

$$\frac{D\mathbf{u}}{Dt} = \frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho} \nabla P + \mathbf{F} + \nu \nabla^2 \mathbf{u}, \quad (22)$$

where  $(D/Dt)$  is the *total* (or *convective*) derivative given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + (\mathbf{u} \cdot \nabla), \quad (23)$$

$\mathbf{u}(\mathbf{x}, t) = (u, v, w)$  is the fluid velocity,  $P(\mathbf{x}, t)$  is the pressure field at a point  $\mathbf{x} = (x, y, z)$ ,  $t$  is time,  $\rho$  is a constant density,  $\mathbf{F}(\mathbf{x}, t)$  is a general body force per unit mass,  $\nu = (\mu/\rho)$  is known as kinematic viscosity,  $\mu$  is called the dynamic viscosity,  $\nabla^2 = \nabla \cdot \nabla$  is the *Laplace operator* and  $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$  is the familiar gradient operator.

In particular, when  $\nu = 0$  ( $\mu = 0$ ), the Navier–Stokes Eq. (22) reduces to the celebrated *Euler equation* of motion for an inviscid fluid

$$\frac{D\mathbf{u}}{Dt} = \frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho} \nabla P + \mathbf{F}. \quad (24)$$

Both the Euler Eq. (24) and the Navier–Stokes Eq. (22) form a closed set when the *equation of mass conservation* (or the *continuity equation*)

$$\operatorname{div} \mathbf{u} = \nabla \cdot \mathbf{u} = 0 \quad (25)$$

is added to (22) or (24) so that there are four equations for four unknown quantities  $u$ ,  $v$ ,  $w$ , and  $P$ . The study of these equations is based on the *continuum hypothesis* which

requires that  $\mathbf{u}$ ,  $p$  and  $\rho$  are continuous function of  $\mathbf{x} = (x, y, z)$  and  $t$ .

Remarkably, this 150-year old system of the Navier–Stokes Eqs. (22) and (25) provided the fundamental basis of modern fluid mechanics. However, there are certain major difficulties associated with the Navier–Stokes equations. First, there are no general results for the Navier–Stokes equations on existence of solutions, uniqueness, regularity, and continuous dependence on the initial conditions. Second, another difficulty arises from the strong nonlinear convective term,  $(\mathbf{u} \cdot \nabla)\mathbf{u}$  in Eq. (22).

### The Euler Equation of Motion in Rectangular Cartesian and Cylindrical Polar Coordinates

With  $\mathbf{F} = (0, 0, -g)$ , where  $g$  is the acceleration due to gravity and constant density  $\rho$ , the three components of the Euler's equation of motion in Cartesian coordinates and the continuity equation are

$$\begin{aligned} \frac{Du}{Dt} &= -\frac{1}{\rho} \frac{\partial P}{\partial x}, & \frac{Dv}{Dt} &= -\frac{1}{\rho} \frac{\partial P}{\partial y}, \\ \frac{Dw}{Dt} &= -\frac{1}{\rho} \frac{\partial P}{\partial z} - g, \end{aligned} \quad (26)$$

where the total derivative is given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z}, \quad (27)$$

and

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \quad (28)$$

The basic assumption in continuum mechanics is that the motion of the fluid can be described mathematically as a topological deformation that depends continuously on the time  $t$ . Consequently, we assume the fluid under consideration to have a boundary surface  $S$ , fixed or moving, which separated it from other media. We consider the case of a body of water with air above it so that  $S$  is the interface between them. We represent that surface  $S$  by an equation  $S(x, y, z, t) = 0$ . The kinematic condition is derived from the fact that the normal fluid velocity of the surface  $(-S_t/|\nabla S|)$  is equal to the normal velocity  $(\mathbf{u} \cdot \mathbf{n} = \mathbf{u} \cdot \nabla S/|\nabla S|)$ , that is,

$$\frac{DS}{Dt} = S_t + (\mathbf{u} \cdot \nabla)S = 0. \quad (29)$$

This means that any fluid particle originally on the boundary surface,  $S$  will remain on it.

It is often convenient to represent the free surface by the equation  $z = h(x, y, t)$  so that the equation for the boundary surface  $S$  is

$$S = z - h(x, y, t) = 0, \tag{30}$$

where  $z$  is independent of other variables.

Thus, the kinematic free surface condition follows from (29) in the form

$$w - (h_t + u h_x + v h_y) = 0 \tag{31}$$

on  $z = h(x, y, t), \quad t > 0.$

Since the gravitational force is only body force which acts in the negative  $z$ -direction, the continuity Eq. (28) and the Euler equation in the form

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla P - g \hat{\mathbf{m}}, \tag{32}$$

where  $\hat{\mathbf{m}}$  is the unit vector in the positive  $z$ -direction, represent the fundamental equations for water wave motion.

In general, the water wave motion is unsteady, and irrotational which physically means that the individual fluid particle do not rotate. Mathematically, this implies that vorticity  $\boldsymbol{\omega} = \text{curl } \mathbf{u} = \mathbf{0}$ . So, there exists a single-valued velocity potential  $\phi$  so that  $\mathbf{u} = \nabla \phi$ , where  $\phi = \phi(\mathbf{x}, t)$ . Consequently, the continuity equation reduces to the Laplace equation

$$\nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 0. \tag{33}$$

So,  $\phi$  is a harmonic function. This is, indeed, a great advantage because the velocity field  $\mathbf{u}$  can be obtained from a single potential function  $\phi$  which satisfies a linear partial differential equation. This equation with prescribed boundary conditions can readily be solved in many simple cases without difficulty.

In view of a vector identity

$$(\mathbf{u} \cdot \nabla) \mathbf{u} = \frac{1}{2} \nabla u^2 - \mathbf{u} \times \boldsymbol{\omega} \tag{34}$$

combined with  $\boldsymbol{\omega} = \text{curl } \mathbf{u} = \mathbf{0}$  and  $\mathbf{u} = \nabla \phi$ , the Euler Eq. (32) may be written as

$$\nabla \left[ \phi_t + \frac{1}{2} (\nabla \phi)^2 + \frac{P}{\rho} + gz \right] = 0. \tag{35}$$

This can be integrated with respect to the space variables to give the equation for pressure  $P$  at every point of the fluid

$$\phi_t + \frac{1}{2} (\nabla \phi)^2 + \frac{P}{\rho} + gz = C(t), \quad t \geq 0, \tag{36}$$

where  $C(t)$  is an arbitrary function of time only ( $\nabla C = 0$ ). Since the pressure gradient affects the flow, a function of  $t$  alone added to the pressure field  $P$  has no effect on the motion. So, without loss of generality, we can set  $C(t) = 0$  in Eq. (36). Thus, the pressure Eq. (36) becomes

$$\phi_t + \frac{1}{2} (\nabla \phi)^2 + \frac{P}{\rho} + gz = 0. \tag{37}$$

This is the so-called the *Bernoulli's equation* which determines the pressure in terms of the velocity potential  $\phi$ . Thus, Eqs. (33) and (37) are used to determine the potential  $\phi$  (hence, three velocity components  $u, v,$  and  $w$ ) and the pressure field  $P$ .

We consider a body of inviscid, incompressible fluid occupying the region  $b(x, y) \leq z \leq h(x, y, t) = h_0 + a\eta(x, y, t)$ , where  $z = b(x, y)$  is the bottom boundary surface and  $z = h_0$  is the undisturbed (initial) typical constant depth,  $a$  is the typical amplitude and  $\eta(x, y, t)$  is nondimensional the free surface elevation that tends to zero as  $t \rightarrow 0$ . We suppose that the bottom boundary  $z = b(x, y)$  is a rigid solid surface.

Since the upper boundary is the surface exposed to a constant atmospheric pressure  $P_a$ , we have  $P = P_a$  on this surface,  $S$ . Thus, Eq. (35) assumes the form

$$\phi_t + \frac{1}{2} (\nabla \phi)^2 + \frac{1}{\rho} P_a + gz = C(t) \tag{38}$$

on  $S, \quad t \geq 0.$

Absorbing  $\frac{1}{\rho} P_a$  and  $C(t)$  into  $\phi_t$ , this equation may be rewritten as

$$\phi_t + \frac{1}{2} (\nabla \phi)^2 + gz = 0, \quad \text{on } S, \quad t \geq 0. \tag{39}$$

Since  $S$  is a upper boundary surface of the fluid, it contains the same fluid particles for all time  $t$ , that is,  $S$  is a material surface. Hence, it follows from (39) that

$$\frac{D}{Dt} \left[ \phi_t + \frac{1}{2} (\nabla \phi)^2 + gz \right] = 0, \tag{40}$$

on  $S, \quad t \geq 0.$

Or, equivalently,

$$\begin{aligned} & \left[ \frac{\partial}{\partial t} + (\nabla \phi \cdot \nabla) \right] \left[ \frac{\partial \phi}{\partial t} + \frac{1}{2} (\nabla \phi)^2 + gz \right] \\ &= \phi_{tt} + 2 \nabla \phi \cdot \nabla (\phi_t) + \frac{1}{2} \nabla \phi \cdot \nabla (\nabla \phi)^2 + g \phi_z \\ &= 0, \quad \text{on } S, \quad t \geq 0. \end{aligned} \tag{41}$$

We next include the effects of *surface tension force* per unit length which does support a pressure difference across

a curved surface so that  $P - P_a = -(T/R)$ , where  $R^{-1} = (R_1^{-1} + R_2^{-1})$  is called the *Gaussian curvature* expressed as the sum of two principal radii of curvatures  $R_1^{-1}$  and  $R_2^{-1}$  given by

$$\begin{aligned} R_1^{-1} &= \frac{\partial}{\partial x} \left[ h_x (1 + h_x^2 + h_y^2)^{-\frac{1}{2}} \right], \\ R_2^{-1} &= \frac{\partial}{\partial y} \left[ h_y (1 + h_x^2 + h_y^2)^{-\frac{1}{2}} \right]. \end{aligned} \quad (42ab)$$

Explicitly,  $R^{-1}$  can be written in terms of  $h(x, y, t)$  and its partial derivatives as

$$\begin{aligned} R^{-1} &= (R_1^{-1} + R_2^{-1}) \\ &= \frac{h_{xx}(1 + h_y^2) - 2h_x h_y h_{xy} + h_{yy}(1 + h_x^2)}{(1 + h_x^2 + h_y^2)^{3/2}}. \end{aligned} \quad (43)$$

For small deviation of the free surface  $z = h(x, y, t)$ , its first partial derivatives  $h_x$  and  $h_y$  are small so that (43) takes the linearized form

$$R^{-1} \approx (h_{xx} + h_{yy}). \quad (44)$$

Consequently, the linearized dynamic condition at the free surface  $z = h$  becomes

$$\phi_t + gz - \frac{T}{\rho}(h_{xx} + h_{yy}) = 0. \quad (45)$$

For an inviscid fluid, like the free surface condition (29), there is a bottom boundary condition which follows from the fact that

$$\frac{D}{Dt} [z - b(x, y, t)] = 0, \quad (46)$$

where  $z = b(x, y, t)$  is the equation of the fixed solid bottom boundary surface. Thus, Eq. (46) assumes the form

$$\begin{aligned} w &= b_t + (\mathbf{u} \cdot \nabla)b = b_t + u b_x + v b_y \\ &\text{on } z = b(x, y, t), \end{aligned} \quad (47)$$

where  $z = b(x, y, t)$  is given. However, there is a class of problems including sediment movement, where  $b$  is not known. For stationary bottom,  $b$  is independent of time so that (47) becomes

$$w = u b_x + v b_y \quad \text{on } z = b(x, y). \quad (48)$$

For one-dimensional problem,  $h = b(x)$  with  $\mathbf{u} = (u, 0)$  so that (48) reduces to the simple bottom condition

$$w = u b'(x) \quad \text{on } z = b(x). \quad (49)$$

It is convenient to introduce a typical amplitude parameter  $a$  by writing the free surface  $z = h(x, y, t)$  in the form

$$h = h_0 + a\eta(x, y, t), \quad (50)$$

where  $h_0$  is the undisturbed depth of water. The pressure field can be rewritten as

$$P = P_a + g\rho(h_0 - z) + (g\rho h_0)p, \quad (51)$$

where  $P_a$  is the constant atmospheric pressure,  $p$  is the pressure variable which measures the deviation from the hydrostatic pressure,  $g\rho(h_0 - z)$ , and  $g\rho h_0$  is the typical pressure scale based on the pressure at depth  $h = h_0$ .

We next introduce two fundamental parameters  $\varepsilon$  and  $\delta$  as

$$\varepsilon = \frac{a}{h_0} \quad \text{and} \quad \delta = \frac{h_0^2}{\lambda^2}, \quad (52)$$

where  $\lambda$  is the typical wavelength of the surface gravity wave,  $\varepsilon$  and  $\delta$  are called the *amplitude* and *long wavelength* parameters respectively.

In terms of typical depth scale  $h_0$ ,  $\lambda$  is the typical wavelength,  $c_0 = \sqrt{g h_0}$  is the typical horizontal velocity scale,  $(\lambda/c_0)$  is the typical time scale, it is also convenient to introduce non-dimensional flow variables denoted by asterisks

$$\begin{aligned} (x^*, y^*) &= \frac{1}{\lambda}(x, y), \quad (z^*, b^*) = \frac{1}{h_0}(z, b), \\ t^* &= \left(\frac{c_0}{\lambda}\right)t, \end{aligned} \quad (53)$$

$$\begin{aligned} (u^*, v^*) &= \frac{1}{c_0}(u, v), \quad w^* = \left(\frac{\lambda c_0}{h_0}\right)w \quad \text{and} \\ p^* &= \frac{P}{g\rho h_0}. \end{aligned} \quad (54)$$

In terms of these nondimensional flow variables and parameters, the Euler Eqs. (26) and the continuity Eq. (28) can be rewritten, dropping the asterisks, in the form

$$\frac{D}{Dt}(u, v) = -\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)p, \quad \delta \frac{Dw}{Dt} = -\frac{\partial p}{\partial z}, \quad (55)$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z},$$

and

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \quad (56)$$

The most general dynamic condition is  $P = P_a - \frac{T}{R}$  on the free surface  $z = h = h_0 + a\eta$ . Without surface tension ( $T = 0$ ), it becomes  $P = P_a$  on  $z = h_0 + a\eta$ . Using the pressure field (51), this free surface dynamic condition reduces to the nondimensional form  $p = \varepsilon\eta$  on  $z = 1 + \varepsilon\eta$ . Thus, the nondimensional kinematic condition (31) and the dynamic condition at the free surface are given by

$$\begin{aligned} w - \varepsilon(\eta_t + u\eta_x + v\eta_y) &= 0, \\ p &= \varepsilon\eta \quad \text{on } z = 1 + \varepsilon\eta. \end{aligned} \quad (57ab)$$

The nondimensional form of the bottom boundary condition (48) remains the unchanged form.

Consistent with the governing equations and free surface boundary conditions, we introduce a set of scaled flow variables

$$(u, v, w, p) \rightarrow \varepsilon(u, v, w, p) \quad (58)$$

so that the Euler Eqs. (55) and the continuity Eq. (56) reduce to the form

$$\frac{D}{Dt}(u, v) = -\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)p, \quad \delta \frac{Dw}{Dt} = -\frac{\partial p}{\partial z}, \quad (59)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0, \quad (60)$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \varepsilon\left(u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}\right). \quad (61)$$

The free surface boundary conditions (57ab) remain the same. The horizontal bottom ( $b = 0$ ) boundary condition is

$$w = 0, \quad \text{on } z = 0, \quad (62)$$

In cylindrical polar coordinates  $(r, \theta, z)$ , the Euler equations and the continuity equation are given by

$$\begin{aligned} \frac{Du}{Dt} - \frac{v^2}{r} &= -\frac{1}{\rho}P_r, & \frac{Dv}{Dt} + \frac{uv}{r} &= -\frac{1}{\rho} \frac{1}{r}P_\theta, \\ \frac{Dw}{Dt} &= -\frac{1}{\rho}P_z - g, \end{aligned} \quad (63)$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial r} + \frac{v}{r}\frac{\partial}{\partial \theta} + w\frac{\partial}{\partial z}, \quad (64)$$

and

$$\frac{1}{r}\frac{\partial}{\partial r}(ru) + \frac{1}{r}\frac{\partial v}{\partial \theta} + \frac{\partial w}{\partial z} = 0. \quad (65)$$

In terms of the above nondimensional flow variables except for  $x$  and  $y$  that are replaced by  $r^* = \frac{1}{\lambda}r$ , the above Eqs. (63)–(65) assume the following nondimensional form, dropping the asterisks,

$$\begin{aligned} \frac{Du}{Dt} - \frac{v^2}{r} &= -\frac{\partial p}{\partial r}, & \frac{Dv}{Dt} + \frac{uv}{r} &= -\frac{1}{r}\frac{\partial p}{\partial \theta}, \\ \delta \frac{Dw}{Dt} &= -\frac{\partial p}{\partial z}, \end{aligned} \quad (66)$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial r} + \frac{v}{r}\frac{\partial}{\partial \theta} + w\frac{\partial}{\partial z}, \quad (67)$$

and

$$\frac{1}{r}\frac{\partial}{\partial r}(ru) + \frac{1}{r}\frac{\partial v}{\partial \theta} + \frac{\partial w}{\partial z} = 0. \quad (68)$$

### Basic Equations of Water Waves with Effects of Surface Tension

We consider an irrotational unsteady motion of an inviscid and incompressible water occupying the region  $b(x, y) \leq z \leq h_0 + a\eta(x, y, t)$  in a constant gravitational field  $g$  which is in the negative  $z$ -direction. The equation of uneven bottom boundary is  $z = b(x, y)$ . Including the effects of surface tension  $T$ , the basic equations, two free surface boundary conditions and the bottom boundary condition for the velocity potential  $\phi = \phi(x, y, z, t)$  and the free surface elevation  $\eta = \eta(x, y, t)$  are

$$\begin{aligned} \nabla^2 \phi &= \phi_{xx} + \phi_{yy} + \phi_{zz} = 0, \\ b(x, y) &\leq z \leq h_0 + a\eta, \quad t > 0, \end{aligned} \quad (69)$$

$$\begin{aligned} \eta_t + (\phi_x \eta_x + \phi_y \eta_y) - \phi_z &= 0, \\ &\text{on } z = h_0 + a\eta, \quad t > 0, \end{aligned} \quad (70)$$

$$\begin{aligned} \phi_t + \frac{1}{2}(\nabla \phi)^2 + g\eta - \frac{T}{\rho}(R_1^{-1} + R_2^{-1}) &= 0, \\ &\text{on } z = h_0 + a\eta, \quad t > 0, \end{aligned} \quad (71)$$

$$\phi_z - (\phi_x b_x + \phi_y b_y) = 0, \quad \text{on } z = b(x, y), \quad (72)$$

where  $R_1^{-1}$  and  $R_2^{-1}$  are given by (42ab),  $h_0$  is the typical depth of water and  $a$  is the typical amplitude of the surface gravity wave.

In water of infinite depth with the origin at the free surface, the bottom boundary condition (72) is replaced by  $(\nabla \phi) \rightarrow 0$  as  $z \rightarrow -\infty$ .

Because of the presence of nonlinear terms in the free surface boundary conditions (70)–(71), the determination of  $\phi$  and  $\eta$  in the general case is a difficult task.

Similarly, we can write the basic equations for the water wave problems in cylindrical polar coordinates.

In terms of the nondimensional flow variables and the parameters  $\varepsilon$  and  $\delta$  stated above, the basic water wave Eqs. (69)–(72) without surface tension can be written in the nondimensional form

$$\delta(\phi_{xx} + \phi_{yy}) + \phi_{zz} = 0, \quad \text{on } b \leq z \leq 1 + \varepsilon\eta, \quad t > 0, \quad (73)$$

$$\delta[\eta_t + \varepsilon(\phi_x\eta_x + \phi_y\eta_y)] - \phi_z = 0, \quad \text{on } z = 1 + \varepsilon\eta, \quad (74)$$

$$\phi_t + \eta + \frac{\varepsilon}{2}(\phi_x^2 + \phi_y^2) + \frac{\varepsilon}{2\delta}\phi_z^2 = 0, \quad \text{on } z = 1 + \varepsilon\eta, \quad (75)$$

$$\phi_z - \delta(\phi_x b_x + \phi_y b_y) = 0, \quad \text{on } z = b(x, y), \quad (76)$$

where  $(\varepsilon/\delta) = (a\lambda^2/h_0^3)$  is another fundamental parameter in water-wave theory.

Luke [47] first explicitly formulated a variational principle for two-dimensional water waves and proved that the basic Laplace equation, free surface and bottom boundary conditions can be derived from the Hamilton principle. We formulate the variational principle for three-dimensional water waves in the form

$$\delta I = \delta \iint_D L \, dx \, dt = 0, \quad (77)$$

where the Lagrangian  $L$  is assumed to be equal to the pressure so that

$$L = -\rho \int_{-h(x,y)}^{\eta(x,t)} \left[ \phi_t + \frac{1}{2}(\nabla\phi)^2 + gz \right] dz, \quad (78)$$

where  $D$  is an arbitrary region in the  $(\mathbf{x}, t)$  space, and  $\phi(\mathbf{x}, z, t)$  is the velocity potential of an unbounded fluid lying between the rigid bottom  $z = -h(x, y)$  and the free boundary surface  $z = \eta(x, y, t)$ . Using the standard procedure in the calculus of variations (see Debnath [19]), the following nonlinear system of equations for the classical water waves can be derived:

$$\nabla^2\phi = 0, \quad -h < z < \eta, \quad -\infty < (x, y) < \infty, \quad (79)$$

$$\eta_t + (\phi_x\eta_x + \phi_y\eta_y) - \phi_z = 0, \quad \text{on } z = \eta, \quad (80)$$

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + gz = 0, \quad \text{on } z = \eta, \quad (81)$$

$$\phi_z + \phi_x h_x + \phi_y h_y = 0, \quad \text{on } z = -h. \quad (82)$$

These results are in perfect agreement with those of Luke for the two-dimensional waves on water of arbitrary but uniform depth  $h$ . In his pioneering work, Whitham [58,59] first developed a general approach to linear and nonlinear dispersive waves using a Lagrangian. It is now well known that most of the general ideas about dispersive waves have originated from the classical problems of water waves.

Making reference to Debnath [19], we first state the solution of the linearized two-dimensional problem of the classical water waves on water of uniform depth  $h$  governed by the following equation, free surface and boundary conditions

$$\phi_{xx} + \phi_{zz} = 0, \quad -h \leq z \leq 0, \quad t > 0, \quad (83)$$

$$\eta_t = \phi_z, \quad \phi_t + g\eta = 0, \quad \text{on } z = 0, \quad t > 0, \quad (84a)$$

$$\phi_z = 0, \quad \text{on } z = -h, \quad t \geq 0. \quad (85)$$

The solutions for  $\phi(x, z, t)$  and  $\eta(x, t)$  representing a sinusoidal wave propagating in the  $x$ -direction are given by

$$\phi(x, z, t) = \text{Re } a \left( \frac{ig}{\omega} \right) \frac{\cosh k(z+h)}{\cosh kh} \exp [i(\omega t - kx)], \quad (86)$$

$$\eta(x, t) = \text{Re } a \exp [i(\omega t - kx)], \quad (87)$$

where  $a = (C\omega/ig) \cosh kh = \max |\eta|$  is the amplitude and  $C$  is an arbitrary constant.

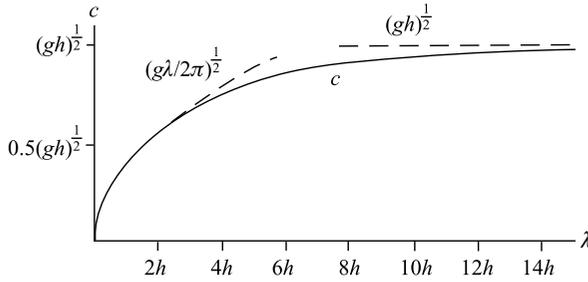
Using (84ab), we obtain the celebrated *dispersion relation* between the frequency  $\omega$  and the wavenumber  $k$  in the form:

$$\omega^2 = gk \tanh kh. \quad (88)$$

Physically, this relation describes the interaction between inertial and gravitational forces. This can also be rewritten in terms of the wave (or phase) velocity,  $c(k)$  as

$$c(k) = \frac{\omega}{k} = (gk^{-1} \tanh kh)^{\frac{1}{2}} = \left[ \left( \frac{g\lambda}{2\pi} \right) \tanh \left( \frac{2\pi h}{\lambda} \right) \right]^{\frac{1}{2}}. \quad (89)$$

This formula shows that the phase velocity  $c(k)$  depends on the gravity  $g$  and depth  $h$  as well as the wavenumber  $k$  or wavelength  $\lambda = (2\pi/k)$ . Thus, water waves of different wavelengths travel with different wave (or phase) velocities. Such waves are called *dispersive*, as time passes, these waves disperse (or spread out) into various group of waves such that each group of waves such that each group would consist of waves having approximately the



**Water Waves and the Korteweg–de Vries Equation, Figure 1**  
 The phase velocity  $c(\lambda)$  against the wavelength  $\lambda$  (from [46], Figure 52, page 217)

same wavelength. Figure 1 showing a plot of the phase velocity  $c$  given by (89) against the wavelength  $\lambda$  reveals a transition between the deep water limit  $c \sim (g\lambda/2\pi)^{\frac{1}{2}}$  (parabolic form) when  $\lambda < (3.5)h$  and the shallow water limit  $c \sim \sqrt{gh}$  for  $\lambda > (14)h$ .

The quantity  $(d\omega/dk)$  represents the velocity of such a group in the direction of propagation and is called the *group velocity*, denoted by  $C(k)$  so that

$$C(k) = \frac{d\omega}{dk} = \left(\frac{g}{2\omega}\right) (\tanh kh + kh \operatorname{sech}^2 kh), \quad (90)$$

which is, using (89)

$$= \frac{1}{2} c(k) [1 + 2kh \operatorname{cosech}(2kh)]. \quad (91)$$

Evidently, the group velocity is, in general, different from the phase velocity.

Two limiting cases are of special interest: (i) Shallow water waves and (ii) Deep water waves.

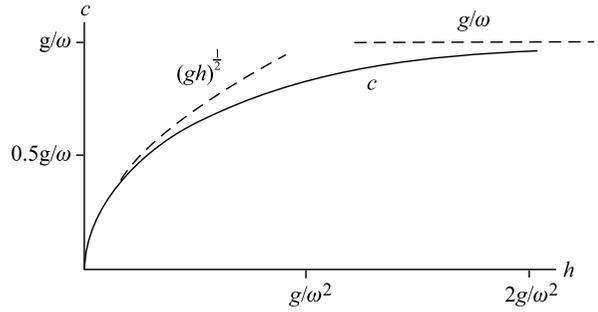
For shallow water, the wavelength;  $\lambda = (2\pi/k)$  is large compared with the depth  $h$  so that  $kh \ll 1$ , and hence,  $\tanh kh \approx kh$  and  $\sin 2kh \approx 2kh$ . In such a case, (88)–(91) reduce to

$$\omega^2 = gk h^2, \quad c(k) = \sqrt{gh} = C(k). \quad (92)$$

Both phase and group velocities are independent of the wavenumber  $k$ . Thus, shallow water waves are nondispersive, and their phase velocity is equal to group velocity. Both vary as the square root of the depth  $h$ .

In the other limiting case dealing with deep water waves, the wavelength is very small compared with the depth so that  $kh \gg 1$ . In the limit as  $kh \rightarrow \infty$ ,  $\tanh kh \rightarrow 1$ ,  $[\cosh k(z+h)/\cosh kh] \rightarrow \exp(kz)$ , and the corresponding solutions for  $\phi$  and  $\eta$  become

$$\begin{aligned} \phi &= \operatorname{Re} \left( \frac{ia g}{\omega} \right) e^{kz} \exp [i(\omega t - kx)] \\ &= \left( \frac{ag}{\omega} \right) e^{kz} \sin(kx - \omega t), \end{aligned} \quad (93)$$



**Water Waves and the Korteweg–de Vries Equation, Figure 2**  
 The phase velocity  $c$  for waves of frequency  $\omega$  on water of depth  $h$  (from [46], Figure 53, page 218)

$$\eta = \operatorname{Re} a \exp [i(\omega t - kx)] = a \cos(kx - \omega t). \quad (94)$$

Evidently, for deep water waves, the dispersion relation (88), the phase velocity (89) and the group velocity (90) become

$$\begin{aligned} \omega^2 = gk = \left( \frac{2\pi g}{\lambda} \right), \quad c(k) = (g/k)^{\frac{1}{2}} = \left( \frac{g\lambda}{2\pi} \right)^{\frac{1}{2}}, \\ C(k) = \frac{1}{2} c(k). \end{aligned} \quad (95abc)$$

Thus, deep water waves are dispersive and their phase velocity is proportional to the square root of their wavelengths. Also, the group velocity is equal to one-half of the phase velocity.

Another equivalent form of the phase velocity  $c(k)$  follows from the dispersion relation (88) in the form

$$c(k) = \frac{\omega}{k} = \left( \frac{\omega^2}{\omega k} \right) = \frac{g}{\omega} \tanh \left( \frac{\omega h}{c} \right). \quad (96)$$

The phase velocity  $c(k)$  for water waves of frequency  $\omega$  against depth  $h$  is shown in Fig. 2. This figure shows a transition between the deep water limit  $c \sim (g/\omega)$  when  $(\omega h/c) \gg 2$  or  $h \geq 2g/\omega^2$  and the shallow water limit  $c \sim \sqrt{gh}$  when  $(\omega h/c) \ll 1$  or  $h \ll g/\omega^2$ .

Similarly, an alternative form of the group velocity follows from the dispersion relation (88) as

$$C(k) = \frac{d\omega}{dk} = \frac{d}{dk}(c k) = c + k \frac{dc}{dk} = c - \lambda \frac{dc}{d\lambda}. \quad (97)$$

Finally, we close the section by adding surface capillary-gravity waves on water of constant depth  $h$  with the free surface at  $z = 0$ . For such waves, the linearized free surface conditions (70)–(71) with the effect of surface tension are modified as follows:

$$\eta_t = \phi_z, \quad \phi_t + g\eta - \frac{T}{\rho} \eta_{xx} = 0 \quad \text{on } z = 0. \quad (98)$$

These conditions can be combined to obtain

$$(\phi_{tt} + g\phi_z) - \frac{T}{\rho} \phi_{xxz} = 0 \quad \text{on } z = 0. \quad (99)$$

In this case, the solutions for the velocity potential  $\phi(x, z, t)$  and the free surface elevation  $\eta(x, t)$  are exactly the same as (86) and (87) on water of constant depth  $h$  or as (93)–(94) for deep water. Evidently, the dispersion relation for capillary-gravity waves on water of constant depth  $h$  is given by

$$\omega^2 = gk \left( 1 + \frac{Tk^2}{\rho g} \right) \tanh kh. \quad (100)$$

The phase velocity is

$$c(k) = \left[ \frac{g}{k} \left( 1 + \frac{Tk^2}{\rho g} \right) \tanh kh \right]^{\frac{1}{2}}. \quad (101)$$

The corresponding dispersion relation for deep water ( $kh \gg 1$ ) is

$$\omega^2 = gk(1 + T^*), \quad (102)$$

where  $T^* = (Tk^2/g\rho) = (4\pi^2 T/g\rho\lambda^2)$  is a parameter giving the relative importance of surface tension and gravity. The phase velocity is

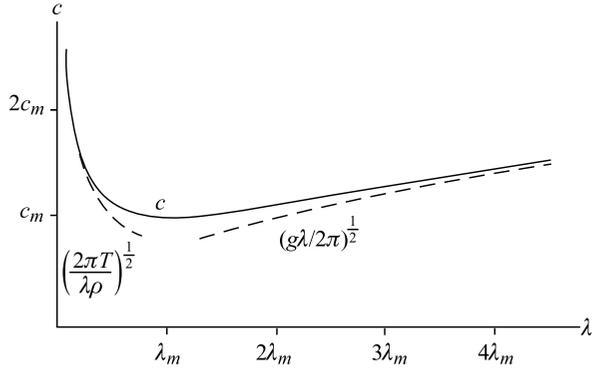
$$c(k) = \left[ \frac{g}{k} (1 + T^*) \right]^{\frac{1}{2}} = \left[ \left( \frac{g\lambda}{2\pi} \right) (1 + T^*) \right]^{\frac{1}{2}}. \quad (103)$$

The phase velocity  $c(k)$  has a minimum value at  $k = k_m = \sqrt{g\rho/T}$  (or  $T^* = 1$ ). The corresponding minimum value for  $c = c_m$  attained at  $k = k_m = \sqrt{g\rho/T}$  (or  $T^* = 1$ ) is

$$c = c_m = \left( \frac{4gT}{\rho} \right)^{\frac{1}{4}} \quad (104)$$

at wavelength  $\lambda = \lambda_m = 2\pi(T/g\rho)^{\frac{1}{2}}$ .

The inequality  $k \ll k_m$  holds for waves to be effectively pure gravity waves, since the surface tension is negligible by comparison. This inequality is equivalent to the wavelength  $\lambda$  to be large compared with  $\lambda_m = (2\pi/k_m) = 2\pi(T/g\rho)^{\frac{1}{2}}$ . The phase velocity  $c(\lambda)$  for capillary-gravity waves against  $\lambda$  in deep water is shown in Fig. 3. This figure shows the transition between the pure capillary-wave value  $(2\pi T/\lambda\rho)^{\frac{1}{2}}$  and the pure gravity-wave value  $(g\lambda/2\pi)^{\frac{1}{2}}$ . Also shown in this figure are (i) the minimum phase velocity  $c_m$  attained at  $\lambda = \lambda_m$ , (ii) the gravity-wave curve ( $T^* = 0$ ) for  $\lambda > \lambda_m$ , and (iii) the capillary-wave curve with no gravitational effect  $g \rightarrow 0$  or  $T^* \rightarrow \infty$ . Figure 3 also shows that for  $\lambda > 4\lambda_m$  the capil-



Water Waves and the Korteweg–de Vries Equation, Figure 3  
The phase velocity  $c(\lambda)$  given by (103) against  $\lambda$  (from [46], Figure 56, page 224)

lary-gravity wave curve is rapidly running closer with the gravity-wave curve. On the other hand, when  $\lambda < \lambda_m$ , the rapid tendency is for wave speeds to increase again so that the capillary-gravity wave curve approaches the capillary wave curve as  $T^* \rightarrow \infty$ , which corresponds to very short waves called *capillary waves* (or *ripples*). In fact, when  $\lambda < \frac{1}{4}\lambda_m$ , results (102) and (103) for  $T^* \rightarrow \infty$  give

$$\omega^2 = \rho^{-1}Tk^3 \quad \text{and} \quad c(k) = (\rho^{-1}Tk)^{\frac{1}{2}}. \quad (105ab)$$

For such capillary waves, surface tension is the only significant restoring force.

### The Stokes Waves and Nonlinear Dispersion Relation

In his 1847 classic paper, Stokes [54] first established the nonlinear solutions for periodic plane waves on deep water. We consider here some of the nonlinear effects neglected in the linearized theory. A more simple approach is to recall the exact free surface dynamic condition (41) with constant atmospheric pressure and the negligible surface tension so that (41) becomes

$$(\phi_{tt} + g\phi_z) + 2\nabla\phi \cdot \nabla\phi_t + \frac{1}{2}\nabla\phi \cdot \nabla(\nabla\phi)^2 = 0, \quad \text{on } z = \eta, \quad \text{for } t \geq 0. \quad (106)$$

This condition is applied on the unknown free surface  $z = \eta$  given by (39), that is,

$$\eta = -\frac{1}{g} \left[ \phi_t + \frac{1}{2}(\nabla\phi)^2 \right]. \quad (107)$$

In the absence of nonlinear terms, results (106) and (107) reduce to (84ab) and (84b), respectively.

We use Taylor series expansions of  $\phi$  and its derivatives from  $z = \eta$  to  $z = 0$  in the form

$$\phi(x, y, \eta, t) = \phi(x, y, 0, t) + \eta \left( \frac{\partial \phi}{\partial z} \right)_{z=0} + \frac{1}{2} \eta^2 \left( \frac{\partial^2 \phi}{\partial z^2} \right)_{z=0} + \dots \quad (108)$$

Using this expansion procedure for each of the derivatives in (106) and (107), we can generate a series of boundary conditions on the surface plane  $z = 0$ . Thus, we obtain first three conditions:

$$L\phi = \phi_{tt} + g\phi_z = 0 + O(\phi^2), \quad (109)$$

$$L\phi + 2\nabla\phi \cdot \nabla\phi_t - \frac{1}{g} \phi_t \frac{\partial}{\partial z}(L\phi) = 0 + O(\phi^3), \quad (110)$$

$$\begin{aligned} L\phi + 2\nabla\phi \cdot \nabla\phi_t &+ \frac{1}{2} \nabla\phi \cdot \nabla(\nabla\phi)^2 \\ &- \frac{1}{g} \phi_t \frac{\partial}{\partial z}(L\phi + 2\nabla\phi \cdot \nabla\phi_t) \\ &- \frac{1}{g} \left[ -\frac{1}{g} \phi_t \phi_{zt} + \frac{1}{2} (\nabla\phi)^2 \right] \frac{\partial}{\partial z}(L\phi) \\ &+ \frac{1}{2g^2} \phi_t^2 \frac{\partial^2}{\partial z^2}(L\phi) \\ &= 0 + O(\phi^4), \end{aligned} \quad (111)$$

where the symbol  $O(\cdot)$  is used to indicate the magnitude of the neglected terms.

If we substitute the first-order velocity potential (93) for deep water in the second-order boundary condition (110), the second-order terms in (110) vanish. Thus, the first-order potential is a solution of the second-order boundary value problem, and we can write

$$\phi = \left( \frac{ga}{\omega} \right) e^{kz} \sin(kx - \omega t) + O(a^3). \quad (112)$$

Similarly, using (108), we can incorporate the second-order effects in  $\eta(x, z, t)$  so that

$$\begin{aligned} \eta(x, z, t) &= -\frac{1}{g} \left[ \phi_t + \frac{1}{2} (\nabla\phi)^2 \right]_{z=\eta} \\ &= -\frac{1}{g} \left[ \phi_t + \frac{1}{2} (\nabla\phi)^2 \right]_{z=0} \\ &+ \eta \frac{\partial}{\partial z} \left[ -\frac{1}{g} \left\{ \phi_t + \frac{1}{2} (\nabla\phi)^2 \right\} \right]_{z=0} + \dots \\ &= -\frac{1}{g} \left[ \phi_t + \frac{1}{2} (\nabla\phi)^2 - \frac{1}{g} \phi_t \phi_{zt} \right]_{z=0} + \dots \quad (113) \end{aligned}$$

Direct substitution of (112) in (113) gives the following second-order result:

$$\begin{aligned} \eta &= a \cos(kx - \omega t) \\ &+ \frac{1}{2} ka^2 \{ 2 \cos^2(kx - \omega t) - 1 \} + \dots \\ &= a \cos \theta + \frac{1}{2} ka^2 \cos 2\theta + \dots, \end{aligned} \quad (114)$$

where the phase  $\theta = kx - \omega t$ . Clearly, the second term in (114) represents the nonlinear effects on the free-surface profile, and it is positive both at the crests  $\theta = 0, 2\pi, 4\pi, \dots$  and at the troughs  $\theta = \pi, 3\pi, 5\pi, \dots$ . The notable feature of this solution (114) is that the wave profile is no longer sinusoidal as shown in Fig. 2.7 of Debnath [19].

The actual shape of this wave profile is a curve known as a *trochoid*: The crests are steeper and the troughs are flatter. This feature becomes accentuated as the wave amplitude is increased.

For the third-order free-surface condition (111), direct substitution of the plane wave potential (112) in the nonlinear terms in (111) eliminates all but one term. Therefore, the free-surface boundary condition for the third-order plane wave solution is

$$L\phi + \frac{1}{2} \nabla\phi \cdot \nabla(\nabla\phi)^2 = 0 + O(\phi^4). \quad (115)$$

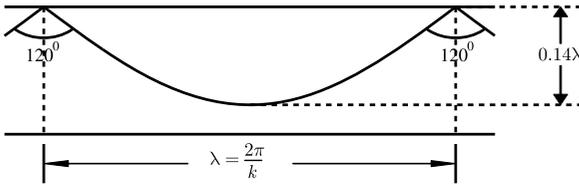
The first-order solution (112) satisfies this third-order boundary condition so that the dispersion relation (95a) includes a second-order effect of the form

$$\omega^2 = gk(1 + a^2k^2) + O(a^3k^3). \quad (116)$$

This remarkable dependence of frequency on wave amplitude is usually known as *amplitude* (or *nonlinear*) *dispersion*. The modified phase velocity expression is

$$\begin{aligned} c(k) = \frac{\omega}{k} &= \left( \frac{g}{k} \right)^{\frac{1}{2}} (1 + k^2 a^2)^{\frac{1}{2}} \\ &\approx \left( \frac{g}{k} \right)^{\frac{1}{2}} \left( 1 + \frac{1}{2} a^2 k^2 \right). \end{aligned} \quad (117)$$

The significant change from the linearized theory is (116) or (117), which confirms that the phase velocity now depends on amplitude  $a$  as well as on wavelength; steep waves travel faster than less steep waves of the same wavelength. Surface gravity waves thus acquire amplitude dispersion as a second-order correction. The dependence of  $c$  on amplitude is known as the *amplitude dispersion* in contrast to the *frequency dispersion* as given by (95a). It may be noted that Stokes' results (114), (116), and (117) can easily be approximated further to obtain solutions for long waves (or shallow water) and for short waves (or deep water).



Water Waves and the Korteweg–de Vries Equation, Figure 4  
The steepest wave profile

We next discuss the phenomenon of breaking of water waves which is one of the most common observable phenomena on an ocean beach. A wave coming from deep ocean changes shape as it moves across a shallow beach. Its amplitude and wavelength also are modified. The wavetrain is very smooth some distance offshore, but as it moves inshore, the front of the wave steepens noticeably until, finally, it breaks. After breaking, waves continue to move inshore as a series of bores or hydraulic jumps, whose energy is gradually dissipated by means of the water turbulence. Of the phenomena common to waves on beaches, breaking is the physically most significant and mathematically least known. In fact, it is one of the most intriguing longstanding problems of water waves.

For waves of small amplitude in deep water, the maximum particle velocity is  $v = a\omega = ack$ . But the basic assumption of small amplitude theory implies that  $\frac{v}{c} = ak \ll 1$ . Therefore, wave breaking can never be predicted by the small amplitude wave theory. That possibility arises only in the theory of finite amplitude waves. It is to be noted that the Stokes expansions are limited to relatively small amplitude and cannot predict the wavetrain of maximum height at which the crests are found to be very sharp. For a wave profile of constant shape moving at a uniform velocity, it can be shown that the maximum total crest angle as the wave begins to break is  $120^\circ$  as shown in Fig. 4.

The upshot of the Stokes analysis reveals that the inclusion of higher-order terms in the representation of the surface wave profile distorts its shape away from the linear sinusoidal curve. The effects of nonlinearity are likely to make crests narrower (sharper) and the troughs flatter as depicted in Figure 2.7 of Debnath [19]. The resulting wave profile more accurately portrays the water waves that are observed in nature. Finally, the sharp crest angle of  $120^\circ$  was first found by Stokes.

On the other hand, in 1865, Rankine conjectured that there exists a wave of extreme height. In a moving reference frame, the Euler equations are Galilean invariant, and the Bernoulli Eq. (36) on the free surface of water with  $\rho = 1$  becomes

$$\frac{1}{2}|\nabla\phi|^2 + gz = E. \tag{118}$$

Thus, this equation represents the conservation of local energy, where the first term is the kinetic energy of the fluid and the second term is the potential energy due to gravity. For the wave of maximum height,  $E = gz_{\max}$ , where  $z_{\max}$  is the maximum height of the fluid. Thus, the velocity is zero at the maximum height so that there will be a stagnation point in the fluid flow. Rankine conjectured that a cusp is developed at the peak of the free surface with a vertical slope so that the angle subtended at the peak is  $120^\circ$  as also conjectured by Stokes [54]. Toland [56] and Amick et al. [4] have proved rigorously the existence of a wave of greatest height and the Stokes conjecture for the wave of extreme form. However, Toland [56] also proved that if the singularity at the peak is *not* a cusp, that is, if there is no vertical slope at the peak of the free surface, then the Stokes remarkable conjecture of the crest angle of  $120^\circ$  is true. Subsequently, Amick et al. [4] confirmed that the singularity at the peak is *not* a cusp. Therefore, the full Euler equations exhibit singularities, and there is a limiting amplitude to the periodic waves.

The nonlinear solutions for plane waves based on systematic power series in the wave amplitude are known as *Stokes expansions*. Stokes [54] showed that the free surface elevation  $\eta$  of a plane wavetrain on deep water can be expanded in powers of the amplitude  $a$  as

$$\eta = a \cos \theta + \frac{1}{2}ka^2 \cos 2\theta + \frac{3}{8}k^2a^3 \cos 3\theta + \dots, \tag{119}$$

where  $\theta = kx - \omega t$  and the square of the frequency is

$$\omega^2 = gk \left( 1 + a^2k^2 + \frac{5}{4}a^4k^4 + \dots \right). \tag{120}$$

A question was raised about the convergence of the Stokes expansion in order to prove the existence of solution representing periodic water waves of permanent form. Considerable attention had been given to this problem by several authors. The problem was eventually resolved by Levi–Civita [43], who proved formally that the Stokes expansion for deep water converges provided the wave steepness ( $ak$ ) is very small. Struik [55] extended the proof of Levi–Civita to small-amplitude waves on water of arbitrary, but constant depth. Subsequently, Kraskovskii [40,41] finally established the existence of permanent periodic waves for all amplitudes less than the extreme value at which the waves assume a sharp-crested form. Although the success of the perturbation expansion as a means of representing waves of finite amplitude depends on the convergence of Stokes’ expansion, convergence *does not* at all imply stability. In spite of the preceding attempts to establish the possibility of finite-am-

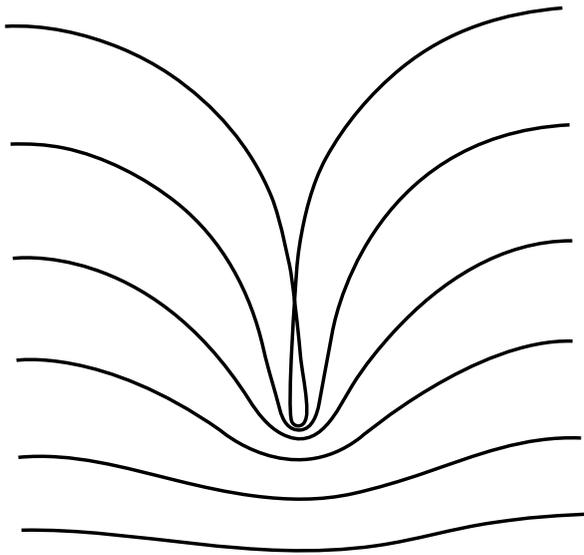
plitude water waves of permanent form, the question of their stability was altogether ignored until the 1960s. The independent work of Lighthill [44,45], Benjamin [5] and Whitham [60,61] finally established that Stokes waves on deep water are definitely *unstable*! This is one of the most remarkable discoveries in the history of the subject of the theory of water waves.

As a complement to the Stokes theory of pure gravity waves, Crapper [16] first discovered the *exact nonlinear* solution for pure progressive capillary waves of arbitrary amplitude by using a complex variable method. He obtained a remarkably new result for the phase velocity

$$c(k) = \left(\frac{kT}{\rho}\right)^{\frac{1}{2}} \left(1 + \frac{\pi^2 H^2}{4\lambda}\right)^{-1/4}, \tag{121}$$

where  $H = 2a$  is the wave height representing the vertical distance between crest and trough. This result clearly shows that the phase velocity decreases as the wave amplitude increases for a fixed wavelength. Result (121) is now known as *Crapper’s nonlinear capillary wave solution*. It has also been shown by Crapper that the capillary wave of greatest height occurs when  $H = 0.73\lambda$ , which has a striking contrast with the corresponding result due to Michell [48] for pure gravity waves,  $H = 0.142\lambda = \frac{1}{7}\lambda$ . Figure 5 exhibits Crapper’s nonlinear capillary wave profiles.

One of the important features of Crapper’s solution is that there is a maximum possible steepness for pure capil-



**Water Waves and the Korteweg–de Vries Equation, Figure 5**  
Surface profile for nonlinear capillary waves up to maximum possible wave,  $(H/\lambda) = 0.73$  (from [16])

lary waves,  $H/\lambda = 0.73$ , at which the waves hit each other. In the case of pure gravity waves, Stokes suggested that the wave steepness would possibly be greatest when the crest actually becomes a point with the maximum included crest angle of  $120^\circ$ . Miche [49] obtained the maximum wave height in water of finite depth  $h$  as

$$\left(\frac{H}{\lambda}\right)_{\max} = (0.14) \tanh kh, \tag{122}$$

which, in very shallow water ( $kh \ll 1$ ), gives

$$\left(\frac{H}{h}\right)_{\max} = 0.88. \tag{123}$$

This result is in excellent agreement with experimental observations. As the ratio of wave amplitude to local water depth increases, waves are gradually deformed and seem to behave more and more like a series of solitary waves or a train of cnoidal waves. As revealed by reliable observations, soon they become unstable and then break, forming a whitecap. However, little is known about the detailed nature of wave breaking except for the fact that it is a typical complicated nonlinear phenomenon. It has also been predicted that swell in shallow water tends to break when the crest-to-trough height is about 0.88 times the local water depth.

**Surface Gravity Waves on a Running Stream in Water**

Debnath and Rosenblat [18] solved the *initial value problem* for the generation and propagation of two-dimensional surface waves at the free surface of a running stream on water of finite depth. With the aid of generalized functions, it is proved that the asymptotic solution of the initial value problem leads to the ultimate steady-state solution and the transient solution *without* the need to resort to the use of a radiation condition at infinity or equivalent device.

The fundamental two-dimensional water wave equations on a running stream with velocity  $U$  in the  $x$ -direction in water of depth  $h$  (see Debnath [19], page 115) are given by

$$\phi_{xx} + \phi_{zz} = 0, \quad -h \leq z \leq 0, \quad -\infty < x < \infty, \quad t > 0, \tag{124}$$

$$\eta_t + U\eta_x - \phi_z = 0, \quad \text{on } z = 0, \quad t > 0, \tag{125}$$

$$\phi_t + U\phi_x + g\eta = -\frac{P}{\rho} p(x)e^{i\omega t}, \quad z = 0, \quad t > 0, \tag{126}$$

$$\phi_z = 0, \quad \text{on } z = -h, \tag{127}$$

where the term on the right hand side of (126) represents the arbitrary applied pressure with frequency  $\omega$ ,  $P$  is a constant and  $p(x)$  is arbitrary function of  $x$ .

Making reference to Debnath [19] for a detailed method of solution and asymptotic analysis of the solution of the initial value problem, we discuss only the *inherent resonant (critical) phenomenon* involved in this problem. For certain limiting values of the three quantities  $U$ ,  $\omega$ , and  $h$ , there exists a critical speed  $U_c$  such that

$$(a) \quad U = U_c = \frac{1}{4} \left( \frac{g}{\omega} \right) \quad \text{for } h \rightarrow \infty, \quad \omega \text{ and } U \text{ are finite.} \quad (128)$$

and

$$(b) \quad U = U_c = \sqrt{gh} \quad \text{for } \omega = 0, \quad h \text{ and } U \text{ are finite.} \quad (129)$$

A special property of (a) and (b) is that in both situations there exists a critical value of the speed  $U$ , above and below which the steady-state wave system has respectively quite a different character; and that when  $U$  assumes its critical value,  $U_c$  the solution is singular due presumably to a breakdown of the linearized theory. More precisely, when  $U > U_c$ , these are two surface waves downstream of the origin traveling with speeds  $(\omega/s_1)$  and  $(\omega/s_2)$ , respectively, in the position  $x$ -direction; there are none upstream, where  $-s_1$  and  $-s_2$  are two roots (see Debnath and Rosenblat [18]) of the equations

$$(\omega + kU) \mp (gk \tanh kh)^{\frac{1}{2}} = 0. \quad (130ab)$$

On the other hand, when  $U < U_c$ , these are two more surface waves which move with speeds  $(\omega/\sigma_1)$  and  $(\omega/\sigma_2)$ , both in the negative  $x$ -direction, where  $\sigma_1$  and  $\sigma_2$  are arising from (130a) at positive values of  $k = \sigma_1$  and  $k = \sigma_2$ . The former of these exists only on the upstream side of the origin, and the latter only on the downstream side. This latter wave thus appears to originate at infinity, but this is only when it is viewed from the moving coordinate system. It is easily verified from equation (130a) that the speed  $(\omega/\sigma_1)$  is always less than  $U$ , the speed of the frame, so that relative to axes at rest this wave travels in a positive direction. Depending on  $U$ ,  $\omega$  and  $h$ , there is a delimiting case when roots  $\sigma_1$  and  $\sigma_2$  coalesce into a double root. This is the critical case which demand a certain relationship between physical quantities involved, which can be combined into velocities  $U$ ,  $\sqrt{gh}$  and  $(g/\omega)$ . Indeed, when  $h \rightarrow \infty$ , we have the case (a), and when  $\omega = 0$ , the case (b) occurs.

When  $\omega = 0$ , the wave system is degenerate. In this case,  $s_1 = s_2 = \sigma_1 = 0$  and the only wave that occurs is the one associated with the root  $\sigma_2$ . As before, it exists on the downstream side, when  $U > U_c$ , only, but is now a steady motion relative to the moving frame. This is the case discussed by Stoker ([53], p. 214) who found a similar behavior.

Combining the results (128) and (129) together, we see that there are three quantities involved having the dimensions of velocity, namely  $U$ ,  $(\frac{g}{\omega})$  and  $\sqrt{gh}$ . In general, all these three quantities are finite, and hence the critical situation can be expressed as a functional relationship between them of the form

$$F \left( U, g/\omega, \sqrt{gh} \right) = 0. \quad (131)$$

This function  $F$  is found and is seen to yield  $U$  as a single-valued function of  $(g/\omega)$  and  $\sqrt{gh}$ , with formulas (128) and (129) emerging as limiting cases.

Debnath and Rosenblat [18] also showed that the free surface elevation  $\eta(x, t)$  becomes singular when roots  $\sigma_1$  and  $\sigma_2$  coalesce into a double root,  $\sigma$ . Their asymptotic evaluation of  $\eta(x, t)$  for large  $|x|$  and  $t$  reveals that it represents two waves where the amplitude of one wave increases like  $x$ , while the amplitude of other wave is of  $t^{\frac{1}{2}}$  as  $t \rightarrow \infty$ . This singular behavior is not unexpected and is in accord with the findings of Stoker [53] and Kaplan [37]. Mathematically, this critical situation corresponds to the coalescence of two roots ( $\sigma_1 = \sigma_2 = \sigma$ ), which is in turn is equivalent to the reinforcement by superposition of two like waves leading to a *resonance-type effect*. Physically, this situation would reveal itself through wave motions of large amplitude, which cannot come within the scope of linearized theory. Consequently, it would be necessary to include nonlinear terms in the original formulation of the problem in order to achieve a mathematically valid and physically reasonable solution.

Akylas [2] developed a nonlinear theory near the critical speed  $U_c = \sqrt{gh}$  under the action of surface pressure  $p(x)$  traveling at a constant speed  $U$ . With slow time scale  $T = \varepsilon t$  and the slow space variable  $X = \varepsilon^{\frac{1}{3}} x$ , where  $\varepsilon = \frac{a}{h}$ , his asymptotic analysis reveals that the nonlinear response is of bounded amplitude  $A(X, t)$  and is governed by a *forced Korteweg and de Vries (KdV)* equation in the form

$$A_T + \gamma A_X - 2AA_X - \frac{1}{6} A_{XXX} = \pi \bar{p}(0)\delta'(X), \quad (132)$$

where  $(U/\sqrt{gh}) = 1 + \gamma \varepsilon^{\frac{2}{3}}$ ,  $\gamma = O(1)$ ,  $\bar{p}(k)$  is the Fourier transform of  $p(X)$  and  $\delta(X)$  is the Dirac delta function.

The main conclusion of this asymptotic analysis is that the far field disturbance is of relatively large amplitude,

but it remains bounded. The main question whether the nonlinear response evolves to solitons, or disperses out or even gives a cnoidal wave, still remains open. However, it was shown by numerical study that Eq. (132) admits a series of solitons that are generated in front of the pressure field. For  $\gamma < 0$ , the linearized solution consists of a periodic sinusoidal wave in  $X < 0$ . The nonlinear evolution of the wave disturbance at the resonance condition ( $\gamma = 0$ ) remains bounded. The major conclusion of this study is that a series of solitary waves (or solitons) is successively generated when  $X < 0$  and propagates in front of the pressure field. This prediction is in theoretical agreement with the nonlinear study of Wu and Wu [64], and also in good agreement with experimental findings of Huang et al. [32] who observed solitons in their experiments involving a ship moving in shallow water.

Akylas [3] also developed a nonlinear theory to extend Debnath and Rosenblat's linearized study for the evolution of surface waves generated by a moving oscillatory pressure near the *critical* (or *resonant*) speed. Based on the slow time scale  $T = \varepsilon t$  and the slow spatial variable  $X = \sqrt{\varepsilon} x$ , his nonlinear asymptotic analysis reveals that the evolution for the wave envelope  $A(X, T)$  is governed by the *forced nonlinear Schrödinger equation*. For details, the reader is referred to Akylas [3] or Debnath [19].

### History of Russell's Solitary Waves and Their Interactions

Historically, John Scott Russell first experimentally observed the solitary wave, a long water wave without change in shape, on the Edinburgh–Glasgow Canal in 1834. He called it the “*great wave of translation*” and then reported his observations at the British Association in his 1844 paper “Report on Waves.” Thus, the solitary wave represents, not a periodic wave, but the propagation of a single isolated symmetrical hump of unchanged form. His discovery of this remarkable phenomenon inspired him further to conduct a series of extensive laboratory experiments on the generation and propagation of such waves. Based on his experimental findings, Russell discovered, empirically, one of the most important relations between the speed  $U$  of a solitary wave and its maximum amplitude  $a$  above the free surface of liquid of finite depth  $h$  in the form

$$U^2 = g(h + a), \quad (133)$$

where  $g$  is the acceleration due to gravity. His experiments stimulated great interest in the subject of water waves and his findings received a strong criticism from G.B. Airy [1] and G.G. Stokes [54]. In spite of his remarkable work on

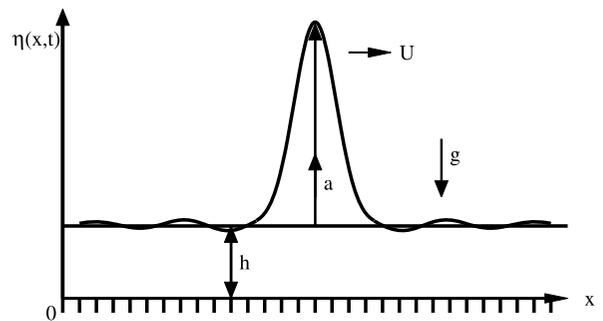
the existence of periodic wavetrains representing a typical feature of nonlinear dispersive wave systems, Stokes' conclusion on the existence of the solitary wave was erroneous.

However, Stokes [54] proposed that the free surface elevation of the plane wavetrains on deep water can be expanded in powers of the wave amplitude. His original result for the dispersion relation in deep water is given by (120). Despite these serious attempts to prove the existence of finite-amplitude water waves of permanent form, the independent question of their stability remained unanswered until the 1960s except for an isolated study by Korteweg and de Vries in 1895 on long surface waves in water of finite depth. But one of the most remarkable discoveries made in the 1960s was that the periodic Stokes waves on sufficiently deep water are *definitely* unstable. This result seems revolutionary in view of the sustained attempts to prove the existence of Stokes waves of finite amplitude and permanent form. Scott Russell's discovery of solitary waves contradicted the theories of water waves due to Airy and Stokes; they raised questions on the existence of Russell's solitary waves and conjectured that such waves cannot propagate in a liquid medium without change of form. It was not until the 1870s that Russell's prediction was finally and independently confirmed by both J. Boussinesq [9,10,11,12] and Lord Rayleigh [51]. From the equations of motion for an inviscid incompressible liquid, they derived formula (133). In fact, they also showed that the Russell's solitary wave profile (see Fig. 6)  $z = \eta(x, t)$  is given by

$$\eta(x, t) = a \operatorname{sech}^2 [\beta(x - Ut)], \quad (134)$$

where  $\beta^2 = 3a \div \{4h^2(h + a)\}$  for any  $a > 0$ .

Although these authors found the  $\operatorname{sech}^2$  solution, which is valid only if  $a \ll h$ , they did not write any equation for  $\eta$  that admits (134) as a solution. However, Boussinesq made much more progress and discovered several



Water Waves and the Korteweg–de Vries Equation, Figure 6  
A solitary wave

new ideas, including a nonlinear evolution equation for such long water waves in the form

$$\eta_{tt} = c^2 \left[ \eta_{xx} + \frac{3}{2} \left( \frac{\eta^2}{h} \right)_{xx} + \frac{1}{3} h^2 \eta_{xxx} \right], \quad (135)$$

where  $c = \sqrt{gh}$  is the speed of the shallow water waves. This is known as the *Boussinesq (bidirectional) equation*, which admits the solution

$$\eta(x, t) = a \operatorname{sech}^2 \left[ (3a/h^3)^{1/2} (x \pm Ut) \right]. \quad (136)$$

This represents solitary waves traveling in both, the positive and negative  $x$ -directions.

More than 60 years later, in 1895, two Dutchmen, D.J. Korteweg (1848–1941) and G. de Vries [39], formulated a mathematical model equation to provide an explanation of the phenomenon observed by Scott Russell. They derived the now-famous equation for the propagation of waves in one direction on the surface water of density  $\rho$  in the form

$$\eta_t = \frac{c}{h} \left[ \left( \varepsilon + \frac{3}{2} \eta \right) \eta_x + \frac{1}{2} \sigma \eta_{XXX} \right], \quad (137)$$

where  $X$  is a coordinate chosen to be moving (almost) with the wave,  $c = \sqrt{gh}$ ,  $\varepsilon$  is a small parameter, and

$$\sigma = h \left( \frac{h^2}{3} - \frac{T}{g\rho} \right) \sim \frac{1}{3} h^3, \quad (138)$$

when the surface tension  $T (\ll \frac{1}{3} g \rho h^2)$  is negligible. Equation (137) is known as the *Korteweg–de Vries (KdV) equation*. This is one of the simplest and most useful nonlinear model equations for solitary waves, and it represents the longtime evolution of wave phenomena in which the steepening effect of the nonlinear term is counterbalanced by the smoothening effect of the linear dispersion.

It is convenient to introduce the change of variables  $\eta = \eta(X^*, t)$  and  $X^* = X + (\varepsilon/h)ct$  which, dropping the asterisks, allows us to rewrite Eq. (137) in the form

$$\eta_t = \frac{c}{h} \left( \frac{3}{2} \eta \eta_x + \frac{1}{2} \sigma \eta_{XXX} \right). \quad (139)$$

Modern developments in the theory and applications of the KdV solitary waves began with the seminal work published as a Los Alamos Scientific Laboratory Report in 1955 by Fermi, Pasta, and Ulam [23] on a numerical model of a discrete nonlinear mass-spring system. In 1914, Debye suggested that the finite thermal conductivity of an anharmonic lattice is due to the nonlinear forces in the springs. This suggestion led Fermi, Pasta, and Ulam to believe that

a smooth initial state would eventually relax to an equipartition of energy among all modes because of nonlinearity. But their study led to the striking conclusion that there is no equipartition of energy among the modes. Although all the energy was initially in the lowest modes, after flowing back and forth among various low-order modes, it eventually returns to the lowest mode, and the end state is a series of recurring states. This remarkable fact has become known as the *Fermi–Pasta–Ulam (FPU) recurrence phenomenon*. Cercignani [13] and later on Palais [50] described the FPU experiment and its relationship to the KdV equation in some detail.

This curious result of the FPU experiment inspired Martin Kruskal and Norman Zabusky [65] to formulate a continuum model for the nonlinear mass-spring system to understand why recurrence occurred. In fact, they considered the initial-value problem for the KdV equation,

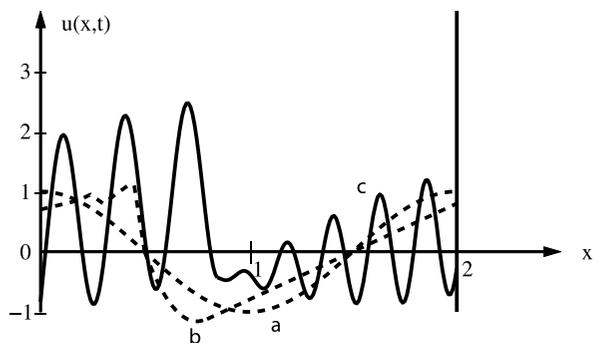
$$u_t + uu_x + \delta u_{xxx} = 0, \quad (140)$$

where  $\delta = \left( \frac{h}{\ell} \right)^2$ ,  $\ell$  is a typical horizontal length scale, with the initial condition

$$u(x, 0) = \cos \pi x, \quad 0 \leq x \leq 2, \quad (141)$$

and the periodic boundary conditions with period 2, so that  $u(x, t) = u(x + 2, t)$  for all  $t$ . Their numerical study with  $\sqrt{\delta} = 0.022$  produced remarkably new interesting results, which are shown in Fig. 7.

They observed that, initially, the wave steepened in regions where it had a negative slope, a consequence of the dominant effects of nonlinearity over the dispersive term,  $\delta u_{xxx}$ . As the wave steepens, the dispersive effect then becomes significant and balances the nonlinearity. At later times, the solution develops a series of *eight* well-defined waves, each like  $\operatorname{sech}^2$  functions with the taller (faster)



**Water Waves and the Korteweg–de Vries Equation, Figure 7** Development of solitary waves: **a** initial profile at  $t = 0$ , **b** profile at  $t = \pi^{-1}$ , and **c** wave profile at  $t = (3.6)\pi^{-1}$  (from [65])

waves ever catching up and overtaking the shorter (slower) waves. These waves undergo nonlinear interaction according to the KdV equation and then emerge from the interaction without change of form and amplitude, but with only a small change in their phases. So, the most remarkable feature is that these waves retain their identities after the nonlinear interaction. Another surprising fact is that the initial profile reappears, very similarly to the FPU recurrence phenomenon. In view of their preservation of shape and the resemblance to the particle-like character of these waves, Kruskal and Zabusky called these solitary waves, *solitons*, like photon, proton, electron, and other terms for elementary particles in physics.

Historically, the famous 1965 paper of Zabusky and Kruskal [65] marked the birth of the new concept of the *soliton*, a name intended to signify particle-like quantities. Subsequently, Zabusky [66] confirmed, numerically, the actual physical interaction of two solitons, and Lax [42] gave a rigorous analytical proof that the identities of two distinct solitons are preserved through the nonlinear interaction governed by the KdV equation. Physically, when two solitons of different amplitudes (and hence, of different speeds) are placed far apart on the real line, the taller (faster) wave to the left of the shorter (slower) wave, the taller one eventually catches up to the shorter one and then overtakes it. When this happens, they undergo a nonlinear interaction according to the KdV equation and emerge from the interaction completely preserved in form and speed with only a phase shift. Thus, these two remarkable features, (i) steady progressive pulse-like solutions and (ii) the preservation of their shapes and speeds, confirmed the particle-like property of the waves and, hence, the definition of the soliton. Subsequently, Gardner et al. [24,25] and Hirota [29,30,31] constructed analytical solutions of the KdV equation that provide the description of the interaction among  $N$  solitons for any positive integral  $N$ . After the discovery of the complete integrability of the KdV equation in 1967, the theory of the KdV equation and its relationship to the Euler equations of motion as an approximate model derived from the theory of asymptotic expansions became of major interest. From a physical point of view, the KdV equation is not only a standard nonlinear model for long water waves in a dispersive medium, it also arises as an approximate model in numerous other fields, including ion-acoustic plasma waves, magnetohydrodynamic waves, and anharmonic lattice vibrations. Experimental confirmation of solitons and their interactions has been provided successfully by Zabusky and Galvin [67], Hammack and Segur [26], and Weidman and Maxworthy [57]. Thus, these discoveries have led, in turn, to extensive theoretical, experimental, and computa-

tional studies over the last 40 years. Many nonlinear model equations have now been found that possess similar properties, and diverse branches of pure and applied mathematics have been required to explain many of the novel features that have appeared.

### The Korteweg–de Vries and Boussinesq Equations

We consider an inviscid liquid of constant mean depth  $h$  and constant density  $\rho$  without surface tension. We assume that the  $(x, y)$ -plane is the undisturbed free surface with the  $z$ -axis positive upward. The free surface elevation above the undisturbed mean depth  $h$  is given by  $z = \eta(x, y, t)$ , so that the free surface is at  $z = H = h + \eta$  and  $z = 0$  is the horizontal rigid bottom (see Fig. 8).

It has already been recognized that the parameters  $\varepsilon = (a/h)$  and  $\kappa = ak$ , where  $a$  is the surface wave amplitude and  $k$  is the wavenumber, must both be small for the linearized theory of surface waves to be valid. To develop the nonlinear shallow water theory, it is convenient to introduce the following nondimensional flow variables based on a different length scale  $h$  (which could be the fluid depth):

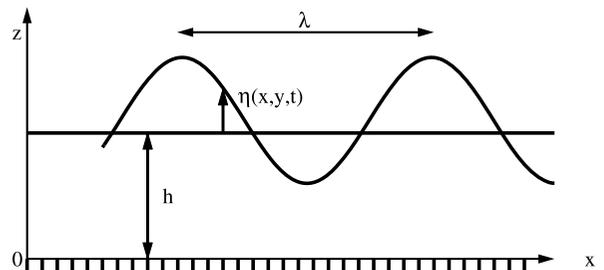
$$(x^*, y^*) = \frac{1}{l}(x, y), \quad z^* = \frac{z}{h}, \quad t^* = \left(\frac{ct}{l}\right), \\ \eta^* = \frac{\eta}{a}, \quad \phi^* = \left(\frac{h}{alc}\right)\phi, \quad (142)$$

where  $c = \sqrt{gh}$  is the typical horizontal velocity (or shallow water wave speed).

We next introduce two fundamental parameters to characterize the nonlinear shallow water waves:

$$\varepsilon = \frac{a}{h} \quad \text{and} \quad \delta = \frac{h^2}{l^2}, \quad (143)$$

where  $\varepsilon$  is called the *amplitude* parameter and  $\sqrt{\delta}$  is called the *long wavelength* or *shallowness* parameter.



Water Waves and the Korteweg–de Vries Equation, Figure 8  
A shallow water wave model

In terms of the preceding nondimensional flow variables and the parameters, the basic equations for water waves (73)–(76) can be written in the nondimensional form, dropping the asterisks:

$$\delta(\phi_{xx} + \phi_{yy}) + \phi_{zz} = 0, \quad 0 \leq z \leq 1 + \varepsilon\eta, \quad (144)$$

$$\frac{\partial \phi}{\partial t} + \frac{\varepsilon}{2}(\phi_x^2 + \phi_y^2) + \frac{\varepsilon}{2\delta}\phi_z^2 + \eta = 0 \quad \text{on } z = 1 + \varepsilon\eta, \quad (145)$$

$$\delta[\eta_t + \varepsilon(\phi_x\eta_x + \phi_y\eta_y)] - \phi_z = 0 \quad \text{on } z = 1 + \varepsilon\eta, \quad (146)$$

$$\phi_z = 0 \quad \text{on } z = 0. \quad (147)$$

It is noted that the parameter  $\kappa = ak$  does not enter explicitly in Eqs. (144)–(147), but an equivalent parameter  $\gamma = (a/l)$  is associated with  $\varepsilon$  and  $\delta$  through  $\gamma = (a/h) \cdot (h/l) = \varepsilon\sqrt{\delta}$ .

If  $\varepsilon$  is small, the terms involving  $\varepsilon$  in (145) and (146) can be neglected to recover the linearized free surface conditions. However, the assumption that  $\delta$  is small might be interpreted as the characteristic feature of the shallow water theory. So, we expand  $\phi$  in terms of  $\delta$  without any assumption about  $\varepsilon$ , and write

$$\phi = \phi_0 + \delta\phi_1 + \delta^2\phi_2 + \dots, \quad (148)$$

and then substitute in (144)–(146). The lowest-order term in (144) is

$$\phi_{0zz} = 0, \quad (149)$$

which, with (147), yields  $\phi_{0z} = 0$ , for all  $z$ , or  $\phi_0 = \phi_0(x, y, t)$ , which indicates that the horizontal velocity components are independent of the vertical coordinate  $z$  in lowest order. Consequently, we use the notation

$$\phi_{0x} = u(x, y, t) \quad \text{and} \quad \phi_{0y} = v(x, y, t). \quad (150ab)$$

The first- and second-order terms in (144) are given by

$$\phi_{0xx} + \phi_{0yy} + \phi_{1zz} = 0, \quad (151)$$

$$\phi_{1xx} + \phi_{1yy} + \phi_{2zz} = 0. \quad (152)$$

Integrating (151) with respect to  $z$  and using (150ab) gives

$$\phi_{1z} = -z(u_x + v_y) + C(x, y, t), \quad (153)$$

where the arbitrary function  $C(x, y, t)$  becomes zero because of the bottom boundary condition (147). Integrating

the resulting Eq. (153), again with respect to  $z$  and omitting the arbitrary constant, we obtain

$$\phi_1 = -\frac{z^2}{2}(u_x + v_y), \quad (154)$$

so that  $\phi_1 = 0$  at  $z = 0$  and  $u$  and  $v$  are then the horizontal velocity components at the bottom boundary.

We next substitute (154) in (151) and (152), and then integrate with condition (147) to determine the arbitrary function. Consequently,

$$\begin{aligned} \phi_{2z} &= \frac{1}{6}z^3 [(\nabla^2 u)_x + (\nabla^2 v)_y], \\ \phi_2 &= \frac{1}{24}z^4 [(\nabla^2 u)_x + (\nabla^2 v)_y], \end{aligned} \quad (155ab)$$

where  $\nabla^2$  is the two-dimensional Laplacian.

We next consider the free surface boundary conditions retaining all terms up to order  $\delta$ ,  $\varepsilon$  in (145), and  $\delta^2$ ,  $\varepsilon^2$ , and  $\delta\varepsilon$  in (146). It turns out that conditions (145) and (146) become

$$\phi_{0t} - \frac{\delta}{2}(u_{tx} + v_{ty}) + \eta + \frac{1}{2}\varepsilon(u^2 + v^2) = 0, \quad (156)$$

$$\begin{aligned} \delta[\{\eta_t + \varepsilon(u\eta_x + v\eta_y)\} + (1 + \varepsilon\eta)(u_x + v_y)] \\ = \frac{\delta^2}{6} [(\nabla^2 u)_x + (\nabla^2 v)_y]. \end{aligned} \quad (157)$$

Differentiating (156) first with respect to  $x$  and then with respect to  $y$  gives two equations:

$$u_t + \varepsilon(uu_x + vv_x) + \eta_x - \frac{1}{2}\delta(u_{txx} + v_{txy}) = 0, \quad (158)$$

$$v_t + \varepsilon(uu_y + vv_y) + \eta_y - \frac{1}{2}\delta(u_{txy} + v_{tyy}) = 0. \quad (159)$$

Simplifying (157) yields

$$\begin{aligned} \eta_t + [u(1 + \varepsilon\eta)]_x + [v(1 + \varepsilon\eta)]_y \\ = \frac{\delta}{6} [(\nabla^2 u)_x + (\nabla^2 v)_y]. \end{aligned} \quad (160)$$

Evidently, Eqs. (158)–(160) represent the nondimensional *shallow water equations*.

Using the fact that  $\phi_0$  is irrotational, that is,  $u_y = v_x$  and neglecting terms  $O(\delta)$  in (158)–(160), we obtain the fundamental shallow water equations

$$u_t + \varepsilon(uu_x + vv_x) + \eta_x = 0, \quad (161)$$

$$v_t + \varepsilon(uv_x + vv_y) + \eta_y = 0, \quad (162)$$

$$\eta_t + [u(1 + \varepsilon\eta)]_x + [v(1 + \varepsilon\eta)]_y = 0. \quad (163)$$

This system of three, coupled, nonlinear equations is closed and admits some interesting and useful solutions for  $u$ ,  $v$ , and  $\eta$ . It is equivalent to the boundary-layer equations in fluid mechanics. Finally, it can be linearized when  $\varepsilon = (a/h) \ll 1$  to obtain the following dimensional equations:

$$\begin{aligned} u_t + g \eta_x &= 0, & v_t + g \eta_y &= 0, \\ \eta_t + h(u_x + v_y) &= 0. \end{aligned} \tag{164abc}$$

Eliminating  $u$  and  $v$  from these equations gives

$$\eta_{tt} = c^2(\eta_{xx} + \eta_{yy}). \tag{165}$$

This is a well-known *two-dimensional wave equation*. It corresponds to the nondispersive shallow water waves that propagate with constant velocity  $c = \sqrt{gh}$ . This velocity is simply the linearized version of  $\sqrt{g(h + \eta)}$ . The wave equation has the simple *d'Alembert solution* representing plane progressive waves

$$\begin{aligned} \eta(x, y, t) = f(k_1x + l_1y - \kappa_1ct) \\ + g(k_2x + l_2y - \kappa_2ct), \end{aligned} \tag{166}$$

where  $f$  and  $g$  are arbitrary functions and  $\kappa_r^2 = (k_r^2 + l_r^2)$ ,  $r = 1, 2$ .

We consider the one-dimensional case retaining both  $\varepsilon$  and  $\delta$  order terms in (158)–(160) so that these equations reduce to the *Boussinesq* [9,10,11,12] equations

$$u_t + \varepsilon uu_x + \eta_x - \frac{1}{2} \delta u_{txx} = 0, \tag{167}$$

$$\eta_t + [u(1 + \varepsilon\eta)]_x - \frac{1}{6} \delta u_{xxx} = 0. \tag{168}$$

On the other hand, Eqs. (161)–(163), expressed in dimensional form, read

$$u_t + uu_x + vv_y + gH_x = 0, \tag{169}$$

$$v_t + uv_x + vv_y + gH_y = 0, \tag{170}$$

$$H_t + (uH)_x + (vH)_y = 0, \tag{171}$$

where  $H = (h + \eta)$  is the total depth and  $H_x = \eta_x$ , since the depth  $h$  is constant.

In particular, the one-dimensional version of the shallow water equations follows from (169)–(171) and is given by

$$u_t + uu_x + gH_x = 0, \tag{172}$$

$$H_t + (uH)_x = 0. \tag{173}$$

This system of approximate shallow water equations is analogous to the exact governing equations of gas dynamics for the case of a compressible flow involving only one space variable (see Riabouchinsky [52]).

It is convenient to rewrite these equations in terms of the wave speed  $c = \sqrt{gH}$  by using  $dc = (g/2c) dH$ , so that they become

$$u_t + uu_x + 2cc_x = 0, \tag{174}$$

$$2c_t + cu_x + 2uc_x = 0. \tag{175}$$

The standard method of characteristics can easily be used to solve (174) and (175). Adding and subtracting these equations allows us to rewrite them in the characteristic form

$$\left[ \frac{\partial}{\partial t} + (c + u) \frac{\partial}{\partial x} \right] (u + 2c) = 0, \tag{176}$$

$$\left[ \frac{\partial}{\partial t} + (c - u) \frac{\partial}{\partial x} \right] (u - 2c) = 0. \tag{177}$$

Equations (176) and (177) show that  $u + 2c$  propagates in the positive  $x$ -direction with velocity  $c + u$ , and  $u - 2c$  travels in the negative  $x$ -direction with velocity  $c - u$ , that is, both  $u + 2c$  and  $u - 2c$  propagate in their respective directions with velocity  $c$  relative to the water. In other words,

$$\begin{aligned} u + 2c &= \text{constant on curves } C_+ \text{ on which } \frac{dx}{dt} = u + c \\ u - 2c &= \text{constant on curves } C_- \text{ on which } \frac{dx}{dt} = u - c \end{aligned} \tag{178ab}$$

where  $C_+$  and  $C_-$  are *characteristic curves* of the system of partial differential Eqs. (172) and (173). A disturbance propagates along these characteristic curves at speed  $c$  relative to the flow speed. The quantities  $(u \pm 2c)$  are called the *Riemann invariants* of the system, and a simple wave is propagating to the right into water of depth  $h$ , that is,  $u - 2c = c_0 = \sqrt{gh}$ . Then, the solution is given by

$$u = f(\xi), \quad x = \xi + \left( c_0 + \frac{3}{2} u \right) t, \tag{179}$$

where  $u(x, t) = f(x)$  at  $t = 0$ . However, we note that

$$u_x = \left( 1 - \frac{3}{2} u_x t \right) f'(\xi),$$

giving

$$u_x = \frac{2f'(\xi)}{2 + 3t f'(\xi)}. \tag{180}$$

Thus, if  $f'(\xi)$  is anywhere less than zero,  $u_x$  tends to infinity as  $t \rightarrow -2/(3f')$ . In terms of the free surface elevation, solution (179) implies that the wave profile progressively distorts itself, and, in fact, any forward-facing portion of such a wave continually steepens, or the higher parts of the wave tend to catch up with lower parts in front of them. Thus, all of these waves, carrying an increase of elevation, invariably break. The breaking of water waves on beaches is perhaps the most common and the most striking phenomenon in nature.

An alternative system equivalent to the nonlinear evolution Eqs. (167) and (168) can be derived from the nonlinear shallow water theory, retaining both  $\varepsilon$  and  $\delta$  order terms with  $\delta < 1$ . This system is also known as the *Boussinesq equations*, which, in dimensional variables, are given by

$$\eta_t + [(h + \eta)u]_x = 0, \tag{181}$$

$$u_t + uu_x + g\eta_x = \frac{1}{3}h^2 u_{xxt}. \tag{182}$$

They describe the evolution of long water waves that move in both positive and negative  $x$ -directions. Eliminating  $\eta$  and neglecting terms smaller than  $O(\varepsilon, \delta)$  gives a single *Boussinesq equation* for  $u(x, t)$  in the form

$$u_{tt} - c^2 u_{xx} + \frac{1}{2}(u^2)_{xt} = \frac{1}{3}h^2 u_{xxtt}. \tag{183}$$

The linearized Boussinesq equation for  $u$  and  $\eta$  follows from (181) and (182) as

$$\left[ \frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} - \frac{1}{3}h^2 \frac{\partial^4}{\partial x^2 \partial t^2} \right] \begin{pmatrix} u \\ \eta \end{pmatrix} = 0. \tag{184}$$

This is in perfect agreement with the infinitesimal wave theory result expanded for small  $kh$ . Thus, the third derivative term in (182) may be identified with the frequency dispersion.

Another equivalent version of the Boussinesq equation is given by

$$\eta_{tt} - c^2 \eta_{xx} = \frac{3}{2} \left( \frac{\eta^2}{h} \right)_{xx} + \frac{1}{3}h^2 \eta_{xxxx}. \tag{185}$$

There are several features of this equation. It is a nonlinear partial differential equation that incorporates the basic idea of nonlinearity and dispersion. Boussinesq obtained three invariant physical quantities,  $Q$ ,  $E$ , and  $M$ , defined by

$$\begin{aligned} Q &= \int_{-\infty}^{\infty} \eta \, dx, & E &= \int_{-\infty}^{\infty} \eta^2 \, dx, \\ M &= \int_{-\infty}^{\infty} \left[ \eta_x^2 - 3 \left( \frac{\eta}{h} \right)^3 \right] dx, \end{aligned} \tag{186}$$

provided that  $\eta \rightarrow 0$  as  $|x| \rightarrow \infty$ . Evidently,  $Q$  and  $E$  represent the volume and the energy of the solitary wave. The third quantity  $M$  is called the *moment of instability*, and the variational problem,  $\delta M = 0$  with  $E$  fixed, leads to the unique solitary-wave solution. Boussinesq also derived the results for the amplitude and volume of a solitary wave of given energy in the form

$$a = \frac{3}{4} \left( \frac{E^{3/2}}{h} \right), \quad Q = 2 h E^{1/3}. \tag{187ab}$$

The former result shows that the amplitude of a solitary wave in a channel varies inversely as the channel depth  $h$ .

The Boussinesq equation can then be written in the normalized form

$$u_{tt} - u_{xx} - \frac{3}{2}(u^2)_{xx} - u_{xxxx} = 0. \tag{188}$$

This particular form is of special interest because it admits inverse scattering formalism. Equation (188) has steady progressive wave solutions in the form

$$u(x, t) = 4k^2 f(X), \quad X = kx - \omega t, \tag{189}$$

where the equation for  $f(X)$  can be integrated to obtain

$$f_{XX} = 6A + (4 - 6B)f - 6f^2, \tag{190}$$

where  $A$  is a constant of integration and

$$\omega^2 = k^2 + k^4(4 - 6B). \tag{191}$$

For the special case  $A = B = 0$ , a single solitary-wave solution is given by

$$f(X) = \operatorname{sech}^2(X - X_0), \tag{192}$$

where  $X_0$  is a constant of integration. This result can be used to construct a solution for a series of solitary waves, spaced  $2\sigma$  apart, in the form

$$f(X) = \sum_{n=-\infty}^{\infty} \operatorname{sech}^2(X - 2n\sigma). \tag{193}$$

This is a  $2\sigma$  periodic function that satisfies (190) for certain values of  $A$  and  $B$ .

We next assume that  $\varepsilon$  and  $\delta$  are comparable, so that all terms  $O(\varepsilon, \delta)$  in (158)–(160) can be retained. For the case of the two-dimensional wave motion ( $v = 0$  and  $\partial/\partial y = 0$ ), these equations become

$$u_t + \eta_x + \varepsilon u u_x - \frac{1}{2}\delta u_{txx} = 0, \tag{194}$$

$$\eta_t + [u(1 + \varepsilon\eta)]_x - \frac{1}{6}\delta u_{xxx} = 0. \tag{195}$$

We now seek steady progressive wave solutions traveling to the positive  $x$ -direction only, so that  $u = u(x - Ut)$  and  $\eta = \eta(x - Ut)$ . With the terms of zero order in  $\varepsilon$  and  $\delta$  and  $U = 1$ , we assume a solution of the form

$$u = \eta + \varepsilon P + \delta Q, \tag{196}$$

where  $P$  and  $Q$  are unknown functions to be determined. Consequently, Eqs. (194) and (195) become

$$(\eta + \varepsilon P + \delta Q)_t + \eta_x + \varepsilon \eta \eta_x - \frac{1}{2} \delta \eta_{txx} = 0, \tag{197}$$

$$\eta_t + \left[ (1 + \varepsilon \eta)(\eta + \varepsilon P + \delta Q) \right]_x - \frac{1}{6} \delta \eta_{xxx} = 0. \tag{198}$$

These equations must be consistent so that we stipulate for the zero order

$$\eta_t = -\eta_x, \quad P = -\frac{1}{4} \eta^2, \quad Q = \frac{1}{3} \eta_{xx} = -\frac{1}{3} \eta_{xt}. \tag{199}$$

We use these results to rewrite both (197) and (198) with the assumption that  $\varepsilon$  and  $\delta$  are of equal order and small enough for their products and squares to be ignored, so that the ratio  $(\varepsilon/\delta) = (a^2/h^3)$  is of the order one. Consequently, we obtain a single equation for  $\eta(x, t)$  in the form

$$\eta_t + \left( 1 + \frac{3}{2} \varepsilon \eta \right) \eta_x + \frac{1}{6} \delta \eta_{xxx} = 0. \tag{200}$$

This is now universally known as the *Korteweg and de Vries equation* as they discovered it in their 1895 seminal work. We point out that  $(\varepsilon/\delta) = a^2/h^3$  is one of the *fundamental* parameters in the theory of nonlinear shallow water waves. Recently, Infeld [33] considered three-dimensional generalizations of the Boussinesq and Korteweg–de Vries equations.

**Solutions of the KdV Equation: Solitons and Cnoidal Waves**

To find solutions of the KdV equation, it is convenient to rewrite it in terms of dimensional variables as

$$\eta_t + c \left( 1 + \frac{3}{2h} \eta \right) \eta_x + \frac{ch^2}{6} \eta_{xxx} = 0, \tag{201}$$

where  $c = \sqrt{gh}$ , and the total depth  $H = h + \eta$ . The first two terms ( $\eta_t + c \eta_x$ ) describe wave evolution at the shallow water speed  $c$ , the third term with coefficient  $(3c/2h)$  represents a nonlinear wave steepening, and the last term with coefficient  $(ch^2/6)$  describes linear dispersion. Thus, the KdV equation is a balance between time evolution, nonlinearity, and linear dispersion. The dimensional velocity  $u$  is obtained from (196) with (199) in the form

$$u = \frac{c}{h} \left( \eta - \frac{1}{4h} \eta^2 + \frac{h}{3} \eta_{xx} \right). \tag{202}$$

We seek a traveling wave solution of (201) in the frame  $X$  so that  $\eta = \eta(X)$  and  $X = x - Ut$  with  $\eta \rightarrow 0$ , as  $|x| \rightarrow \infty$ , where  $U$  is a constant speed. Substituting this solution in (201) gives

$$(c - U)\eta' + \frac{3c}{2h} \eta \eta' + \frac{ch^2}{6} \eta''' = 0, \tag{203}$$

where  $\eta' = d\eta/dX$ . Integrating this equation with respect to  $X$  yields

$$(c - U)\eta + \frac{3c}{4h} \eta^2 + \frac{ch^2}{6} \eta'' = A, \tag{204}$$

where  $A$  is an integrating constant.

We multiply this equation by  $2\eta'$  and integrate again to obtain

$$(c - U)\eta^2 + \left( \frac{c}{2h} \right) \eta^3 + \left( \frac{ch^2}{6} \right) \left( \frac{d\eta}{dX} \right)^2 = 2A\eta + B, \tag{205}$$

where  $B$  is also a constant of integration.

We now consider a special case when  $\eta$  and its derivatives tend to zero at infinity and  $A = B = 0$ , so that (205) gives

$$\left( \frac{d\eta}{dX} \right)^2 = \frac{3}{h^3} \eta^2 (a - \eta), \tag{206}$$

where

$$a = 2h \left( \frac{U}{c} - 1 \right). \tag{207}$$

The right-hand side of (206) vanishes at  $\eta = 0$  and  $\eta = a$ , and the exact solution of (206) represents Russell's solitary wave in the form

$$\eta(X) = a \operatorname{sech}^2(bX), \quad b = \left( \frac{3a}{4h^3} \right)^{1/2}. \tag{208ab}$$

Thus, the explicit form of the solution is

$$\eta(x, t) = a \operatorname{sech}^2 \left[ \left( \frac{3a}{4h^3} \right)^{1/2} (x - Ut) \right], \tag{209}$$

where the velocity of the wave is

$$U = c \left( 1 + \frac{a}{2h} \right). \tag{210}$$

This is an *exact* solution of the KdV equation for all  $(a/h)$ ; however, the equation is derived with the approximation

$(a/h) \ll 1$ . The solution (209) is called a *soliton* (or *solitary wave*) describing a single hump of height  $a$  above the undisturbed depth  $h$  and tending rapidly to zero away from  $X = 0$ . The solitary wave propagates to the right with velocity  $U (> c)$ , which is directly proportional to the amplitude  $a$  and has width  $b^{-1} = (3a/4h^3)^{-1/2}$ , that is,  $b^{-1}$  is inversely proportional to the square root of the amplitude  $a$ . Another significant feature of the soliton solution is that it travels in the medium without change of shape, which is hardly possible without retaining  $\delta$ -order terms in the governing equation. A solitary wave profile has already been shown in Fig. 6.

In the general case, when both  $A$  and  $B$  are nonzero, (205) can be rewritten as

$$\frac{h^3}{3} \left( \frac{d\eta}{dX} \right)^2 = -\eta^3 + 2h \left( \frac{U}{c} - 1 \right) \eta^2 + \frac{2h}{c} (2A\eta + B) = F(\eta), \quad (211)$$

where  $F(\eta)$  is a cubic with simple zeros.

We seek a real bounded solution for  $\eta(X)$ , which has a minimum value zero and a maximum value  $a$  and oscillates between the two values. For bounded solutions, all three zeros  $\eta_1, \eta_2, \eta_3$  must be real. Without loss of generality, we set  $\eta_1 = 0$  and  $\eta_2 = a$ . Hence, the third zero must be negative so that  $\eta_3 = -(b - a)$  with  $b > a > 0$ . With these choices,  $F(\eta) = \eta(a - \eta)(\eta - a + b)$  and Eq. (211) assumes the form

$$\frac{h^3}{3} \left( \frac{d\eta}{dX} \right)^2 = \eta(a - \eta)(\eta - a + b), \quad (212)$$

where

$$U = c \left( 1 + \frac{2a - b}{2h} \right), \quad (213)$$

which is obtained by comparing the coefficients of  $\eta^2$  in (211) and (212).

Writing  $a - \eta = p^2$ , it follows from Eq. (212) that

$$\left( \frac{3}{4h^3} \right)^{1/2} dX = \frac{dp}{[(a - p^2)(b - p^2)]^{1/2}}. \quad (214)$$

Substituting  $p = \sqrt{a} q$  in (214) gives the standard elliptic integral of the first kind (see Dutta and Debnath [22] and Helal and Molines [28])

$$\left( \frac{3b}{4h^3} \right)^{1/2} X = \int_0^q \frac{dq}{[(1 - q^2)(1 - m^2 q^2)]^{1/2}}, \quad m = \left( \frac{a}{b} \right)^{1/2}, \quad (215)$$

and then, function  $q$  can be expressed in terms of the Jacobian elliptic function,  $sn(z, m)$

$$q(X, m) = sn \left[ \left( \frac{3b}{4h^3} \right)^{1/2} X, m \right], \quad (216)$$

where  $m$  is the modulus of  $sn(z, m)$ .

Finally,

$$\begin{aligned} \eta(X) &= a \left[ 1 - sn^2 \left\{ \left( \frac{3b}{4h^3} \right)^{1/2} X \right\} \right] \\ &= a cn^2 \left[ \left( \frac{3b}{4h^3} \right)^{1/2} X \right], \end{aligned} \quad (217)$$

where  $cn(z, m)$  is also the Jacobian elliptic function with a period  $2K(m)$ , where  $K(m)$  is the complete elliptic integral of the first kind defined by

$$K(m) = \int_0^{\pi/2} (1 - m^2 \sin^2 \theta)^{-1/2} d\theta, \quad (218)$$

and  $cn^2(z) + sn^2(z) = 1$ .

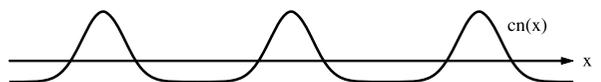
It is important to note that  $cn z$  is periodic, and hence,  $\eta(X)$  represents a train of periodic waves in shallow water. Thus, these waves are called *cnoidal waves* with wavelength

$$\lambda = 2 \left( \frac{4h^3}{3b} \right)^{1/2} K(m). \quad (219)$$

The upshot of this analysis is that solution (217) represents a nonlinear wave whose shape and wavelength (or period) all depend on the amplitude of the wave. A typical cnoidal wave is shown in Fig. 9. Sometimes, the cnoidal waves with slowly varying amplitude are observed in rivers. More often, wavetrains behind a weak bore (called an *undular bore*) can be regarded as cnoidal waves. Two limiting cases are of special physical interest: (i)  $m \rightarrow 0$  and (ii)  $m \rightarrow 1$ .

In the first case,  $sn z \rightarrow \sin z$ ,  $cn z \rightarrow \cos z$  as  $m \rightarrow 0$  ( $a \rightarrow 0$ ). This corresponds to small-amplitude waves where the linearized KdV equation is appropriate. So, in this limiting case, the solution (217) becomes

$$\begin{aligned} \eta(x, t) &= \frac{1}{2} a [1 + \cos(kx - \omega t)], \\ k &= \left( \frac{3b}{h^3} \right)^{1/2}, \end{aligned} \quad (220)$$



Water Waves and the Korteweg–de Vries Equation, Figure 9  
A cnoidal wave

where the corresponding dispersion relation is

$$\omega = Uk = ck \left( 1 - \frac{1}{6} k^2 h^2 \right). \tag{221}$$

This corresponds to the first two terms of the series expansion of  $(gk \tanh kh)^{1/2}$ . Thus, these results are in perfect agreement with the linearized theory.

In the second limiting case,  $m \rightarrow 1 (a \rightarrow b)$ ,  $cn z \rightarrow$  sech  $z$ . Thus, the cnoidal wave solution tends to the classical KdV solitary-wave solution where the wavelength  $\lambda$ , given by (219), tends to infinity because  $K(a) = \infty$  and  $K(0) = \pi/2$ . The solution identically reduces to (209) with (210).

We next report the numerical computation of the KdV Eq. (201) due to Berezin and Karpman [8]. In terms of new variables defined by

$$x^* = x - ct, \quad t^* = t, \quad \eta^* = \left( \frac{3c}{2h} \right) \eta, \tag{222}$$

omitting the asterisks, Eq. (201) becomes

$$\eta_t + \eta \eta_x + \beta \eta_{xxx} = 0, \tag{223}$$

where  $\beta = (\frac{1}{6}) ch^2$ .

We examine the numerical solution of (223) with the initial condition

$$\eta(x, 0) = \eta_0 f \left( \frac{x}{\ell} \right), \tag{224}$$

where  $\eta_0$  is constant and  $f(\xi)$  is a nondimensional function characterizing the initial wave profile. It is convenient to introduce the dimensionless variables

$$\xi = \frac{x}{\ell}, \quad \tau = \frac{\eta_0 t}{\ell}, \quad u = \frac{\eta}{\eta_0} \tag{225}$$

so that Eqs. (223) and (224) reduce to

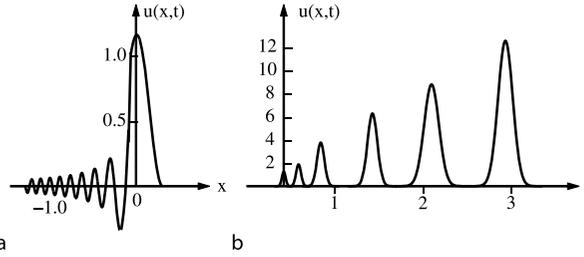
$$u_\tau + uu_\xi + \sigma^{-2} u_{\xi\xi\xi} = 0, \tag{226}$$

$$u(\xi, 0) = f(\xi), \tag{227}$$

where the dimensionless parameter  $\sigma$  is defined by  $\sigma = \ell(\eta_0/\beta)^{1/2}$ .

Berezin and Karpman [8] obtained the numerical solution of (226) with the Gaussian initial pulse of the form  $u(\xi, 0) = f(\xi) = \exp(-\xi^2)$  and values of the parameter  $\sigma = 1.9$  and  $\sigma = 16.5$ . Their numerical solutions are shown in Fig. 10.

As shown in Fig. 10 for case (a), the perturbation splits into a soliton and a wavepacket. In case (b), there are six solitons. It is readily seen that the peaks of the solitons lie



**Water Waves and the Korteweg–de Vries Equation, Figure 10**  
The solutions of the KdV equation  $u(x, t)$  for large values of  $t$  with the values of the similarity parameter  $\sigma$ : **a**  $\sigma = 1.9$ , **b**  $\sigma = 16.5$  (from [8])

nearly on a straight line. This is due to the fact that the velocity of the soliton is proportional to its amplitude, so that the distances traversed by the solitons would also be proportional to their amplitudes.

Zabusky’s [66] numerical investigation of the interaction of two solitons reveals that the taller soliton, initially behind, catches up to the shorter one, they undergo a nonlinear interaction and, then, emerge from the interaction without any change in shape and amplitude. The end result is that the taller soliton reappears in front and the shorter one behind. This is essentially strong computational evidence of the stability of solitons.

Using the transformation

$$x^* = \varepsilon \beta (x - ct), \quad t^* = \varepsilon^3 t, \quad \eta^* = (\alpha \varepsilon^2)^{-1} \eta$$

with  $\alpha\beta (= 3c/2h) = 6$  and  $\beta^3 (= ch^2/6) = 1$ , we write the KdV Eq. (201) in the normalized form, dropping the asterisks,

$$\eta_t + 6\eta\eta_x + \eta_{xxx} = 0. \tag{228}$$

We next seek a steady progressive wave solution of (228) in the form

$$\eta = 2k^2 f(X), \quad X = kx - \omega t. \tag{229}$$

Then, the equation for  $f(X)$  can be integrated once to obtain

$$f''(X) = 6A + (4 - 6B)f - 6f^2, \tag{230}$$

where  $A$  is a constant of integration and the frequency  $\omega$  is given by

$$\omega = k^3(4 - 6B). \tag{231}$$

A single-soliton solution corresponds to the special case  $A = B = 0$  and is given by

$$f(X) = \text{sech}^2(X - X_0), \tag{232}$$

where  $X_0$  is a constant.

Thus, for a series of solitons spaced  $2\sigma$  apart, we write

$$f(X) = \sum_{n=-\infty}^{\infty} \operatorname{sech}^2(X - 2n\sigma). \tag{233}$$

This is a  $2\sigma$  periodic function that satisfies (230) for some  $A$  and  $B$ .

The general elliptic function solution of (228) can be obtained from the integral of (230), which can be written as

$$f_x^2 = -4C + 12Af + (4 - 6B)f^2 - 4f^3, \tag{234}$$

where  $C$  is a constant of integration. Various asymptotic and numerical results lead to the relations (see Whitham [63])

$$C = -\frac{1}{2} \frac{dA(\sigma)}{d\sigma} = \frac{1}{4} \frac{dB(\sigma)}{d\sigma^2}, \tag{235ab}$$

and the cubic in (234) can be factorized as

$$-C + 3Af + \left(1 - \frac{3}{2}B\right)f^2 - f^3 = (f_1 - f)(f - f_2)(f - f_3), \tag{236}$$

where  $f_r(\sigma)$  ( $r = 1, 2, 3$ ) are determined from  $A(\sigma)$ ,  $B(\sigma)$ , and  $C(\sigma)$ . If we set  $f_1 > f_2 > f_3$  and, then, the periodic solution oscillates between  $f_1$  at  $X = 0$  and  $f_2$  at  $X = \sigma$ , one particular form of the solution is given by

$$f(X) = f_2 + (f_1 - f_2)cn^2(\sqrt{(f_1 - f_3)} X), \tag{237}$$

where the modulus  $m$  of  $cn(z, m)$  is given by

$$m^2 = \left(\frac{f_1 - f_2}{f_1 - f_3}\right). \tag{238}$$

Thus, it follows from (233) and (237) that the following identity holds:

$$f_2 + (f_1 - f_2)cn^2(\sqrt{(f_1 - f_3)} X) = \sum_{n=-\infty}^{\infty} \operatorname{sech}^2(X - 2n\sigma), \tag{239}$$

which can be verified by comparing the periods and poles of the two sides.

Finally, the higher-order modified KdV equation

$$v_t + (p + 1)v^p v_x + v_{xxxx} = 0, \quad p > 2, \tag{240}$$

admits single-soliton solutions in the form

$$v(x, t) = a \operatorname{sech}^{2/p}(kx - \omega t). \tag{241}$$

However, in view of the fractional powers of the sech function, it seems, perhaps, unlikely that there will be any simple superposition formula.

### Derivation of the KdV Equation from the Euler Equations

This problem was discussed in Sect. “The Korteweg–de Vries and Boussinesq Equations” by using the Laplace equation for the velocity potential under the assumption that  $\delta = O(\varepsilon)$  as  $\varepsilon \rightarrow 0$ . Here we follow Johnson [35] to present another derivation of the KdV equation from the Euler equation in (1 + 1) dimensions. This approach can be generalized to derive higher dimensional KdV equations.

We consider the problem of surface gravity waves which propagate in the positive  $x$ -direction over stationary water of constant depth. The associated Euler Eqs. (59) and the continuity Eq. (60) in (1 + 1) dimensions are given by

$$u_t + \varepsilon(uu_x + wu_z) = -p_x, \tag{242}$$

$$\delta [w_t + \varepsilon(uw_x + ww_z)] = -p_z, \tag{243}$$

$$u_x + w_z = 0. \tag{244}$$

The free surface and bottom boundary conditions are obtained from (57ab) and (62) in the form

$$w = \eta_t + \varepsilon u \eta_x, \quad p = \eta, \quad \text{on } z = 1 + \varepsilon \eta, \tag{245}$$

$$w = 0 \quad \text{on } z = 0. \tag{246}$$

It can easily be shown that, for any  $\sqrt{\delta}$  as  $\varepsilon \rightarrow 0$ , there exists a region in the  $(x, t)$ -space where there is a balance between nonlinearity and dispersion which leads to the KdV equation. The region of interest is defined by a scaling of the independent flow variables as

$$x \rightarrow \sqrt{\frac{\delta}{\varepsilon}} x \quad \text{and} \quad t \rightarrow \sqrt{\frac{\delta}{\varepsilon}} t, \tag{247}$$

for any  $\varepsilon$  and  $\sqrt{\delta}$ . In order to ensure consistency in the continuity equation, it is necessary to introduce a scaling of  $w$  by

$$w \rightarrow \sqrt{\frac{\varepsilon}{\delta}} w. \tag{248}$$

Consequently, the net effect of the scalings is to replace  $\delta$  by  $\varepsilon$  in Eqs. (242)–(246) so that they become

$$u_t + \varepsilon(uu_x + wu_z) = -p_x, \tag{249}$$

$$\varepsilon [w_t + \varepsilon(uw_x + ww_z)] = -p_z, \tag{250}$$

$$u_x + w_z = 0, \tag{251}$$

$$w = \eta_t + \varepsilon u \eta_x, \quad \text{and} \quad p = \eta \quad \text{on } z = 1 + \varepsilon \eta, \tag{252}$$

$$w = 0 \quad \text{on } z = 0. \tag{253}$$

In the limit as  $\varepsilon \rightarrow 0$ , the first-order approximation of Eqs. (249) and (252) gives

$$\eta = p, \quad 0 \leq z \leq 1, \quad \text{and} \quad u_t + \eta_x = 0. \tag{254}$$

It then follows from (251) that  $w = -z u_x$  which satisfies (253). The boundary condition (252) leads to  $\eta_t + u_x = 0$  on  $z = 1$ , which can be combined with (254) to obtain the linear wave equation

$$\eta_{tt} - \eta_{xx} = 0. \tag{255}$$

For waves propagating in the positive  $x$ -direction, we introduce the far-field variables

$$\xi = x - t \quad \text{and} \quad \tau = \varepsilon t, \tag{256}$$

so that  $\xi = O(1)$  and  $\tau = O(1)$  give the far-field region of the problem. This is the region where nonlinearity balances the dispersion to produce the KdV equation.

With the choice of the transformations (256), Eqs. (249)–(253) can be rewritten in the form

$$-u_\xi + \varepsilon(u_\tau + uu_\xi + wu_z) = -p_\xi, \tag{257}$$

$$\varepsilon[-w_\xi + \varepsilon(w_\tau + uw_\xi + ww_z)] = -p_z, \tag{258}$$

$$u_\xi + w_z = 0, \tag{259}$$

$$w = -\eta_\xi + \varepsilon(\eta_\tau + u\eta_\xi), \quad p = \eta, \quad \text{on } z = 1 + \varepsilon\eta, \tag{260}$$

$$w = 0 \quad \text{on } z = 0. \tag{261}$$

We seek an asymptotic series expansion of the solutions of the system (257)–(261) in the form

$$\eta(\xi, \tau, \varepsilon) = \sum_{n=0}^{\infty} \varepsilon^n \eta_n(\xi, \tau), \tag{262}$$

$$q(\xi, \tau, z; \varepsilon) = \sum_{n=0}^{\infty} \varepsilon^n q_n(\xi, \tau, z),$$

where  $q$  (and the corresponding  $q_n$ ) denotes each of the variables  $u, w$ , and  $p$ .

Consequently, the leading-order problem is given by

$$u_{0\xi} = p_{0\xi}, \quad p_{0z} = 0, \quad u_{0\xi} + w_{0z} = 0, \tag{263}$$

$$p_0 = \eta_0, \quad w + \eta_{0\xi} = 0 \quad \text{on } z = 1, \tag{264}$$

$$w = 0 \quad \text{on } z = 0. \tag{265}$$

These leading-order equations give

$$p_0 = \eta_0, \quad u_0 = \eta_0, \quad w_0 + z\eta_{0\xi} = 0, \quad 0 \leq z \leq 1, \tag{266}$$

with  $u_0 = 0$  whenever  $\eta_0 = 0$ . The boundary condition on  $w_0$  at  $z = 1$  is automatically satisfied.

Using the Taylor series expansion of  $u, w$ , and  $p$  about  $z = 1$ , the two surface boundary conditions on  $z = 1 + \varepsilon\eta$  are rewritten on  $z = 1$  and, hence, take the form

$$p_0 + \varepsilon\eta_0 p_{0z} + \varepsilon p_1 = \eta_0 + \varepsilon\eta_1 + O(\varepsilon^2) \quad \text{on } z = 1. \tag{267}$$

$$w_0 + \varepsilon\eta_0 w_{0z} + \varepsilon w_1 = -\eta_{0\xi} - \varepsilon\eta_{1\xi} + \varepsilon(\eta_{0\tau} + u_0\eta_{0\xi}) + O(\varepsilon^2) \quad \text{on } z = 1. \tag{268}$$

These conditions are to be used together with (257), (258), and (261).

The equations in the next order are given by

$$-u_{1\xi} + u_{0\tau} + u_0 u_{0\xi} + w_0 u_{0z} = -p_{1\xi}, \quad p_{1z} = w_{0\xi}, \tag{269}$$

$$u_{1\xi} + w_{1z} = 0, \tag{270}$$

$$p_1 + \eta_0 p_0 = \eta_1, \quad w_1 + \eta_0 + w_{0z} = -\eta_{1\xi} + \eta_{0\tau} + u_0 \eta_{0\xi} \quad \text{on } z = 1, \tag{271}$$

$$w_1 = 0 \quad \text{on } z = 0. \tag{272}$$

Noting that

$$u_{0z} = 0, \quad p_{0z} = 0, \quad \text{and} \quad w_{0z} = -\eta_{0\xi}, \tag{273}$$

we obtain

$$p_1 = \frac{1}{2}(1 - z^2)\eta_{0\xi\xi} + \eta_1, \tag{274}$$

and therefore,

$$w_{1z} = -u_{1\xi} = -p_{1\xi} - u_{0\tau} - u_0 u_{0\xi} = -[\eta_{1\xi} + \frac{1}{2}(1 - z^2)\eta_{0\xi\xi\xi} + \eta_{0\tau} + \eta_0 \eta_{0\xi}]. \tag{275}$$

Finally, we find

$$w_1 = -\left[\eta_{1\xi} + \eta_{0\tau} + \eta_0 + \eta_{0\xi} + \frac{1}{2}\eta_{0\xi\xi\xi}\right]z + \frac{1}{6}z^3\eta_{0\xi\xi\xi}, \tag{276}$$

which satisfies the bottom boundary condition on  $z = 0$ .

The free surface boundary condition on  $z = 1$  gives

$$\begin{aligned} (w_1)_{z=1} &= -(\eta_{1\xi} + \eta_{0\tau} + \eta_0\eta_{0\xi} + \frac{1}{2}\eta_{0\xi\xi\xi}) + \frac{1}{6}\eta_{0\xi\xi\xi} \\ &= -\eta_{1\xi} + \eta_{0\tau} + 2\eta_0\eta_{0\xi}, \end{aligned} \tag{277}$$

which yields the equation for  $\eta_0(\xi, \tau)$  in the form

$$\eta_{0\tau} + \frac{3}{2}\eta_0\eta_{0\xi} + \frac{1}{6}\eta_{0\xi\xi\xi} = 0. \tag{278}$$

This is the *Korteweg–de Vries (KdV) equation*, which describes nonlinear plane gravity waves propagating in the  $x$ -direction. The exact solution of the general initial-value problem for the KdV equation can be obtained provided the initial data decay sufficiently rapidly as  $|\xi| \rightarrow \infty$ . We may raise the question of whether the asymptotic expansion for  $\eta$  (and hence, for the other flow variables) is uniformly valid as  $|\xi| \rightarrow \infty$  and as  $\tau \rightarrow \infty$ . For the case of  $\tau \rightarrow \infty$ , this question is difficult to answer because the equations for  $\eta_n (n \geq 1)$  are not easy to solve. However, all the available numerical evidence suggests that the asymptotic expansion of  $\eta$  is indeed uniformly valid as  $\tau \rightarrow \infty$  (at least for solutions which satisfy  $\eta \rightarrow 0$  as  $|\xi| \rightarrow \infty$ ). From a physical point of view, if the waves are allowed to propagate indefinitely, then other physical effects cannot be neglected. In the case of real water waves, the most common physical effects include viscous damping. In practice, the viscous damping seems to be sufficiently weak to allow the dispersive and nonlinear effects to dominate before the waves eventually decay completely.

### Two-Dimensional and Axisymmetric KdV Equations

It is well known that the KdV equation describes nonlinear plane waves which propagate only in the  $x$ -direction. However, there are many physical situations in which waves move on a two-dimensional surface. So it is natural to include both  $x$ - and  $y$ -directions with the appropriate balance of dispersion and nonlinearity. One of the simplest examples is the classical two-dimensional linear wave equation

$$u_{tt} = c^2(u_{xx} + u_{yy}), \tag{279}$$

which describes the propagation of long waves.

This equation has a solution in the form

$$u(\mathbf{x}, t) = a \exp [i(\omega t - \boldsymbol{\kappa} \cdot \mathbf{x})], \tag{280}$$

where  $a$  is the wave amplitude,  $\omega$  is the frequency,  $\mathbf{x} = (x, y)$ , and the wavenumber vector is  $\boldsymbol{\kappa} = (k, \ell)$ .

The dispersion relation is given by

$$\omega^2 = c^2(k^2 + \ell^2). \tag{281}$$

For waves that propagate predominantly in the  $x$ -direction, the wavenumber component  $\ell$  becomes small so that the approximate phase velocity is given by

$$c_p = c \left( 1 + \frac{1}{2} \frac{\ell^2}{k^2} \right) \text{ as } \ell \rightarrow 0. \tag{282}$$

It follows from (282) that the phase velocity suffers from a small correction provided by the wavenumber component  $\ell$  in the  $y$ -direction. In order to ensure that this small correction is the same order as the dispersion and nonlinearity, it is necessary to require  $\ell = O(\sqrt{\varepsilon})$  or  $\ell^2 = O(\varepsilon)$ . This requirement can be incorporated in the governing equations by a scaling of the flow variables as

$$y \rightarrow \sqrt{\varepsilon} y \quad \text{and} \quad v \rightarrow \sqrt{\varepsilon} v, \tag{283}$$

and using the same far-field transformations as (256).

Consequently, Eqs. (59)–(62) and the free surface condition (57ab) with the parameters  $\delta$  replaced by  $\varepsilon$  reduce to the form

$$-u_\xi + \varepsilon(u_\tau + uu_\xi + \varepsilon v u_y + w u_z) = -p_\xi, \tag{284}$$

$$-v_\xi + \varepsilon(v_\tau + uv_\xi + \varepsilon v v_y + w v_z) = -p_y, \tag{285}$$

$$\begin{aligned} -\varepsilon [-w_\xi + \varepsilon(w_\tau + uw_\xi + \varepsilon v w_y + w w_z)] \\ = -p_z, \end{aligned} \tag{286}$$

$$-u_\xi + \varepsilon v_y + w_z = 0, \tag{287}$$

$$\begin{aligned} w = -\eta_\xi + \varepsilon(\eta_\tau + uu_\xi + \varepsilon v \eta_y) \\ \text{and } p = \eta \text{ on } z = 1 + \varepsilon \eta, \end{aligned} \tag{288}$$

$$w = 0 \quad \text{on } z = 0. \tag{289}$$

We seek the same asymptotic series solution (262) valid as  $\varepsilon \rightarrow 0$  without any change of the leading order problem except that the flow variables involve  $y$  so that

$$\begin{aligned} p_0 = \eta_0, \quad u_0 = \eta_0, \quad w_0 = -z \eta_{0\xi}, \\ 0 \leq z \leq 1, \end{aligned} \tag{290}$$

$$v_{0\xi} = \eta_{0y}. \tag{291}$$

At the next order, the only difference from Sect. “[Solutions of the KdV Equation: Solitons and Cnoidal Waves](#)” is in the continuity equation which becomes

$$w_{1z} = -u_{1\xi} - v_{0y}. \tag{292}$$

This change leads to the following equation:

$$\begin{aligned} w_1 = -(\eta_{1\xi} + \eta_{0\tau} + \eta_0\eta_{0\xi} + \frac{1}{2}\eta_{0\xi\xi\xi} + v_{0y})z \\ + \frac{1}{6}z^3\eta_{0\xi\xi\xi}. \end{aligned} \tag{293}$$

Using Eq. (291), the final result is the equation for the leading-order representation of the surface wave in the form

$$2 \eta_{0\tau} + 3 \eta_0 \eta_{0\xi} + \frac{1}{3} \eta_{0\xi\xi} + v_{0y} = 0. \tag{294}$$

Consequently, differentiating (294) with respect to  $\xi$  and replacing  $v_{0\xi}$  by  $\eta_{0y}$  give the evolution equation for  $\eta_0(\xi, \tau, y)$  in the form

$$(2 \eta_{0\tau} + 3 \eta_0 \eta_{0\xi} + \frac{1}{3} \eta_{0\xi\xi\xi})_{\xi} + \eta_{0yy} = 0. \tag{295}$$

This is the *two-dimensional* or, more precisely, the  $(1 + 2)$ -dimensional KdV equation. Obviously, when there is no  $y$ -dependence, (295) reduces to the KdV Eq. (278). The two-dimensional KdV equation is also known as the *Kadomtsev–Petviashvili (KP) equation* (Kadomtsev and Petviashvili, [36]). This is another very special completely integrable equation, and it has an exact analytical solution that represents obliquely crossing nonlinear waves. Physically, any number of waves cross obliquely and interact nonlinearly. In particular, the nonlinear interaction of three waves leads to a *resonance phenomenon*. Such an interaction becomes more pronounced, leading to a strongly nonlinear interaction as the wave configuration is more nearly that of parallel waves. This situation can be interpreted as one in which the waves interact nonlinearly over a large distance so that a strong distortion is produced among these waves. The reader is referred to Johnson’s [35] book for more detailed information on oblique interaction of waves.

We now consider the Euler Eqs. (63) and the continuity Eq. (65) in cylindrical polar coordinates  $(r, \theta, z)$ . It is convenient to use the nondimensional flow variables, parameters, and scaled variables similar to those defined by (53), (54), and (58) with  $r = \ell r^*$  where  $r^*$  is a nondimensional variable so that the polar form of Euler Eqs. (66), continuity Eq. (65), the free surface boundary conditions (57ab) in nondimensional cylindrical polar form become

$$\begin{aligned} \frac{D u}{D t} - \frac{\varepsilon v^2}{r} &= -\frac{\partial p}{\partial r}, & \frac{D v}{D t} + \frac{\varepsilon u v}{r} &= -\frac{1}{r} \frac{\partial p}{\partial \theta}, \\ \delta \frac{D w}{D t} &= -\frac{\partial p}{\partial z}, \end{aligned} \tag{296}$$

$$\frac{1}{r} \frac{\partial}{\partial r}(r u) + \frac{1}{r} \frac{\partial v}{\partial \theta} + \frac{\partial w}{\partial z} = 0, \tag{297}$$

$$\begin{aligned} w &= \eta_t + \varepsilon \left( u \eta_r + \frac{v}{r} \eta_{\theta} \right) \\ &\text{and } p = \eta \text{ on } z = 1 + \varepsilon \eta, \end{aligned} \tag{298}$$

$$w = 0 \text{ on } z = 0, \tag{299}$$

where

$$\frac{D}{D t} = \frac{\partial}{\partial t} + \varepsilon \left( u \frac{\partial}{\partial r} + \frac{v}{r} \frac{\partial}{\partial \theta} + w \frac{\partial}{\partial z} \right), \tag{300}$$

and  $\varepsilon$  and  $\delta$  are defined by (52).

For the case of axisymmetric wave motions ( $\frac{\partial}{\partial \theta} = 0$ ), the governing equations become

$$u_t + \varepsilon(uu_r + wu_z) = -p_r, \tag{301}$$

$$\delta [w_t + \varepsilon(uw_r + ww_z)] = -p_z, \tag{302}$$

$$u_r + \frac{u}{r} + w_z = 0, \tag{303}$$

$$w = \eta_t + \varepsilon u \eta_r \text{ and } p = \eta \text{ on } z = 1 + \varepsilon \eta, \tag{304}$$

$$w = 0 \text{ on } z = 0. \tag{305}$$

In the limit as  $\varepsilon \rightarrow 0$ , the linearized version of Eqs. (296)–(300) become

$$u_t = -p_r, \quad v_t = -\frac{1}{r} p_{\theta}, \quad \delta w_t = -p_z, \tag{306}$$

$$\frac{1}{r} \frac{\partial}{\partial r}(r u) + \frac{1}{r} \frac{\partial v}{\partial \theta} + \frac{\partial w}{\partial z} = 0, \tag{307}$$

$$w = \eta_t \text{ and } p = \eta \text{ on } z = 1, \tag{308}$$

$$w = 0 \text{ on } z = 0, \tag{309}$$

For long waves ( $\delta \rightarrow 0$ ), Eqs. (306)–(309) lead to the classical wave equation

$$\eta_{tt} = \eta_{rr} + \frac{1}{r} \eta_r + \frac{1}{r^2} \eta_{\theta\theta}. \tag{310}$$

For axisymmetric surface waves, the wave Eq. (310) becomes

$$\eta_{tt} = \eta_{rr} + \frac{1}{r} \eta_r. \tag{311}$$

This can be solved by using the Hankel transform (see Debnath and Bhatta [21]).

It is convenient to introduce the characteristic variable  $\xi = r - t$  for outgoing waves and  $R = \alpha r$  so that  $\alpha \rightarrow 0$  which corresponds to large radius  $r$ . In other words,  $R = O(1)$ ,  $\alpha \rightarrow 0$  gives  $r \rightarrow \infty$ . The Eq. (311) reduces to the form

$$2 \eta_{\xi R} + \frac{1}{R} \eta_{\xi} + \alpha \left( \eta_{RR} + \frac{1}{R} \eta_R \right) = 0. \tag{312}$$

In the limit as  $\alpha \rightarrow 0$ , it follows that

$$\sqrt{R} \eta_{\xi} = g(\xi), \tag{313}$$

where  $g(\xi)$  is an arbitrary function of  $\xi$ .

For outgoing waves, the correct solution takes the form for  $\alpha \rightarrow 0 (r \rightarrow \infty)$ ,

$$\eta = \frac{1}{\sqrt{R}} \int g(\xi) d\xi = \frac{1}{\sqrt{R}} f(\xi), \tag{314}$$

where  $\eta = 0$  when  $f = 0$ .

This shows that the amplitude of axisymmetric waves decays as the radius  $r \rightarrow \infty (R \rightarrow \infty)$ . This behavior is totally different from the derivation of the KdV equation where the amplitude is uniformly  $O(\varepsilon)$ . In the present axisymmetric case, the amplitude decreases as the radius  $r$  increases so that there is no far-field region where the balance between nonlinearity and dispersion occurs. In other words, the amplitude is so small that nonlinear terms play no role at the leading order. However, there exists a scaling of the flow variables which leads to the *axisymmetric (concentric) KdV equation* as shown by Johnson [35].

We recall the axisymmetric Euler equations of motion and boundary conditions (301)–(305) and introduce scalings in terms of large radial variable  $R$  (see Johnson [35]),

$$\xi = \frac{\varepsilon^2}{\delta} (r - t) \quad \text{and} \quad R = \frac{\varepsilon^6}{\delta^2} r. \tag{315}$$

We next apply the transformations of the flow variables

$$(\eta, p, u, w) = \frac{\varepsilon^3}{\delta} \left( \eta^*, p^*, u^*, \frac{\varepsilon^2}{\delta} w^* \right), \tag{316}$$

where large distance/time is measured by the scale  $(\delta^2/\varepsilon^6)$  so that

$$\left( \frac{\delta^2}{\varepsilon^6} \right)^{-\frac{1}{2}} = \left( \frac{\varepsilon^3}{\delta} \right),$$

which represents the scale of the amplitude of the waves. It is important to point out that the original wave amplitude parameter  $\varepsilon$  can now be interpreted based on the amplitude of the wave for  $r = O(1)$  and  $t = O(1)$ . Consequently, the governing equations and the boundary conditions (301)–(305) become, dropping the asterisks in the variables,

$$-u_\xi + \alpha(uu_\xi + wu_z + \alpha u u_R) = -(p_\xi + \alpha p_R), \tag{317}$$

$$\alpha [-w_\xi + \alpha(uw_\xi + ww_z + \alpha u w_R)] = -p_z, \tag{318}$$

$$u_\xi + w_z + \alpha \left( u_R + \frac{1}{R} u \right) = 0, \tag{319}$$

$$w = -\eta_\xi + \alpha(u\eta_\xi + \alpha u \eta_R), \quad p = \eta \tag{320}$$

on  $z = 1 + \alpha\eta$ ,

$$w = 0 \quad \text{on} \quad z = 0, \tag{321}$$

where  $\alpha = (\delta^{-1}\varepsilon^4)$  is a new parameter. These equations are similar in structure to those discussed above with parameter  $\varepsilon$ , which is now replaced by  $\alpha$  in (317)–(321) so that the limit as  $\alpha \rightarrow 0$  is required. This requirement is satisfied (for example  $\varepsilon \rightarrow 0$  with  $\delta$  fixed), and the scaling introduced by (315) describes the region where the appropriate balance occurs so that the wave amplitude in this region is  $O(\alpha)$ .

We now seek an asymptotic series solution in the form

$$q(\xi, R, z) = \sum_{n=0}^{\infty} \alpha^n q_n(\xi, R, z), \quad \alpha \rightarrow 0, \tag{322}$$

where  $q$  represents each of  $\eta, u, w$ , and  $p$ .

In the leading order, we obtain the familiar equations

$$p_0 = \eta_0, \quad u_0 = \eta_0, \quad w_0 = -z \eta_{0\xi}, \tag{323}$$

$0 \leq z \leq 1.$

It follows from the continuity Eq. (319) that

$$w_{1z} = -u_{1z} - \left( u_{0R} + \frac{1}{R} u_0 \right). \tag{324}$$

Without any more algebraic calculation, it turns out that  $\eta_0(\xi, R)$  satisfies the nonlinear evolution equation

$$2 \eta_{0R} + \frac{1}{R} \eta_0 + 3 \eta_0 \eta_{0\xi} + \frac{1}{3} \eta_{0\xi\xi\xi} = 0. \tag{325}$$

This is usually referred to as the *axisymmetric (concentric) KdV equation* which includes a new term  $R^{-1}\eta_0$ . We may use the large time variable  $\tau = (\delta^{-2}\varepsilon^6)t$  so that  $R = \tau + \alpha\xi \approx \tau$  (see Johnson [35] or [34]).

Johnson [34] derived a new concentric KdV equation which incorporates weak dependence on the angular coordinate  $\theta$ . In the derivation of KP Eq. (295),  $\sqrt{\varepsilon}$  was chosen as the scaling parameter in the  $y$ -direction. Similarly, the appropriate scaling on the angular variable  $\theta$  may be chosen as  $\sqrt{\alpha}$ . In the derivation of the concentric KdV Eq. (325), the parameter  $\alpha$  plays the role of  $\varepsilon$  which is used as the small parameter in the asymptotic solution of the KdV equation.

Following the work of Johnson [34,35], we choose the variables  $\xi$  and  $R$  defined by (315) and the scaled  $\theta$  variable as

$$\theta = \sqrt{\alpha} \theta^* = (\delta^{-1}\varepsilon^2) \theta^*, \tag{326}$$

which introduces a small angular deviation from purely concentric effects. We also use the scaled velocity component in the  $\theta$ -direction as

$$v = (\delta^{-3/2}\varepsilon^5)v^*, \tag{327}$$

so that the scalings on  $u = \phi_r$  and  $v = \frac{1}{r} \phi_\theta$  are found to be consistent.

Using (315)–(316), Eqs. (296)–(300), and following Johnson [35], the equation in cylindrical polar coordinates can be obtained in the form:

$$\left(2\eta_{0R} + \frac{1}{R}\eta_0 + 3\eta_0\eta_{0\xi} + \frac{1}{3}\eta_{0\xi\xi\xi}\right)_\xi + \frac{1}{R^2}\eta_{0\theta\theta} = 0. \tag{328}$$

This is known as *Johnson’s (or nearly concentric KdV) equation*, as it was first derived by Johnson [34]. In the absence of the  $\theta$ -dependence, Eq. (328) reduces to the concentric KdV Eq. (325).

### The Nonlinear Schrödinger Equation and Solitary Waves

We describe below that the nonlinear modulation of a quasi-monochromatic wave described by the nonlinear Schrödinger equation. To take into account the nonlinearity and the modulation in the far-field approximation, the wavenumber  $k$  and frequency  $\omega$  in the linear dispersion relation are replaced by  $k - i\frac{\partial}{\partial x}$  and  $\omega + i\frac{\partial}{\partial t}$ , respectively. It is convenient to use the nonlinear dispersion relation in the form

$$D\left(k - i\frac{\partial}{\partial x}, \omega + i\frac{\partial}{\partial t}, |A|^2\right)A = 0. \tag{329}$$

We consider the case of a weak nonlinearity and a slow variation of the amplitude, and hence, the amplitude  $A$  is assumed to be a slowly varying function of space and time. We next expand (329) with respect to  $|A|^2$ ,  $-i\frac{\partial}{\partial x}$ , and  $i\frac{\partial}{\partial t}$  to obtain

$$D(k, \omega, 0) - i\left(D_k\frac{\partial}{\partial x} - D_\omega\frac{\partial}{\partial t}\right)A - \frac{1}{2}\left(D_{kk}\frac{\partial^2}{\partial x^2} - 2D_{k\omega}\frac{\partial^2}{\partial x\partial t} + D_{\omega\omega}\frac{\partial^2}{\partial t^2}\right)A + \frac{\partial D}{\partial |A|^2}|A|^2A = 0, \tag{330}$$

where the first term  $D(k, \omega, 0)$  corresponds to the linear dispersion so that  $D(k, \omega, 0) = 0$  represents the linear dispersion relation.

Introducing the transformation  $x^* = x - Ct$ ,  $t^* = t$ , assuming that  $A = O(\varepsilon)$ ,  $\frac{\partial}{\partial x^*} = O(\varepsilon)$ , and  $\frac{\partial}{\partial t^*} = O(\varepsilon^2)$ , retaining all terms up to  $O(\varepsilon^3)$ , and dropping the asterisks, we find that

$$i\frac{\partial A}{\partial t} + p\frac{\partial^2 A}{\partial x^2} + q|A|^2A = 0, \tag{331}$$

where

$$p = \frac{1}{2}\left(\frac{dC}{dk}\right), \quad q = \frac{1}{D_\omega}\left(\frac{\partial D}{\partial |A|^2}\right), \tag{332}$$

and the following results

$$C = -\frac{D\omega}{Dk}, \quad \frac{dC}{dk} = (D_{kk} + 2CD_{\omega k} + C^2D_{\omega\omega})/D_\omega, \tag{333}$$

have been used with  $D$  given by (329).

Equation (331) is known as the *cubic nonlinear Schrödinger (NLS) equation*. When the last cubic nonlinear term is neglected, Eq. (331) reduces to the corresponding *linear Schrödinger (NLS) equation* for  $A(x, t)$  in the form

$$i\frac{\partial A}{\partial t} + \frac{1}{2}\omega''(k)\frac{\partial^2 A}{\partial x^2} = 0, \tag{334}$$

when  $p = \frac{1}{2}C'(k) = \frac{1}{2}\omega''(k)$  is used.

More explicitly, if the nonlinear dispersion relation is given by

$$\omega = \omega(k, a^2), \tag{335}$$

and if we expand  $\omega$  in a Taylor series about  $k = k_0$  and  $|a|^2 = 0$ , we obtain

$$\omega \approx \omega_0 + (k - k_0)\omega'_0 + \frac{1}{2}(k - k_0)^2\omega''_0 + \left(\frac{\partial\omega}{\partial |a|^2}\right)_{|a|^2=0}|a|^2. \tag{336}$$

Replacing  $(\omega - \omega_0)$  by  $i\left(\frac{\partial}{\partial t}\right)$ ,  $k - k_0$  by  $-i\left(\frac{\partial}{\partial x}\right)$ , and assuming that the resulting operators act on the amplitude function  $a(x, t)$ , it turns out that  $a(x, t)$  satisfies the equation

$$i(a_t + \omega'_0 a_x) + \frac{1}{2}\omega''_0 a_{xx} + \gamma|a|^2 a = 0, \tag{337}$$

where

$$\gamma = -\left(\frac{\partial\omega}{\partial |a|^2}\right)_{|a|^2=0} = \text{constant}. \tag{338}$$

Equation (337) is known as the *cubic nonlinear Schrödinger equation*. If we choose a frame of reference moving with the linear group velocity  $\omega'_0$ , that is,  $\xi = x - \omega'_0 t$  and  $\tau = t$ , the term involving  $a_x$  will drop out from (337), and therefore, the amplitude  $a(x, t)$  satisfies the normalized *nonlinear Schrödinger equation*

$$i a_\tau + \frac{1}{2}\omega''_0 a_{\xi\xi} + \gamma|a|^2 a = 0. \tag{339}$$

The corresponding dispersion relation is given by

$$\omega = \frac{1}{2}\omega''_0 k^2 - \gamma a^2. \tag{340}$$

According to the stability criterion established by Whitham [62], the wave modulation is stable if  $\gamma\omega_0'' < 0$  or unstable if  $\gamma\omega_0'' > 0$ .

To study the solitary wave solution, it is convenient to use the NLS equation in the standard form

$$i\psi_t + \psi_{xx} + \gamma|\psi|^2\psi = 0, \quad -\infty < x < \infty, \quad t > 0. \quad (341)$$

We then seek waves of permanent form by assuming the solution

$$\psi = f(X)e^{i(mX-nt)}, \quad X = x - Ut, \quad (342)$$

for some functions  $f$  and constant wave speed  $U$  to be determined, and where  $m, n$  are constants.

Substitution of (342) in (341) gives

$$f'' + i(2m - U)f' + (n - m^2)f + \gamma|f|^2f = 0. \quad (343)$$

We eliminate  $f'$  by setting  $2m - U = 0$ , and then write  $n = m^2 - \alpha$ , so that  $f$  can be assumed real. Thus, Eq. (343) becomes

$$f'' - \alpha f + \gamma f^3 = 0. \quad (344)$$

Multiplying this equation by  $2f'$  and integrating, we find that

$$f'^2 = A + \alpha f^2 - \frac{\gamma}{2}f^4 = F(f), \quad (345)$$

where  $F(f) = (\alpha_1 - \alpha_2 f^2)(\beta_1 - \beta_2 f^2)$ , so that  $\alpha = -(\alpha_1\beta_2 + \alpha_2\beta_1)$ ,  $A = \alpha_1\beta_1$ ,  $\gamma = -2(\alpha_2\beta_2)$ , and the  $\alpha$ 's and  $\beta$ 's are assumed to be real and distinct.

Evidently, it follows from (345) that

$$X = \int_0^f \frac{df}{\sqrt{(\alpha_1 - \alpha_2 f^2)(\beta_1 - \beta_2 f^2)}}. \quad (346)$$

Setting  $(\alpha_2/\alpha_1)^{1/2}f = u$  in this integral (346), we deduce the following elliptic integral of the first kind (see Dutta and Debnath [22] and Helal and Molines [28]):

$$\sigma X = \int_0^f \frac{df}{\sqrt{(1-u^2)(1-\kappa^2 u^2)}}, \quad (347)$$

where  $\sigma = (\alpha_2\beta_1)^{1/2}$  and  $\kappa = (\alpha_1\beta_2)/(\beta_1\alpha_2)$ .

Thus, the final solution can be expressed in terms of the Jacobian elliptic function

$$u = sn(\sigma X, \kappa). \quad (348a)$$

Thus, the solution for  $f(X)$  is given by

$$f(X) = \left(\frac{\alpha_1}{\alpha_2}\right)^{1/2} sn(\sigma X, \kappa). \quad (348b)$$

In particular, when  $A = 0$ ,  $\alpha > 0$ , and  $\gamma > 0$ , we obtain a solitary wave solution. In this case, (345) can be rewritten

$$\sqrt{\alpha}X = \int_0^f \left\{f^2 \left(1 - \frac{\gamma}{2\alpha}f^2\right)\right\}^{-1/2} df. \quad (349)$$

Substitution of  $(\gamma/2\alpha)^{1/2}f = \text{sech } \theta$  in this integral gives the exact solution

$$f(X) = \left(\frac{2\alpha}{\gamma}\right)^{1/2} \text{sech}[\sqrt{\alpha}(x - Ut)]. \quad (350)$$

This represents a *solitary wave* solution that propagates without change of shape with constant velocity  $U$ . Unlike the solitary wave solution of the KdV equation, the amplitude and the velocity of the wave are independent parameters. It is noted that the solitary wave exists only for the unstable case ( $\gamma > 0$ ). This means that small modulations of the unstable wavetrain lead to a series of solitary waves.

The well-known nonlinear dispersion relation (116) for deep water waves is

$$\omega = \sqrt{gk} \left(1 + \frac{1}{2}a^2k^2\right). \quad (351)$$

Therefore,

$$\omega'_0 = \frac{\omega_0}{2k_0}, \quad \omega''_0 = \frac{\omega_0}{4k_0^2}, \quad \text{and } \gamma = -\frac{1}{2}\omega_0k_0^2, \quad (352)$$

and the NLS equation for deep water waves is obtained from (337) in the form

$$i \left(a_t + \frac{\omega_0}{2k_0}a_x\right) - \frac{\omega_0}{8k_0^2}a_{xx} - \frac{1}{2}\omega_0k_0^2|a|^2a = 0. \quad (353)$$

The normalized form of this equation in a frame of reference moving with the linear group velocity  $\omega'_0$  is

$$i a_t - \left(\frac{\omega_0}{8k_0^2}\right) a_{xx} - \frac{1}{2}\omega_0k_0^2|a|^2a = 0. \quad (354)$$

Since  $\gamma\omega_0'' = (\omega_0^2/8) > 0$ , this equation confirms the instability of deep water waves. This is one of the most remarkable recent results in the theory of water waves.

We next discuss the uniform solution and the solitary wave solution of the NLS Eq. (354). We look for solutions in the form

$$a(x, t) = A(X) \exp(i\gamma^2 t), \quad X = x - \omega'_0 t, \quad (355)$$

and substitute this in Eq. (354) to obtain the following equation:

$$A_{XX} = \frac{8k_0^2}{\omega_0} \left(\gamma^2 A + \frac{1}{2}\omega_0k_0^2A^3\right). \quad (356)$$

We multiply this equation by  $2A_X$  and, then, integrate to find

$$A_X^2 = - \left( A_0^4 m'^2 + \frac{8}{\omega_0} \gamma^2 k_0^2 A^2 + 2k_0^4 A^4 \right) = (A_0^2 - A^2)(A^2 - m'^2 A_0^3), \quad (357)$$

where  $A_0^4 m'^2$  is an integrating constant,  $2k_0^4 = 1$ ,  $m'^2 = 1 - m^2$ , and  $A_0^2 = 4\gamma^2/\omega_0 k_0^2(m^2 - 2)$ , which relates  $A_0$ ,  $\gamma$ , and  $m$ .

Finally, we rewrite Eq. (357) in the form

$$A_0^2 dX = \frac{dA}{\left[ \left( 1 - \frac{A^2}{A_0^2} \right) \left( \frac{A^2}{A_0^2} - m'^2 \right) \right]^{1/2}}, \quad (358a)$$

or, equivalently,

$$A_0(X - X_0) = \int^t \frac{ds}{\left[ (1 - s^2)(s^2 - m'^2) \right]^{1/2}}, \quad s = (A/A_0). \quad (358b)$$

This can readily be expressed in terms of the Jacobi  $dn$  function (see Dutta and Debnath [22] and Helal and Moulins [28]):

$$A = A_0 dn [A_0(X - X_0), m], \quad (359)$$

where  $m$  is the modulus of the  $dn$  function.

In the limit,  $m \rightarrow 0$ ,  $dn z \rightarrow 1$ , and  $\gamma^2 \rightarrow -\frac{1}{2}\omega_0 k_0^2 A_0^2$ . Hence, the solution becomes

$$a(x, t) = A(t) = A_0 \exp \left( -\frac{1}{2} i \omega_0 k_0^2 A_0^2 t \right). \quad (360)$$

On the other hand, when  $m \rightarrow 1$ ,  $dn z \rightarrow \operatorname{sech} z$ , and  $\gamma^2 \rightarrow -\frac{1}{4}\omega_0 k_0^2 A_0^2$ . Therefore, the solitary wave solution is

$$a(x, t) = A_0 \exp \left( -\frac{i}{4} \omega_0 k_0^2 A_0^2 t \right) \cdot \operatorname{sech} [A_0(x - \omega_0' t - X_0)]. \quad (361)$$

An analysis of this section reveals several remarkable features of the nonlinear Schrödinger equation. Like the KdV equation, the nonlinear Schrödinger equation is the lowest-order approximation for weakly and strongly nonlinear dispersive waves system. This equation can also be used to investigate instability phenomena in many other physical systems. Like the various forms of the KdV equation, the NLS equation arises in many physical problems, including nonlinear water waves and ocean waves, waves in plasma, propagation of heat pulses in a solid, self-trapping phenomena in nonlinear optics, nonlinear waves in a fluid-filled viscoelastic tube, and various nonlinear instability phenomena in fluids and plasmas.

### Whitham's Equations of Nonlinear Dispersive Waves

To describe a slowly varying nonlinear and nonuniform oscillatory wavetrain in a dispersive medium, we assume the existence of a one-dimensional solution in the form (see Whitham [62] or Debnath [20]), so that

$$\phi(x, t) = a(x, t) \exp \{ i\theta(x, t) \} + c.c., \quad (362)$$

where  $c.c.$  stands for the complex conjugate and  $a(x, t)$  is the complex amplitude (see Debnath [20]). The phase function  $\theta(x, t)$  is given by

$$\theta(x, t) = xk(x, t) - t\omega(x, t), \quad (363)$$

and  $k$ ,  $\omega$ , and  $a$  are slowly varying functions of space variable  $x$  and time  $t$ .

Because of the slow variations of  $k$  and  $\omega$ , it is reasonable to assume that these quantities still satisfy the linear Dispersion relation of the form

$$\omega = W(k). \quad (364)$$

Differentiating (363) with respect to  $x$  and  $t$ , respectively, we obtain

$$\theta_x = k + \{x - t W'(k)\} k_x, \quad (365)$$

$$\theta_t = -W(k) + \{x - t W'(k)\} k_t. \quad (366)$$

In the neighborhood of stationary points defined by  $W'(k) = (x/t) > 0$ , these equations become

$$\theta_x = k(x, t) \quad \text{and} \quad \theta_t = -\omega(x, t). \quad (367ab)$$

These results can be used as a definition of *local wavenumber* and *local frequency* of a slowly varying nonlinear wavetrain.

In view of (367), relation (364) gives a nonlinear partial differential equation for the phase  $\theta(x, t)$  in the form

$$\frac{\partial \theta}{\partial t} + W \left( \frac{\partial \theta}{\partial x} \right) = 0. \quad (368)$$

The solution of this equation determines the geometry of the wave pattern.

We eliminate  $\theta$  from (367ab) to obtain the equation

$$\frac{\partial k}{\partial t} + \frac{\partial \omega}{\partial x} = 0. \quad (369)$$

This is known as the *Whitham equation* for the conservation of waves, where  $k$  represents the *density* of waves and  $\omega$  is the *flux* of waves.

Using the dispersion relation (364), Eq. (369) gives

$$\frac{\partial k}{\partial t} + C(k) \frac{\partial k}{\partial x} = 0, \tag{370}$$

where  $C(k) = W'(k)$  is the group velocity. This represents the simplest nonlinear wave (hyperbolic) equation for the propagation of wavenumber  $k$  with group velocity  $C(k)$ .

Equations (370) and (364) reveal that  $\omega$  also satisfies the first-order, nonlinear wave (hyperbolic) equation

$$\frac{\partial \omega}{\partial t} + W'(k) \frac{\partial \omega}{\partial x} = 0. \tag{371}$$

It follows from Eqs. (370) and (371) that both  $k$  and  $\omega$  remain constant on the characteristic curves defined by

$$\frac{dx}{dt} = W'(k) = C(k) \tag{372}$$

in the  $(x, t)$ -plane. Since  $k$  or  $\omega$  is constant on each characteristic, the characteristics are straight lines with slope  $C(k)$ .

Finally, it follows from the preceding analysis that any constant value of the phase  $\theta$  propagates according to  $\theta(x, t) = \text{constant}$ , and hence,

$$\theta_t + \left(\frac{dx}{dt}\right) \theta_x = 0, \tag{373}$$

which gives, by (367ab),

$$\frac{dx}{dt} = -\frac{\theta_t}{\theta_x} = \frac{\omega}{k} = c(k). \tag{374}$$

Thus, the phase of the waves propagates with the phase speed  $c(k)$ . On the other hand, Eq. (370) ensures that the wavenumber  $k$  propagates with group velocity  $C(k) = (d\omega/dk) = W'(k)$ .

We next follow Whitham [62] or Debnath [20] to investigate how wave energy propagates in a dispersive medium. We consider the following integral involving the square of the wave amplitude (energy) between any two points  $x = x_1$  and  $x = x_2 (0 < x_1 < x_2)$ :

$$Q(t) = \int_{x_1}^{x_2} |A|^2 dx = \int_{x_1}^{x_2} AA^* dx \tag{375}$$

$$= 2\pi \int_{x_1}^{x_2} \frac{F(k)F^*(k)}{t|W''(k)|} dx, \tag{376}$$

which, due to a change of variable  $x = tW'(k)$ ,

$$= 2\pi \int_{k_1}^{k_2} F(k)F^*(k)dk, \tag{377}$$

where  $x_r = tW'(k_r), r = 1, 2$ .

When  $k_r$  is kept fixed as  $t$  varies,  $Q(t)$  remains constant so that

$$\begin{aligned} 0 &= \frac{dQ}{dt} = \frac{d}{dt} \int_{x_1}^{x_2} |A|^2 dx \\ &= \int_{x_1}^{x_2} \frac{\partial}{\partial t} |A|^2 dx + |A|^2_2 W'(k_2) - |A|^2_1 W'(k_1). \end{aligned} \tag{378}$$

In the limit, as  $x_2 - x_1 \rightarrow 0$ , this result reduces to the first-order, partial differential equation

$$\frac{\partial}{\partial t} |A|^2 + \frac{\partial}{\partial x} [W'(k)|A|^2] = 0. \tag{379}$$

This represents the equation for the conservation of wave energy where  $|A|^2$  and  $|A|^2 W'(k)$  are the energy density and energy flux, respectively. It also follows that the energy propagates with group velocity  $W'(k)$ . It has been shown that the wavenumber  $k$  also propagates with the group velocity. Thus, the group velocity plays a double role.

The preceding analysis reveals another important fact that (364), (369), and (379) constitute a closed set of equations for the three functions  $k, \omega$ , and  $A$ . Indeed, these are the fundamental equations for nonlinear dispersive waves and are known as *Whitham's equations*.

In his pioneering work on nonlinear dispersive waves, Whitham [62] formulated a more general energy equation based on the amount of energy  $Q(t)$  between two points  $x_1$  and  $x_2$

$$Q(t) = \int_{x_1}^{x_2} g(k)A^2 dx, \tag{380}$$

where  $g(k)$  is an arbitrary proportionality factor associated with the square of the amplitude and energy.

In a new coordinate system moving with the group velocity, that is, along the rays  $x = C(k)t$ , Eq. (380) reduces to the form

$$Q(t) = 2\pi \int_{x_1}^{x_2} g(k)F(k)F^*(k)dk, \tag{381}$$

where  $\omega = \Omega(k)$  and  $\Omega''(k) > 0$  and  $k_1$  and  $k_2$  are defined by  $x_1 = C(k_1)t$  and  $x_2 = C(k_2)t$ , respectively.

Using the principle of conservation of energy, that is, stating that the energy between the points  $x_1$  and  $x_2$  traveling with the group velocities  $C(k_1)$  and  $C(k_2)$  remains invariant, it turns out from (380) that

$$\begin{aligned} \frac{dQ}{dt} &= \int_{x_1}^{x_2} \frac{\partial}{\partial t} \{g(k)A^2\} dx + g(k_2)C(k_2)A^2(x_2, t) \\ &\quad - g(k_1)C(k_1)A^2(x_1, t) = 0, \end{aligned} \tag{382}$$

which is, in the limit as  $x_2 - x_1 \rightarrow 0$ ,

$$\frac{\partial}{\partial t} \{g(k)A^2\} + \frac{\partial}{\partial x} \{g(k)C(k)A^2\} = 0. \tag{383}$$

This may be treated as the more general energy equation. Quantities  $g(k)A^2$  and  $g(k)C(k)A^2$  represent the *energy density* and the *energy flux*, so that they are proportional to  $|A|^2$  and  $C(k)|A|^2$ , respectively. The flux of energy propagates with the group velocity  $C(k)$ . Hence, the group velocity has a double role: it is the propagation velocity for the wavenumber  $k$  and for the energy  $g(k)|A|^2$ .

Thus, (369) and (383) are known as *Whitham’s conservation equations* for nonlinear dispersive waves. The former represents the conservation of wave-number  $k$  and the latter is the conservation of energy (or more generally, the conservation of wave action). Whitham also derived the conservation equations more rigorously from a general and extremely powerful approach that is now known as *Whitham’s averaged variational principle*.

For slowly varying dispersive wavetrains, the solution maintains the elementary form  $u = \Phi(\theta, a)$ , but  $\omega, \kappa$ , and  $a$  are no longer constants, so that  $\theta$  is not a linear function of  $x_i$  and  $t$ . The local wavenumber and local frequency are defined by

$$k_i = \frac{\partial \theta}{\partial x_i}, \quad \omega = -\frac{\partial \theta}{\partial t}. \tag{384ab}$$

The *Whitham averaged Lagrangian* over the phase of the integral of the Lagrangian  $L$  is defined by

$$\mathcal{L}(\omega, \kappa, a, \mathbf{x}, t) = \frac{1}{2\pi} \int_0^{2\pi} L \, d\theta \tag{385}$$

and is calculated by assuming the uniform periodic solution  $u = \Phi(\theta, a)$  in  $L$ . Whitham first formulated the *averaged variational principle* in the form

$$\delta \iint \mathcal{L} \, d\mathbf{x} \, dt = 0, \tag{386}$$

to derive the equations for  $\omega, \kappa$ , and  $a$ .

It is noted that the dependence of  $\mathcal{L}$  on  $\mathbf{x}$  and  $t$  reflects possible nonuniformity of the medium supporting the wave motion. In a uniform medium,  $\mathcal{L}$  is independent of  $\mathbf{x}$  and  $t$ , so that the Whitham function  $\mathcal{L} = \mathcal{L}(\omega, \kappa, a)$ . However, in a uniform medium, some additional variables also appear only through their derivatives. They represent potentials whose derivatives are important physical quantities.

The Euler equations resulting from the independent variations of  $\delta a$  and  $\delta \theta$  in (386) with  $\mathcal{L} = \mathcal{L}(\omega, \kappa, a)$  are

$$\delta a : \mathcal{L}_a(\omega, \kappa, a) = 0, \tag{387}$$

$$\delta \theta : \frac{\partial}{\partial t} \mathcal{L}_\omega - \frac{\partial}{\partial x_i} \mathcal{L}_{\kappa_i} = 0. \tag{388}$$

The  $\theta$ -eliminant of (384ab) gives the consistency equations

$$\frac{\partial k_i}{\partial t} + \frac{\partial \omega}{\partial x_i} = 0, \quad \frac{\partial k_i}{\partial x_j} - \frac{\partial k_j}{\partial x_i} = 0. \tag{389ab}$$

Thus, (387)–(389) represent the *Whitham equations* for describing the slowly varying wavetrain in a nonuniform medium and constitute a closed set from which the triad  $\omega, \kappa$ , and  $a$  can be determined.

In linear problems, the Lagrangian  $L$ , in general, is a quadratic in  $u$  and its derivatives. Hence, if  $\Phi(\theta) = a \cos \theta$  is substituted in (385),  $\mathcal{L}$  must always take the form

$$\mathcal{L}(\omega, \kappa, a) = D(\omega, \kappa)a^2, \tag{390}$$

so that the dispersion relation ( $\mathcal{L}_a = 0$ ) must take the form

$$D(\omega, \kappa) = 0. \tag{391}$$

We note that the stationary value of  $\mathcal{L}$  is, in fact, zero for linear problems. In the simple case,  $L$  equals the difference between kinetic and potential energy. This proves the well-known principle of equipartition of energy, stating that average potential and kinetic energies are equal.

### Whitham’s Instability Analysis of Water Waves

Section “The Nonlinear Schrödinger Equation and Solitary Waves” deals with Whitham’s new remarkable variational approach to the theory of slowly varying, nonlinear, dispersive waves. Based upon Whitham’s ideas and, especially, Whitham’s fundamental dispersion Eq. (388), Lighthill [44,45] developed an elegant and remarkably simple result determining whether very gradual – not necessarily small – variations in the properties of a wavetrain are governed by hyperbolic or elliptic partial differential equations. A general account of Lighthill’s work with special reference to the instability of wavetrains on deep water was described by Debnath [19]. This section is devoted to the Whitham instability theory with applications to water waves.

According to Whitham’s nonlinear dispersive wave theory  $\mathcal{L}_a = 0$  gives a dispersion relation that depends on wave amplitude  $a$  has the form

$$\omega = \omega(k, a), \tag{392}$$

where equations for  $k$  and  $a$  are no longer uncoupled and constitute a system of partial differential equations. The first important question is whether these equations are hyperbolic or elliptic. This can be answered by a standard and simple method combined with Whitham’s conservation Eqs. (389a) and (383). For moderately small amplitudes, we use the Stokes expansion of  $\omega$  in terms of  $k$  and  $a^2$  in the form

$$\omega(k) = \omega_0(k) + \omega_2(k)a^2 + \dots \tag{393}$$

We substitute this result in (389a) and (383), replace  $\omega'(k)$  by its linear value  $\omega'_0(k)$ , and retain the terms of order  $a^2$  to obtain the equations for  $k$  and  $a^2$  in the form

$$\frac{\partial k}{\partial t} + [\omega'(k) + \omega'_2(k)a^2] \frac{\partial k}{\partial x} + \omega_2(k) \frac{\partial a^2}{\partial x} = 0, \quad (394)$$

$$\frac{\partial a^2}{\partial t} + \omega'_0(k) \frac{\partial a^2}{\partial x} + \omega''_0(k)a^2 \frac{\partial k}{\partial x} = 0. \quad (395)$$

Neglecting the term  $O(a^2)$ , these equations can be rewritten as

$$\frac{\partial k}{\partial t} + \omega'_0 \frac{\partial k}{\partial x} + \omega_2 \frac{\partial a^2}{\partial x} = 0, \quad (396)$$

$$\frac{\partial a^2}{\partial t} + \omega'_0(k) \frac{\partial a^2}{\partial x} + \omega''_0 a^2 \frac{\partial k}{\partial x} = 0. \quad (397)$$

These describe the modulations of a linear dispersive wavetrain and represent a coupled system due to the nonlinear dispersion relation (393) exhibiting the dependence of  $\omega$  on both  $k$  and  $a$ . In matrix form, these equations read

$$\begin{pmatrix} \omega'_0 & \omega_2 \\ \omega''_0 a^2 & \omega'_0 \end{pmatrix} \begin{pmatrix} \frac{\partial k}{\partial x} \\ \frac{\partial a^2}{\partial x} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial k}{\partial t} \\ \frac{\partial a^2}{\partial t} \end{pmatrix} = 0. \quad (398)$$

Hence, the eigenvalues  $\lambda$  are the roots of the determinant equation

$$|a_{ij} - \lambda b_{ij}| = \begin{vmatrix} \omega'_0 - \lambda & \omega_2 \\ \omega''_0 a^2 & \omega'_0 - \lambda \end{vmatrix} = 0, \quad (399)$$

where  $a_{ij}$  and  $b_{ij}$  are the coefficient matrices of (398). This determinant equation gives the characteristic velocities

$$\begin{aligned} \lambda &= \left( \frac{dx}{dt} \right) = C(k) \\ &= C_0(k) \pm a [\omega_2(k)\omega''_0(k)]^{1/2} + O(a^2), \end{aligned} \quad (400ab)$$

where  $C_0(k) = \omega'_0(k)$  is the linear group velocity, and, in general,  $\omega''_0(k) \neq 0$  for dispersive waves. The equations are hyperbolic or elliptic depending on whether  $\omega_2(k)\omega''_0(k) > 0$  or  $< 0$ .

In the hyperbolic case, the characteristics are real, and the double characteristic velocity splits into two separate velocities and provides a generalization of the group velocity of nonlinear dispersive waves. In fact, the characteristic velocities (400ab) are used to define the *nonlinear group velocities*. The splitting of the double characteristic velocity into two different velocities is one of the most significant results of the Whitham theory. This means that any initial disturbance of finite extent will eventually split into two separate disturbances. This prediction is radically different from that of the linearized theory, where an initial distur-

bance may suffer from a distortion due to dependence of the linear group velocity  $C_0(k) = \omega'_0(k)$  on the wavenumber  $k$ , but would *never split* into two. Another significant consequence of nonlinearity in the hyperbolic case is that compressive modulations will suffer from gradual distortion and steepening in the typical hyperbolic manner discussed earlier. This leads to the multiple-valued solutions and hence, eventually, breaking of waves.

In the elliptic case ( $\omega_2, \omega''_0 < 0$ ), the characteristics are imaginary. This leads to *ill-posed problems* in the theory of nonlinear wave propagation. Any small sinusoidal disturbances in  $a$  and  $k$  may be represented by solutions of the form  $\exp[ia\{x - C(k)t\}]$ , where  $C(k)$  is given by (400ab) for the unperturbed values of  $a$  and  $k$ . Thus, when  $C(k)$  is complex, these disturbances will grow exponentially in time, and hence, the periodic wavetrains become definitely *unstable*.

An application of this analysis to the Stokes waves on deep water reveals that the associated dispersion equation is elliptic in this case. For waves on deep water, the dispersion relation is

$$\omega = \sqrt{gk} \left( 1 + \frac{1}{2} a^2 k^2 \right) + O(a^4). \quad (401)$$

This result is compared with the Stokes expansion (393) to give  $\omega_0(k) = \sqrt{gk}$  and  $\omega_2(k) = \frac{1}{2} k^2 \sqrt{gk}$ . Hence,  $\omega''_0 \omega_2 < 0$ , the velocities (400ab) are complex, and the Stokes waves on deep water are definitely *unstable*. The instability of deep water waves came as a real surprise to researchers in the field in view of the very long and controversial history of the subject. The question of instability went unrecognized for a long period of time, even though special efforts have been made to prove the existence of a permanent shape for periodic water waves for all amplitudes less than the critical value at which the waves assume a sharp-crested form. However, Lighthill's [45] theoretical investigation into the elliptic case and Benjamin and Feir's [6] theoretical and experimental findings have provided conclusive evidence of the instability of Stokes waves on deep water.

For more details of nonlinear dispersive wave phenomena, the reader is also referred to Debnath [19].

In his pioneering work on nonlinear water waves, Whitham [62] observed that the neglect of dispersion in the nonlinear shallow water equations leads to the development of multivalued solutions with a vertical slope, and hence, eventually breaking occurs. It seems clear that the third derivative dispersive term in the KdV equation produces the periodic and solitary waves which are not found in the shallow water theory. However, the KdV equation cannot describe the observed symmetrical peaking of the

crests with a finite angle. On the other hand, the Stokes waves include full effects of dispersion, but they are limited to small amplitude, and describe neither the solitary waves nor the peaking phenomenon.

Although both breaking and peaking are without doubt involved in the governing equations of the exact potential theory, Whitham [62] developed a mathematical equation that can include all these phenomena. It has been shown earlier that the breaking of shallow water waves is governed by the nonlinear equation

$$\eta_t + c_0 \eta_x + d \eta \eta_x = 0, \quad d = 3c_0(2h_0)^{-1}. \quad (402)$$

On the other hand, the linear equation corresponding to a general linear dispersion relation

$$\frac{\omega}{k} = c(k) \quad (403)$$

is given by the integrodifferential equation in the form

$$\eta_t + \int_{-\infty}^{\infty} K(x-s)\eta_s(s,t)ds = 0, \quad (404)$$

where the kernel  $K$  is given by the inverse Fourier transform of  $c(k)$ :

$$K(x) = \mathcal{F}^{-1}\{c(k)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikx} c(k) dk. \quad (405)$$

Whitham combined the above ideas to formulate a new class of *nonlinear nonlocal* equations

$$\eta_t + d \eta \eta_x + \int_{-\infty}^{\infty} K(x-s)\eta_s(s,t)ds = 0. \quad (406)$$

This is well known as the *Whitham equation*, which can, indeed, describe symmetric waves that propagate without change of shape and peak at a critical height, as well as asymmetric waves that invariably break.

Once a wave breaks, it usually continues to travel in the form of a bore as observed in tidal waves. The weak bores have a smooth but oscillatory structure, whereas the strong bores have a structure similar to turbulence with no coherent oscillations. Since the region where waves break is a zone of high energy dissipation, it is natural to include a second derivative dissipative term in the KdV equation to obtain

$$\eta_t + c_0 \eta_x + d \eta \eta_x + \mu \eta_{xxx} - \nu \eta_{xx} = 0, \quad (407)$$

where  $\mu = \frac{1}{6}c_0 h_0^2$ . This is known as the *KdV–Burgers equation*, which also arises in nonlinear acoustics for fluids with gas bubbles (see [38]).

Whitham's [62] equations for the slow modulation of the wave amplitude  $a$  and the wavenumber  $k$  in the case of two-dimensional deep water waves are given by

$$\frac{\partial}{\partial t} \left( \frac{a^2}{\omega_0} \right) + \frac{\partial}{\partial x} \left( C \frac{a^2}{\omega_0} \right) = 0, \quad (408)$$

$$\frac{\partial k}{\partial t} + \frac{\partial \omega}{\partial x} = 0, \quad (409)$$

where  $\omega_0 = \sqrt{gk}$  is the first-order approximation for the wave frequency  $\omega(k)$  and  $C = (g/2\omega_0)$  is the group velocity.

Chu and Mei [14,15] observed that certain terms of the dispersive type, neglected in Whitham's equations to the same order of approximation, must be included to extend the validity of these equations. Whitham's theory is based on the direct use of Stokes' dispersion relation for a uniform wavetrain,

$$\omega = \omega_0 \left( 1 + \frac{1}{2} \varepsilon^2 a^2 k^2 \right), \quad (410)$$

whereas Chu and Mei added terms of higher derivatives and dispersive type to the expression for  $\omega$ , so that

$$\omega = \omega_0 \left[ 1 + \varepsilon^2 \left( \frac{1}{2} a^2 k^2 + \left\{ \left( \frac{a}{\omega_0} \right)_{tt} \div 2\omega_0 a \right\} \right) \right]. \quad (411)$$

They used the expression (411) to transform (408) and (409) in a frame of reference moving with the group velocity  $C(k)$  and obtained the following nondimensional equations:

$$\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x} (a^2 \phi_x) = 0, \quad (412)$$

$$-2 \frac{\partial^2 \phi}{\partial x \partial t} + \frac{\partial}{\partial x} \left[ -\phi_x^2 + \frac{a^2}{4} + \frac{a_{xx}}{16a} \right] = 0, \quad (413)$$

where we have used Chu and Mei's result  $W = -2\phi_x$  and  $\phi$  is a small phase variation. Integrating (413) with respect to  $x$  and setting the constant of integration to be zero gives

$$\phi_t + \frac{1}{2} \phi_x^2 - \frac{1}{8} a^2 - \frac{a_{xx}}{32a} = 0. \quad (414)$$

A transformation  $\Psi = a \exp(4i\phi)$  is used to simplify (412) and (414), which reduces to the *nonlinear Schrödinger equation*

$$i \Psi_t + \frac{1}{8} \Psi_{xx} + \frac{1}{2} \Psi |\Psi|^2 = 0. \quad (415)$$

This equation has also been derived and exploited by several authors, including Benney and Roskes [7], Hasimoto and Ono [27], and Davey and Stewartson [17] to examine the nonlinear evolution of Stokes' waves on water.

An alternative way to study nonlinear evolution of two- and three-dimensional wavepackets is to use the method of multiple scales in which the small parameter  $\varepsilon$  is explicitly built into the expansion procedure. The small parameter  $\varepsilon$  characterizes the wave steepness. This method has been employed by several authors in various fields and has also been used by Hasimoto and Ono [27] and Davey and Stewartson [17].

### The Benjamin–Feir Instability of the Stokes Water Waves

One of the simplest solutions of the nonlinear Schrödinger Eq. (354) is given by (360), that is,

$$A(t) = A_0 \exp\left(-\frac{1}{2}i\omega_0 k_0^2 A_0^2 t\right), \quad (416)$$

where  $A_0$  is a constant. This essentially represents the fundamental component of the Stokes wave. We consider a perturbation of (416) and express it in the form

$$a(x, t) = A(t) [1 + B(x, t)], \quad (417)$$

where  $B(x, t)$  is the perturbation function. Substituting this result in (354) gives

$$\begin{aligned} & i(1+B)A_t + iAB_t - \left(\frac{\omega_0}{8k_0^2}\right)AB_{xx} \\ &= \frac{1}{2}\omega_0 k_0^2 A_0^2 [(1+B) + BB^*(1+B) \\ & \quad + (B+B^*)B + (B+B^*)]A, \end{aligned} \quad (418)$$

where  $B^*(x, t)$  is the complex conjugate of the perturbed function  $B(x, t)$ . Neglecting squares of  $B$ , Eq. (418) reduces to

$$iB_t - \left(\frac{\omega_0}{8k_0^2}\right)B_{xx} = \frac{1}{2}\omega_0 k_0^2 A_0^2 (B + B^*). \quad (419)$$

We look for a solution for perturbed quantity  $B(x, t)$  in the form

$$\begin{aligned} B(x, t) = & B_1 \exp(\Omega t + i\ell x) \\ & + B_2 \exp(\Omega^* t - i\ell x), \end{aligned} \quad (420)$$

where  $B_1$  and  $B_2$  are complex constants,  $\ell$  is a real wavenumber, and  $\Omega$  is a growth rate (possibly a complex quantity) to be determined. Substituting the solution for  $B$  in (419) yields a pair of coupled equations:

$$\left(i\Omega + \frac{\omega_0 \ell^2}{8k_0^2}\right)B_1 - \frac{1}{2}\omega_0 k_0^2 A_0^2 (B_1 + B_2^*) = 0, \quad (421)$$

$$\left(i\Omega^* + \frac{\omega_0 \ell^2}{8k_0^2}\right)B_2 - \frac{1}{2}\omega_0 k_0^2 A_0^2 (B_1^* + B_2) = 0. \quad (422)$$

We take the complex conjugate of (422) to transform it into the form

$$\left(-i\Omega + \frac{\omega_0 \ell^2}{8k_0^2}\right)B_2^* - \frac{1}{2}\omega_0 k_0^2 A_0^2 (B_1 + B_2^*) = 0. \quad (423)$$

The pair of linear homogeneous Eqs. (421) and (423) for  $B_1$  and  $B_2^*$  admits a nontrivial eigenvalue for  $\Omega$  provided

$$\begin{vmatrix} i\Omega + \frac{\omega_0 \ell^2}{8k_0^2} - \frac{1}{2}\omega_0 k_0^2 A_0^2 & -\frac{1}{2}\omega_0 k_0^2 A_0^2 \\ -\frac{1}{2}\omega_0 k_0^2 A_0^2 & i\Omega + \frac{\omega_0 \ell^2}{8k_0^2} - \frac{1}{2}\omega_0 k_0^2 A_0^2 \end{vmatrix} = 0, \quad (424)$$

which is equivalent to

$$\Omega^2 = \frac{1}{2} \left(\frac{\omega_0 \ell}{2k_0}\right)^2 \left(k_0^2 A_0^2 - \frac{\ell^2}{8k_0^2}\right). \quad (425)$$

The growth rate  $\Omega$  is purely imaginary or real (and positive) depending on whether  $\ell^2 > 8k_0^4 A_0^2$  or  $\ell^2 < 8k_0^4 A_0^2$ . The former case represents a wave solution for  $B$ , and the latter corresponds to the Benjamin–Feir (or *modulational instability*) with a criterion in terms of the nondimensional wavenumber  $\tilde{\ell} = (\ell/k_0)$  as

$$\tilde{\ell}^2 < 8k_0^2 A_0^2. \quad (426)$$

Thus, the range of instability is given by

$$0 < \tilde{\ell} < \tilde{\ell}_c = 2\sqrt{2} k_0 A_0. \quad (427)$$

Since  $\Omega$  is a function of  $\tilde{\ell}$ , the maximum instability occurs at  $\tilde{\ell} = \tilde{\ell}_{\max} = 2k_0 A_0$ , with a maximum growth rate given by

$$(\text{Re } \Omega)_{\max} = \frac{1}{2}\omega_0 k_0^2 A_0^2. \quad (428)$$

To establish the connection with the Benjamin–Feir instability [6], we have to find the velocity potential for the fundamental wave mode multiplied by  $\exp(kz)$ . It turns out that the term proportional to  $B_1$  is the upper sideband, whereas that proportional to  $B_2$  is the lower sideband. The main conclusion of this analysis is that Stokes water waves are definitely *unstable*.

### Future Directions

Although this chapter has been devoted to water waves and the Korteweg and de Vries equations, there are some challenging problems dealing with solitary waves envelopes in the wake of a ship in oceans, and the KdV equation with variable coefficients in an ocean of variable depth. Despite some recent progress, there is no complete theory for ships in waves that takes into account of the effects of the finite ship volume, and possible interactions

between Kelvin wake and soliton envelopes. In order to respond to experimental results with a rough bottom, theories based on an empirical formula for the bottom stress have been developed, but they are not yet completely satisfactory when compared to experiments.

Of the systems that conserve energy, some are completely integrable in the sense of solitary wave theory, but many including those most important for applications are not. In some cases, useful analytical techniques are currently available, and others are waiting to be discovered. So there is a need for research in the area of nonlinear lattices. When we deal with real world problems, the conservation of energy does not hold because of dissipative effects and forcing terms. Consequently, the KdV equation and other relevant evolution equations need modification by including terms that describe such effects. So, there are challenging problems dealing with these modified evolution equations, their methods of solutions and the qualitative and quantitative understanding of the solutions. On the other hand, prospects for research in the transverse coupling between solitary waves and parallel systems are indeed bright.

## Bibliography

### Primary Literature

1. Airy GB (1845) Tides and Waves. In: Encyclopedia Metropolitana, Article 192
2. Akylas TR (1984) On the excitation of long nonlinear water waves by a moving pressure distribution. *J Fluid Mech* 141:455–466
3. Akylas TR (1984) On the excitation of nonlinear water waves by a moving pressure distribution oscillatory at resonant frequency. *Phys Fluids* 27:2803–2807
4. Amick CJ, Fraenkel LE, Toland JF (1982) On the Stokes Conjecture for the wave of extreme form. *Acta Math Stockh* 148:193–214
5. Benjamin TB (1967) Instability of periodic wave trains in nonlinear dispersive systems. *Proc Roy Soc London A* 299:59–75
6. Benjamin TB, Feir JE (1967) The disintegration of wavetrains on deep water. Part 1. Theory. *J Fluid Mech* 27:417–430
7. Benney DJ, Roskes G (1969) Wave instabilities. *Stud Appl Math* 48:377–385
8. Berezin YA, Karpman VI (1966) Nonlinear evolution of disturbances in plasmas and other dispersive media. *Soviet Phys JETP* 24:1049–1056
9. Boussinesq MJ (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation se propageant dans un canal rectangulaire. *Comptes Rendus* 72:755–759
10. Boussinesq MJ (1871) Théorie generale des mouvements qui sont propagés dans un canal rectangulaire horizontal. *Comptes Rendus* 72:755–759
11. Boussinesq MJ (1872) Théorie des ondes et des rumeurs qui se propagent le long d'un canal rectagulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. *J Math Pures Appl ser 2/17*:55–108
12. Boussinesq MJ (1877) Essai sur la théorie des eaux courants. *Mémoires Acad Sci Inst Fr ser 2* 23:1–630
13. Cercignani C (1977) Solitons. Theory and application. *Rev Nuovo Cimento* 7:429–469
14. Chu VH, Mei CC (1970) On slowly-varying Stokes waves. *J Fluid Mech* 41:873–887
15. Chu VH, Mei CC (1971) The nonlinear evolution of Stokes waves in deep water. *J Fluid Mech* 47:337–351
16. Crapper GD (1957) An exact solution for progressive capillary waves of arbitrary amplitude. *J Fluid Mech* 2:532–540
17. Davey A, Stewartson K (1974) On three-dimensional packets of surface waves. *Proc Roy Soc London A* 338:101–110
18. Debnath L, Rosenblat S (1969) The ultimate approach to the steady state in the generation of waves on a running stream. *Quart J Mech Appl Math XXII*:221–233
19. Debnath L (1994) *Nonlinear Water Waves*. Academic Press, Boston
20. Debnath L (2005) *Nonlinear Partial Differential Equations for Scientists and Engineers (Second Edition)*. Birkhauser, Boston
21. Debnath L, Bhatta D (2007) *Integral Transforms and Their Applications. Second Edition*. Chapman Hall/CRC Press, Boca Raton
22. Dutta M, Debnath L (1965) *Elements of the Theory of Elliptic and Associated Functions with Applications*. World Press Pub. Ltd., Calcutta
23. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems I. Los Alamos Report LA 1940. In: Newell AC (ed) *Lectures in Applied Mathematics*. American Mathematical Society, Providence 15:143–156
24. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the KdV equation. *Phys Rev Lett* 19:1095–1097
25. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equation and generalizations, VI, Methods for exact solution. *Comm Pure Appl Math* 27:97–133
26. Hammack JL, Segur H (1974) The Korteweg–de Vries equation and water waves. Part 2. Comparison with experiments. *J Fluid Mech* 65:289–314
27. Hasimoto H, Ono H (1972) Nonlinear modulation of gravity waves. *J Phys Soc Jpn* 35:805–811
28. Helal MA, Molines JM (1981) Nonlinear internal waves in shallow water. A Theoretical and Experimental study. *Tellus* 33:488–504
29. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. *Phys Rev Lett* 27:1192–1194
30. Hirota R (1973) Exact envelope-soliton solutions of a nonlinear wave equation. *J Math Phys* 14:805–809
31. Hirota R (1973) Exact N-Solutions of the wave equation of long waves in shallow water and in nonlinear lattices. *J Math Phys* 14:810–814
32. Huang DB, Sibul OJ, Webster WC, Wehausen JV, Wu DM, Wu TY (1982) Ship moving in a transcritical range. *Proc. Conf. on Behavior of Ships in Restricted Waters (Varna, Bulgaria)* 2:26-1–26-10
33. Infeld E (1980) On three-dimensional generalizations of the Boussinesq and Korteweg–de Vries equations. *Quart Appl Math XXXVIII*:277–287
34. Johnson RS (1980) Water waves and Korteweg–de Vries equations. *J Fluid Mech* 97:701–719

35. Johnson RS (1997) *A Modern Introduction to the Mathematical Theory of Water Waves*. Cambridge University Press, Cambridge
36. Kadomtsev BB, Petviashvili VI (1970) On the stability of solitary waves in weakly dispersive media. *Sov Phys Dokl* 15:539–541
37. Kaplan P (1957) The waves generated by the forward motion of oscillatory pressure distribution. In: *Proc. 5th Midwest Conf. Fluid Mech*, pp 316–328
38. Karpman VI (1975) *Nonlinear Waves in Dispersive Media*. Pergamon Press, London
39. Korteweg DJ and de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Phil Mag* (5) 39:422–443
40. Kraskovskii YP (1960) On the theory of steady waves of not all small amplitude (in Russian). *Dokl Akad Nauk SSSR* 130:1237–1252
41. Kraskovskii YP (1961) On the theory of permanent waves of finite amplitude (in Russian). *Zh Vychisl Mat Fiz* 1:836–855
42. Lax PD (1968) Integrals of nonlinear equations of evolution and solitary waves. *Comm. Pure Appl Math* 21:467–490
43. Levi-Civita T (1925) Détermination rigoureuse des ondes permanentes d'amplitude finie. *Math Ann* 93:264–314
44. Lighthill MJ (1965) Group velocity. *J Inst Math Appl* 1:1–28
45. Lighthill MJ (1967) Some special cases treated by the Whitham theory. *Proc Roy Soc London A* 299:38–53
46. Lighthill MJ (1978) *Waves in Fluids*. Cambridge University Press, Cambridge
47. Luke JC (1967) A variational principle for a fluid with a free surface. *J Fluid Mech* 27:395–397
48. Michell JH (1893) On the highest waves in water. *Phil Mag* 36:430–435
49. Miche R (1944) Mouvements ondulatoires de la mer en profondeur constants ou décroissante. *Forme limite de la houle lors de son déferlement. Application aux digues maritimes*. *Ann Ponts Chaussées* 114:25–78, 131–164, 270–292, 369–406
50. Palais RS (1997) The symmetries of solitons. *Bull Amer Math Soc* 34:339–403
51. Rayleigh L (1876) On waves. *Phil Mag* 1:257–279
52. Riabouchinsky D (1932) Sur l'analogie hydraulique des mouvements d'un fluide compressible, Institut de France, Académie des Sciences. *Comptes Rendus* 195:998
53. Stoker JJ (1957) *Water Waves*. Interscience, New York
54. Stokes G (1847) On the theory of oscillatory waves. *Trans Camb Phil Soc* 8:197–229
55. Struik DJ (1926) Détermination rigoureuse des ondes irrotationnelles périodiques dans un canal à profondeur finie. *Math Ann* 95:595–634
56. Toland JF (1978) On the existence of a wave of greatest height and Stokes' conjecture. *Proc. Roy. Soc. Lond. A* 363:469–485
57. Weidman PD, Maxworthy T (1978) Experiments on strong interactions between solitary waves. *J Fluid Mech* 85:417–431
58. Whitham GB (1965) A general approach to linear and nonlinear dispersive waves using a Lagrangian. *J Fluid Mech* 22:273–283
59. Whitham GB (1965) Nonlinear dispersive waves, *Proc Roy Soc London A* 283:238–261
60. Whitham GB (1967) Nonlinear dispersion of water waves. *J Fluid. Mech* 27:399–412
61. Whitham GB (1967) Variational methods and application to water waves. *Proc Roy Soc London A* 299:6–25
62. Whitham GB (1974) *Linear and Nonlinear Waves*. John Wiley, New York
63. Whitham GB (1984) Comments on periodic waves and solitons. *IMA J Appl Math* 32:353–366
64. Wu DM, Wu TY (1982) Three-dimensional nonlinear long waves due to a moving pressure. In: *Proc 14th Symp. Naval Hydrodyn. National Academy of Sciences, Washington DC*, pp 103–129
65. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence of initial states. *Phys Rev Lett* 15:240–243
66. Zabusky NJ (1967) A synergetic approach to problems of nonlinear dispersive wave propagation and interaction. In: Ames WF (ed) *Proc. Symp. on Nonlinear Partial Differential Equations*. Academic Press, Boston. pp 223–258
67. Zabusky NJ, Galvin CJ (1971) Shallow water waves, the Korteweg–de Vries equation and solitons. *J Fluid Mech* 48:811–824

### Books and Reviews

- Ablowitz MJ, Clarkson PA (1991) *Solitons, Nonlinear Evolution Equations and Inverse Scattering*. Cambridge University Press, Cambridge
- Bhatnagar PL (1979) *Nonlinear Waves in One-Dimensional Dispersive Systems*. Oxford University Press, Oxford
- Crapper GD (1984) *Introduction to Water Waves*. Ellis Horwood Limited, Chichester
- Debnath L (1983) *Nonlinear Waves*. Cambridge University Press, Cambridge
- Dodd RK, Eilbeck JC, Gibbons JD, Morris HC (1982) *Solitons and Nonlinear Wave Equations*. Academic Press, London
- Drazin PG, Johnson RS (1989) *Solitons: An Introduction*. Cambridge University Press, Cambridge
- Infeld E, Rowlands G (1990) *Nonlinear Waves, Solitons and Chaos*. Cambridge University Press, Cambridge
- Lamb H (1932) *Hydrodynamics*, 6th edition. Cambridge University Press, Cambridge
- Leibovich S, Seebass AR (1972) *Nonlinear Waves*. Cornell University Press, Ithaca
- Myint-U T, Debnath L (2007) *Linear Partial Different Equations for Scientists and Engineers* (fourth edition). Birkhauser, Boston
- Russell JS (1844) Report on waves. In: Murray J (ed) *Report on the 14th Meeting of the British Association for the Advancement of Science*. London 311–390

---

## Wavelets, Introduction to

EDWARD ABOUFADEL

Department of Mathematics,  
Grand Valley State University, Allendale, USA

Wavelets emerged in the late 1980s as a valuable new tool in science and engineering, and as a topic for fruitful mathematical research. Wavelet transforms are now regularly applied in areas such as image processing, statistical analysis, and seismic research. Wavelets are at the heart of

the WSQ standard used by the United States' Federal Bureau of Investigation to compress fingerprint images [1]. They have been implemented in the Red professional digital video camera [2] and they are now being used in an attempt to analyze paintings by famous artists [3]. In the past twenty years, related “-lets” tools have also emerged, such as ridgelets and shearlets (see ► [Curvelets and Ridgelets](#)). The 1992 book by Daubechies [4] is a classic in the field.

Wavelet analysis shares with Fourier analysis the key idea of having a basis of functions with which to analyze signals and other functions. While Fourier analysis uses sine and cosine functions (or complex exponentials) – which are smooth and periodic – wavelet analysis utilizes functions that tend to be rough and localized (see ► [Wavelets and the Lifting Scheme](#) for examples of wavelet functions). To create the basis in Fourier analysis, the frequencies of the standard functions are varied. However, to create a system of wavelet functions, a scaling function is modified by translations and dilations, leading to various resolutions and hence the term *multiresolution* that is frequently used in the field. There are both discrete and continuous Fourier transforms, and we have the same for wavelets (see ► [Comparison of Discrete and Continuous Wavelet Transforms](#)). There are also analogues of the Fast Fourier Transform. Ironically, much of the derivation of results in wavelets uses the Fourier transform as a theoretical tool.

To be more specific, the creation of a wavelet basis starts with a *scaling function*  $\phi$  and a *mother wavelet*  $\psi$ . (The scaling function is sometimes called the “father wavelet”.) These functions tend to either have compact support (meaning they are equal to 0 outside of a bounded interval), or decay exponentially outside of a relatively small part of their domain. Translations and dilations of the mother wavelet are created in the form  $\psi(2^n t - k)$ , where  $t$  and  $k$  are integers. The resolution varies with  $n$ , with larger positive values of  $n$  corresponding to finer resolutions. By varying  $k$ , the support of the function can be translated, allowing localized analysis of other functions or signals.

Wavelet transformations are described and applied in multiple ways. As functions, wavelets can be studied with either the real line or the complex plane as a domain. High pass and low pass wavelet filters can be applied to tease out details in signals or images (see ► [Popular Wavelet Families and Filters and Their Use](#)). To think of wavelet transformations as a combination of “prediction” and “update” steps is the main idea behind the lifting scheme (see ► [Wavelets and the Lifting Scheme](#)). In addition, working with Fourier transforms of wavelet functions lead to various polyphase results (see ► [Multiwavelets](#), for example).

As suggested above, the use of wavelets in various applications has contributed to their popularity. In statistics, wavelets are used to denoise signals and estimate properties of zero-mean stochastic processes (see ► [Statistical Applications of Wavelets](#)). Wavelet techniques have been combined with methods from partial differential equations to clean up images, or fill in damaged parts of an image (which is called *inpainting*) (see ► [Wavelets and PDE Techniques in Image Processing, A Quick Tour of](#)). Because images from astronomy tend to be isotropic (that is, having similar shapes or values in all directions, such as a sphere), wavelet methods have also been useful in that discipline (see ► [Numerical Issues When Using Wavelets](#)).

Wavelet analysis is preferred to Fourier analysis in many applications. One reason is that the relative roughness of the wavelet functions is a better match than trigonometric functions are for real data. This leads to a rather sparse representation of this data in the wavelet domain, which is helpful for compression and noise reduction. The Fourier transform breaks down a function into different frequencies; consequently this representation is based on averages of the function over its whole domain. Wavelets give you information about a function at different scales and different locations simultaneously. The localized nature of wavelet basis functions leads to better identification and elimination of local behavior in data such as spikes, edges, and other discontinuities. In addition, the nature of wavelets allows analysis at multiple resolutions, from coarse to fine.

A fertile area of research in wavelets is to consider extensions where the domain and/or range of the wavelet functions are multi-dimensional. For instance, since images are two-dimensional objects, a natural area to investigate are situations where the domain is  $\mathbf{R}^2$  (see ► [Bivariate \(Two-dimensional\) Wavelets](#)). Basic wavelet families are orthogonal, but they don't need to be (see ► [Numerical Issues When Using Wavelets](#)). In order to work with radial images such as retinal scans [5], or anisotropic images (which are long and narrow, or tall and thin), other wavelet tools have been developed (see ► [Curvelets and Ridgelets](#)). Multiwavelets are functions on the real line whose range is two or more copies of the complex plane (see ► [Multiwavelets](#)). Finally, there is an intimate connection between wavelets and splines (see ► [Multivariate Splines and Their Applications](#)).

For the reader unfamiliar with wavelets, sections I, II, and III of the ► [Wavelets and the Lifting Scheme](#) article, along with the glossary of the ► [Comparison of Discrete and Continuous Wavelet Transforms](#) article, are good places to begin. A familiarity with the Fourier transform is critical to read many of the articles in this section.

## Bibliography

1. Bradley J, Brislawn C, Hopper T (1993) The FBI wavelet/scalar quantization standard for gray-scale fingerprint image compression. In: Proc Conf Visual Info Process II Proc SPIE, vol 1961. pp 293–304
2. RED Digital Cinema Camera Company [www.red.com](http://www.red.com)
3. Lyu S, Rockmore D, Farid H (2004) A digital technique for art authentication. Proc Nat Acad Sci 101:17006–17010
4. Daubechies I (1992) Ten lectures on wavelets. SIAM
5. Feng P et al (2007) Enhancing retinal image by the Contourlet transform. Pattern Recognit Lett 28:516–522

## Wavelets and the Lifting Scheme

ANDERS LA COUR–HARBO<sup>1</sup>, ARNE JENSEN<sup>2</sup>

<sup>1</sup> Department of Electronics Systems, Aalborg University, Aalborg East, Denmark

<sup>2</sup> Department of Mathematical Sciences, Aalborg University, Aalborg East, Denmark

### Article Outline

Glossary

Definition of the Subject

Introduction

Lifting

Prediction and Update

Interpretation

Lifting and Filter Banks

Making Lifting Steps from Filters

Bibliography

### Glossary

**Lifting step** An equation describing the transformation of an odd (even) sample of a signal by means of a linear combination of even (odd) samples, respectively. Changing an odd sample is sometimes called a prediction step, while changing an even sample is called an update step.

**Discrete wavelet transform (DWT)** Transformation of a discrete (sampled) signal into another discrete signal by means of a wavelet basis. The transform can be accomplished in a number of ways, typically either by a two channel filter bank or by lifting steps.

**Filter bank** A series of bandpass filters, which separates the input signal into a number of components, each with a distinct range of frequencies of the original signal.

**z-Transform** A mapping of a vector  $\mathbf{x} = \{x[k]\}$  to a function in the complex plane by

$$X(z) = \sum_k x[k]z^{-k}.$$

For a vector with a finite number of non-zero entries  $X(z)$  is a Laurent polynomial.

**Filter** A linear map, which maps a signal  $\mathbf{x}$  with finite energy to another signal with finite energy. In the time domain it is given by convolution with a vector  $\mathbf{h}$ , which in the frequency domain is equal to  $H(z)X(z)$ . The vector  $\mathbf{h}$  is called the impulse response (IR) of the filter (or sometimes the filter taps), and  $H(e^{j\omega})$  the transfer function (or sometimes the frequency response). If  $\mathbf{h}$  is a finite sequence, then  $\mathbf{h}$  is called a FIR (finite impulse response) filter. An infinite  $\mathbf{h}$  is then called an IIR filter. We only consider FIR filters in this article. We restrict ourselves to filters with real coefficients.

**Filter taps** The entries in the vector that by convolution with a signal gives a filtering of the signal. The filter taps are also called the impulse response of the filter.

**Laurent polynomial** A polynomial in the variables  $z$  and  $z^{-1}$ . Some examples:  $3z^{-2} + 2z$ ,  $z^{-3} - 2z^{-1}$ , and  $1 + z + z^3$ . The  $z$ -transform of a FIR filter  $\mathbf{h}$  is a Laurent polynomial of the form

$$H(z) = \sum_{k=k_b}^{k_e} h[k]z^{-k}, \quad k_b \leq k_e.$$

This is in contrast to ordinary polynomials, where we only have non-negative powers of  $z$ . Assuming that  $h[k_e] \neq 0$  and  $h[k_b] \neq 0$ , the degree of a Laurent polynomial is defined as  $|h| = k_e - k_b$ .

**Monomial** A Laurent polynomial of degree 0, for example  $3z^7$ ,  $z^{-8}$ , or 12.

### Definition of the Subject

The objective of this article is to give a concise introduction to the discrete wavelet transform (DWT) based on a technique called *lifting*. The lifting technique allows one to give an elementary, but rigorous, definition of the DWT, with modest requirements on the reader. A basic knowledge of linear algebra and signal processing will suffice. The lifting based definition is equivalent to the usual filter bank based definition of the DWT. The article does not discuss applications in any detail. The reader is referred to other articles in this collection.

The DWT was introduced in the second half of the eighties, through the work of Y. Meyer, I. Daubechies, S. Mallat, and many other researchers. The lifting technique was introduced by W. Sweldens in 1996, and the equivalence with the filter bank approach was established by I. Daubechies and W. Sweldens in an article published in 1998 [2], but available in preprint form from 1996.

This article is based on the book *Ripples in Mathematics – The Discrete Wavelet Transform* [3] with kind permission of Springer Science and Business Media. For more information on the background and history of wavelets, we again refer to other articles in this collection.

## Introduction

We begin the exposition of the lifting technique with a simple, yet very descriptive example. It will shortly become apparent how this is related to lifting and to the discrete wavelet transform. We take a digital signal consisting of eight samples (this idea is due to Mulcahy [5], and we use his example, with a minor modification)

56, 40, 8, 24, 48, 48, 40, 16.

We choose to believe that these numbers are not random, but have some structure that we want to extract. Perhaps there is some correlation between a number and its immediate successor. To reveal such a correlation we will take the numbers in pairs and compute the mean, and the difference between the first member of the pair and the computed mean. The result of this computation is shown as the second row in Table 1. This row contains the four means followed by the four differences, the latter being typeset in boldface. We then leave the four differences unchanged and apply the mean and difference computations to the first four entries of the second row. We repeat this procedure once more on the third row. The fourth row then contains a first entry, which is the mean of the original eight numbers, and the seven calculated differences. The boldface entries in the table are called the details of the signal.

It is important to observe that no information has been lost in this transformation of the first row into the fourth row. We can see this by reversing the calculation. Beginning with the last row, we compute the first two entries in the third row as  $32 = 35 + (-3)$  and  $38 = 35 - (-3)$ . In the same way the first four entries in the second row are calculated as  $48 = 32 + (16)$ ,  $16 = 32 - (16)$ ,  $48 = 38 + (10)$ , and finally  $28 = 38 - (10)$ . Repeating this procedure we get the first row in the table.

### Wavelets and the Lifting Scheme, Table 1

Mean and difference computation. Differences are in *boldface type*

56	40	8	24	48	48	40	16
48	16	48	28	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>
32	38	<b>16</b>	<b>10</b>	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>
35	<b>-3</b>	<b>16</b>	<b>10</b>	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>

So, what is the purpose of these computations? The four signals contain the same information, just in different ways, so how do we gain anything from this change of representation of the signal? If our assumption (pairwise equality of samples) is correct, the four means in the second row will equal the original numbers, and the four differences will be zero. If further, the original numbers are equal in groups of four, the two means in the third row will equal the original quadruples, and the differences will be zero. Finally, if all eight original samples are equal, the mean in row four will be equal to all eight samples, and the seven differences will be zero.

Apparently, our assumption is not correct, as the differences are not zero. However, the numbers in the fourth row are generally smaller than the original numbers. This indicates that while the numbers are not equal, they are ‘similar’, leaving us with small differences. We can say that we have achieved some kind of loss-free compression by reducing the dynamic range of the signal. By loss-free we mean that we can transform back to the original signal without any information being lost in the process. As a simple measure of compression we can count the number of digits used to represent the signal. The first row contains 15 digits. The last row contains 12 digits and two negative signs. So in this example the compression is not very large. But it is easy to give other examples, where the compression can be substantial.

We see in this example that the pair 48, 48 do fit our assumption of equality, and therefore the difference is zero. Suppose that after transformation we find that many ‘difference entries’ are zero. Then we can store the transformed signal more efficiently by only storing the non-zero entries (and their locations).

## Processing the Transformed Signal

Let us now suppose that we are willing to accept a certain loss of quality in the signal, if we can get a higher rate of compression. One technique for lossy compression is called thresholding. We choose a threshold and decide to set equal to zero all entries with an absolute value less than this threshold value. Let us in our example choose 4 as the threshold. This means that we in Table 1 replace the entry  $-3$  by **0** and then perform the reconstruction. The result is in Table 2, left hand part.

The original and transformed signals are both shown on the left in Fig. 1. We have chosen to join the sample points by straight line segments to get a good visualization of the signals. Clearly the two graphs differ very little. Now let us perform a more drastic compression. This time we choose the threshold equal to 9. The computations are

### Wavelets and the Lifting Scheme, Table 2

Reconstruction with threshold 4 (*top*) and threshold 9 (*bottom*)

59	43	11	27	45	45	37	13
51	19	45	25	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>
35	35	<b>16</b>	<b>10</b>	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>
35	<b>0</b>	<b>16</b>	<b>10</b>	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>

51	51	19	19	45	45	37	13
51	19	45	25	<b>0</b>	<b>0</b>	<b>0</b>	<b>12</b>
35	35	<b>16</b>	<b>10</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>12</b>
35	<b>0</b>	<b>16</b>	<b>10</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>12</b>

given on the right in Table 2, and the graphs are plotted on the right in Fig. 1. Notice that the peaks in the original signal have been flattened. We also note that the signal now is represented by only four non-zero entries.

There are several variations of the transformation. We could have stored differences instead of ‘half-differences’, or we could have used the difference between the second element of the pair and the computed average. The first choice will lead to boldface entries in the tables that can be obtained from the computed ones by multiplication by a factor  $-2$ . The second variant is obtained by multiplication by  $-1$ .

The ‘mean-difference’ transformation can be performed on any signal of even length, since all we need is the possibility of taking pairs. If we want to be able to keep transforming the signal until there is one signal mean left, we need a signal of length  $2^N$ , and it will lead to a table with  $N + 1$  rows, where the first row is the original signal. If the given signal has a length different from a power of 2, then we will have to do some additional operations on the signal to compensate for that. One possibility is to add sam-

ples with value zero to one or both ends of the signal until a length of  $2^N$  is achieved. This is called zero padding.

### Lifting

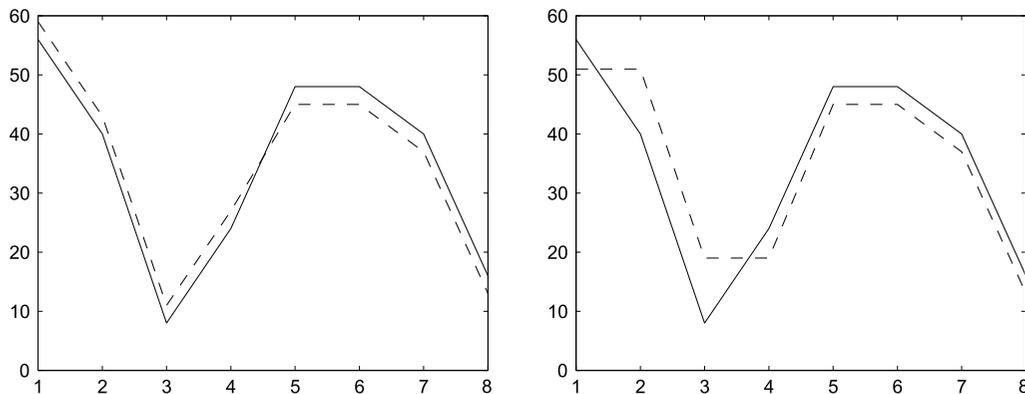
The transformation performed by successively calculating means and differences of a signal is an example of the discrete wavelet transform. Specifically, the transform introduced in the example is called the Haar transform, and occasionally also the first of the Daubechies transforms. It can be undone by simply reversing the steps performed, and it provides a number of different representations of the same signal. Actually, all discrete wavelet transforms can be realized by a similar procedure, where the only modification from the Haar transform is the way, in which we compute the means part and differences part of the transformed signal. The procedure is known as ‘lifting’ in the literature, and was introduced by Wim Sweldens in 1996 in a series of papers [8,9].

To fully appreciate the concept of lifting we need to elaborate on the example in the previous section. We have a discrete signal of real (or complex) numbers

$$x[0], x[1], x[2], x[3], \dots, x[N - 1].$$

Our convention is to start indexing at zero. In implementations one may need to adapt this to the programming language used. Here we assume the signal to have finite length  $N$ , but this is not a restriction. The theory also applies to infinite signals, provided they have finite energy. In the mathematical (and signal processing) literature finite energy means

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 < \infty,$$



Wavelets and the Lifting Scheme, Figure 1

Original signal (*solid*) and modified signal (*dashed*) with threshold 4 (*left*) and 9 (*right*)

and the set of such signals is usually denoted by  $\ell^2(\mathbf{Z})$ . Sometimes we use the mathematical term sequence instead of signal, and we also use the term vector, in particular in connection with use of results from linear algebra.

### Introducing Prediction and Update

Let us now return to the example in the Introduction. We took a pair of numbers  $a, b$  and computed the mean, and the difference between the first entry and the mean

$$s = \frac{a + b}{2}, \quad (1)$$

$$d = a - s. \quad (2)$$

Alternatively, almost the same result can be achieved in the following way

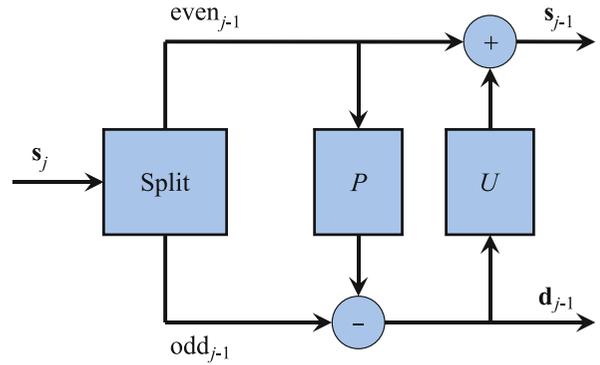
$$d = a - b, \quad (3)$$

$$s = \frac{a + b}{2} = \frac{d}{2} + b, \quad (4)$$

except we now have the difference rather than half the difference. This mathematical formulation of the Haar transform, albeit quite simple, brings us to the definition of lifting. The two operations, mean and difference, can be viewed as special cases of more general operations. Remember that we assumed that there is some correlation between two successive samples, and we therefore computed the difference. If two samples are almost equal, the difference is, of course, small. Thus one can use the first sample to predict the second sample, the prediction being that they are equal. If  $s$  is a good prediction, the difference between them is small. Thus we can call (3) a *prediction step*. We can use other and more sophisticated prediction steps than one based on just the previous sample. We will give a few examples of this later.

We also calculated the mean of the two samples. This step can be viewed in two ways. Either as an operation, which preserves some properties of the original signal (later we shall see how the mean value or the energy of a signal is preserved during transformation), or as an extraction of essential features of the signal. The latter viewpoint is based on the fact that the pair-wise mean values contain the overall structure of the signal, but with only half the number of samples. We use the term *update step* for the this operation, and (4) is the update step in the Haar transform. Just as the prediction operation, the update operation can be more sophisticated than just calculating the mean.

The prediction and update operations are shown in Fig. 2. The notation and the setup is a little different from



Wavelets and the Lifting Scheme, Figure 2

The three steps in a lifting building block. Note that the minus sign means 'the signal from the left minus the signal from the top'

the Introduction; we start with a finite sequence  $s_j$  of length  $2^j$  and end with two sequences  $s_{j-1}$  and  $d_{j-1}$ , each of length  $2^{j-1}$ . Let us explain the steps in Fig. 2.

**split** The entries are sorted into the even and the odd entries. It is important to note that we do this only to explain the algorithm. In (effective) implementations the entries are not moved or separated.

**prediction** If the signal contains some structure, then we can expect correlation between a sample and its nearest neighbors. In our first example the prediction is that the signal is constant. More elaborately, given the value at the sample number  $2n$ , we predict that the value at sample  $2n + 1$  is the same. We then compute the difference and in the figure let it replace the value at  $2n + 1$ , since this value is not needed anymore. In our notation this is

$$d_{j-1}[n] = s_j[2n + 1] - s_j[2n].$$

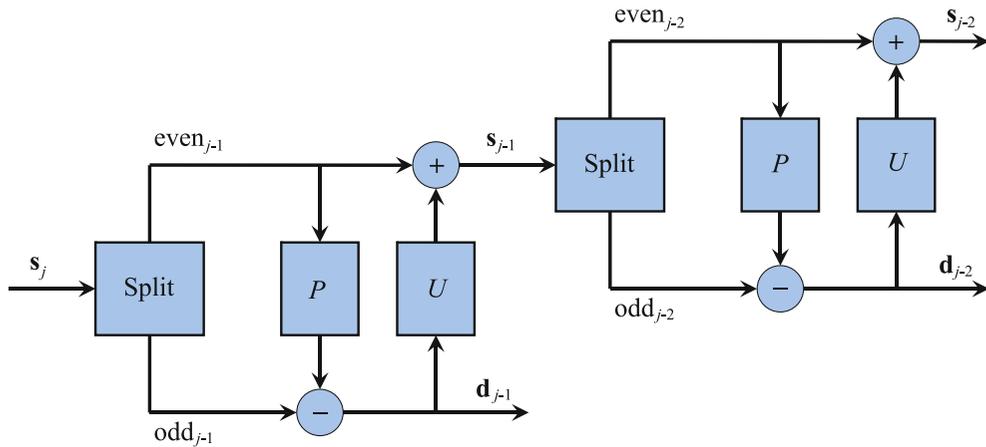
In general, the idea is to have a prediction procedure  $P$  and then compute the difference signal as

$$d_{j-1} = \text{odd}_{j-1} - P(\text{even}_{j-1}). \quad (5)$$

Thus in the  $d$  signal each entry is one odd sample minus some prediction based on a number of even samples.

**update** Given an even entry we predicted that the next odd entry has the same value, and stored the difference. We then update our even entry to reflect our knowledge of the signal. In the example above we replaced the even entry by the average. In our notation

$$s_{j-1}[n] = s_j[2n] + d_{j-1}[n]/2.$$



Wavelets and the Lifting Scheme, Figure 3  
Two step discrete wavelet transform

In general we decide on an updating procedure, and then compute

$$s_{j-1} = \text{even}_{j-1} + U(\mathbf{d}_{j-1}). \tag{6}$$

The algorithm described here is called one step lifting. It requires the choice of a prediction procedure  $P$ , and an update procedure  $U$ .

**Multiple Transforms**

The discrete wavelet transform is obtained by combining a number of lifting steps. As in the example in Table 1 we keep the computed differences  $\mathbf{d}_{j-1}$  and use the average sequence  $s_{j-1}$  as input for one more lifting step. This two step procedure is illustrated in Fig. 3. Obviously, we can apply more lifting steps if we want to transform the signal further. In general, we start with a signal  $s_j$  of length  $2^j$  and we can repeat the transformation  $j$  times until we have a single number  $s_0[0]$  and  $j$  difference sequences  $\mathbf{d}_{j-1}$  to  $\mathbf{d}_0$ . If we let  $j = 3$  and use the Haar transform  $s_0[0]$  will be the mean value of the eight entries in the original sequence, and  $\mathbf{d}_0$ ,  $\mathbf{d}_1$ , and  $\mathbf{d}_2$  will be the numbers in boldface in Table 1. In lifting step notation this table becomes as Table 3.

We have previously motivated the prediction operation with the reduction in dynamic range of the signal obtained in using differences rather than the original values, potentially leading to good compression of a signal. The update procedure has not yet been clearly motivated. The update performed in the first example was

$$s_{j-1}[n] = s_j[2n] + \frac{1}{2}d_{j-1}[n] = \frac{1}{2}(s_j[2n] + s_j[2n + 1]).$$

It turns out that this update operation preserves the mean value. The consequence is that all the  $s$  sequences have the

Wavelets and the Lifting Scheme, Table 3  
This is Table 1 in the lifting step notation

$s_3[0]$	$s_3[1]$	$s_3[2]$	$s_3[3]$	$s_3[4]$	$s_3[5]$	$s_3[6]$	$s_3[7]$
$s_2[0]$	$s_2[1]$	$s_2[2]$	$s_2[3]$	$d_2[0]$	$d_2[1]$	$d_2[2]$	$d_2[3]$
$s_1[0]$	$s_1[1]$	$d_1[0]$	$d_1[1]$	$d_2[0]$	$d_2[1]$	$d_2[2]$	$d_2[3]$
$s_0[0]$	$d_0[0]$	$d_1[0]$	$d_1[1]$	$d_2[0]$	$d_2[1]$	$d_2[2]$	$d_2[3]$

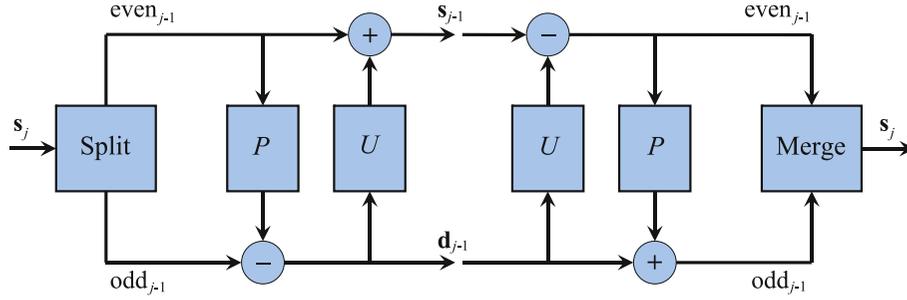
same mean value. It is easy to verify in the case of the example in Table 1, since

$$\frac{56 + 40 + 8 + 24 + 48 + 48 + 40 + 16}{8} = \frac{48 + 16 + 48 + 28}{4} = \frac{32 + 38}{2} = 35.$$

It is not difficult to see that this holds for any sequence  $s$  of length  $2^j$ . In particular,  $s_0[0]$  equals the mean value of the original samples  $s_j[0], \dots, s_j[2^j - 1]$ .

**Inverse Transform**

A lifting step is easy to undo. In fact, looking at Fig. 2 it is clear that we can go backwards from  $s_{j-1}$  and  $\mathbf{d}_{j-1}$  to  $s_j$  simply by first ‘undoing’ the update step and then ‘undoing’ the prediction step. This is accomplished by using the same update step as in the direct lifting step, but subtracting instead of adding, followed by the same prediction step, but using addition instead of subtraction. Finally, we need to merge the odd and even samples. The direct lifting step and the corresponding inverse lifting step are shown in Fig. 4. This merging step, where the sequences  $\text{even}_{j-1}$  and  $\text{odd}_{j-1}$  are merged to form the sequence  $s_j$ , is shown to explain the algorithm. It is not performed in implementations, since the entries are not reordered in the first place.



Wavelets and the Lifting Scheme, Figure 4  
Direct and inverse lifting step

Mathematically, the inverse lifting step is accomplished by using the formulas in opposite order, and isolating the odd and even samples. In our example the inverse lifting step becomes

$$b = s - d/2, \quad (7)$$

$$a = d + b, \quad (8)$$

where (7) is actually (4) with  $b$  isolated, and (8) is (3) with  $a$  isolated. In the lifting step notation this is first (6) rewritten as

$$s_j[2n] = s_{j-1}[n] - d_{j-1}[n]/2$$

to undo the update step, and then (5) rewritten as

$$s_j[2n + 1] = d_{j-1}[n] + s_j[2n]$$

to undo the prediction step. For the general lifting step

$$\mathbf{d}_{j-1} = \text{odd}_{j-1} - P(\text{even}_{j-1})$$

$$s_{j-1} = \text{even}_{j-1} + U(\mathbf{d}_{j-1})$$

the inversion is accomplished by the steps

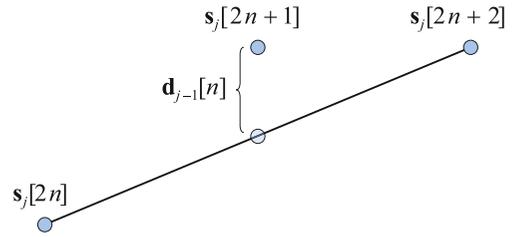
$$\text{even}_{j-1} = s_{j-1} - U(\mathbf{d}_{j-1})$$

$$\text{odd}_{j-1} = \mathbf{d}_{j-1} + P(\text{even}_{j-1}).$$

That is all there is to the inverse transform.

### Prediction and Update

Assuming that a signal is constant was useful in the first example to demonstrate the concept of lifting. However, this assumption is hardly the best for most signals. Fortunately, there are many other possible prediction and update procedures to accommodate various signal assumptions. We will present two examples of other lifting steps. The first example is the next logical step from the Haar transform when we assume the signal to be stepwise linear rather than just constant. The second example is the Daubechies 4 transform, which is not based on a specific



Wavelets and the Lifting Scheme, Figure 5

The linear prediction uses two even samples to predict one odd sample

assumption about the signal, but is taken from the classical wavelet theory and adapted to the lifting step method.

### Linear Prediction

Instead of basing the prediction on the assumption that the signal is constant, we will now base it on the assumption that the signal is linear. We really do mean an affine signal, but we stick to the commonly used term 'linear.' Thus a linear signal is in this context a signal of the form  $s_j[n] = \alpha n + \beta$ , that is, all the samples of the signal lie on a straight line. For a given odd entry  $s_j[2n + 1]$  we now need the two nearest even neighbors to predict the value (remember that for the constant signal we needed only one neighbor). The predicted value is  $\frac{1}{2}(s_j[2n] + s_j[2n + 2])$ , the mean of the two even samples. This value is the circle on the middle of the line in Fig. 5.

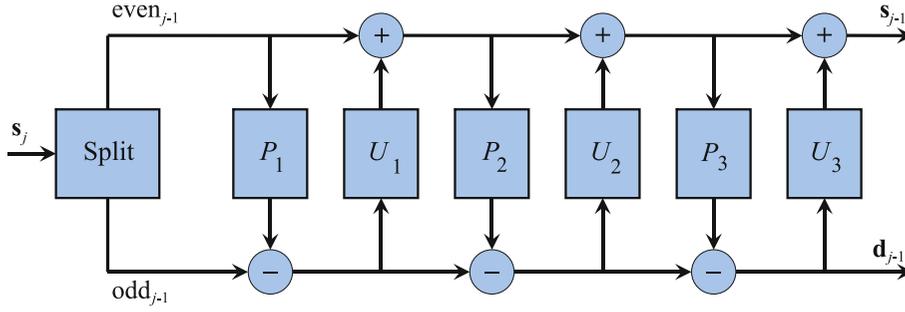
The correction is then the difference between what we predict the middle sample to be and what it actually is

$$d_{j-1}[n] = s_j[2n + 1] - \frac{1}{2}(s_j[2n] + s_j[2n + 2]).$$

We decide to base the update procedure on the two most recently computed differences. We take it of the form

$$s_{j-1}[n] = s_j[2n] + A(d_{j-1}[n - 1] + d_{j-1}[n]),$$

where  $A$  is a constant to be determined. In the first exam-



Wavelets and the Lifting Scheme, Figure 6  
Multiple prediction and update steps

ple we had the property

$$\sum_n s_{j-1}[n] = \frac{1}{2} \sum_n s_j[n], \quad (9)$$

that is, the mean was preserved. (Recall that  $s_j$  has length  $2^j$  and  $s_{j-1}$  has length  $2^{j-1}$ . This explains the factor  $\frac{1}{2}$ .) We would like to have the same property here. Let us first rewrite the expression for  $s_{j-1}[n]$  above,

$$\begin{aligned} s_{j-1}[n] &= s_j[2n] + Ad_{j-1}[n-1] + Ad_{j-1}[n] \\ &= s_j[2n] + A(s_j[2n-1] - \frac{1}{2}s_j[2n-2] - \frac{1}{2}s_j[2n]) \\ &\quad + A(s_j[2n+1] - \frac{1}{2}s_j[2n] - \frac{1}{2}s_j[2n+2]). \end{aligned}$$

Using this expression, and gathering even and odd terms, we get

$$\sum_n s_{j-1}[n] = (1-2A) \sum_n s_j[2n] + 2A \sum_n s_j[2n+1].$$

To satisfy (9) we must choose  $A = \frac{1}{4}$ . Summarizing, we have the following two steps

$$d_{j-1}[n] = s_j[2n+1] - \frac{1}{2}(s_j[2n] + s_j[2n+2]), \quad (10)$$

$$s_{j-1}[n] = s_j[2n] + \frac{1}{4}(d_{j-1}[n-1] + d_{j-1}[n]). \quad (11)$$

The transform in this example also has the property.

$$\sum_n ns_{j-1}[n] = \frac{1}{4} \sum_n ns_j[n]. \quad (12)$$

Note that there is a misprint in this formula in [3]. We say that the transform preserves the *first moment* of the sequence. The mean is also called the *zeroth moment* of the sequence. Finally, we want to note that this transform is known in the literature as CDF(2,2). The origin of this name is explained later.

In the above presentation we have simplified the notation by not specifying where the finite sequences start and end, thereby for the moment avoiding keeping track of the ranges of the variables. In other words, we have considered our finite sequences as infinite, adding zeroes before and after the given entries. In implementations one has to keep track of these things, but doing so now would obscure the simplicity of the lifting procedure. For more details on transformation of finite signals, see Chap. 10 in [3].

#### Daubechies 4 Wavelet

We now turn to the second example, the Daubechies 4 wavelet. This transform is a bit different from what we have seen so far. So to adapt this transform to the lifting step method we need to first take a closer look at the construction of the lifting step. Looking at Fig. 4 once more we see that we can add another prediction step after the update step without changing the lifting concept; we still have even and odd samples, we still have  $s_{j-1}$  and  $d_{j-1}$  as output, and we can still invert the lifting step by reversing the order of the prediction and update steps. In fact, we can add as many prediction and update steps as we want, and still have the same lifting concept. As an illustration of this Fig. 6 shows a direct transform consisting of three pairs of prediction and update operations. If we have two prediction and only one update, but insist on having them in pairs (this is occasionally useful in the theory), we can always add an operation of either type, which does nothing.

It turns out that this generalization is crucial in applications. There are many important transforms, where the steps do not occur in pairs. The Daubechies 4 transform is an example, where there is a  $U$  operation followed by a  $P$  operation and another  $U$  operation. Furthermore, in the last two steps, in (16) and (17), we add a new type of operation which is called normalization, or sometimes rescaling. The resulting algorithm is applied to a signal  $s_j$

as follows

$$s_{j-1}^{(1)}[n] = s_j[2n] + \sqrt{3}s_j[2n+1] \quad (13)$$

$$d_{j-1}^{(1)}[n] = s_j[2n+1] - \frac{1}{4}\sqrt{3}s_{j-1}^{(1)}[n] - \frac{1}{4}(\sqrt{3}-2)s_{j-1}^{(1)}[n-1] \quad (14)$$

$$s_{j-1}^{(2)}[n] = s_{j-1}^{(1)}[n] - d_{j-1}^{(1)}[n+1] \quad (15)$$

$$s_{j-1}[n] = \frac{\sqrt{3}-1}{\sqrt{2}}s_{j-1}^{(2)}[n] \quad (16)$$

$$d_{j-1}[n] = \frac{\sqrt{3}+1}{\sqrt{2}}d_{j-1}^{(1)}[n]. \quad (17)$$

Since there is more than one  $U$  operation, we have used superscripts on the  $\mathbf{s}$  and  $\mathbf{d}$  signals in order to tell them apart. Note that in the normalization steps we have

$$\frac{\sqrt{3}-1}{\sqrt{2}} \cdot \frac{\sqrt{3}+1}{\sqrt{2}} = 1.$$

Normalization does not change the information on the signal, but is merely a means to have certain transform properties satisfied.

The Daubechies 4 equations are derived from the classical wavelet theory by a procedure called Euclidean factorization, see Sect. “[Lifting and Filter Banks](#)”. While the Daubechies 4 transform does have a well-described origin in mathematics (see for instance Ten Lectures on Wavelets [1] by Daubechies), it becomes somewhat unclear in the lifting step version, what exactly this transform can provide. In particular, starting with an update step does not conform with our previous exposition of the lifting approach. However, this transform has one important property not possessed by the linear prediction transform CDF(2,2), or indeed any predicting lifting steps based on polynomial interpolation; in a frequency interpretation of the wavelet transforms the Daubechies 4 behaves somewhat nicer than CDF(2,2). This is because the former is an orthogonal transform, while the latter is a biorthogonal transform. We refer to other wavelet articles in this collection for more information on this subject.

While the Daubechies 4 is lacking a proper interpretation in terms of prediction and update, and thus may appear to be less appealing to include in the lifting concept, it's ability to be written as lifting step is in fact a consequence of the pleasant fact that all wavelet transforms (not matter their origin and design criteria) can be written as lifting steps. The two examples then show that some lifting steps comes from a sample-by-sample design method,

while others may come from completely different design methods (the Daubechies transform was designed within the field of filter theory to have certain nice frequency related properties). This article touches upon the filter subject, see Sect. “[Lifting and Filter Banks](#)”. For a more in-depth coverage of the subject, we refer the reader to other articles in this collection.

To find the inverse transform we have to use the prescription given above. We do the steps in reverse order and with the signs reversed. The normalization is undone by multiplication by the inverse constants. The result is

$$d_{j-1}^{(1)}[n] = \frac{\sqrt{3}-1}{\sqrt{2}}d_{j-1}[n] \quad (18)$$

$$s_{j-1}^{(2)}[n] = \frac{\sqrt{3}+1}{\sqrt{2}}s_{j-1}[n] \quad (19)$$

$$s_{j-1}^{(1)}[n] = s_{j-1}^{(2)}[n] + d_{j-1}^{(1)}[n+1] \quad (20)$$

$$s_j[2n+1] = d_{j-1}^{(1)}[n] + \frac{1}{4}\sqrt{3}s_{j-1}^{(1)}[n] + \frac{1}{4}(\sqrt{3}-2)s_{j-1}^{(1)}[n-1] \quad (21)$$

$$s_j[2n] = s_{j-1}^{(1)}[n] - \sqrt{3}s_j[2n+1]. \quad (22)$$

Incidentally, this transform illustrates one of the problems that has to be faced in computer implementations. For example, to compute  $d_{j-1}^{(1)}[0]$  we need to know  $s_{j-1}^{(1)}[0]$  and  $s_{j-1}^{(1)}[-1]$ . But to compute  $s_{j-1}^{(1)}[-1]$  one needs the values  $s_j[-2]$  and  $s_j[-1]$ , which are not defined. The easiest solution to this problem is to use zero padding to get a sample at this index value (zero padding means that all undefined samples are defined to be 0). There exists other more sophisticated methods, but that topic is beyond the scope of this article.

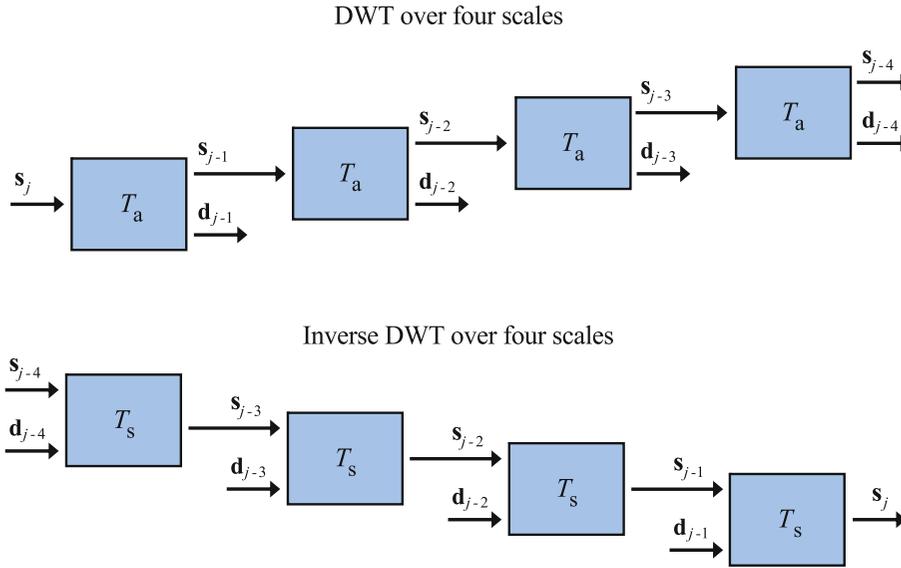
Finally, let us repeat our first example in the above notation. We also add a normalization step. In this form the transform is known as the Haar transform in the literature (we used this term previously, but in fact the scaling needs to be including for the transform to be the true Haar transform). The direct transform is

$$d_{j-1}^{(1)}[n] = s_j[2n+1] - s_j[2n] \quad (23)$$

$$s_{j-1}^{(1)}[n] = s_j[2n] + \frac{1}{2}d_{j-1}^{(1)}[n] \quad (24)$$

$$s_{j-1}[n] = \sqrt{2}s_{j-1}^{(1)}[n] \quad (25)$$

$$d_{j-1}[n] = \frac{1}{\sqrt{2}}d_{j-1}^{(1)}[n] \quad (26)$$



Wavelets and the Lifting Scheme, Figure 7  
DWT and inverse DWT over four scales

and the inverse transform is given by

$$d_{j-1}^{(1)}[n] = \sqrt{2}d_{j-1}[n] \tag{27}$$

$$s_{j-1}^{(1)}[n] = \frac{1}{\sqrt{2}}s_{j-1}[n] \tag{28}$$

$$s_j[2n] = s_{j-1}^{(1)}[n] - \frac{1}{2}d_{j-1}^{(1)}[n] \tag{29}$$

$$s_j[2n + 1] = s_j[2n] + d_{j-1}^{(1)}[n]. \tag{30}$$

We note that this transform can be applied to a signal of length  $2^j$  without using zero padding. It turns out to be essentially the only transform with this property.

**Discrete Wavelet Transform**

We have now established a method for transforming one signal into another signal that has a natural division into two parts of equal length. This transform is the discrete wavelet transform (DWT). In many cases one is not interested in the actual implementation of the transform, though, and the DWT is in those cases often shown as a box with one input and two outputs. The direct transform is sometimes called ‘analysis’, and we will represent such a direct transform by the symbol  $T_a$  (‘a’ for analysis).



In our notation the input would be  $s_j$  and the outputs would be  $s_{j-1}$  and  $d_{j-1}$ . The inverse transform is then nat-

urally shown as another box with two inputs and one output, and is sometimes called ‘synthesis’. Therefore, the inverse transform is represented by the symbol  $T_s$ .

The contents of the  $T_a$  box could be the direct Haar transform as given by (23)–(26), and the contents of the  $T_s$  box could be the inverse Haar transform as given by (27)–(30). Obviously, we must make sure to use the inverse transform corresponding to the applied direct transform. Otherwise, the results will be meaningless.

We can now combine these building blocks to get a multi-scale discrete wavelet transforms. We perform the transform over a certain number of scales  $k$ , meaning that we combine  $k$  of the building blocks as shown in Fig. 3 in the case of 2 scales, and with the upper-most diagram in Fig. 7 in the case of 4 scales. In the latter figure we use the building block representation of the individual steps.

It is possible to apply the DWT to the difference output of the transform as well, instead of just the average (or means) output. This is known as the wavelet packet transform, and is commonly found in the literature, and is often used in practical implementations. This concept introduces new possibilities and methods. We refer to other articles presenting this subject.

**Interpretation**

The main topic for this collection of articles is ‘Wavelets’. This term refers to a function, often denoted by  $\psi$  in the wavelet literature, that takes many shapes and determines the properties of the wavelet transform. So what is the

relation between a series of lifting steps and this wavelet function, and is it useful to know this relation, if one just wants to use the wavelet transform? After all, the lifting step method works quite well without any apparent need for knowledge of the wavelet function.

It turns out that while the transform can be implemented easily with just a few equations, the understanding of what is actually happening with a signal, when it is transformed, relies heavily on a proper interpretation of the transform. Here we will present an interpretation based on linear algebra, which will provide some useful insights. However, this presentation is incomplete; we state some results from the general theory, and illustrate them with explicit computations, but we do not discuss the general theory in detail, since this is beyond the scope of this article.

In the following discussion we will continue to use the Haar transform example from the Introduction, partly because it allows for simple and yet convincing computations, partly because we believe that the reader has become familiar with this example by now. We will end with an interpretation of the Daubechies 4 transform.

**Using Linear Algebra**

We saw above that the first of the eight numbers in the transformed signal in the first example in the Introduction, that is, the number 35, had a special meaning; it was the mean of the entire signal. This property (though not the number itself) was actually independent of the signal; for any signal that first number  $s_0[0]$  will always be the mean of the signal. In fact, it was a property derived from the transform, and it provided an interpretation of the number 35 in the transformed signal. Similar interpretations can be made of the other numbers in the transformed signal.

First, we examine the original example in Table 1 once more.

56	40	8	24	48	48	40	16
48	16	48	28	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>
32	38	<b>16</b>	<b>10</b>	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>
35	<b>-3</b>	<b>16</b>	<b>10</b>	<b>8</b>	<b>-8</b>	<b>0</b>	<b>12</b>

This table was constructed from top to bottom by multiple Haar transforms, and the ‘signal mean’ property was derived using equations, see for instance (9). However, we can reach the same conclusion by examining signal samples, and by starting at the bottom. To isolate the property of interest here, we start with a bottom row consisting of zeros at the entries we are not interested in interpreting, and a 1 at the entry we would like to interpret, that is, the

first of the eight numbers. Then we do a three scale synthesis of the signal and get

1	1	1	1	1	1	1	1
1	1	1	1	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
1	1	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
1	<b>0</b>						

Note that the normalization steps (25) and (26) has been omitted to make the table easier to read (this does not interfere with the interpretation, even though this fact may not be clear at this point). It is obvious that if we had started with any other number than 1, this number would then be the one repeated eight times in the top row after the three scale synthesis. This computation indicates that the entry at location 0 (recall that our indexing starts with 0) in the fourth row has an equal contribution from all eight original samples, and thus is the mean value.

Let us repeat this procedure with the remaining seven entries in the fourth row. Here are the first two, having a 1 at entry number 1 and 2.

1	1	1	1	-1	-1	-1	-1
1	1	-1	-1	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
1	-1	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
0	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>

1	1	-1	-1	0	0	0	0
1	-1	0	0	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
0	0	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
0	<b>0</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>

Examining the first row in the left-most table reveals that the value at entry number 1, i.e. the value 1 in the fourth row, can be interpreted as the mean of the first four sample minus the mean of the last four sample. In filter terms this is roughly equal to a band pass filtering just above DC, i.e. DC does not contribute to entry number 1, and neither would any frequency higher than a single cycle.

The result of all eight computations are now represented using notation from linear algebra; they are inserted as columns in a matrix

$$W_s^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{bmatrix}, \tag{31}$$

which is denoted by  $\mathbf{W}$  with a subscript  $s$  for synthesis for the following reason. A signal of length 8 is a vector in the vector space  $\mathbf{R}^8$ . The process described above of reconstructing the signal, whose transform is one of the canonical basis vectors in  $\mathbf{R}^8$ , is the same as finding the columns in the matrix (with respect to the canonical basis) of the three scale synthesis transform. This is just the well known definition from linear algebra of the matrix of a linear transform. This in turn means that applying this matrix to any length 8 signal performs the inverse Haar transform. For example, multiplying it with the fourth row of Table 1 (regarded as a column vector) produces the first row of that same table.

The matrix of the direct three scale transform is obtained in a similar fashion; by computing the transforms of the eight canonical basis vectors in  $\mathbf{R}^8$ . In other words, we start with the signal  $[1, 0, 0, 0, 0, 0, 0, 0]$  and carry out the three scale transform as shown here.

1	0	0	0	0	0	0	0
$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0	0	0
$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{2}$	0	0	0
$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	0	$\frac{1}{2}$	0	0	0

Computing the other seven transforms and inserting the resulting signals as column vectors gives

$$\mathbf{W}_a^{(3)} = \begin{bmatrix} \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}. \quad (32)$$

Multiplying the matrices we find  $\mathbf{W}_s^{(3)} \cdot \mathbf{W}_a^{(3)} = \mathbf{I}$  and  $\mathbf{W}_a^{(3)} \cdot \mathbf{W}_s^{(3)} = \mathbf{I}$ , where  $\mathbf{I}$  denotes the  $8 \times 8$  identity matrix. This is the linear algebra formulation of perfect reconstruction, or of the invertibility of the three scale transform.

Another interesting property to note here is that the structures of the two matrices are quite similar. The transpose of the one matrix bears a striking resemblance to the other matrix. In fact, if the above computations of the direct and inverse transforms of the canonical basis vectors had used the normalization steps, we would get an orthogonal matrix and its transpose, which by elementary linear algebra equals its inverse. Any wavelet transform with this property is called an orthogonal transform (i. e. all columns are

orthogonal to each other and all have norm 1). All transforms in the Daubechies family have this property, and so do other well-known families like Coiflets and Symlets. The CDF families are not orthogonal (but rather bi-orthogonal), and consequently the structure of the CDF matrices is slightly more complicated.

It is clear that analogous constructions can be carried out for signal of length  $2^j$  and transforms to all scales  $k = 1, \dots, j$ . The linear algebra point of view is useful in understanding the theory. However, if one is trying to carry out numerical computations, then it is a bad idea to use the matrix formulation. The direct  $k$  scale wavelet transform using the lifting steps requires a number of operations (additions, multiplications, etc.) on the computer, which is proportional to the length  $L$  of the signal. If we perform the transform using its matrix, then in general a number of operations proportional to  $L^2$  is needed. For longer signals this makes a huge difference for the computation time.

### From Transform to Wavelets

Let us now consider the column vectors in (31) as a sampling of continuous, piecewise constant signals on the interval  $[0; 1]$ . These continuous signals are shown in Fig. 8. In this figure the signals have been group according to their origin. Recall that the first column came from a three scale synthesis of  $[1, 0, 0, 0, 0, 0, 0, 0]$ . This is equivalent to letting  $\mathbf{s}_0 = 1$  and  $\mathbf{d}_0 = 0$ ,  $\mathbf{d}_1 = 0$ , and  $\mathbf{d}_2 = 0$  and using a three scale inverse DWT as shown in Fig. 7. The four right-most graphs in Fig. 8 can be obtained by inverse transforms where  $\mathbf{d}_2$  equals  $[1, 0, 0, 0]$ ,  $[0, 1, 0, 0]$ ,  $[0, 0, 1, 0]$ , and  $[0, 0, 0, 1]$ , respectively, and  $\mathbf{s}_0 = 0$ ,  $\mathbf{d}_0 = 0$ , and  $\mathbf{d}_1 = 0$  in all four cases.

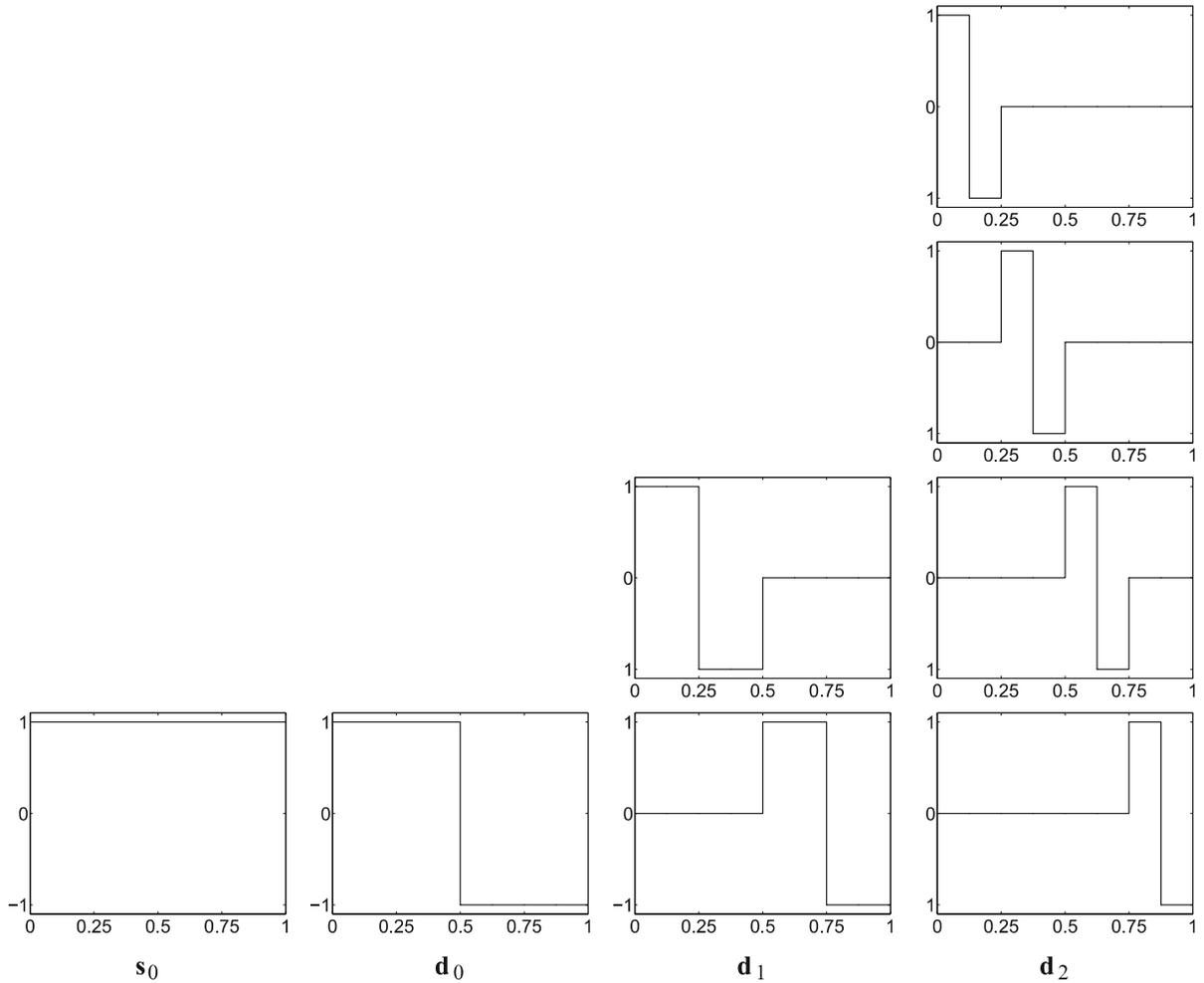
The  $\mathbf{s}$  group consists of only one signal, which is constant. The  $\mathbf{d}$  groups all consists of signals with the same shape (positive signal value followed by negative signal value). The three  $\mathbf{d}$  differ by a factor 2 scaling in the  $x$ -axis direction, and the signals the  $\mathbf{d}_1$  and  $\mathbf{d}_2$  groups differ by translation along the  $x$ -axis.

This pattern can be contained in a single equation. First, define the function

$$\psi(t) = \begin{cases} 1, & t \in [0; \frac{1}{2}[, \\ -1, & t \in [\frac{1}{2}; 1], \\ 0, & \text{otherwise.} \end{cases} \quad (33)$$

This is in fact the  $\mathbf{d}_0$  signal. Now all the other  $\mathbf{d}$  signal can be obtained by

$$\psi_{j,k}(t) = \psi(2^j t - k),$$



Wavelets and the Lifting Scheme, Figure 8

Graphs showing the 8 column vectors in (31) on the interval [0; 1]. The graphs are sorted according to the multiscale decomposition shown in Fig. 7

where  $j$  is equal to 0, 1, or 2 for the groups  $\mathbf{d}_0$ ,  $\mathbf{d}_1$ , and  $\mathbf{d}_2$ , and  $k$  ranges from 0 to  $2^j - 1$ . The  $\psi$  function is the wavelet. Sometimes it is called the mother wavelet to signify that a wavelet transform consists of numerous similar functions originating from one function. If we include the normalization step the equation reaches its final form

$$\psi_{j,k}(t) = 2^{-j/2} \psi(2^j t - k).$$

This framework does not include the  $\mathbf{s}_0$ . However, another equation combines  $\mathbf{s}_0$  and  $\mathbf{d}_0$ . We noticed previously that while the vector  $[1, 1, 1, 1, 1, 1, 1, 1]$ , the first column in (31), represents the mean of the original signal, the vector  $[1, 1, 1, 1, -1, -1, -1, -1]$  is mean of the first half of the original signal minus the mean of the other half of the signal. This link can be expressed with continuous functions.

First, define

$$\phi(t) = \begin{cases} 1, & t \in [0; 1], \\ 0, & \text{otherwise,} \end{cases}$$

that is, the function corresponding to the continuous  $\mathbf{s}_0$  signal. This function is called the scaling function in wavelet literature. Now we can link  $\mathbf{s}_0$  and  $\mathbf{d}_0$  with

$$\psi(t) = \phi(2t) - \phi(2t - 1). \tag{34}$$

This equation is called the two-scale equation, and is found in virtually any literature presenting the wavelet theory. Its general form is

$$\psi(t) = \sum_n g_n \phi(2t - n), \tag{35}$$

and it is valid for all orthogonal wavelet transforms. For

the Haar transform we have  $g_0 = 1$  and  $g_1 = -1$  and  $g_n = 0$  for all indices other than 0 and 1. The  $g_n$  for the Daubechies 4 transform are

$$g_0 = -\frac{1 - \sqrt{3}}{4\sqrt{2}}, \quad g_1 = \frac{3 - \sqrt{3}}{4\sqrt{2}},$$

$$g_2 = -\frac{3 + \sqrt{3}}{4\sqrt{2}}, \quad g_3 = \frac{1 + \sqrt{3}}{4\sqrt{2}}.$$

There is also a two-scale equation that produces the scaling function.

$$\phi(t) = \sum_n h_n \phi(2t - n). \tag{36}$$

For orthogonal transforms  $\mathbf{h}$  is obtained from  $\mathbf{g}$  by  $h_n = (-1)^n g_{1-n}$ , that is, reverse sequence order with alternating change of sign and index shift by 1.

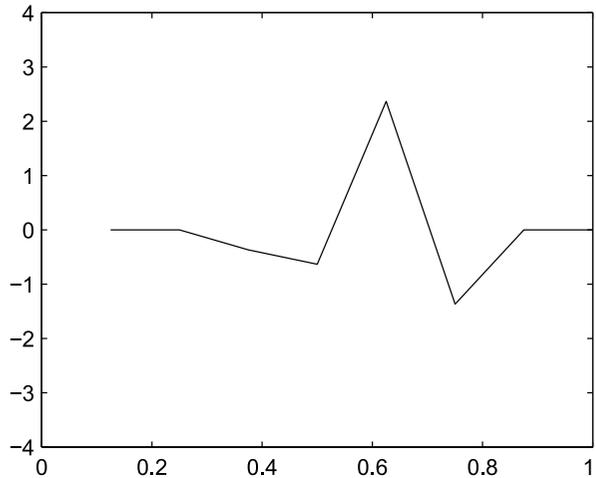
**Wavelet and Scaling Function for Daubechies 4**

The appearance of the Haar wavelet function through transform synthesis in the previous section may appear to be a special property of the simple Haar transform. However, the principle applies to all wavelets, and the method for generating the wavelet function can be used to obtain any wavelet, at least as a graph. Actually, in many cases no closed form of the wavelet exist. This includes the Daubechies family.

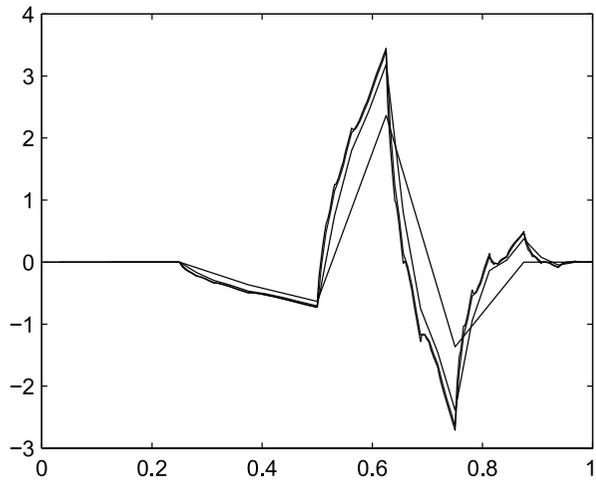
Thus, to see the shape of the Daubechies 4 wavelet we start again with an ‘all but one entry is zero’-signal. We saw that in fact when obtaining the wavelet graph it does not matter which entry is non-zero, as long as it is not the first entry. The only difference in the output is the dilation and translation of the resulting function. As an example we therefore choose  $[0, 0, 0, 0, 0, 1, 0, 0]$ . We perform a inverse three scale transform using the inverse Daubechies 4 Equations (18)–(22) and plot the result, see Fig. 8. Note that the problem with finite signals needs to be addressed in this inverse transform. For our purpose zero padding is sufficient.

This figure contains very little information. But let us now repeat the procedure for vectors of length 8, 32, 128, and 512, applied to a vector with a single 1 as its sixth entry. This requires inverse transforms over 3, 5, 7, and 9 scales, respectively. We fit each transform to the same interval, as if we have a finer and finer resolution. The result is shown in Fig. 9.

This figure shows that the graphs rapidly approach a limiting graph, as we increase the length of the vector. This is a result that can be established rigorously, but it is not easy to do so, and it is beyond the scope of this article.



Wavelets and the Lifting Scheme, Figure 9  
Inverse Daubechies 4 of  $[0, 0, 0, 0, 0, 1, 0, 0]$  over three scales

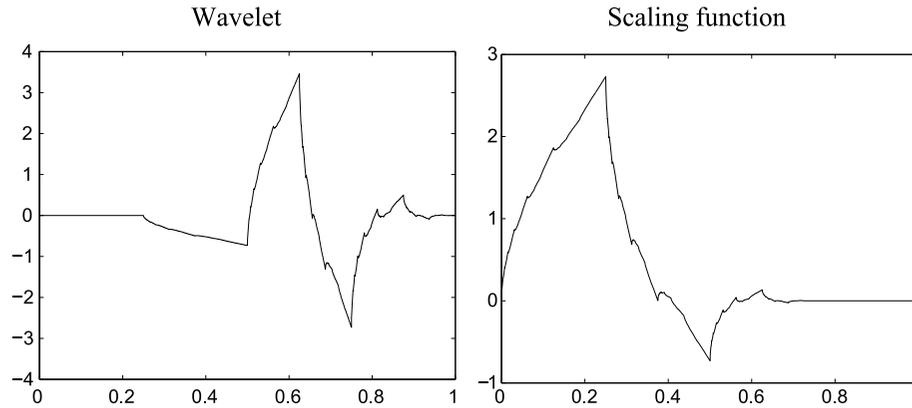


Wavelets and the Lifting Scheme, Figure 10  
Inverse Daubechies 4 of sixth basis vector, length 8, 32, 128 and 512

One can interpret the limiting function in Fig. 9 as a function whose values, sampled at appropriate points, represent the entries in the inverse transform of a vector of length  $2^N$ , with a single 1 as its sixth entry. For  $N$  just moderately large, say  $N = 12$ , this is a very good approximation to the actual value. See Fig. 10 for the result for  $N = 12$ , i. e. a vector of length 4096.

**Wavelet and Scaling Function for CDF(2,2)**

Finally, let us repeat the procedure for the CDF(2,2) transform in (10) and (11), and at the same time illustrate how the wavelet is translated depending on the placement of



Wavelets and the Lifting Scheme, Figure 11

Inverse Daubechies 4 of sixth and first basis vectors, both length 4096. The result is the Daubechies 4 wavelet on the *left* and the scaling function on the *right*

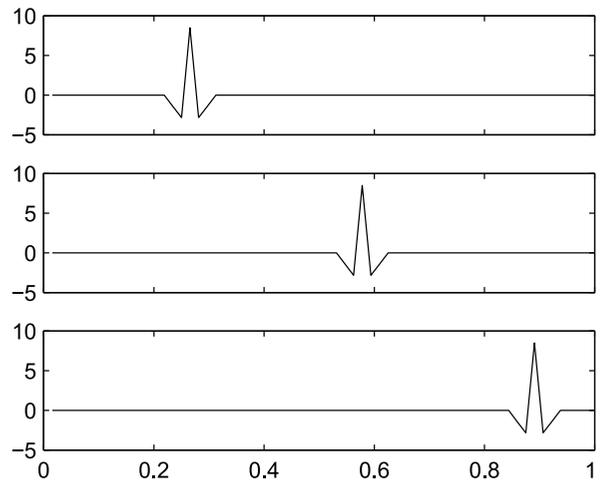
the 1 in the otherwise 0 vector. An example is given in Fig. 11. The difference in the graphs in Fig. 11 and Fig. 10 is striking. It reflects the result that the Daubechies 4 wavelet has very little regularity (it is not differentiable), whereas the CDF(2,2) wavelet is a piecewise linear function.

Recall that the Haar and Daubechies 4 transform are orthogonal transforms, which means that in linear algebra terms the transpose of the synthesis matrix equals the analysis matrix. A consequence of this is that the same wavelet and scaling function are used for analysis and for synthesis. This is not the case for CDF(2,2), which is a biorthogonal transform. Therefore, it has one synthesis pair of scaling function and wavelets, and an analysis pair. If we use the forward transform instead of the inverse we can obtain the analysis functions. They are shown in Fig. 12. These functions are quite complicated, and it is interesting to see that while the analysis wavelet and scaling function are very simple functions (we have not shown the scaling function of CDF(2,2) though) the inverse of that same transform have some rather complex wavelet and scaling functions.

As an end remark notice that all CDF(2,2) functions are symmetrical. This is a special property that can only be achieved by biorthogonal wavelets. Orthogonal wavelets can come close to symmetry, but they can never become completely symmetrical. An almost symmetrical family has been constructed, called the Symlets, see for example [4].

### The General Case

The above computations lead us to the conclusion that there are just two functions underlying the direct transform, and another two functions underlying the inverse



Wavelets and the Lifting Scheme, Figure 12

Inverse CDF(2,2) of three basis vectors of length 64, entry 40, or 50, or 60, equal to 1 and the remaining entries equal to zero. The result is the same function (the wavelet) with different translations

transform, in the sense that if we take sufficiently long vectors, say  $2^N$ , and perform a  $k$  scale transform, with  $k$  large, then we get values that are sampled values of one of the underlying functions. More precisely, inverse transforms of unit vectors with a one in places from 1 to  $2^{N-k}$  yield translated copies of the scaling function. Inverse transforms of unit vectors with a one in places from  $2^{N-k} + 1$  to  $2^{N-k+1}$  yield translated copies of the wavelet. Finally, inverse transforms of unit vectors with a one at places from  $2^{N-k+1} + 1$  to  $2^N$  yield scaled and translated copies of the wavelet.

These results are strictly correct only in a limiting sense, and they are not easy to establish. There is one fur-

the complication which we have omitted to state clearly. If one performs the procedure above with a 1 close to the start or end of the vector, then there will in general be some strange effects, depending on how the transform has been implemented. We refer to these as boundary effects. They depend on how one makes up for missing samples in computations near the start or end of a finite vector, the so-called boundary corrections. We have already mentioned zero padding as one of the correction methods. This is what we have used indirectly in plotting for example Fig. 10, where we have taken a vector with a one at place 24, and zeroes everywhere else.

Readers interested in a rigorous treatment of the interpretation of the transforms given here, and with the required mathematical background, are referred to the literature, for example the books by I. Daubechies [1], S. Mallat [4], and M. Vetterli-J. Kovačević [10,11]. Note that these books base their treatment of the wavelet transforms on the concepts of multiresolution analysis and filter theory rather than lifting.

### Lifting and Filter Banks

So far, the lifting technique has been carried out in the time domain; the correlation of samples was based on odd and even entries, interpretation was based on the signal in the time domain, and every reference to the signal was made by  $\mathbf{s}$  and  $\mathbf{d}$ , i. e. time representations of the signals. However, there is much to be learned about wavelets and about lifting when turning to the frequency domain.

The lifting method is a special case of a standard frequency-based signal processing method called filter banks. A filter bank is a series of bandpass filters which separates the input signal into a number of components, each with a distinct range of frequencies from the original signal. Often, in a filter bank, the filters are designed such that no information is lost, that is, it is possible to reconstruct the original signal from the components. The outputs of a filter bank are called subband signals, and it has as many subbands as there are filters in the filter bank.

The following introduction to lifting in the frequency domain requires some knowledge of the  $z$ -transform as well as basic Fourier theory. To ease the reading, we briefly state some important relations for the  $z$ -transform. The transform maps a signal  $\mathbf{x} = \{x[n]\}$  to a function defined in the complex plane

$$X(z) = \sum_{n \in \mathbb{Z}} x[n]z^{-n} .$$

This is equivalent to the Fourier transform, when  $z = e^{j\omega}$ , and we use capital  $X$  to denote the  $z$ -transform of  $\mathbf{x}$ . In the

following exposition we will need up and down sampling by two of a signal. In the time domain this is accomplished by inserted zeros between all samples and by removing every other sample, respectively. Explicitly, if  $\mathbf{x} = \{x[n]\}$ , then the down sampled sequence  $\mathbf{x}_{2\downarrow}$  is given by  $x_{2\downarrow}[n] = x[2n]$ . The up sampled sequence  $\mathbf{x}_{2\uparrow}$  is given by  $x_{2\uparrow}[n] = x[n/2]$ , if  $n$  is even, and  $x_{2\uparrow}[n] = 0$ , if  $n$  is odd.

In the  $z$ -transform up sampling of  $X(z)$  is accomplished by  $X(z^2)$ , and down sampling is accomplished by

$$\frac{1}{2}(X(z^{1/2}) + X(-z^{1/2})) . \quad (37)$$

Finally, we note that in this section all signals have finite length, and thus the  $z$ -transform has a finite number of non-zero terms. The  $z$ -transform is therefore a Laurent polynomial, which is a polynomial in the variables  $z$  and  $z^{-1}$ .

### Lifting in the $z$ -Transform Representation

The lifting method consists of a three basic operations; it is composed of splitting the signal in odd and even samples, predicting an odd sample, and updating an even sample. This is shown in Fig. 3. The splitting is given in the  $z$ -transform representation by

$$X(z) = X_0(z^2) + z^{-1}X_1(z^2) , \quad (38)$$

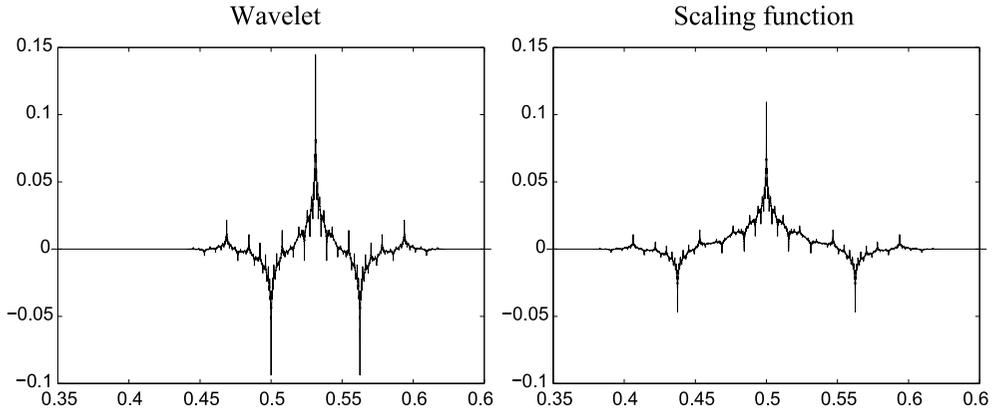
where

$$X_0(z) = \sum_n x[2n]z^{-n} = \frac{1}{2}(X(z^{1/2}) + X(-z^{1/2})) , \quad (39)$$

$$\begin{aligned} X_1(z) &= \sum_n x[2n+1]z^{-n} \\ &= \frac{1}{2}z^{1/2}(X(z^{1/2}) - X(-z^{1/2})) , \quad (40) \end{aligned}$$

are the  $z$ -transforms of the even and odd samples. Note that the second equality in both formulas comes from (37). We represent this decomposition by the left side of the diagram in Fig. 13. In the diagram the time shift is inserted prior to the down sampling, rather than after as in (40), since this gives multiplication with  $z$  instead of  $z^{1/2}$ . The decomposition in (38) is called a polyphase decomposition (with two components).

The inverse operation is obtained by reading Eq. (38) from right to left. The equation tells us that we can obtain  $X(z)$  from  $X_0(z)$  and  $X_1(z)$  by first up sampling the two components by 2, then shifting  $X_1(z^2)$  one time unit right (by multiplication by  $z^{-1}$ ), and finally adding the two components. We represent this reconstruction by the



Wavelets and the Lifting Scheme, Figure 13  
Wavelet and Scaling function for CDF(2,2)

right hand side of the diagram in Fig. 13. Notice that as with the lifting method this inversion simply ‘undoes’ the changes made to the signal. The transform pair represented in Fig. 13 is sometimes in the literature called the ‘lazy’ wavelet transform and its inverse.

**Prediction and Update Steps** Let us now see how we can implement the prediction step from Sect. “Lifting” in the  $z$ -transform representation. The prediction technique was to form a linear combination of the even entries and then subtract the result from the odd entry under consideration. The linear combination was formed independently of the index of the odd sample under consideration, and based only on the relative location of the even entries. For example, in the CDF(2,2) transform the first step in (10) can be implemented as  $X_1(z) - T(z)X_0(z)$ , where  $T(z) = \frac{1}{2}(1+z)$ . This is because  $T(z)X_0(z)$  is the convolution  $\mathbf{t} * \mathbf{x}_0$  in the time domain, which is exactly a linear combination of the even entries with weights  $t_n$ . Explicitly,

$$\begin{aligned} X_1(z) - T(z)X_0(z) &= X_1(z) - \frac{1}{2}(1+z) \sum_n x[2n]z^{-n} \\ &= X_1(z) - \frac{1}{2} \sum_n x[2n]z^{-n} + \frac{1}{2} \sum_n x[2n]z^{-n+1} \\ &= \sum_n x[2n+1] - \sum_n \frac{1}{2}(x[2n] + x[2n+2])z^{-n} \\ &= \sum_n (x[2n+1] - \frac{1}{2}(x[2n] + x[2n+2])) z^{-n}, \end{aligned}$$

which is the  $z$ -transform representation of the right hand side of (10). The transition is described by matrix multiplication as in

$$\begin{bmatrix} X_0(z) \\ X_1(z) - T(z)X_0(z) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -T(z) & 1 \end{bmatrix} \begin{bmatrix} X_0(z) \\ X_1(z) \end{bmatrix}.$$

Here we use  $2 \times 2$  matrices whose entries are Laurent polynomials. An entirely analogous computation shows that if we define  $S(z) = \frac{1}{4}(1+z^{-1})$ , then the update step in (11) is implemented in the  $z$ -transform representation as multiplication by the matrix

$$\begin{bmatrix} 1 & S(z) \\ 0 & 1 \end{bmatrix}.$$

The normalization step, as for example given in (26) and (25), can be implemented by multiplication by a matrix of the form

$$\begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix},$$

where  $K > 0$  is a constant.

**Entire Transform** Since every prediction step is subtraction of a linear combination of odd samples, any prediction step can be implemented by means of multiplication by a matrix of the form

$$\mathbf{P}(z) = \begin{bmatrix} 1 & 0 \\ -T(z) & 1 \end{bmatrix}. \quad (41)$$

The same result applies to any update step, which can be implemented by multiplication by a matrix of the form

$$\mathbf{U}(z) = \begin{bmatrix} 1 & S(z) \\ 0 & 1 \end{bmatrix}. \quad (42)$$

Here  $T(z)$  and  $S(z)$  are both Laurent polynomials.

The general one scale DWT is a combination of multiple prediction and updates steps and one normalization step. This is shown in Fig. 6 (although the normalization

is not included in the figure). In the  $z$ -transform representation this becomes

$$\mathbf{H}(z) = \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & S_N(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -T_N(z) & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & S_1(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -T_1(z) & 1 \end{bmatrix}. \quad (43)$$

The order of the factors is determined by the order in which the steps are applied. First a prediction step, then an update step, repeated  $N$  times, and then finally the normalization step.

An important property of the DWT implemented via lifting steps was the invertibility of the transform, as illustrated for example in Fig. 4. It is easy to verify that we have

$$\mathbf{P}(z)^{-1} = \begin{bmatrix} 1 & 0 \\ T(z) & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{U}(z)^{-1} = \begin{bmatrix} 1 & -S(z) \\ 0 & 1 \end{bmatrix}. \quad (44)$$

We note that the inverse matrix is again a matrix with Laurent polynomials as entries. Thus, the inversion of the transform is accomplished simply by inverting each  $2 \times 2$  matrix and reversing their order.

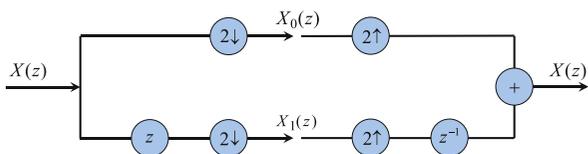
$$\mathbf{H}^{-1}(z) = \mathbf{G}(z) = \begin{bmatrix} 1 & 0 \\ T_1(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_1(z) \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 \\ T_N(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_N(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} K^{-1} & 0 \\ 0 & K \end{bmatrix}. \quad (45)$$

Multiplying out the matrices in the product defining  $\mathbf{H}(z)$  in (43), we get a  $2 \times 2$  matrix with entries, which are Laurent polynomials. We use the notation

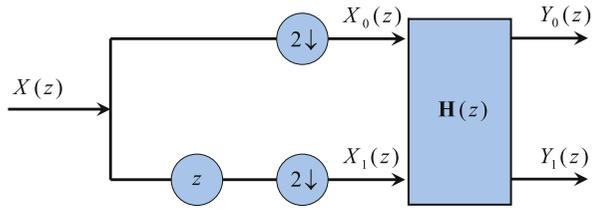
$$\mathbf{H}(z) = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix}, \quad (46)$$

for such a general matrix. Written in matrix notation the DWT is given as

$$\begin{bmatrix} Y_0(z) \\ Y_1(z) \end{bmatrix} = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} \begin{bmatrix} X_0(z) \\ X_1(z) \end{bmatrix}, \quad (47)$$



**Wavelets and the Lifting Scheme, Figure 14**  
 Splitting in even and odd components by means of up sampling followed by reconstruction of the signal from the odd and even components



**Wavelets and the Lifting Scheme, Figure 15**  
 One scale DWT in the  $z$ -representation as given in (47). This is the polyphase representation and it comes directly from the lifting step method

and we can then represent the implementation of the complete one scale DWT in the  $z$ -transform representation by the diagram in Fig. 15.

This representation of a two channel filter bank, without any reference to lifting, is in the signal analysis literature called the polyphase representation, see for example [7,10,11].

**Two Channel Filter Banks with Perfect Reconstruction**

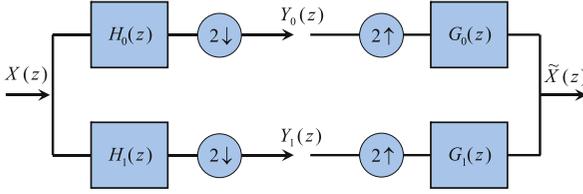
We have now seen how the time-based lifting method has an equivalent representation in the frequency domain. Lifting allows for perfect reconstruction of the signal after transformation, and we will now use this property (which is preserved in the frequency representation) to better understand the link between lifting and filter banks. We assume the reader is familiar with the concept of filtering and filter banks. Details on filtering can be found in any book on signal analysis, for example in [6,7]. The filter bank approach to wavelets is the one used in most introductions to the subject.

A two channel filter bank starts with two analysis filters, here denoted by the filter taps  $\mathbf{h}_0$  and  $\mathbf{h}_1$ , and two synthesis filters, denoted by  $\mathbf{g}_0$  and  $\mathbf{g}_1$ . Usually the filters with index 0 are chosen to be low pass filters, and the filters with index 1 to be high pass filters. The analysis and synthesis parts of the filter bank are shown in Fig. 15.

The analysis part transforms the input  $X(z)$  to the output pair  $Y_0(z)$  and  $Y_1(z)$ . The synthesis part then transforms this pair to the output  $\tilde{X}(z)$ . The filtering scheme is said to have the perfect reconstruction property, if  $X(z) = \tilde{X}(z)$  for any  $X(z)$ .

The main point that we will now establish is that the two Figs. 15 and 16 do indeed show the same thing, and that it is possible to get the one from the other.

**Conditions for Perfect Reconstruction** To do this we first analyze which conditions are needed on the four filters in Fig. 15 in order to obtain the perfect reconstruction



Wavelets and the Lifting Scheme, Figure 16

Two channel analysis and synthesis. This is the standard filter bank representation

property. Filtering by  $\mathbf{h}_0$  transforms  $X(z)$  to  $H_0(z)X(z)$ , and we then use (38) to down sample by two. Thus we have

$$Y_0(z) = \frac{1}{2} \left( H_0(z^{1/2})X(z^{1/2}) + H_0(-z^{1/2})X(-z^{1/2}) \right), \quad (48)$$

$$Y_1(z) = \frac{1}{2} \left( H_1(z^{1/2})X(z^{1/2}) + H_1(-z^{1/2})X(-z^{1/2}) \right). \quad (49)$$

Up sampling by two followed by filtering by the  $G$ -filters, and addition of the results leads to a reconstructed signal

$$\tilde{X}(z) = G_0(z)Y_0(z^2) + G_1(z)Y_1(z^2). \quad (50)$$

Perfect reconstruction means that  $\tilde{X}(z) = X(z)$ . We combine the above expressions and then regroup terms to get

$$\begin{aligned} \tilde{X}(z) = & \frac{1}{2} [G_0(z)H_0(z) + G_1(z)H_1(z)]X(z) \\ & + \frac{1}{2} [G_0(z)H_0(-z) + G_1(z)H_1(-z)]X(-z). \end{aligned}$$

To fulfill the condition  $\tilde{X}(z) = X(z)$  we need

$$G_0(z)H_0(z) + G_1(z)H_1(z) = 2, \quad (51)$$

$$G_0(z)H_0(-z) + G_1(z)H_1(-z) = 0. \quad (52)$$

These conditions mean that the four filters cannot be chosen independently, if we want to have perfect reconstruction. To determine what conditions these equations impose on  $H$  and  $G$  it is convenient to write them as a matrix equation,

$$\begin{bmatrix} H_0(z) & H_1(z) \\ H_0(-z) & H_1(-z) \end{bmatrix} \begin{bmatrix} G_0(z) \\ G_1(z) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}. \quad (53)$$

We want to solve this equation, that is, determine  $G_0(z)$  and  $G_1(z)$  as functions of  $H_0(z)$  and  $H_1(z)$ . Since the matrix is invertible (it performs a DWT that can be inverted), we can use Cramer's rule to get

$$G_0(z) = \frac{\begin{vmatrix} 2 & H_1(z) \\ 0 & H_1(-z) \end{vmatrix}}{d(z)} = 2d(z)^{-1}H_1(-z).$$

The determinant  $d(z)$  of an invertible  $2 \times 2$  matrix with Laurent polynomials as entries is a monomial (see [3]),

which in this case has the property that  $d(-z) = -d(z)$ . Therefore, the determinant can be assumed to be of the form  $d(z) = \frac{1}{2}C^{-1}z^{-2k-1}$  for some integer  $k$  and some real constant  $C$ . Thus,

$$G_0(z) = \frac{\begin{vmatrix} 2 & H_1(z) \\ 0 & H_1(-z) \end{vmatrix}}{\frac{1}{2}C^{-1}z^{-2k-1}} = Cz^{2k+1}H_1(-z), \quad (54)$$

$$G_1(z) = \frac{\begin{vmatrix} H_0(z) & 2 \\ H_0(-z) & 0 \end{vmatrix}}{\frac{1}{2}C^{-1}z^{-2k-1}} = -Cz^{2k+1}H_0(-z). \quad (55)$$

These equations show that we can choose either the  $H$ -filter pair or the  $G$ -filter pair. We will assume that we have filters  $H_0$  and  $H_1$ , subject to the condition that

$$\begin{aligned} d(z) = H_0(z)H_1(-z) - H_0(-z)H_1(z) \\ = \frac{1}{2}C^{-1}z^{-2k-1} \end{aligned} \quad (56)$$

for some integer  $k$  and nonzero constant  $C$ . Then  $G_0$  and  $G_1$  are determined by the Eqs. (54) and (55), which implies that they are unique up to a scaling factor and an odd shift in time. Note that this argument is valid for any two channel filter bank, subject to the condition (56).

### Lifting and Two Channel Filter Banks with Perfect Reconstruction Are Equivalent

Many presentations of the wavelet transform start with a two channel filter bank with the perfect reconstruction property. The analysis part is then used to define the direct one scale DWT, and the synthesis part is used for reconstruction. It is an important result that the filtering approach, and the one based on lifting, actually are identical. Thus any set of lifting steps leads to a two channel filter bank with perfect reconstruction, and, quite remarkably, vice versa. This means that they are just two different ways of describing the same transformation from  $X(z)$  to  $Y_0(z)$  and  $Y_1(z)$ . This equivalence will be the subject of the remainder of the article.

The first step is to show that the analysis step in Fig. 15 coming from the two channel filter bank is equivalent to the analysis step summarized in Fig. 14 and in (47), i. e. coming from the lifting method. Thus, we want to find the equations relating the coefficients in the  $\mathbf{H}$  matrix (46) and the filters  $H_0$ ,  $H_1$ ,  $G_0$ , and  $G_1$ . The analysis step by both methods should yield the same result. To avoid the square root terms we compare the results after up sampling by two. We start with the equality

$$\begin{bmatrix} Y_0(z^2) \\ Y_1(z^2) \end{bmatrix} = \mathbf{H}(z^2) \begin{bmatrix} \frac{1}{2}(X(z) + X(-z)) \\ \frac{1}{2}z(X(z) - X(-z)) \end{bmatrix},$$

where the left hand side is from the filter bank approach (48) and (49), and the right hand side from the lifting approach with polyphase matrix (46) and up sampled  $X_0(z)$  and  $X_1(z)$ . The first equation can then be written as

$$\begin{aligned} \frac{1}{2}(H_0(z)X(z) + H_0(-z)X(-z)) = \\ H_{00}(z^2)\frac{1}{2}(X(z)+X(-z)) + H_{01}(z^2)\frac{1}{2}z(X(z)-X(-z)). \end{aligned}$$

This leads to the relation

$$H_0(z) = H_{00}(z^2) + zH_{01}(z^2).$$

The relation for  $H_1$  is found analogously, and then the relations for  $G_0$  and  $G_1$  can be found using the perfect reconstruction conditions (54) and (55) in the two cases. The results are summarized here.

$$H_0(z) = H_{00}(z^2) + zH_{01}(z^2), \quad (57)$$

$$H_1(z) = H_{10}(z^2) + zH_{11}(z^2), \quad (58)$$

$$G_0(z) = G_{00}(z^2) + z^{-1}G_{01}(z^2), \quad (59)$$

$$G_1(z) = G_{10}(z^2) + z^{-1}G_{11}(z^2). \quad (60)$$

Note the difference in the decomposition of the  $H$ -filters and the  $G$ -filters. Thus in the polyphase representation we use

$$\begin{aligned} \mathbf{H}(z) &= \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix}, \\ \mathbf{G}(z) &= \mathbf{H}(z)^{-1} = \begin{bmatrix} G_{00}(z) & G_{10}(z) \\ G_{01}(z) & G_{11}(z) \end{bmatrix}. \end{aligned}$$

Note the placement of entries in  $\mathbf{G}(z)$ , which differs from the usual notation for matrices. The requirement of perfect reconstruction in the polyphase formulation was the requirement that  $\mathbf{G}(z)$  should be the inverse of  $\mathbf{H}(z)$ .

**From Filters to Lifting Steps** It is easy to start with a filter bank and derive  $\mathbf{H}(z)$  using (57) and (58). However, going the other way is somewhat more tricky; given  $\mathbf{H}(z)$  it is necessary to factorize it into a number of  $2 \times 2$  matrices of a particular structure to obtain the lifting steps. The remarkable result is that this is always possible. This result was obtained by I. Daubechies and W. Sweldens in 1998 in the paper [2]. The factorization result for  $2 \times 2$  matrices with Laurent entries was previously known in the mathematical literature. The importance of the paper by I. Daubechies and W. Sweldens lies in its impact on signal processing.

**Theorem 1(Daubechies and Sweldens 1998)** *Assume that  $\mathbf{H}(z)$  is a  $2 \times 2$  matrix of Laurent polynomials, normalized to  $\det \mathbf{H}(z) = 1$ . Then there exists a constant  $K \neq 0$  and Laurent polynomials  $S_1(z), \dots, S_N(z), T_1(z), \dots, T_N(z)$ , such that*

$$\begin{aligned} \mathbf{H}(z) &= \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & S_N(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ T_N(z) & 1 \end{bmatrix} \cdots \\ &\begin{bmatrix} 1 & S_1(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ T_1(z) & 1 \end{bmatrix}. \quad (61) \end{aligned}$$

The normalization  $\det \mathbf{H}(z) = 1$  in the theorem can always be satisfied. As seen above, in the general case  $\det \mathbf{H}(z) = Cz^{2k+1}$ . One can always get the determinant equal to one by scaling and an odd shift in time. Therefore the result is stated with  $\det \mathbf{H}(z) = 1$  without loss of generality.

The proof of this theorem is constructive. It gives an algorithm for finding the Laurent polynomials  $S_1(z), \dots, S_N(z), T_1(z), \dots, T_N(z)$  in the factorization. It is important to note that the factorization is not unique. Once we have a factorization, we can translate it into lifting steps.

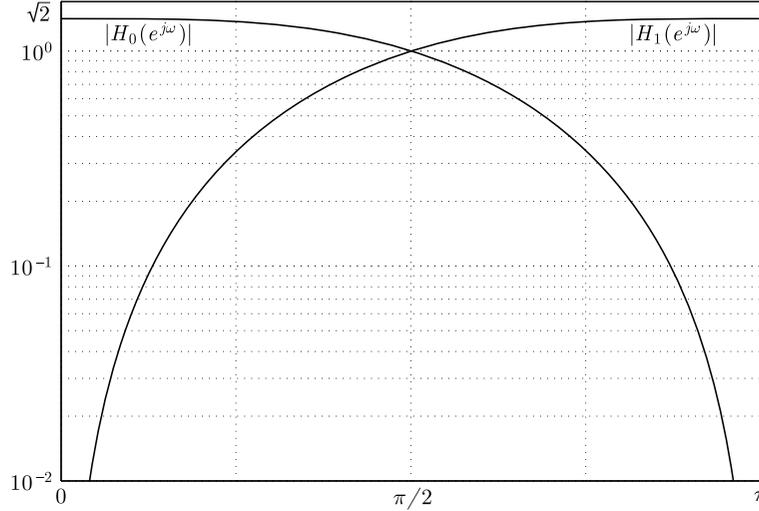
The advantage of the lifting approach, compared to the filter approach, is that it is very easy to find perfect reconstruction filters  $H_0, H_1, G_0$ , and  $G_1$ . It is just a matter of multiplying the lifting steps as in (43), and then assemble the filters according to the Eqs. (57)–(60). There can also be algorithmic advantages in implementing a transform using lifting steps, since it may result in fewer arithmetic operations than a filter implementation.

This approach should be contrasted with the traditional signal analysis approach, where one tries to find (approximate numerical) solutions to the Eqs. (51) and (52), using for example spectral factorization. The weakness in constructing a transform based solely on the lifting technique is that it is based entirely on considerations in the time domain. Sometimes it is desirable to design filters with certain properties in the frequency domain, and once filters have been constructed in the frequency domain, we can use the constructive proof of the theorem to derive a lifting implementation.

We should mention that the numerical stability of transforms designed using lifting can be difficult to analyze.

### Examples of Lifting Steps in the Frequency Domain

We will now see how the Daubechies 4 transform looks in the frequency domain, that is, what frequency response we would get when passing a signal through the Daubechies



**Wavelets and the Lifting Scheme, Figure 17**  
Frequency response of the Daubechies 4 lifting steps

4 transform. First, we need to find  $\mathbf{H}(z)$  by multiplying the lifting steps originating in the Daubechies 4 Eqs. (13) through (17). Each equation has an equivalent lifting step matrix (41) or (42). The first of the Daubechies 4 equations

$$s_{j-1}^{(1)}[n] = s_j[2n] + \sqrt{3}s_j[2n+1]$$

is an update step, and it simply multiplies the odd sample with  $\sqrt{3}$  and add the result to the even sample. In the  $z$ -transform this is achieved by multiplying with

$$\begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix},$$

which makes this matrix the first transform step. The second equation

$$d_{j-1}^{(1)}[n] = s_j[2n+1] - \frac{1}{4}\sqrt{3}s_{j-1}^{(1)}[n] - \frac{1}{4}(\sqrt{3}-2)s_{j-1}^{(1)}[n-1]$$

is a prediction step that uses the present value (index  $n$ ) and the first previous value (index  $n-1$ ) of the signal. This is achieved in the  $z$ -transform by

$$\begin{bmatrix} 1 & & 0 \\ -\frac{\sqrt{3}}{4} & -\frac{\sqrt{3}-2}{4}z^{-1} & 1 \end{bmatrix}.$$

The third equation converts to  $z$ -transform similarly. The two scaling Equation (16) and (17) are joined in one scaling matrix

$$\begin{bmatrix} \frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix}.$$

Now, the polyphase matrix  $\mathbf{H}(z)$  can be written

$$\begin{aligned} \mathbf{H}(z) &= \begin{bmatrix} \frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{3}}{4} & -\frac{\sqrt{3}-2}{4}z^{-1} \end{bmatrix} \begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{3-\sqrt{3}}{4\sqrt{2}}z + \frac{1+\sqrt{3}}{4\sqrt{2}} & \frac{1-\sqrt{3}}{4\sqrt{2}}z + \frac{3+\sqrt{3}}{4\sqrt{2}} \\ -\frac{3+\sqrt{3}}{4\sqrt{2}} & -\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-1} + \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} \end{bmatrix} \\ &= \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} \end{aligned}$$

Thus, from (57) and (58) it follows that the low and high pass filter taps (in  $z$ -transform) are given by

$$\begin{aligned} H_0(z) &= H_{00}(z^2) + zH_{01}(z^2) \\ &= \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3+\sqrt{3}}{4\sqrt{2}}z + \frac{3-\sqrt{3}}{4\sqrt{2}}z^2 + \frac{1-\sqrt{3}}{4\sqrt{2}}z^3, \\ H_1(z) &= H_{10}(z^2) + zH_{11}(z^2) \\ &= -\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-2} + \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} - \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1+\sqrt{3}}{4\sqrt{2}}z. \end{aligned}$$

To determine the frequency response of these transfer functions, we simply replace  $z$  with  $e^{j\omega}$ , which gives the Fourier transform of the filter taps. We then plot  $|H_0(e^{j\omega})|$  and  $|H_1(e^{j\omega})|$  as function of  $\omega$  to show the amplitude characteristics of the transfer functions  $H_0(z)$  and  $H_1(z)$  on the unit circle in the complex plane. This plot is shown in Fig. 16.

### Making Lifting Steps from Filters

There are basically three ways of representing the building block in a DWT: (1) The transform can be represented by a pair of filters (usually low pass and high pass filters) satisfying the perfect reconstruction conditions, or it can be given as lifting steps, which are either given (2) in the time domain as a set of equations, or (3) in the frequency domain as a factored matrix of Laurent polynomials.

As an example the Daubechies 4 transform can be given in these three forms, as shown below.

#### Equation form:

$$s^{(1)}[n] = S[2n] + \sqrt{3}S[2n + 1], \quad (62)$$

$$d^{(1)}[n] = S[2n + 1] - \frac{1}{4}\sqrt{3}s^{(1)}[n] - \frac{1}{4}(\sqrt{3} - 2)s^{(1)}[n - 1], \quad (63)$$

$$s^{(2)}[n] = s^{(1)}[n] - d^{(1)}[n + 1], \quad (64)$$

$$\tilde{s}[n] = \frac{\sqrt{3} - 1}{\sqrt{2}}s^{(2)}[n], \quad (65)$$

$$\tilde{d}[n] = \frac{\sqrt{3} + 1}{\sqrt{2}}d^{(1)}[n]. \quad (66)$$

#### Matrix form:

$$\mathbf{H}(z) = \begin{bmatrix} \frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix}. \quad (67)$$

#### Filter form:

$$\mathbf{h} = \frac{1}{4\sqrt{2}} [1 + \sqrt{3}, 3 + \sqrt{3}, 3 - \sqrt{3}, 1 - \sqrt{3}], \quad (68)$$

$$\mathbf{g} = \frac{1}{4\sqrt{2}} [1 - \sqrt{3}, -3 + \sqrt{3}, 3 + \sqrt{3}, -1 - \sqrt{3}]. \quad (69)$$

The equation form was presented in (13) through (17). Converting this to the matrix form has not been shown directly. However, using (41), (42), and (43) it is not difficult to derive the matrices from the equations. It is not difficult either to obtain the filter taps  $\mathbf{h}$  and  $\mathbf{g}$  from the matrix form. Simply multiply the matrix to get the polyphase matrix (46), and the odd entries in  $\mathbf{h}$  will be the coefficient in  $H_{00}(z)$ , and the even entries the coefficients of  $H_{01}(z)$ . Likewise, the odd and even entries of  $\mathbf{g}$  is the coefficients of  $H_{10}(z)$  and  $H_{11}(z)$ .

Now, the hard part is to go in the other direction. Suppose we have the filter taps and want to find the lifting steps (in either time or frequency domain). That it is theoretical possible to obtain the lifting steps was made evident with Theorem 1. Now we want to show how to actually obtain the lifting steps.

### The Euclidean Algorithm

The Euclidean algorithm is usually first presented as an algorithm for finding the greatest common divisor of two integers. But it can be applied to many other analogous problems. One application is to finding the greatest common divisor of two Laurent polynomials. This turns out to be the key step in factorizing a polyphase matrix.

Take two Laurent polynomials  $a(z)$  and  $b(z) \neq 0$  with  $|a(z)| \geq |b(z)|$ . Then there always exist a Laurent polynomial  $q(z)$ , the quotient, with  $|q(z)| = |a(z)| - |b(z)|$ , and a Laurent polynomial  $r(z)$ , the remainder, with  $|r(z)| < |b(z)|$ , such that

$$a(z) = b(z)q(z) + r(z). \quad (70)$$

We use the notation  $q(z) = a(z)/b(z)$  as 'division disregarding the remainder'. Now, let  $a_0(z) = a(z)$  and  $b_0(z) = b(z)$ , and iterate the following steps starting from  $n = 0$

$$a_{n+1}(z) = b_n(z),$$

$$b_{n+1}(z) = a_n(z) - q_{n+1}(z)b_n(z)$$

which in matrix form is

$$\begin{bmatrix} a_{n+1}(z) \\ b_{n+1}(z) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -q_{n+1}(z) \end{bmatrix} \begin{bmatrix} a_n(z) \\ b_n(z) \end{bmatrix},$$

where  $q_{n+1}(z) = a_n(z)/b_n(z)$ . Let  $N$  denote the smallest integer with  $N \leq |b(z)| + 1$ , for which  $b_N(z) = 0$ . Then  $a_N(z)$  is a greatest common divisor for  $a(z)$  and  $b(z)$ . In short, for two Laurent polynomials  $a(z)$  and  $b(z)$  with the given constraints we have

$$\begin{bmatrix} a_N(z) \\ 0 \end{bmatrix} = \prod_{n=N}^1 \begin{bmatrix} 0 & 1 \\ 1 & -q_n(z) \end{bmatrix} \begin{bmatrix} a(z) \\ b(z) \end{bmatrix}.$$

We note that there is no unique greatest common divisor for  $a(z)$  and  $b(z)$ , since if  $d(z)$  divides both  $a(z)$  and  $b(z)$ , and is of maximal degree, then  $\alpha z^m d(z)$  is also a divisor of the same degree.

Note the order of the terms in this product, as given by the limits in the product. Inverting the matrices and moving them to the other side, we get (products in reverse order, as shown by the limits)

$$\begin{bmatrix} a(z) \\ b(z) \end{bmatrix} = \prod_{n=1}^N \begin{bmatrix} q_n(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_N(z) \\ 0 \end{bmatrix}. \quad (71)$$

Now, take  $a(z) = H_{00}(z)$  and  $b(z) = H_{01}(z)$  from the polyphase matrix (46). We then get

$$\begin{bmatrix} H_{00}(z) \\ H_{01}(z) \end{bmatrix} = \prod_{n=1}^N \begin{bmatrix} q_n(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} K \\ 0 \end{bmatrix}. \tag{72}$$

The greatest common divisor  $a_N(z)$  of  $H_{00}(z)$  and  $H_{01}(z)$  is a constant. This is because of the determinant assumption

$$H_{00}(z)H_{11}(z) - H_{01}(z)H_{10}(z) = 1,$$

since  $a_N(z)$  divides the left hand side, it also divides the right hand side. Thus,  $a_N(z) = Kz^\ell$ . By shifting the indices in the  $z$ -transform of  $H_0(z)$  with  $\ell$  steps, the greatest common divisor becomes a constant.

Observing that

$$\begin{aligned} \begin{bmatrix} q(z) & 1 \\ 1 & 0 \end{bmatrix} &= \begin{bmatrix} 1 & q(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q(z) & 1 \end{bmatrix}, \end{aligned}$$

and replacing

$$\begin{bmatrix} K \\ 0 \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix},$$

and letting  $M = \lceil N/2 \rceil$ , we can rewrite (72) to

$$\begin{aligned} &\begin{bmatrix} H_{00}(z) & H'_{10}(z) \\ H_{01}(z) & H'_{11}(z) \end{bmatrix} \\ &= \prod_{n=1}^M \begin{bmatrix} 1 & q_{2n-1}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_{2n}(z) & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix}, \end{aligned}$$

where these equations define  $H'_{10}(z)$  and  $H'_{11}(z)$ . If  $N$  is odd, let  $q_{2M}(z) = 0$ . Transposing both sides yields

$$\begin{aligned} &\begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H'_{10}(z) & H'_{11}(z) \end{bmatrix} \\ &= \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \prod_{n=M}^1 \begin{bmatrix} 1 & q_{2n}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_{2n-1}(z) & 1 \end{bmatrix}. \tag{73} \end{aligned}$$

which clearly resembles the factorization in Theorem 1, which we are trying to achieve. All we need to do now is to find out how  $H'_{10}(z)$  and  $H'_{11}(z)$  are connected with  $H_{10}(z)$  and  $H_{11}(z)$ . By using the fact that both polyphase matrices (46) and (73) must have determinant 1, it is possible to show that

$$\begin{bmatrix} 1 & 0 \\ -t(z) & 1 \end{bmatrix} \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H'_{10}(z) & H'_{11}(z) \end{bmatrix} = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix}$$

for

$$t(z) = H'_{10}(z)H_{11}(z) - H'_{11}(z)H_{10}(z). \tag{74}$$

Consequently, we have the relation

$$\begin{aligned} \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} &= \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -K^2 t(z) & 1 \end{bmatrix} \\ &\quad \prod_{n=M}^1 \begin{bmatrix} 1 & q_{2n}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_{2n-1}(z) & 1 \end{bmatrix}, \tag{75} \end{aligned}$$

which by a suitable re-indexing of the  $q$  polynomials (and at the same time making  $K^2 t(z)$  one of them), determines the  $S_n(z)$  and  $T_n(z)$  in Theorem 1.

### Practical Implementation of the Euclidean Algorithm

We now have a step-by-step method for obtaining the lifting steps from any filter. It goes as follows.

1. Find the  $z$ -transform  $H_0(z)$  of the low pass filter taps  $\mathbf{h}$ .
2. Determine  $H_{00}(z)$  and  $H_{01}(z)$  by means of (57).
3. Assign  $a_0(z) = H_{00}(z)$  and  $b_0(z) = H_{01}(z)$ , and let  $n = 0$ .
4. Determine the quotient  $q_{n+1}(z) = a_n(z)/b_n(z)$ . There may be multiple solutions to this step. Choose one.
5. Determine the remainder as  $b_{n+1} = a_n(z) - q_{n+1}(z)b_n(z)$ .
6. Assign  $a_{n+1}(z) = b_n(z)$ .
7. Increment  $n$ . If  $b_n \neq 0$ , go to step 3.
8. Let  $K = a_n(z)$ . Determine  $H'_{10}(z)$  and  $H'_{11}(z)$  using (73).
9. Find the  $z$ -transform  $H_1(z)$  of  $\mathbf{g}$  such that it is index-shifted an odd number compared to  $H_0(z)$ .
10. Determine  $H_{10}(z)$  and  $H_{11}(z)$  by means of (58).
11. Determine  $t(z)$  from (74).
12. Let  $T_n(z) = q_{2n-1}(z)$  for  $n = 1, \dots, M$ .
13. Let  $S_n(z) = q_{2n}(z)$  for  $n = 1, \dots, M$ .
14. Let  $T_{M+1}(z) = -K^2 t(z)$ .
15. If  $K$  has non-zero power of  $z$ , you may discard this. Alternatively, choosing the right  $z$ -transform in step 2 will give a 'zero power'  $K$ .

This procedure will provide all the ingredients for building the lifting step matrices. Note that this will always give a prediction step as the first matrix, and as such this procedure seems unable to generate lifting steps for, say, Daubechies 4 which starts with an update step. However, in the polyphase representation prediction and update steps are merely linear combinations of even and odd samples applied to odd and even samples, and as such predictions and updates can be interchanged simply by an odd shift of the  $z$ -transform of the signal.

### Factoring Daubechies 4 into Lifting Steps

We now give an example of creating lifting steps using the algorithm presented in the previous section. The example is factorization of Daubechies 4. This means that we show how to get the matrix form (67) from the filter tap form (68) and (69). We will follow the step-bystep method laid out in the previous section.

The Daubechies 4 filter taps are given by

$$\mathbf{h} = \begin{bmatrix} \frac{1+\sqrt{3}}{4\sqrt{2}} & \frac{3+\sqrt{3}}{4\sqrt{2}} & \frac{3-\sqrt{3}}{4\sqrt{2}} & \frac{1-\sqrt{3}}{4\sqrt{2}} \end{bmatrix}. \quad (76)$$

In the  $z$ -transform in step 1 we are free to choose a shift. For instance, we could choose

$$H_0(z) = h[0] + h[1]z + h[2]z^2 + h[3]z^3. \quad (77)$$

We know that eventually this choice will provide a monomial  $K$  that may or may not have zero power. As we will see later, this choice does in fact lead to a real number  $K$ .  $H_0(z)$  must be separated into even and odd indices. This happens in step 2. Combining this with assignment to  $a_0(z)$  and  $b_0(z)$  in step 3 we get

$$\begin{aligned} a_0(z) &= H_{00}(z) = h[0] + h[2]z \\ &= \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z, \end{aligned} \quad (78)$$

$$\begin{aligned} b_0(z) &= H_{01}(z) = h[1] + h[3]z \\ &= \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z. \end{aligned} \quad (79)$$

Note that the concept of even and odd applies to the chosen  $z$ -transform of the filters taps (77), not the filter tap vector itself (76). Then step 4 is to find  $q_1(z)$ . Since  $a_0(z)$  and  $b_0(z)$  have the same degree, the quotient is a monomial. Matching the  $z$  coefficients yields

$$\begin{aligned} q_1(z) &= \frac{z \text{ coefficient of } a_0(z)}{z \text{ coefficient of } b_0(z)} = \frac{\frac{3-\sqrt{3}}{4\sqrt{2}}}{\frac{1-\sqrt{3}}{4\sqrt{2}}} = \frac{3-\sqrt{3}}{1-\sqrt{3}} \\ &= \frac{(3-\sqrt{3})(1+\sqrt{3})}{(1-\sqrt{3})(1+\sqrt{3})} = \frac{2\sqrt{3}}{-2} = -\sqrt{3}. \end{aligned}$$

The remainder is then

$$\begin{aligned} b_1(z) &= a_0(z) - b_0(z)q_1(z) \\ &= \left( \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z \right) \\ &\quad - \left( \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z \right) \cdot (-\sqrt{3}) \\ &= \frac{1+\sqrt{3}+3\sqrt{3}+3}{4\sqrt{2}} = \frac{1+\sqrt{3}}{\sqrt{2}}, \end{aligned}$$

thus completing step 5. In step 6 we assign

$$a_1(z) = b_0(z) = \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z.$$

In step 7 we increment  $n = 0$  to  $n = 1$ , and since  $b_1 \neq 0$  we repeat from step 4. This time the quotient has degree 1, since  $b_1(z)$  is one degree less than  $a_1(z)$ . More specifically,  $q_2(z)$  must be on the form  $c + dz$ . Further, since the degree of the remainder decreases, and  $|b_1(z)| = 0$ , the remainder must be zero this time. So

$$\begin{aligned} b_1(z)q_2(z) &= \frac{1+\sqrt{3}}{\sqrt{2}}(c + dz) = a_1(z) \\ &= \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z. \end{aligned}$$

Thus,

$$\begin{aligned} c &= \frac{\frac{3+\sqrt{3}}{4\sqrt{2}}}{\frac{1+\sqrt{3}}{\sqrt{2}}} = \frac{3+\sqrt{3}}{4(1+\sqrt{3})} = \frac{\sqrt{3}}{4} \\ d &= \frac{\frac{1-\sqrt{3}}{4\sqrt{2}}}{\frac{1+\sqrt{3}}{\sqrt{2}}} = \frac{1-\sqrt{3}}{4(1+\sqrt{3})} = \frac{\sqrt{3}-2}{4}. \end{aligned}$$

Therefore,

$$\begin{aligned} q_2(z) &= \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4}z, \quad b_2(z) = 0, \\ a_2(z) &= b_1(z) = \frac{1+\sqrt{3}}{\sqrt{2}}. \end{aligned}$$

Now, this time in step 7 we have  $b_2(z) = 0$ . Thus, we continue to step 8, assign

$$K = \frac{1+\sqrt{3}}{\sqrt{2}},$$

and use (73) to find  $H'_{10}$  and  $H'_{11}$ .

$$\begin{aligned} &\begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H'_{10}(z) & H'_{11}(z) \end{bmatrix} \\ &= \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & q_2(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_1(z) & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sqrt{3}+1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & \frac{\sqrt{3}-2}{4}z + \frac{\sqrt{3}}{4} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\sqrt{3} & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z & \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z \\ \frac{3-\sqrt{3}}{\sqrt{2}} & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix}. \end{aligned} \quad (80)$$

Now that we have  $H'_{10}$  and  $H'_{11}$ , we need  $H_{10}$  and  $H_{11}$  in order to determine  $t(z)$ . First, we need the  $z$ -transform of the high pass filter taps. The high pass filter  $\mathbf{g}$  is the time reversed of  $\mathbf{g}$  with alternating signs. From (76) we therefore get

$$\mathbf{g} = \mathbf{H}_1 = \begin{bmatrix} -\frac{1-\sqrt{3}}{4\sqrt{2}} & \frac{3-\sqrt{3}}{4\sqrt{2}} & -\frac{3+\sqrt{3}}{4\sqrt{2}} & \frac{1+\sqrt{3}}{4\sqrt{2}} \end{bmatrix}. \quad (81)$$

The  $z$ -transform of  $\mathbf{g}$  must have the index shifted by 1 (or some other odd number) compared to  $H_0(z)$ . For instance,

$$H_1(z) = g[0]z^{-2} + g[1]z^{-1} + g[2] + g[3]z, \quad (82)$$

where  $g[3] = h[0]$  now is coefficient for an odd power, whereas  $h[0]$  is coefficient for an even power in  $H_0(z)$ . Splitting this in even and odd coefficients (where even and odd again relates to the  $z$ -transform), we get from (58) that

$$H_{10}(z) = g[0]z^{-1} + g[2] = -\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-1} - \frac{3+\sqrt{3}}{4\sqrt{2}},$$

$$H_{11}(z) = g[1]z^{-1} + g[3] = \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} + \frac{1+\sqrt{3}}{4\sqrt{2}},$$

thus completing step 11. We now insert these  $H_{10}$  and  $H_{11}$  together with  $H'_{10}$  and  $H'_{11}$  from (80) into (74).

$$\begin{aligned} t(z) &= H'_{10}(z)H_{11}(z) - H'_{11}(z)H_{10}(z) \\ &= \frac{\sqrt{3}-3}{\sqrt{2}} \left( \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} + \frac{1+\sqrt{3}}{4\sqrt{2}} \right) \\ &\quad - \frac{\sqrt{3}-1}{\sqrt{2}} \left( -\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-1} - \frac{3+\sqrt{3}}{4\sqrt{2}} \right) \\ &= \frac{1}{8} \left( (\sqrt{3}-3)(3-\sqrt{3})z^{-1} + (\sqrt{3}-3)(1+\sqrt{3}) \right. \\ &\quad \left. + (\sqrt{3}-1)(1-\sqrt{3})z^{-1} + (\sqrt{3}-1)(3+\sqrt{3}) \right) \\ &= \frac{z^{-1}}{8} \left( (-9+6\sqrt{3}-3) + (-1+2\sqrt{3}-3) \right) \\ &= (\sqrt{3}-2)z^{-1}, \end{aligned}$$

which completes step 11. We can now determine the extra matrix we need, as shown in (75),

$$-K^2 t(z) = -\left( \frac{1+\sqrt{3}}{\sqrt{2}} \right)^2 (\sqrt{3}-2)z^{-1} = z^{-1}.$$

It is now possible to determine all the matrix lifting steps that makes up the polyphase matrix, which is step 12, 13, and 14. We have  $M = 1$ , so we get one  $S$  matrix and two  $T$  matrices.

$$T_1(z) = q_1(z) = -\sqrt{3}$$

$$S_1(z) = q_2(z) = \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4}z$$

$$T_2(z) = -K^2 t(z) = z^{-1}.$$

Since  $K$  is a real number, we do not need to shift the powers of  $z$  (as mentioned in step 15), and we can now write the complete factorization of the polyphase matrix  $\mathbf{H}(z)$  as shown in (61) from Theorem 1.

$$\mathbf{H}(z) = \begin{bmatrix} \frac{\sqrt{3}+1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4}z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\sqrt{3} & 1 \end{bmatrix}.$$

But, this is fact not (67). The reason is that the presented algorithm for finding the lifting steps always will start with a prediction step, whereas the factorization in (67) starts with an update step. So while  $\mathbf{H}(z)$  in (67) indeed is equal to  $\mathbf{H}(z)$  above, the factorizations are not the same. This is a consequence of the fact that the distinction of the lifting matrices into update and prediction is useful for interpretation of the transform results, but not necessary for the implementation of the transform.

Actually, this is not the only 'non-uniqueness feature' of the wavelet transform. If instead of (77) and (82) we had chosen

$$H_0(z) = h[0]z^{-3} + h[1]z^{-2} + h[2]z^{-1} + h[3],$$

$$H_1(z) = g[0]z^1 + g[1]z^2 + g[2]z^3 + g[3]z^4,$$

the even coefficients from before are now odd, and vice versa. That means that the relation between  $\mathbf{h}$  and  $\mathbf{g}$  changes signs, so for instance,  $g[3] = -h[0]$  now, rather than  $g[3] = h[0]$ . The resulting factorization becomes

$$\mathbf{H}(z) = \begin{bmatrix} -\frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & -\frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -\frac{\sqrt{3}}{4}z^{-1} - \frac{\sqrt{3}+2}{4}z^{-2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \sqrt{3}z & 1 \end{bmatrix}.$$

There are more examples of factorizations in Jensen and la Cour-Harbo [3] and in the original paper by Sweldens and Daubechies [2].

## Bibliography

1. Daubechies I (1992) Ten Lectures on Wavelets. CBSM-NSF Regional Conference Series in Applied Mathematics, vol 60. SIAM, Philadelphia
2. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4(3):245-267
3. Jensen A, la Cour-Harbo A (2001) Ripples in mathematics - the discrete wavelet transform. Springer, Heidelberg
4. Mallat S, Wavelet A (1998) Tour of signal processing. Academic Press Inc, San Diego

5. Mulcahy C (1996) Plotting and scheming with wavelets. *Math Mag* 69(5):323–343
6. Oppenheim AV, Schaffer R (1975) *Digital signal processing*. Prentice Hall, Upper Saddle River
7. Oppenheim AV, Schaffer R, Buck JR (1999) *Discrete-time signal processing*. Prentice Hall, Upper Saddle River
8. Sweldens W (1996) The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl Comput Harmon Anal* 3(2):186–200
9. Sweldens W (1997) The lifting scheme: A construction of second generation wavelets. *SIAM J Math Anal* 29(2):511–546
10. Vetterli M, Kovačević J (1995) *Wavelets and subband coding*. Prentice-Hall, Englewood Cliffs
11. Vetterli M, Kovačević J (2007) *Wavelets and subband coding*, 2nd edn. <http://waveletsandsubbandcoding.org>, accessed on Sept. 5th, 2008

## Wavelets and PDE Techniques in Image Processing, A Quick Tour of

HAO-MIN ZHOU<sup>1</sup>, TONY F. CHAN<sup>2</sup>, JIANHONG SHEN<sup>3</sup>

<sup>1</sup> School of Mathematics, Georgia Institute of Technology, Atlanta, USA

<sup>2</sup> Department of Mathematics, University of California, Los Angeles, USA

<sup>3</sup> Barclays Capital, New York, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Wavelets in Image Processing](#)

[PDE Techniques](#)

[Wavelet Based Variational PDE Methods](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Wavelets** Wavelets are selected functions that generate orthonormal bases of the square integrable function space  $L^2$  (or more generally, frames of  $L^p$  spaces) by using dilations and translations. The basis functions have certain locality, such as compact support or fast decay property. And they are usually organized according to different scales or resolutions, which are called Multi-Resolution Analysis (MRA). Fast wavelet transforms are filtering procedures that compute the projection of any given function onto a wavelet basis.

**Digital images** Digital images usually refer to  $n$  dimensional data arrays recorded by optical or other imaging devices, such as digital cameras, Radar, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI). They can also be generated by computer graphics software. Most digital images in the literature are 2- or 3-dimensional.

**Image restoration** Image restoration rebuilds high-quality images from given images that are corrupted or polluted during acquisition or transmission processes. The most commonly seen restoration tasks are denoising and deblurring. Denoising is to remove random perturbations to individual pixel values. Deblurring is to remove the unwanted correlation between nearby pixels and to recover the original clear images.

**Image compression** Compression converts images from  $n$  dimensional data arrays into “0” and “1” bit streams so that they can be stored or transmitted more efficiently. There are two types of compression, lossy and lossless, depending on whether information is permanently lost or recoverable, respectively. Many of the commonly used compression algorithms, such as the ones used by international image compression standards JPEG and JPEG2000, are transform-based compression, which consists of three basic steps: transform pixel values into frequency coefficients, quantization of the frequency coefficients, and coding to convert them into bit streams.

**Image segmentation** Segmentation partitions images into subregions (segments), on which images share similar features. Each region often corresponds to the image of an individual object in the 3-dimensional world.

**Image inpainting** Inpainting is an artistic word referring to filling in missing image information on damaged regions, e. g., scratches and damages in precious old photos, old Hollywood films, and ancient paintings. The objective of digital image inpainting is to fill in the missing information automatically and meaningfully.

### Definition of the Subject

Explosive information has dominated nearly all aspects of modern society, science and technology. Visualization is one of the most direct and preferable ways to observe information carried by data, which are often massive in size and uncertain in data quality. To better reveal the information, especially when it is hidden, implicit, or corrupted, data must first be properly processed. In achieving this, image processing, which includes many differ-

ent tasks such as compression, restoration, inpainting, segmentation, pattern recognition and registration, has played a critical role. Historically, it has been viewed as a branch of signal processing, and many classical methods are adopted from traditional Fourier-based signal processing algorithms. In the past couple of decades, numerous new competing methods have emerged. Among them, wavelets, variational and PDE techniques, and stochastic methods have demonstrated outstanding performance due to their special properties. For instance, wavelets have become the dominant tool in image processing because of their multiresolution structure, energy concentration ability and fast transform algorithms. The popularity of variational PDE techniques is driven by their extraordinary properties in understanding and manipulating geometrical features. These new techniques have contributed to new observations, understandings, and discoveries in science and technology. Furthermore, many of these methods have been successfully applied to applications in the fields of medical, physical sciences, engineering, and even everyday life.

## Introduction

Digital image processing analyzes or extracts certain information from digital images, which are often viewed as 2- or multi-dimensional data sets in mathematics. Each element in the data sets is called a pixel. Typical image processing tasks include segmentation, restoration, pattern recognition, analysis, compression, registration and motion detection [41,46]. Image processing has a wide range of applications including communication, computer vision, acoustics, satellite imaging, medical and industrial diagnosis and many more.

Image processing tasks often require large-scale computations, mainly due to the large amount of data to be processed. A typical gray scale still image with moderate resolution, such as  $1024 \times 1024$ , has over a million pixels. The size of a color image is three times as large given the same resolution. A video sequence usually consists of over 24 color frames per second with each frame being a still image. A multi-spectral image contains a collection of several (usually more than 3) monochrome images of the same scene, each of which is taken with a different wavelength by a different sensor. In addition, many applications, such as airport screening and unmanned vehicle navigation, require real time response. All of these demand efficient and reliable algorithms.

Traditional image processing methods are mainly based on Fourier/wavelets or statistical approaches. The best example is the current international image compression

standards, JPEG and JPEG2000, which are based on discrete cosine transform (DCT) and wavelet transforms. For this reason, more images are stored using their wavelet coefficients. The tremendous success of wavelets in image processing is due to their positive properties, including multiresolution data structures, fast transform algorithms and superb energy concentration ability, which allows one to approximate functions (images) using only a relative small number of coefficients.

Thousands of researchers have devoted their efforts to the development of wavelet theory, analysis, and algorithms in different applications. Groundbreaking contributions include Meyer's wavelet theory [53], Daubechies' compact support orthogonal wavelets [33], Mallat's multiresolution analysis [49,50], Shapiro's progressive zero tree image coding algorithm [63], and many other works cited in books such as [28,34,44,51], and [64].

Roughly speaking, wavelet transforms can express any square integrable functions by superpositions of wavelet basis functions, which are generated by dilations and translations from a few (if not a single) wavelet functions. The summation coefficients are called wavelet coefficients, which are standard  $L^2$  inner products between wavelets and the given functions. Wavelet transforms are realized by filtering procedures. Usually, wavelet coefficients are classified into two types: low or high frequencies. Low frequency coefficients correspond to certain kinds of weighted local averages of the data values. High frequency coefficients are related to certain order derivatives. Therefore, high frequency coefficients are small for smooth functions and large for functions containing discontinuities.

In applications, it is inevitable that some of the wavelet coefficients, especially the high frequencies, are unavailable due to intentional or involuntary reasons. For instance, in wavelet-based image compression, insignificant (smaller in magnitude) high frequency coefficients are discarded on purpose to save more storage space. In lossy channel communication, coefficients are lost or damaged during the transmission due to unwanted disturbances. Obviously, with incomplete wavelet coefficients, one cannot synthesize the exact original functions. Many problematic issues could arise as a result. One that has drawn the most attention is the assertion that oscillations are generated near discontinuities. This is the famous Gibbs' phenomenon in mathematics and edge artifacts in image processing.

Several directions have been taken to improve the performance of wavelet based image processing methods by reducing the Gibbs' oscillations, and by better preserving geometrical information in images. One strategy involves

the use of nonlinear thresholding procedures to allocate more storage resource to significant coefficients. Well-known examples include translation invariant denoising methods [31], wavelet hard thresholding, and wavelet shrinkage (also called soft thresholding) [37].

Another strategy is to build new geometry friendly wavelet-like multiresolution representations, such as ridgelets [8], curvelets [9], beamlets [36], bandelets [57] and many more recent developments. With geometry directly incorporated into the construction of multiresolution representations, it is expected that the decompositions have better performance near discontinuities.

The third direction is to modify the existing wavelet transforms so that fewer large high frequency coefficients are generated near discontinuities. Thus, less information is truncated in the thresholding process. Many methods have been proposed, such as Harten's remarkable general multiresolution framework [42] and its recent developments [2], the adaptive lifting scheme [30], and the adaptive Essential Non-Oscillatory (ENO) wavelet transforms [25,26]. Many recent contributions are collected in [65].

In a different direction, PDE techniques for image processing, pioneered by Mumford–Shah's segmentation functional [55], Rudin–Osher–Fatemi's Total Variation (TV) restoration [59], and Perona–Malik's anisotropic diffusion [58], have emerged more recently. Due to their outstanding properties in handling geometrical information, different variational PDE models and methods have been proposed and studied for a variety of image processing goals, such as affine scale space [62], fundamental equations for image processing [1], total variation image analysis [16], active contour for segmentation [12,23], blind deconvolution [24], image interpolation and inpainting [4, 5,18,20,52], restoration [17,19], and compression [27,38]. The field is significantly enriched, and many books have been published in recent years (see [3,13,21,54,56,61,66] and references therein).

Given the developments in both wavelets and PDE techniques in image processing, it is natural to think of combining their advantages to gain more benefits in the applications, especially when geometrical features are important. Well designed wavelet PDE methods can retain the good properties of wavelets, such as multiresolution and fast algorithms. Meanwhile, they are able to use PDE concepts, such as gradients, curvatures to capture, control and manipulate the geometrical information to achieve image processing goals in more systematic manners. There are quite a few examples that have demonstrated the combined advantages in different applications [10,15,27,29, 39,48].

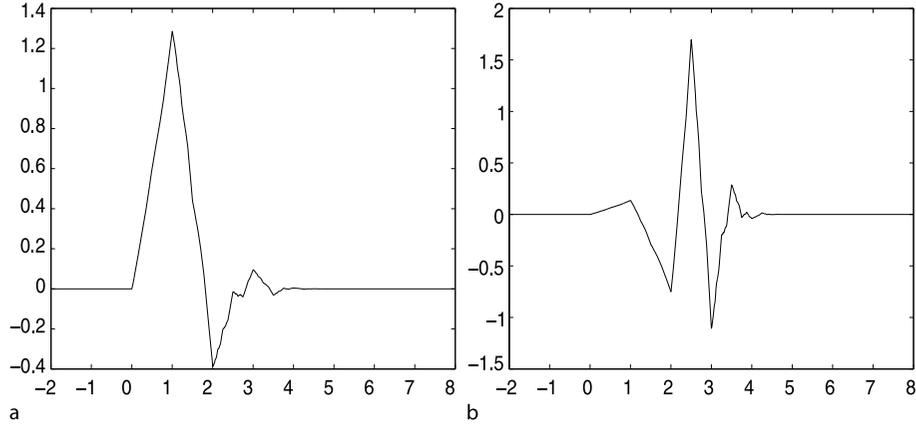
In this paper, it is not our intention to give a complete survey on either wavelets or PDE techniques in image processing. Instead, we will focus on a recent trend that combines them together. To be self-contained, we start with a brief introduction to wavelets, and followed by PDE techniques in image processing. We hope to use selected topics based on our experience to help readers, especially beginners, to know some basic models and a few commonly used methodologies on the subject. The rest of the paper is arranged as follows. Section “Wavelets in Image Processing” is a brief introduction to wavelets and their applications in image processing. Section “PDE Techniques” presents some well known PDE models in image processing. In Sect. “Wavelet Based Variational PDE Methods” we give some new developments of combining wavelets and PDE techniques. A concise list of future directions is stated in the end.

## Wavelets in Image Processing

Historically, Fourier decompositions, which express any given square integrable function by superpositions of sinusoidal functions, have been the major tool for image processing due to their efficient representations and fast Fourier transforms (FFT). This is particularly true for 1-D signals, such as audio sequences. However, all Fourier basis functions have global supports, which implies that any local change in the given function has to result in a global change in the representations. For this reason, Fourier bases are not efficient to represent local information, such as discontinuities. The well-known Gibbs' phenomena is an exhibition of this limitation. Unfortunately, most salient features, such as edges and corners in images, are local and discontinuous. Thus, all Fourier-based methods for image processing suffer from the ringing artifacts.

Facing this shortcoming, it is highly desirable to have efficient representations which can better handle local information, especially discontinuities. Or more precisely, the basis functions should have local support or fast decay properties so that any local perturbation will only cause changes in a small neighborhood but not to far away places. To a certain extent, wavelets are designed to fill up this expectation and have gained unsurpassed success in many applications of image processing.

After several decades of intensive studies, wavelets have been developed into a very rich mathematical theory. There are many different types of wavelets such as Meyer's wavelets, spline wavelets, and bi-orthogonal wavelets. Here, we present a very brief introduction based on Daubechies' compact supported wavelets and their connections to compression and denoising.



**Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 1**  
**a** The scaling function for Daubechies-6 wavelet. **b** The corresponding wavelet

### Wavelets

Wavelets can be viewed as orthonormal bases of the square integrable function space  $L^2(\mathbf{R})$ . It starts with carefully selected scaling function  $\phi(x)$  and corresponding wavelet  $\psi(x)$  defined on finite support  $[0, l]$ , where  $l$  is a positive integer. We refer to [34] for the detailed selection procedure for  $\phi(x)$  and  $\psi(x)$ . Many commonly used software such as MATLAB have built-in routines for the scaling and wavelet functions already.

The functions  $\phi(x)$  and  $\psi(x)$  satisfy the dilation equations (also called two-scale relations or refinement equations in some literature):

$$\phi(x) = \sqrt{2} \sum_{s=0}^l c_s \phi(2x - s), \quad (1)$$

and

$$\psi(x) = \sqrt{2} \sum_{s=0}^l h_s \phi(2x - s), \quad (2)$$

where the  $c_s$ 's and  $h_s$ 's are constants, called low- and high-pass filters, respectively. To give examples, the famous Haar wavelet selects

$$\phi(x) = \begin{cases} 1 & x \in [0, 1) \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\psi(x) = \begin{cases} 1 & x \in [0, \frac{1}{2}) \\ -1 & x \in [\frac{1}{2}, 1) \\ 0 & \text{otherwise,} \end{cases}$$

which are step functions. We also plot the scaling and wavelet functions of Daubechies-6 in Fig. 1.

Using dilation and translation, one can form families of functions from  $\phi(x)$  and  $\psi(x)$ , as follows,

$$\phi_{j,k}(x) = 2^{\frac{j}{2}} \phi(2^j x - k), \quad (3)$$

and

$$\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k), \quad (4)$$

where  $(j, k)$  are integers. Then the collection of  $\psi_{j,k}(x)$  form an orthonormal basis of  $L^2(\mathbf{R})$ . This means that for any given function  $f(x) \in L^2(\mathbf{R})$ , one has

$$f(x) = \sum_{j,k} \langle f(x), \psi_{j,k}(x) \rangle \psi_{j,k}(x), \quad (5)$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard  $L^2(\mathbf{R})$  inner product defined by

$$\langle f(x), g(x) \rangle = \int_{\mathbf{R}} f(x)g(x)dx.$$

There are many desirable properties for the scaling functions and wavelets. Among them, locality and oscillations are the most cited common features in all wavelets. Literally speaking, they make wavelets behave like localized small waves, which also explains the origination of the name.

The locality, which often refers to compact support or fast decay properties, enables wavelets to decompose or approximate functions locally. This satisfies the desire of many applications, particularly in image processing.

A good mathematical way to describe the oscillatory nature of wavelets is to use their vanishing moment property, which means

$$\int \psi(x)x^j dx = 0, \quad j = 0, 1, \dots, p-1, \quad (6)$$

where  $p$  is a positive integer. In this case, the wavelet  $\psi(x)$  is said to have  $p$  vanishing moments. The more vanishing moments, the more oscillations in wavelets in general.

Locality and oscillation together have been the main driving engines for the success of wavelets in many applications.

### Multi-Resolution Analysis

The success of wavelets also relies on their connection to Multi-Resolution Analysis, introduced by Mallat [49,50].

Consider the subspace of  $L^2(\mathbf{R})$  defined by the scaling function  $\phi_{j,k}(x)$ ,

$$V_j = \text{span}\{\phi_{j,k}(x), k \in \mathbf{Z}\},$$

for every fixed  $j$ . The dilation Eq. (1) implies that the subspaces form an ordered chain,

$$\cdots \subseteq V_{j-1} \subseteq V_j \subseteq V_{j+1} \subseteq V_{j+2} \cdots, j \in \mathbf{Z},$$

which also satisfies

$$\overline{\lim_{j \rightarrow \infty} V_j} = L^2(\mathbf{R}), \lim_{j \rightarrow -\infty} V_j = 0.$$

Here, larger indexes  $j$  correspond to finer resolutions or scales.

Similarly, one can define the subspaces generated by wavelets  $\psi_{j,k}(x)$ ,

$$W_j = \text{span}\{\psi_{j,k}(x), k \in \mathbf{Z}\}.$$

The dilation Eq. (2) implies the following connection between  $W_j$  and  $V_j$ ,

$$V_j = V_{j-1} \oplus W_{j-1}, j \in \mathbf{Z}. \quad (7)$$

Therefore  $L^2(\mathbf{R})$  can be decomposed into,

$$L^2(\mathbf{R}) = V_J \oplus \sum_{j>J} W_j = \sum_{j=-\infty}^{\infty} W_j,$$

where  $J$  is an arbitrary reference resolution level. Consequently,  $f(x) \in L^2(\mathbf{R})$  can be decomposed into a multi-resolution representation as,

$$f(x) = \sum_k \alpha_{J,k} \phi_{J,k}(x) + \sum_{j>J,k} \beta_{j,k} \psi_{j,k}(x), \quad (8)$$

where  $\alpha_{j,k} = \langle f(x), \phi_{j,k}(x) \rangle$  is called a low frequency (or scaling) coefficient, and  $\beta_{j,k} = \langle f(x), \psi_{j,k}(x) \rangle$  is a high frequency (or wavelet) coefficient. Without causing confusion, we call them wavelet coefficients for simplicity in this paper.

The decomposition (7) and the dilation Eqs. (1), (2) lead to the following filtering and down-sampling procedures to compute the coarser scale wavelet coefficients

from the finer scale coefficients,

$$\alpha_{j,k} = \sum_{s=0}^l c_s \alpha_{j+1,2k+s}, \quad (9)$$

and

$$\beta_{j,k} = \sum_{s=0}^l h_s \alpha_{j+1,2k+s}. \quad (10)$$

These are the famous fast wavelet transforms.

Apparently, fast wavelet transforms involve only the coefficients and can be started if one knows the low frequency coefficients  $\{\alpha_{I,k}\}$  on a certain fine resolution  $I$ . Then, it is natural to ask how to obtain  $\{\alpha_{I,k}\}$ . Theoretically,  $\{\alpha_{I,k}\}$  must be computed by  $\langle f(x), \phi_{I,k}(x) \rangle$  according to the definition. However, they are often replaced by the point-wise values  $f(x_k)$  in practice, even though such an action is called a wavelet crime in [64]. The replacement makes sense when the function  $f(x)$  is smooth and the resolution  $I$  is fine enough, because the low frequency coefficients  $\alpha_{I,k}$ , which are the weighted local averages of  $f(x)$ , are very close approximations to the point-wise values.

The above described wavelet transforms are for 1-D functions. Wavelet transforms for 2-D images are realized by simple tensor product in practice. More precisely, 2-D transforms are obtained by performing column-wise 1-D transforms followed by row-wise 1-D transforms.

### Wavelet Thresholding and Image Processing

The wavelet representations (8) provide a mechanism to approximate functions in a multi-resolution fashion. For instance, the  $j$ th scale (resolution) approximation is simply defined as:

$$\begin{aligned} f_j(x) &= \sum_k \alpha_{j,k} \phi_{j,k}(x) \\ &= \sum_k \alpha_{J,k} \phi_{J,k}(x) + \sum_{J<i<j,k} \beta_{i,k} \psi_{i,k}(x). \end{aligned} \quad (11)$$

This multi-resolution approximation satisfies a standard error bound,

$$\|f(x) - f_j(x)\| \leq C 2^{-jp} \|f^{(p)}(x)\|, \quad (12)$$

where  $C$  is a constant independent of  $j$ . It is obvious that the error is controlled by the vanishing moment  $p$ , the norm of the  $p$ th derivative of  $f(x)$ , and the resolution  $j$ . Better approximations with more detailed information can be easily obtained by adding more terms for the finer resolutions. Many have argued that this convenient zoom-in and zoom-out multi-resolution approximation is by far one of the best mathematical models that mimic human perceptions.

In addition to the multi-resolution structure, the success of wavelet decomposition in image processing also depends on the sparsity of the wavelet coefficients. Simple integration by parts can show that high frequency wavelet coefficients satisfy  $|\beta_{j,k}| = |f^{(p)}(x)|O(2^{-jp})$ , which suggests that the wavelet coefficients are small if the function  $f(x)$  is smooth enough. For images containing many smooth regions, such as most of the natural scenery images, it is easy to observe that a large number of the high frequency coefficients are insignificant, and therefore can be ignored in applications. Thresholdings are mathematical procedures that realize this observation.

Loosely speaking, thresholding is setting selected wavelet coefficients to be zero. There are many different types of thresholdings. In fact, the  $j$ th scale approximation is one of them. It is constructed by ignoring all the scales higher than the given resolution  $j$ . This is often called linear thresholding, because the procedure is linear. Other nonlinear data dependent thresholdings, including commonly used hard and soft thresholdings, can achieve much better performance in image processing.

The hard thresholding simply sets any wavelet coefficients whose magnitudes are smaller than a given tolerance  $\epsilon$  to be zero, i. e.

$$\tilde{\beta}_{j,k} = \begin{cases} \beta_{j,k} & |\beta_{j,k}| > \epsilon \\ 0 & |\beta_{j,k}| \leq \epsilon. \end{cases}$$

A similar formula holds for the low frequency coefficients too.

The soft wavelet thresholding is slightly different from the hard thresholding. It is a shrinkage procedure. In addition to setting the coefficients whose magnitudes are smaller than the tolerance to zero, it reduces the magni-

tudes of other coefficients by  $\epsilon$  as well,

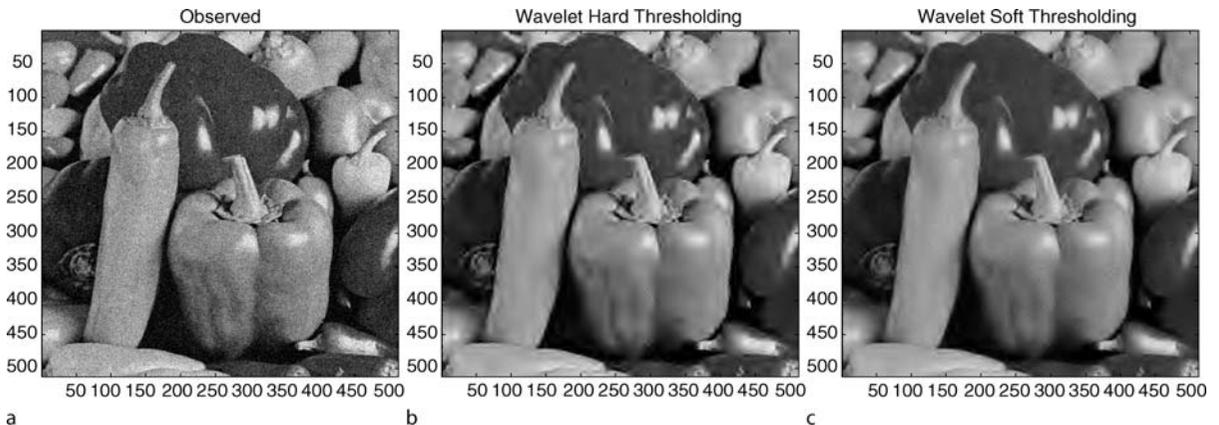
$$\tilde{\beta}_{j,k} = \begin{cases} \text{sign}(\beta_{j,k})(|\beta_{j,k}| - \epsilon) & |\beta_{j,k}| > \epsilon \\ 0 & |\beta_{j,k}| \leq \epsilon, \end{cases}$$

where the  $\text{sign}(\cdot)$  is the signum function.

The selection of the threshold  $\epsilon$  has also been investigated by many groups. Among many proposed strategies, Donoho–Johnstone’s *SQTWOLOG* [37] and Stein’s unbiased risk estimate have been widely used.

Thresholding procedures have accomplished remarkable success in image processing, especially in compression. It is easy to understand that wavelet thresholdings are useful in this application because one does not have to store the coefficients that are zero. However, it is more subtle in practical compression schemes. The problem is that not only does one need to remember the non-zero wavelet coefficients, but also their locations. The location information may occupy more storage space than the coefficients if they are recorded in a naive way. Shapiro’s zero tree scheme [63] introduces a tree structure for wavelet coefficients based on their multiresolution property. A branch of the tree can be represented by a single bit ‘0’ if all coefficients in the branch are zero. This is used in conjunction with thresholdings to achieve very efficient compression. Many well known state-of-the-art compression methods, such as Set Partitioning in Hierarchical Trees (SPIHT) [60] and Group Test Wavelet (GTW) [45] compression algorithms, are based on the zero tree idea.

Simple thresholdings also provide fast and effective methods for noise removal. They have found many successful applications in communications, military and medical images. In Fig. 2, we display the denoising effects of wavelet hard (b) and soft (c) thresholdings of a test image with additive white noise (a).



Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 2

a Test image corrupted by white noise. b Denoised image by hard thresholding. c Denoised image by soft thresholding

From a mathematical point of view, the success of thresholdings can be explained by their connections to optimizations. It has been shown that many thresholding results are optimal in a certain sense. In layman's terms, those thresholding results are best under certain criteria. For example, let us assume that the hard thresholding reconstruction

$$\tilde{f}(x) = \sum_k \tilde{\alpha}_{j,k} \phi_{j,k}(x) + \sum_{j,k} \tilde{\beta}_{j,k} \psi_{j,k}(x)$$

has  $M$  nonzero wavelet coefficients. Then  $\tilde{f}(x)$  is the minimizer of the following optimization problem,

$$\min_g \|f - g\|_2, \quad \text{subject to} \\ g \text{ has at most } M \text{ nonzero wavelet coefficients.}$$

This leads to the conclusion that the hard thresholding gives the best  $M$ -term approximation in  $L^2(\mathbf{R})$  among all possible combinations.

In a more general setting as discussed in [15] and [67], it is proved that the soft thresholding gives the minimizer of the following optimization problem

$$\min_g \{ \|f - g\|_2 + 2\epsilon \|g\|_{B_1^1(L^1)} \},$$

where  $B_1^1(L^1)$  is a Besov space. And the linear thresholding gives an approximate minimizer of the following optimization problem,

$$\min_g \{ \|f - g\|_2 + 2\epsilon \|g\|_{W^m(L^2)} \},$$

where  $W^m(L^2)$  is a Sobolev space. We refer readers to [15] for a detailed discussion.

## PDE Techniques

Compared to wavelets, modern PDE techniques in image processing have appeared more recently, even though some traditional image processing methods can be interpreted from PDE perspective. For instance, the classical Gaussian filter for image denoising is accomplished by convolving the noisy image  $u_0$  with the Gaussian kernel (also called heat kernel in literature)  $G(x, t) = \frac{1}{2\pi t} \exp(-\frac{x^2}{2t})$ ,

$$u = G * u_0 = \int u_0(y) G(x - y, t) dy. \quad (13)$$

This denoised image  $u$  is actually the solution  $u(x, t)$  of the following diffusion PDE,

$$u_t(x, t) = D\Delta u(x, t), \quad u(x, 0) = u_0(x), \quad (14)$$

where  $\Delta$  is the Laplace operator, and  $D = 1/2$  is diffusive coefficient.

Modern PDE techniques have drawn great attention and reached remarkable success in the past two decades.

This is due to their extraordinary ability to handle geometrical features, which are lacking in traditional statistical or Fourier/wavelet based approaches. Two different strategies are commonly used to design PDE techniques for different image processing goals.

1. Construct PDE-based evolution processes and incorporate geometry in the equations.
2. Pose image processing tasks in variational framework and derive corresponding Euler-Lagrange equations to compute the minimizers.

In both strategies, image processing goals are achieved by solving PDE's. Next, we use a few well-known examples to demonstrate these two strategies.

## Anisotropic Diffusion for Denoising

Image denoising removes unwanted disturbances in images. Very often, those disturbances, such as white noise and pepper-and-salt noise, are highly localized and oscillatory. This makes it harder to separate noise from edges, which are also local and discontinuous. As an anti-oscillation procedure, diffusion is a natural selection for denoising. As mentioned earlier, the classical Gaussian filter for denoising is equivalent to the linear isotropic diffusion (14). However, it has been observed in both experimental and theoretical studies that isotropic diffusion unavoidably smears sharp edges, corners and other geometrical features embedded in  $u_0$  while filtering out noise. This is because it treats all orientations identically and never recognizes the presence of spatially coherent discontinuities – edges. In addition, the larger the diffusive coefficient  $D$ , the quicker the smoothing out.

To remedy this drawback, Perona–Malik [58] proposed using anisotropic diffusion instead,

$$u_t = \nabla \cdot (D(x, u, \nabla u) \nabla u). \quad (15)$$

The diffusivity coefficient  $D$  is data dependent and must sense the existence of edges, so that the PDE stops diffusion across the discontinuities. For this purpose, it is desirable to have  $D$  satisfying the following requirements,

$$D = \begin{cases} \text{large,} & \text{when } |\nabla u| \text{ is small on intra-regions,} \\ \text{small,} & \text{when } |\nabla u| \text{ is large near edges.} \end{cases} \quad (16)$$

Therefore, the evolution only smooths out the oscillations away from edges but not across them. In [58],  $D$  is selected as

$$D = g(|\nabla u|^2),$$

where  $g$  is a smooth positive concave function satisfying  $g(+\infty) = 0$ . For example,  $g$  can be taken as

$$g(|\nabla u|^2) = e^{-\frac{|\nabla u|^2}{2\sigma^2}}, \quad \text{or} \quad g(|\nabla u|^2) = \frac{1}{1 + b|\nabla u|^2},$$

where  $\sigma$  and  $b > 0$  are constants.

In practice, the anisotropic diffusion (15) must face a challenge on how to compute the coefficient  $D$  robustly. This may be troublesome especially in the beginning of the diffusion process when  $u_0$  contains highly oscillatory noise, because  $|\nabla u|$  is large almost everywhere, so  $D$  is small everywhere. Thus, the diffusion is not effective in removing noise. To overcome this difficulty, the use of a mollified image in  $g$  has been proposed in [14], which takes the form as

$$u_t = \nabla \cdot (g(|\nabla(G_\sigma * u)|^2)\nabla u), \quad u(x, 0) = u_0(x),$$

where  $G_\sigma$  is again the Gaussian kernel.

Along the lines of anisotropic diffusion, much more research has been done including the well known general axiomatic scale-space theory in [1]. We refer readers to [21,66] for more discussion.

### Total Variation Image Denoising

A different viewpoint for denoising is to reduce the uncorrelated local oscillations in images. Mathematically speaking, total variation (TV) is a quantity that measures oscillations in functions. It is intuitive that oscillatory noise greatly increases the TV norm. Naturally, one can think of denoising as reducing the total variations of images. In fact, this observation leads to the famous TV model proposed by Rudin–Osher–Fatemi [59],

$$\min_u \int |\nabla u| dx \quad \text{subject to} \quad \|u - u_0\|_2 \leq \sigma, \quad (17)$$

where  $\sigma$  is related to the noise level. The objective functional is to reduce oscillations in the reconstruction, and the constraint term is a fitting requirement. This optimization problem can be read as to find the least oscillatory image within a small ball of radius  $\sigma$  centered at the noisy image  $u_0$ .

The model is often re-formulated as a non-constraint minimization problem as

$$\min_u \int |\nabla u| dx + \frac{\lambda}{2} \|u - u_0\|_2, \quad (18)$$

where  $\lambda \geq 0$  is a Lagrange multiplier, which is the factor that balances the competition between oscillations and fidelity. The smaller the  $\lambda$ , the fewer details in the denoised images. In extreme situations, the solution for (18) is a flat constant when  $\lambda$  is zero, or is the noisy image  $u_0$  when  $\lambda$  is infinite.

The most outstanding advantage of TV denoising model (18) is that it allows sharp edges to be preserved in the reconstruction. This implies that TV model has the ability to reduce small oscillations (noise) without penalizing the edges. This feature has been well understood in the context of computational fluid dynamics (CFD), especially in shock capturing, where TV semi norm is intensively used. In fact, the authors of [59] are also experts in CFD, and it is no doubt that (17) is inspired by numerical shock capturing.

Another attraction of TV denoising is its geometrical properties. For functions with finite TV semi norms, this can be seen clearly through an equivalent coarea formula,

$$\int |\nabla u| dx = \int_{-\infty}^{+\infty} \int_{\{u=\gamma\}} ds dy.$$

Here the term  $\int_{\{u=\gamma\}}$  is the length of the level set  $\{u = \gamma\}$ . The TV semi norm is obtained by integrating along all level contours of  $\{u = \gamma\}$  for all values of  $\gamma$ . This suggests that TV semi norm controls both the size of the jumps and the geometry of the level sets.

The geometric connection of TV minimization is more visible if we analyze the optimization (18) by calculus of variation. The standard theory shows that the minimizer must satisfy the following Euler–Lagrange equation,

$$-\nabla \cdot \left( \frac{\nabla u}{|\nabla u|} \right) + \lambda(u - u_0) = 0. \quad (19)$$

The first term, the functional derivative of TV semi norm, is precisely the curvature of the image, which makes the method more geometric friendly. For noisy pixels, the jumps are isolated and their curvature is large. They will be wiped out much quicker than the edges that are coherent jumps with relatively smaller curvature.

The best known, but not necessarily the most efficient, algorithm to solve (18) is the gradient descent method, which introduces an artificial time to form an evolution PDE,

$$u_t = \nabla \cdot \left( \frac{\nabla u}{|\nabla u|} \right) - \lambda(u - u_0). \quad (20)$$

Compared to (15), the gradient descent of TV minimization (20) is also an anisotropic diffusion with a degenerate diffusive coefficient  $D = 1/|\nabla u|$ . And it satisfies the anisotropic diffusion requirement (16). In particular, if an edge is sharp,  $D$  will be zero and no diffusion is performed across the edge. In practice, to prevent numerical blow-up caused by  $|\nabla u| = 0$  in the denominator, it is often replaced by  $\sqrt{|\nabla u|^2 + \epsilon}$ , where  $\epsilon$  is a small positive number. Actually, this replacement can be derived from variational framework too.

An interesting and natural question is why would one want to use TV semi-norm in (18) instead of  $\int |\nabla u|^2 dx$ , which is the famous Sobolev  $H_1$  semi-norm in PDE's. In fact, a very similar calculation can show that the  $H_1$  minimization leads to exactly the isotropic diffusion (14), which loses the geometrical properties.

### Variational Models for Image Segmentation

The purpose of image segmentation is to divide an image into regions within which the image has similar features, such as intensity values, texture pattern, or belonging to same objects. Segmentation is a crucial building block for many high level image processing and vision tasks such as object detection, recognition, and tracking. Obviously, one image may produce different partitions because of different segmentation criteria. This non-uniqueness nature, which is also true for many other image processing tasks, makes the segmentation problem very challenging. There is extensive literature on the subject and many methods have been proposed using different strategies. For example, the celebrated intensity-edge mixture model is statistic-based [40], while the widely-used active contour (also called snake) model [47] uses variational framework. We take the well-known Mumford–Shah segmentation model [55] and Chan–Vese region-based active contour model [23] as examples to demonstrate how mathematical formulations and computational strategies can contribute to segmentation.

The original Mumford–Shah segmentation model is stated in a variational format,

$$\min \lambda_1 \int_{\Gamma} ds + \frac{\lambda_2}{2} \int_{\Omega \setminus \Gamma} |\nabla u|^2 dx + \frac{\lambda_3}{2} \int_{\Gamma} (u - u_0)^2 dx, \quad (21)$$

where  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  are three constants,  $u$  is the partitions with different segments.  $\Omega$  is the region where the image is defined and  $\Gamma$  is the interior boundary separating different segments. The first term is the length of the interior boundary curves. The second term is isotropic diffusion within each homogeneous region. Similar to the TV minimization model (18), the third term is the fitting term. From formulation (21), the segmentation is achieved by balancing the competitive three terms. Different ratios among  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  give different partitions.

The Mumford–Shah model has many desirable properties and is very general. Many other known models can be viewed as special cases of it. However, it also faces serious computational challenges because the partition boundary  $\Gamma$  is unknown. And thus the first term involving

line integral along  $\Gamma$  has no easy way to compute. To ease the challenges, many other models are proposed for better computation properties. Among them, Chan–Vese's *active contour without edge* model [23] has gained remarkable success due to its simplicity and robustness.

Assume that  $C$  is a closed curve partitioning the segments. The model is designed to move  $C$  so that the following energy is minimized,

$$\min \lambda_1 \cdot \text{Length}(\Gamma) + \lambda_2 \cdot \text{Area}(\text{inside}(\Gamma)) + \lambda_3 \int_{\text{inside}(\Gamma)} (u_0 - c_1)^2 dx + \lambda_4 \int_{\text{outside}(\Gamma)} (u_0 - c_2)^2 dx, \quad (22)$$

where  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  are positive fixed parameters. This model uses piecewise constant approximations inside and outside the partition curves. If one picks  $\lambda_2 = 0$ , (22) becomes the minimal partition model which is a special case for (21).

The Chan–Vese model (22) can also be formulated in a level set framework, and it leads to a fast and robust computation method, which sparkles a large amount of follow-up researches in using level set-based active contour methods for segmentations in different applications.

### PDE Method for Image Inpainting

Image inpainting, or its mathematical synonym image interpolation, fills in missing or damaged image regions based on known surrounding information. It is a very fundamental problem having numerous prior work in existence. It also shares common ground with many other image processing tasks, such as image replacement, error concealment, edge completion and image editing. Here we only use 1) a third order nonlinear inpainting PDE by Bertalmio et al. [6], 2) a variational inpainting model by Chan–Shen [20], as two examples to illustrate how modern mathematics is used for this traditional labor-intensive task, because image inpainting used to be done by hand.

Similar to segmentation, image inpainting is an inverse problem having possible multiple solutions. It is obvious that when information is missing, different people may have different ways to patch information to the regions, and all of them may look reasonable. However, it is commonly agreed that the inpainted regions must have consistent geometrical features and texture patterns with their surroundings. For this reason, many of the inpainting methods are based on geometrical interpolations or extrapolations. One example is the remarkable third order inpainting PDE introduced by Bertalmio et al. [6]. In fact, the term *image inpainting* was first used by them, and

the work has stimulated a wave of interest in inpainting related problems.

The PDE is given as

$$u_t = \nabla(\Delta u) \cdot \nabla^\perp u, \quad (23)$$

where  $\nabla^\perp$  is the orthogonal gradient direction (isophote direction as called in the original paper). The idea behind (23) is a brilliant intuition of information transport along broken level lines (isophotes). The PDE is solved only inside the inpainting regions with proper boundary conditions based on the gray values and isophote directions. It is discovered later that (23) actually connects to the famous Navier–Stokes equations in CFD [5].

The Chan–Shen’s inpainting model tackles the problem from a different angle. It starts with a variational principle inspired by TV restoration model [59],

$$\min_u \int_{\Omega} |\nabla u| dx + \frac{\lambda}{2} \int_{\Omega \setminus D} (u - u_0)^2 dx, \quad (24)$$

where  $D$  is the inpainting region. Similar to (17), a straightforward interpretation of this model is that the minimizer  $u$  is the least oscillatory image that is close enough to the given image  $u_0$  outside the inpainting region. For the regions inside of  $D$ , the restored  $u$  has no restriction except matching its surroundings in a least oscillatory fashion. This model can lead to a nonlinear data dependent PDE similar to (19) and can be solved numerically. The results are impressive and much follow-up work has been performed to analyze the model and extend it to include more sophisticated measurements such as Euler’s Elastica into consideration for better curve treatments [18].

### Wavelet Based Variational PDE Methods

As discussed in the previous sections, both wavelets and PDE techniques have been used extensively in image processing and achieved tremendous success in numerous applications. Their success is based on different properties of both approaches. Wavelets have multi-resolution data structure, energy concentration (sparsity) and fast algorithms. PDE techniques are geometrically friendly and often tied to variational principles. A closer look at both approaches can easily reveal that those properties do not overlap, and one cannot be used to replace the other. In a certain sense, they are complementary to each other, and it seems natural to combine the advantages of both to gain benefits. In fact, many research efforts have been put forward in this direction.

There are two different strategies that have been explored to merge PDE techniques with wavelets,

- (a) Use computational PDE skills to modify the standard wavelet transforms to form new transforms having better geometric properties.
- (b) Design new wavelet-based variational models for different image processing tasks.

In this section, we will select a few examples to demonstrate both strategies.

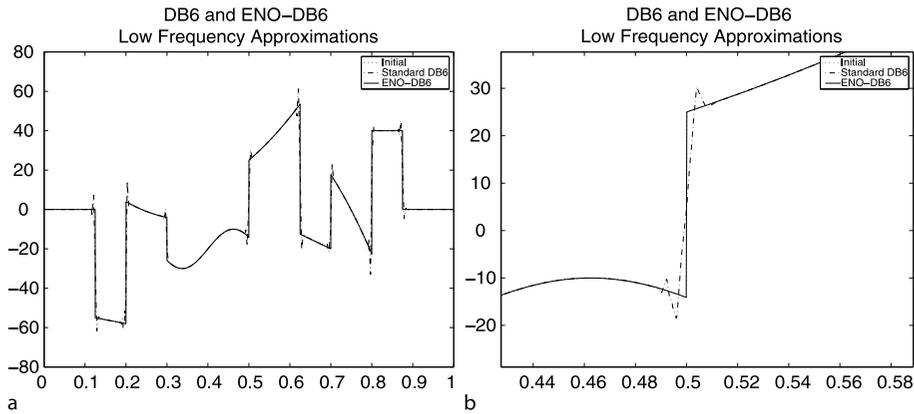
### ENO-Wavelet Transforms

As mentioned earlier in Sect. “Wavelets in Image Processing”, it is a well-known fact that Fourier-based algorithms suffer Gibbs’ oscillations. Wavelets can remarkably reduce the severity of oscillations due to their locality. But they still exist unless one retains all discontinuity related coefficients, which is not practical in many applications. To improve the image quality, one needs to reduce the oscillations by lowering the threshold  $\epsilon$  in thresholding procedures. As a consequence, more coefficients, especially edge-related ones, must be retained, which is why a majority of the storage is allocated to edge-related coefficients in JPEG2000.

To further improve the performance and reduce the ringing artifacts, it is desirable to design wavelet-like transforms so that fewer significant high-frequency coefficients are generated. One way to achieve this goal is to incorporate geometrical features in the design of wavelet-like basis (or redundant frame) functions so that discontinuous functions can be more effectively represented. Many efforts have been proposed, such as curvelets [9], beamlets [36], and bandelets [57] to name a few. A different approach is to reduce the wavelet filter length. This is based on the fact that the larger the supports of wavelets, the longer the filters, and the worse the Gibbs’ oscillations. Then, one can adaptively use shorter wavelet filters near the vicinities of edges. The adaptive lifting scheme proposed in [30] employs this idea.

ENO-wavelet transforms approach the problem from a different angle. It is inspired by the original Harten’s multi-resolution framework [42], which has a profound impact on many new methods in the field. ENO-wavelet transforms borrow a key idea, the one-sided interpolation strategy, from ENO schemes for shock capturing [43]. Different from the fore-mentioned methods, which adapt the filters or basis functions to better fit the data, ENO-wavelet transforms change the data near edge areas and feed them into the same standard wavelet filters. The data is changed in a special way so that the filters do not see the discontinuities.

Let us imagine that we filter around a jump discontinuity. A high-frequency coefficient is large if the high



### Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 3

**a** The comparison between 4-level ENO-Daubechies-6 (solid line) and standard Daubechies-6 (dash-dotted line) approximations. **b** A zoom-in of the picture on the left near a discontinuity. Standard Daubechies-6 generates oscillations near discontinuities, but the ENO-Daubechies-6 does not

pass filter is convolved with data across the jump. However, one can extend the data from both left and right sides in smooth ways and feed the extended data to the filters. Then the high-frequency coefficients are small as the high pass filter only sees the smooth data on both sides. Of course, one may immediately question the fact that near the jump region, we have actually two different pieces of data overlapping in the area. In fact, this is a serious issue, in that it causes a double storage problem, which means we have doubled the number of wavelet coefficients in the jump region. And this is directly against the goals of many image processing tasks, especially image compression. Fortunately, the problem can be avoided by a strategy called coarse-level extrapolation, which extends the data in such ways that some of the jump-related wavelet coefficients are predictable and do not need to be memorized. And the storage can be reduced to the same as that of standard wavelet transforms. We refer to [25] for detailed ENO-wavelet transform algorithms. Here we just point out the main ideas and some important results.

ENO-wavelet transforms can be used as functional replacements of standard wavelet transforms. Indeed, ENO-wavelet transforms perform standard wavelet transforms if no discontinuity is detected. ENO-wavelet transforms retain the essential properties and advantages of standard wavelet transforms, such as energy concentration, multiresolution framework and fast transform algorithms, all without any edge artifacts. They also achieve uniform approximation accuracy up to the discontinuities. If  $\hat{f}_j(x)$  is the  $j$ th resolution approximation to  $f(x)$  by using ENO-wavelet transforms, then

$$\|\hat{f}_j(x) - f(x)\| \leq C2^{-jp} \|f^{(p)}(x)\|_{\Omega \setminus \Gamma}, \quad (25)$$

where  $\Gamma$  is the set of discontinuous points. It is worth noting that the error (12) for standard wavelet transforms depends on the  $p$ th derivative of  $f(x)$  on the entire region  $\Omega$ , which is unbounded if the discontinuous set  $\Gamma$  is not empty. In contrast, the error for ENO-wavelet transforms (25) depends on  $f^{(p)}(x)$  only on the domain  $\Omega$  excluding  $\Gamma$ . This ensures that ENO-wavelet transforms perform uniformly accurate, regardless of the presence of discontinuities, which is probably the best result one may expect. In Fig. 3, we show a comparison between ENO-wavelets and standard wavelets.

### Wavelet Based Minimal Energy Methods for Denoising

As discussed in Sect. "PDE Techniques", anisotropic diffusion and total variation minimization for image denoising have a tremendous ability to extract image features, especially edges, for better image quality preservation. However, it is also commonly recognized that such PDE techniques often post higher computational demands, because numerical solutions for nonlinear PDE's need to be computed iteratively. To achieve reasonable solutions, many iterations must be performed. This has been a major criticism of PDE techniques, especially when one compares them with wavelets, which have ultra fast filtering algorithms.

To retain the capability in feature extraction while eliminating the need of iterations for anisotropic diffusion, there have been efforts to formulate geometric friendly energy minimizations in wavelet spaces so that the minimizers can be obtained directly from wavelet coefficients without iterations. In fact, as discussed in Subsect. "Wavelet Thresholding and Image Processing", classical wavelet

thresholdings, including linear, hard and soft thresholdings, have corresponding energy optimizations in certain functional spaces. But those minimization problems are not built to handle geometrical features. In [29], Chui–Wang suggested a new geometrical energy minimization in wavelet space given as

$$\min E_\lambda(\rho, \beta) = \lambda E_i(\rho, \beta) + \frac{1}{2} \|\beta - \alpha\|^2, \quad (26)$$

where  $E_i(\rho, \beta)$  is a selected internal energy which can be expressed by the wavelet coefficients  $\beta$ . The second term is the standard  $L^2$  fitting requirement. In their original paper, a blended internal energy is chosen as

$$E_i(\rho, \beta) = \sum_{j,k} (\rho(m_{j,k}(p)) + \rho(\beta_{j,k}^d)), \quad (27)$$

where  $m_{j,k}(p) = (|\beta_{j,k}^h|^p + |\beta_{j,k}^v|^p)^{1/p}$ , and  $\rho(s) = |s|$ . In (27), the notations  $\beta^d$ ,  $\beta^h$  and  $\beta^v$  are 2-D tensor product wavelet coefficients along diagonal, horizontal and vertical directions respectively. It is clear that the energy functional is resolution, orientation and spatial dependent. In this way, the energy functional can “see” the corners and edges in wavelet spaces because those geometrical structures create correlated wavelet coefficients along diagonal, horizontal and vertical directions respectively. When  $p = 2$ , the minimizers of (26) and (27) can be attained explicitly from the wavelet coefficients as

$$(\beta_\lambda^*)_{j,k}^{h,v} = (\beta^0)_{j,k}^{h,v} \left( 1 - \frac{\lambda}{m_{j,k}^0} \right)_+, \quad (28)$$

and

$$(\beta_\lambda^*)_{j,k}^d = \text{sign}(\beta)_{j,k}^{0,d} \left( |\beta_{j,k}^{0,d} - \lambda| \right)_+, \quad (29)$$

where  $(\cdot)_+$  denotes the nonnegative value function.

Along this line, there has been a recent trend in the computational harmonic analysis community to design data dependent nonlinear filters based on PDE techniques, for example, the adaptive digital TV filter presented in [19]. More recently, Chui and collaborators have proposed a new anisotropic filtering strategy based on ideas of finding approximate solutions of anisotropic diffusion equations discussed in Subject. “[Anisotropic Diffusion for Denoising](#)”. Their method realizes image denoising by one sweep of nonlinear filtering.

### Diffusion Wavelets

Diffusion wavelets has been proposed by Coifman and collaborators in [32]. It is a different way to generalize classical wavelets using PDE and geometry concepts. The goal is to construct a multiresolution analysis framework on

general geometric structures, such as manifolds, graphs or even discrete point sets, so that image processing tasks can be performed for functions defined on these structures.

As discussed in Subject. “[Wavelets](#)”, standard wavelet multi-resolutions are based on dilation and translations. However, this is often impossible for general data structures, especially when little geometric information is known. To overcome this difficulty, diffusion wavelets use dyadic powers of a diffusion operator  $T$  (with  $\|T\| < 1$ ), such as the heat operator defined on the general data structure to create scales. The following two properties are crucial for constructing diffusion wavelets. One is that the spectral of high powers of  $T$  decay faster as the power gets higher. Consequently, one can use a few leading eigenfunctions of  $T^j$  ( $j$  large) to approximate the range spaces of  $T^j$  accurately.

The other property is that applying higher powers of  $T$  to local functions, such as Dirac delta functions defined on a point in a discrete data set produces smoother functions with wider supports, because  $T$  is a diffusion process. After a non-trivial process involving orthonormalization, which we refer to [32] for details, one can construct a multi-resolution analysis based on  $T^{2^j}$  ( $j \in \mathbf{Z}^+$ ). Especially, once the multi-resolution analysis is formed,  $T^{2^j}$  can be expressed in a highly compressed format.

There are many potential applications, such as in data mining and learning theory. Here, we pick the following simple example to illustrate their usage. Let us consider computing the inverse of Laplacian  $(I - T)^{-1}$  applied to an arbitrary vector  $f$  defined on a general data structure. The operator  $(I - T)^{-1}$  is a deblurring process commonly seen in image restorations. It is well known that

$$(I - T)^{-1}f = \sum_{k=1}^{+\infty} T^k f.$$

Define

$$S_K = \sum_{k=1}^{2^K} T^k,$$

then an approximation to  $(I - T)^{-1}$  can be achieved by

$$S_{K+1}f = (S_K + T^{2^K} S_K)f = \prod_{k=0}^K (I + T^{2^k})f.$$

Since the powers  $T^{2^k}$  have been compressed in the multi-resolution analysis and can be efficiently applied to  $f$ ,  $(I - T)^{-1}f$  is computed efficiently.

### TV Wavelet Inpainting

Wavelet inpainting, or more generally wavelet interpolation, refers to the problem of filling in missing or dam-



in which the retained coefficients are considered to be correct and will not be alerted.

**Model I:**

$$\begin{aligned} \min_{\beta_{j,k}: (j,k) \in I} F(u, z) &= \int_{\mathbb{R}^2} |\nabla_x u(\beta, x)| dx \\ &= \text{TV}(u(\beta, x)), \end{aligned} \quad (30)$$

where  $I$  is the inpainting index region in wavelet domain, and  $u(\beta, x)$  has the wavelet transform:

$$u(\beta, x) = \sum_{j,k} \beta_{j,k} \psi_{j,k}(x).$$

For noisy images, since every coefficient is also corrupted by noise. Then one has to modify (denoise) them too.

**Model II:**

$$\begin{aligned} \min_{\beta_{j,k}} F(u, z) &= \int_{\mathbb{R}^2} |\nabla_x u(\beta, x)| dx \\ &+ \sum_{(j,k)} \lambda_{j,k} (\beta_{j,k} - \alpha_{j,k})^2, \end{aligned} \quad (31)$$

and the parameter  $\lambda_{(j,k)}$  is zero if  $(j, k) \in I$ ; otherwise, it equals a positive constant  $\lambda$ .

Clearly, these two models are inspired by TV denoising model for their exceptional ability of handling geometries. Both models recover the wavelet coefficients so that the restored images are least oscillatory while matching the known information. The key difference is that the arguments are restricted to the inpainting regions  $I$  only for Model I, so the dimension of unknowns is the number of coefficients in  $I$ . While in Model II, the parameter  $\lambda$  is taken to be zero in the inpainting regions  $I$  in the wavelet domain, in contrast to the standard denoising and compression models, where  $\lambda$  is usually taken to be a constant everywhere. This difference essentially puts no constraint on the missing wavelet coefficients so that they can change freely. In Fig. 4, we show an example of wavelet inpainting by the two models.

### Compressive Sampling

Compressive sampling [7], also goes by the name compressed sensing [35], is an emerging theory addressing the sampling problem in image and signal processing. In information theory, the classical Shannon–Nyquist sampling theorem states that “Exact reconstruction of a continuous-time baseband signal from its samples is possible if the signal is bandlimited and the sampling frequency is greater than twice the signal bandwidth”. More precisely, bandlimited signals refer to functions whose Fourier fre-

quencies are in a bounded interval. And the sampling theorem says that a bandlimited signal can be fully reconstructed from its evenly spaced samples, provided that the sampling rate must exceed twice the maximum frequency in the bandlimited signal. This rate is often called Nyquist rate.

Compressive sampling considers a different scenario. In the simplest case, let us assume that a signal  $f(t)$  is sparse in the frequency space or any other convenient spaces. For instance, the signal consists of only a few Fourier terms, i. e.

$$f(t) = \sum_{j=1}^n \beta_j e^{-ik_j t},$$

where  $n$  is an integer much smaller than the signal resolution  $N$ ,  $\{k_j\}$ 's are frequencies, and  $i$  the imaginary unit. But the actual values of frequency  $k_j$  are not known. Obviously,  $f(t)$  is a bandlimited signal. Without loss of generality, we assume  $k_n$  is the highest frequency. Then Shannon–Nyquist sampling theory requires at least  $2k_n$  evenly spaced observations to exactly reconstruct  $f(t)$ . However, since the value  $k_n$  is not known, the actual number of samples (resolution  $N$ ) needed may be much higher than  $2k_n$ .

Compressive sampling asks a different question: Can one exactly recover a sparse signal  $f(t)$  using a small number of samples  $\{f(t_j)\}_1^m$  observed at randomly selected time  $t_j$  ( $j = 1, \dots, m$ )? Here  $m$  may be much smaller than  $N$ . Given the knowledge of sparsity, ideally one can convert this problem into the following minimization problem,

$$\min \|\beta\|_{l_0}, \quad \text{subject to } (F^* \beta)(t_j) = f(t_j), \quad (32)$$

where  $\beta$  is a vector containing the Fourier coefficients of the reconstructed signal,  $F$  is the Fourier matrix, and  $\|\cdot\|_{l_0}$  is the  $l_0$  norm of a discrete sequence, which is defined as the number of nonzero elements. Then  $F^*$  gives the inverse Fourier transform, and  $(F^* \beta)$  is the reconstructed signal.  $l_0$  minimization (32) finds the sparsest reconstruction  $(F^* \beta)$  among all possible functions that agree with the observations  $f(t_j)$ .

However,  $l_0$  minimization (32) is essentially a large non-convex integer optimization problem, which is computationally prohibitive. Then compressive sampling suggests that it is still possible to exactly recover  $f(t)$  from the samples  $\{f(t_j)\}_1^m$ . The exact reconstruction is realized by the following  $l_1$  optimization,

$$\min \|\beta\|_{l_1}, \quad \text{subject to } (F^* \beta)(t_j) = f(t_j). \quad (33)$$

In other words, the compressive sampling achieves exact recovery by finding the signal having the smallest  $l_1$  norm

in frequency space among all signals matching the sample values  $f(t_j)$  at  $t_j$ .

There are many reasons to select  $l_1$  norm in the optimization, including the remarkable mathematical insights given in [11]. We do not intend to present their results here. Instead, we list the following two reasons that are more intuitive and may explain the essence of  $l_1$  optimization in sparse recovery.

1.  $l_1$  norm exhibits interesting sparsity in many applications. In other words,  $l_1$  minimization in frequency space intends to drive more frequencies to zero.
2.  $l_1$  norm has the least index  $p$  among all  $l_p$  norms that are convex.

The convex property ensures that (33) may be solved efficiently by the standard convex optimization algorithms.

It is believed that compressive sampling may have many implications. One of the most attractive potentials is that it suggests the possibility of new data acquisition protocols that translate analog information into digital form with fewer sensors than what was considered necessary. There are many interesting studies on how to advance the theory and applications, and even design new hardware to realize the implications. We refer to [7] for more information on the subject.

It is worth noting that even though TV wavelet inpainting and compressive sampling are developed independently, there is an interesting connection between them. For instance, the derivative of a piecewise constant image,  $\nabla u$ , can be viewed as sparse in the pixel space. If one makes measurements in the wavelet space, then Model I (30) is the  $l_1$  minimization of the derivative in the pixel space with constraints in the wavelet space, which fits well in the framework of compressive sampling. In this sense, they are complementary to each other, and can be viewed as dual formulations.

### Future Directions

Driven by rapidly developing imaging sciences and technologies, the last couple of decades have witnessed the tremendous success of wavelets and PDE techniques in mathematical image processing. Many researchers have been working in the field, and exciting new developments are constantly reported. However, compared to the even faster growing demands, there still is a long way to go to meet the ever-increasing expectations. The following is just a very short list of directions that are being or shall be pursued in the near future.

- (1) Developing more sophisticated models, methods to better preserve features for images, or general data sets

in higher dimensions, such as video or hyper-spectral images. Merging traditional wavelets and PDE techniques seems to be a promising direction. For example, developing wavelets and PDE models for segmentation is interesting. To our knowledge, it has not been attempted yet.

- (2) New applications in high-level vision, such as pattern recognition, auto navigation and tracking, which demands better understanding of the problems and more accurate extraction of the connections among data sets.
- (3) Robust and efficient implementation strategies to compute the solutions of mathematical image processing models, especially those involving solutions of nonlinear PDE's.

The list is based on the authors' experience and reflects our own perspective. Certainly it does not cover all aspects of this large field. Interested readers are encouraged to read up-to-date literature to follow the latest advancements on the subject.

### Acknowledgments

Research supported in part by grants ONR-N00014-06-1-0345, NSF CCF-0430077, CCF-0528583, DMS-0610079, DMS-0410062 and CAREER Award DMS-0645266, NIH U54 RR021813, and STTR Program from TechFinity Inc.

### Bibliography

1. Alvarez L, Guichard F, Lions PL, Morel JM (1993) Axioms and Fundamental Equations of Image Processing. Arch Ration Mech Anal 16:200–257
2. Arandiga F, Donat R (2000) Nonlinear Multiscale Decompositions: The approach of A Harten. Numer Algorithm 23:175–216
3. Aubert G, Kornprobst P (2001) Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations, vol 147. Springer
4. Ballester C, Bertalmio M, Caselles V, Sapiro G, Verdera J (2001) Filling-in by Joint Interpolation of Vector Fields and Grey Levels. IEEE Trans Image Process 10(8):1200–1211
5. Bertalmio M, Bertozzi AL, Sapiro G (2001) Navier–Stokes, fluid dynamics, and image and video inpainting. IEEE Conference on Computer Vision and Pattern Recognition Dec Kauai, Hawaii
6. Bertalmio M, Sapiro G, Caselles V, Ballester C (1999) Image Inpainting, Tech Report. ECE-University of Minnesota
7. Candès E (2006) Compressive Sampling. Proceedings of the International Congress of Mathematicians, Madrid
8. Candès E, Donoho D (1999) Ridgelets: a Key to Higher-dimensional Intermittency? Phil Trans R Soc Lond A 357(1760):2495–2509
9. Candès E, Donoho D (1999) Curvelets: A Surprisingly Effective Nonadaptive Representation of Objects with Edges. Tech Report. Dept of Stat, Stanford University

10. Candès E, Romberg J, Tao T (2006) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Trans Inform Theory* 52:489–509
11. Candès E, Tao T (2006) Near Optimal Signal Recovery From Random Projections and Universal Encoding Strategies. *IEEE Trans Inform Theory* 52:5406–5425
12. Caselles V, Kimmel R, Sapiro G (1997) On Geodesic Active Contours. *Int J Comput Vis* 22(1):61–79
13. Caselles V, Morel JM, Sapiro G, Tannenbaum A (1998) Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis. *IEEE Tran Image Proc* 7(3)
14. Catte F, Lions PL, Morel JM, Coll T (1992) Image Selective Smoothing and Edge Detection by Nonlinear Diffusion. *SIAM J Numer Anal* 29:182–193
15. Chambolle A, DeVore R, Lee N, Lucier B (1998) Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal Through Wavelet Shrinkage. *IEEE Tran Image Proc* 7(3):319–333
16. Chambolle A, Lions PL (1997) Image Recovery via Total Variational Minimization and Related Problems. *Numer Math* 76:167–188
17. Chan TF, Golub GH, Mulet P (1996) A Nonlinear Primal-Dual Method for Total Variation-Based Image Restoration In: *ICAOS'96, 12th International Conference on Analysis and Optimization of Systems: Images, Wavelets and PDEs*, Paris, 26–28 June, 1996. Lecture Notes in Control and Information Sciences, vol 219. Springer, pp 241–252
18. Chan TF, Kang SH, Shen J (2002) Euler's Elastica and Curvature Based inpainting. *SIAM J Appl Math* 63(2):564–592
19. Chan TF, Osher S, Shen J (2001) The Digital TV Filter and Nonlinear Denoising. *Trans IEEE Image Process* 10(2):231–241
20. Chan TF, Shen J (2002) Mathematical Models for Local Non-Texture Inpainting. *SIAM J Appl Math* 62(3) 1019–1043
21. Chan TF, Shen J (2005) *Image Processing and Analysis - Variational, PDE, Wavelet and Stochastic Methods*. SIAM Publisher, Philadelphia
22. Chan TF, Shen J, Zhou HM (2006) Total Variation Wavelet Inpainting. *J Math Imaging Vis* 25(1):107–125
23. Chan TF, Vese L (2001) Active Contour Without Edges. *IEEE Trans Image Process* 10(2):266–277
24. Chan TF, Wong CK (1998) Total Variation Blind Deconvolution. *IEEE Trans Image Process* 7:370–375
25. Chan TF, Zhou HM (2002) ENO-wavelet Transforms for Piecewise Smooth Functions. *SIAM J Numer Anal* 40(4):1369–1404
26. Tony Chan F, Zhou HM (2003) ENO-wavelet Transforms and Some Applications In: *Stoeckler J, Welland GV (eds) Beyond Wavelets*. Academic Press, New York
27. Chan TF, Zhou HM (2000) Optimal Constructions of Wavelet Coefficients Using Total Variation Regularization in Image Compression, CAM Report, No 00–27. Dept of Math, UCLA
28. Chui CK (1997) *Wavelet: A Mathematical Tool for Signal Analysis*. SIAM, Philadelphia
29. Chui CK, Wang J (2007) Wavelet-based Minimal-Energy Approach to Image Restoration. *Appl Comp Harmon Anal* 23(1):114–130
30. Claypoole P, Davis G, Sweldens W, Baraniuk R (1999) Nonlinear Wavelet Transforms for Image Coding. *Correspond. Author: Baraniuk, Dept of Elec and Comp Sci; Submit to IEEE Transactions on Image Processing* 12(12):1449–1459
31. Coifman R, Donoho D (1995) Translation invariant de-noising In: *Antoniadis A, Oppenheim G (eds) Wavelets and Statistics*. Springer, New York, pp 125–150
32. Coifman R, Maggioni M (2006) Diffusion Wavelets. *Appl Comp Harm Anal* 21(1):53–94
33. Daubechies I (1988) Orthonormal Bases of Compactly Supported Wavelets. *Comm Pure Appl Math* 41:909–996
34. Daubechies I (1992) *Ten Lectures on Wavelets*. SIAM, Philadelphia
35. Donoho D (2004) *Compressed Sensing*. Tech Report, Stanford University
36. Donoho D, Huo X (2002) Beamlets and multiscale image analysis. In: *Barth TJ, Chan T, Haimes R (eds) Multiscale and Multiresolution Methods. Lecture Notes in Computational Science and Engineering*, vol 20. Springer, pp 149–196
37. Donoho D, Johnstone I (1995) Adapting to Unknown Smoothness via Wavelet Shrinkage. *J Amer Stat Assoc* 90:1200–1224
38. Dugatkin D, Zhou HM, Chan TF, Effros M (2002) Lagrangian Optimization of A Group Testing for ENO Wavelets Algorithm. *Proceedings to the 2002 Conference on Information Sciences and Systems*, Princeton University, New Jersey, 20–22 March, 2002
39. Durand S, Froment J (2001) Artifact Free Signal Denoising with Wavelets In: *Proceedings of ICASSP'01*, vol 6. pp 3685–3688
40. Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans Patt Anal Mach Intell* 6:721–741
41. Gonzalez RC, Woods RE (1993) *Digital Image Processing*. Addison Wesley, Reading
42. Harten A (1994) Multiresolution Representation of Data, II General Framework. Dept of Math, UCLA, CAM Report 94-10
43. Harten A, Engquist B, Osher S, Chakravarthy S (1987) Uniformly High Order Essentially Non-Oscillatory Schemes, III. *J Comput Phys* 71:231–303
44. Hernandez E, Weiss G (1996) *A First Course on Wavelets*. CRC Press, Boca Raton
45. Hong ES, Ladner RE (2000) Group testing for image compression. *IEEE Proceedings of the Data Compression Conference, Snowbird, UT, March 2000*, pp 2–12
46. Jain AK (1989) *Fundamentals of Digital Image Processing*. Prentice Hall, Englewood Cliffs
47. Kass M, Witkin A, Terzopoulos D (1987) Snakes: Active Contour Models. *Int J Comput Vis* 1:321–331
48. Malgouyres F, Guichard F (2001) Edge Direction Preserving Image Zooming: a Mathematical and Numerical Analysis. *SIAM J Num Anal* 39(1):1–37
49. Mallat S (1989) Multiresolution Approximation and Wavelet Orthonormal Bases of  $L^2(\mathbb{R})$ . *Trans Amer Math Soc* 315:69–87
50. Mallat S (1989) A Theory of Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans PAMI* 11:674–693
51. Mallat S (1998) *A Wavelet Tour of Signal Processing*. Academic Press, San Diego
52. Masnou M, Morel J (1998) Level-lines Based Disocclusion. *IEEE ICIP* 3:259–263
53. Meyer Y (1992) *Wavelets and operators*, Advanced mathematics. Cambridge University Press, Cambridge
54. Morel JM, Solimini S (1994) *Variational Methods in Image Segmentation*. Birkhauser, Boston
55. Mumford D, Shah J (1989) Optimal Approximation by Piecewise Smooth Functions and Associated Variational Problems. *Comm, Pure Appl Math* 42:577–685

56. Osher S, Fedkiw R (2002) *Level Set Methods and Dynamic Implicit Surfaces*. Springer, New York
57. Le Pennec E, Mallat S (2000) *Image Compression with Geometrical Wavelets In: IEEE Conference on Image Processing (ICIP), Vancouver, September 2000*
58. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. *IEEE T Pattern Anal* 12(7):629–639
59. Rudin L, Osher S, Fatemi E (1992) Nonlinear Total Variation Based Noise Removal Algorithms. *Phys D* 60:259–268
60. Said A, Pearlman W (1996) A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans Circuits Syst Video Technol* 6(3):243–250
61. Sapiro G (2001) *Geometric Partial Differential Equations and Image Processing*. Cambridge University Press, Cambridge
62. Sapiro G, Tannenbaum A (1993) Affine Invariant Scale-Space. *Internet J Comput Vis* 11:25–44
63. Shapiro J (1993) Embedded Image Coding Using Zerotrees of Wavelet Coefficients. *IEEE Trans Signal Process* 41(12):3445–3462
64. Strang G, Nguyen T (1996) *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley
65. Welland GV (ed) (2003) *Beyond Wavelets*. Academic Press, New York
66. Weickert J (1998) *Anisotropic Diffusion in Image Processing*. ECI Series, Teubner, Stuttgart
67. Xu J, Osher S (2006) Iterative Regularization and Nonlinear Inverse Scale Space Applied to Wavelet Based Denoising. *UCLA CAM Report 06–11, March 2006*

## Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis

KELIN WANG<sup>1,2</sup>, YAN HU<sup>2</sup>, JIANGHENG HE<sup>1</sup>

<sup>1</sup> Pacific Geoscience Centre, Geological Survey of Canada, Sidney, Canada

<sup>2</sup> School of Earth and Ocean Sciences, University of Victoria, Victoria, Canada

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Stable and Critical Coulomb Wedges](#)

[Dynamic Coulomb Wedge](#)

[Stress Drop and Increase in a Subduction Earthquake](#)

[Tsunamigenic Coseismic Seafloor Deformation](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Subduction zone earthquake cycle** Megathrust fault, the interface between the two converging lithospheric

plates at a subduction zone, moves in a stick-slip fashion. In the “stick” phase, the fault is locked or slips very slowly, allowing elastic strain energy to be accumulated in both plates around the fault. Every few decades or centuries, the fault breaks as high-rate “slip” to release the strain energy, causing a large or great earthquake, usually accompanied with a tsunami. An interseismic period and the ensuing earthquake together is called a subduction zone earthquake cycle. The word cycle by no means implies periodicity. Neighboring segments of the same subduction zone may exhibit different temporal patterns of earthquake cycles.

**Accretionary wedge (prism)** At some subduction zones, as one plate subducts beneath the other, some sediment is scraped off the incoming plate and accreted to the leading edge of the upper plate. Because of its wedge shape, the accreted sedimentary body is called the accretionary wedge (or accretionary prism). If all the sediment on the incoming plate is subducted, there is still a sedimentary wedge in the frontal part of the upper plate, but it is usually very small and consists of sediments derived from the upper plate by surface erosion.

**Coulomb plasticity** Coulomb plasticity is a macroscopic, continuum description of the most common type of permanent deformation of Earth materials such as sand, soil, and rock at relatively low temperature and pressure and is widely used in civil engineering and Earth science. In detail, the deformation mechanism is actually brittle shear failure, with or without emitting elastic wave energy. The macroscopic yield criterion is the Coulomb’s law, in which shear strength increases linearly with confining pressure. If the strength does not change with permanent deformation, the material is said to be perfectly plastic. Note that in Earth science the word plasticity is also used to indicate thermally activated creep, but it is very different from the meaning used herein.

**Velocity-weakening and strengthening** These are macroscopic descriptions of dynamic frictional behavior of contact surfaces. Velocity-weakening, featuring a net decrease in frictional strength with increasing slip rate, is the necessary condition for a fault to produce earthquakes. It differs from slip-weakening in that a velocity-weakened fault will regain its strength when the slip slows down or stops. Velocity-strengthening is the opposite of velocity-weakening and is the necessary condition for a fault to resist earthquake rupture. Detailed physical processes on the contact surfaces or within the fault zones controlling their frictional behavior are still being investigated.

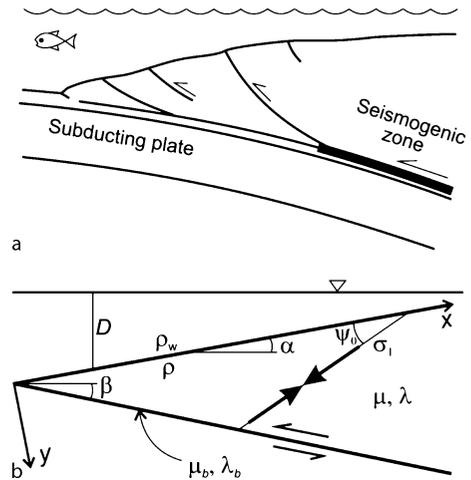
## Definition of the Subject

Mechanics of wedge-shaped geological bodies such as accretionary prisms at subduction zones and fold-and-thrust belts at collision zones is of great scientific interest, mainly because it enables us to use the observed morphology and deformation of the wedge-shaped bodies to constrain properties of the thrust faults underlying them. Davis et al. [12] drew analogy between these wedge-shaped bodies and the sand wedge in front of a moving bulldozer and established a mathematical model. Their work triggered wide applications of wedge mechanics to geology. The fundamental process described in wedge mechanics is how gravitational force, in the presence of a sloping surface, is balanced by basal stress and internal stress. The internal state of stress depends on the rheology of the wedge. The most commonly assumed wedge rheology for geological problems is perfect Coulomb plasticity [12], and the model based on this rheology is referred to as the Coulomb wedge model.

The Coulomb wedge model was designed to explain geological processes of timescale of hundreds of thousands of years. The state of stress is understood to be an average over time, and the wedge is assumed to be in a critical state, that is, uniformly at the Coulomb yield stress. In the application of the model to the sedimentary wedges at subduction zones, attention is now being paid to the temporal changes in stress and pore fluid pressure associated with great subduction earthquakes which have recurrence intervals of decades to centuries. To account for the short-term stress changes, Wang and Hu [50] expanded the Coulomb wedge model by introducing the elastic – perfectly Coulomb plastic rheology. The expanded Coulomb wedge model links long-term geology with coseismic processes and provides a new perspective for the study of subduction zone earthquakes, tsunami generation, frontal wedge structure, and forearc hydrogeology.

## Introduction

Sloping surfaces are commonly dealt with in engineering problems such as dam design and landslide hazard mitigation. In the presence of a sloping surface, materials under gravitational force tend to “flow” downhill and thus generate tensile stress, but whether collapse actually occurs depends on the strength of the material. Materials used in construction can easily sustain stresses caused by the presence of a vertical surface, but a material of no shear strength such as stationary water cannot support any surface slope. A geological wedge, such as the accretionary prism at a subduction zone (see Fig. 1a), overlies a dipping fault, and hence its internal stress is controlled by the



**Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 1**

**a** Schematic illustration of a subduction zone accretionary wedge. **b** Coordinate system used in this article ( $x, y$ ). The surface of the wedge and the subduction fault are simplified to be planar, with slope angle  $\alpha$  and dip  $\beta$ , respectively, and the  $x$  axis is aligned with the upper surface.  $\psi_0$  is the angle between the maximum compressive stress  $\sigma_1$  and the upper surface.  $D$  is water depth.  $\rho$  and  $\rho_w$  are densities of the wedge material and overlying water, respectively.  $\mu, \lambda, \mu_b$  and  $\lambda_b$  are as defined in Eqs. (2) and (3)

strength of the fault as well. If the basal fault is a thrust fault and is strong relative to the strength of the wedge material, the wedge can undergo compressive failure. Wedge mechanics thus consists of two aspects: the frictional behavior of the basal fault and the deformation of the wedge itself. Given a wedge with density  $\rho$ , surface slope angle  $\alpha$ , and basal dip  $\beta$  (see Fig. 1b), Elliot [16], Chapple [8], and later workers all considered the following equations of force balance (exact form varies between publications depending on the assumed coordinate systems)

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} - \rho g \sin \alpha = 0, \quad (1a)$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \rho g \cos \alpha = 0, \quad (1b)$$

where  $g$  is gravitational acceleration,  $\sigma_x$  and  $\sigma_y$  are normal stresses, and  $\tau_{xy}$  is shear stress.

It was Davis et al. [12] who introduced Coulomb plasticity into wedge mechanics. Coulomb plasticity was initially proposed by French engineer Coulomb in 1773 to describe the mechanical strength of soil and sand. The Coulomb failure criterion in its simplest form is

$$\tau = S - \mu(\sigma_n + P) = S - \mu \bar{\sigma}_n, \quad (2)$$

where  $\tau$  is shear strength,  $\sigma_n$  is normal stress,  $S$  is cohesion,

and  $\mu = \tan \varphi$  is the coefficient of internal friction with  $\varphi$  called the internal friction angle.  $P$  is the pressure of fluids present in the pore space between solid grains and in various small fractures and is loosely referred to as the pore fluid pressure. Note that the effective stress  $\bar{\sigma}_n = \sigma_n + P$  is normal stress with  $P$  subtracted. The plus sign is due to the custom in mechanics (except rock mechanics) that compressive stress is defined to be negative, but pressure, although also compressive, is defined to be positive. Here the plane on which the stress is evaluated is oriented in any arbitrary direction, but failure will start on the set of planes that meets the above criterion. This is a generalization of the Coulomb friction criterion for a fixed fault plane

$$\tau_b = S_b - \mu_b(\sigma_{n_b} + P_b) = S_b - \mu_b \bar{\sigma}_{n_b}, \quad (3)$$

where  $S_b$  is the cohesion of the fault,  $\mu_b = \tan \varphi_b$  is its friction coefficient, and  $P_b$  is fluid pressure along the fault.  $S_b$  is usually negligibly small and almost always taken to be zero, and  $\mu_b$  is normally significantly lower than  $\mu$ . A well developed fault such as a plate boundary fault is a zone of finite thickness filled with gouge material, so that the “friction” described by (3) or other friction laws is actually the shear deformation of the gouge in the fault zone, and  $P_b$  is actually pore fluid pressure of the gouge. Fault gouge is often made very weak by the presence of hydrous minerals [5,42], such that the collective strength of the fault zone material is much less than the strength of the rocks on both sides. Another process that may weaken the fault is that the local hydrogeological regime may dynamically maintain  $P_b$  in the fault zone to stay higher than  $P$  on both sides [14]. For both Coulomb plasticity and Coulomb friction, strength increases with depth because of the increasing pressure thus normal stress.

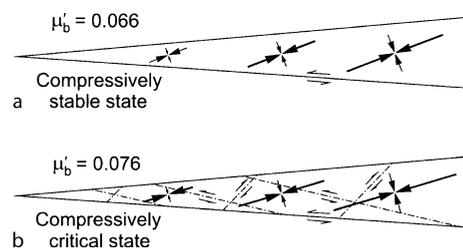
It has long been recognized that Coulomb plasticity, featuring strong depth dependence, applies to the shallow part of Earth’s lithosphere. The most common example is the use of Byerlee’s law of rock friction [6] to describe brittle strength in “Christmas-tree”-like vertical strength-depth profiles of the lithosphere. The Byerlee’s law is an empirical Coulomb friction law. By assuming that faults, i. e., potential failure planes, are oriented in all directions, we regard the brittle part of the lithosphere as being Coulomb plastic. This example also illustrates how a system of numerous discrete structures can be regarded as a continuum at a much larger scale. Similarly, although a geological wedge actually consists of numerous blocks divided by fractures, Coulomb plasticity can be used to describe its overall rheology. However, specific values of friction parameters for submarine wedges may be quite different from those in the Byerlee’s law.

The Coulomb wedge model explains how the ge-

ometry (taper) of the wedge is controlled by the interplay between the gravitational force, the strength of the wedge material, and the strength of the basal fault. The wedge strength and fault strength are both strongly influenced by pore fluid pressure, and the most popular application of the model is to estimate pore fluid pressure from observed wedge geometry. Since the work of Davis et al. [12], more rigorous analytical solutions have been derived [9,10,18,50,55], and some extensions have been proposed, e. g., [2,3,17,54]. Lithospheric scale numerical models are often used to study the evolution of geological wedges including such effects as erosion and sedimentation in collision or subduction zone settings, e. g., [20,53]. Utilizing the bulldozer – sand wedge analogy, important physical insights have been obtained from sandbox experiments [7,11,29,30,31,35,41,52]. For a list of applications of the Coulomb wedge model to subduction-zone accretionary prisms, see [50].

### Stable and Critical Coulomb Wedges

Depending on the state of stress, a Coulomb wedge can be at a critical state, that is, everywhere at Coulomb failure, or a stable (also referred to as supercritical) state, that is, everywhere not at failure (see Fig. 2). The taper of a critical wedge will not change if the stress does not change; note that a critical wedge of stable geometry should not be confused with a stable wedge. Stress solutions have been derived for both critical and stable states. Here we only summarize the simplest, exact solution for a cohesionless wedge ( $S = 0$ ) derived in the coordinate system shown in Fig. 1b, because of the convenience of its application as compared to other solutions. The wedge is assumed to be elastic – perfectly Coulomb plastic. If it is at the critical



### Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 2

An example to show how stresses in an elastic – perfectly Coulomb-plastic wedge, with  $\alpha = 5^\circ$ ,  $\beta = 4^\circ$ ,  $\mu = 0.6$ , and  $\lambda = 0.86$ , are affected by basal friction  $\mu'_b = \mu_b(1 - \lambda_b)$ . Converging arrows represent principal stresses, with the larger one being  $\sigma_1$ . **a** Compressively stable state. **b** Compressively critical state. *Dot-dashed lines* are plastic slip lines (potential failure planes)

state, it obeys (2); if it is at the stable state, it obeys the Hooke's law of elasticity. The basal thrust fault obeys (3). The wedge is assumed to be under water of density  $\rho_w$  and depth  $D$  (a function of  $x$ ; see Fig. 1b). By defining a Hubbert-Rubey fluid pressure ratio within the wedge [9]

$$\lambda = \frac{P - \rho_w g D}{-\sigma_y - \rho_w g D}, \quad (4)$$

and a similar parameter along the basal fault [51]

$$\lambda_b = \frac{P_b - \rho_w g D + \lambda H}{-\sigma_y - \rho_w g D + H}, \quad (5a)$$

where

$$H = \frac{\sigma_y - \sigma_n}{1 - \lambda}, \quad (5b)$$

and assuming  $S_b = 0$ , we can rewrite (3) into

$$\tau_b = -\mu_b \bar{\sigma}_{n,b} = -\mu_b'' \bar{\sigma}_n, \quad (6a)$$

where

$$\mu_b'' = \frac{1 - \lambda_b}{1 - \lambda} \mu_b = \frac{\mu_b'}{1 - \lambda}. \quad (6b)$$

The strength of the basal fault is represented by  $\mu_b' = \mu_b(1 - \lambda_b)$  which always appears as a single parameter and is commonly referred to as the effective friction coefficient. The hydrological process in the fault zone may

differ from that in the wedge and thus cause a sharp gradient in fluid pressure across the wedge base. By allowing  $\lambda_b$  to be different from  $\lambda$ , we use a pressure discontinuity to approximate the sharp gradient. Because stress is continuous across the basal fault, the discontinuity in pore fluid pressure leads to a discontinuity in effective stress. The second equality of Eq. (6a) shows the relation between the effective stress along the fault ( $\bar{\sigma}_{n,b}$ ) and the effective stress just above the fault ( $\bar{\sigma}_n$ ). The establishment of (6) allows the following exact stress solution to be derived. In this expression, all stress components are normalized by  $\rho g y$  (e.g.,  $\bar{\sigma}'_x = \bar{\sigma}_x / \rho g y$ ).

$$\bar{\sigma}'_x = -m(1 - \lambda) \cos \alpha, \quad (7a)$$

$$\bar{\sigma}'_y = -(1 - \lambda) \cos \alpha, \quad (7b)$$

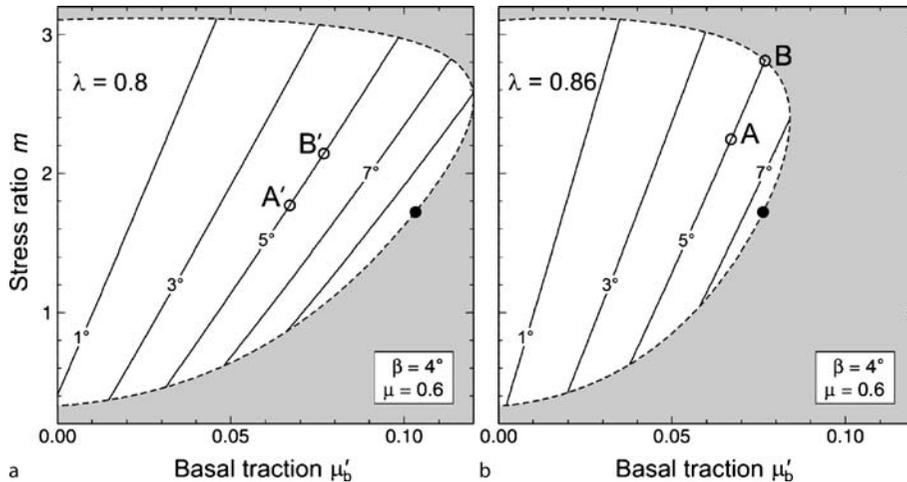
$$\tau'_{xy} = (1 - \rho') \sin \alpha, \quad (7c)$$

where  $\rho' = \rho_w / \rho$ , and the effective stress ratio  $m = \bar{\sigma}_x / \bar{\sigma}_y$  depends on whether the wedge is stable or critical. If the wedge is in a stable state (elastic) [50],

$$m = 1 + \frac{2(\tan \alpha' + \mu_b'')}{\sin 2\theta(1 - \mu_b'' \tan \theta)} - \frac{2 \tan \alpha'}{\tan \theta}, \quad (8)$$

where  $\theta = \alpha + \beta$ , and,

$$\tan \alpha' = \frac{1 - \rho'}{1 - \lambda} \tan \alpha. \quad (9)$$



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 3

Effective stress ratio  $m$  (Eq. (7)) as a function of basal friction  $\mu_b' = \mu_b(1 - \lambda_b)$  for wedges of the same basal dip ( $\beta = 4^\circ$ ) but different surface slope angles ( $\alpha$ ) as labelled on the curves (solid lines). a and b are for two different pore fluid pressure ratio values within the wedge. Each curve is terminated at the extensionally critical state at a lower  $\mu_b'$  and the compressively critical state at a higher  $\mu_b'$ . The end points (connected by a dashed line) outline the stable region (white). No solution exists outside this region. The solid circle marks the state in which the surface slope is at the angle of repose. It divides the line of critical states (dashed line) into the compressive part (above) and extensional part (below). Open circles in b labelled A and B mark the states shown in Fig. 2a and b, respectively. State A' in (a) is for the same wedge with the same basal friction as state A in (b) except for a lower pore fluid pressure ratio, and state B' in (a) corresponds to state B in (b) in the same fashion. Comparison of B with B' shows how an increase in pore fluid pressure weakens the wedge

The angle  $\psi_0$  between the most compressive stress  $\sigma_1$  and the upper surface is uniform (see Fig. 1b). A more general solution for a purely elastic wedge can be found in [23]. If the wedge is in a critical state (perfectly plastic) [9]

$$m = m^c = 1 + \frac{2 \tan \alpha'}{\tan 2\psi_0^c}, \quad (10)$$

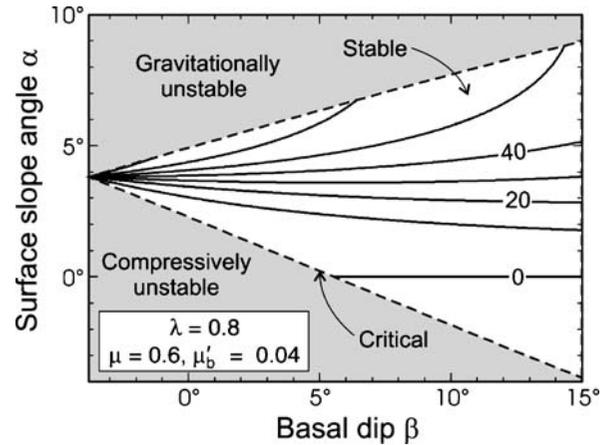
where  $\psi_0^c$  is the value of  $\psi_0$  in the critical state and is given by the following relation

$$\frac{\sin \varphi \sin 2\psi_0^c}{1 - \sin \varphi \cos 2\psi_0^c} = \tan \alpha'. \quad (11)$$

In the above expressions, superscript c indicates critical state. If the wedge has a cohesion that is proportional to depth, (10) and (11) will be modified only slightly [50,55].

A wedge of fixed geometry has two  $m^c$  values. The lower one defines the extensionally critical state, in which the wedge is on the verge of gravitational collapse. This occurs if friction along the basal fault is very low relative to the strength of the wedge material. The higher one defines the compressively critical state, in which the wedge is everywhere on the verge of thrust failure. If  $m$  lies between these two critical values, the wedge is in a stable state and only experiences elastic deformation. A change of basal friction  $\mu'_b$  will cause a change in  $m$  and thus may potentially cause the wedge to switch between the stable and critical states (see Fig. 3). An example of a wedge being in a stable or compressively critical state as controlled by basal friction is shown in Fig. 2. The plastic slip lines (potential failure planes) in the critical wedge (see Fig. 2b) are reminiscent of the out-of-sequence faults in a real accretionary prism (see Fig. 1a).

If we fix the values of all material properties, the critical-wedge solution ( $m = m^c$ ) defines a relation between  $\alpha$  and  $\beta$  representing all possible geometries of a critically tapered wedge (dashed line in Fig. 4). This is a very commonly used diagram, in which the lower branch of the  $\alpha - \beta$  curve represents compressively critical states, and the upper branch represents extensionally critical states. Combinations of  $\alpha$  and  $\beta$  outside of the stability region comprise unstable geometries and cannot exist in steady state. If sedimentary wedges of subduction zones are compressively critical, their observed  $\alpha - \beta$  pairs should line up with the lower branch. However, it has been shown that most of them fall in the stable or even extensionally unstable region of this type of diagram [30,44]. To fit observations, we need to move the lower branch upward by a significant amount. In order to do this, we need to assume either a weaker wedge or a stronger basal fault, or both. This is more simply illustrated by Fig. 3. Given wedge geometry and internal friction, if the state of the wedge is to



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 4

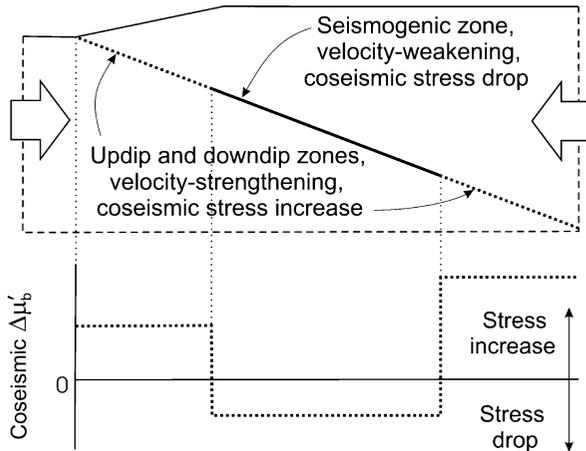
Possible wedge geometry ( $\alpha$  and  $\beta$ ) given material strength and basal friction. The dashed line indicates critical wedge geometry, with the upper and lower branches representing extensionally and compressively critical states, respectively. Contours of  $\psi_0$  (in degrees) are shown in the stable region (white)

be changed from stable to compressively critical, we need to have a higher friction ( $\mu'_b$ ) along the basal fault (i. e., greater stress coupling) and/or higher pore fluid pressure within the wedge (i. e., greater  $\lambda$  weakening the wedge material by reducing effective pressure). Wang and Hu [50] proposed that higher  $\mu'_b$  and  $\lambda$  can occur at the time of a great earthquake and introduced the concept of dynamic Coulomb wedge.

### Dynamic Coulomb Wedge

The concept of dynamic Coulomb wedge is based on the widely recognized frictional behavior of subduction faults. Ignoring the presence of along-strike variations of frictional properties, we can summarize the frictional behavior in a simplified cross-section view (see Fig. 5). The seismogenic zone exhibits velocity-weakening behavior: It weakens in response to high-rate slip, resulting in slip instability, that is, earthquakes. The segments updip and downdip of the seismogenic zone exhibit velocity-strengthening: They strengthen at the time of the earthquake to develop higher stress to resist rupture, but they may slip aseismically after the earthquake to relieve the high stress attained during the earthquake. We assume that the actively deforming sedimentary wedge overlies the updip segment (also see Fig. 1a).

Microscopic mechanisms for the velocity-weakening behavior of the seismogenic zone and the velocity-strengthening behavior of the aseismic zones are subjects of intense research. The aseismic behavior of the deeper



**Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 5**

Schematic illustration of the subduction zone model considered in this work. Large arrows represent interseismic strain accumulation. An earthquake is represented by a sudden decrease in the effective friction coefficient  $\mu'_b$  of the seismogenic zone by  $\Delta\mu'_b$ . Coseismic strengthening of the updip and downdip zones is represented by a sudden increase in their  $\mu'_b$  values

part of any fault is intuitively easy to comprehend; higher temperature at greater depths increasingly enhances viscous deformation and inhibits brittle faulting. There are different physical explanations for the velocity-weakening behavior of the seismogenic zone as summarized in [37]. The mechanism responsible for the velocity-strengthening behavior of the updip segment is yet to be identified, although it is widely accepted that the presence of clay minerals has something to do with it [24,33,34,45,46]. Laboratory experiments indicate that dilatancy of granular fault zone material during fast slip can lead to velocity-strengthening [32]. We think part of the reason for the velocity-strengthening behavior of the updip segment may be its inability to localize into a very thin slip zone. Seismic rupture occurs along very thin slip zones of a few millimeters thickness that are parts of a thicker fault zone [40]. Fast slip of the updip segment, if triggered by the rupture of the deeper seismogenic zone, may tend to drag along fault zone materials over a more distributed band and thus meet greater resistance. This view is different from the velocity-strengthening process described by laboratory-derived rate- and state-dependent friction laws [15,39] in which dynamic changes in the thickness of the slip zone plays no role.

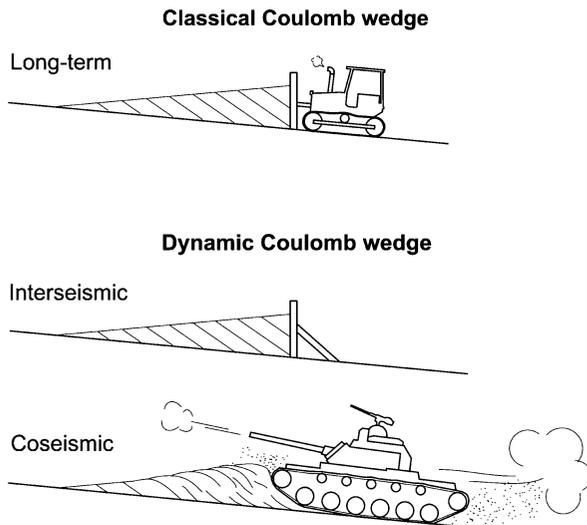
Regardless of the microscopic mechanisms, the downdip variation of the frictional behavior is expected to bring direct consequences to wedge deformation. In an earthquake, thrust motion of the seismogenic zone causes

the frontal wedge to be pushed from behind, and velocity-strengthening of the updip megathrust segment gives rise to higher stress at its base. If the wedge is originally in a stable state, the coseismic strengthening of the basal fault may increase the  $m$  value in (7a) from subcritical to critical. After the earthquake, when the seismogenic zone has returned to a locked state, the stress along the updip segment will relax. The decrease in  $\mu'_b$  leads to a smaller  $m$ , and therefore the frontal wedge returns to a stable state.

If we ignore the change in fluid pressure, the above described process can be seen as the stress ratio  $m$  moving up-and-down along one of the solid lines in Fig. 3 in response to changes in  $\mu'_b$  in earthquake cycles. During a big earthquake, it will hit the upper end of the line ( $m^c$ ). The state of stress for a stable wedge before the earthquake ( $m < m^c$ ) and that for a compressively critical wedge during the earthquake ( $m = m^c$ ) are illustrated by examples in Fig. 2a and b, respectively. Fluid pressure variation should not be ignored, however. For example, the same basal friction as shown in Fig. 2b (also see point B in Fig. 3b) will not drive the wedge to failure if the pore fluid pressure is lower, as shown by point B' in Fig. 3a.

During an earthquake, the sudden compression of the frontal wedge will cause its internal pore fluid pressure to increase, coseismically weakening the wedge material. By comparing Fig. 3a and b, we can see that if the pore fluid pressure ratio in the wedge is higher, the increase in basal friction required to push the wedge into a critical state can be smaller. We may envisage the following scenario. The pore fluid pressure in a frontal wedge may decrease to some degree over the interseismic period due to fluid drainage through fractures and stress relaxation, and the mechanical state of the wedge before an earthquake can be represented by point A' in Fig. 3a as opposed to point A in Fig. 3b. An earthquake not only causes the basal friction to increase but also the pore fluid pressure within the wedge to rise, such that the wedge enters a critical state represented by point B in Fig. 3b. Therefore, the coseismic strengthening of the basal fault and coseismic weakening of the wedge both work toward bringing the wedge to failure.

All previous applications of the Coulomb wedge model to subduction zones assume  $m = m^c$ . The dynamic Coulomb wedge model of [50] explains the meaning of this long-term  $m^c$ : At least as an end-member scenario, it is the value of  $m$  briefly achieved in numerous large earthquakes. The “average” basal stress that determines the wedge geometry in long-term Coulomb wedge models is actually the peak stress achieved at the time of large earthquakes. Thus, the common illustration of the peaceful scene of a bulldozer pushing a sand wedge in classical



### Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 6

Cartoon illustrating the difference between the classical Coulomb wedge model and the dynamic Coulomb wedge model for subduction zone accretionary prisms. In the classical wedge model,  $m = m^c$ . In the dynamic wedge model,  $m < m^c$  in the interseismic period but  $m = m^c$  at the time of a large earthquake. See Eqs. (7) and (10) for definition of  $m$  and  $m^c$

Coulomb wedge papers (see Fig. 6a) should be modified to reflect the unpleasant reality of the world (see Fig. 6b).

### Stress Drop and Increase in a Subduction Earthquake

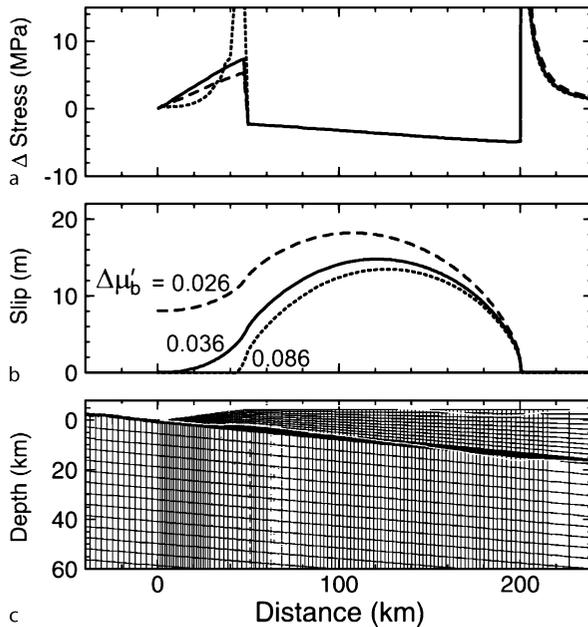
It is important to know the possible amount of stress increase in the frontal wedge for a given earthquake. The increase cannot be arbitrarily large; it is limited by the level of the “push” on the wedge from behind during an earthquake. For this purpose, a numerical model of a larger scale embracing the essential components of the subduction fault as shown in Fig. 5 must be considered, because the stress interaction between the frontal wedge and the material overlying the seismogenic zone cannot be handled by the analytical Coulomb wedge solutions. For an illustration, we consider the following model geometry, representative of most subduction zones. The subduction fault has a constant dip  $\beta = 4^\circ$ . The frontal 50 km of the upper plate has a surface slope  $\alpha = 5^\circ$ , representing the sedimentary wedge, and the rest of the upper plate has a flat surface. We wish to focus on the process of stress transfer from the seismogenic zone to the updip segment during an earthquake, and a static, uniform, and purely elastic model suffices. For simplicity, the effect of pore fluid pressure change on deformation is neglected.

We use a 2D plane-strain finite element model and simulate Coulomb friction along the fault using the method of Lagrange-multiplier Domain Decomposition [48]. The model boundaries are set sufficiently far away so that the model resembles a half-space. For numerical stability, we invoke gravity (assuming a rock density of  $2800 \text{ kg/m}^3$  only when determining yield stress along the fault but exclude it from the deformation calculation. The effect of gravity on coseismic elastic deformation is very small and is neglected in most earthquake cycle deformation models, but gravity is important in the calculation of frictional slip of the fault.

We first generate a pre-stress field by moving the remote seaward and landward model boundaries toward each other against a locked fault. At this stage of “interseismic” strain accumulation, we use a  $\mu'_b$  of 0.04 for the seismogenic and updip segments and 0.004 for the deeper segment. The low strength of the subduction fault is based on the weak fault argument as summarized in Wang and Hu (2006). The nearly zero strength of the deeper part represents a relaxed state after a long time of locking of the seismogenic zone. However, the absolute strength of the fault has no effect on our results. It is the incremental change in fault strength ( $\Delta\mu'_b$ ) at the time of the earthquake that is relevant. A negative  $\Delta\mu'_b$  represents the net effect of weakening, and a positive  $\Delta\mu'_b$  represents the net effect of strengthening. The velocity-dependent evolution of  $\Delta\mu'_b$  through time is not explicitly simulated.

Three examples are shown in Fig. 7. In all cases,  $\Delta\mu'_b = -0.01$  is assigned to the 150-km wide seismogenic zone. This value is chosen to produce an average stress drop of a few MPa (see Fig. 7a), typical of values observed for great subduction earthquakes. The stress drop releases elastic strain energy initially stored in the system, leading to fault slip that represents an earthquake rupture. The deepest segment is assigned a sufficiently large positive  $\Delta\mu'_b$  so that it cannot slip. The examples differ in the  $\Delta\mu'_b$  values assumed for their 50 km wide updip segment, which is the coseismic increase in basal friction in the dynamic Coulomb wedge model.

*Example 1* No trench-breaking rupture (solid line in Fig. 7b). In this case, the strengthening of the updip segment is  $\Delta\mu'_b = 0.036$ . This particular value of  $\Delta\mu'_b$  creates a situation in which the entire updip segment is at failure but is just short of breaking the trench. This is the minimum value of the  $\Delta\mu'_b$  of the updip segment required to prevent trench-breaking rupture and is denoted  $\Delta\mu'_{b_t}$ . The value of  $\Delta\mu'_{b_t}$  depends on the product of the stress drop and the area of the seismogenic zone, a quantity we refer to as “force drop”. That is, if the upper edge of the



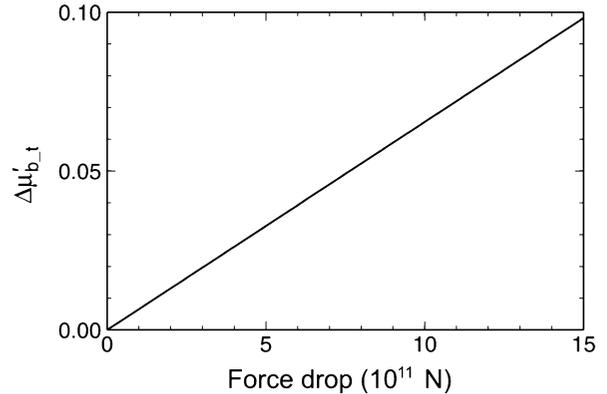
**Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 7**

Three examples of the stress transfer model. The examples differ in the coseismic strength increase of the updip fault segment, indicated in (b) as  $\Delta\mu'_b$ . **a** Shear stress drop (or increase) along the fault. **b** Slip distribution along the fault. **c** Central portion of the finite element mesh (*thin lines*). The “*thick line*” along the plate interface is actually a group of very densely spaced elements. *Thick gray lines* indicate deformed fault and surface after the earthquake (exaggerated by a factor of 2000)

seismogenic zone is fixed, increasing its downdip width or stress drop gives the same result. For the same model geometry as used for this example,  $\Delta\mu'_{b,t}$  as a function of force drop per unit strike length is shown in Fig. 8, assuming the upper edge of the seismogenic zone is fixed. Using a different model geometry or position of the seismogenic zone upper edge will change the slope of this function.

**Example 2** Trench-breaking rupture (dashed line in Fig. 7b). Given the same force drop in the seismogenic zone, if  $\Delta\mu'_b$  of the updip segment is less than  $\Delta\mu'_{b,t}$ , the rupture will break the trench. Knowing whether coseismic trench-breaking rupture exists or is common awaits future seafloor monitoring observations. This example shows that a trench-breaking rupture does not necessarily indicate the updip segment exhibits velocity-weakening. Conceivably, the slip of the velocity-strengthening updip segment may be slower than that of the seismogenic zone and may not generate much seismic waves.

**Example 3** Fully buried rupture (dotted line in Fig. 7b). If  $\Delta\mu'_b$  of the updip segment is greater than  $\Delta\mu'_{b,t}$ , rup-



**Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 8**

Minimum increase in  $\Delta\mu'_b$  of the updip megathrust segment (denoted  $\Delta\mu'_{b,t}$ ) required to prevent trench-breaking rupture as a function of force drop along the seismogenic zone for the model shown in Fig. 7c, with the upper edge of the seismogenic zone fixed at 50 km from the trench

ture may only extend into its lower part. For a very high  $\Delta\mu'_b$ , most of the segment does not slip at all, because a tiny portion immediately updip of the seismogenic zone is sufficient to stop the rupture. This is just the buried-rupture scenario of the crack model commonly used in earthquake simulation [21]. Because most of the updip segment is “protected” and does not experience coseismic stress increase, this scenario is not applicable to the dynamic Coulomb wedge model. A very large stress increase just updip of the seismogenic zone is considered unrealistic.

These examples show the consequences of changes in the strength of the basal fault of the frontal wedge resulting from the rupture of the seismogenic zone. Whether the given strength increase  $\Delta\mu'_b$  can drive the wedge from a stable state into a critical state depends on two factors. First, it depends on the value of  $\mu'_b$  before the earthquake. The value of 0.04 used in the above examples is only one of the numerous possible values. If  $\mu'_b$  is already near a critical value, that is,  $m$  in Fig. 3 for a given  $\alpha$  is already near  $m^c$ , a small increase will do. Conceivably,  $\mu'_b$  before an earthquake may be relatively high if the strengthened state of the fault caused by the previous earthquake has not fully relaxed. Second, given  $\mu'_b$ , it depends on the strength of the wedge material. A weaker wedge becomes critical at a lower  $\Delta\mu'_b$ . The average internal friction value  $\mu$  of an actively deforming frontal prism is lower than the rest of the lithosphere because of its low degree of consolidation and high degree of fracturing. The pore fluid pressure within it, represented by  $\lambda$  in the Coulomb wedge model, may increase due to coseismic compression of the prism,

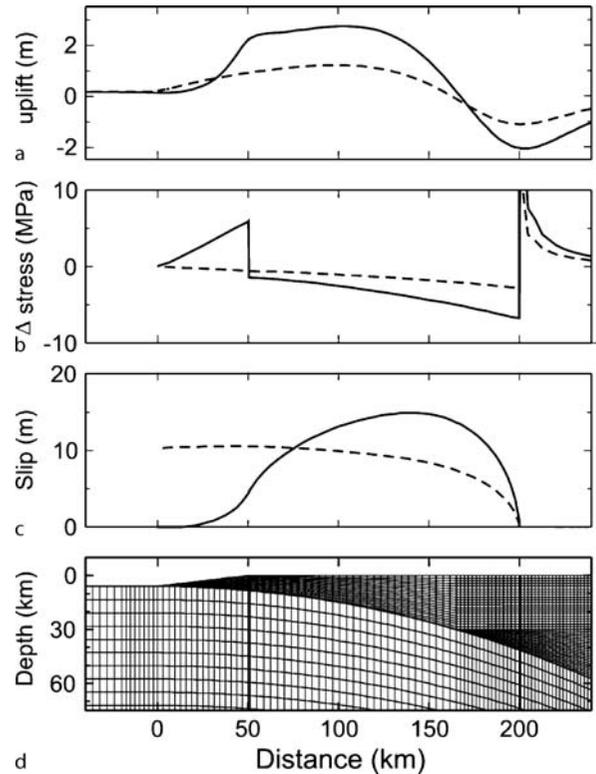
further weakening the wedge, as discussed in Sect. “Dynamic Coulomb Wedge”.

### Tsunamigenic Coseismic Seafloor Deformation

To understand the process of tsunami generation by a great subduction zone earthquake, we must know how the seafloor deforms at the time of the earthquake. Despite the dramatic worldwide improvement of geodetic, seismological, and oceanographic monitoring networks over the past few decades, our knowledge of coseismic seafloor deformation (CSD) is surprisingly poor and is based almost entirely on theoretical models. The problem is the rarity of near-field observations. Except for seafloor pressure sensor records at the time of the M8.2 Tokachi-oki earthquake of 2003 [1] and continuous GPS measurements from islands very near the Sumatra trench at the time of the M8.7 Nias-Simeulue earthquake of 2005 [4,22], most observations are made at sites too remotely located to resolve reliably CSD within about 100 km of the trench. The lack of near-field CSD information causes severe nonuniqueness in the inversion of tsunami, seismic, and geodetic data to determine coseismic slip patterns of the shallow part of the subduction fault, for which the only cure is to introduce a priori constraints on the basis of theoretical models. Until the situation is improved by the establishment of seafloor monitoring systems, we must continue to resort to what we are able to deduce from these models.

The largest uncertainty in our knowledge of the processes that control CSD is how coseismic slip along the subduction fault varies in the downdip direction [21,49]. It may overshadow uncertainties in our knowledge of the timescale of the deformation and the spatial variation of mechanical properties of the rock medium. In comparison, along-strike variations of the coseismic slip, usually described in terms of “asperities”, are much easier to determine using high-density terrestrial monitoring networks. Theoretical models discussed above can help us understand the downdip slip distribution. Using the same type of model as shown in Fig. 5 and Fig. 7 but with a realistic, curved fault, we illustrate how the frictional behavior of the updip segment affects the CSD (see Fig. 9). The simulated earthquakes in the two shown examples have the same “size”, quantified by the seismic moment – the product of rigidity, slip area, and average slip, but they cause very different CSD patterns.

**Example 4** No trench-breaking rupture (solid line in Fig. 9a). This model is similar to the first example in the preceding section in that the updip segment is assumed to strengthen by  $\Delta\mu'_{b,t}$ , and the rupture is on the verge of breaking the trench.



**Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 9**

Two examples showing how the frictional behavior of the updip segment (see Fig. 5) affects CSD. In one example (solid line), the segment strengthens by  $\Delta\mu'_{b,t}$ , and in the other example (dashed line), the segment weakens and thus becomes part of the seismogenic zone. The simulated earthquakes in both examples have the same seismic moment. **a** Surface uplift. **b** Stress drop (or increase) along the fault. **c** Slip distribution along the fault. **d** Central part of the finite element mesh (thin lines). The two vertical “thick lines” at distances 50 km and 200 km bracketing the seismogenic zone of the model of no trench-breaking rupture are actually groups of very densely spaced elements

**Example 5** Full trench-breaking rupture (dashed line in Fig. 9a). In this model, there is no velocity-strengthening updip segment, and the seismogenic zone extends to the trench. The stress drop of the resultant much wider seismogenic zone features a monotonic increase from the trench.

The model of full trench-breaking rupture yields much smaller vertical CSD than does the model of no trench-breaking rupture. The reason is two-fold. First, without a velocity-strengthening updip segment to resist rupture, the maximum slip occurs in the most shallowly dipping near-trench part of the fault where seafloor displacement is predominantly horizontal. This effect would not be

obvious had a straight fault geometry been used in the model. Second, without the resistance of an updip segment, the upper plate does not experience horizontal compression and the resultant vertical expansion. If we rescale the two models so that they have the same maximum slip, the model of full trench-breaking rupture will have a greater seismic moment but still a much lower vertical CSD [49]. This result demonstrates the importance of the frictional behavior of the shallowest fault segment in affecting seafloor uplift. However, it addresses only one aspect of tsunamigenic CSD. Many other factors contribute to tsunami generation. For example, although the full trench-breaking model yields very low vertical CSD, its horizontal CSD may contribute to tsunami generation. If the seafloor slopes at angle  $\alpha$ , its horizontal motion  $D$  in the slope direction should raise the seafloor by  $D \tan \alpha$  relative to a fixed water column above, an effect addressed by Tanioka and Satake [43]. The speed of the coseismic slip is also an important factor in tsunami generation. In some rare cases, the rupture is too low to generate much seismic wave energy yet fast enough to generate rather large tsunamis, giving rise to a class of earthquake called tsunami earthquakes [27].

Elastic deformation of the ocean floor as discussed above is the primary cause of tsunami generation during subduction earthquakes, but inelastic deformation can be locally important. For example, although the lower continental slope of active margins is on average the expression of a critically tapered Coulomb wedge, seafloor topography at these margins is rugged at smaller scales due to sedimentation, erosion, and deformation processes, and where the local surface slope is sufficiently high earthquake shaking may trigger gravitational failure. Such submarine “landslides” may have a locally significant effect on tsunami generation. Another potentially important inelastic process is the coseismic activation of out-of-sequence thrust faults (splay faults) in the accretionary prism (Fig. 1a). Splay faults are much more steeply dipping, and their thrust motion will serve to “redirect” the low-angle slip of the megathrust to a higher angle and thus may greatly enhance local seafloor uplift and contribute to tsunami generation [19,36]. As mentioned in the Introduction, from the continuum perspective, such faulting is a manifestation of Coulomb plasticity. At the local scale, it is actually frictional sliding of a contact surface with elastically deforming rocks on both sides. By comparison of the splay faults schematically illustrated in Fig. 1a and the potential failure planes of the critical wedge in Fig. 2b, we can see that some of the splay faults are oriented optimally for thrust failure if the frontal wedge is compressed during a megathrust earthquake.

## Future Directions

The connection between wedge mechanics and great earthquakes and tsunamis at subduction zones is an emerging new field of study. It leads to challenges in both theoretical development and experimental design and thus excellent research opportunities. We need better constraints on how stresses along different downdip segments of the subduction fault evolve with time throughout an earthquake cycle and how the evolution impacts wedge and seafloor deformation. A number of outstanding questions are to be addressed: Can we constrain the updip limit of the seismogenic zone using wedge morphology? What is the timescale of stress relaxation along the updip segment of the megathrust after an earthquake? Does the seismogenic zone stay locked in the interseismic period? How does pore fluid pressure evolve in an earthquake cycle? How is the transfer of material from the incoming plate to the upper plate (accretion), from the subducted plate to the upper plate (underplating), or from the upper plate to the subducted plate (tectonic erosion) accomplished? What determines the dominant mode of material transfer? What does the spatial change in wedge morphology tell us about changes in the mechanical state of the wedge and the megathrust fault? These questions should be put in the proper context of larger-scale processes such as the viscoelastic relaxation of the mantle following a megathrust earthquake and the deformation of the subducting plate in earthquake cycles [47].

Sandbox experiments designed to study wedge mechanics and dynamic friction experiments designed to study fault mechanics are traditionally separate research activities addressing processes of vastly different timescales. The linkage between subduction earthquakes and submarine wedge evolution suggests the need to combine these experiments. Rapid motion used to simulate earthquakes has begun to be introduced into sandbox experiments [38].

The most promising type of field observation is continuous monitoring of deformation, such as strain and tilt, and fluid pressure using submarine borehole and seafloor observatories. Seafloor elevation change in response to the 2003 Tokachi-oki, northeast Japan, earthquake (M8.2), continuously recorded by two seafloor pressure sensors, clearly indicated coseismic strengthening behavior of the shallowest segment of the subduction fault [1]. Formation fluid pressure changes detected at subsea borehole observatories at the Nankai Trough subduction zone, southwest Japan, have been interpreted to indicate transient aseismic motion of a part of the locked seismogenic zone and/or dynamics of the incoming plate [13]. A number of very-low-

frequency earthquakes have been remotely detected within the Nankai Trough accretionary prism using land-based seismic networks [25], revealing the need for near-field observation using seafloor systems. Submarine monitoring in conjunction with land-based monitoring at subduction zones that are currently in different phases of earthquake cycles will allow us to understand the evolution of fault and wedge stresses during the interseismic period. In this regard, cabled seafloor monitoring networks including borehole observatories, being designed or implemented at different subduction zones [26,28] will surely yield valuable data in the near future.

### Acknowledgments

We thank EE Davis, N Kukowski, SE Lallemand, and K Satake for reviewing the article and providing valuable comments. This work is Geological Survey of Canada contribution 20070221.

### Bibliography

#### Primary Literature

- Baba T, Hirata K, Hori T, Sakaguchi H (2006) Offshore geodetic data conducive to the estimation of the afterslip distribution following the 2003 Tokachi-oki earthquake. *Earth Planet Sci Lett* 241:281–292
- Barr TD, Dahlen FA (1990) Constraints on friction and stress in the Taiwan fold-and-thrust belt from heat flow and geochronology. *Geology* 18:111–115
- Breen NA, Orange DL (1992) The effects of fluid escape on accretionary wedges 1. Variable porosity and wedge convexity. *J Geophys Res* 97:9265–9275
- Briggs RW, Sieh K, Meltzner AJ, Natawidjaja D, Galetzka J, Suwargadi B, Hsu Y-J, Simons M, Hananto N, Suprihanto I, Prayudi D, Avouac J-P, Prawirodirdjo L, Bock Y (2006) Deformation and slip along the Sunda megathrust in the great 2005 Nias-Simeulue earthquake. *Science* 311:1897–1901
- Brown K, Kopf A, Underwood MB, Weinberger JL (2003) Compositional and fluid pressure controls on the state of stress on the Nankai subduction thrust: A weak plate boundary. *Earth Planet Sci Lett* 241:589–603
- Byerlee JD (1978) Friction of rocks. *Pure Appl Geophys* 116:615–626
- Byrne DE, Davis DM, Sykes LR (1988) Local and maximum size of thrust earthquakes and the mechanics of the shallow region of subduction zones. *Tectonics* 7:833–857
- Chapple WM (1978) Mechanics of thin-skinned fold-and-thrust belts. *Geol Soc Am Bull* 89:1189–1198
- Dahlen FA (1984) Noncohesive critical Coulomb wedges: An exact solution. *J Geophys Res* 89:10125–10133
- Dahlen FA, Suppe J, Davis DM (1984) Mechanics of fold-and-thrust belts and accretionary wedges: Cohesive Coulomb theory. *J Geophys Res* 89:10087–10101
- Davis DM (1990) Accretionary mechanics with properties that vary in space and time. In: Debout GE et al. (eds) *Subduction: Top to Bottom*. AGU Monograph 96, Washington, DC, pp 39–48
- Davis DM, Suppe J, Dahlen FA (1983) Mechanics of fold-and-thrust belts and accretionary wedges. *J Geophys Res* 88:1153–1172
- Davis EE, Becker K, Wang K, Obara K, Ito Y (2006) A discrete episode of seismic and aseismic deformation of the Nankai subduction zone accretionary prism and incoming Philippine Sea plate. *Earth Planet Sci Lett* 242:73–84
- Dewhurst DN, Clennell MB, Brown KM, Westbrook GK (1996) Fabric and hydraulic conductivity of sheared clays. *Géotechnique* 46:761–768
- Dieterich JH (1979) Modeling of rock friction: 1. Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Elliot D (1976) The motion of thrust sheets. *J Geophys Res* 81:949–963
- Enlow RL, Koons PO (1998) Critical wedges in three dimensions: Analytical expressions from Mohr–Coulomb constrained perturbation analysis. *J Geophys Res* 103:4897–4914
- Fletcher RC (1989) Approximate analytical solutions for a cohesive fold-and-thrust wedge: Some results for lateral variation in wedge properties and for finite wedge angle. *J Geophys Res* 94:10347–10354
- Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. *J Geophys Res* 84:2303–2314
- Fuller CW, Willett SD, Brandon MT (2006) Formation of forearc basins and their influence on subduction zone earthquakes. *Geology* 34:65–68
- Geist EL, Dmowska R (1999) Local tsunamis and distributed slip at the source. *Pure Appl Geophys* 154:485–512
- Hsu Y-J, Simons M, Avouac J-P, Galetzka J, Sieh K, Chlieh M, Natawidjaja D, Prawirodirdjo L, Bock Y (2006) Frictional afterslip following the 2005 Nias-Simeulue earthquake, Sumatra. *Science* 312:1921–1926
- Hu Y, Wang K (2006) Bending-like behavior of wedge-shaped thin elastic fault blocks. *J Geophys Res* 111; doi:10.1029/2005JB003987
- Hyndman RD, Wang K (1993) Thermal constraints on the zone of major thrust earthquake failure: The Cascadia subduction zone. *J Geophys Res* 98:2039–2060
- Ito Y, Obara K (2006) Dynamic deformation of the accretionary prism excites very low frequency earthquakes. *Geophys Res Lett* 33; doi:10.1029/2005GL025270
- Juniper K, Bornhold B, Barnes C (2006) NEPTUNE Canada community science experiments. *Eos Transactions, American Geophysical Union* 87(52), Fall Meeting Supplement: Abstract OS34F-04
- Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Interior* 6:246–259
- Kaneda Y (2006) The advanced dense ocean floor observatory network system for mega-thrust earthquakes and tsunamis in the Nankai Trough – precise real-time observatory and simulating phenomena of earthquakes and tsunamis. *Eos Transactions, American Geophysical Union* 87(52), Fall Meeting Supplement: Abstract OS34F-01
- Kukowski N, von Hune R, Malavieille J, Lallemand SE (1994) Sediment accretion against a buttress beneath the Peruvian continental margin at 12° S as simulated with sandbox modeling. *Geol Rundsch* 83:822–831
- Lallemand SE, Schnürle P, Malavieille J (1994) Coulomb theory

- applied to accretionary and nonaccretionary wedges: Possible causes for tectonic erosion and/or frontal accretion. *J Geophys Res* 99:12033–12055
31. Lohrmann J, Kukowski N, Adam J, Oncken O (2003) The impact of analogue material properties on the geometry, kinematics, and dynamics of convergent sand wedges. *J Struct Geol* 25:1691–1711
  32. Morone C (1998) Laboratory-derived friction laws and their application to seismic faulting. *Annu Rev Earth Planet Sci* 26:649–696
  33. Moore DE, Lockner DA (2007) Friction of the smectite clay montmorillonite: A review and interpretation of data. In: Dixon T, Moore JC (eds) *The Seismogenic Zone of Subduction Thrust Faults*, Columbia University Press, New York
  34. Moore JC, Saffer D (2001) Updip limit of the seismogenic zone beneath the accretionary prism of southwest Japan: An effect of diagenetic to low grade metamorphic processes and increasing effective stress. *Geology* 29:183–186
  35. Mourgues R, Cobbold PR (2006) Thrust wedges and fluid overpressures: Sandbox models involving pore fluids. *J Geophys Res* 111; doi:10.1029/2004JB003441
  36. Park JO, Tsuru T, Kodaira S, Cummins PR, Kaneda Y (2002) Splay fault branching along the Nankai subduction zone. *Science* 297:1157–1160
  37. Rice J (2006) Heating and weakening of faults during earthquake slip. *J Geophys Res* 111; doi:10.1029/2005JB004006
  38. Rosenau M, Melnick D, Brookhagen B, Ehtler HP, Oncken O, Strecker MR (2006) About the relationship between forearc anatomy and megathrust earthquakes. *Eos Transactions, American Geophysical Union* 87(52), Fall Meeting Supplement: Abstract T12C-04
  39. Ruina A (1983) Slip instability and state variable friction laws. *J Geophys Res* 88:10359–10370
  40. Sibson RH (2003) Thickness of the seismic slip zone. *Bull Seismol Soc Am* 93:1169–1178
  41. Smit JHW, Brun JP, Sokoutis D (2003) Deformation of brittle-ductile thrust wedges in experiments and nature. *J Geophys Res* 108; doi:10.1029/2002JB002190
  42. Takahashi M, Mizoguchi K, Kitamura K, Masuda K (2007) Effects of clay content on the frictional strength and fluid transport property of faults. *J Geophys Res* 112; doi:10.1029/2006JB004678
  43. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. *Geophys Res Lett* 23:861–864
  44. von Huene R, Ranero CR (2003) Subduction erosion and basal friction along the sediment-starved convergent margin off Antofagasta, Chile. *J Geophys Res* 108; doi:10.1029/2001JB001569
  45. Vrolijk P (1990) On the mechanical role of smectite in subduction zones. *Geology* 18:703–707
  46. Wang CY (1980) Sediment subduction and frictional sliding in a subduction zone. *Geology* 8:530–533
  47. Wang K (2007) Elastic and viscoelastic models of subduction earthquake cycles. In: Dixon T, Moore JC (Eds) *The Seismogenic Zone of Subduction Thrust Faults*, Columbia University Press, New York
  48. Wang K, He J (1999) Mechanics of low-stress forearcs: Nankai and Cascadia. *J Geophys Res* 104:15191–15205
  49. Wang K, He J (2007) Effects of Frictional Behaviour and Geometry of Subduction Fault on Coseismic Seafloor Deformation. *Bull Seismol Soc Am* 98:571–579
  50. Wang K, Hu Y (2006) Accretionary prisms in subduction earthquake cycles: The theory of dynamic Coulomb wedge. *J Geophys Res* 111; doi:10.1029/2005JB004094
  51. Wang K, He J, Hu Y (2006) A note on pore fluid pressure ratios in the Coulomb wedge theory. *Geophys Res Lett* 33; doi:10.1029/2006GL027233
  52. Wang WH, Davis DM (1996) Sandbox model simulation of forearc evolution and noncritical wedges. *J Geophys Res* 101:11329–11339
  53. Willett S, Beaumont C, Fullsack P (1993) Mechanical model for the tectonics of doubly vergent compressional orogens. *Geology* 21:371–374
  54. Xiao HB, Dahlen FA, Suppe J (1991) Mechanics of extensional wedges. *J Geophys Res* 96:10301–10318
  55. Zhao W L, Davis DM, Dahlen FA, Suppe J (1986) Origin of convex accretionary wedges: Evidence from Barbados. *J Geophys Res* 91:10246–10258

### Books and Reviews

- Dahlen FA (1990) Critical taper model of fold-and-thrust belts and accretionary wedges. *Annu Rev Earth Planet Sci* 18:55–99
- Dixon T, Moore JC (eds) *The Seismogenic Zone of Subduction Thrust Faults*, Columbia University Press, New York
- Scholz CH (2003) *The Mechanics of Earthquakes and Faulting*, 2nd edn. Cambridge University Press, Cambridge, 471 p

---

## World Wide Web, Graph Structure

LADA A. ADAMIC  
School of Information and the Center for the Study of Complex Systems, University of Michigan, Ann Arbor, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Graph Structure  
Algorithms  
Subsets of the Web Graph  
Future Directions  
Bibliography

### Glossary

**Graph** A set of nodes (vertices) connected by links (edges, arcs). In the Web graph, the nodes are webpages, and the edges are the hyperlinks between them.

**Indegree** The number incoming edges to a node; in the case of the Web, it is the number of webpages pointing to a page.

**Outdegree** The number of outgoing edges; in the case of the Web, it is the number of webpages a webpage points to.

**Strongly connected component** A set of webpages such that any page can be reached from any other page by following hyperlinks.

**Weakly connected component** A set of webpages such that any page can be reached from any other page by treating the hyperlinks as undirected.

**URL** A unique resource locator that corresponds to an online information source.

**Hypertext transfer protocol (HTTP)** The communications protocol that allows web clients to communicate with web servers.

**Web server** A program that responds to requests for web pages.

**Webpage** An information resource, identified by a URL, that is usually but not necessarily in the HTML (Hypertext Markup Language) format. A webpage may be *static*, meaning that it is stored on the server as a document, or *dynamic*, meaning that it is generated dynamically at the point that it is requested by the browser, using scripts and/or back-end databases.

**Domain** A name that identifies one or more IP addresses, e. g. umich.edu.

**Top level domain** The suffix of the domain name, sometimes corresponding to the purpose of the website or the country of origin for the website, e. g. “edu”, “com”, “gov”, “uk”, “cn”, etc.

**Website** A collection of webpages that is hosted on one or more web servers and that share a common root URL, e. g. “www.springer.com”.

**Randomized network** A network that preserves the degrees of each node relative to the original, but the edges themselves are rewired.

## Definition of the Subject

In the period of a few short years, the World Wide Web has become the primary information source. Billions of webpages, tied together by hyperlinks, constitute the world’s largest publicly accessible store of data, and are accessed by hundreds of millions of people worldwide. Understanding how the webpages are connected to form the Web graph is important for building comprehensive and accurate information retrieval systems, as well as characterizing real-world phenomena that are reflected in its content and link structure.

## Introduction

The Web is becoming the single most indispensable source

of information. A massive graph of billions of pages, the Web is continuously growing and evolving. People use the Web to do everything from reading the news, socializing, making purchases, paying their bills, watching videos, playing games, and finding romance. Given how pervasive the Web is in our daily lives, it is hard to believe that it is a relatively recent invention.

The World Wide Web was invented by Tim Berners Lee in 1990 at CERN in Geneva, Switzerland. Berners’ idea was to combine hypertext – the linking of documents via terms – with the internet, allowing users to navigate between documents that are distributed and accessible across the globe. Starting with the first site created at CERN in 1991, the Web grew to 130 sites by 1993 [1]. By 1997 there were around a million sites [1], and by 2006, they numbered 100 million, about half of them active [2].

Such massive growth is only possible through the distributed contribution by many individuals. Thus online information is not contributed by a few vetted experts, nor is access limited to a few individuals. Rather, almost anyone can contribute to online content, especially with the proliferation of easy-to-use web authoring tools. One might expect then that there is a great deal of randomness in how pages are linked to one another, and how users navigate those links. Contrary to this, there are a number of strong regularities both in the structure of the Web and the pattern of access by its users. These regularities can be discovered when one studies the graph structure of the Web.

## Graph Structure

The Web is a graph. The webpages are the nodes or vertices, and the hyperlinks are the directed links or edges between them. Characterizing the Web graph has led to improved crawling, sampling, and ranking algorithms. It has also led to novel insights about the underlying human dynamics that are reflected in the link patterns of the Web.

## Empirical Measurements

Given the enormity of the Web graph, only a handful of projects, WebBase [3] and WebGraph [4,5] among them, have sought to gather comprehensive crawls gathering hundreds of millions of the web’s hundreds of billions of pages and have made them available to the research community. Early characterizations of the Web graph were made from Web crawls made by commercial search engines [6,7], and major search engine companies continue to host a variety of research on the Web graph [8,9,10,11].

**Degree Distributions** One of the most striking features of the Web graph that was discovered early on is the power

law distribution of incoming hyperlinks. While billions of webpages, comprising a substantial portion of the web, receive few or no hyperlinks, a few very popular pages have attracted millions. More precisely, the distribution of incoming links is a power-law (also known as a Zipf's law or Pareto distribution) [12,13,14]

$$p(k) \sim k^{-\alpha} \quad (1)$$

The indegree distribution has shown remarkable consistency across many different studies, having an exponent  $\alpha$  of 2.1 [7,15]. The outdegree distribution has a steeper power law exponent measured at  $\alpha = 2.5$  [15] and  $\alpha = 2.7$  [7]. The steeper power law exponent corresponds to a smaller variance in the number of hyperlinks stemming from a web page as opposed to pointing to it; while a very useful or informative webpage such as a major search engine or news site can attract millions of hyperlinks, it is rather impractical, due to limitations in content length, for a single webpage to contain a very large number of hyperlinks.

Power law degree distributions have been shown to not be preserved for some subsamples of the Web graph, however. Pennock et al. [16] found that for sets of company, university, newspaper, and scientist homepages, distributions of inlinks deviated for power-laws, especially in the number of pages of low degree. Intuitively, one might expect a university homepage, or a newspaper homepage to attract at least a few links, producing a peak in the distribution away from the minimum. This is in contrast to power law distributions, which are scale-free, appearing the same on all scales, e. g. 10 to 100 links or 100 to 1000 links, and having no "typical" degree.

Much research has gone into deriving generative models that can produce the observed degree distributions. We will return to this topic in Subject. "Generative Models".

**Degree Correlations** Degree correlation, also termed assortativity, characterizes to what extent high degree nodes link to other high degree nodes. In the case of the Web graph, a question may be whether pages that receive many links, link to other pages with high indegree. In an undirected version of the nd.edu webpage network, the webpages were found to be mildly disassortative with a correlation of  $-0.065$  in the combined undirected degree of vertices [17]. This webpage network, crawled in 1999, consists of 325,729 documents and 1,469,680 hyperlinks. We can resolve this correlation further by considering pairwise the in and outdegrees of both the page that is giving the link, and the one that is being linked to. Using the nd.edu domain data set, we find that there is a slight negative correlation between both the indegree and

World Wide Web, Graph Structure, Table 1  
Assortativity of webpages in the nd.edu domain

		to	
		indegree	outdegree
from	indegree	$-0.023$	$0.256$
	outdegree	$-0.062$	$-0.014$

outdegree of the webpage containing the hyperlink and the indegree of the page it is pointing to. This indicates that there is a slight tendency of link-rich pages to link to less link-rich pages and less popular pages, and link-poor pages to link to link-rich and popular pages. Interestingly, in this domain, there was a strong *positive* correlation between the indegree of the linking page and the outdegree of the linked page. This may be an indication that the most cited pages themselves link to link-rich pages and so serve a funneling function. The above is summarized in Table 1. The correlation between the pages' own indegree and outdegree was positive ( $\rho = 0.244$ ), indicating that link rich pages themselves tended to receive more links.

**Connected Components and Bow-Tie Structure** The bow-tie structure describes the connectivity of the Web graph as a whole [7]. The middle knot of the bow-tie is the largest strongly connected component (LSCC). The component is called strongly connected because within it, any page can be reached from any other by following directed links. The OUT component consists of those pages that can be reached from the LSCC, but do not have paths leading into the LSCC. The IN component consist of those pages that have paths leading into the LSCC, but are not reachable from the LSCC. The remainder of the graph is composed of tubes, tendrils, and islands. Tubes connect the IN component to the OUT component, bypassing the LSCC. Tendrils are those pages that can either be reached from the IN component, or lead to the OUT component, but are not in tubes or the LSCC itself. Islands are connected components of vertices that are not connected via any links to the other components.

Each of these four regions (LSCC, IN, OUT, and the rest) of the Web's bow-tie accounts for roughly a fourth of the entire graph. This implies that if two pages  $A$  and  $B$  are selected at random, there is only roughly a 1 in 4 chance that there is a directed chain of hyperlinks leading from  $A$  to  $B$ . The reasoning is the following: most often, if there is a path from  $A$  to  $B$ ,  $A$  is either in the IN or LSCC components, while  $B$  is in the LSCC or OUT components. This is true in approximately  $1/2 \times 1/2 = 1/4$  of the cases. By comparison, about 90% of the webpages are

in the largest connected component if the links are treated as undirected.

The bow-tie structure was also observed to exist for subsets of webpages corresponding to a given topic, so that the web appears self-similar on several scales [18]. This structure has important implications for a crawler attempting to traverse the web in its entirety. By starting at a single webpage, the crawler will likely not be able to reach significant portions of the Web graph by following directed hyperlinks. Instead, it may adopt a strategy of discovering web pages to start from by scanning IP addresses to see if they are hosting a web server.

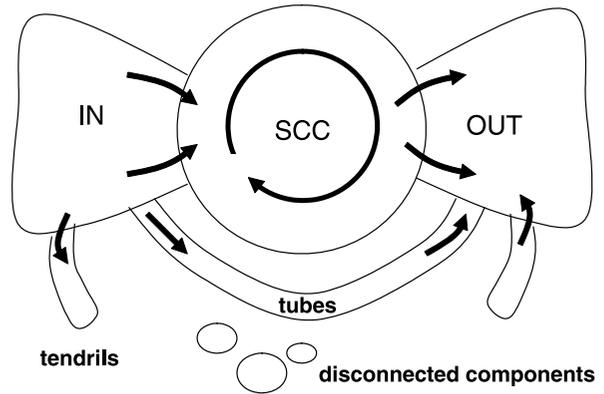
**Shortest Paths and Diameter** In modeling user behavior and web crawler performance, it is useful to know not just whether  $B$  is reachable via a directed path of hyperlinks from  $A$ , but how many hyperlinks must be traversed. Albert et al. [19] found for the above mentioned nd.edu domain crawl, that the average shortest path over all pairs of vertices was 11.2. This closely matched the value of 11.6 predicted by their preferential attachment model, which will be described in Subsect. “Preferential Attachment and the BA Model”. The model produces an average shortest path  $\langle d \rangle$  of

$$\langle d \rangle = 0.35 + 2.06 \log(N) \quad (2)$$

The model likewise provides a close prediction of 17.5 for a graph of the size of the sample of 203 million nodes described by Broder et al. [7]. The measured shortest path between the 24% of the pairs for this sample that were *reachable* was 16. As discussed in Subsect. “Connected Components and Bow-Tie Structure”, the other three quarters of the pairs of web pages have no directed path connecting them, and hence their distance is infinite. On the other hand, treating the graph as undirected produces an average shortest path of just 6 hops for the 90% of the webpages that are connected.

Another quantity used to describe the width of the Web is the diameter, the maximal shortest path between any two connected vertices. Broder et al. [7] measured a diameter of at least 28 in the central core and over 500 for the graph as a whole, indicating that while most reachable pairs can be connected in just a small number of hops, others are far removed.

**Reciprocity, Clustering and Motifs** Reciprocity is simply a measure of how often when one webpage links to another, that page links back. The measure should take into account the expected number of reciprocated links based simply on the link density  $\bar{a} = m/N/(N-1)$ , where  $N$  is



World Wide Web, Graph Structure, Figure 1

A schematic of the bow-tie structure of the web

the number of nodes and  $m$  is the number of edges. An unbiased form of the reciprocity measure is given by

$$\rho = \frac{m_{\text{bd}}/m - \bar{a}}{1 - \bar{a}} \quad (3)$$

where  $m_{\text{bd}}$  is the number of bidirectional links [20]. For the crawl of pages in the nd.edu domain,  $\rho = 0.52$  [19].

Clustering or transitivity quantifies how often webpages that are linked to a common page are linked to each other as well. The measure, usually applied to an undirected version of the network, is given by the following equation.

$$C = \frac{6 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \quad (4)$$

A 1998 Alexa crawl of 260,000 websites yielded a clustering coefficient of 0.014 at the site level (0.11 using the Watts Strogatz clustering coefficient [35]), significantly higher than a randomized version of the network [6].

The clustering coefficient captures the degree of transitivity in the graph, but does not reveal the configuration of those triads. Triad motifs [21] shed further insight into the local structure of the Web graph. There are 13 triad motifs, and together their abundance compared to randomized versions of the web graph constitute a *triad significance profile* [22]. The significance profile SP is a normalized vector

$$\text{SP}_i = \frac{Z_i}{(\sum_i Z_i^2)^{1/2}} \quad (5)$$

of  $z$  scores measuring the statistical significance of the number  $n_i^{\text{web}}$  of observed motifs relative to the randomized counterparts of the network

$$Z_i = \frac{n_i^{\text{web}} - \langle N_i^{\text{rand}} \rangle}{\sigma_i^{\text{rand}}} \quad (6)$$

where  $\langle N_i^{\text{rand}} \rangle$  and  $\sigma_i^{\text{rand}}$  are the mean and standard deviation of the frequency of occurrence of motif  $i$  in randomized versions of the network.

The motif profile of Web graphs is like that of social networks, with closed triads and reciprocated links more frequent than motifs without closure and with unreciprocated links. The most underexpressed profile is that of a webpage that has reciprocated links with two other webpages, but those two webpages do not link to one another. The most highly expressed motif is that of the fully reciprocated closed triad. The probability of such a triad occurring in a random graph is very small, but it occurs rather frequently in the Web graph where clusters of related pages link to one another reciprocally.

### Time Evolution

The web is growing exponentially, with pages being continuously added, modified, and deleted. Addition and deletion clearly affect which nodes are present in the graph, while modifications in content [23,24] possibly result in link addition and deletion. Cho and Garcia-Molina [25] found that 30% of a sample of 720,000 pages disappeared in an interval of a month. Douglis et al. [26] studied 950,000 web page requests from a server, and found that for pages that were accessed more than once, 60% had changed one or more of their hyperlinks. Making the task of discovering new pages easier is the fact that 85–95% of new pages appear on existing sites [27].

### Web Surfing

An important aspect of the Web graph is how it is navigated. Huberman et al. [28] discovered the *Law of Web Surfing*, that describes the distribution in the number of links a user will access on a particular website. This distribution was found to hold for web surfing patterns of AOL users across a million web sites, users accessing the Xerox website, and students, faculty, and staff at Georgia Tech's College of Computing. The underlying process was argued to be the following. The value  $V_L$  of the  $L$ th page in a click stream where  $L$  links were followed is assumed to be related to the previous page, but also to deviate from it in a random way.

$$V_L = V_{L-1} + \epsilon_L \quad (7)$$

where  $\epsilon_L$  is a normally distributed random variable. The individual will continue to surf until the expected discounted value of information to be found on future pages is less than the cost (in time) of continuing. This maps to an option pricing model in finance, and the distribution of the number of links a user follows before the expected

utility of continuing falls below a threshold is given by

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} \exp\left[\frac{-\lambda(L - \mu)^2}{2\mu^2 L}\right] \quad (8)$$

with a mean number of clicks  $E(L) = \mu$  and variance  $\text{Var}(L) = \mu^3 \lambda$ ,  $\lambda$  being a scale parameter. This heavy tailed distribution accurately describes users' surfing behavior on the Web graph: many follow just a link or two, while a few explore dozens or even hundreds of links in a single session.

### Crawling

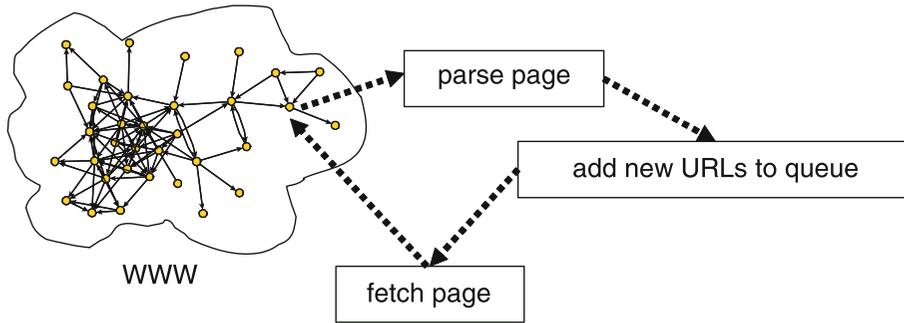
The web is only partly navigated through following hyperlinks. Users jump from one part of the Web graph to another by querying search engines and selecting one or more pages that match their query. In order to deliver web pages matching particular user queries, search engines must first crawl the Web graph. Web crawlers have two tasks: revisiting previously crawled pages in order to refresh them and discovering new pages by following new links. The crawler needs to prioritize both the order in which it will crawl the new content and how often it will recrawl pages it has previously crawled, in part to discover links to new content [27].

**Search Engine Coverage** It is not possible for search engines to crawl the entire Web. There are infinite websites that will return valid and unique content for an infinite number of URLs. Still, it is possible to explore the overlap in search engine coverage of the content that should be covered by the search engines. In 1998 Lawrence and Giles set out to estimate the size of the “indexable” web by issuing queries to multiple search engines [29]. By measuring the overlap in the search results and assuming that each search engine samples the web independently, they estimated the total size of the Web at that point to be at least 320 million pages, with no single search engine achieving more than a third of the coverage.

### Generative Models

Generative models help describe the underlying processes that are shaping the Web, and therefore can be used to make predictions about the future evolution of its properties.

**Preferential Attachment and the BA Model** The World Wide Web inspired the Barabasi–Albert preferential attachment model of network growth [15,30], that has since been applied to networks in many other domains, including biological, social, and technological. As discussed in



World Wide Web, Graph Structure, Figure 2

A basic schematic of a crawler traversing webpages. As new pages are discovered, they are parsed, and their links added to the queue of pages to be visited if they had not been seen before

Subject. “Degree Distributions” Albert et al. observed that the distribution of incoming hyperlinks for webpages was highly skewed. They devised a very simple model of preferential attachment, which had previously been explored in other networked and non-networked contexts [31,32,33], wherein a new page is added at each timestep  $t$ , and attaches to  $m$  existing webpages, each with probability  $\Pi(k_i)$  proportional to that page’s current degree  $k_i$ .

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \quad (9)$$

An intuitive argument for why this is so is that the more popular a web resource is, the more likely an author of a new webpage is to learn about it in the course of browsing and searching the Web, and the more likely she is to cite it.

The links are treated as undirected, and the model yields a scale-free indegree distribution with a power law exponent  $\alpha = 3$ , independent of  $m$ . While reproducing some characteristics of the Web graph, such as a scale free distribution and short average path lengths, the model does not capture others. The power law exponent of 3 is much higher than the consistently observed  $\alpha = 2.1$  for the Web. Its clustering coefficient depends on the size of the graph as  $N^{-0.75}$ , and is much smaller than that of the Web graph. It also predicts that the degree of each node depends on its time of introduction  $i$  as follows

$$k_i(t) = m \left( \frac{t}{t_i} \right)^\beta, \quad \beta = \frac{1}{2}. \quad (10)$$

However, there is only very weak correlation between the age of a webpage or website on the web and its indegree [34].

Extensions of the Barabasi–Albert model were proposed to overcome its limitations. Rewiring of existing edges with the endpoint being chosen preferentially pro-

duces power law distributions with cut-offs as well as exponents in the range  $2 \leq \alpha < \infty$ . Other models have introduced aging factors, where webpages lose attractiveness over time [36], which produced cutoffs in the power law distributions. In the fitness model, webpages have a different inherent attractiveness  $\eta_i$  [37], with preferential attachment combining this attractiveness with the degree  $k_i$  so that  $\Pi_i \sim \eta_i * k_i$ . Interestingly, it was found that non-linear preferential attachment [38] where

$$\Pi(k_i) \sim k_i^\nu \quad (11)$$

does not always produce power law degree distributions when  $\nu \neq 1$  [39].

**Copying Models** The basic preferential attachment models do not explicitly explain why a new webpage would link to existing webpages in proportion to their indegree. Such models also do not capture features such as clustering, since the only factor influencing the likelihood of being linked to is the indegree of the page. In copying models, both clustering and preferential attachment are produced [40,41,42,43] as a result of links being copied, without content being explicitly taken into account. New pages are added over time, and each page selects an existing page, termed a *prototype* or *ambassador* page, and copies several of the links on that page. The process leads to preferential attachment, since the probability that any given page is linked to is proportional to the probability that it is the neighbor of the ambassador node, which is in turn proportional to its indegree. In the case where the new page links to the ambassador page in addition to copying some of its links, the process also produces clustering; each copied link results in a closed triad: the new page  $c$  links to the ambassador page  $a$ ,  $a$  links to  $b$ , and  $c$  copies that link and also links to  $b$ . In the case where the new page does not link to the ambassador page, the process produces bi-cliques, sets of pages that link to a second set of pages.

Kumar et al. [42] constructed a stochastic copying model where new pages are created linearly or exponentially in time. Each page has a fixed outdegree. With probability  $p_r$ , the  $i$ th link will point to a webpage at random, otherwise it will copy the  $i$ th link of the ambassador node. In the case of linear addition of pages in time, the probability  $p(k, t)$  that a node has degree  $k$  at time  $t$  is given by a power law:

$$p(k, t) \sim tk^{-(2-p_r)/(1-p_r)} \quad (12)$$

where the power-law exponent  $\alpha$  of the degree distribution is bounded from below by 2 in the case of pure copying, and grows steeper as the proportion  $p_r$  of random links is increased. The model also produces a large number of bipartite cliques, commensurate with empirically observed proportions.

Related to the copying model is the forest fire model, where the copying continues recursively to all nodes that are encountered by following links [43]. The process is analogous to a forest fire which spreads from tree to tree. Two probabilities are assigned. With forward burning probability  $p$ , the outgoing links from page are copied. With probability  $q$  incoming links are copied. This last step models something akin to a webpage author searching for other pages that link to the current page and deciding to link to them as well.

1. As in the copying model, a new node chooses an ambassador node uniformly at random from the set of all nodes
2. Two geometrically distributed random numbers,  $x$  and  $y$  are generated with means  $p/(1-p)$  and  $q/(1-q)$  respectively.  $x$  out-links and  $y$  in-links are copied from the current node (if  $x$  or  $y$  are greater than the total number, all the links are copied).
3. Step 2 is applied recursively to each of the nodes the new node now has links to. Cycles are disallowed.

Although the forest fire model is so far analytically intractable, it has several properties that match the web. Heavy-tailed indegree distributions occur because in copying links from other nodes, new nodes are attaching preferentially to existing nodes in proportion to their indegree. Heavy-tailed outdegree distributions are produced by the cascading process of recursive link copying. Many cascades will terminate already with the first node, but some will continue for many steps. Community structure and clustering are produced, as in the basic stochastic copying model, because the link copying leads to triadic closure, as a new webpage links to a page and some of its neighbors, and those neighbors' neighbors, etc.

**Content-Based Models** The above generative models have not explicitly modeled the textual content of the webpages. The intuition behind content-based or similarity-driven models is that webpages will not only link to popular pages, but also to topically related webpages (or both). Menczer [44] used a cosine similarity measure

$$s(p_1, p_2) = \frac{\sum_k w_{kp_1} w_{kp_2}}{\sqrt{(\sum_k w_{kp_1}^2) (\sum_k w_{kp_2}^2)}} \quad (13)$$

where  $w_{kp}$  is some weight function, typically, a TFIDF (term frequency, inverse document frequency) weight for the term  $k$  in page  $p$ . In 110,000 webpages sampled from the Open Directory, Menczer found that lexical similarity closely correlates with webpages being clustered (either linking directly or overlapping in their neighbors). He proposed a content based generative network model to take into account that once two pages exceed a threshold overlap ( $s > s^*$ ) in their content, their probability of being linked levels off. Below that threshold, the probability decreases as a power-law, so that while neighboring pages tend to be lexically similar, there is a long tail of lexically distant pages that are neighbors in the Web graph. As in the Barabasi–Albert model, at each time step  $t$  a new page is added and links to  $m$  existing pages. But rather than linking to any page in proportion to their degree, it will only do so below a threshold lexical similarity. Above that threshold the probability that a new page appearing at time  $t$  will link to an existing page  $i$  with degree  $k(i)$  will be proportional to the lexical similarity:

$$Pr(p, t) = \begin{cases} \frac{k(i)}{mt} & \text{if } s(p_i, p_t) > s^* \\ c_1 ([s(p_i, p_t)]^{-1} - 1)^{-\alpha} & \text{otherwise.} \end{cases} \quad (14)$$

The model produces a degree distribution that deviates from a pure powerlaw, but closely matches the sample webpages from the Open Directory.

## Algorithms

### Ranking

Most search engines utilize the structure of the Web graph to rank their search results. Some structural node properties, such as indegree, are purely local and utilize only the nearest neighbors of the webpage. Others, such as PageRank, utilize the entire graph, and yet converge within a short number of iterations.

**Indegree** The simplest link based ranking of webpages is based on their indegree,

$$s(k) \sim k \quad (15)$$

where each inlink serves as a popularity vote. This measure is susceptible to spamming. One can create webpages and websites with relative ease whose primary purpose is to boost the ranking of a particular target page.

### PageRank

The PageRank algorithm was developed by Larry Page, one of the co-founders of the Google search engine [45]. PageRank is less susceptible to spamming because it takes into account not just how many pages link to a particular target page, but also how many pages link to those pages etc. It does so by simulating a random walk on the Web graph. This PageRank score represents the proportion of time a random web surfer would spend at each webpage, traversing the graph by following random links from a succession of web pages. Roughly speaking, PageRank corresponds to the probability distribution given by the principal eigenvector of the normalized webgraph adjacency matrix. To prevent the random walk from being trapped in a region of the graph (remember that only a fraction of the pages is in the LSCC), a damping factor  $d$  (also termed a teleportation probability) is introduced. With probability  $d$  the surfer follows a random link from the page she finds herself on. With probability  $(1 - d)$  the surfer jumps to a random node in the Web graph. Let  $A$  be the adjacency matrix for a Web graph with  $n$  pages. Then the matrix  $B$  corresponding to the Markov chain is given by

$$b_{ij} = d \frac{a_{ij}}{\sum_i a_{ij}} + \frac{1-d}{n}. \quad (16)$$

Rather than solving for the eigenvector exactly, one can obtain a close approximation by applying a power iteration technique, iteratively multiplying an initial vector  $x$  by  $B$

$$x = Bx. \quad (17)$$

The convergence of the iterations depends on the initial vector and the damping factor, but is tractable even for very large web crawls. The damping factor is typically set to  $d = 0.85$ . For most applications, PageRank values are precomputed for all the webpages at once, and are combined with a textual match score to a user's query and other factors in determining the final ranking of webpages on any given search engine result page. Although PageRank is designed to take into account more than direct links, studies have found it to be highly correlated with indegree when applied to the Web graph [46].

Personalization of PageRank can be achieved by biasing the random jump toward pages a particular user has visited already, rather than all webpages uniformly at ran-

dom [47]. Since pages that link to one another are likely to be topically similar, the personalized PageRank algorithm will boost the scores of pages in the areas of the web the user has previously explored, and hence bias the search results toward pages of topical interest to the user.

**Application to Unbiased Sampling** During a crawl, the probability of encountering a link to a page is roughly proportional to the PageRank of that page, since the PageRank represents the random walk probability of finding the page. Therefore crawling will produce a biased sample where popular pages are more likely to be included. In order to create an unbiased sample, one can simply select pages in inverse proportion to their estimated PageRank [48].

### HITS

The HITS (Hypertext Inducted Topic Selection) algorithm was developed by Jon Kleinberg, and starts with a smaller subset of query results, rather than with the entire web graph [49]. Instead of than assigning a webpage a single score, as PageRank does, HITS assigns two scores, a hub score  $y$  and authority score  $x$ . A good authority is a page with authoritative information, and a good hub points to good authorities. Hub may constitute good starting points for a user to explore an information space, while authorities are likely to contain definitive information.

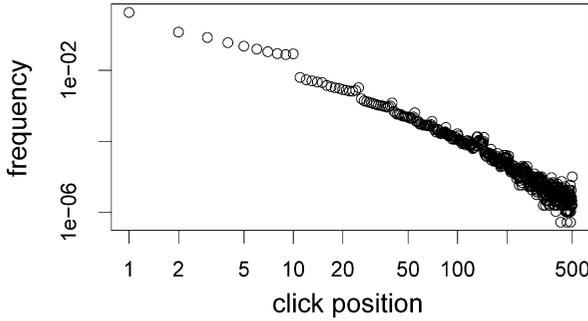
In the first step of the HITS algorithm, an initial root set of webpages (e. g. 200 pages) is retrieved based on a textual match. Some of these pages, containing text relevant, though not necessarily authoritative, on the subject of the query, can be expected to link to prominent authorities on the subject, and some will be linked to by hubs. This root set is therefore expanded to a base set that includes (up to a cutoff of e. g. 1000–3000) pages linked to by the root set and linking to the root set. At this point, links between pages located on the same website are removed, as they are likely to serve a navigational purpose and don't necessarily confer authority.

Much like the PageRank scores, hub scores  $x$  and authority scores  $y$  can be computed iteratively, after being initialized to uniform constants. For a page  $j$ , the value  $x_j$  is updated to be the sum of  $y_i$  over all the pages  $i$  that link to  $j$ .

$$x_j = \sum_{i \text{ such that } i \rightarrow j} y_i \quad (18)$$

Similarly, the hub score  $y_i$  of a page  $i$  is updated according to the authority scores  $x_j$  of the pages it points to.

$$y_i = \sum_{j \text{ such that } i \rightarrow j} x_j \quad (19)$$



World Wide Web, Graph Structure, Figure 3

Frequency of clicks on different search result positions in 20 million queries issued by 600,000 users of the AOL search engine, from March to May 2006 [50]

The above can also be computed using a power iteration technique, with the authority score  $x$  corresponding to the principal eigenvector of  $A^T A$ , while  $y$  corresponds to the principal eigenvector of  $A A^T$ ,  $A$  being the adjacency matrix of the base set:

$$\begin{aligned} x &\leftarrow A^T y \leftarrow A^T A x \\ y &\leftarrow A x \leftarrow A A^T y. \end{aligned}$$

Unlike PageRank, where the scores for all pages are precomputed and subsequently used for all queries, HITS is computed at query-time.

### Search Engine Bias

Since most search engines utilize a link-based ranking algorithm to order their search results, there is the possibility that they further bias web traffic toward already popular sites on a given topic. As shown in Fig. 3, 90% of the search results users click on are on the first page, with 43% being the very top search result. In part the higher rate of click-throughs for highly ranked results is due to users' tendency to scan content in order, but in part it is also due to search engines' ability to rank relevant content successfully. This may inhibit newer, but high quality pages, or pages presenting a different perspective, from receiving attention from search engine users.

While search engines drive traffic to the most heavily linked sites, the search engine users spread themselves out much more widely by searching for a wide variety of queries. In a 2005 study, Fortunato et al. [51] found that a majority of queries to Google return less than 30,000 hits (representing less than one millionth of the entire set of the web crawled by Google at that time). Hence, while some web pages have tens of thousands and even millions of in-links, they do not grab all of the search engine users' attention.

### Finding Communities in the Web Graph

Webpages on similar topics naturally form densely connected regions of the Web graph, termed communities. Various algorithms have been developed to discover web communities based on the link structure alone, without considering the textual content or starting from a given seed set. Kleinberg et al. [52] *trawled* for emerging cyber-communities by identifying bipartite cores: sets of  $i$  pages that all cite  $j$  authoritative pages. Of thousands of such cores identified, the vast majority corresponded to groups of related pages. This approach has been extended to finding larger, but incomplete bipartite cores [53]. Communities can also be identified by performing the HITS expansion step from a root set that is comprised of a core [54].

Flake et al. [55] used maximum-flow minimum-cut algorithms to discover communities, the definition of a community being such that any page within the community has more links within the community than outside of it. The Web graph is augmented by creating an artificial source  $s$  with infinite capacity edges routed to all seed vertices  $S$  for which one would like to find a community. All preexisting edges are made bidirectional and rescaled to a heuristically chosen constant  $k$ . A virtual sink is connected to all vertices except the seed vertices and virtual source. A maximum flow procedure is used to produce a residual-flow graph (the graph of edges that have excess capacity). All vertices that are accessible from  $s$  through the residual graph satisfy the definition of community.

Many other network based clustering algorithms exist, some of which have been applied to subsets of the Web. Some, based on the concept of modularity, divide the network into communities so as to maximize the within-community edges compared to what one would expect for a random assignment of pages to communities. The modularity can be expressed as:

$$Q = \frac{1}{4m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \quad (20)$$

where  $A$  is the adjacency matrix,  $m$  is the total number of edges in the network,  $k_i$  is the degree of vertex  $i$  and  $c_i$  is the community  $i$  is assigned to. The modularity can be maximized through a greedy hierarchical algorithm that scales up to millions of nodes [56], a simulated annealing approach [57], or by directly computing the eigenvectors of the modularity matrix  $B$

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}. \quad (21)$$

Beyond identifying topically related sets of pages, community finding algorithms may be applied to identify spam

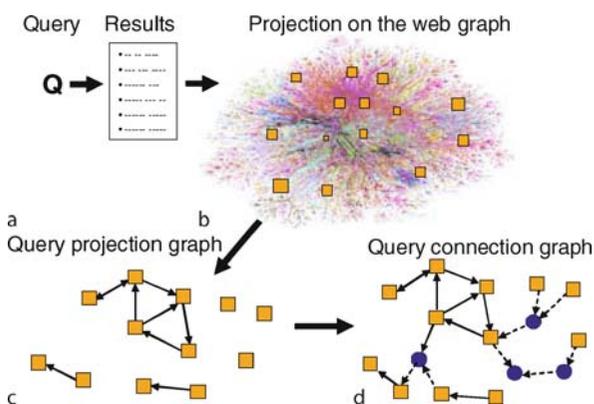
hosts and pages, since pages within so called link farms tend to link to one another much more often than to legitimate sites [9,58].

### Subsets of the Web Graph

The Web contains various interesting subgraphs, reflecting different topics, and social and collaborative activity. Here we discuss the graph properties of just a few of those subsets.

#### Query Connection Graphs

Subgraphs corresponding to search engine results for given queries can be informative both regarding the quality of the search results and the likelihood that a user will reformulate the query, making it more specific or more general [59]. The process of constructing query connection graphs is shown in Fig. 4. A *projection graph* is constructed, consisting of the search results and the hyperlinks between them. A *connection graph* is constructed by adding the minimum number of other webpages necessary to reconnect the projection graph. One can also construct the two graphs on the domain level, where a domain encompasses all of e.g. “springer.com” or “umich.edu”, and two domains share an edge if a page in one domain links to a page in the other. As one might expect, domain query projection graphs are much more densely linked than projection graphs on individual webpages. A complete domain graph of the web from February 2006 contained 39 million domain names and 720 million directed edges, with an average shortest path of 4 and the largest



World Wide Web, Graph Structure, Figure 4

The process of constructing a query connection graph. A search engine is used to retrieve web pages relevant to the query. The pages are projected onto the Web graph. The projection subgraph contains only the search results and the hyperlinks between them. The connection subgraph also includes the nodes (webpages) that are minimally needed to reconnect the pages

weakly connected component containing 99% of the nodes. Whether at the page or domain level, when compared with human ratings of the quality of a set of pages for a given query, the sets that had more tightly connected projection graphs (and correspondingly smaller connection graphs) were good predictors of quality of the pages.

The success of a user’s search is only in part due to the accuracy of the search engine ranking algorithm. Another component is the skill of the user in formulating a query. By looking at the query subgraphs of issued queries, one can predict whether a user will reformulate a query. Queries with result sets that are more tightly knit and contain a few central high degree nodes are less likely to be reformulated. On the other hand, if a query projection graph lacks central, high degree nodes, the search engine user is more likely to reformulate the query. However, if the largest connected component constitutes e.g. 10 of the top 20 search results, the query may be too specialized, and the user is more likely to reformulate the query to be more general.

#### Weblogs

A rapidly growing part of the Web are weblogs (blogs), webpages containing time-stamped posts. Blogs are interconnected through direct links from one post to another, which may occur when bloggers discuss the same topic. Blogs also frequently include a sidebar of links to other blogs and websites. Millions are used simply as personal journals, but many specialize in particular topics, such as politics or technology. Some blogs today attract as much attention as mainstream media sites [60]. Kumar et al. [61] studied the profiles of 1.3 million blogs on livejournal.com, one of the most popular blogging sites. Livejournal gives its bloggers the opportunity to specify other bloggers who are their friends, basic demographic information, and to join groups based on their interests. The network is highly clustered, with a clustering coefficient of  $C = 0.2$  (meaning that 20% of the time friends of the same blogger are friends themselves). This high clustering is consistent with the observation that 70% of the connections correspond to one of three factors: shared interest, same 5 year age group, and geographic location.

The geographic location information of livejournal users allowed for a very large scale study of the small world phenomenon [62]. It had been known for decades, ever since the 1960s, when Stanley Milgram conducted his famous small world experiment [63], that people are able to form short chains of acquaintances to reach a target individual using only information about their immediate network neighbors. Subsequently, Kleinberg [64] proved that

for certain spatial network topologies, individuals would be able to use a simple greedy algorithm, choosing an acquaintance closest to the target as the next person in the chain, to connect to targets in a short number of steps. This holds true for LiveJournal, where the probability of two LiveJournal users being connected is inversely proportional to the number of people who lived between them, and this spatial distribution of connections allows for successful formation of small world chains [62].

Blogs are continuously being updated and this temporal evolution is one of their most interesting aspects. Citations patterns between blogs change rapidly, even on the scale of a few months [76]. Kumar et al. [65] found that as blogging activity grew in popularity from 2002 to 2003, the addition of links became bursty – indicating that activity was centered around communities and events relevant to them. The degree distribution appeared to be power-law and asymptoting to a power-law exponent of 2.1. This inequality in link distribution was noted for the subset of political blogs [66,67] in the United States. The subgraph of political blogs also showed a strong tendency of blogs to link predominantly to blogs of similar political leanings [68,69].

The time resolved nature of blog entries enable researchers to track individual *memes*, topics that are contagiously spreading as information cascades from blog to blog. Several models [70,71] were developed to identify topics and their spread through the blog network. In a general cascade model [71] of the directed and weighted transmission graph of blogs, the weight of each edge corresponds to the probability that a blog both reads and copies an item from another blog, and this weight is learned through a machine learning algorithm from observed mentions by the two blogs of previous topics. By tacking 7,000 topics over a set of blogs, Kempe et al. [71] found that the amount of information flowing through blogs corresponded to the ranking of those blogs by blog search engines. Similarly, ordering the edges by the transmission probability on each edge closely corresponded to blogs that were directly linked through their blogroll links (links to other blogs that occur on the sidebar of the page). Adar et al. [72] used the presence of direct links between blogs, and overlap in text and links previously cited, to predict the path of information flow among a set of blogs mentioning the same URL. Leksovec et al. [73] developed an algorithm for near optimal selection of a set of blogs such that information cascades are detected quickly. Because of the skewed nature of participation and attention in the blogosphere, only a small fraction of blogs need to be monitored in order to capture a large portion of the activity.

Although blog content is highly distributed, and to a certain extent collaborative, with blogs filtering and generating news at rates often outpacing the mainstream media, the content is highly redundant, often inaccurate, and poorly interlinked. In contrast, Wikipedia is an example of collaboratively generated content that is accurate, non-redundant, and carefully structured.

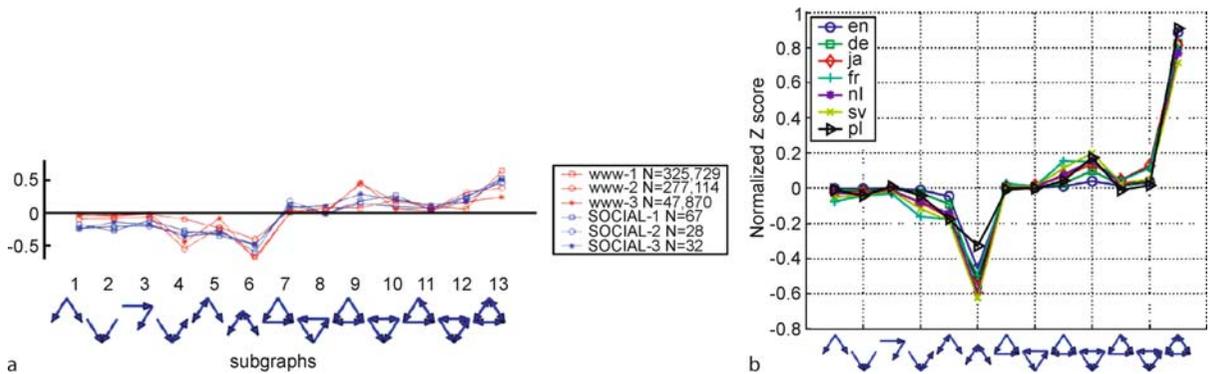
## Wikipedia

Wikipedia is a web-based encyclopedia that is collaboratively written by volunteers all around the world. Since its creation in 2001, Wikipedia has grown rapidly into one of the most prominent reference sources, its pages frequently appearing among the top search engine results. As of August 2007, there were 75,000 active contributors, with nearly 2 million articles in English alone. A study of the 30 largest language Wikipedias in January 2005 yielded degree distribution exponents ( $\gamma_{in} = 2.15 \pm 0.13$  and  $\gamma_{out} = 2.57 \pm 0.27$ ) reflective of the web overall.

Unlike the web, however, between 85 and 97 percent of the pages in the various languages belong to the strongly connected component [74] – indicating that Wikipedia content is more consistently bound than the WWW at large. It is also demonstrating densification, with the number of links growing superlinearly with the number of Wikipedia pages as  $L \sim N^\alpha$ , with  $\alpha = 1.14 \pm 0.05$ . Similar growth patterns, with more links than nodes being added over time, have been observed in the physical internet, patent and scientific citation networks [75] as well as blog citation networks [76].

The Wikipedia networks display a high degree of reciprocity, where one article linking to another frequently corresponds to another linking back. Using the unbiased mutual reciprocity measure specified in Eq. (3), the degree of reciprocity for the different language Wikipedias is  $\rho = 0.32 \pm 0.05$ , lower than for the crawl of pages in the nd.edu domain ( $\rho = 0.52$ ) [19,20]. The Wikipedia also exhibits a high degree of clustering, with related articles linking together in triads. Specifically, the motif profile closely follows that of the web overall and falls in the same triad significance superfamily.

Beyond its structural characteristics, Wikipedia has been an ideal medium to validate the preferential attachment model. It was used to verify that the probability of acquiring new links is in fact proportional to the number of links already present on the page [77]. Such time resolved data had previously been difficult to obtain, but Wikipedia provides a complete history of edits for download. Two interesting observations were made. For pages with moderate numbers of inlinks and outlinks, the proportion is



World Wide Web, Graph Structure, Figure 5

**a** Web motif profile **b** Wikipedia motif profile. The motif profiles of several web samples and social networks compared to profiles found in the different language Wikipedias. The x axis depicts the possible triads, and the y axis represents normalized z-scores comparing the frequency of each motif to randomized versions of the networks

slightly sublinear, but close to unity ( $\nu = 0.9$  in Eq. (11)). Second, the probability reaches a peak, where pages that already contain many links are again not as likely to add more, and those that are already heavily referenced do not tend to attract more links. This could represent both mature pages and mature topics, where the links have already been established.

## Future Directions

### Web 2.0

The Wikipedia and the blogosphere are just two of the many exciting developments in Web 2.0. Web 2.0 refers to collaborative and interactive applications that are fast becoming integrated into the Web. Rather than simply being consumers of information that they retrieve on the web, web users now can easily contribute content through tools such as blogs and wikis. Unlike simple web pages, that are generated by a single author or through a script from a database, Web 2.0 allows users to interact with one another by modifying pages. For example, bloggers can interact with one another by leaving comments on each others posts. Wikipedians interact by collaboratively editing the same Wikipedia entries, and participating in discussions on talk pages dedicated to those entries. Many content providers for text, images, and video allow users to provide reviews in the form of ratings and comments on the content, or to respond by adding content of their own.

In addition to reflecting a change in the way that web content is generated, Web 2.0 sites allow the users to consume web content in a collaborative way. Through RSS feeds, users receive updates of web content rather than having to remember to visit individual sites. They can tag pages with keywords for easy subsequent retrieval,

not just for themselves, but for others interested in the same topic. Through collaborative filtering, users can learn about products, news articles, and media that their friends enjoyed. The best recommenders need not be friends however. Collaborative tools such as RSS readers, bookmarking websites, and content providers can automatically identify users with like interests and make recommendations based on items that those users enjoyed. With Web 2.0 technologies, one can view the Web not only as a proliferation of information and content, but also as giant, distributed machinery that allows the most relevant, personalized content to bubble to the top. At the same time that the Web may have been producing a problem through over-production and over-participation, it has solved it through the emergence of a collection of exceedingly simple yet effective filtering tools.

At this stage, it is unclear whether social networking sites, which allow users to explicitly specify who their friends are, will become platforms that other content providers will be integrated into, or whether many different websites will be able to integrate an open, portable social network the users can bring with them. In either case, the Web graph may soon appear rather different, with webpages becoming mere containers of segmented, time-stamped, personalized content. One of the latest trends in combining content is that of mashups. One can use a mashup to overlay, for example, rental property listings or restaurant reviews on a neighborhood map. These mashups will be especially useful on location and user-aware devices.

As our lives become easier to track, from the sites we browse, to the people we know, to our physical movements tracked by cellphones, we may find our lives increasingly caught in the Web. Most people will readily trade

such information for improved access to personalized and geo-specific services and information. It does, however, also raise interesting privacy concerns, since the Web is a highly distributed storage and dissemination medium.

### Mobile Web

Another forcing function for Web innovation is the fact that many individuals access the Web from small mobile devices. For many individuals in developing countries, this may be their primary way of accessing the Web. This has led to the development of web services tailored to the mobile Web.

Access patterns from mobile devices have been shown to follow the same regularities as general web browsing [78]. This holds despite the fact that Web access from such devices presents challenges due to limited download speed and small display size [79]. It also presents opportunities, as users are able to contribute content wherever they may find themselves. Moblogging (mobile blogging) involves updating one's blog instantaneously with images and videos as they are captured. Currently several news sites encourage the contribution of such materials. The mobile Web also allows users to access Web content, such as maps, traffic and tourist information when they are on the go. Again this may change the structure of the Web graph, as its content is delivered in small, personalized chunks to a variety of devices. Even so, one would expect its main features to remain: billions of hyperlinked pieces of information, showing strong regularities that are a reflection of the real world that is now inextricably tied to the Web.

### Bibliography

#### Primary Literature

1. Gray M (1997) Web growth summary. [www.mit.edu/~mkgray/net/web-growth-summary.html](http://www.mit.edu/~mkgray/net/web-growth-summary.html). Accessed 9 March 2008
2. Walton M (2006) Web reaches new milestone: 100 million sites. [www.cnn.com/2006/TECH/internet/11/01/100millionwebsites/index.html](http://www.cnn.com/2006/TECH/internet/11/01/100millionwebsites/index.html). Accessed 9 March 2008
3. Cho J, Garcia-Molina H, Haveliwala T, Lam W, Paepcke A, Raghavan S, Wesley G (2006) Stanford webbase components and applications. *ACM Trans Inter Tech* 6(2):153–186
4. Boldi P, Vigna S (2004) The webgraph framework I: compression techniques. In: *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, pp 595–602
5. Boldi P, Codenotti B, Santini M, Vigna S (2004) UbiCrawler: a scalable fully distributed Web crawler. *Softw Pract Experience* 34(8):711–726
6. Adamic LA (1999) The Small World Web. *Proc ECDL* 99:443–452
7. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the Web. *Comput Netw* 33(1–6):309–320
8. Joshi A, Kumar R, Reed B, Tomkins A (2007) Anchor-based proximity measures. In: *Proceedings of the 16th International World Wide Web Conference*, Banff, Canada
9. Fetterly D, Manasse M, Najork M (2004) Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In: *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*, Paris, France, pp 1–6
10. Zhang B, Li H, Liu Y, Ji L, Xi W, Fan W, Chen Z, Ma WY (2005) Improving web search results using affinity graph. In: *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, pp 504–511
11. Broder AZ, Lempel R, Maghoul F, Pedersen J (2006) Efficient PageRank approximation via graph aggregation. *Inf Retr* 9(2):123–138
12. Adamic LA, Huberman BA (2002) Zipfs law and the internet. *Glottometrics* 3:143–150
13. Mitzenmacher M (2003) A brief history of generative models for power law and lognormal distributions. *Internet Math* 1(2):226–251
14. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46(5):323–351
15. Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. *Science* 286(5439):509
16. Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL (2002) Winners don't take all: Characterizing the competition for links on the web. *Proc Natl Acad Sci* 99(8):5207
17. Newman MEJ (2002) Assortative Mixing in Networks. *Phys Rev Lett* 89(20):208701
18. Dill S, Kumar R, McCurley KS, Rajogopalan S, Sivakumar D, Tomkins A (2002) Self-Similarity In the Web. *ACM Trans Internet Technol* 2(3):205–223
19. Albert R, Jeong H, Barabási AL (1999) Internet: Diameter of the world-wide web. *Nature* 401:130–131. doi:10.1038/43601
20. Garlaschelli D, Loffredo MI (2004) Patterns of link reciprocity in directed networks. *Phys Rev Lett* 93(26):268701
21. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298(5594):824–827
22. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of designed and evolved networks. *Science* 303(5663):1538–1542
23. Brewington BE, Cybenko G (2000) How dynamic is the Web? *Computer Netw* 33(1–6):257–276
24. Fetterly D, Manasse M, Najork M, Wiener JL (2004) A large-scale study of the evolution of Web pages. *Softw Pract Experience* 34(2):213–237
25. Cho J, Garcia-Molina H (2003) Estimating frequency of change. *ACM Trans Internet Technol* 3(3):256–290
26. Douglass F, Feldmann A, Krishnamurthy B, Mogul J (1997) Rate of change and other metrics: a live study of the world wide web. *USENIX Symposium on Internet Technologies and Systems*, vol 119
27. Dasgupta A, Ghosh A, Kumar R, Olston C, Pandey S, Tomkins A (2007) The discoverability of the web. In: *Proceedings of the 16th International Conference on World Wide Web*, Banff, Canada, pp 421–430
28. Huberman BA, Pirolli PLT, Pitkow JE, Lukose RM (1998) Strong Regularities in World Wide Web Surfing. *Science* 280(5360):95

29. Lawrence S, Giles CL (1998) Searching the World Wide Web. *Science* 280(5360):98
30. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74(1):47–97
31. Simon HA (1955) On a Class of Skew Distribution Functions. *Biometrika* 42(3/4):425–440
32. Yule GU (1925) A Mathematical Theory of Evolution, Based on the Conclusions of Dr. JC Willis, FRS. *Philos Trans Roy Soc Lond Ser B, Containing Papers of a Biological Character* 213:21–87
33. Price DS (1976) A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inf Sci* 27(5–6):292–306
34. Adamic LA, Huberman BA (2000) Power-law distribution of the world wide web. *Science* 287(5461):2115a
35. Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393(6684):440–442
36. Albert R, Barabási A-L (2000) Topology of Evolving Networks: Local Events and Universality. *Phys Rev Lett* 85(24):5234–5237
37. Dorogovtsev SN, Mendes JFF (2000) Scaling behaviour of developing and decaying networks. *Europhys Lett* 52(1):33–39
38. Bianconi G, Barabasi AL (2001) Competition and multiscaling in evolving networks. *Europhys Lett* 54(4):436–442
39. Jeong H, Neda Z, Barabasi AL (2003) Measuring preferential attachment in evolving networks. *Europhys Lett* 61(4):567–572
40. Vázquez A (2003) Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys Rev E* 67(5):56104
41. Jackson MO, Rogers BW (2007) Meeting Strangers and Friends of Friends: How Random Are Social Networks? *Am Econ Rev* 97(3):890–915
42. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E (2000) Stochastic models for the web graph. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, p 57
43. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. *ACM Trans Knowl Discov Data (TKDD)* 1(1)
44. Menczer F (2002) Growing and navigating the small world Web by local content. *Proc Natl Acad Sci* 99(22):14014–14019
45. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. *Comput Netw ISDN Syst* 30(1–7):107–117
46. Litvak N, Volkovich Y, Donato D (2007) Determining factors behind the pagerank log-log plot. *Lecture notes in computer science* 4863:108
47. Haveliwala Taher H (2002) Topic-sensitive pagerank. In: WWW ‘02: Proceedings of the 11th International Conference on World Wide Web, New York, pp 517–526
48. Henzinger MR, Heydon A, Mitzenmacher M, Najork M (2000) On near-uniform URL sampling. *Comput Netw* 33(1–6):295–308
49. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. *J ACM* 46(5):604–632
50. Pass G, Chowdhury A, Torgeson C (2006) A picture of search. In: *Infoscale ‘06, Hong Kong*. In: Proceedings of the 1st international conference on Scalable information systems. ACM, New York, pp 1
51. Fortunato S, Flammini A, Menczer F, Vespignani A (2006) Topical interests and the mitigation of search engine bias. *Proc Natl Acad Sci* 103(34):12684
52. Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The Web as a Graph: Measurements, Models, and Methods. *Computing and Combinatorics: 5th Annual International Conference, Cocoon’99, Tokyo, Japan*
53. Dourisboure Y, Geraci F, Pellegrini M (2007) Extraction and classification of dense communities in the web. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, pp 461–470
54. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Extracting large-scale knowledge bases from the web. In: Proceedings of the 25th Very Large Data Bases Conference, Edinburgh, UK, pp 639–650
55. Flake GW, Lawrence S, Giles CL, Coetzee FM (2002) Self-organization and identification of Web communities. *Computer* 35(3):66–70
56. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. *Phys Rev E* 70(6):66111
57. Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *J Stat Mech Theor Exp* 9:P09008
58. Castillo C, Donato D, Gionis A, Murdock V, Silvestri F (2007) Know your neighbors: Web spam detection using the web topology. In: Proceedings of SIGIR, pp 423–430. ACM Press, Amsterdam
59. Leskovec J, Dumais S, Horvitz E (2007) Web projections: learning from contextual subgraphs of the web. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, pp 471–480
60. Anderson C (2006) *The long tail*. Hyperion, New York
61. Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and evolution of blogspace. *Commun ACM* 47(12):35–39
62. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. *Proc Natl Acad Sci* 102(33):11623–11628
63. Travers J, Milgram S (1969) An experimental study of the small world problem. *Sociometry* 32:425–443
64. Kleinberg J (2000) Navigation in a small world. *Nature* 406(6798):845
65. Kumar R, Novak J, Raghavan P, Tomkins A (2005) On the bursty evolution of blogspace. *World Wide Web* 8(2):159–178
66. Hindman M, Tsioutsoulis K, Johnson JA (2003) *Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web*. Annual Meeting of the Midwest Political Science Association
67. Drezner DW, Farrell H (2004) The power and politics of blogs. Download at <http://www.danieldrezner.com/research/blogpaperfinal.pdf>
68. Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, Aug 21–25, Chicago, IL, pp 36–43
69. Newman MEJ (2006) Modularity and community structure in networks. *Proc Natl Acad Sci* 103(23):8577–8582
70. Gruhl D, Guha R, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: WWW ‘04: Proceedings of the 13th International Conference on World Wide Web. ACM Press, New York, pp 491–501
71. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pp 137–146
72. Adar E, Zhang L, Adamic LA, Lukose RM (2004) Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem, New York, NY*

73. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pp 420–429
74. Zlatic V, Bozicevic M, Stefancic H, Domazet M (2006) Wikipedias: Collaborative web-based encyclopedias as complex networks. *Phys Rev E* 74(1):016115
75. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Conference on Knowledge Discovery in Data, Chicago, IL, USA, pp 177–187
76. Shi X, Tseng B, Adamic LA (2007) Looking at the Blogosphere Topology through Different Lenses. Proceedings of ICWSM'07. Boulder, CO, USA
77. Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G (2006) Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Phys Rev E* 74(3):036116
78. Halvey M, Keane MT, Smyth B (2006) Mobile web surfing is the same as web surfing. *Commun ACM* 49(3):76–81
79. Buyukkokten O, Garcia-Molina H, Paepcke A, Winograd T (2000) Power browser: efficient web browsing for pdas. In: CHI '00: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, New York, pp 430–437

### Books and Reviews

- Aiello W, Broder A, Janssen J, Milius E (eds) (2006) Algorithms and Models for the WebGraph: Proceedings of the 4th International Workshop, WAW 2006, Banff, Canada, 30 Nov–1 Dec
- Bonato A (2005) A Survey of Models of the Web graph. In: López-Ortiz A, Hamel A (eds) Combinatorial and Algorithmic Aspects of Networking. Lecture Notes in Computer Science, vol 3405. Springer, Berlin, pp 159
- Bonato A, Chung FRK (eds) (2007) Algorithms and Models for the WebGraph: Proceedings of the 5th International Workshop, WAW 2007, San Diego, CA, USA
- Caldarelli G (2007) Technological Networks: Internet and WWW in Scale-Free Networks: Complex Webs in Nature and Technology. Oxford University Press, Oxford
- Huberman BA (2001) The Laws of the Web: Patterns in the Ecology of Information. MIT, Cambridge