Computational Reservoir Geomechanics
Derek Elsworth
UPC, Qingdao, China
May 31 – June 4, 2018
http://www.ems.psu.edu/~elsworth/courses/comp_res_geomechs/2018/
http://www.youtube.com/derekelsworth

Day 1 (Thursday May 31st)
1. Reactive Flow and Permeability Dynamics – I　　　　　　　　[09:00-10:15]
2. Reactive Flow and Permeability Dynamics – II　　　　　　　　[10:30-11:45]
　　　------ Lunch -----　　　　　　　　　　　　　　　　　　　[11:45-14:30]
3. Introduction to Computational Reservoir Geomechanics [1:1]　[Self-Study]
4. Fluid Flow and Pressure Diffusion [2:-]
　　a. Finite Element Methods [2:1]　　　　　　　　　　　　　[14:30-15:45]
　　b. Conservation Equations and Galerkin Approximation [2:2]　[16:00-17:15]

Day 2 (Friday June 1st)
　　　----- Free-morning　　----- 　　　　　　　　　　　　　[09:00-11:45]
1. Fluid Flow and Pressure Diffusion, Continued [2:-]
　　a. 2D Triangular Constant Gradient Elements [2:3]　　　　　[14:30-15:45]
　　b. 1D Isoparametric Elements [2:4]　　　　　　　　　　　　[16:00-17:15]

Day 3 (Saturday June 2nd)
1. Geomechanics of Coal and Gas Shales　　　　　　　　　　　[09:00-10:15]
2. Fluid Flow and Pressure Diffusion, Continued [2:-]
　　a. 2D Isoparametric Elements and Numerical Integration [2:5]　[10:30-11:45]
　　　------ Lunch -----　　　　　　　　　　　　　　　　　　　[11:45-14:30]
　　b. Transient Behavior – "Mass" Matrices [2:6]　　　　　　　[14:30-15:45]
　　c. Transient Behavior – Integration in Time [2:7]　　　　　　[16:00-17:15]

Day 4 (Sunday June 3rd)
1. Gas Fracturing in Unconventional Reservoirs　　　　　　　　[09:00-10:15]
2. Mass Transport [3:-]
　　a. Conservation of Mass and 1D Models [3:1]　　　　　　　[10:30-11:45]
　　　------ Lunch -----　　　　　　　　　　　　　　　　　　　[11:45-14:30]
　　b. 2D Constant Gradient Elements [3:2]　　　　　　　　　　[14:30-15:45]
　　c. Sorption and Reactive Transport [3:3]　　　　　　　　　　[16:00-17:15]
3. Momentum Transport [4:-]
　　a. Self Study – Fluids, Navier-Stokes Equations [4:1]　　　　[Self-Study]

Day 5 (Monday June 4th)
1. Solid Mechanics [5:-]
　　a. Propagation, Proppant Transport and Conductivity of HFs　[09:00-10:15]
　　b. 1D and 2D Elements [5:1]　　　　　　　　　　　　　　　[10:30-11:45]
　　c. Self Study – Constitutive Equations [5:2]　　　　　　　　[Self-Study]
　　d. Self Study – Preamble for Coupled Systems [5:3]　　　　　[Self-Study]
　　　------ Lunch -----　　　　　　　　　　　　　　　　　　　[11:45-14:30]
2. "Coupled" Multiphysics Systems [6:-]
　　a. Dual-Porosity/Dual-Permeability Models [6:1]　　　　　　[14:30-15:45]
　　b. Coupled Hydro-Mechanical Models [6:2]　　　　　　　　　[16:00-17:15]
　　c. Self-Study – ComSol Models for HM Coupling [6:3]　　　　[Self-Study]
　　d. Self-Study – EGEEfem Models for HM Coupling [6:4]　　　[Self-Study]

EGEE 520 - Computational Geomechanics

Coupled Processes of Flow, Transport and Deformation

Derek Elsworth

Part I - Introduction

1 Introduction

[1:1] Introduction

    Overview of Syllabus

    Attributes of the Class

        FEM but also SPH/BEM/DEM/LBM

        Physics-based approach

        1D through 3D models

        Coupled multiphysics

            CFD and solid mechanics (momentum)

            Energy, Fluid, mass flows

        History of FEM

        Principal attributes of FEM

Part II - Finite Element Methods

2 Fluid Flow

[2:1] Fluid Flow and Pressure Diffusion

    Recap of FEM

    Comsol Applied to Flow

    1D Element

[2:2] Fluid Flow and Pressure Diffusion

    Recap

    Conservation of Mass

    Galerkin Formulation

    1D Element and Analysis

[2:3] Fluid Flow and Pressure Diffusion

    Recap

    2D Triangular (Constant Gradient) Elements

        Derivation

        Example

        EGEEfem

[2:4] Fluid Flow and Pressure Diffusion

    Recap – with EGEEfem

    2D Isoparametric Elements

        Concept

        1D example

        Numerical integration

[2:5] Fluid Flow and Pressure Diffusion

    Recap

    Isoparametric Elements

        Numerical integration

        2D and 3D elements

1
Introduction

[1:1] Introduction

Overview of Syllabus

Attributes of the Class

    FEM but also SPH/BEM/DEM/LBM

    Physics-based approach

    1D through 3D models

    Coupled multiphysics

        CFD and solid mechanics (momentum)

        Energy, Fluid, mass flows

    History of FEM

    Principal attributes of FEM

Chapter 1

Mathematical Foundations

1.1 TENSORS AND CONTINUUM MECHANICS

Continuum mechanics deals with physical quantities which are independent of any particular coordinate system that may be used to describe them. At the same time, these physical quantities are very often specified most conveniently by referring to an appropriate system of coordinates. Mathematically, such quantities are represented by *tensors*.

As a mathematical entity, a tensor has an existence independent of any coordinate system. Yet it may be specified in a particular coordinate system by a certain set of quantities, known as its *components*. Specifying the components of a tensor in one coordinate system determines the components in any other system. Indeed, the *law of transformation* of the components of a tensor is used here as a means for defining the tensor. Precise statements of the definitions of various kinds of tensors are given at the point of their introduction in the material that follows.

The physical laws of continuum mechanics are expressed by tensor equations. Because tensor transformations are linear and homogeneous, such tensor equations, if they are valid in one coordinate system, are valid in any other coordinate system. This *invariance* of tensor equations under a coordinate transformation is one of the principal reasons for the usefulness of tensor methods in continuum mechanics.

1.2 GENERAL TENSORS. CARTESIAN TENSORS. TENSOR RANK.

In dealing with general coordinate transformations between arbitrary curvilinear coordinate systems, the tensors defined are known as *general tensors*. When attention is restricted to transformations from one homogeneous coordinate system to another, the tensors involved are referred to as *Cartesian tensors*. Since much of the theory of continuum mechanics may be developed in terms of Cartesian tensors, the word "tensor" in this book means "Cartesian tensor" unless specifically stated otherwise.

Tensors may be classified by *rank*, or *order*, according to the particular form of the transformation law they obey. This same classification is also reflected in the number of components a given tensor possesses in an n-dimensional space. Thus in a three-dimensional Euclidean space such as ordinary physical space, the number of components of a tensor is 3^N, where N is the order of the tensor. Accordingly a tensor of *order zero* is specified in any coordinate system in three-dimensional space by *one* component. Tensors of order zero are called *scalars*. Physical quantities having magnitude only are represented by scalars. Tensors of *order one* have *three* coordinate components in physical space and are known as *vectors*. Quantities possessing both magnitude and direction are represented by vectors. *Second-order* tensors correspond to *dyadics*. Several important quantities in continuum mechanics are represented by tensors of rank two. Higher order tensors such as *triadics*, or tensors of order three, and *tetradics*, or tensors of order four are also defined and appear often in the mathematics of continuum mechanics.

1.3 VECTORS AND SCALARS

Certain physical quantities, such as force and velocity, which possess both magnitude and direction, may be represented in a three-dimensional space by *directed line segments* that obey the *parallelogram law of addition.* Such directed line segments are the geometrical representations of first-order tensors and are called *vectors.* Pictorially, a vector is simply an arrow pointing in the appropriate direction and having a length proportional to the magnitude of the vector. *Equal vectors* have the same direction and equal magnitudes. A *unit* vector is a vector of unit length. The *null* or *zero* vector is one having zero length and an unspecified direction. The *negative* of a vector is that vector having the same magnitude but opposite direction.

Those physical quantities, such as mass and energy, which possess magnitude only are represented by tensors of order zero which are called *scalars.*

In the *symbolic,* or *Gibbs* notation, vectors are designated by bold-faced letters such as **a, b,** etc. Scalars are denoted by italic letters such as a, b, λ, etc. Unit vectors are further distinguished by a caret placed over the bold-faced letter. In Fig. 1-1, arbitrary vectors **a** and **b** are shown along with the unit vector **ê** and the pair of equal vectors **c** and **d.**

Fig. 1-1

The magnitude of an arbitrary vector **a** is written simply as a, or for emphasis it may be denoted by the vector symbol between vertical bars as $|a|$.

1.4 VECTOR ADDITION. MULTIPLICATION OF A VECTOR BY A SCALAR

Vector addition obeys the *parallelogram law,* which defines the vector sum of two vectors as the diagonal of a parallelogram having the component vectors as adjacent sides. This law for vector addition is equivalent to the *triangle rule* which defines the sum of two vectors as the vector extending from the tail of the first to the head of the second when the summed vectors are adjoined head to tail. The graphical construction for the addition of **a** and **b** by the parallelogram law is shown in Fig. 1-2(a). Algebraically, the addition process is expressed by the vector equation

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a} = \mathbf{c} \qquad (1.1)$$

Vector subtraction is accomplished by addition of the negative vector as shown, for example, in Fig. 1-2(b) where the triangle rule is used. Thus

$$\mathbf{a} - \mathbf{b} = -\mathbf{b} + \mathbf{a} = \mathbf{d} \qquad (1.2)$$

The operations of vector addition and subtraction are commutative and associative as illustrated in Fig. 1-2(c), for which the appropriate equations are

$$(\mathbf{a} + \mathbf{b}) + \mathbf{g} = \mathbf{a} + (\mathbf{b} + \mathbf{g}) = \mathbf{h} \qquad (1.3)$$
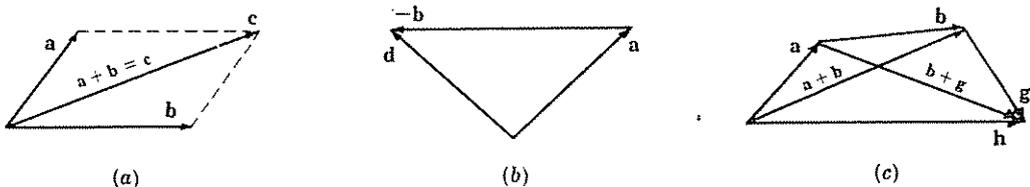

(a)   (b)   (c)

Fig. 1-2

Multiplication of a *vector* by a *scalar* produces in general a new vector having the same direction as the original but a different length. Exceptions are multiplication by zero to produce the null vector, and multiplication by unity which does not change a vector. Multiplication of the vector **b** by the scalar m results in one of the three possible cases shown in Fig. 1-3, depending upon the numerical value of m.

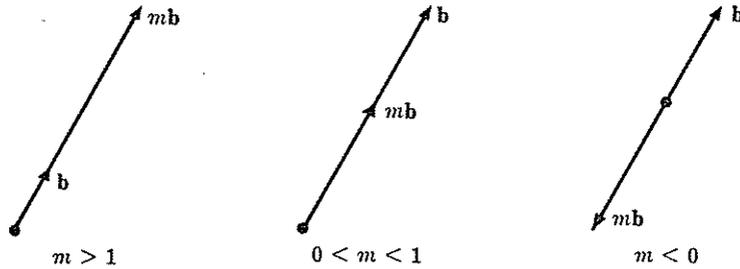$$m > 1 \qquad\qquad 0 < m < 1 \qquad\qquad m < 0$$

Fig. 1-3

Multiplication of a vector by a scalar is associative and distributive. Thus

$$m(n\mathbf{b}) = (mn)\mathbf{b} = n(m\mathbf{b}) \tag{1.4}$$

$$(m+n)\mathbf{b} = (n+m)\mathbf{b} = m\mathbf{b} + n\mathbf{b} \tag{1.5}$$

$$m(\mathbf{a}+\mathbf{b}) = m(\mathbf{b}+\mathbf{a}) = m\mathbf{a} + m\mathbf{b} \tag{1.6}$$

In the important case of a vector multiplied by the reciprocal of its magnitude, the result is a *unit vector* in the direction of the original vector. This relationship is expressed by the equation

$$\hat{\mathbf{b}} = \mathbf{b}/b \tag{1.7}$$

1.5 DOT AND CROSS PRODUCTS OF VECTORS

The *dot* or *scalar product* of two vectors **a** and **b** is the scalar

$$\lambda = \mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} = ab \cos\theta \tag{1.8}$$

in which θ is the smaller angle between the two vectors as shown in Fig. 1-4(a). The dot product of **a** with a unit vector $\hat{\mathbf{e}}$ gives the projection of **a** in the direction of $\hat{\mathbf{e}}$.

$$(a) \qquad\qquad\qquad\qquad\qquad (b)$$
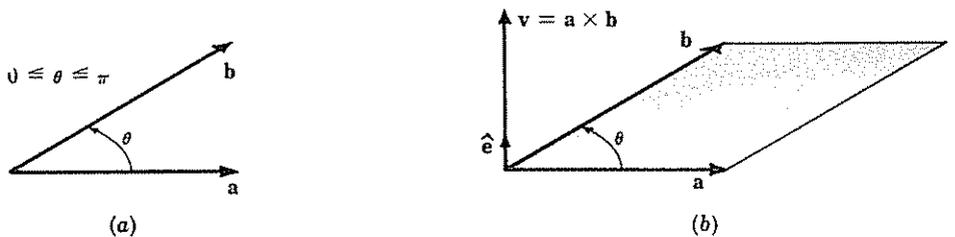
Fig. 1-4

The *cross* or *vector product* of **a** into **b** is the vector **v** given by

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a} = (ab \sin\theta)\,\hat{\mathbf{e}} \tag{1.9}$$

in which θ is the angle less than 180° between the vectors **a** and **b**, and $\hat{\mathbf{e}}$ is a unit vector perpendicular to their plane such that a right-handed rotation about $\hat{\mathbf{e}}$ through the angle θ carries **a** into **b**. The magnitude of **v** is equal to the area of the parallelogram having **a** and **b** as adjacent sides, shown shaded in Fig. 1-4(b). The cross product is not commutative.

The *scalar triple product* is a dot product of two vectors, one of which is a cross product.

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{b} \times \mathbf{c} = \lambda \qquad (1.10)$$

As indicated by (1.10) the dot and cross operation may be interchanged in this product. Also, since the cross operation must be carried out first, the parentheses are unnecessary and may be deleted as shown. This product is sometimes written [abc] and called the *box product*. The magnitude λ of the scalar triple product is equal to the volume of the parallelepiped having $\mathbf{a}, \mathbf{b}, \mathbf{c}$ as coterminous edges.

The *vector triple product* is a cross product of two vectors, one of which is itself a cross product. The following identity is frequently useful in expressing the product of $\mathbf{a}$ crossed into $\mathbf{b} \times \mathbf{c}$.

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c} = \mathbf{w} \qquad (1.11)$$

From (1.11), the product vector $\mathbf{w}$ is observed to lie in the plane of $\mathbf{b}$ and $\mathbf{c}$.

1.6 DYADS AND DYADICS

The *indeterminate vector product* of $\mathbf{a}$ and $\mathbf{b}$, defined by writing the vectors in juxtaposition as $\mathbf{ab}$ is called a *dyad*. The indeterminate product is not in general commutative, i.e. $\mathbf{ab} \neq \mathbf{ba}$. The first vector in a dyad is known as the *antecedent*, the second is called the *consequent*. A *dyadic* D corresponds to a tensor of order two and may always be represented as a finite sum of dyads

$$\mathsf{D} = \mathbf{a}_1\mathbf{b}_1 + \mathbf{a}_2\mathbf{b}_2 + \cdots + \mathbf{a}_N\mathbf{b}_N \qquad (1.12)$$

which is, however, never unique. In symbolic notation, dyadics are denoted by bold-faced sans-serif letters as above.

If in each dyad of (1.12) the antecedents and consequents are interchanged, the resulting dyadic is called the *conjugate dyadic* of D and is written

$$\mathsf{D}_c = \mathbf{b}_1\mathbf{a}_1 + \mathbf{b}_2\mathbf{a}_2 + \cdots + \mathbf{b}_N\mathbf{a}_N \qquad (1.13)$$

If each dyad of D in (1.12) is replaced by the dot product of the two vectors, the result is a scalar known as the *scalar of the dyadic* D and is written

$$\mathsf{D}_s = \mathbf{a}_1 \cdot \mathbf{b}_1 + \mathbf{a}_2 \cdot \mathbf{b}_2 + \cdots + \mathbf{a}_N \cdot \mathbf{b}_N \qquad (1.14)$$

If each dyad of D in (1.12) is replaced by the cross product of the two vectors, the result is called the *vector of the dyadic* D and is written

$$\mathsf{D}_v = \mathbf{a}_1 \times \mathbf{b}_1 + \mathbf{a}_2 \times \mathbf{b}_2 + \cdots + \mathbf{a}_N \times \mathbf{b}_N \qquad (1.15)$$

It can be shown that $\mathsf{D}_c, \mathsf{D}_s$ and D_v are independent of the representation (1.12).

The indeterminate vector product obeys the distributive laws

$$\mathbf{a}(\mathbf{b} + \mathbf{c}) = \mathbf{ab} + \mathbf{ac} \qquad (1.16)$$

$$(\mathbf{a} + \mathbf{b})\mathbf{c} = \mathbf{ac} + \mathbf{bc} \qquad (1.17)$$

$$(\mathbf{a} + \mathbf{b})(\mathbf{c} + \mathbf{d}) = \mathbf{ac} + \mathbf{ad} + \mathbf{bc} + \mathbf{bd} \qquad (1.18)$$

and if λ and μ are any scalars,

$$(\lambda + \mu)\mathbf{ab} = \lambda\mathbf{ab} + \mu\mathbf{ab} \qquad (1.19)$$

$$(\lambda\mathbf{a})\mathbf{b} = \mathbf{a}(\lambda\mathbf{b}) = \lambda\mathbf{ab} \qquad (1.20)$$

If **v** is any vector, the dot products $\mathbf{v}\cdot\mathbf{D}$ and $\mathbf{D}\cdot\mathbf{v}$ are the vectors defined respectively by

$$\mathbf{v}\cdot\mathbf{D} = (\mathbf{v}\cdot\mathbf{a}_1)\mathbf{b}_1 + (\mathbf{v}\cdot\mathbf{a}_2)\mathbf{b}_2 + \cdots + (\mathbf{v}\cdot\mathbf{a}_N)\mathbf{b}_N = \mathbf{u} \qquad (1.21)$$

$$\mathbf{D}\cdot\mathbf{v} = \mathbf{a}_1(\mathbf{b}_1\cdot\mathbf{v}) + \mathbf{a}_2(\mathbf{b}_2\cdot\mathbf{v}) + \cdots + \mathbf{a}_N(\mathbf{b}_N\cdot\mathbf{v}) = \mathbf{w} \qquad (1.22)$$

In (1.21) **D** is called the *postfactor,* and in (1.22) it is called the *prefactor.* Two dyadics **D** and **E** are *equal* if and only if for every vector **v**, either

$$\mathbf{v}\cdot\mathbf{D} = \mathbf{v}\cdot\mathbf{E} \quad \text{or} \quad \mathbf{D}\cdot\mathbf{v} = \mathbf{E}\cdot\mathbf{v} \qquad (1.23)$$

The *unit dyadic,* or *idemfactor* **I**, is the dyadic which can be represented as

$$\mathbf{I} = \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1 + \hat{\mathbf{e}}_2\hat{\mathbf{e}}_2 + \hat{\mathbf{e}}_3\hat{\mathbf{e}}_3 \qquad (1.24)$$

where $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$ constitute any orthonormal basis for three-dimensional Euclidean space (see Section 1.7). The dyadic **I** is characterized by the property

$$\mathbf{I}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{I} = \mathbf{v} \qquad (1.25)$$

for all vectors **v**.

The cross products $\mathbf{v}\times\mathbf{D}$ and $\mathbf{D}\times\mathbf{v}$ are the dyadics defined respectively by

$$\mathbf{v}\times\mathbf{D} = (\mathbf{v}\times\mathbf{a}_1)\mathbf{b}_1 + (\mathbf{v}\times\mathbf{a}_2)\mathbf{b}_2 + \cdots + (\mathbf{v}\times\mathbf{a}_N)\mathbf{b}_N = \mathbf{F} \qquad (1.26)$$

$$\mathbf{D}\times\mathbf{v} = \mathbf{a}_1(\mathbf{b}_1\times\mathbf{v}) + \mathbf{a}_2(\mathbf{b}_2\times\mathbf{v}) + \cdots + \mathbf{a}_N(\mathbf{b}_N\times\mathbf{v}) = \mathbf{G} \qquad (1.27)$$

The dot product of the dyads **ab** and **cd** is the dyad defined by

$$\mathbf{ab}\cdot\mathbf{cd} = (\mathbf{b}\cdot\mathbf{c})\mathbf{ad} \qquad (1.28)$$

From (1.28), the dot product of any two dyadics **D** and **E** is the dyadic

$$\mathbf{D}\cdot\mathbf{E} = (\mathbf{a}_1\mathbf{b}_1 + \mathbf{a}_2\mathbf{b}_2 + \cdots + \mathbf{a}_N\mathbf{b}_N)\cdot(\mathbf{c}_1\mathbf{d}_1 + \mathbf{c}_2\mathbf{d}_2 + \cdots + \mathbf{c}_N\mathbf{d}_N)$$

$$= (\mathbf{b}_1\cdot\mathbf{c}_1)\mathbf{a}_1\mathbf{d}_1 + (\mathbf{b}_1\cdot\mathbf{c}_2)\mathbf{a}_1\mathbf{d}_2 + \cdots + (\mathbf{b}_N\cdot\mathbf{c}_N)\mathbf{a}_N\mathbf{d}_N = \mathbf{G} \qquad (1.29)$$

The dyadics **D** and **E** are said to be *reciprocal* of each other if

$$\mathbf{E}\cdot\mathbf{D} = \mathbf{D}\cdot\mathbf{E} = \mathbf{I} \qquad (1.30)$$

For reciprocal dyadics, the notation $\mathbf{E} = \mathbf{D}^{-1}$ and $\mathbf{D} = \mathbf{E}^{-1}$ is often used.

Double dot and cross products are also defined for the dyads **ab** and **cd** as follows,

$$\mathbf{ab} : \mathbf{cd} = (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) \quad = \lambda, \quad \text{a scalar} \qquad (1.31)$$

$$\mathbf{ab} \overset{\times}{\cdot} \mathbf{cd} = (\mathbf{a}\times\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) \quad = \mathbf{h}, \quad \text{a vector} \qquad (1.32)$$

$$\mathbf{ab} \overset{\cdot}{\times} \mathbf{cd} = (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\times\mathbf{d}) \quad = \mathbf{g}, \quad \text{a vector} \qquad (1.33)$$

$$\mathbf{ab} \overset{\times}{\times} \mathbf{cd} = (\mathbf{a}\times\mathbf{c})(\mathbf{b}\times\mathbf{d}) = \mathbf{uw}, \text{ a dyad} \qquad (1.34)$$

From these definitions, double dot and cross products of dyadics may be readily developed. Also, some authors use the double dot product defined by

$$\mathbf{ab}\cdots\mathbf{cd} = (\mathbf{b}\cdot\mathbf{c})(\mathbf{a}\cdot\mathbf{d}) = \lambda, \quad \text{a scalar} \qquad (1.35)$$

A dyadic **D** is said to be *self-conjugate,* or *symmetric,* if

$$\mathbf{D} = \mathbf{D}_c \qquad (1.36)$$

and *anti-self-conjugate,* or *anti-symmetric,* if

$$\mathbf{D} = -\mathbf{D}_c \qquad (1.37)$$

Every dyadic may be expressed uniquely as the sum of a symmetric and anti-symmetric dyadic. For the arbitrary dyadic **D** the decomposition is

$$\mathbf{D} = \tfrac{1}{2}(\mathbf{D} + \mathbf{D}_c) + \tfrac{1}{2}(\mathbf{D} - \mathbf{D}_c) = \mathbf{G} + \mathbf{H} \qquad (1.38)$$

for which $\qquad$ $\mathbf{G}_c = \frac{1}{2}(\mathbf{D}_c + (\mathbf{D}_c)_c) = \frac{1}{2}(\mathbf{D}_c + \mathbf{D}) = \mathbf{G}$ (symmetric) $\qquad$ (1.39)

and $\qquad$ $\mathbf{H}_c = \frac{1}{2}(\mathbf{D}_c - (\mathbf{D}_c)_c) = \frac{1}{2}(\mathbf{D}_c - \mathbf{D}) = -\mathbf{H}$ (anti-symmetric) $\qquad$ (1.40)

Uniqueness is established by assuming a second decomposition, $\mathbf{D} = \mathbf{G}^* + \mathbf{H}^*$. Then

$$\mathbf{G}^* + \mathbf{H}^* = \mathbf{G} + \mathbf{H} \qquad (1.41)$$

and the conjugate of this equation is

$$\mathbf{G}^* - \mathbf{H}^* = \mathbf{G} - \mathbf{H} \qquad (1.42)$$

Adding and subtracting (1.41) and (1.42) in turn yields respectively the desired equalities, $\mathbf{G}^* = \mathbf{G}$ and $\mathbf{H}^* = \mathbf{H}$.

1.7 COORDINATE SYSTEMS. BASE VECTORS. UNIT VECTOR TRIADS

A vector may be defined with respect to a particular coordinate system by specifying the *components* of the vector in that system. The choice of coordinate system is arbitrary, but in certain situations a particular choice may be advantageous. The reference system of coordinate axes provides units for measuring vector magnitudes and assigns directions in space by which the orientation of vectors may be determined.

The well-known *rectangular Cartesian coordinate system* is often represented by the mutually perpendicular axes, $Oxyz$ shown in Fig. 1-5. Any vector $\mathbf{v}$ in this system may be expressed as a linear combination of three arbitrary, nonzero, noncoplanar vectors of the system, which are called *base vectors*. For base vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and suitably chosen scalar coefficients λ, μ, ν the vector $\mathbf{v}$ is given by

$$\mathbf{v} = \lambda\mathbf{a} + \mu\mathbf{b} + \nu\mathbf{c} \qquad (1.43)$$

Base vectors are by hypothesis linearly independent, i.e. the equation

$$\lambda\mathbf{a} + \mu\mathbf{b} + \nu\mathbf{c} = 0 \qquad (1.44)$$

is satisfied only if $\lambda = \mu = \nu = 0$. A set of base vectors in a given coordinate system is said to constitute a *basis* for that system.
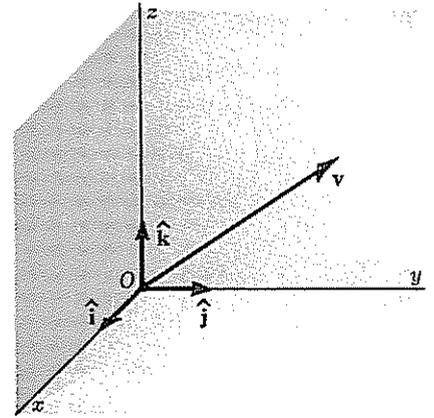
Fig. 1-5

The most frequent choice of base vectors for the rectangular Cartesian system is the set of unit vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ along the coordinate axes as shown in Fig. 1-5. These base vectors constitute a right-handed *unit vector triad*, for which

$$\hat{\mathbf{i}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}}, \quad \hat{\mathbf{j}} \times \hat{\mathbf{k}} = \hat{\mathbf{i}}, \quad \hat{\mathbf{k}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}} \qquad (1.45)$$

and

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{i}} = \hat{\mathbf{j}} \cdot \hat{\mathbf{j}} = \hat{\mathbf{k}} \cdot \hat{\mathbf{k}} = 1$$

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{j}} = \hat{\mathbf{j}} \cdot \hat{\mathbf{k}} = \hat{\mathbf{k}} \cdot \hat{\mathbf{i}} = 0 \qquad (1.46)$$

Such a set of base vectors is often called an *orthonormal basis*.

In terms of the unit triad $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, the vector $\mathbf{v}$ shown in Fig. 1-6 below may be expressed by

$$\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}} \qquad (1.47)$$

in which the Cartesian components

$$v_x = \mathbf{v} \cdot \hat{\mathbf{i}} = v \cos \alpha$$

$$v_y = \mathbf{v} \cdot \hat{\mathbf{j}} = v \cos \beta$$

$$v_z = \mathbf{v} \cdot \hat{\mathbf{k}} = v \cos \gamma$$

are the projections of $\mathbf{v}$ onto the coordinate axes. The unit vector in the direction of $\mathbf{v}$ is given according to (1.7) by

$$\hat{\mathbf{e}}_v = \mathbf{v}/v$$

$$= (\cos \alpha)\hat{\mathbf{i}} + (\cos \beta)\hat{\mathbf{j}} + (\cos \gamma)\hat{\mathbf{k}} \quad (1.48)$$

Since $\mathbf{v}$ is arbitrary, it follows that any unit vector will have the *direction cosines* of that vector as its *Cartesian components*.

In Cartesian component form the dot product of $\mathbf{a}$ and $\mathbf{b}$ is given by

$$\mathbf{a} \cdot \mathbf{b} = (a_x\hat{\mathbf{i}} + a_y\hat{\mathbf{j}} + a_z\hat{\mathbf{k}}) \cdot (b_x\hat{\mathbf{i}} + b_y\hat{\mathbf{j}} + b_z\hat{\mathbf{k}})$$

$$= a_xb_x + a_yb_y + a_zb_z \quad (1.49)$$

Fig. 1-6

For the same two vectors, the cross product $\mathbf{a} \times \mathbf{b}$ is

$$\mathbf{a} \times \mathbf{b} = (a_yb_z - a_zb_y)\hat{\mathbf{i}} + (a_zb_x - a_xb_z)\hat{\mathbf{j}} + (a_xb_y - a_yb_x)\hat{\mathbf{k}} \quad (1.50)$$

This result is often presented in the determinant form

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix} \quad (1.51)$$

in which the elements are treated as ordinary numbers. The triple scalar product may also be represented in component form by the determinant

$$[\mathbf{abc}] = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix} \quad (1.52)$$

In Cartesian component form, the dyad $\mathbf{ab}$ is given by

$$\mathbf{ab} = (a_x\hat{\mathbf{i}} + a_y\hat{\mathbf{j}} + a_z\hat{\mathbf{k}})(b_x\hat{\mathbf{i}} + b_y\hat{\mathbf{j}} + b_z\hat{\mathbf{k}})$$

$$= a_xb_x\hat{\mathbf{i}}\hat{\mathbf{i}} + a_xb_y\hat{\mathbf{i}}\hat{\mathbf{j}} + a_xb_z\hat{\mathbf{i}}\hat{\mathbf{k}}$$

$$+ a_yb_x\hat{\mathbf{j}}\hat{\mathbf{i}} + a_yb_y\hat{\mathbf{j}}\hat{\mathbf{j}} + a_yb_z\hat{\mathbf{j}}\hat{\mathbf{k}}$$

$$+ a_zb_x\hat{\mathbf{k}}\hat{\mathbf{i}} + a_zb_y\hat{\mathbf{k}}\hat{\mathbf{j}} + a_zb_z\hat{\mathbf{k}}\hat{\mathbf{k}} \quad (1.53)$$

Because of the *nine* terms involved, (1.53) is known as the *nonion form* of the dyad $\mathbf{ab}$. It is possible to put any dyadic into nonion form. The nonion form of the idemfactor in terms of the unit triad $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ is given by

$$\mathbf{I} = \hat{\mathbf{i}}\hat{\mathbf{i}} + \hat{\mathbf{j}}\hat{\mathbf{j}} + \hat{\mathbf{k}}\hat{\mathbf{k}} \quad (1.54)$$

In addition to the rectangular Cartesian coordinate system already discussed, curvilinear coordinate systems such as the cylindrical (R, θ, z) and spherical (r, θ, ϕ) systems shown in Fig. 1-7 below are also widely used. Unit triads $(\hat{\mathbf{e}}_R, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z)$ and $(\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_\phi)$ of base vectors illustrated in the figure are associated with these systems. However, the base vectors here do not all have fixed directions and are therefore, in general, functions of position.
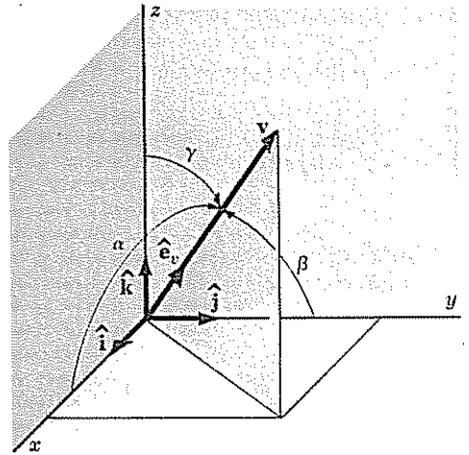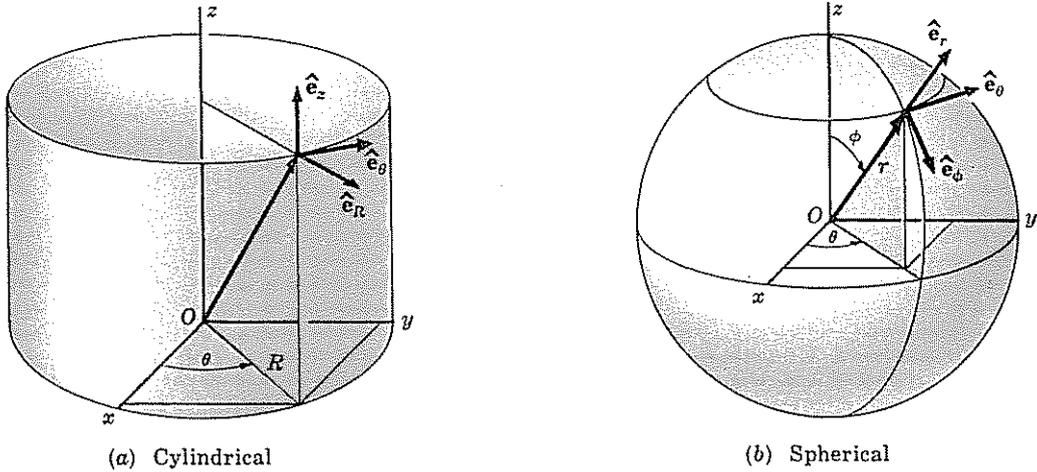
(a) Cylindrical

(b) Spherical

Fig. 1-7

1.8 LINEAR VECTOR FUNCTIONS. DYADICS AS LINEAR VECTOR OPERATORS

A vector **a** is said to be a function of a second vector **b** if **a** is determined whenever **b** is given. This functional relationship is expressed by the equation

$$\mathbf{a} = \mathbf{f}(\mathbf{b}) \tag{1.55}$$

The function **f** is said to be linear when the conditions

$$\mathbf{f}(\mathbf{b}+\mathbf{c}) = \mathbf{f}(\mathbf{b}) + \mathbf{f}(\mathbf{c}) \tag{1.56}$$

$$\mathbf{f}(\lambda\mathbf{b}) = \lambda\mathbf{f}(\mathbf{b}) \tag{1.57}$$

are satisfied for all vectors **b** and **c**, and for any scalar λ.

Writing **b** in Cartesian component form, equation (1.55) becomes

$$\mathbf{a} = \mathbf{f}(b_x\hat{\mathbf{i}} + b_y\hat{\mathbf{j}} + b_z\hat{\mathbf{k}}) \tag{1.58}$$

which, if **f** is linear, may be written

$$\mathbf{a} = b_x\mathbf{f}(\hat{\mathbf{i}}) + b_y\mathbf{f}(\hat{\mathbf{j}}) + b_z\mathbf{f}(\hat{\mathbf{k}}) \tag{1.59}$$

In (1.59) let $\mathbf{f}(\hat{\mathbf{i}}) = \mathbf{u}$, $\mathbf{f}(\hat{\mathbf{j}}) = \mathbf{v}$, $\mathbf{f}(\hat{\mathbf{k}}) = \mathbf{w}$, so that now

$$\mathbf{a} = \mathbf{u}(\hat{\mathbf{i}}\cdot\mathbf{b}) + \mathbf{v}(\hat{\mathbf{j}}\cdot\mathbf{b}) + \mathbf{w}(\hat{\mathbf{k}}\cdot\mathbf{b}) = (\mathbf{u}\hat{\mathbf{i}} + \mathbf{v}\hat{\mathbf{j}} + \mathbf{w}\hat{\mathbf{k}})\cdot\mathbf{b} \tag{1.60}$$

which is recognized as a dyadic-vector dot product and may be written

$$\mathbf{a} = \mathbf{D}\cdot\mathbf{b} \tag{1.61}$$

where $\mathbf{D} = \mathbf{u}\hat{\mathbf{i}} + \mathbf{v}\hat{\mathbf{j}} + \mathbf{w}\hat{\mathbf{k}}$. This demonstrates that any linear vector function **f** may be expressed as a dyadic-vector product. In (1.61) the dyadic **D** serves as a *linear vector operator* which operates on the *argument* vector **b** to produce the *image* vector **a**.

1.9 INDICIAL NOTATION. RANGE AND SUMMATION CONVENTIONS

The components of a tensor of any order, and indeed the tensor itself, may be represented clearly and concisely by the use of the *indicial notation*. In this notation, letter indices, either subscripts or superscripts, are appended to the *generic* or *kernel* letter representing the tensor quantity of interest. Typical examples illustrating use of indices are the tensor symbols

$$a_i, \quad b^j, \quad T_{ij}, \quad F_i{}^j, \quad \epsilon_{ijk}, \quad R^{pq}$$

In the "mixed" form, where both subscripts and superscripts appear, the dot shows that j is the second index.

Under the rules of indicial notation, a letter index may occur either *once* or *twice* in a given term. When an index occurs unrepeated in a term, that index is understood to take on the values $1, 2, \ldots, N$ where N is a specified integer that determines the *range* of the index. Unrepeated indices are known as *free* indices. The tensorial rank of a given term is equal to the number of free indices appearing in that term. Also, correctly written tensor equations have the same letters as free indices in every term.

When an index appears *twice* in a term, that index is understood to take on all the values of its range, and the resulting terms *summed*. In this so-called *summation convention*, repeated indices are often referred to as *dummy indices*, since their replacement by any other letter not appearing as a free index does not change the meaning of the term in which they occur. In general, no index occurs more than twice in a properly written term. If it is absolutely necessary to use some index more than twice to satisfactorily express a certain quantity, the summation convention must be suspended.

The number and location of the free indices reveal directly the exact tensorial character of the quantity expressed in the indicial notation. Tensors of *first order* are denoted by kernel letters bearing *one free index*. Thus the arbitrary vector **a** is represented by a symbol having a single subscript or superscript, i.e. in one or the other of the two forms,

$$a_i, \quad a^i$$

The following terms, having only one free index, are also recognized as first-order tensor quantities:

$$a_{ij}b_j, \quad F_{ikk}, \quad R^p_{.qp}, \quad \epsilon_{ijk}u_j v_k$$

Second-order tensors are denoted by symbols having *two* free indices. Thus the arbitrary dyadic **D** will appear in one of the three possible forms

$$D^{ij}, \quad D_i^{.j} \quad \text{or} \quad D^i_{.j}, \quad D_{ij}$$

In the "mixed" form, the dot shows that j is the second index. Second-order tensor quantities may also appear in various forms as, for example,

$$A_{ijip}, \quad B^{ij}_{..jk}, \quad \delta_{ij}u_k v_k$$

By a logical continuation of the above scheme, *third-order* tensors are expressed by symbols with *three* free indices. Also, a symbol such as λ which has no indices attached, represents a scalar, or tensor of zero order.

In ordinary physical space a *basis* is composed of three, noncoplanar vectors, and so any vector in this space is completely specified by its three components. Therefore the range on the index of a_i, which represents a vector in physical three-space, is $1, 2, 3$. Accordingly the symbol a_i is understood to represent the three components a_1, a_2, a_3. Also, a_i is sometimes interpreted to represent the ith component of the vector or indeed to represent the vector itself. For a range of three on both indices, the symbol A_{ij} represents nine components (of the second-order tensor (dyadic) **A**). The tensor A_{ij} is often presented explicitly by giving the nine components in a square array enclosed by large parentheses as

$$A_{ij} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \tag{1.62}$$

In the same way, the components of a first-order tensor (vector) in three-space may be displayed explicitly by a row or column arrangement of the form

$$a_i = (a_1, a_2, a_3) \quad \text{or} \quad a_i = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \tag{1.63}$$

In general, for a range of N, an nth order tensor will have N^n components.

The usefulness of the indicial notation in presenting systems of equations in compact form is illustrated by the following two typical examples. For a range of three on both i and j the indicial equation

$$x_i = c_{ij} z_j \tag{1.64}$$

represents in expanded form the three equations

$$\begin{aligned} x_1 &= c_{11} z_1 + c_{12} z_2 + c_{13} z_3 \\ x_2 &= c_{21} z_1 + c_{22} z_2 + c_{23} z_3 \\ x_3 &= c_{31} z_1 + c_{32} z_2 + c_{33} z_3 \end{aligned} \tag{1.65}$$

For a range of two on i and j, the indicial equation

$$A_{ij} = B_{ip} C_{jq} D_{pq} \tag{1.66}$$

represents, in expanded form, the four equations

$$\begin{aligned} A_{11} &= B_{11}C_{11}D_{11} + B_{11}C_{12}D_{12} + B_{12}C_{11}D_{21} + B_{12}C_{12}D_{22} \\ A_{12} &= B_{11}C_{21}D_{11} + B_{11}C_{22}D_{12} + B_{12}C_{21}D_{21} + B_{12}C_{22}D_{22} \\ A_{21} &= B_{21}C_{11}D_{11} + B_{21}C_{12}D_{12} + B_{22}C_{11}D_{21} + B_{22}C_{12}D_{22} \\ A_{22} &= B_{21}C_{21}D_{11} + B_{21}C_{22}D_{12} + B_{22}C_{21}D_{21} + B_{22}C_{22}D_{22} \end{aligned} \tag{1.67}$$

For a range of three on both i and j, (1.66) would represent nine equations, each having nine terms on the right-hand side.

1.10 SUMMATION CONVENTION USED WITH SYMBOLIC NOTATION

The summation convention is very often employed in connection with the representation of vectors and tensors by *indexed base vectors* written in the symbolic notation. Thus if the rectangular Cartesian axes and unit base vectors of Fig. 1-5 are relabeled as shown by Fig. 1-8, the arbitrary vector **v** may be written

$$\mathbf{v} = v_1 \hat{\mathbf{e}}_1 + v_2 \hat{\mathbf{e}}_2 + v_3 \hat{\mathbf{e}}_3 \tag{1.68}$$

in which v_1, v_2, v_3 are the rectangular Cartesian components of **v**. Applying the summation convention to (1.68), the equation may be written in the abbreviated form

$$\mathbf{v} = v_i \hat{\mathbf{e}}_i \tag{1.69}$$

where i is a summed index. The notation here is essentially *symbolic*, but with the added feature of the *summation convention*. In such a "combination" style of notation, tensor character is not given by the *free indices rule* as it is in true indicial notation.
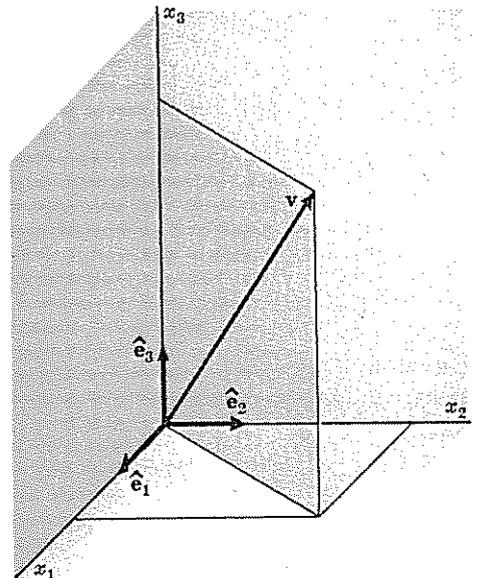

Fig. 1-8

Second-order tensors may also be represented by summation on indexed base vectors. Accordingly the dyad **ab** given in nonion form by *(1.53)* may be written

$$\mathbf{ab} = (a_i\hat{\mathbf{e}}_i)(b_j\hat{\mathbf{e}}_j) = a_ib_j\hat{\mathbf{e}}_i\hat{\mathbf{e}}_j \tag{1.70}$$

It is essential that the sequence of the base vectors be preserved in this expression. In similar fashion, the nonion form of the arbitrary dyadic **D** may be expressed in compact notation by

$$\mathbf{D} = D_{ij}\hat{\mathbf{e}}_i\hat{\mathbf{e}}_j \tag{1.71}$$

1.11 COORDINATE TRANSFORMATIONS. GENERAL TENSORS

Let x^i represent the arbitrary system of coordinates x^1, x^2, x^3 in a three-dimensional Euclidean space, and let θ^i represent any other coordinate system $\theta^1, \theta^2, \theta^3$ in the same space. Here the numerical superscripts are labels and not exponents. Powers of x may be expressed by use of parentheses as in $(x)^2$ or $(x)^3$. The letter superscripts are indices as already noted. The *coordinate transformation equations*

$$\theta^i = \theta^i(x^1, x^2, x^3) \tag{1.72}$$

assign to any point (x^1, x^2, x^3) in the x^i system a new set of coordinates $(\theta^1, \theta^2, \theta^3)$ in the θ^i system. The functions θ^i relating the two sets of variables (coordinates) are assumed to be single-valued, continuous, differentiable functions. The determinant

$$J = \begin{vmatrix} \dfrac{\partial\theta^1}{\partial x^1} & \dfrac{\partial\theta^1}{\partial x^2} & \dfrac{\partial\theta^1}{\partial x^3} \\[2mm] \dfrac{\partial\theta^2}{\partial x^1} & \dfrac{\partial\theta^2}{\partial x^2} & \dfrac{\partial\theta^2}{\partial x^3} \\[2mm] \dfrac{\partial\theta^3}{\partial x^1} & \dfrac{\partial\theta^3}{\partial x^2} & \dfrac{\partial\theta^3}{\partial x^3} \end{vmatrix} \tag{1.73}$$

or, in compact form,

$$J = \left|\frac{\partial\theta^i}{\partial x^j}\right| \tag{1.74}$$

is called the *Jacobian* of the transformation. If the Jacobian does not vanish, *(1.72)* possesses a unique inverse set of the form

$$x^i = x^i(\theta^1, \theta^2, \theta^3) \tag{1.75}$$

The coordinate systems represented by x^i and θ^i in *(1.72)* and *(1.75)* are completely general and may be any curvilinear or Cartesian systems.

From *(1.72)*, the differential vector $d\theta^i$ is given by

$$d\theta^i = \frac{\partial\theta^i}{\partial x^j}\, dx^j \tag{1.76}$$

This equation is a prototype of the equation which defines the class of tensors known as *contravariant vectors*. In general, a set of quantities b^i associated with a point P are said to be the components of a *contravariant tensor of order one* if they transform, under a coordinate transformation, according to the equation

$$b'^i = \frac{\partial\theta^i}{\partial x^j}\, b^j \tag{1.77}$$

where the partial derivatives are evaluated at P. In *(1.77)*, b^j are the components of the tensor in the x^j coordinate system, while b'^i are the components in the θ^i system. In general

tensor theory, contravariant tensors are recognized by the use of superscripts as indices. It is for this reason that the coordinates are labeled x^i here rather than x_i, but it must be noted that it is only the differentials dx^i, and not the coordinates themselves, which have tensor character.

By a logical extension of the tensor concept expressed in (1.77), the definition of *contravariant tensors of order two* requires the tensor components to obey the transformation law

$$B'^{ij} \;=\; \frac{\partial \theta^i}{\partial x^r}\frac{\partial \theta^j}{\partial x^s}\,B^{rs} \tag{1.78}$$

Contravariant tensors of third, fourth and higher orders are defined in a similar manner.

The word *contravariant* is used above to distinguish that class of tensors from the class known as *covariant* tensors. In general tensor theory, covariant tensors are recognized by the use of subscripts as indices. The prototype of the *covariant vector* is the partial derivative of a scalar function of the coordinates. Thus if $\phi = \phi(x^1, x^2, x^3)$ is such a function,

$$\frac{\partial \phi}{\partial \theta^i} \;=\; \frac{\partial \phi}{\partial x^j}\frac{\partial x^j}{\partial \theta^i} \tag{1.79}$$

In general, a set of quantities b_i are said to be the components of a *covariant tensor of order one* if they transform according to the equation

$$b'_i \;=\; \frac{\partial x^j}{\partial \theta^i}\,b_j \tag{1.80}$$

In (1.80), b'_i are the covariant components in the θ^i system, b_i the components in the x_i system. *Second-order covariant tensors* obey the transformation law

$$B'_{ij} \;=\; \frac{\partial x^r}{\partial \theta^i}\frac{\partial x^s}{\partial \theta^j}\,B_{rs} \tag{1.81}$$

Covariant tensors of higher order and *mixed tensors*, such as

$$T'^{r}_{\cdot sp} \;=\; \frac{\partial \theta^r}{\partial x^m}\frac{\partial x^n}{\partial \theta^s}\frac{\partial x^q}{\partial \theta^p}\,T^{m}_{\cdot nq} \tag{1.82}$$

are defined in the obvious way.

1.12 THE METRIC TENSOR. CARTESIAN TENSORS

Let x^i represent a system of rectangular Cartesian coordinates in a Euclidean three-space, and let θ^i represent any system of rectangular or curvilinear coordinates (e.g. cylindrical or spherical coordinates) in the same space. The vector $\mathbf{x}$ having Cartesian components x^i is called the *position vector* of the arbitrary point $P(x^1, x^2, x^3)$ referred to the rectangular Cartesian axes. The square of the differential element of distance between neighboring points $P(\mathbf{x})$ and $Q(\mathbf{x} + d\mathbf{x})$ is given by

$$(ds)^2 \;=\; dx^i\,dx^i \tag{1.83}$$

From the coordinate transformation

$$x^i \;=\; x^i(\theta^1, \theta^2, \theta^3) \tag{1.84}$$

relating the systems, the distance differential is

$$dx^i \;=\; \frac{\partial x^i}{\partial \theta^p}\,d\theta^p \tag{1.85}$$

and therefore (1.83) becomes

$$(ds)^2 \;=\; \frac{\partial x^i}{\partial \theta^p}\frac{\partial x^i}{\partial \theta^q}\, d\theta^p\, d\theta^q \;=\; g_{pq}\, d\theta^p\, d\theta^q \tag{1.86}$$

where the second-order tensor $g_{pq} = (\partial x^i/\partial \theta^p)(\partial x^i/\partial \theta^q)$ is called the *metric tensor*, or *fundamental tensor* of the space. If θ^i represents a rectangular Cartesian system, say the x'^i system, then

$$g_{pq} \;=\; \frac{\partial x^i}{\partial x'^p}\frac{\partial x^i}{\partial x'^q} \;=\; \delta_{pq} \tag{1.87}$$

where δ_{pq} is the *Kronecker delta* (see Section 1.13) defined by $\delta_{pq}=0$ if $p \neq q$ and $\delta_{pq}=1$ if $p=q$.

Any system of coordinates for which the squared differential element of distance takes the form of (1.83) is called a system of *homogeneous coordinates*. Coordinate transformations between systems of homogeneous coordinates are *orthogonal transformations*, and when attention is restricted to such transformations, the tensors so defined are called *Cartesian tensors*. In particular, this is the case for transformation laws between systems of rectangular Cartesian coordinates with a common origin. For Cartesian tensors there is no distinction between contravariant and covariant components and therefore it is customary to use subscripts exclusively in expressions representing Cartesian tensors. As will be shown next, in the transformation laws defining Cartesian tensors, the partial derivatives appearing in general tensor definitions, such as (1.80) and (1.81), are replaced by constants.

1.13. TRANSFORMATION LAWS FOR CARTESIAN TENSORS. THE KRONECKER DELTA. ORTHOGONALITY CONDITIONS

Let the axes $Ox_1x_2x_3$ and $Ox_1'x_2'x_3'$ represent two rectangular Cartesian coordinate systems with a common origin at an arbitrary point O as shown in Fig. 1-9. The primed system may be imagined to be obtained from the unprimed by a rotation of the axes about the origin, or by a reflection of axes in one of the coordinate planes, or by a combination of these. If the symbol a_{ij} denotes the cosine of the angle between the ith primed and jth unprimed coordinate axes, i.e. $a_{ij} = \cos(x_i', x_j)$, the relative orientation of the individual axes of each system with respect to the other is conveniently given by the table

Fig. 1-9

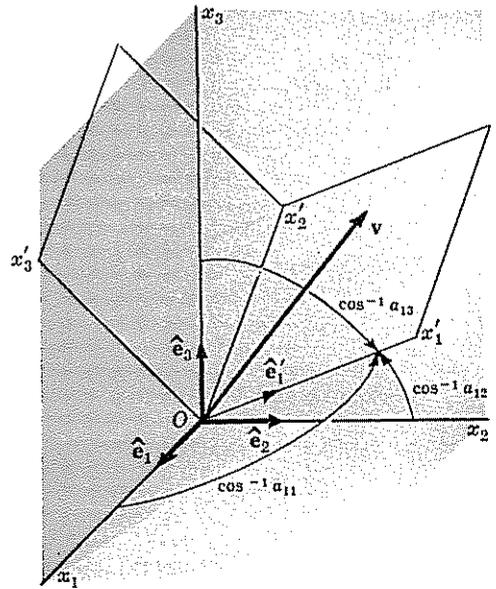|	x_1	x_2	x_3
x_1'	a_{11}	a_{12}	a_{13}
x_2'	a_{21}	a_{22}	a_{23}
x_3'	a_{31}	a_{32}	a_{33}

or alternatively by the transformation tensor

$$\mathbf{A} \;=\; \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

From this definition of a_{ij}, the unit vector $\hat{e}_1'$ along the x_1' axis is given according to (1.48) and the summation convention by

$$\hat{e}_1' = a_{11}\hat{e}_1 + a_{12}\hat{e}_2 + a_{13}\hat{e}_3 = a_{1j}\hat{e}_j \tag{1.88}$$

An obvious generalization of this equation gives the arbitrary unit base vector $\hat{e}_i'$ as

$$\hat{e}_i' = a_{ij}\hat{e}_j \tag{1.89}$$

In component form, the arbitrary vector $\mathbf{v}$ shown in Fig. 1-9 may be expressed in the unprimed system by the equation

$$\mathbf{v} = v_j\hat{e}_j \tag{1.90}$$

and in the primed system by

$$\mathbf{v} = v_i'\hat{e}_i' \tag{1.91}$$

Replacing $\hat{e}_i'$ in (1.91) by its equivalent form (1.89) yields the result

$$\mathbf{v} = v_i'a_{ij}\hat{e}_j \tag{1.92}$$

Comparing (1.92) with (1.90) reveals that the vector components in the primed and unprimed systems are related by the equations

$$v_j = a_{ij}v_i' \tag{1.93}$$

The expression (1.93) is the *transformation law* for first-order Cartesian tensors, and as such is seen to be a special case of the general form of first-order tensor transformations, expressed by (1.80) and (1.77). By interchanging the roles of the primed and unprimed base vectors in the above development, the inverse of (1.93) is found to be

$$v_i' = a_{ij}v_j \tag{1.94}$$

It is important to note that in (1.93) the free index on a_{ij} appears as the second index. In (1.94), however, the free index appears as the first index.

By an appropriate choice of dummy indices, (1.93) and (1.94) may be combined to produce the equation

$$v_j = a_{ij}a_{ik}v_k \tag{1.95}$$

Since the vector $\mathbf{v}$ is arbitrary, (1.95) must reduce to the identity $v_j = v_j$. Therefore the coefficient $a_{ij}a_{ik}$, whose value depends upon the subscripts j and k, must equal 1 or 0 according to whether the numerical values of j and k are the same or different. The *Kronecker delta*, defined by

$$\delta_{ij} = \begin{cases} 1 & \text{for } i=j \\ 0 & \text{for } i \neq j \end{cases} \tag{1.96}$$

may be used to represent quantities such as $a_{ij}a_{ik}$. Thus with the help of the Kronecker delta the conditions on the coefficient in (1.95) may be written

$$a_{ij}a_{ik} = \delta_{jk} \tag{1.97}$$

In expanded form, (1.97) consists of nine equations which are known as the *orthogonality* or *orthonormality conditions* on the direction cosines a_{ij}. Finally, (1.93) and (1.94) may also be combined to produce $v_i = a_{ij}a_{kj}v_k'$ from which the orthogonality conditions appear in the alternative form

$$a_{ij}a_{kj} = \delta_{ik} \tag{1.98}$$

A linear transformation such as (1.93) or (1.94), whose coefficients satisfy (1.97) or (1.98), is said to be an *orthogonal transformation*. Coordinate axes rotations and reflections of the axes in a coordinate plane both lead to orthogonal transformations.

The Kronecker delta is sometimes called the *substitution operator*, since, for example,

$$\delta_{ij}b_j = \delta_{i1}b_1 + \delta_{i2}b_2 + \delta_{i3}b_3 = b_i \qquad (1.99)$$

and, likewise,

$$\delta_{ij}F_{ik} = \delta_{1j}F_{1k} + \delta_{2j}F_{2k} + \delta_{3j}F_{3k} = F_{jk} \qquad (1.100)$$

It is clear from this property that the Kronecker delta is the indicial counterpart to the symbolic idemfactor I, which is given by (1.54).

According to the transformation law (1.94), the dyad u_iv_j has components in the primed coordinate system given by

$$u'_i v'_j = (a_{ip}u_p)(a_{jq}v_q) = a_{ip}a_{jq}u_p v_q \qquad (1.101)$$

In an obvious generalization of (1.101), any second-order Cartesian tensor T_{ij} obeys the transformation law

$$T'_{ij} = a_{ip}a_{jq}T_{pq} \qquad (1.102)$$

With the help of the orthogonality conditions it is a simple calculation to invert (1.102), thereby giving the transformation rule from primed components to unprimed components:

$$T_{ij} = a_{pi}a_{qj}T'_{pq} \qquad (1.103)$$

The transformation laws for first and second-order Cartesian tensors generalize for an Nth order Cartesian tensor to

$$T'_{ijk\ldots} = a_{ip}a_{jq}a_{km}\ldots T_{pqm\ldots} \qquad (1.104)$$

1.14 ADDITION OF CARTESIAN TENSORS. MULTIPLICATION BY A SCALAR

Cartesian tensors of the same order may be added (or subtracted) component by component in accordance with the rule

$$A_{ijk\ldots} \pm B_{ijk\ldots} = T_{ijk\ldots} \qquad (1.105)$$

The sum is a tensor of the same order as those added. Note that like indices appear in the same sequence in each term.

Multiplication of every component of a tensor by a given scalar produces a new tensor of the same order. For the scalar multiplier λ, typical examples written in both indicial and symbolic notation are

$$b_i = \lambda a_i \quad \text{or} \quad \mathbf{b} = \lambda\mathbf{a} \qquad (1.106)$$

$$B_{ij} = \lambda A_{ij} \quad \text{or} \quad \mathbf{B} = \lambda\mathbf{A} \qquad (1.107)$$

1.15 TENSOR MULTIPLICATION

The *outer product* of two tensors of arbitrary order is the tensor whose components are formed by multiplying each component of one of the tensors by every component of the other. This process produces a tensor having an order which is the sum of the orders of the factor tensors. Typical examples of outer products are

(a) $a_i b_j = T_{ij}$ $\qquad$ (c) $D_{ij}T_{km} = \Phi_{ijkm}$

(b) $v_i F_{jk} = \alpha_{ijk}$ $\qquad$ (d) $\epsilon_{ijk}v_m = \Theta_{ijkm}$

As indicated by the above examples, outer products are formed by simply setting down the factor tensors in juxtaposition. (Note that a dyad is formed from two vectors by this very procedure.)

Contraction of a tensor with respect to two free indices is the operation of assigning to both indices the same letter subscript, thereby changing these indices to dummy indices. Contraction produces a tensor having an order two less than the original. Typical examples of contraction are the following.

(a) Contractions of T_{ij} and $u_i v_j$

$$T_{ii} = T_{11} + T_{22} + T_{33}$$

$$u_i v_i = u_1 v_1 + u_2 v_2 + u_3 v_3$$

(b) Contractions of $E_{ij} a_k$

$$E_{ij} a_j = b_i$$

$$E_{ij} a_i = c_j$$

$$E_{ii} a_k = d_k$$

(c) Contractions of $E_{ij} F_{km}$

$$E_{ij} F_{im} = G_{jm} \qquad E_{ij} F_{kk} = P_{ij}$$

$$E_{ij} F_{ki} = H_{jk} \qquad E_{ij} F_{jm} = Q_{im}$$

$$E_{ii} F_{km} = K_{km} \qquad E_{ij} F_{kj} = R_{ik}$$

An *inner product* of two tensors is the result of a contraction, involving one index from each tensor, performed on the outer product of the two tensors. Several inner products important to continuum mechanics are listed here for reference, in both the indicial and symbolic notations.

Outer Product	Inner Product	
	Indicial Notation	Symbolic Notation
1. $a_i b_j$	$a_i b_i$	$\mathbf{a} \cdot \mathbf{b}$
2. $a_i E_{jk}$	$a_i E_{ik} = f_k$	$\mathbf{a} \cdot \mathbf{E} = \mathbf{f}$
	$a_i E_{ji} = h_j$	$\mathbf{E} \cdot \mathbf{a} = \mathbf{h}$
3. $E_{ij} F_{km}$	$E_{ij} F_{jm} = G_{im}$	$\mathbf{E} \cdot \mathbf{F} = \mathbf{G}$
4. $E_{ij} E_{km}$	$E_{ij} E_{jm} = B_{im}$	$\mathbf{E} \cdot \mathbf{E} = (\mathbf{E})^2$

Multiple contractions of fourth-order and higher tensors are sometimes useful. Two such examples are

1. $E_{ij} F_{km}$ contracted to $E_{ij} F_{ij}$,   or $\mathbf{E} : \mathbf{F}$

2. $E_{ij} E_{km} E_{pq}$ contracted to $E_{ij} E_{jm} E_{mq}$, or $(\mathbf{E})^3$

1.16 VECTOR CROSS PRODUCT. PERMUTATION SYMBOL. DUAL VECTORS

In order to express the cross product $\mathbf{a} \times \mathbf{b}$ in the indicial notation, the third-order tensor ϵ_{ijk}, known as the *permutation symbol* or *alternating tensor*, must be introduced. This useful tensor is defined by

$$
\epsilon_{ijk} = \begin{cases}
1 & \text{if the values of } i, j, k \text{ are an } even\ permutation \text{ of } 1, 2, 3 \text{ (i.e. if they appear in sequence as in the arrangement 1 2 3 1 2).} \\
-1 & \text{if the values of } i, j, k \text{ are an } odd\ permutation \text{ of } 1, 2, 3 \text{ (i.e. if they appear in sequence as in the arrangement 3 2 1 3 2).} \\
0 & \text{if the values of } i, j, k \text{ are } not \text{ a permutation of } 1, 2, 3 \text{ (i.e. if two or more of the indices have the same value).}
\end{cases}
$$

From this definition, the cross product $\mathbf{a} \times \mathbf{b} = \mathbf{c}$ is written in indicial notation by

$$\epsilon_{ijk}a_jb_k = c_i \qquad (1.108)$$

Using this relationship, the box product $\mathbf{a} \times \mathbf{b} \cdot \mathbf{c} = \lambda$ may be written

$$\lambda = \epsilon_{ijk}a_ib_jc_k \qquad (1.109)$$

Since the same box product is given in the form of a determinant by (1.52), it is not surprising that the permutation symbol is frequently used to express the value of a 3×3 determinant.

It is worthwhile to note that ϵ_{ijk} obeys the tensor transformation law for third order Cartesian tensors as long as the transformation is a *proper* one ($\det a_{ij} = 1$) such as arises from a rotation of axes. If the transformation is *improper* ($\det a_{ij} = -1$), e.g. a reflection in one of the coordinate planes whereby a right-handed coordinate system is transformed into a left-handed one, a minus sign must be inserted into the transformation law for ϵ_{ijk}. Such tensors are called *pseudo-tensors*.

The *dual vector* of an arbitrary second-order Cartesian tensor T_{ij} is defined by

$$v_i = \epsilon_{ijk}T_{jk} \qquad (1.110)$$

which is observed to be the indicial equivalent of $\mathbf{T}_v$, the "vector of the dyadic $\mathbf{T}$", as defined by (1.15).

1.17　MATRICES. MATRIX REPRESENTATION OF CARTESIAN TENSORS

A rectangular array of elements, enclosed by square brackets and subject to certain laws of combination, is called a *matrix*. An $M \times N$ matrix is one having M (horizontal) rows and N (vertical) columns of elements. In the symbol A_{ij}, used to represent the typical element of a matrix, the first subscript denotes the row, the second subscript the column occupied by the element. The matrix itself is designated by enclosing the typical element symbol in square brackets, or alternatively, by the *kernel* letter of the matrix in *script*. For example, the $M \times N$ matrix $\mathcal{A}$, or $[A_{ij}]$ is the array given by

$$\mathcal{A} = [A_{ij}] = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1N} \\ A_{21} & A_{22} & \ldots & A_{2N} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_{M1} & A_{M2} & \ldots & A_{MN} \end{bmatrix} \qquad (1.111)$$

A matrix for which $M = N$, is called a *square matrix*. A $1 \times N$ matrix, written $[a_{1k}]$, is called a *row matrix*. An $M \times 1$ matrix, written $[a_{k1}]$, is called a *column matrix*. A matrix having only zeros as elements is called the *zero matrix*. A square matrix with zeros everywhere except on the main diagonal (from A_{11} to A_{NN}) is called a *diagonal matrix*. If the nonzero elements of a diagonal matrix are all *unity*, the matrix is called the *unit* or *identity matrix*. The $N \times M$ matrix $\mathcal{A}^T$, formed by interchanging rows and columns of the $M \times N$ matrix $\mathcal{A}$, is called the *transpose matrix* of $\mathcal{A}$.

Matrices having the same number of rows and columns may be *added* (or subtracted) *element* by *element*. Multiplication of the matrix $[A_{ij}]$ by a scalar λ results in the matrix $[\lambda A_{ij}]$. The product of two matrices, $\mathcal{A}\mathcal{B}$, is defined only if the matrices are *conformable*, i.e. if the *prefactor* matrix $\mathcal{A}$ has the same number of columns as the *postfactor* matrix $\mathcal{B}$ has rows. The product of an $M \times P$ matrix multiplied into a $P \times N$ matrix is an $M \times N$ matrix. Matrix multiplication is usually denoted by simply setting down the matrix symbols in juxtaposition as in

$$\mathcal{A}\mathcal{B} = \mathcal{C} \quad \text{or} \quad [A_{ij}][B_{jk}] = [C_{ik}] \qquad (1.112)$$

Matrix multiplication is not, in general, commutative: $\mathcal{A}\mathcal{B} \neq \mathcal{B}\mathcal{A}$.

A square matrix $\mathcal{A}$ whose determinant $|A_{ij}|$ is zero is called a *singular matrix*. The *cofactor* of the element A_{ij} of the square matrix $\mathcal{A}$, denoted here by A_{ij}^{*}, is defined by

$$A_{ij}^{*} = (-1)^{i+j}M_{ij} \qquad (1.113)$$

in which M_{ij} is the *minor* of A_{ij}; i.e. the determinant of the square array remaining after the row and column of A_{ij} are deleted. The *adjoint* matrix of $\mathcal{A}$ is obtained by replacing each element by its cofactor and then interchanging rows and columns. If a square matrix $\mathcal{A} = [A_{ij}]$ is non-singular, it possesses a unique *inverse matrix* $\mathcal{A}^{-1}$ which is defined as the adjoint matrix of $\mathcal{A}$ divided by the determinant of $\mathcal{A}$. Thus

$$\mathcal{A}^{-1} = \frac{[A_{ji}^{*}]}{|\mathcal{A}|} \qquad (1.114)$$

From the inverse matrix definition (1.114) it may be shown that

$$\mathcal{A}^{-1}\mathcal{A} = \mathcal{A}\mathcal{A}^{-1} = \mathcal{J} \qquad (1.115)$$

where $\mathcal{J}$ is the *identity matrix*, having ones on the principal diagonal and zeros elsewhere, and so named because of the property

$$\mathcal{J}\mathcal{A} = \mathcal{A}\mathcal{J} = \mathcal{A} \qquad (1.116)$$

It is clear, of course, that $\mathcal{J}$ is the matrix representation of δ_{ij}, the Kronecker delta, and of I, the unit dyadic. Any matrix $\mathcal{A}$ for which the condition $\mathcal{A}^{T} = \mathcal{A}^{-1}$ is satisfied is called an *orthogonal matrix*. Accordingly, if $\mathcal{A}$ is orthogonal,

$$\mathcal{A}^{T}\mathcal{A} = \mathcal{A}\mathcal{A}^{T} = \mathcal{J} \qquad (1.117)$$

As suggested by the fact that any dyadic may be expressed in the nonion form (1.53), and, equivalently, since the components of a second-order tensor may be displayed in the square array (1.62), it proves extremely useful to represent second-order tensors (dyadics) by square, 3×3 matrices. A first-order tensor (vector) may be represented by either a 1×3 row matrix, or by a 3×1 column matrix. Although every Cartesian tensor of order two or less (dyadics, vectors, scalars) may be represented by a matrix, not every matrix represents a tensor.

If both matrices in the product $\mathcal{A}\mathcal{B} = \mathcal{C}$ are 3×3 matrices representing second-order tensors, the multiplication is equivalent to the inner product expressed in indicial notation by

$$A_{ij}B_{jk} = C_{ik} \qquad (1.118)$$

where the range is three. Expansion of (1.118) duplicates the "row by column" multiplication of matrices wherein the elements of the ith row of the prefactor matrix are multiplied in turn by the elements of the kth column of the postfactor matrix, and these products summed to give the element in the ith row and kth column of the product matrix. Several such products occur repeatedly in continuum mechanics and are recorded here in the various notations for reference and comparison.

(a) *Vector dot product*

$$\mathbf{a}\cdot\mathbf{b} = \mathbf{b}\cdot\mathbf{a} = \lambda \qquad [a_{1j}][b_{j1}] = [\lambda]$$

$$a_{i}b_{i} = b_{i}a_{i} = \lambda \qquad [a_{1}, a_{2}, a_{3}]\begin{bmatrix} b_{1} \\ b_{2} \\ b_{3} \end{bmatrix} = [a_{1}b_{1} + a_{2}b_{2} + a_{3}b_{3}] \qquad (1.119)$$

(b) Vector-dyadic dot product

$$\mathbf{a} \cdot \mathbf{E} = \mathbf{b} \qquad a\mathcal{E} = \mathcal{B}$$

$$a_i E_{ij} = b_j \qquad [a_{1i}][E_{ij}] = [b_{1j}]$$

$$[a_1, a_2, a_3] \begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{bmatrix} = \begin{matrix} [a_1 E_{11} + a_2 E_{21} + a_3 E_{31}, \\ a_1 E_{12} + a_2 E_{22} + a_3 E_{32}, \\ a_1 E_{13} + a_2 E_{23} + a_3 E_{33}] \end{matrix} \qquad (1.120)$$

(c) Dyadic-vector dot product

$$\mathbf{E} \cdot \mathbf{a} = \mathbf{c} \qquad \mathcal{E}a = c$$

$$E_{ij} a_j = c_i \qquad [E_{ij}][a_{j1}] = [c_{i1}]$$

$$\begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} a_1 E_{11} + a_2 E_{12} + a_3 E_{13} \\ a_1 E_{21} + a_2 E_{22} + a_3 E_{23} \\ a_1 E_{31} + a_2 E_{32} + a_3 E_{33} \end{bmatrix} \qquad (1.121)$$

1.18 SYMMETRY OF DYADICS, MATRICES AND TENSORS

According to (1.36) (or (1.37)), a dyadic **D** is said to be symmetric (anti-symmetric) if it is equal to (the negative of) its conjugate **D**$_c$. Similarly the second-order tensor D_{ij} is *symmetric* if

$$D_{ij} = D_{ji} \qquad (1.122)$$

and is *anti-symmetric*, or *skew-symmetric*, if

$$D_{ij} = -D_{ji} \qquad (1.123)$$

Therefore the decomposition of D_{ij} analogous to (1.38) is

$$D_{ij} = \tfrac{1}{2}(D_{ij} + D_{ji}) + \tfrac{1}{2}(D_{ij} - D_{ji}) \qquad (1.124)$$

or, in an equivalent abbreviated form often employed,

$$D_{ij} = D_{(ij)} + D_{[ij]} \qquad (1.125)$$

where parentheses around the indices denote the symmetric part of D_{ij}, and square brackets on the indices denote the anti-symmetric part.

Since the interchange of indices of a second-order tensor is equivalent to the interchange of rows and columns in its matrix representation, a square matrix $\mathcal{A}$ is symmetric if it is equal to its transpose $\mathcal{A}^T$. Consequently a symmetric 3×3 matrix has only six independent components as illustrated by

$$\mathcal{A} = \mathcal{A}^T = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} \qquad (1.126)$$

An anti-symmetric matrix is one that equals the *negative* of its transpose. Consequently a 3×3 anti-symmetric matrix $\mathcal{B}$ has zeros on the main diagonal, and therefore only three independent components as illustrated by

$$\mathcal{B} = -\mathcal{B}^T = \begin{bmatrix} 0 & B_{12} & B_{13} \\ -B_{12} & 0 & B_{23} \\ -B_{13} & -B_{23} & 0 \end{bmatrix} \qquad (1.127)$$

Symmetry properties may be extended to tensors of higher order than two. In general, an arbitrary tensor is said to be symmetric with respect to a pair of indices if the value of the typical component is unchanged by interchanging these two indices. A tensor is anti-symmetric in a pair of indices if an interchange of these indices leads to a change of sign without a change of absolute value in the component. Examples of symmetry properties in higher-order tensors are

(a) $R_{ijkm} = R_{ikjm}$ (symmetric in k and j)

(b) $\epsilon_{ijk} = -\epsilon_{kji}$ (anti-symmetric in k and i)

(c) $G_{ijkm} = G_{jimk}$ (symmetric in i and j; k and m)

(d) $\beta_{ijk} = \beta_{ikj} = \beta_{kji} = \beta_{jik}$ (symmetric in all indices)

1.19 PRINCIPAL VALUES AND PRINCIPAL DIRECTIONS OF SYMMETRIC SECOND-ORDER TENSORS

In the following analysis, only symmetric tensors with real components are considered. This simplifies the mathematics somewhat, and since the important tensors of continuum mechanics are usually symmetric there is little sacrifice in this restriction.

For every symmetric tensor T_{ij}, defined at some point in space, there is associated with each direction (specified by the unit normal n_i) at that point, a vector given by the inner product

$$v_i = T_{ij}n_j \qquad (1.128)$$

Here T_{ij} may be envisioned as a linear vector operator which produces the vector v_i conjugate to the direction n_i. If the direction is one for which v_i is parallel to n_i, the inner product may be expressed as a scalar multiple of n_i. For this case,

$$T_{ij}n_j = \lambda n_i \qquad (1.129)$$

and the direction n_i is called a *principal direction*, or *principal axis* of T_{ij}. With the help of the identity $n_i = \delta_{ij}n_j$, (1.129) can be put in the form

$$(T_{ij} - \lambda \delta_{ij})n_j = 0 \qquad (1.130)$$

which represents a system of three equations for the four unknowns, n_i and λ, associated with each principal direction. In expanded form, the system to be solved is

$$(T_{11} - \lambda)n_1 + T_{12}n_2 + T_{13}n_3 = 0$$
$$T_{21}n_1 + (T_{22} - \lambda)n_2 + T_{23}n_3 = 0 \qquad (1.131)$$
$$T_{31}n_1 + T_{32}n_2 + (T_{33} - \lambda)n_3 = 0$$

Note first that for every λ, the trivial solution $n_i = 0$ satisfies the equations. The purpose here, however, is to obtain non-trivial solutions. Also, from the homogeneity of the system (1.131) it follows that no loss of generality is incurred by restricting attention to solutions for which $n_i n_i = 1$, and this condition is imposed from now on.

For (1.130) or, equivalently, (1.131) to have a non-trivial solution, the determinant of coefficients must be zero, that is,

$$|T_{ij} - \lambda \delta_{ij}| = 0 \qquad (1.132)$$

Expansion of this determinant leads to a cubic polynomial in λ, namely,

$$\lambda^3 - I_T \lambda^2 + II_T \lambda - III_T = 0 \qquad (1.133)$$

which is known as the *characteristic equation* of T_{ij}, and for which the scalar coefficients,

$$I_T = T_{ii} = \text{tr } T_{ij} \text{ (trace of } T_{ij}) \tag{1.134}$$

$$II_T = \tfrac{1}{2}(T_{ii}T_{jj} - T_{ij}T_{ij}) \tag{1.135}$$

$$III_T = |T_{ij}| = \det T_{ij} \tag{1.136}$$

are called the first, second and third *invariants*, respectively, of T_{ij}. The three roots of the cubic (*1.133*), labeled $\lambda_{(1)}, \lambda_{(2)}, \lambda_{(3)}$, are called the *principal values* of T_{ij}. For a symmetric tensor with real components, the principal values are real; and if these values are distinct, the three principal directions are mutually orthogonal. When referred to principal axes, both the tensor array and its matrix appear in diagonal form. Thus

$$\mathbf{T} = \begin{pmatrix} \lambda_{(1)} & 0 & 0 \\ 0 & \lambda_{(2)} & 0 \\ 0 & 0 & \lambda_{(3)} \end{pmatrix} \quad \text{or} \quad T = \begin{bmatrix} \lambda_{(1)} & 0 & 0 \\ 0 & \lambda_{(2)} & 0 \\ 0 & 0 & \lambda_{(3)} \end{bmatrix} \tag{1.137}$$

If $\lambda_{(1)} = \lambda_{(2)}$, the tensor has a diagonal form which is independent of the choice of $\lambda_{(1)}$ and $\lambda_{(2)}$ axes, once the principal axis associated with $\lambda_{(3)}$ has been established. If all principal values are equal, any direction is a principal direction. If the principal values are ordered, it is customary to write them as $\lambda_{(I)}, \lambda_{(II)}, \lambda_{(III)}$ and to display the ordering as in $\lambda_{(I)} > \lambda_{(II)} > \lambda_{(III)}$.

For principal axes labeled $Ox_1^*x_2^*x_3^*$, the transformation from $Ox_1x_2x_3$ axes is given by the elements of the table

	x_1	x_2	x_3
x_1^*	$a_{11} = n_1^{(1)}$	$a_{12} = n_2^{(1)}$	$a_{13} = n_3^{(1)}$
x_2^*	$a_{21} = n_1^{(2)}$	$a_{22} = n_2^{(2)}$	$a_{23} = n_3^{(2)}$
x_3^*	$a_{31} = n_1^{(3)}$	$a_{32} = n_2^{(3)}$	$a_{33} = n_3^{(3)}$

in which $n_i^{(j)}$ are the direction cosines of the jth principal direction.

1.20 POWERS OF SECOND-ORDER TENSORS. HAMILTON-CAYLEY EQUATION

By direct matrix multiplication, the square of the tensor T_{ij} is given as the inner product $T_{ik}T_{kj}$; the cube as $T_{ik}T_{km}T_{mj}$; etc. Therefore with T_{ij} written in the diagonal form (*1.137*), the nth power of the tensor is given by

$$(\mathbf{T})^n = \begin{pmatrix} \lambda_{(1)}^n & 0 & 0 \\ 0 & \lambda_{(2)}^n & 0 \\ 0 & 0 & \lambda_{(3)}^n \end{pmatrix} \quad \text{or} \quad T^n = \begin{bmatrix} \lambda_{(1)}^n & 0 & 0 \\ 0 & \lambda_{(2)}^n & 0 \\ 0 & 0 & \lambda_{(3)}^n \end{bmatrix} \tag{1.138}$$

A comparison of (*1.138*) and (*1.137*) indicates that T_{ij} and all its integer powers have the same principal axes.

Since each of the principal values satisfies (*1.133*), and because of the diagonal matrix form of T^n given by (*1.138*), the tensor itself will satisfy (*1.133*). Thus

$$T^3 - I_T T^2 + II_T T - III_T \mathcal{I} = 0 \tag{1.139}$$

in which $\mathcal{I}$ is the identity matrix. This equation is called the *Hamilton-Cayley equation*. Matrix multiplication of each term in (*1.139*) by T produces the equation,

$$T^4 = I_T T^3 - II_T T^2 + III_T T \tag{1.140}$$

Combining (1.140) and (1.139) by direct substitution,

$$T^4 = (\mathrm{I}_T^2 - \mathrm{II}_T)T^2 + (\mathrm{III}_T - \mathrm{I}_T\mathrm{II}_T)T + \mathrm{I}_T\mathrm{III}_T\mathcal{I} \tag{1.141}$$

Continuation of this procedure yields the positive powers of T as linear combinations of T^2, T and $\mathcal{I}$.

1.21 TENSOR FIELDS. DERIVATIVES OF TENSORS

A *tensor field* assigns a tensor $T(x, t)$ to every pair (x, t) where the position vector x varies over a particular region of space and t varies over a particular interval of time. The tensor field is said to be continuous (or differentiable) if the components of $T(x, t)$ are continuous (or differentiable) functions of x and t. If the components are functions of x only, the tensor field is said to be *steady*.

With respect to a rectangular Cartesian coordinate system, for which the position vector of an arbitrary point is

$$x = x_i\hat{e}_i \tag{1.142}$$

tensor fields of various orders are represented in indicial and symbolic notation as follows,

    (a) scalar field: $\qquad\qquad \phi = \phi(x_i, t) \quad$ or $\quad \phi = \phi(x, t) \tag{1.143}$

    (b) vector field: $\qquad\qquad v_i = v_i(x, t) \quad$ or $\quad v = v(x, t) \tag{1.144}$

    (c) second-order tensor field:

$$T_{ij} = T_{ij}(x, t) \quad \text{or} \quad T = T(x, t) \tag{1.145}$$

Coordinate differentiation of tensor components with respect to x_i is expressed by the differential operator $\partial/\partial x_i$, or briefly in indicial form by ∂_i, indicating an operator of tensor rank one. In symbolic notation, the corresponding symbol is the well-known differential vector operator ∇, pronounced *del* and written explicitly

$$\nabla = \hat{e}_i\frac{\partial}{\partial x_i} = \hat{e}_i\partial_i \tag{1.146}$$

Frequently, partial differentiation with respect to the variable x_i is represented by the *comma-subscript convention* as illustrated by the following examples.

    (a) $\dfrac{\partial\phi}{\partial x_i} = \phi_{,i}$ (d) $\dfrac{\partial^2 v_i}{\partial x_j\,\partial x_k} = v_{i,jk}$

    (b) $\dfrac{\partial v_i}{\partial x_i} = v_{i,i}$ (e) $\dfrac{\partial T_{ij}}{\partial x_k} = T_{ij,k}$

    (c) $\dfrac{\partial v_i}{\partial x_j} = v_{i,j}$ (f) $\dfrac{\partial^2 T_{ij}}{\partial x_k\,\partial x_m} = T_{ij,km}$

From these examples it is seen that the operator ∂_i produces a tensor of order one higher if i remains a free index ((a) and (c) above), and a tensor of order one lower if i becomes a dummy index ((b) above) in the derivative.

Several important differential operators appear often in continuum mechanics and are given here for reference.

$$\operatorname{grad}\phi = \nabla\phi = \frac{\partial\phi}{\partial x_i}\hat{e}_i \quad \text{or} \quad \partial_i\phi = \phi_{,i} \tag{1.147}$$

$$\operatorname{div}v = \nabla\cdot v \qquad\qquad \text{or} \quad \partial_i v_i = v_{i,i} \tag{1.148}$$

$$\operatorname{curl}v = \nabla\times v \qquad\qquad \text{or} \quad \epsilon_{ijk}\partial_j v_k = \epsilon_{ijk}v_{k,j} \tag{1.149}$$

$$\nabla^2\phi = \nabla\cdot\nabla\phi \qquad\qquad \text{or} \quad \partial_{ii}\phi = \phi_{,ii} \tag{1.150}$$

1.22 LINE INTEGRALS. STOKES' THEOREM

In a given region of space the vector function of position, $\mathbf{F} = \mathbf{F}(\mathbf{x})$, is defined at every point of the piecewise smooth curve C shown in Fig. 1-10. If the *differential tangent vector* to the curve at the arbitrary point P is $d\mathbf{x}$, the integral

$$\int_C \mathbf{F} \cdot d\mathbf{x} \;\;\equiv\;\; \int_{\mathbf{x}_A}^{\mathbf{x}_B} \mathbf{F} \cdot d\mathbf{x} \qquad\qquad (1.151)$$

taken along the curve from A to B is known as the *line integral* of F along C. In the indicial notation, (1.151) becomes

$$\int_C F_i \, dx_i \;\;\equiv\;\; \int_{(x_i)_A}^{(x_i)_B} F_i \, dx_i \qquad\qquad (1.152)$$
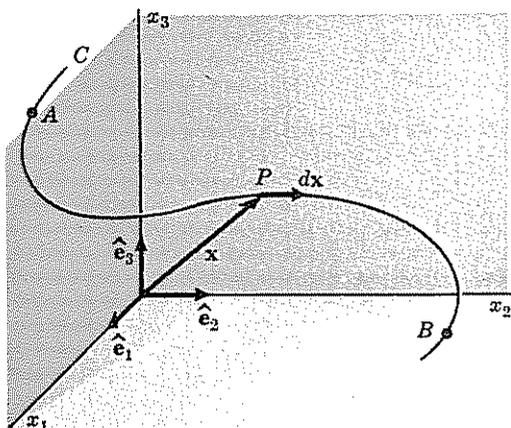

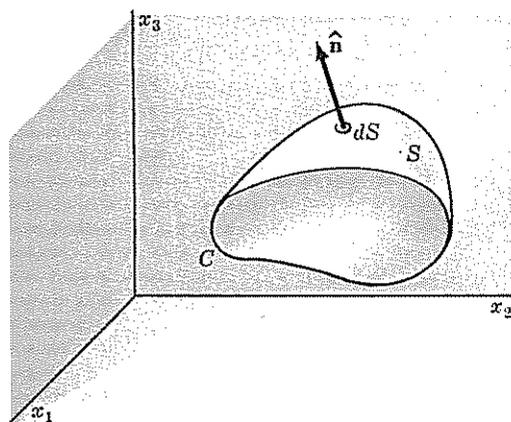Fig. 1-10   Fig. 1-11

Stokes' theorem says that the line integral of $\mathbf{F}$ taken around a closed reducible curve C, as pictured in Fig. 1-11, may be expressed in terms of an integral over any two-sided surface S which has C as its boundary. Explicitly,

$$\oint_C \mathbf{F} \cdot d\mathbf{x} \;\;=\;\; \int_S \hat{\mathbf{n}} \cdot (\nabla \times \mathbf{F}) \, dS \qquad\qquad (1.153)$$

in which $\hat{\mathbf{n}}$ is the unit normal on the positive side of S, and dS is the differential element of surface as shown by the figure. In the indicial notation, (1.153) is written

$$\oint_C F_i \, dx_i \;\;=\;\; \int_S n_i \epsilon_{ijk} F_{k,j} \, dS \qquad\qquad (1.154)$$

1.23 THE DIVERGENCE THEOREM OF GAUSS

The *divergence theorem of Gauss* relates a volume integral to a surface integral. In its traditional form the theorem says that for the vector field $\mathbf{v} = \mathbf{v}(\mathbf{x})$,

$$\int_V \operatorname{div} \mathbf{v} \, dV \;\;=\;\; \int_S \hat{\mathbf{n}} \cdot \mathbf{v} \, dS \qquad\qquad (1.155)$$

where $\hat{\mathbf{n}}$ is the outward unit normal to the bounding surface S, of the volume V in which the vector field is defined. In the indicial notation, (1.155) is written

$$\int_V v_{i,i} \, dV \;\;=\;\; \int_S v_i n_i \, dS \qquad\qquad (1.156)$$

The divergence theorem of Gauss as expressed by (1.156) may be generalized to incorporate a tensor field of any order. Thus for the arbitrary tensor field $T_{ijk\ldots}$ the theorem is written

$$\int_V T_{ijk\ldots,p} \, dV \;\;=\;\; \int_S T_{ijk\ldots} n_p \, dS \qquad\qquad (1.157)$$

Summary of Notation – Diffusion Equation

Tensor:
$$A\frac{\partial c}{\partial t} + \nabla \cdot (-D\nabla c) = R \quad \text{with} \quad \nabla = \begin{Bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{Bmatrix} \quad and \quad \nabla \cdot \nabla = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$$

Matrix:
$$A\dot{c} - \underline{\nabla}^T D \underline{\nabla} c = R \quad \text{with} \quad \underline{\nabla} = \begin{Bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{Bmatrix} \quad and \quad \nabla \cdot \nabla = \underline{\nabla}^T \underline{\nabla} = \nabla^2$$

Indicial:

$$A\frac{\partial c}{\partial t} - D\left(\frac{\partial^2 c}{\partial x_1^2} + \frac{\partial^2 c}{\partial x_2^2} + \frac{\partial^2 c}{\partial x_3^2}\right) = R \quad or \quad A\frac{\partial c}{\partial t} - D\left(\frac{\partial^2 c}{\partial x_i^2}\right) = R \quad or \quad A\dot{c} - Dc_{i,ii} = R \quad or \quad A\frac{\partial c}{\partial t} - Ddiv(gradc) = R$$

Epanded:
$$A\frac{\partial c}{\partial t} - D\left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2}\right) = R$$

Advective-Diffusive Flows

$$A\frac{\partial c}{\partial t} + \nabla \cdot (-D\nabla c) = R - \mathbf{v} \cdot \nabla c$$

Momentum Transfer - Fluid Mechanics – Navier-Stokes Equations (Incompressible)

Tensor:

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v} \cdot \nabla)\mathbf{v} = \mathbf{F} - \nabla P + \mu \nabla^2 \mathbf{v} \quad \text{with} \quad \nabla^2 = \nabla \cdot \nabla = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$$
$$\nabla \cdot \mathbf{v} = 0$$

Expanded:

$$\rho\frac{\partial}{\partial t}\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} + \rho[v_x\frac{\partial}{\partial x} + v_y\frac{\partial}{\partial y} + v_z\frac{\partial}{\partial z}]\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} - \begin{bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{bmatrix}P + \mu[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}]\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}$$

$$[\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}] = 0$$

Momentum Transfer - Solid Mechanics (strain positive in extension)

$$-\nabla \cdot (c\nabla \mathbf{u}) = \mathbf{F}$$

$$-G\nabla^2 u_i - \frac{G}{1-2\nu} u_{k,ki} - \alpha p_{,i} = F_i$$

CONSERVATION OF MASS

Conservation Equation:
$$\frac{d\rho}{dt} + \frac{d}{dx}(\rho v_x) + \frac{d}{dy}(\rho v_y) = 0 \qquad (1)$$

Incompressible Fluid: $\rho = \text{constant}$ (in space and in time)

$$\therefore \ d\rho/dt = 0 \qquad d/dx(\rho v_x) = \rho \, dv_x/dx + v_x \, \overset{0}{\cancel{d\rho/dx}}$$

$$\therefore \qquad \frac{dv_x}{dx} + \frac{dv_y}{dy} = 0 \qquad (2)$$

Slightly Compressible Fluid (Darcian): $(v^2/2g \rightarrow \text{small})$

$$d\rho/dt = \underbrace{d\rho/dP} \ dP/dt \qquad \begin{array}{l} P = \text{reduced pressure} \\ P = p - \rho g z \end{array} \qquad (3)$$

$$v_x = -\frac{k}{\mu}\frac{dP}{dx} \qquad (4)$$

Alternately: $\dfrac{d\rho}{dt} = \dfrac{d}{dt}\left(\rho\dfrac{\forall}{\forall}\right) = \dfrac{1}{V}\dfrac{d}{dt}(\rho\forall) = \dfrac{1}{\forall}\rho\dfrac{d\forall}{dt} + \dfrac{1}{\forall}\forall\overset{0}{\cancel{\dfrac{d\rho}{dt}}} = \rho\dfrac{1}{\forall}\dfrac{d\forall}{dP}\dfrac{dP}{dt}$

$$(5)$$

Combining (4) and (5) into (1)

$$\cancel{\rho}\ \frac{1}{\forall}\frac{d\forall}{dP}\frac{dP}{dt}\ \underset{\underbrace{}}{} = \cancel{\rho}\ \frac{k}{\mu}\frac{d}{dx}\frac{dP}{dx} + \cancel{\rho}\ \frac{k}{\mu}\frac{d}{dy}\frac{dP}{dy} \qquad (6).$$
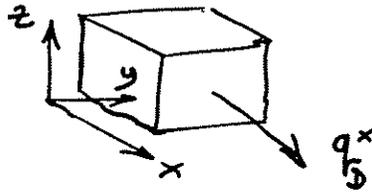
Compressibility, β

$$\beta \, dP/dt = k/\mu \, \nabla^2 P$$

ENERGY CONSERVATION EQUATIONS

$D^* =$ thermal conductivity

$q_D^x =$ Darcy flux in x-direction

CONSERVATION OF ENERGY

$$\frac{\partial(\rho c T)}{\partial t} = -\frac{\partial}{\partial x}\left(q_{Thermal}^x\right) \quad (1)$$

Energy accumulation = Balance (In - Out)

Diffusive flux - Fourier's Law

$$q_{T\,diffusive}^x = -D^* \frac{dT}{dx} \quad (2)$$
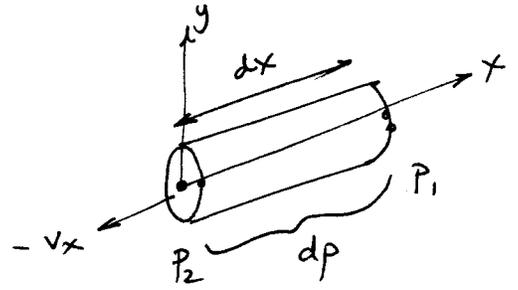
Advective flux.

$$q_{T\,advective}^x = \rho c\, q_D^x\, T \quad (3)$$

Substituting (2) and (3) into (1) gives $\left(\text{noting that } \frac{\partial(\rho c)}{\partial t} = 0\right)$ sometimes

$$\rho c \frac{\partial T}{\partial t} = D^* \frac{\partial^2 T}{\partial x^2} - \rho c\, q_D^x \frac{\partial T}{\partial x} \quad \left(\text{for 1-D problems}\right)$$

DARCY'S LAW

$$v_x = -k \frac{\partial p}{\partial x}$$

Conservation of Mass:

$$\frac{\partial \rho}{\partial t} = - \nabla \cdot \rho \underset{\sim}{v}$$

Eulerian:

$$\frac{\partial \rho}{\partial t} = -\left(v_x \frac{\partial \rho}{\partial x} + v_y \frac{\partial \rho}{\partial y} + v_z \frac{\partial \rho}{\partial z} \right) - \rho \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right)$$

$$v_x = -k \frac{\partial p}{\partial x}$$

Lagrangian:

$$\frac{D\rho}{Dt} = -\rho \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right)$$

$$\frac{D\rho}{Dt} = +\rho k \left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} \right)$$

CONSERVATION OF MOMENTUM

Navier - Stokes Equations — (Steady state).

Momentum: $\rho (v \cdot \nabla) v \doteq -\nabla P + \mu \nabla^2 v$

Convective Accn ┊ Pressure grad   Viscous Forces

$(ma = F)$

$P = p - \rho g z$ (reduced pressure)

For n-dimensional problem, have $n+1$ unknowns: $v_x, v_y + P$ 2-D
                                                    $v_x, v_y, v_z + P$ 3-D

∴ need another eqn

Mass: $\nabla \cdot v = 0$

Simplifications!

(Convective accn → 0) $\rho (v \cdot \nabla) v = 0$ ∴ $\nabla P = \mu \nabla^2 v$

(Stokes Flow).

Inviscid flow ($\mu \to 0$) $\mu \nabla^2 v = 0$ ∴ $\rho (v \cdot \nabla) v = -\nabla P$

(Euler).

Finite Element Method

Chapter 2

Mathematical Concepts and Weighted Residual Techniques

2.1 Introduction

The present chapter begins with a brief resumé of the mathematical equations governing the motion of viscous incompressible fluids. Some closed from solutions of these equations have been presented in many well known texts and will not be discussed. However, the necessary steps in the transformation of these equations into a form suitable for the application of the *F.E.M.* is considered in some detail.

Quite frequently, when utilising a finite difference approach, the governing equations are first written in terms of the basic variables — stream function and vorticity. The pressure distribution is then evaluated subsequent to solving for these variables. Whilst a similar approach is possible using the *F.E.M.*, the authors have, however, followed a policy of solving for the *primitive* variables u,v,p, which are the local point values of velocity in the x cartesian coordinate, y cartesian coordinate directions and the pressure, respectively. Once these primitive variables are evaluated then the distribution of the stream function, vorticity, tractive force etc. can be readily evaluated.

2.2 Two dimensional form of the governing equations

The governing equations are those normally quoted in the literature[1] and a detailed derivation is omitted. However, there are some salient features of the equations which bear repetition and therefore a general outline is included.

2.2.1 Conservation of mass

Equating the quantity of mass entering and leaving an elemental volume, the non-steady flow of a compressible fluid in two dimensions is governed by,

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0 \qquad (2.1)$$

where ρ is the mass density and t represents time.

For an incompressible fluid, $\rho =$ constant, and (2.1) reduces to,

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \qquad (2.2)$$

Since the primitive variables are employed the above equation should be satisfied explicitly, pointwise, everywhere within a flow domain.
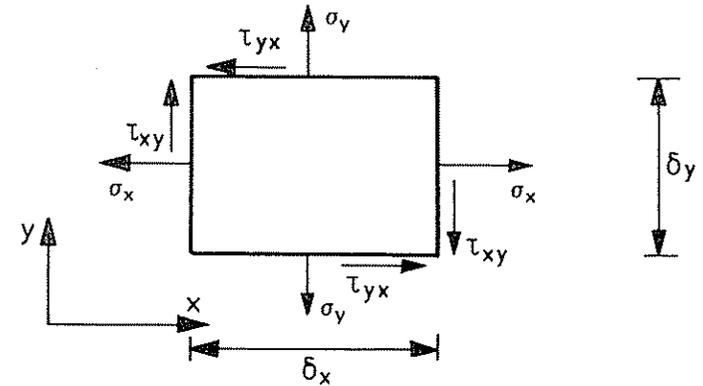
Fig. 2.1   Stress notation and distribution on an elemental area

2.2.2 Conservation of momentum

The conservation of momentum, again obtained by examining the faces on an elemental area of fluid, can be written[2],

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = F_x + \left(\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y}\right) \qquad (2.3)$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = F_y + \left(\frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yx}}{\partial x}\right) \qquad (2.4)$$

where $\sigma_x,\sigma_y,\tau_{xy},\tau_{yx}$ are stress components, Fig. 2.1 and F_x,F_y are body forces in the x,y directions respectively. For a Newtonian fluid these stresses can be related to local pressure and rate of strain via Poisson's Constitutive Law[2]

(For compressible Fluid)

$$\sigma_x = -p + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) + 2\mu\frac{\partial u}{\partial x} \qquad (a)$$

$$\sigma_y = -p + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) + 2\mu\frac{\partial v}{\partial y} \qquad (b)$$

$$\tau_{xy} = \tau_{yx} = \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \qquad (c)$$

See (2.8) for INC

where μ is the molecular viscosity and $\lambda = -\frac{2}{3}\mu$ when the pressure is assumed to be equal, but opposite in sign, to the normal stresses, i.e. the Stokes postulation[2].

Utilising (2.5), (2.3) and (2.4) a form of the Navier Stokes equation assuming constant viscosity is,

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=\frac{1}{\rho}F_x-\frac{1}{\rho}\frac{\partial p}{\partial x}+\frac{1}{3}v\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)+v\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \qquad \text{(a)}$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=\frac{1}{\rho}F_y-\frac{1}{\rho}\frac{\partial p}{\partial y}+\frac{1}{3}v\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)+v\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right) \qquad \text{(b)}$$

$$(2.6)$$

in which $v=\frac{\mu}{\rho}$.

If the fluid is assumed incompressible, $\left(\dfrac{\partial u}{\partial x}+\dfrac{\partial v}{\partial y}\right)=0$, then (2.6) reduces to,

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=\frac{1}{\rho}F_x-\frac{1}{\rho}\frac{\partial p}{\partial x}+v\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \qquad \text{(a)}$$

and

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=\frac{1}{\rho}F_y-\frac{1}{\rho}\frac{\partial p}{\partial y}+v\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right) \qquad \text{(b)}$$

$$(2.7)$$

Two points are worthy of note at this juncture. The first is that if the incompressibility condition is invoked the stress equations now become,

$$\sigma_x=-p+2\mu\frac{\partial u}{\partial x} \qquad \text{(a)}$$

$$\sigma_y=-p+2\mu\frac{\partial v}{\partial y} \qquad \text{(b)}$$

and

$$\tau_{xy}=\tau_{yx}=\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right) \qquad \text{(c)}$$

$$(2.8)$$

which can be used to evaluate the local stress at boundaries or along a line within the fluid. Although the program developed in later chapters is specifically formulated to solve the steady state Navier-Stokes equations, clearly equations of the form,

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=\frac{1}{\rho}F_x-\frac{1}{\rho}\frac{\partial p}{\partial x} \qquad \text{(a)}$$

$$\text{(EULER)}$$

$$\text{(b)}$$

$$u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=\frac{1}{\rho}F_y-\frac{1}{\rho}\frac{\partial p}{\partial y}$$

$$(2.9)$$

when the fluid is assumed inviscid, and, on ignoring the convective terms,

$$0=\frac{1}{\rho}F_x-\frac{1}{\rho}\frac{\partial p}{\partial x}+v\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \qquad \text{(a)}$$

$$\text{(STOKES)}$$

$$0=\frac{1}{\rho}F_y-\frac{1}{\rho}\frac{\partial p}{\partial y}+v\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right) \qquad \text{(b)}$$

$$(2.10)$$

can also be analysed.

The general *steady state* equations, which are analysed in detail are,

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=\frac{1}{\rho}F_x-\frac{1}{\rho}\frac{\partial p}{\partial x}+v\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \qquad \text{(a)}$$

$$u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=\frac{1}{\rho}F_y-\frac{1}{\rho}\frac{\partial p}{\partial y}+v\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right) \qquad \text{(b)}$$

$$(2.11)$$

where both the convective terms and viscous forces are retained. In the coordinate system normally adopted the '*y*' direction corresponds to the vertical. Usually, the body force $F_x=0$ and $F_y=-\rho g$ per unit volume of fluid where g is the gravitational acceleration. For the examples cited in the text both F_x and F_y are assumed to be zero, although provision is made for their inclusion in the computer program.

2.2.3 Vorticity-stream function form of the governing equations

As stated earlier a form of the governing equations which can be used when an analysis is conducted by either the finite difference or finite element method is commonly called the *vorticity-stream function* formulation. The essential steps in the derivation of these equations will now be outlined. Eliminating the pressure from (2.11) by differentiating (2.11a) with respect to y and (2.11b) with respect to x, adding and introducing the definition for vorticity,

$$\omega=-\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right) \qquad (2.12)$$

we obtain the generalised steady state momentum equation in terms of vorticity,

$$u\frac{\partial \omega}{\partial x}+v\frac{\partial \omega}{\partial y}=v\left(\frac{\partial^2 \omega}{\partial x^2}+\frac{\partial^2 \omega}{\partial y^2}\right) \qquad (2.13)$$

Defining the velocities in terms of a stream function, ψ,

$$u = \frac{\partial \psi}{\partial y} \qquad \text{(a)}$$

and

$$v = -\frac{\partial \psi}{\partial x} \qquad \text{(b)}$$

$$\text{(2.14)}$$

such that the continuity equation is satisfied automatically and the vorticity can be re-defined,

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = -\omega \qquad \text{(2.15)}$$

such that (2.13) becomes,

$$\frac{\partial \psi}{\partial y} \cdot \frac{\partial}{\partial x}(\nabla^2 \psi) - \frac{\partial \psi}{\partial x} \cdot \frac{\partial}{\partial y}(\nabla^2 \psi) = \nu \nabla^4 \psi \qquad \text{(2.16)}$$

When analysing for the spatial distribution of stream function and vorticity the following form of the equations are widely employed,

$$\nabla^2 \psi = -\omega \qquad \text{(2.17)}$$

and re-writing (2.13),

$$\nu \nabla^2 \omega = \frac{\partial \psi}{\partial y} \cdot \frac{\partial \omega}{\partial x} - \frac{\partial \psi}{\partial x} \cdot \frac{\partial \omega}{\partial y} \qquad \text{(2.18)}$$

Some early solutions, utilising the finite element method, to these equations were published by Baker[3] and Cheng[4]. It must be noted, however, that this choice of variables, as opposed to the primitive form, has the associated problem of defining vorticity boundary conditions.

Although the methods adopted for solving equations (2.17) and (2.18) is not outlined in this text, it is useful to note that if the velocity distribution is known then both values of stream function and vorticity can be evaluated quite readily from (2.12) and (2.15).

2.3 Axisymmetric flow

Flow of a fluid through pipes is a particularly common occurrence. This quasi-three dimensional situation can be described by equations similar to those already present for two dimensional flow, providing there is no rotation about the axis of symmetry.

Adopting a right hand cylindrical coordinate system, Fig. 2.2, where x is measured along the longitudinal axis of the duct, r measured radially and Φ the azimuth angle on a plane normal to the longitudinal axis.
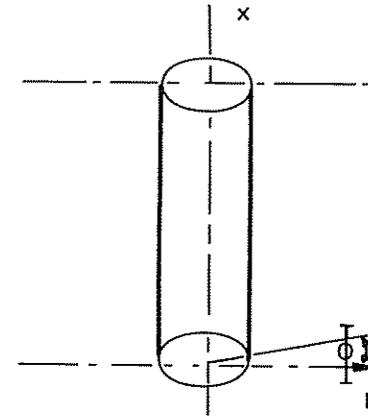
Fig. 2.2   Cylindrical co-ordinate system — axisymmetric flow

2.3.1 Conservation of mass

Assuming the flow to be unidirectional along the x axis such that all variations with respect to Φ are zero. The steady state equation for axisymmetric incompressible flow is,

$$\frac{\partial v}{\partial r} + \frac{v}{r} + \frac{\partial u}{\partial x} = 0 \qquad \text{(2.19)}$$

where u denotes the velocity in the x, axial, direction and v in the orthogonal direction r. This, as in (2.2), involves only two primitive variables.

2.3.2 Conservation of momentum

Again assuming steady state incompressible flow the equations depicting conservation of momentum are[1],

$$\rho\left(u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial r}\right) = F_x + \frac{1}{r}\left(\frac{\partial}{\partial x}(r\sigma_x) + \frac{\partial}{\partial r}(r\tau_{rx})\right) \qquad \text{(2.20)}$$

$$\rho\left(u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial r}\right) = F_r + \frac{1}{r}\left(\frac{\partial}{\partial x}(r\tau_{xr}) + \frac{\partial}{\partial r}(r\sigma_r)\right) - \frac{\sigma_\Phi}{r} \qquad \text{(2.21)}$$

The body forces in the x and r directions are now represented by F_x and F_r respectively. As before the stresses can be written as,

$$\sigma_x = -p + 2\mu\frac{\partial u}{\partial x} \qquad \text{(a)}$$

$$\sigma_r = -p + 2\mu\frac{\partial v}{\partial r} \qquad \text{(b)}$$

$$\sigma_\Phi = -p + 2\mu\frac{v}{r} \qquad \text{(c)}$$

and

$$\tau_{xr} = \tau_{rx} = \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial r}\right) \tag{d}$$

$$\tag{2.22}$$

Combining (2.21) and (2.22) the required form of the momentum equation is,

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial r} = \frac{1}{\rho}F_x - \frac{1}{\rho}\frac{\partial p}{\partial x} + v\left(\frac{\partial^2 u}{\partial x^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial r^2}\right) \tag{a}$$

and

$$u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial r} = \frac{1}{\rho}F_y - \frac{1}{\rho}\frac{\partial p}{\partial r} + v\left(\frac{\partial^2 v}{\partial x^2} + \frac{1}{r}\frac{\partial v}{\partial r} - \frac{v^2}{r} + \frac{\partial^2 v}{\partial r^2}\right) \tag{b}$$

$$\tag{2.23}$$

which are, basically, quite similar to (2.11). It can be stated, therefore, that the principles developed for solving two dimensional problems would be equally applicable to axisymmetric flow and this facility is included in the programs subsequently presented.

The remaining quantities usually required, the stream function and vorticity, can, once the velocity is known, be evaluated using the following definitions,

$$u = \frac{1}{r}\frac{\partial \psi}{\partial r} \tag{a}$$

$$v = -\frac{1}{r}\frac{\partial \psi}{\partial x} \tag{b}$$

$$\tag{2.24}$$

and

$$\frac{1}{r}\left(\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{r}\frac{\partial \psi}{\partial r} + \frac{\partial^2 \psi}{\partial r^2}\right) = -w = \left(\frac{\partial u}{\partial r} - \frac{\partial v}{\partial x}\right) \tag{2.25}$$

2.4 Method of weighted residuals

Having defined the governing equations the method chosen for solution depends, largely, on the physical problem being analysed. If the flow domain and boundary conditions are well posed then an analytical solution could well be possible. For the majority of flow problems of practical interest, however, the flow domain is geometrically complex and recourse has to be made to an approximate method which may then be amenable to direct analysis. The authors have chosen to limit discussions to one, the Method of Weighted Residuals, which has been used quite extensively in the field of fluid mechanics.
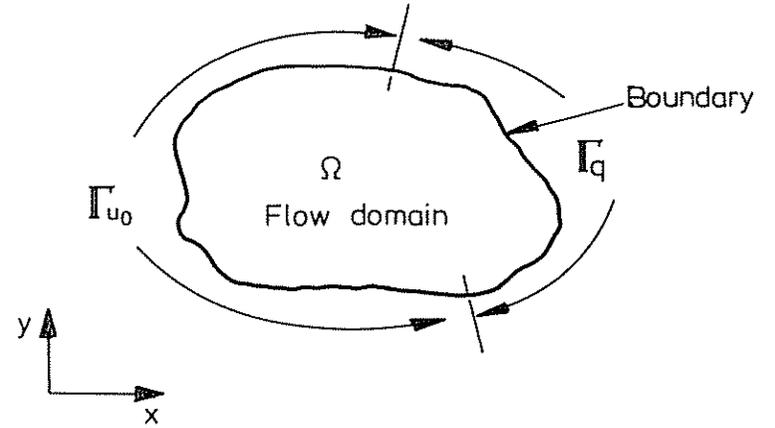
Fig. 2.3   Definition of flow domain and boundary type

Weighted residual methods are, in essence, numerical techniques which can be used to solve a single or set of partial differential equations. Consider such a set, say representative of (2.2) and (2.11),

$$\mathcal{L}(\mathbf{u}) = \oint \tag{2.26}$$

in a domain Ω, Fig. 2.3, where $\mathbf{u}$ is the exact solution and may represent a single variable or a column vector of variables. The two prevalent type boundary conditions are,

essential (Dirichlet) $G(\mathbf{u}) = u_0$ on Γ_{u_0} (a)

where the value of the variable is prescribed, and

natural (Neumann) $S(\mathbf{u}) = q$ on Γ_q (b)

$$\tag{2.27}$$

where at least the first order gradient in the variable is prescribed.

The relevance and full explanation of each type boundary condition will become apparent when considering a specific example. The first step in the application of the weighted residual procedure is to assume that $\mathbf{u}$ can be approximated over the whole domain by,

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i \beta_i \tag{2.28}$$

where α are functions described in terms of independent variables, such as spatial coordinates (x,y), and β are undetermined parameters.

Utilising this approximation and incorporating (2.28) in (2.26) results in an error or *residual*, ε, such that

$$\varepsilon = \mathscr{L}(\hat{u}) - \int \neq 0 \tag{2.29}$$

where ε is exactly zero when $\hat{u} = u$ i.e. an exact solution is possible.

In order to make ε identically zero a set of 'arbitrary' weighting functions, W, are employed such that over the whole domain, Ω,

$$\int_{\Omega} W\varepsilon \, d\Omega = 0 \tag{2.30}$$

If the number of unknown parameters is s and there are s linearly independent weighting functions and (2.30) can be re-written,

$$\int_{\Omega} W_k \varepsilon \, d\Omega = \int_{\Omega} W_k (\mathscr{L}(\hat{u}) - \int) \, d\Omega = 0 \quad k = 1,2,3\ldots s \tag{2.31}$$

The only limitation, at this stage, placed on W_k is that this must be, positive, single valued and finite.

There are a number of ways in which the above concepts can be utilised to transform the differential equations into a form where finite element techniques can be adopted with effect. These have been expounded in various texts[5,6,7,8] and the deliberate policy of confining for simplicity, the present text to one method will again be invoked and only the Galerkin[9] method will be considered.

2.4.1 The Galerkin weighted residual method

Before embarking on the main objective of this section a further brief introduction must be given to the commonly adopted concept of trial or shape functions in a finite element context.

The technique of defining approximated values of the required variable via a discrete summation was introduced in (2.28). The approximate values were defined in terms of some functions α and discrete values β. This applied over the whole domain under consideration in which s refers to the total number of discrete values. If we now refine this concept and subdivide the domain into elements, Fig. 2.4, the variable value within that subregion can now be defined in terms of discrete values on the boundary of or within that region,

$$\hat{u} = \sum_{i=1}^{n} N_i \beta_i \tag{2.32}$$

where N are a set of trial functions written in terms of local coordinates associated with n discrete values within or on the boundary of an element. Each element will, normally, possess a unique set of equations and β is now confined to each element.
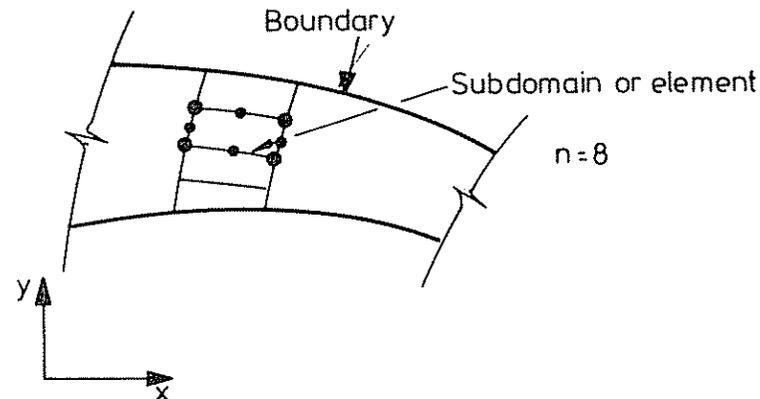

Fig. 2.4   Definition of subdomain or element

The residual now becomes,

$$\varepsilon = \mathscr{L}\left(\sum_{i=1}^{n}(N_i \beta_i)\right) - \int \tag{2.33}$$

such that (2.31) can be rewritten,

$$\int_{\Omega} W_k \left(\mathscr{L}\left(\sum_{i=1}^{n} N_i \beta_i\right) - \int\right) d\Omega = 0 \quad k = 1,2\ldots s \tag{2.34}$$

In the Galerkin method the same approximating functions are used for the weighting and trial functions, i.e. $W_k = N_k$ and the generalised equation is,

$$\int_{\Omega} N_k \left(\mathscr{L}\left(\sum_{i=1}^{n} N_i \beta_i\right) - \int\right) d\Omega = 0 \tag{2.35}$$

where orthogonalisation has been effected with the same functions.

Example: Flow between parallel plates

The example chosen is that of flow between infinite parallel plates, Fig. 2.5. which has well known exact analytical solutions. The flow is assumed to be fully developed and subject to the following boundary conditions,

$$u\left(x, -\frac{h}{2}\right) = 0, \quad u\left(x, \frac{h}{2}\right) = 0$$

$$v = 0 \text{ for all } x \text{ and } -\frac{h}{2} \leqslant y \leqslant \frac{h}{2}$$
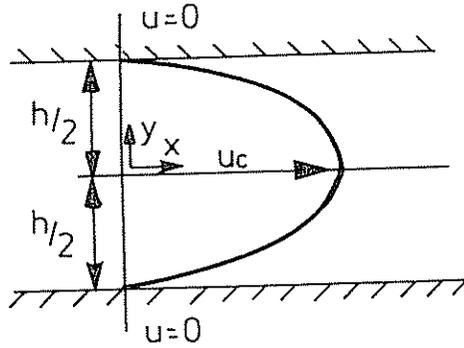
$$\tag{E2.1.1}$$

Fig. 2.5 Laminar flow between stationary parallel plates

The continuity equation,

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{E2.1.2}$$

together with the steady state momentum equation,

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + v\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{E2.1.3}$$

are the governing equations for the laminar flow under consideration. For the steady state fully developed conditions imposed then the convective terms are zero and (E2.1.3) can be written,

$$0 = -\frac{1}{\rho}\frac{\partial p}{\partial x} + v\frac{\partial^2 u}{\partial y^2} \tag{E2.1.4}$$

Integrating twice with respect to y and applying the boundary conditions (E2.1.1), equation (E2.1.4) becomes,

$$u = -\frac{1}{2\mu}\frac{\partial p}{\partial x}\left(\frac{h^2}{4} - y^2\right) \tag{E2.1.5}$$

and

$$u_c = -\frac{h^2}{8\mu} \cdot \frac{\partial p}{\partial x} \tag{E2.1.6}$$

where u_c is the centre-line velocity.

A trial function which leads to exact answers for the current example is,

$$v = \begin{bmatrix} 1 & y^2 \end{bmatrix} \cdot \begin{Bmatrix} \alpha_1 \\ \alpha_2 \end{Bmatrix} \qquad \begin{aligned} N_1 &= 1 \\ N_2 &= y^2 \end{aligned} \tag{E2.1.7}$$

where both α_1 and α_2 are, as yet, unknown constants.

The Galerkin weighted residual leads to,

$$\int_{-h/2}^{h/2} W_i\left(-\frac{1}{\rho}\cdot\frac{\partial p}{\partial x} + v\frac{\partial^2 u}{\partial x^2}\right)\mathrm{d}y = 0 \tag{E2.1.8}$$

inserting, when

$$\begin{cases} i=1, & W_1 = 1 \\ i=2, & W_2 = y^2 \end{cases}$$

For the condition $i=1$, $N_1 = 1$ gives

$$\int_{-h/2}^{+h/2}\left(-\frac{1}{\rho}\cdot\frac{\partial p}{\partial x} + 2v\alpha_2\right)\mathrm{d}y = 0 \tag{E2.1.9a}$$

and when $i=2$, $N_2 = y^2$

$$\int_{-h/2}^{+h/2}\left(-\frac{1}{\rho}\cdot\frac{\partial p}{\partial x} + 2v\alpha_2\right)y^2\,\mathrm{d}y = 0 \tag{E2.1.9b}$$

On integrating (E2.1.9a) or (E2.1.9b), this trivial example results in,

$$-\frac{1}{\rho}\frac{\partial p}{\partial x} + 2v\alpha_2 = 0$$

and

$$\alpha_2 = \frac{1}{2\rho v}\frac{\partial p}{\partial x} = \frac{1}{2\mu}\frac{\partial p}{\partial x} \tag{E2.1.10}$$

such that

$$u = \alpha_1 + \frac{1}{2\mu}\frac{\partial p}{\partial x}y^2 \tag{E2.1.11}$$

The other term in equation (E2.1.11) can be evaluated using the boundary conditions,

$$u = 0, \qquad y = \pm\frac{h}{2}$$

which gives,

$$\alpha_1 = \frac{1}{2\mu}\frac{\partial p}{\partial x}\frac{h^2}{4}$$

and (E2.1.11) becomes,

B

$$u = \left(y^2 - \frac{h^2}{4} \right) \frac{1}{2\mu} \frac{\partial p}{\partial x} \tag{E2.1.12}$$

which, as expected, is the exact solution.

Repeating the same example but now choosing an arbitrary function which simply satisfies the boundary conditions, e.g.

$$u = \alpha \cos\left(\frac{\pi y}{h} \right) \tag{E2.1.13}$$

the equivalent equation to (2.31) is,

$$\int_{-h/2}^{h/2} \left(-\frac{1}{\rho} \frac{\partial p}{\partial x} - \alpha \frac{\pi^2}{h^2} \cos\left(\frac{\pi y}{h} \right) \right) \cos\left(\frac{\pi y}{h} \right) \, dy = 0 \tag{E2.1.14}$$

after integration and imposing the limits indicated,

$$\alpha = \frac{4h^2}{\pi^3 \mu} \frac{\partial p}{\partial x} \tag{E2.1.15}$$

which results in the velocity distribution

$$u = \frac{4h^2}{\pi^3 \mu} \frac{\partial p}{\partial x} \cos\left(\frac{\pi y}{h} \right) \tag{E2.1.16}$$

Introducing numerical values,

$$h = 0.1 \text{ metres}$$

$$\frac{\partial p}{\partial x} = -5.0 \times 10^{-3} \ N/m^2$$

and

$$\mu = 10^{-3} \ Ns/m^2$$

leads to the comparison shown in Table 2.1 between the exact and the approximate solution.

It is evident from Table 2.1 that even with a very crude approximation quite reasonable results can be obtained.

The above example was confined to the case where the operators are self adjoint and only essential boundary conditions imposed. Generally, both the trial and weighting functions must be such that the $(k-1)^{th}$ derivative is continuous, where k is the order of differentiation of governing differential equation. For the example problem chosen this can be demonstrated by considering Fig. 2.6 where the original trial function results in an integrable second order differential. It is immediately apparent that this is a minimum requirement. Such a function is said to be C_1 continuous. Generally a pth order derivative would require C_{p-1} continuity for the resulting weighted residual to be integrable.

Table 2.1

| | Velocity m/sec | |
| | Exact | Weighted residual |
y		
0.0	6.25×10^{-3}	6.45×10^{-3}
0.01	6.0×10^{-3}	6.13×10^{-3}
0.02	5.25×10^{-3}	5.22×10^{-3}
0.03	4.00×10^{-3}	3.79×10^{-3}
0.04	2.25×10^{-3}	1.99×10^{-3}
0.05	0.0	0.0

Second derivatives square integrable
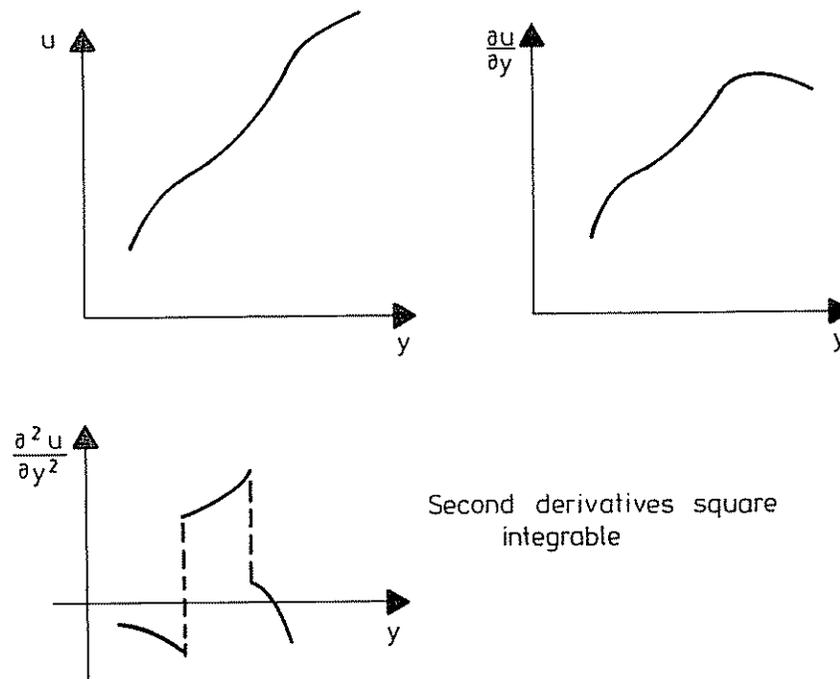
Fig. 2.6

The above requirement leads to the conclusion that obvious advantages would be gained if the order of the governing equation were reduced. This would result in a lower order requirement in both the trial and weighting functions. This is exploited in the following section.

2.5 'Weak' formulation of the governing equations

Starting again with the Galerkin form of the weighted residual process applied to the general operator,

$$\iint_\Omega (\ \mathscr{L}(\mathbf{u}) - \wp)W_i \ \mathrm{d}x \ \mathrm{d}y = 0 \qquad (2.36)$$

or

$$\iint_\Omega \varepsilon W_i \ \mathrm{d}x \ \mathrm{d}y = 0 \qquad (2.37)$$

subject to the usual essential and natural boundary conditions.

The sequence required to reduce the order of a governing equation can be illustrated by considering the second order heat conduction equation for a homogeneous conducting medium,

$$\frac{\partial}{\partial x}\left(K\frac{\partial \varphi}{\partial x}\right) + \frac{\partial}{\partial y}\left(K\frac{\partial \varphi}{\partial y}\right) + Q = 0 \qquad (2.38)$$

subject to the following boundary conditions

essential $\varphi = \bar{\varphi}$ on boundary Γ_ϕ

and

natural $q = \dot{q}$ on boundary Γ_q

$$(2.39)$$

The weighted residual form of this equation is,

$$\iint_\Omega W_i\left(\frac{\partial}{\partial x}\left(K\frac{\partial \varphi}{\partial x}\right) + \frac{\partial}{\partial y}\left(K\frac{\partial \varphi}{\partial y}\right) + Q\right) \mathrm{d}x \ \mathrm{d}y \qquad (2.40)$$

Integrating (2.40) by parts with respect to φ and W_i results in,

$$-\iint_\Omega\left(\frac{\partial W_i}{\partial x}K\frac{\partial \varphi}{\partial x} + \frac{\partial W_i}{\partial y}K\frac{\partial \varphi}{\partial y} - QW_i\right) \mathrm{d}x \ \mathrm{d}y + \int_\Gamma W_iK\frac{\partial \varphi}{\partial n} = 0 \quad (2.41)$$

where Γ represents the complete boundary. (2.41) can be re-written,

$$\iint_\Omega\left(\frac{\partial W_i}{\partial x}K\frac{\partial \varphi}{\partial x} + \frac{\partial W_i}{\partial y}K\frac{\partial \varphi}{\partial y} - QW_i\right) \mathrm{d}x \ \mathrm{d}y$$

$$(2.42)$$

$$-\int_{\Gamma_\bullet} W_iK\frac{\partial \varphi}{\partial n} \ \mathrm{d}\Gamma - \int_{\Gamma_\bullet} W_iK\frac{\partial \varphi}{\partial n} \ \mathrm{d}\Gamma = 0$$

In equation (2.42) the second integral is re-written as,

$$\int_{\Gamma_\bullet} W_i\dot{q} \ \mathrm{d}\Gamma$$

In the above equations

$$\frac{\partial \varphi}{\partial n} \equiv l_x\frac{\partial \varphi}{\partial x} + l_y\frac{\partial \varphi}{\partial y}$$

where l_x, l_y are the components of the unit outward normal vectors at the boundary.

The boundary integral terms in (2.42) require some further explanation before proceeding to demonstrate the application by example. Clearly, the physical significance of $\dot{q}$ can be interpreted as an outward flux. On common faces between two elements the nett flux will, if the distribution of φ is correct, be zero. Therefore, for elements within a domain under consideration the nett effect of the third terms in equation (2.42) is zero and can, for all intents and purposes, be ignored. On that part of the boundary where the values of φ are defined,

$$\varphi = \bar{\varphi}$$

then the third term becomes redundant since the equation would be eliminated from the solution technique. On these boundaries the flux can, however, be evaluated from

$$\int_{\Gamma_\bullet} W_iK\frac{\partial \varphi}{\partial n} \ \mathrm{d}\Gamma \qquad (2.43)$$

Therefore, without loss of generality, (2.42) can be re-written,

$$\iint_\Omega\left(\frac{\partial W_i}{\partial x}K\frac{\partial \varphi}{\partial x} + \frac{\partial W_i}{\partial y}K\frac{\partial \varphi}{\partial y} - QW_i\right) \mathrm{d}x \ \mathrm{d}y - \int_{\Gamma_\bullet} W_i\dot{q} \ \mathrm{d}\Gamma = 0 \qquad (2.44)$$

where the boundary integral is only retained on boundaries where a flux type boundary condition is imposed.

Example   Flow between parallel plates — weak formation

The main objective of the present example is to introduce the weak formulation incorporating the 'gradient' type boundary condition. Again consider the Couette type flow where each wall is stationary, Fig. 2.5, utilising the same trial functions,

$$u = [1 \ \ y^2]\begin{Bmatrix}\alpha_1 \\ \alpha_2\end{Bmatrix} \qquad \begin{matrix}N_1 = 1 \\ N_2 = y^2\end{matrix} \Big\} \qquad (E2.2.1)$$

with the boundary conditions,

essential $u = u_c$ at $y = 0$

natural $\dfrac{\partial u}{\partial n} = \dfrac{\partial u}{\partial y} = \dot{q}$ at $y = \dfrac{h}{2}$

Invoking the Galerkin weighted residual approach the relevant weak formulation leads to

$$\int_0^{h/2} \left(\frac{1}{\rho} \frac{\partial p}{\partial x} W_i + v \frac{\partial W_i}{\partial y} \frac{\partial u}{\partial y} \right) dy - \int_{\Gamma_q} v W_i \dot{q} \, d\Gamma = 0 \qquad \text{(E2.2.2)}$$

which, when (E2.2.1) is used becomes,

$$\int_0^{h/2} \left(\frac{1}{\rho} \frac{\partial p}{\partial x} W_i + v \frac{\partial W_i}{\partial y} 2\alpha_2 y \right) dy - \int_{\Gamma_q} v W_i \dot{q} \, d\Gamma = 0 \qquad \text{(E2.2.3)}$$

when

$$i = 1, \quad \frac{\partial W_i}{\partial y} = 0$$

and, for this condition the *L.H.S.* of equation (E2.2.3) gives,

$$\int_0^{h/2} \left(\frac{1}{\rho} \frac{\partial p}{\partial x} 1 + 0 \right) dy - \int_{\Gamma_q} v \dot{q} \, d\Gamma = 0 \qquad \text{(E2.2.4)}$$

Integrating and applying the limits of integration,

$$\frac{h}{2\rho} \frac{\partial p}{\partial x} - \int_{\Gamma_q} v \dot{q} \, d\Gamma = 0 \qquad \text{(E2.2.5)}$$

When

$$i = 2, \quad \frac{\partial W_i}{\partial y} = 2y$$

and equation (E2.2.3) now gives,

$$\int_0^{h/2} \left(\frac{1}{\rho} \frac{\partial p}{\partial x} y^2 + v 4\alpha_2 y^2 \right) dy - \int_{\Gamma_q} v y^2 \dot{q} \, d\Gamma = 0 \qquad \text{(E2.2.6)}$$

which, upon integration, results in

$$\frac{h^3}{24\rho} \frac{\partial p}{\partial x} + \frac{v}{6} h^3 \alpha_2 - \frac{h^2}{4} \int_{\Gamma_q} v \dot{q} \, d\Gamma = 0 \qquad \text{(E2.2.7)}$$

Adding equations (E2.2.5) and (E2.2.7) leads to a general expression of the

$$\frac{h}{2\rho} \frac{\partial p}{\partial x} \left(1 + \frac{h^2}{12} \right) + \frac{v}{6} h^3 \alpha_2 - v \left(\frac{h^2}{4} + 1 \right) \frac{\partial u}{\partial y} \bigg|_{0,h/2} = 0 \qquad \text{(E2.2.8)}$$

in which the last term refers to the boundary at $0,h/2$. Using the same variable values as before i.e.

$$h = 0.1 \text{ metres}$$

$$\frac{\partial p}{\partial x} = -5.0 \times 10^{-3} \ N/m^2$$

and

$$\mu = 10^{-3} \ Ns/m^2$$

with the additional boundary conditions,

$$u = 6.25 \times 10^{-3} \text{ m/sec at } y = 0$$

and

$$\frac{\partial u}{\partial y} = -0.25 \text{ at } y = h/2$$

Substituting these values into equation (E2.2.8) the value of α_2 is found to be -2.5. Note that this could have been obtained from (E2.2.7) only.

Substituting into (E2.2.1) and using the velocity boundary condition at $y = 0$, $\alpha_1 = 6.25 \times 10^{-3}$. The general equation for the velocity is, therefore,

$$u = 6.25 \times 10^{-3} - 2.5 \ y^2 \qquad \text{(E2.2.9)}$$

which, as expected, yields the exact answers.

The question still remains, however, regarding the compatability of results when a lower order trial function is utilised in conjunction with the 'weak' formulation. This can be demonstrated by assuming an equation of the form,

$$u = [1 \quad y] \cdot \begin{Bmatrix} \alpha_1 \\ \alpha_2 \end{Bmatrix} \qquad \text{(E2.2.10)}$$

which would be too low an order when the weak formulation is not utilised. Using the weak formulation this gives rise to the equation,

$$\int_0^{h/2} \left(\frac{1}{\rho} \frac{\partial p}{\partial x} y + v \alpha_2 \right) dy - v \frac{h}{2} \frac{\partial u}{\partial y} \bigg|_{0,h/2} = 0 \qquad \text{(E2.2.11)}$$

Table 2.2   Comparison of exact velocity profile and weak formulation with minimum order of trial function

y	Velocity m/sec	
	Exact	Weak weighted residual
0.0	6.25×10^{-3}	6.25×10^{-3}
0.01	6.0×10^{-3}	5.00×10^{-3}
0.02	5.25×10^{-3}	3.75×10^{-3}
0.03	4.00×10^{-3}	2.5×10^{-3}
0.04	2.25×10^{-3}	1.25×10^{-3}
0.05	0.0	0.0

or

$$\frac{h}{4\mu}\frac{\partial p}{\partial x} + v\alpha_2 - v\frac{\partial u}{\partial y}\bigg|_{0,h/2} = 0 \qquad (E2.2.12)$$

Substituting values we have,

$$\alpha_2 = -0.125$$

and

$$\alpha_1 = 6.25 \times 10^{-3}$$

The resulting equation for the velocity distribution, now linear, is

$$u = 6.25 \times 10^{-3} - 0.125y \qquad (E2.2.13)$$

A comparison with the exact velocity distribution is shown in Table 2.2, which illustrates the considerable errors which have been incurred when a linear profile is assumed.
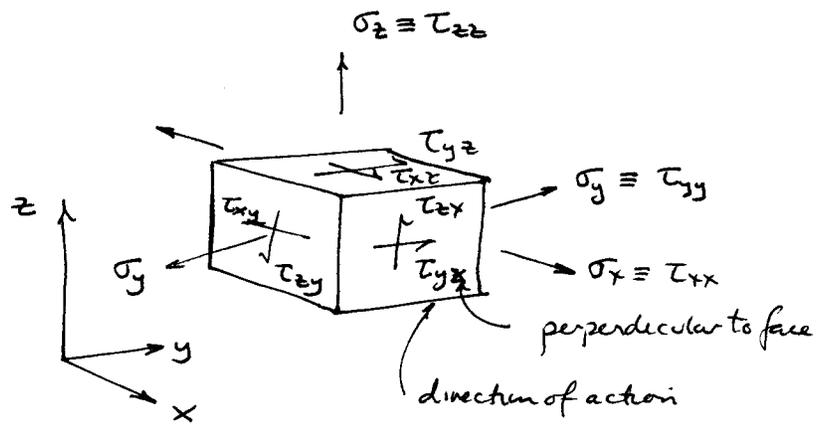
The concept of 'weak' formulation can be extended to include higher order equations, for instance the biharmonic equation[9], where the natural boundary conditions assume considerable importance. Further reading on this topic is left, however, to the interested reader. The stage has now been reached where the weighted residual technique and the *F.E.M.* can be combined leading to a general integrable form of equation where both trial and shape functions are defined explicitly.

References

1. SCHLICHTING, H. "Boundary Layer Theory", McGraw Hill, New York, 1960.
2. GOLDSTEIN, S. (Ed.) "Modern Developments in Fluid Dynamics", Oxford Press, 1938.
3. BAKER, A. J. "A Finite Element Solution Algorithm for Viscous Incompressible Fluid Dynamics", *Int. Journ. Num. Meth. in Eng.*, Vol. 6, 1973.
4. CHENG, R. T. "Numerical Solution of the Navier-Stokes Equations by the Finite Element Method", *Physics of Fluids*, 15, 1972.

5. HINTON, E. and OWEN, D. R. J. "Finite Element Programming", Academic Press, 1977.
6. HINTON, E. and OWEN, D. R. J. "An Introduction to Finite Element Computations", Pineridge Press, 1980.
7. HINTON, E. and OWEN, D. R. J. "A Simple Guide to Finite Elements", Pineridge Press, 1980.
8. TAYLOR, C. and MORGAN, K. (Eds.) "Recent Advances in Numerical Methods in Fluids", Pineridge Press, 1980.
9. CRANDALL, S. H. "Engineering Analysis", McGraw Hill, 1966.

EQUILIBRIUM EQUATIONS

$\sigma_z \equiv \tau_{zz}$

τ_{yz}

τ_{xt}

τ_{zx}

$\sigma_y \equiv \tau_{yy}$

τ_{xy}

τ_{zy}

σ_y

τ_{yx}

$\sigma_x \equiv \tau_{xx}$

perpendicular to face

direction of action

Force balance yields:

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial t} = \rho g_x + \frac{\partial^2 v_x}{\partial t^2}$$

(3 equations)

Consistent with conservation of momentum:

$$\underbrace{\frac{\partial}{\partial t} \rho v_x + \left(\frac{\partial}{\partial x} \rho v_x v_x + \frac{\partial}{\partial y} \rho v_y v_x + \frac{\partial}{\partial t} \rho v_z v_x \right)}_{\rho \, D v_x / Dt}$$

$$= - \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial t} \right) - \frac{\partial p}{\partial x} + \rho g_x$$

$$\rho \, \cancel{\frac{D v_x}{Dt}}^{\,0} = - \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial t} \right) - \cancel{\frac{\partial p}{\partial x}}^{\,0} + \rho g_x$$

HOOKE'S LAW
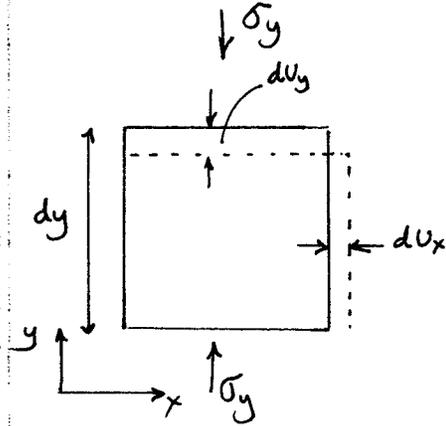
$$\varepsilon_y = \frac{\partial u_y}{\partial y} \doteq \frac{\Delta u_y}{\Delta y}$$

$$\nu = -\varepsilon_x \frac{E}{\sigma_y}$$
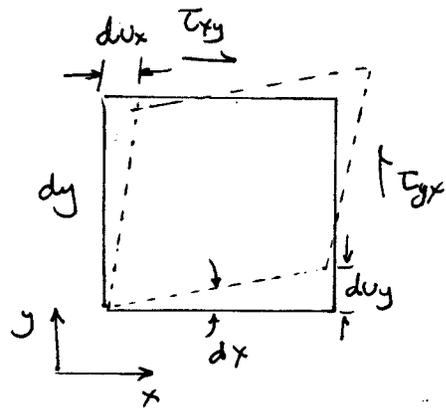
$$\gamma_{xy} = \left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \right)$$

$$\varepsilon_x = \frac{1}{E} \left[\sigma_x - \nu(\sigma_y + \sigma_z) \right]$$

$$\varepsilon_y = \frac{1}{E} \left[\sigma_y - \nu(\sigma_x + \sigma_z) \right]$$

$$\varepsilon_z = \frac{1}{E} \left[\sigma_z - \nu(\sigma_y + \sigma_x) \right]$$

$$\gamma_{xy} = \tau_{xy}/G$$

$$\gamma_{yz} = \tau_{yz}/G$$

$$\gamma_{zx} = \tau_{zx}/G$$

$$\left. \right\} \quad G = \frac{E}{2(1+\nu)}$$

Create inverse relations: $\quad \mathcal{E} = f(\sigma, p) \quad \longrightarrow \quad \sigma = f(\mathcal{E}, p)$

$$\sigma_m = \tfrac{1}{3}(\sigma_x + \sigma_y + \sigma_z)$$

Define volume strain:

$$\mathcal{E}_x = \tfrac{1}{E}\left[\sigma_x - \nu(\sigma_y + \sigma_z)\right]$$

$$\mathcal{E}_y = \tfrac{1}{E}\left[\sigma_y - \nu(\sigma_x + \sigma_z)\right]$$

$$\mathcal{E}_z = \tfrac{1}{E}\left[\sigma_z - \nu(\sigma_x + \sigma_y)\right]$$

$$\mathcal{E}_v = \tfrac{1}{E}\left[3\sigma_m - 2\nu(3\sigma_m)\right]$$

$$\mathcal{E}_v = \tfrac{3}{E}\left[\sigma_m - 2\nu\sigma_m\right]$$

$$\mathcal{E}_v = \frac{3\sigma_m}{E}\left[1 - 2\nu\right]$$

$$\sigma_m = \frac{\mathcal{E}_v \, E}{3(1 - 2\nu)}$$

Rewrite Hooke's Law:

$$\mathcal{E}_x = \tfrac{1}{E}\left[\sigma_x \qquad - \nu(\sigma_y + \sigma_z)\right]$$

$$\mathcal{E}_x = \tfrac{1}{E}\left[\sigma_x + \nu\sigma_x - \nu\sigma_x - \nu(\sigma_y + \sigma_z)\right]$$

$$\mathcal{E}_x = \tfrac{1}{E}\left[(1+\nu)\sigma_x - \nu(\sigma_x + \sigma_y + \sigma_z)\right] = \tfrac{1}{E}\left[(1+\nu)\sigma_x - \nu(3\sigma_m)\right]$$

Rearrange in terms of σ_x:

Substitute for σ_m

$$\frac{\mathcal{E}_x E}{(1+\nu)} + \frac{\nu \, 3\sigma_m}{(1+\nu)} = \sigma_x$$

$$G = \frac{E}{2(1+\nu)}$$

$$\frac{2}{2}\frac{\mathcal{E}_x E}{(1+\nu)} + \frac{2}{2}\frac{3\nu}{(1+\nu)}\frac{\mathcal{E}_v E}{3(1-2\nu)} = \sigma_x$$

$$2G\,\mathcal{E}_x + \frac{2G \, \nu \, \mathcal{E}_v}{(1-2\nu)} = \sigma_x$$

Solid Mechanics (from J C Sheng, J Liu, WC Zhu, D Elsworth, in submittal)

GOVERNING EQUATIONS

Behavior is defined in terms of mechanical equilibrium, with components included to represent the heat and fluid transport in a porous medium.

1   Mechanical Equilibrium

For an elastic medium the constitutive relation (Hooke's Law) is defined in terms of the total stress σ_{ij} (positive for tension), strain ε_{ij}, pore fluid pressure change p (negative for suction) and temperature change T as

$$\sigma_{ij} = 2G\varepsilon_{ij} + \frac{2Gv}{1-2v}\varepsilon_{kk}\delta_{ij} - \alpha p\delta_{ij} - K'\alpha_T T\delta_{ij}, \tag{1}$$

in which G is the shear modulus, v is the drained Poisson's ratio, δ_{ij} is the Kronecker delta defined as 1 for $i = j$ and 0 for $i \neq j$, K' ($= 2G(1+v)/3(1-2v)$) is the drained bulk modulus of the medium, α_T is coefficient of volumetric expansion of the bulk medium under constant pore pressure and stress (˚C^{-1}), the parameter α (≤ 1) is Biot's coefficient which depends on the compressibility of the constituents and can be defined as

$$\alpha = 1 - \frac{K'}{K_s} = \frac{3(v_u - v)}{B(1-2v)(1+v_u)}, \tag{2}$$

where K_s is the effective bulk modulus of the solid constituent, and the effective stress is defined as $\sigma'_{ij} = \sigma_{ij} + \alpha p\delta_{ij}$.

Using compact notation, the equations of equilibrium and the strain-displacement relations can be expressed as

$$\sigma_{ij,j} + F_i = 0 \tag{3}$$

and

$$\varepsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}) \tag{4}$$

respectively. Where F_i and u_i ($i = x, y, z$) are the components of the net body force and displacement in the i-direction. From eqns (1) and (4), a modified Navier equation may

be derived via eqn (3), in terms of displacement under a combination of changes of applied stresses, pore fluid pressures, and temperature as

$$Gu_{i,jj} + \frac{G}{1-2v}u_{j,ji} - \alpha p_{,i} - K'\alpha_T T_{,i} + F_i = 0 .$$ (5)

2   Flow Equation

For a porous solid filled with an interstitial and freely diffusing pore fluid, where solid and fluid are assumed in thermal equilibrium, the rate of change of volume V caused by changes of temperature, pore fluid pressure, and strains can be expressed as (Zhou et al., 1998)

$$\frac{1}{V}\frac{\partial V}{\partial t} = \frac{\partial \varepsilon_v}{\partial t} = -\nabla \cdot q_l + [\phi\alpha_l + (1-\phi)\alpha_s]\frac{\partial T}{\partial t} - \left(\frac{\phi}{\beta_l} + \frac{1-\phi}{K_s}\right)\frac{\partial p}{\partial t} + \frac{1}{3K_s}\sigma_{ij}\delta_{ij}$$ (6)

where t is time (s), ε_v is the volume stain ($= \varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}$), q_l is the water flux/unit area (m/s), ϕ is the porosity in a general continuum, α_l is the coefficient of volumetric thermal expansion of the liquid (°C^{-1}), α_s is the coefficient of volumetric thermal expansion of the solid matrix (°C^{-1}), and β_l is the bulk modulus of the pore fluid (Pa). Rearrangement of eqn (6) results in the fluid mass conservation equation

$$\nabla \cdot q_l = -\frac{\partial \varepsilon_v}{\partial t} + [\phi\alpha_l + (1-\phi)\alpha_s]\frac{\partial T}{\partial t} - \left(\frac{\phi}{\beta_l} + \frac{1-\phi}{K_s}\right)\frac{\partial p}{\partial t} + \frac{1}{3K_s}\sigma_{ij}\delta_{ij} .$$ (7)

By neglecting effects of thermal-osmosis, the constitutive relation for fluid diffusion can be expressed by Darcy's law, as,

$$q_l = -\kappa\nabla(p + \rho_l gz)$$ (8)

where z is the vertical coordinate, κ is the coefficient of permeability [m^4/(N·s)] with $\kappa = k/\mu_l$, where μ_l is the dynamic fluid viscosity (N·s·m^{-2}), k is the intrinsic permeability in a general continuum (m^2), ρ_l is the liquid density (kg/m^3), and g is gravitational acceleration (m/s^2). Substitution of eqns (8) and (1) into eqn (7) results in

$$c_1\frac{\partial \varepsilon_v}{\partial t} - c_2\frac{\partial T}{\partial t} + c_3\frac{\partial p}{\partial t} = \nabla \cdot [\kappa(\nabla p + \rho_l g\nabla z)]$$ (9)

where

$$c_1 = 1 - \frac{K'}{K_s} = \frac{3(v_u - v)}{B(1+v_u)(1-2v)} ,$$

$$c_2 = \phi\alpha_l + (1-\phi)\alpha_s - \frac{\alpha_T K'}{K_s}, \tag{10}$$

$$c_3 = \frac{\phi}{\beta_l} + \frac{1-\phi}{K_s} = \frac{9(1-2\nu_u)(\nu_u - \nu)}{2GB^2(1-2\nu)(1+\nu_u)^2}, \tag{11}$$

3   Energy Conservation Equation

By neglecting thermal-filtration effects, the constitutive relation for heat diffusion is given by Fourier's law as

$$q_T = -\lambda_M \nabla T \tag{11}$$

where q_T is the heat flux transmitted by conduction in the fluid-solid mixture, with

$$\lambda_M = (1-\phi)\lambda_s + \phi\lambda_l . \tag{12}$$

Here, λ_s and λ_l are the thermal conductivities of the solid (rock) and liquid [J/(s·m·°C)] components. Due to the assumption of thermal equilibrium between the fluid and solid phases, the heat energy balance equation over an REV can be expressed in terms of a single equation which neglects the terms representing the interconvertibility of thermal and mechanical energy (Zhou et al., 1998; Noorishad and Tsang, 1996; Kurashige, 1989)

$$(\rho C)_M \frac{\partial T}{\partial t} - (T_0 + T)a_l\beta_l\nabla q_l - (T_0 + T)K'\alpha_T \frac{\partial \varepsilon_v}{\partial t} = -\nabla \cdot q_T - \nabla \cdot (\rho_0 H q_l) \tag{13}$$

where T_0 is the absolute reference temperature in the stress-free state (K), ρ_0 is the reference mass density, H represents the specific enthalpy of the pore fluid, $(\rho C)_M$ is the specific heat capacity of the fluid-filled medium, defined as $(\rho C)_M = \phi(\rho_l C_l) + (1-\phi)(\rho_s C_s)$, where ρ_s is the mass density of the rock matrix (kg/m³), and C_l and C_s are the fluid and solid specific heat constants at constant volume (J·kg⁻¹·°C⁻¹).

The first term on the left-hand side of eqn (13) represents the rate of internal heat energy change per unit volume due to an increase in temperature. The second term represents a heat sink due to thermal dilatation of the fluid. The last term represents a heat sink due to thermal expansion of the medium. For a small variation of temperature (the temperature changes (T) are small compared to the absolute ambient temperature),

$T_0 + T \approx T_0$, this term is identical to that given by Biot (1956). The second and third terms on the left-hand side of eqn (13) represent the thermoporoelastic coupling in the heat energy balance equation (Zhou et al., 1998). The last term on the right-hand side of eqn (13) represents the convective heat flux (the transportation of enthalpy by fluid flow through pores).

We assume that heat exchange between the solid matrix and the pore fluid is rapid in comparison with the global heat and fluid diffusion processes. Thus, the local heat equilibrium is established (Kurashige, 1989) as,

$$H = (\rho C)_M T / (\phi \rho_0).$$ (14)

Substitution of eqns (11) and (14) into eqn (13) results in

$$(\rho C)_M \frac{\partial T}{\partial t} + (T_0 + T)\alpha_l \beta_l \nabla \cdot (\kappa \nabla p + \rho_l g \nabla z)$$

$$- (T_0 + T)K'\alpha_T \frac{\partial \varepsilon_v}{\partial t} - \frac{(\rho C)_M}{\phi} \kappa (\nabla p + \rho_l g \nabla z) \cdot \nabla T = \lambda_M \nabla \cdot q_T$$ (15)

The last term on the left-hand side of eqn (15) represents the convective heat flux.

Equations (5), (9) and (15) represent a set of fully coupled non-linear equations governing the thermo-poroelastic response of a saturated medium. The equations account for thermodynamically coupled heat and mass transfer, mechanical and thermal compressibility of the constituents, and importantly in this work, convective heat flow.

4 Initial and Boundary Conditions

The triply coupled THM physics of the system is defined through equations (5), (9) and (15). For completeness, standard boundary conditions and initial conditions are defined as follows.

4.1 Boundary conditions

Stress-displacement conditions for the mechanical analysis are defined as

$$\mathbf{u}(\mathbf{x},t) = \overline{\mathbf{u}}(\mathbf{x},t), \quad t \in [0,\infty),$$ (16)

$$\sigma(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x}) = \overline{\mathbf{F}}(\mathbf{x},t), \; t \in [0, \infty). \tag{17}$$

Fluid flow is defined in terms of boundary conditions representing:

The Dirichlet condition: $p(\mathbf{x},t) = \overline{p}(\mathbf{x},t), \; t \in [0, \infty). \tag{18}$

The Neumann condition: $\kappa \cdot (\nabla p - \rho_l \boldsymbol{g}) \cdot \mathbf{n}(\mathbf{x}) = \overline{Q}_l(\mathbf{x},t), \; t \in [0, \infty). \tag{19}$

And likewise for heat transport:

The Dirichlet condition: $T(\mathbf{x},t) = \overline{\overline{T}}(\mathbf{x},t), \; t \in [0, \infty). \tag{20}$

The Neumann condition: $\lambda_M \nabla T \cdot \mathbf{n}(\mathbf{x}) = \overline{Q}_h(\mathbf{x},t), \; t \in [0, \infty). \tag{21}$

where $\mathbf{n}$ is the outward unit normal vector on the domain boundary.

4.2 Initial conditions

Initial conditions for the mechanical, flow and thermal analyses are defined as

$$\mathbf{u}(\mathbf{x},0) = 0 \text{ on } V, \tag{22}$$

$$\sigma(\mathbf{x},0) = 0 \text{ on } V, \tag{23}$$

$$p(\mathbf{x},0) = 0 \text{ on } V, \tag{24}$$

$$T(\mathbf{x},0) = 0 \text{ on } V. \tag{25}$$

The dependent variables, $\mathbf{u}$, p, and T, represent incremental deviations from the strain-free state assumed by the above choice of initial conditions. The quantity V represents the volume under consideration.

REFERENCES

Biot, M.A., 1956. Thermoelasticity and irreversible thermodynamics. J of applied physics, 12:155-164.

Cleary, M.P., 1977. Fundamental solutions for a fluid-saturated porous solid. Int J Solids Structures. 13:785

Detournay, E., and Cheng A. H-D, 1988. Poroelastic response of a borehole in a non-hydrostatic stress field. Int J Rock Mech Min Sci & Geomech Abstr. 25(3): 171-182.

Jing L, Tsang C-F, Stephansson O. 1995. DECOVALEX-An international cooperative research project on mathematical models of coupled THM processes for safety

analysis of radioactive waste repositories. Int. J. Rock Mech. Min. Sci. 32(5): 387-398.

Kurashige, M., 1989. A thermoelastic theory of fluid-filled porous materials. Int J Solids Structures. 25(9):1039-1052.

Noorishad J. and Tsang C-F. 1996. Coupled thermohydroelasticity phenomena in variably saturated fractured porous rocks-Formulation and numerical solution. In Coupled Thermo-Hydro-Mechanical Processes of Fractured Media, ed. O Stephansson, L Jing, CF Tsang, p.93-134. Amsterdam: Elsevier Science Publishers. 575pp.

Rice J.R. and Cleary, M.P., 1976. Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. Rev. Geophys. Space Phys. 14:227-241.

Zhou, Y., Rajapakse, R., and Graham, J., 1998. A coupled thermoporoelastic model with thermo-osmosis and thermal-filtration. Int J Solids Structures. 35(34-35):4659-4683.

SIMILARITIES BETWEEN FLUID & SOLID MECHANICS

	Fluid Mechs.	Solid Mechs.

CONSERVATION:

Momentum: (Eulerian Ref frame)

$$\frac{\partial}{\partial t}\rho v_x = -\left(\frac{\partial}{\partial x}\rho v_x v_x + \frac{\partial}{\partial y}\rho v_y v_x + \ldots\right)$$

$$-\left(\frac{\partial}{\partial x}\tau_{xx} + \frac{\partial}{\partial y}\tau_{yx} + \ldots\right) - \frac{\partial p}{\partial x} + \rho g_x$$

(Lagrangian Ref frame)

$$\rho\frac{Dv_x}{Dt} = -\frac{\partial p}{\partial x} - \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z}\right) + \rho g_x \qquad (3)$$

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} = \rho g_x + \rho\frac{\partial^2}{\partial t} \qquad (3)$$

Mass:

$$\frac{\partial \rho}{\partial t} = -\nabla\cdot\rho v$$

(Eulerian)

$$\frac{\partial \rho}{\partial t} + \left(v_x\frac{\partial \rho}{\partial x} + \ldots\right) = -\rho\left(\frac{\partial v_x}{\partial x} + \ldots\right)$$

$$\varepsilon_x = \frac{\partial v_x}{\partial x} \qquad \gamma_{xy} = \frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x}$$

(Lagrangian)

$$\frac{D\rho}{Dt} = -\rho(\nabla\cdot v) \qquad (1)$$

$$\frac{\partial^2 \varepsilon_x}{\partial y^2} + \frac{\partial^2 \varepsilon_y}{\partial x^2} = \frac{\partial^2 \gamma_{xy}}{\partial x\partial y} \qquad (6)$$

CONSTITUTIVE:

Linear:

$$\sigma_x = -p + \lambda\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}\right) + 2\mu\frac{\partial v_x}{\partial x} \qquad (6)$$

Poisson Eqn. $\quad \lambda = -\frac{2}{3}\mu$

$$\sigma_x = 2g\varepsilon_x + \lambda(\varepsilon_x + \varepsilon_y + \varepsilon_z) \qquad (6)$$

$$\left(g = \frac{E}{2(1+v)} \quad ; \quad \lambda = \frac{2gv}{(1-2v)}\right)$$

Failure:

$$\sigma_1 = N\sigma_3 + (1-N)p + 2cN^{1/2}$$

$$\left(N = (1+\sin\phi)/(1-\sin\phi)\right)$$

Variables:

$$v_x ; v_y ; v_z ; p$$

$$\text{or} \equiv \underbrace{\dot{u}_x ; \dot{u}_y ; \dot{u}_z}_{\text{velocities}} ; p$$

$$\underbrace{u_x ; u_y ; u_z}_{\text{displacements}} ; p^{?}$$

v = Poisson ratio

g = shear modulus

λ = Lamé constant; c = cohesion

$\lambda = \frac{2}{3}\mu$

μ = dynamic viscosity

GENERAL FORM OF FINITE ELEMENT EQUATIONS

2nd order PDE in space:

$$\frac{\partial^2 u}{\partial t^2} \longrightarrow \int_V \underline{a}^T \underline{D} \, \underline{a} \, dv \; \boxed{\underline{u}}$$

$$u = \underline{b} \, u$$

$$\varepsilon = \underline{a} \, u$$

$$\sigma = \underline{D} \, \varepsilon$$

1st order PDE in space:

$$v \frac{\partial c}{\partial t} \longrightarrow \int_V \underline{b}^T \underline{v} \, \underline{a} \, dv \; \boxed{\underline{c}}$$

$$c = \underline{b} \, c$$

$$c_1 = \frac{\partial c}{\partial t} \qquad c_{,} = \underline{a} \, c$$

$$\underline{v} = [v_x \, ; \, v_y]$$

0th order PDE in space:

$$S_s \frac{\partial h}{\partial t} \longrightarrow S_s \int_V \underline{b}^T \underline{b} \, dv \; \boxed{\dot{\underline{h}}}$$

$$h = \underline{b} \, h$$

A historical outline of matrix structural analysis: a play in three acts

C.A. Felippa *

Department of Aerospace Engineering Sciences and Center for Aerospace Structures, University of Colorado, Boulder, CO 80309-0429, USA

Abstract

The evolution of matrix structural analysis (MSA) from 1930 through 1970 is outlined. Highlighted are major contributions by Collar and Duncan, Argyris, and Turner, which shaped this evolution. To enliven the narrative the outline is configured as a three-act play. Act I describes the pre-WWII formative period. Act II spans a period of confusion during which matrix methods assumed bewildering complexity in response to conflicting demands and restrictions. Act III outlines the cleanup and consolidation driven by the appearance of the direct stiffness method, through which MSA completed morphing into the present implementation of the finite element method (FEM). No attempt is made at chronicling the more complex history of FEM itself. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Matrix structural analysis; Finite elements; History; Displacement method; Force method; Direct stiffness method; Duality

1. Introduction

Who first wrote down a stiffness or flexibility matrix?

The question was posed in a 1995 paper [1]. The educated guess was "somebody working in the aircraft industry of Britain or Germany, in the late 1920s or early 1930s". Since then the writer has examined reports and publications of that time. These trace the origins of matrix structural analysis (MSA) to the aeroelasticity group of the National Physics Laboratory (NPL) at Teddington, a town that has now become a suburb of greater London.

The present paper is an expansion of the historical vignettes in Section 4 of [1]. It outlines the major steps in the evolution of MSA by highlighting the fundamental contributions of four individuals: Collar, Duncan, Argyris and Turner. These contributions are lumped into three milestones:

Creation: Beginning in 1930 Collar and Duncan formulated discrete aeroelasticity in matrix form. The first two journal papers on the topic appeared in 1934–1935 [2,3] and the first book, coauthored with Frazer, in 1938 [4]. The representation and terminology for discrete dynamical systems is essentially that used today.

Unification: In a series of journal articles appearing in 1954 and 1955 [5] Argyris presented a formal unification of force and displacement methods (FDM) using dual energy theorems. Although practical applications of the duality proved ephemeral, this work systematized the concept of assembly of structural system equations from elemental components.

FEMinization: In 1959 Turner proposed [6] the direct stiffness method (DSM) as an efficient and general computer implementation of the then embryonic, and as yet unnamed, finite element method (FEM). This technique, fully explained in a follow-up article [7], naturally encompassed structural and continuum models, as well as nonlinear, stability and dynamic simulations. By 1970 DSM had brought about the demise of the classical

---

* Tel.: +1-303-492-6547; fax: +1-303-492-4990.

E-mail address: carlos@titan.colorado.edu (C.A. Felippa).

force method (CFM), and become the dominant implementation in production-level FEM programs.

These milestones help dividing MSA history into three periods. To enliven and focus the exposition these will be organized as three acts of a play, properly supplemented with a "matrix overture" prologue, two interludes and a closing epilogue. Here is the program:

Prologue – Victorian artifacts: 1858–1930.
Act I – gestation and birth: 1930–1938.
Interlude I – WWII blackout: 1938–1947.
Act II – the matrix forest: 1947–1956.
Interlude II – questions: 1956–1959.
Act III – answers: 1959–1970.
Epilogue – revisiting the past: 1970–date.

Act I, as well as most of the prologue, takes place in the UK. The following events feature a more international cast.

2. Background and terminology

Before departing for the theater, this Section offers some general background and explains historical terminology. Readers familiar with the subject should skip to Section 3.
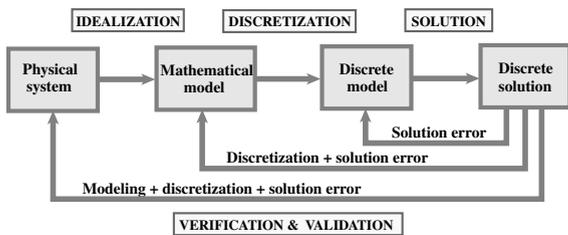

Fig. 1. Flowchart of model-based simulation (MBS) by computer.

The overall schematics of model-based simulation (MBS) by computer is flowcharted in Fig. 1. For mechanical systems such as structures the FEM is the most widely used discretization and solution technique. Historically the ancestor of FEM is MSA, as illustrated in Fig. 2. The morphing of the MSA from the pre-computer era – as described for example in the first MSA book [4] – into the first programmable computers took place, in wobbly gyrations, during the transition period herein called Act II. Following a confusing interlude, the young FEM begin to settle, during the early 1960s, into the configuration shown on the right of Fig. 2. Its basic components have not changed since 1970.

MSA and FEM stand on three legs: mathematical models, matrix formulation of the discrete equations, and computing tools to do the numerical work. Of the three legs the latter is the one that has undergone the most dramatic changes. The "human computers" of the 1930s and 1940s morphed by stages into programmable computers of analog and digital type. The matrix formulation moved like a pendulum. It begins as a simple displacement method in Act I, reaches bewildering complexity in Act II and goes back to conceptual simplicity in Act III.

Unidimensional structural models have changed little: a 1930 beam is still the same beam. The most noticeable advance is that pre-1955 MSA, following classical Lagrangian mechanics, tended to use spatially discrete energy forms from the start. The use of space-continuum forms as basis for multidimensional element derivation was pioneered by Argyris [5], successfully applied to triangular geometries by Turner et al. [8], and finalized by Melosh [9] and Irons [10,11] with the precise statement of compatibility and completeness requirements for FEM.

Matrix formulations for MSA and FEM have been traditionally classified by the choice of primary unknowns. These are those solved for by the human or digital computer to determine the system state. In the
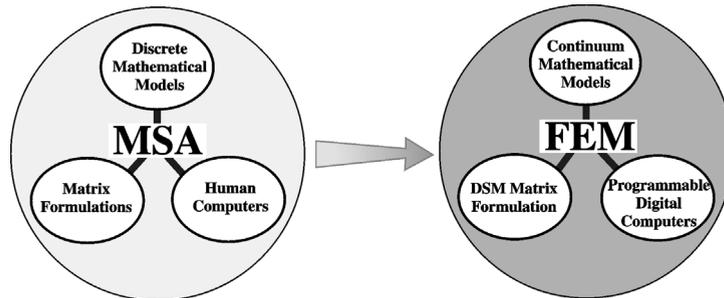

Fig. 2. Morphing of the pre-computer MSA (before 1950) into the present FEM. On the left "human computer" means computations under direct human control, possibly with the help of analog devices (slide rule) or digital devices (desk calculator). The FEM configuration shown on the right settled by the mid 1960s.

displacement method (DM) these are physical or generalized displacements. In the CFM these are amplitudes of redundant force (or stress) patterns. (The qualifier "classical" is important because there are other versions of the force method (FM) which select for example stress function values or Lagrange multipliers as unknowns.) There are additional methods that involve combinations of displacements, forces and/or deformations as primary unknowns, but these have no practical importance in the pre-1970 period covered here.

Appropriate mathematical names for the DM are range-space method or primal method. This means that the primary unknowns are the same type as the primary variables of the governing functional. Appropriate names for the CFM are null-space method, adjoint method, or dual method. This means that the primary unknowns are of the same type of the adjoint variables of the governing functional, which in structural mechanics are forces. These names are not used in the historical outline, but are useful in placing more recent developments, as well as nonstructural FEM applications, within a general framework.

The terms stiffness method and flexibility method are more diffuse names for the displacement and force methods, respectively. Generally speaking these apply when stiffness and flexibility matrices, respectively, are important part of the modeling and solution process.

3. Prolog – Victorian artifacts: 1858–1930

Matrices – or "determinants" as they were initially called – were invented in 1858 by Cayley at Cambridge, although Gibbs (the co-inventor, along with Heaviside, of vector calculus) claimed priority for the German mathematician Grassmann. Matrix algebra and matrix calculus were developed primarily in the UK and Germany. Its original use was to provide a compact language to support investigations in mathematical topics such as the theory of invariants and the solution of algebraic and differential equations. For a history of these early developments the monograph by Muir [12] is unsurpassed. Several comprehensive treatises in matrix algebra appeared in the late 1920s and early 1930s [13–15].

Compared to vector and tensor calculus, matrices had relatively few applications in science and technology before 1930. Heisenberg's 1925 matrix version of quantum mechanics was a notable exception, although technically it involved infinite matrices. The situation began to change with the advent of electronic desk calculators, because matrix notation provided a convenient way to organize complex calculation sequences. Aeroelasticity was a natural application because the stability analysis is naturally posed in terms of determinants of matrices that depend on a speed parameter.

The nonmatrix formulation of discrete structural mechanics can be traced back to the 1860s. By the early 1900s the essential developments were complete. A readable historical account is given by Timoshenko [16]. Interestingly enough, the term "matrix" never appears in this book.

4. Act I – gestation and birth: 1930–1938

In the decade of World War I aircraft technology begin moving toward monoplanes. Biplanes disappeared by 1930. This evolution meant lower drag and faster speeds but also increased disposition to flutter. In the 1920s aeroelastic research began in an international scale. Pertinent developments at the NPL are well chronicled in a 1978 historical review article by Collar [17], from which the following summary is extracted.

4.1. The source papers

The aeroelastic work at the Aerodynamics Division of NPL was initiated in 1925 by R.A. Frazer. He was joined in the following year by W.J. Duncan. Two years later, in August 1928, they published a monograph on flutter [18], which came to be known as "The Flutter Bible" because of its completeness. It laid out the principles on which flutter investigations have been based since. In January 1930 A.R. Collar joined Frazer and Duncan to provide more help with theoretical investigations. Aeroelastic equations were tedious and error prone to work out in long hand. Here are Collar's own words [17, p. 17] on the motivation for introducing matrices:

> "Frazer had studied matrices as a branch of applied mathematics under Grace at Cambridge; and he recognized that the statement of, for example, a ternary flutter problem in terms of matrices was neat and compendious. He was, however, more concerned with formal manipulation and transformation to other coordinates than with numerical results. On the other hand, Duncan and I were in search of numerical results for the vibration characteristics of airscrew blades; and we recognized that we could only advance by breaking the blade into, say, 10 segments and treating it as having 10 degrees of freedom. This approach also was more conveniently formulated in matrix terms, and readily expressed numerically. Then we found that if we put an approximate mode into one side of the equation, we calculated a better approximation on the other; and the matrix iteration procedure was born. We published our method in two papers in Phil. Mag. [2,3]; the first, dealing with conservative

systems, in 1934 and the second, treating damped systems, in 1935. By the time this had appeared, Duncan had gone to his Chair at Hull''.

The aforementioned papers appear to be the earliest journal publications of MSA. These are amazing documents: clean and to the point. They do not feel outdated. Familiar names appear: mass, flexibility, stiffness, and dynamical matrices. The matrix symbols used are $[m]$, $[f]$, $[c]$ and $[D] = [c]^{-1}[m] = [f][m]$, respectively, instead of the $\mathbf{M}$, $\mathbf{F}$, $\mathbf{K}$ and $\mathbf{D}$ in common use today. A general inertia matrix is called $[a]$. As befit the focus on dynamics, the DM is used. Point-mass displacement degrees of freedom are collected in a vector $\{x\}$ and corresponding forces in vector $\{P\}$. These are called $[q]$ and $[Q]$, respectively, when translated to generalized coordinates.

The notation was changed in the book [4] discussed below. In particular matrices are identified in Ref. [4] by capital letters without surrounding brackets, in more agreement with the modern style; for example mass, damping and stiffness are usually denoted by A, B and C, respectively.

4.2. The matrix structural analysis source book

Several papers on matrices followed, but apparently the traditional publication vehicles were not viewed as suitable for description of the new methods. At that stage Collar notes [17, p. 18] that

"Southwell (Sir Richard Southwell, the "father" of relaxation methods) suggested that the authors of the various papers should be asked to incorporate them into a book, and this was agreed. The result was the appearance in November 1938 of "Elementary Matrices" published by Cambridge University Press [4]; it was the first book to treat matrices as a branch of applied mathematics. It has been reprinted many times, and translated into several languages, and even now after nearly 40 years [this was written in 1975], stills sells in hundreds of copies a year – mostly paperback. The interesting thing is that the authors did not regard it as particularly good; it was the book we were instructed to write, rather than the one we would have liked to write''.

The writer has copies of the 1938 and 1963 printings. No changes other than minor fixes are apparent. Unlike the source papers [2,3] the book feels dated. The first 245 pages are spent on linear algebra and ODE-solution methods that are now standard part of engineering and science curricula. The numerical methods, oriented to desk calculators, are obsolete. That leaves the modeling and application examples, which are not coherently interweaved. No wonder that the authors were not happy about the book. They had followed Southwell's "merging" suggestion too literally. Despite these flaws its direct and indirect influence during the next two decades was significant. Being first excuses imperfections.

The book focuses on dynamics of a complete airplane and integrated components such as wings, rudders or ailerons. The concept of structural element is primitive: take a shaft or a cantilever and divide it into segments. The assembled mass, stiffness or flexibility is given directly. The source of damping is usually aerodynamic. There is no static stress analysis; pre-WWII aircraft were overdesigned for strength and typically failed by aerodynamic or propulsion effects.

Readers are reminded that in aeroelastic analysis stiffness matrices are generally unsymmetric, being the sum of a symmetric elastic stiffness and an unsymmetric aerodynamic stiffness. This clean decomposition does not hold for flexibility matrices because the inverse of a sum is not the sum of inverses. The treatment of [4] includes the now called load-dependent stiffness terms, which represent another first.

On reading the survey articles by Collar [17,19] one cannot help being impressed by the lack of pretension. With Duncan he had created a tool for future generations of engineers to expand and improve upon. Yet he appears almost apologetic: "I will complete the matrix story as briefly as possible" [17 p. 17]. The NPL team members shared a common interest: to troubleshoot problems by understanding the physics, and viewed numerical methods simply as helpers.

5. Interlude I – WWII blackout: 1938–1947

Interlude I is a "silent period" taken to extend from the book [4] to the first journal publication on the matrix FM for aircraft [20]. Aeroelastic research continued. New demands posed by high strength materials, higher speeds, combat maneuvers, and structural damage survival increased interest in stress analysis. For the beam-like skeletal configurations of the time, the traditional flexibility-based methods such as CFM were appropriate. Flexibilities were often measured experimentally by static load tests, and fitted into the calculations. Punched-card computers and relay calculators were increasingly used, and analog devices relied upon to solve ODEs in guidance and ballistics. Precise accounts of MSA work in aerospace are difficult to trace because of publication restrictions. The blackout was followed by a 2–3 year hiatus until those restrictions were gradually lifted, R&D groups restaffed, and journal pipelines refilled.

6. Act II – the matrix forest: 1947–1956

As Act II starts MSA work is still mainly confined to the aerospace community. But the focus has shifted from dynamics to statics, and especially stress, buckling, fracture and fatigue analysis. Turbines, supersonic flight and rocket propulsion brought forth thermomechanical effects. The Comet disasters forced attention on stress concentration and crack propagation effects due to cyclic cabin pressurization. Failsafe design gained importance. In response to these multiple demands aircraft companies staffed specialized groups: stress, aerodynamics, aeroelasticity, propulsion, avionics, and so on. A multilevel management structure with well defined territories emerged.

The transition illustrated in Fig. 2 starts, driven by two of the legs supporting MSA: new computing resources and new mathematical models. The matrix formulation merely reacts.

6.1. Computers become machines

The first electronic commercial computer: Univac I, manufactured by a division of Remington–Rand, appeared during summer 1951. The six initial machines were delivered to US government agencies [21]. That model was joined in 1952 by the Univac 1103, a scientific-computation oriented machine built by ERA, a R–R acquisition. This was the first computer with a drum memory. T.J. Watson Sr., founder of IBM, had been once quoted as saying that six electronic computers would satisfy the needs of the planet. Turning around from that prediction, IBM launched the competing 701 model in 1953.

Big aircraft companies began purchasing or leasing these expensive wonders by 1954. But this did not mean immediate access for everybody. The behemoths had to be programmed in machine or assembly code by specialists, who soon formed computer centers allocating and prioritizing cycles. By 1956 structural engineers were still likely to be using their slides rules, Marchants and punched card equipment. Only after the 1957 appearance of the first high level language (Fortran I, offered on the IBM 704) were engineers and scientists able (and allowed) to write their own programs.

6.2. The matrix CFM takes center stage

In static analysis the nonmatrix version of the CFM had enjoyed a distinguished reputation since the source contributions by Maxwell, Mohr and Castigliano. The method provides directly the internal forces, which are of paramount interest in stress-driven design. It offers considerable scope of ingenuity to experienced structural engineers through clever selection of redundant force

systems. It was routinely taught to aerospace, civil and mechanical engineering students.

Success in hand-computation dynamics depends on "a few good modes". Likewise, the success of CFM depends crucially on the selection of good redundant force patterns. The structures of pre-1950 aircraft were a fairly regular lattice of ribs, spars and panels, forming beam-like configurations. If the panels are ignored, the selection of appropriate redundants was well understood. Panels were modeled conservatively as inplane shear-force carriers, circumventing the difficulties of two-dimensional elasticity. With some adjustments and experimental validations, sweptback wings of high aspect ratio were eventually fitted into these models.

A matrix framework was found convenient to organize the calculations. The first journal article on the matrix CFM, which focused on sweptback wing analysis, is by Levy [20], followed by publications of Rand [22], Langefors [23], Wehle and Lansing [24] and Denke [25]. The development culminates in the article series of Argyris [5] discussed in Section 6.5.

6.3. The delta wing challenge

The DM continued to be used for vibration and aeroelastic analysis, although as noted above this was often done by groups separated from stress and buckling analysis. A new modeling challenge entered in the early 1950s: delta wing structures. This rekindled interest in stiffness methods.

The traditional approach to obtain flexibility and stiffness matrices of unidimensional structural members such as bars and shafts is illustrated in Fig. 3. The governing differential equations are integrated, analytically or numerically, from one end to the other. The end quantities, grouping forces and displacements, are thereby connected by a transition matrix. Using simple algebraic manipulations three more matrices shown in Fig. 3 can be obtained: deformational flexibility, deformational stiffness and free–free stiffness. This well known technique has the virtue of reducing the number of unknowns since the integration process can absorb structural details that are handled in the present FEM with multiple elements.

Notably absent from the scheme of Fig. 3 is the free–free flexibility. This was not believed to exist since it is the inverse of the free–free stiffness, which is singular. A general closed-form expression for this matrix as a Moore–Penrose generalized stiffness inverse was not found until recently [26,27].

Modeling delta wing configurations required two-dimensional panel elements of arbitrary geometry, of which the triangular shape, illustrated in Fig. 4, is the simplest and most versatile. Efforts to follow the ODE-integration approach lead to failure. (One particularly bizarre proposal, for solving exactly the wrong problem,
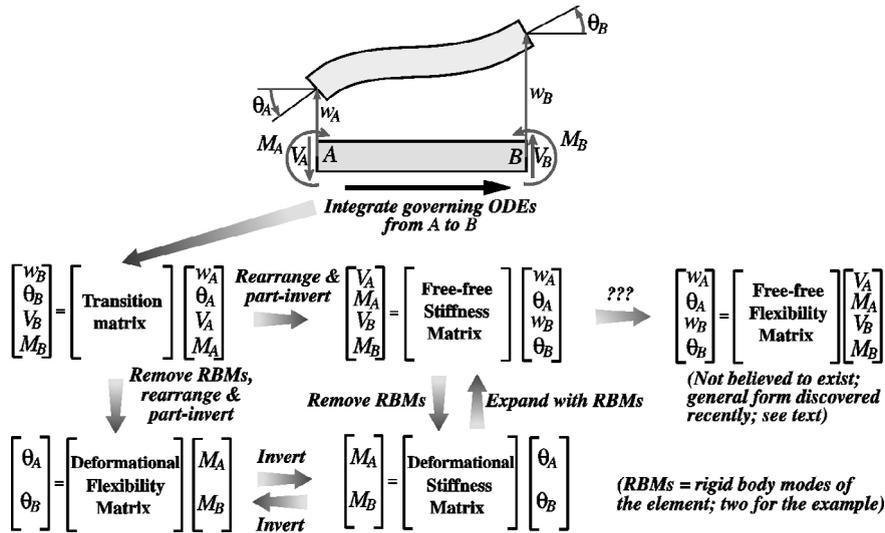
Fig. 3. Transition, flexibility and stiffness matrices for unidimensional linear structural elements, such as the plane beam depicted here, can be obtained by integrating the governing differential equations, analytically or numerically, over the member to relate end forces and displacements. Clever things were done with this "method of lines" approach, such as including intermediate supports or elastic foundations.
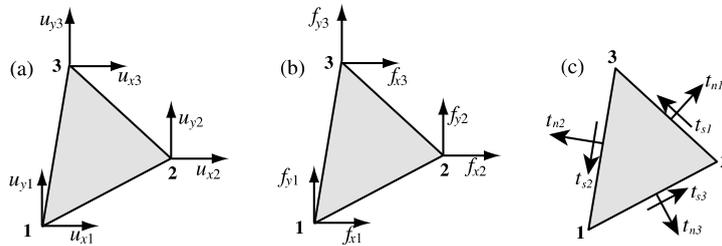

Fig. 4. Modeling delta wing configurations required panel elements of arbitrary geometry such as the triangles depicted here. The traditional ODE-based approach of Fig. 3 was tried by some researchers who (seriously) proposed finding the corner displacements in (a) produced by the concentrated corner forces in (b) on a supported triangle from the elasticity equations solved by numerical integration! Bad news: those displacements are infinite. Interior fields assumptions were inevitable, but problems persisted. A linear inplane displacement field is naturally specified by corner displacements, whereas a constant membrane force field is naturally defined by edge tractions (c). Those quantities "live" on different places. The puzzle was first solved in Ref. [8] by lumping edge tractions to node forces on the way to the free–free stiffness matrix.

is mentioned for fun in the label of Fig. 4.) This motivated efforts to construct the stiffness matrix of the panel directly. The first attempt in this direction is by Levy [28]; this was only partly successful but was able to illuminate the advantages of the stiffness approach.

The article series by Argyris [5] contains the derivation of the 8×8 free–free stiffness of a flat rectangular panel using bilinear displacement interpolation in Cartesian coordinates. But that geometry was obviously inadequate to model delta wings. The landmark contribution of Turner, Clough, Martin and Topp [8] finally succeeded in directly deriving the stiffness of a triangular panel. Clough [29] observes that this paper represents

the delayed publication of 1952–1953 work at Boeing. It is recognized as one of the two sources of present FEM implementations, the second being the DSM discussed later. Because of the larger number of unknowns compared to CFM, competitive use of the DM in stress analysis had necessarily to wait until computers become sufficiently powerful to handle hundreds of simultaneous equations.

6.4. Reduction fosters complexity

For efficient digital computation on present computers, data organization (in terms of fast access as well

as exploitation of sparseness, vectorization and parallelism) is of primary concern whereas raw problem size, up to certain computer-dependent bounds, is secondary. But for hand calculations minimal problem size is a key aspect. Most humans cannot comfortably solve by hand linear systems of more than 5 or 6 unknowns by direct elimination methods, and 5–10 times that through problem-oriented "relaxation" methods. The first-generation digital computers improved speed and reliability, but were memory strapped. For example the Univac I had 1000 45-bit words and the IBM 701, 2048 36-bit words. Clearly solving a full system of 100 equations was still a major challenge.

It should come as no surprise that problem reduction techniques were paramount throughout this period, and exerted noticeable influence until the early 1970s. In static analysis reduction was achieved by elaborated functional groupings of static and kinematic variables. Most schemes of the time can be understood in terms of the following classification:

$$
\begin{array}{l}
\text{Generalised forces}
\begin{cases}
\text{primary}
\begin{cases}
\text{applied forces } \mathbf{f}_a \\
\text{redundant forces } \mathbf{y}
\end{cases} \\
\text{secondary}
\begin{cases}
\text{condensable forces } \mathbf{f}_c = 0 \\
\text{support reactions } \mathbf{f}_s
\end{cases}
\end{cases} \\
\text{Generalised displacemetns}
\begin{cases}
\text{primary}
\begin{cases}
\text{applied displacements } \mathbf{u}_a \\
\text{redundant displacements } \mathbf{z}
\end{cases} \\
\text{secondary}
\begin{cases}
\text{condensable displacements } \mathbf{u}_c \\
\text{support conditions } \mathbf{u}_s = 0
\end{cases}
\end{cases}
\end{array}
\tag{1}
$$

Here applied forces are those acting with nonzero values, that is, the ones visibly drawn as arrows by an engineer or instructor. In reduction-oriented thinking zero forces on unloaded degrees of freedom are classified as condensable because they can be removed through static condensation techniques. Similarly, nonzero applied displacements were clearly differentiated from zero displacements arising from support conditions because the latter can be thrown out while the former must be retained. Redundant displacements, which are the counterpart of redundant forces, have been given many names, among them "kinematically indeterminate displacements" and "kinematic deficiencies".

Matrix formulation evolved so that the unknowns were the force redundants $\mathbf{y}$ in the CFM and the displacement redundants $\mathbf{z}$ in the DM. Partitioning matrices in accordance to (1) fostered exuberant growth culminating in the matrix forest that characterizes works of this period.

To a present day FEM programmer familiar with the DSM, the complexity of the matrix forest would strike as madness. The DSM master equations can be assembled without functional labels. Boundary conditions are applied on the fly by the solver. But the computing

limitations of the time must be kept in mind to see the method in the madness.

6.5. Two paths through the forest

A series of articles published by J.H. Argyris in four issues of Aircraft Engrg. during 1954 and 1955 collectively represents the second major milestone in MSA. In 1960 the articles were collected in a book, entitled "Energy Theorems and Structural Analysis" [5]. Part I, sub-entitled General Theory, reprints the four articles, whereas Part II, which covers additional material on thermal analysis and torsion, is co-authored by Argyris and Kelsey. Both authors are listed as affiliated with the Aerospace Department of the Imperial College at London.

The dual objectives of the work, stated in the preface, are "to generalize, extend and unify the fundamental energy principles of elastic structures" and "to describe in detail practical methods of analysis of complex structures – in particular for aeronautical applications".

The first objective succeeds well, and represents a key contribution toward the development of continuum-based models. Part I carefully merges classical contributions in energy and work methods with matrix methods of discrete structural systems. The coverage is methodical, with numerous illustrative examples. The exposition of the FM for wing structures reaches a level of detail unequaled for the time.

The DM is then introduced by duality – called "analogy" in this work:

"The analogy between the developments for the flexibilities and stiffnessess ... shows clearly that parallel to the analysis of structures with forces as unknowns there must be a corresponding theory with deformations as unknowns".

This section credits Ostenfeld [30] with being the first to draw attention to the parallel development. The duality is exhibited in a striking form in Table II, in which both methods are presented side by side with simply an exchange of symbols and appropriate rewording. The steps are based on the following decomposition of internal deformation states $\mathbf{g}$ and force patterns $\mathbf{p}$:

$$\mathbf{p} = \mathbf{B}_0\mathbf{f}_a + \mathbf{B}_1\mathbf{y}, \qquad \mathbf{g} = \mathbf{A}_0\mathbf{u}_a + \mathbf{A}_1\mathbf{z}, \qquad (2)$$

where the notation of [1] is used. Here $\mathbf{B}_i$ and $\mathbf{A}_i$ denote system equilibrium and compatibility matrices, respectively. The vector symbols on the right reflect a particular choice of the force–displacement decomposition (1), with kinematic deficiencies taken to be the condensable displacements: $\mathbf{z} \equiv \mathbf{u}_c$.

This unification exerted significant influence over the next decade, particularly on the European community. An excellent textbook exposition is that of Pestel and Leckie [31]. This book covers both paths, following Argyris' framework, in Chapters 9 and 10, using 83 pages and about 200 equations. These chapters are highly recommended to understand the organization of numeric and symbolic hand computations in vogue at that time, but it is out of print. Still in print (by Dover) is the book by Przemieniecki [32], which describes the DM and CFM paths in two Chapters: 6 and 8. The DM coverage is strongly influenced, however, by the DSM; thus duality is only superficially used.

6.6. Dubious duality

One key application of the duality in Ref. [5] was to introduce the DM by analogy to the then better known CFM. Although done with good intentions this approach did not anticipate the forthcoming development of continuum-based finite elements through stiffness methods. These are naturally derived directly from the total potential energy principle via shape functions, a technique not fully developed until the mid-1960s.

The side by side presentation of Table II of Ref. [5] tried to show that CFM and DM were going through exactly the same sequence of steps. Some engineers, eventually able to write Fortran programs, concluded that the methods had similar capabilities and selecting one or the other was a matter of taste. (Most structures groups, upholding tradition, opted for the CFM.) But the few engineers who tried implementing both noticed a big difference. And that was before the DSM, which has no dual counterpart under the decomposition (2), appeared.

The paradox is explained in Section 4 of Ref. [1]. It is also noted there that Eqs. (2) is not a particularly useful state decomposition. A better choice is studied in Section 2 of that paper; that one permits all known methods of classical MSA, including the DSM, to be derived for skeletal structures as well as for a subset of continuum models.

7. Interlude II – questions: 1956–1959

Interlude I was a silent period dominated by the war blackout. Interlude II is more vocal: a time of questions.

An array of methods, models, tools and applications is now on the table, and growing. Solid-state computers, Fortran, ICBMs, the first satellites. So many options. Stiffness or flexibility? Forces or displacements? Do transition matrix methods have a future? Is the CFM–DM duality a precursor to general-purpose programs that will simulate everything? Will engineers be allowed to write those programs?

As convenient milestone this outline takes 1959, the year of the first DSM paper, as the beginning of Act III. Arguments and counter-arguments raised by the foregoing questions will linger, however, for two more decades into diminishing circles of the aerospace community.

8. Act III – answers: 1959–1970

The curtain of Act III lifts in Aachen, Germany. On 6 November 1959, M.J. Turner, head of the Structural Dynamics Unit at Boeing and an expert in aeroelasticity, presented the first paper on the DSM to an AGARD Structures and Materials Panel meeting [6]. (AGARD is NATO's Advisory Group for Aeronautical Research and Development, which had sponsored workshops and lectureships since 1952. Bound proceedings or reports are called AGARDographs.)

8.1. A path outside the forest

No written record of Ref. [6] seem to exist. Nonetheless it must have produced a strong impression since published contributions to the next (1962) panel meeting kept referring to it. By 1960 the method had been applied to nonlinear problems [33] using incremental techniques. In July 1962 Turner et al. presented an expanded version of the 1959 paper, which appeared in an AGARDograph volume published by Pergamon in 1964 [7]. Characteristic of Turner's style, the introduction goes directly to the point:

> "In a paper presented at the 1959 meeting of the AGARD Structures and Materials Panel in Aachen, the essential features of a system for numerical analysis of structures, termed the DSM, were described. The characteristic feature of this particular version of the DM is the assembly procedure, whereby the stiffness matrix for a composite structure is generated by direct addition of matrices associated with the elements of the structure".

The DSM is explained in six text lines and three equations:

> "For an individual element e the generalized nodal force increments $\{\Delta X^e\}$ required to maintain a set

of nodal displacement increments $\{\Delta u\}$ are given by a matrix equation

$$\{\Delta X^e\} = K^e\{\Delta u\} \qquad (3)$$

in which K^e denotes the stiffness matrix of the individual element. Resultant nodal force increments acting on the complete structure are

$$\{\Delta X\} = \sum\{\Delta X^e\} = K\{\Delta u\} \qquad (4)$$

wherein K, the stiffness of the complete structure, is given by the summation

$$K = \sum K^e \qquad (5)$$

which provides the basis for the matrix assembly procedure noted earlier''.

Knowledgeable readers will note a notational glitch. For Eq. (5) to be correct matrix equations, K^e must be an element stiffness fully expanded to global (in that paper: "basic reference") coordinates, a step that is computationally unnecessary. A more suggestive notation used in present DSM expositions is $K = \sum (L^e)^T K^e L^e$, in which L^e are Boolean localization matrices. Note also the use of Δ in front of u and X and their identification as "increments". This simplifies the extension to nonlinear analysis, as outlined in the next paragraph:

"For the solution of linear problems involving small deflections of a structure at constant uniform temperature which is initially stress-free in the absence of external loads, the matrices K^e are defined in terms of initial geometry and elastic properties of the materials comprising the elements; they remain unchanged throughout the analysis. Problems involving nonuniform heating of redundant structures and/or large deflections are solved in a sequence of linearized steps. Stiffness matrices are revised at the beginning of each step to account for charges in internal loads, temperatures and geometric configurations''.

Next are given some computer implementation details, including the first ever mention of user-defined elements:

"Stiffness matrices are generally derived in local reference systems associated with the elements (as prescribed by a set of subroutines) and then transformed to the basic reference system. It is essential that the basic program be able to accommodate arbitrary additions to the collection of subroutines as new elements are encountered. Associated with these are a set of subroutines for generation of stress matrices S^e relating matrices of stress components σ^e in the local reference system of nodal displacements:

$$\{\sigma^e\} = S^e\{\bar{u}\} \qquad (6)$$

The vector $\{\bar{u}\}$ denotes the resultant displacements relative to a local reference system which is attached to the element. ... Provision should also be made for the introduction of numerical stiffness matrices directly into the program. This permits the utilization and evaluation of new element representations which have not yet been programmed. It also provides a convenient mechanism for introducing local structural modifications into the analysis''.

The assembly rule in Eqs. (3)–(5) is insensitive to element type. It work the same way for a 2-node bar, or a 64-node hexahedron. To do dynamics and vibration one adds mass and damping terms. To do buckling one adds a geometric stiffness and solves the stability eigenproblem, a technique first explained in [33]. To do nonlinear analysis one modifies the stiffness in each incremental step. To apply multipoint constraints the paper [7] advocates a master-slave reduction method.

Some computational aspects are missing from this paper, notably the treatment of simple displacement boundary conditions, and the use of sparse matrix assembly and solution techniques. The latter were first addressed in Wilson's thesis work [34,35].

8.2. The fire spreads

DSM is a paragon of elegance and simplicity. The writer is able to teach the essentials of the method in three lectures to graduate and undergraduate students alike. Through this path the old MSA and the young FEM achieved smooth confluence. The matrix formulation returned to the crispness of the source papers [2,3]. A widely referenced correlation study by Gallagher [36] helped dissemination. Computers of the early 1960s were finally able to solve hundreds of equations. In an ideal world, structural engineers should have quickly razed the forest and embraced DSM.

It did not happen that way. The world of aerospace structures split. DSM advanced first by word of mouth. Among the aerospace companies, only Boeing and Bell (influenced by Turner and Gallagher, respectively) had made major investments in DSM by 1965. Among academic institutions the Civil Engineering Department at Berkeley become a DSM evangelist through Clough, who made his students – including the writer – use DSM in their thesis work. These codes were freely disseminated into the non-aerospace world since 1963. Martin introduced the DSM at Washington, and Zienkiewicz,

influenced by Clough, at Swansea. The first textbook on FEM [37], which appeared in 1967, makes no mention of force methods. By then the application to nonstructural field problems (thermal, fluids, electromagnetics,...) had begun, and again the DSM scaled well into the brave new world.

8.3. The final test

Legacy CFM codes continued, however, to be used at many aerospace companies. The split reminds one of Einstein's answer when he was asked about the reaction of the old-guard school to the new physics: "we did not convince them; we outlived them". Structural engineers hired in the 1940s and 1950s were often in managerial positions in the 1960s. They were set in their ways. How can duality fail? All that is needed are algorithms for having the computer select good redundants automatically. Substantial effort was spent in those "structural cutters" during the 1960s [32,38].

That tenacity was eventually put to a severe test. The 1965 NASA request-for-proposal to build the NASTRAN finite element system called for the simultaneous development of Displacement and Force versions [39]. Each version was supposed to have identical modeling and solution capabilities, including dynamics and buckling. Two separate contracts, to MacNeal–Schwindler and Martin–Marietta, were awarded accordingly. Eventually the development of the Force version was cancelled in 1969. The following year may be taken as closing the transition depicted in Fig. 2, and as marking the end of the FM as a serious contender for general-purpose FEM programs.

9. Epilogue – revisiting the past: 1970–date

Has MSA, now under the wider umbrella of FEM, attained a final form? This seems the case for general-purpose FEM programs, which by now are truly "1960 heritage" codes.

Resurrection of the CFM for special uses, such as optimization, was the subject of a speculative technical note by the writer [40]. This was motivated by concerted efforts of numerical analysts to develop sparse null-space methods [41–45]. That research appears to have been abandoned by 1990. Section 2 of [26] elaborates on why, barring unexpected breakthroughs, a resurrection of CFM is unlikely.

A more modest revival involves the use of non-CFM flexibility methods for multilevel analysis. The structure is partitioned into substructures and then into subdomains, each of which is processed by DSM; but the subdomains are connected by Lagrange multipliers that physically represent node forces. A key driving application is massively parallel processing in which subdomains are mapped on distributed-memory processors and the force-based interface subproblem solved iteratively by finite element tearing and interconnecting (FETI) methods [46]. Another set of applications include inverse problems such as system identification and damage detection. Pertinent references and a historical sketch may be found in a recent article [47].

The true duality for structural mechanics is now known to involve displacements and stress functions, rather than displacements and forces. This was discovered by Fraeijs de Veubeke in the 1970s [48]. Although extendible beyond structures, the potential of this idea remains largely unexplored.

10. Concluding remarks

The patient reader who has reached this final section may have noticed that this is a critical overview of MSA history, rather than a recital of events. It reflects personal interpretations and opinions. There is no attempt at completeness. Only what are regarded as major milestones are covered in some detail. Furthermore there is only spotty coverage of the history of FEM itself as well as its computer implementation; this is the topic of an article under preparation for Applied Mechanics Reviews.

In particular, contributions from the 1938–1947 "Interlude" period will be examined in more detail in that review, including some largely forgotten publications pointed out by readers of a draft of this article. To date the best summary of the early history of FEM from circa 1800 B.C. (Egyptian contributions to geometry) through 1970, is given in Chapter 1 of the textbook by Martin and Carey [49].

This article can be hopefully instructive in two respects. First, matrix methods now in disfavor may come back in response to new circumstances. An example is the resurgence of flexibility methods in massively parallel processing. A general awareness of the older literature helps. Second, the sweeping victory of DSM over the befuddling complexity of the "matrix forest" period illustrates the virtue of Occam's proscription against multiplying entities: when in doubt chose simplicity. This dictum is relevant to the present confused state of computational mechanics.

Acknowledgements

Jennings and Wilson, as well as the reviewers, are gratefully acknowledged.

References

[1] Felippa CA. Parametrized unification of matrix structural analysis: classical formulation and d-connected elements. Finite Elem Anal Des 1995;21:45–74.

[2] Duncan WJ, Collar AR. A method for the solution of oscillations problems by matrices. Phil Magn 1934; 17(7):865.

[3] Duncan WJ, Collar AR. Matrices applied to the motions of damped systems. Phil Magn 1935;19(7):197.

[4] Frazer RA, Duncan WJ, Collar AR. Elementary matrices and some applications to dynamics and differential equations. 1st ed. Cambridge: Cambridge University Press; 1938 (7th paperback printing 1963).

[5] Argyris JJ, Kelsey S. Energy theorems and structural analysis. London: Butterworths; 1960 (Part I reprinted from Aircraft Engrg. 26 Oct–Nov, 1954 and 27 April–May, 1955).

[6] Turner MJ. The direct stiffness method of structural analysis. Structural and Materials Panel Paper, AGARD Meeting, Aachen, Germany, 1959.

[7] Turner MJ, Martin HC, Weikel RC. Further development and applications of the stiffness method AGARD structures and materials panel, Paris, France, July 1962. In: Fraeijs de Veubeke BM, editor. AGARDograph 72: Matrix Methods of Structural Analysis, Pergamon Press, Oxford, 1964. p. 203–66.

[8] Turner MJ, Clough RW, Martin HC, Topp LJ. Stiffness and deflection analysis of complex structures. J Aero Sci 1956;23:805–24.

[9] Melosh RJ. Bases for the derivation of matrices for the direct stiffness method. AIAA J 1963;1:1631–7.

[10] Irons BM. Comments on matrices for the direct stiffness method by R.J. Melosh. AIAA J 1964;2:403.

[11] Irons BM. Engineering application of numerical integration in stiffness methods. AIAA J 1966;4:2035–7.

[12] Muir T. The history of determinants in the historical order of development, vols I–IV, MacMillan, London. p. 1906–23.

[13] Turnbull HW. The theory of determinants, matrices and invariants. London: Blackie & Sons; 1929 (reprinted by Dover Publications, 1960).

[14] MacDuffee CC. The theory of matrices. Berlin: Springer; 1933 (Chelsea Pub. Co., New York, 1946).

[15] Muir T, Metzler WJ. A treatise on the theory of determinants. London: Longmans and Green; 1933.

[16] Timoshenko SP. History of strength of materials. New York: McGraw-Hill; 1953 (Dover edition, 1983).

[17] Collar AR. The first fifty years of aeroelasticity. Aerospace 1978:5(2);12–20.

[18] Frazer RA., Duncan WJ. The Flutter of Airplane Wings, Reports & Memoranda 1155, Aeronautical Research Committee, London, 1928.

[19] Collar AR. Aeroelasticity, retrospect and prospect. J Roy Aero Soc 1959;63(577):1–17.

[20] Levy S. Computation of influence coefficients for aircraft structures with discontinuities and sweepback. J Aero Sci 1947;14:547–60.

[21] Ceruzzi PE. A history of modern computing. Cambridge: MIT Press; 1998.

[22] Rand T. An approximate method for computation of stresses in sweptback wings. J Aero Sci 1951;18:61–3.

[23] Langefors B. Analysis of elastic structures by matrix coefficients, with special regard to semimonocoque structures. J Aero Sci 1952;19:451–8.

[24] Wehle LB, Lansing W. A method for reducing the analysis of complex redundant structures to a routine procedure. J Aero Sci 1952;19:677–84.

[25] Denke PH. A matrix method of structural analysis. Proceedings of the Second US National Congress on Applied Mechanics, ASCE, 1954. p. 445–57.

[26] Felippa CA, Park KC. A direct flexibility method. Comput Meth Appl Mech Engng 1997;149:319–37.

[27] Felippa CA, Park KC, Justino MR. The construction of free–free flexibility matrices as generalized stiffness inverses. Comput Struct 1998;68:411–8.

[28] Levy S. Structural analysis and influence coefficients for delta wings. J Aero Sci 1953;20:677–84.

[29] Clough RW. The finite element method – a personal view of its original formulation. In: K. Bell editor. From Finite Elements to the Troll Platform – the Ivar Holand 70th Anniversary volume. Tapir, Trondheim, Norway, 1994. p. 89–100.

[30] Ostenfeld A. Die deformation methode. Berlin: Springer; 1926.

[31] Pestel EC, Leckie FA. Matrix methods in elastomechanics. New York: McGraw-Hill; 1963.

[32] Przemieniecki JS. Theory of matrix structural analysis. New York: McGraw-Hill; 1968 (Dover edition 1986).

[33] Turner MJ, Dill EH, Martin HC, Melosh RJ. Large deflection analysis of complex structures subjected to heating and external loads. J Aero Sci 1960;27:97–107.

[34] Wilson EL. Finite element analysis of two-dimensional structures, PhD Dissertation, Department of Civil Engineering, University of California at Berkeley, 1963.

[35] Wilson EL. Automation of the finite element method – a historical view. Finite Elem Anal Des 1993;13:91–104.

[36] Gallaguer RH. A correlation study of methods of matrix structural analysis. Oxford: Pergamon; 1964.

[37] Zienkiewicz OC, Cheung YK. The finite element method in structural and solid mechanics. London: McGraw-Hill; 1967.

[38] Robinson J. Structural matrix analysis for the engineer. New York: Wiley; 1966.

[39] MacNeal RH. The MacNeal Schwendler Corporation: The First Twenty Years, Gardner Litograph, Buena Park, CA, 1988.

[40] Felippa CA. Will the force method come back?. J Appl Mech 1987;54:728–9.

[41] Berry MW, Heath MT, Kaneko I, Lawo M, Plemmons RJ, Ward RC. An algorithm to compute a sparse basis of the null space. Numer Math 1985;47:483–504.

[42] Kaneko I, Plemmons RJ. Minimum norm solutions to linear elastic analysis problems. Int J Numer Meth Engng 1984;20:983–98.

[43] Gilbert JR, Heath MT. Computing a sparse basis for the null space. SIAM J Alg Disc Meth 1987;8:446–59.

[44] Coleman TF, Pothen A. The null space problem: II. Algorithms. SIAM J Alg Disc Meth 1987;8:544–63.

[45] Plemmons RJ, White RE. Substructuring methods for computing the nullspace of equilibrium matrices. SIAM J Matrix Anal Appl 1990;1:1–22.

[46] Farhat C, Roux FX. Implicit parallel processing in structural mechanics. Computat Mechs Adv 1994;2(1):1–124.

[47] Park KC, Felippa CA. A variational principle for the formulation of partitioned structural systems. Int J Numer Meth Engng 2000;47:395–418.

[48] Fraeijs de Veubeke BM. Stress function approach. In: Geradin M, editor. Proceedings of the World Congress on Finite Element Methods, October 1975, Woodlands, England. Reprinted In: Fraeijs de Veubeke BM, Memorial volume of selected papers, Sitthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1980. p. 663–715.

[49] Martin HC, Carey GC. Introduction to finite element analysis: theory and application. New York: McGraw-Hill; 1973.

APPROXIMATION TECHNIQUES

2.1 Methods of Weighted Residual

Methods of weighted residual are useful to obtain approximate solutions to a differential governing equation. In order to explain the methods, we consider the following sample problem:

$$\begin{cases} \frac{d^2u}{dx^2} - u = -x, & 0 < x < 1 \\ u(0) = 0, \text{ and } u(1) = 0 \end{cases} \tag{2.1.1}$$

The first step in the methods of weighted residual is to assume a trial function which contains unknown coefficients to be determined later. For example, a trial function, $\tilde{u} = ax(1 - x)$, is selected as an approximate solution to Eq. (2.1.1). Here, ~ denotes an approximate solution which is usually different from the exact solution. The trial function is chosen here such that it satisfies the boundary conditions (i.e., $\tilde{u}(0) = 0$ and $\tilde{u}(1) = 0$), and it has one unknown coefficient a to be determined.

In general, accuracy of an approximated solution is dependent upon proper selection of the trial function. However, a simple form of trial function is selected for the present example to show the basic procedure of the methods of weighted residual. Once a trial function is selected, residual is computed by substituting the trial function into the differential equation. That is, the residual R becomes

$$R = \frac{d^2\tilde{u}}{dx^2} - \tilde{u} + x = -2a - ax(1 - x) + x \tag{2.1.2}$$

Because $\tilde{u}$ is different from the exact solution, the residual does not vanish for all values of x within the domain. The next step is to determine the unknown constant a such that the chosen test function best approximates the exact solution. To this end, a test (or weighting) function w is selected and the weighted average of the residual over the problem domain is set to zero. That is,

$$I = \int_0^1 wR \, dx = \int_0^1 w \left(\frac{d^2\tilde{u}}{dx^2} - \tilde{u} + x \right) dx$$

$$= \int_0^1 w\{-2a - ax(1 - x) + x\} dx = 0 \tag{2.1.3}$$

Table 2.1.1 Comparison of Solution to Eq. (2.1.1) at x=0.5

Exact Solution	Collocation	Least Squares	Galerkin
0.0566	0.0556	0.0576	0.0568

The next step is to decide the test function. The resulant approximate solution differs depending on the test function. The methods of weighted residual can be classified based on how the test function is determined. Some of the methods of weighted residual are explained below. Readers may refer to Refs [1-3] for other methods.

1. Collocation Method. The Dirac delta function, $\delta(x - x_i)$, is used as the test function, where the sampling point x_i must be within the domain, $0 < x_i < 1$. In other words,

$$w = \delta(x - x_i) \qquad (2.1.4)$$

Let $x_i = 0.5$ and we substitute the test function into the weighted residual, Eq. (2.1.3), to find $a = 0.2222$. Then, the approximate solution becomes $\tilde{u} = 0.2222x(1 - x)$.

2. Least Squares Method. The test function is determined from the residual such that

$$w = \frac{dR}{da} \qquad (2.1.5)$$

Applying Eq.(2.1.5) to Eq. (2.1.2) yields $w = -2 - x(1-x)$. Substitution of the test function into Eq. (2.1.3) results in $a = 0.2305$. Then $\tilde{u} = 0.2305x(1 - x)$.

3. Galerkin's Method. For Galerkin's method, the test function comes from the chosen trial function. That is,

$$w = \frac{d\tilde{u}}{da} \qquad (2.1.6)$$

For the present trial function, $w = x(1 - x)$. Applying this test function to Eq. (2.1.3) gives $a = 0.2272$ so that $\tilde{u} = 0.2272x(1 - x)$. Comparison of these three approximate solutions to the exact solution at $x = 0.5$ is provided in Table 2.1.1. As seen in the comparison, all three methods result in reasonably accurate approximate solutions to Eq. (2.1.1).

In order to improve the approximate solutions, we can add more terms to the previously selected trial function. For example, another trial function is $\tilde{u} = a_1 x(1 - x) + a_2 x^2(1 - x)$. This trial function has two unknown constants to be determined. Computation of the residual using the present trial function yields

$$R = a_1(-2 - x + x^2) + a_2(2 - 6x - x^2 + x^3) + x \qquad (2.1.7)$$

We need the same number of test functions as that of unknown constants so that the constants can be determined properly. Table 2.1.2 summarizes how to determine test

Table 2.1.2   Test Functions for Methods of Weighted Residual

Method	Description
Collocation	$w_i = \delta(x - x_i), \quad i = 1, 2, ..., n$ where x_i is a point within the domain
Least Squares	$w_i = \partial R/\partial a_i, \quad i = 1, 2, ..., n,$ where R is the residual and a_i is an unknown coefficient in the trial function
Galerkin	$w_i = \partial \tilde{u}/\partial a_i, \quad i = 1, 2, ..., n$ where $\tilde{u}$ is the selected trial function

functions for a chosen trial function which has n unknowns coefficients. Application of Table 2.1.2 to the present trial function results in the following test functions for each method.

$$\text{Collocation Method}: \ w_1 = \delta(x - x_1), \quad w_2 = \delta(x - x_2) \tag{2.1.8}$$

$$\text{Least Squares Method}: \ w_1 = -2 - x + x^2, \quad w_2 = 2 - 6x - x^2 + x^3 \tag{2.1.9}$$

$$\text{Galerkin's Method}: \ w_1 = x(1 - x), \quad w_2 = x^2(1 - x) \tag{2.1.10}$$

For the collocation method, x_1 an x_2 must be selected such that the resultant weighted residual, i.e. Eq. (2.1.3), can produce two independent equations to determine unknowns a_1 and a_2 uniquely. The least squares method produces a symmetric matrix regardless of a chosen trial function. Example 2.1.1 shows symmetry of the matrix resulting from the least squares method. Galerkin's method does not result in a symmetric matrix when it is applied to Eq. (2.1.1). However, Galerkin's method may produce a symmetric matrix under certain conditions as explained in the next section.

♣ **Example 2.1.1**   A differential equation is written as

$$L(u) = f \tag{2.1.11}$$

where L is a linear differential operator. A trial solution is chosen such that

$$\tilde{u} = \sum_{i=1}^{n} a_i g_i \tag{2.1.12}$$

in which g_i is a known function in terms of the spatial coordinate system and it is assumed to satisfy boundary conditions. Substitution of Eq. (2.1.12) into Eq.

(2.1.11) and collection of terms with the same coefficient a_i yield the residual as seen below;

$$R = \sum_{i=1}^{n} a_i h_i + p \tag{2.1.13}$$

Here, h_i and p are functions in terms of the spatial coordinate system. Test functions for the least squares method are

$$w_j = h_j, \quad j = 1, 2, \ldots, n \tag{2.1.14}$$

The weighted average of the residual over the domain yields the matrix equation

$$I = \int_{\Omega} w_j R \, d\Omega = \sum_{i=1}^{n} A_{ij} a_i - b_j = 0, \qquad j = 1, 2, \ldots, n \tag{2.1.15}$$

where

$$A_{ij} = \int_{\Omega} h_i h_j d\Omega \tag{2.1.16}$$

Equation (2.1.16) shows that $A_{ij} = A_{ji}$ (symmetry). ‡

2.2 Weak Formulation

We consider the previous sample problem, Eq. (2.1.1), again. The formulation described in the preceeding section is called the *strong formulation* of the weighted residual method. The strong formulation requires evaluation of $\int_0^1 w(\partial^2 \tilde{u}/\partial x^2) dx$, which includes the highest order of derivative term in the differential equation. The integral must have a non-zero finite value to yield a meaningful approximate solution to the differential equation. This means a trial function should be differentiable twice and its second derivative should not vanish.

So as to reduce the requirement for a trial function in terms of order of differentiability, integration by parts is applied to the strong formulation. Then Eq. (2.1.3) becomes

$$\begin{aligned} I &= \int_0^1 w \left(\frac{d^2 \tilde{u}}{dx^2} - \tilde{u} + x \right) dx \\ &= \int_0^1 \left(-\frac{dw}{dx} \frac{d\tilde{u}}{dx} - w\tilde{u} + xw \right) dx + \left[w \frac{d\tilde{u}}{dx} \right]_0^1 = 0 \end{aligned} \tag{2.2.1}$$

As seen in Eq. (2.2.1), the trial function needs the first order differentiation instead of the second order differentiation. As a result, the requirement for the trial function is reduced for Eq. (2.2.1). This formulation is called the *weak formulation*.

Weak formulation has an advantage for Galerkin's method where test functions are obtained directly from the selected trial function. If a governing differential equation is the self-adjoint operator, Galerkin's method along with the *weak formulation*
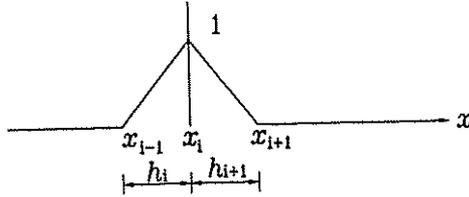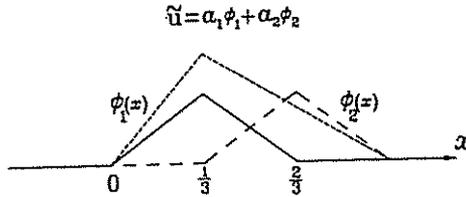
Figure 2.3.1 Piecewise Linear Functions

Figure 2.3.2 Piecewise Linear Trial Function

results in a symmetric matrix in terms of unknown coefficients of the trial function. Using a trial function $\tilde{u} = ax(1 - x)$ for the *weak formulation*, Eq. (2.2.1) results in the same solution as obtained from the *strong formulation* as expected. However, when a piecewise function is selected as a trial function, we see the advantage of the *weak formulation* over the *strong formulation*.

2.3 Piecewise Continuous Trial Function

Regardless of the weak or strong formulation, the accuracy of an approximate solution so much depends on the chosen trial function. However, assuming a proper trial function for the unknown exact solution is not an easy task. This is especially true when the unknown exact solution is expected to have a large variation over the problem domain, the domain has a complex shape in two-dimensional or three-dimensional problems, and/or the problem has complicated boundary conditions. In order to overcome these problems, a trial function can be described using piecewise continuous functions.

Consider piecewise linear functions in an one-dimensional domain as defined below:

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h_i & \text{for } x_{i-1} \leq x \leq x_i \\ (x_{i+1} - x)/h_{i+1} & \text{for } x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases} \qquad (2.3.1)$$

The function defined in Eq. (2.3.1) is plotted in Fig. 2.3.1 and Example 2.3.1 illustrates the use of the function as a trial function.

♣ **Example 2.3.1**   Consider the same problem as given in Eq. (2.1.1). It is rewritten here

$$\begin{cases} \frac{d^2u}{dx^2} - u = -x, & 0 < x < 1 \\ u(0) = 0, \text{ and } u(1) = 0 \end{cases} \tag{2.1.1}$$

The weak formulation is also rewritten as below:

$$I = \int_0^1 w\left(\frac{d^2\tilde{u}}{dx^2} - \tilde{u} + x\right)dx$$

$$= \int_0^1 \left(-\frac{dw}{dx}\frac{d\tilde{u}}{dx} - w\tilde{u} + xw\right)dx + \left[w\frac{d\tilde{u}}{dx}\right]_0^1 = 0 \tag{2.2.1}$$

A trial function is chosen such that $\tilde{u} = a_1\phi_1(x) + a_2\phi_2(x)$ in which a_1 and a_2 are unknown constants to be determined, and ϕ_1 and ϕ_2 are defined as below:

$$\phi_1(x) = \begin{cases} 3x, & 0 \le x \le \frac{1}{3} \\ 2 - 3x, & \frac{1}{3} \le x \le \frac{2}{3} \\ 0, & \frac{2}{3} \le x \le 1 \end{cases} \tag{2.3.2}$$

$$\phi_2(x) = \begin{cases} 0, & 0 \le x \le \frac{1}{3} \\ 3x - 1, & \frac{1}{3} \le x \le \frac{2}{3} \\ 3 - 3x, & \frac{2}{3} \le x \le 1 \end{cases} \tag{2.3.3}$$

$\phi_1(x)$ and $\phi_2(x)$ are plotted in Fig. 2.3.2. For the present trial function, the problem domain is divided into three subdomains and two piecewise linear functions are used. Of course, more piecewise functions can be used along with more subdomains to improve accuracy of the approximate solution. The trial function can be rewritten as

$$\tilde{u} = \begin{cases} a_1(3x), & 0 \le x \le \frac{1}{3} \\ a_1(2 - 3x) + a_2(3x - 1), & \frac{1}{3} \le x \le \frac{2}{3} \\ a_2(3 - 3x), & \frac{2}{3} \le x \le 1 \end{cases} \tag{2.3.4}$$

Use of Galerkin's method yields the following test functions

$$w_1 = \begin{cases} 3x, & 0 \le x \le \frac{1}{3} \\ 2 - 3x, & \frac{1}{3} \le x \le \frac{2}{3} \\ 0, & \frac{2}{3} \le x \le 1 \end{cases} \tag{2.3.5}$$

and

$$w_2 = \begin{cases} 0, & 0 \le x \le \frac{1}{3} \\ 3x - 1, & \frac{1}{3} \le x \le \frac{2}{3} \\ 3 - 3x, & \frac{2}{3} \le x \le 1 \end{cases} \tag{2.3.6}$$

Averaged weighted residuals are

$$I_1 = \int_0^1 \left(-\frac{dw_1}{dx}\frac{d\tilde{u}}{dx} - w_1\tilde{u} + xw_1\right)dx = 0 \tag{2.3.7}$$

$$I_2 = \int_0^1 \left(-\frac{dw_2}{dx}\frac{d\tilde{u}}{dx} - w_2\tilde{u} + xw_2\right)dx = 0 \tag{2.3.8}$$

where $[w\frac{d\bar{u}}{dx}]_0^1$ is omitted because $w_1(0)=w_1(1)=w_2(0)=w_2(1)=0$. Substitution of both trial and test functions into Eq. (2.3.7) and Eq. (2.3.8) respectively gives

$$I_1 = \int_0^{\frac{1}{3}} [-3(3a_1) - 3x(3a_1x) + x(3x)]dx+$$

$$\int_{\frac{1}{3}}^{\frac{2}{3}} [3(-3a_1 + 3a_2) - (2 - 3x)(2a_1 - 3a_1x + 3a_2x - a_2) \qquad (2.3.9)$$

$$+ x(2 - 3x)]dx + \int_{\frac{2}{3}}^1 0dx$$

$$= -6.222a_1 + 2.9444a_2 + 0.1111 = 0$$

$$I_2 = \int_0^{\frac{1}{3}} 0dx + \int_{\frac{1}{3}}^{\frac{2}{3}} [-3(-3a_1 + 3a_2)$$

$$- (3x - 1)(2a_1 - 3a_1x + 3a_2x - a_2) + x(3x - 1)]dx+ \qquad (2.3.10)$$

$$\int_{\frac{2}{3}}^1 [3(-3a_2) - (3 - 3x)(3a_2 - 3a_2x) + x(3 - 3x)]dx$$

$$= 2.9444a_1 - 6.2222a_2 + 0.2222 = 0$$

Solutions for a_1 and a_2 are $a_1 = 0.0488$ and $a_2 = 0.0569$ from Eq. (2.3.9) and Eq. (2.3.10). That is, the approximate solution is $\bar{u} = 0.0448\phi_1(x) + 0.0569\phi_2(x)$. If the trial function Eq. (2.3.4) were used for the strong formulation Eq. (2.1.3), it would not give a reasonable, approximate solution because $\frac{d^2\bar{u}}{dx^2}$ vanishes completely over the domain. ‡

2.4 Galerkin's Finite Element Formulation

As seen in the previous section, use of piecewise continuous functions for the trial function has advantages. As we increase the number of subdomains for the piecewise functions, we can represent a complex function by using sum of simple piecewise linear functions. Later, the subdomains are called finite elements. From now on, ~ used to denote a trial function is omitted unless there is any confusion.

This section shows how to compute weighted residual in a systematic manner using finite elements and piecewise continuous functions. In the previous section, the piecewise continuous functions were defined in terms of the generalized coefficients (i.e. a_1, a_2, etc.). For a systematic formulation, the piecewise continuous functions are defined in terms of nodal variables.

Consider a subdomain or a finite element shown in Fig. 2.4.1. The element has two nodes, one at each end. At each node, the corresponding coordinate value (x_i or x_{i+1}) and the nodal variable (u_i or u_{i+1}) are assigned. Let us assume the unknown trial function to be
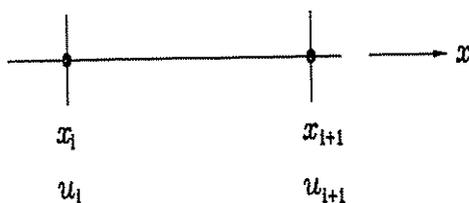
$$u = c_1x + c_2 \qquad (2.4.1)$$

Figure 2.4.1 Two-Node Linear Element

We want to express Eq. (2.4.1) in terms of nodal variables. In other words, c_1 and c_2 need to be replaced by u_i and u_{i+1}. To this end, we evaluate u at $x = x_i$ and $x = x_{i+1}$. Then

$$u(x_i) = c_1 x_i + c_2 = u_i \tag{2.4.2}$$

$$u(x_{i+1}) = c_1 x_{i+1} + c_2 = u_{i+1} \tag{2.4.3}$$

Solving Eq. (2.4.2) and Eq. (2.4.3) simultaneously for c_1 and c_2 gives

$$c_1 = \frac{u_{i+1} - u_i}{x_{i+1} - x_i} \tag{2.4.4}$$

$$c_2 = \frac{u_i x_{i+1} - u_{i+1} x_i}{x_{i+1} - x_i} \tag{2.4.5}$$

Substitution of Eq. (2.4.4) and Eq. (2.4.5) back into Eq. (2.4.1) and rearrangement of the resultant expression result in

$$u = H_1(x) u_i + H_2(x) u_{i+1} \tag{2.4.6}$$

where

$$H_1(x) = \frac{x_{i+1} - x}{h_i} \tag{2.4.7}$$

$$H_2(x) = \frac{x - x_i}{h_i} \tag{2.4.8}$$

$$h_i = x_{i+1} - x_i \tag{2.4.9}$$

Equation (2.4.6) gives an expression for the variable u in terms of nodal variables, and Eq. (2.4.7) and Eq. (2.4.8) are called linear shape functions. The shape functions are plotted in Fig. 2.4.2. These functions have the following properties:

1. The shape function associated with node i has a unit value at node i and vanishes at other nodes. That is,

$$H_1(x_i) = 1, \ H_1(x_{i+1}) = 0, \ H_2(x_i) = 0, \ H_2(x_{i+1}) = 1 \tag{2.4.10}$$
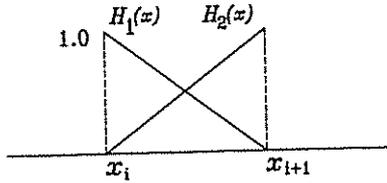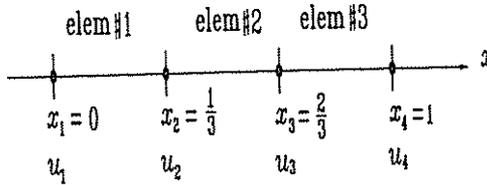
Figure 2.4.2   Linear Shape Functions

Figure 2.4.3   Finite Element Mesh With 3 Linear Elements

2.   Sum of all shape functions is unity.

$$\sum_{i=1}^{2} H_i(x) = 1 \qquad (2.4.11)$$

These are important properties for shape functions. The first property, Eq. (2.4.10), states that the variable u must be equal to the corresponding nodal variable at each node (i.e. $u(x_i) = u_i$ and $u(x_{i+1}) = u_{i+1}$ as enforced in Eq. (2.4.2) and Eq. (2.4.3)). The second property, Eq. (2.4.11), tells that the variable u can represent a uniform solution within the element. If the solution remains constant within the element, $u = u_i = u_{i+1}$. Substitution of this condition into Eq. (2.4.6) gives

$$u = \{H_1(x) + H_2(x)\}u_i = u_i \qquad (2.4.12)$$

Equation (2.4.12) results in the second property of shape functions, Eq. (2.4.10).

♣ **Example 2.4.1**   We solve the same problem as given in Example 2.3.1 using the linear finite elements. The weighted residual can be written as

$$I = \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} \left(-\frac{dw}{dx}\frac{du}{dx} - wu + xw \right) dx + \left[u'w \right]_0^1 = 0 \qquad (2.4.13)$$

for n elements. If the problem domain is discretized into three equal size of elements, i.e. $n = 3$, Fig. 2.4.3 shows the corresponding finite element mesh. Consider the ith element (i.e. $i=1$, 2, or 3). The integral for this element is

$$\int_{x_i}^{x_{i+1}} \left(-\frac{dw}{dx}\frac{du}{dx} - wu + xw \right) dx \qquad (2.4.14)$$

The trial function u is expressed as

$$u = H_1(x)u_i + H_2(x)u_{i+1} \qquad (2.4.6)$$

and test functions for Galerkin's method are $w_1 = H_1(x)$ and $w_2 = H_2(x)$. Putting these u and w into Eq. (2.4.13) gives

$$-\int_{x_i}^{x_{i+1}} \left(\begin{Bmatrix} H_1' \\ H_2' \end{Bmatrix} [H_1' H_2'] + \begin{Bmatrix} H_1 \\ H_2 \end{Bmatrix} [H_1 H_2] \right) dx \begin{Bmatrix} u_i \\ u_{i+1} \end{Bmatrix}$$

$$+\int_{x_i}^{x_{i+1}} x \begin{Bmatrix} H_1 \\ H_2 \end{Bmatrix} dx \qquad (2.4.15)$$

where H_i' denotes $\frac{dH_i(x)}{dx}$ and H_i is given in Eq. (2.4.7) and Eq. (2.4.8). Computation of these integrals finally yields

$$-\begin{bmatrix} \frac{1}{h_i} + \frac{h_i}{3} & -\frac{1}{h_i} + \frac{h_i}{6} \\ -\frac{1}{h_i} + \frac{h_i}{6} & \frac{1}{h_i} + \frac{h_i}{3} \end{bmatrix} \begin{Bmatrix} u_i \\ u_{i+1} \end{Bmatrix} + \begin{Bmatrix} \frac{h_i}{6}(x_{i+1} + 2x_i) \\ \frac{h_i}{6}(2x_{i+1} + x_i) \end{Bmatrix} \qquad (2.4.16)$$

For each element, Eq. (2.4.16) can be written as

Element #1

$$\begin{bmatrix} -3.111 & 2.9444 \\ 2.9444 & -3.111 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} + \begin{Bmatrix} 0.0185 \\ 0.0370 \end{Bmatrix} \qquad (2.4.17)$$

Element #2

$$\begin{bmatrix} -3.111 & 2.9444 \\ 2.9444 & -3.111 \end{bmatrix} \begin{Bmatrix} u_2 \\ u_3 \end{Bmatrix} + \begin{Bmatrix} 0.0741 \\ 0.0926 \end{Bmatrix} \qquad (2.4.18)$$

Element #3

$$\begin{bmatrix} -3.111 & 2.9444 \\ 2.9444 & -3.111 \end{bmatrix} \begin{Bmatrix} u_3 \\ u_4 \end{Bmatrix} + \begin{Bmatrix} 0.1296 \\ 0.1481 \end{Bmatrix} \qquad (2.4.19)$$

As shown in Eq. (2.4.13), we need to sum Eqs (2.4.17) through (2.14.19). Each element has different nodes associated with it. As a result, we expand each equation such that the equation has a matrix and a vector of size m which is the total number of degrees of freedom in the system. For the present problem, $m = 4$. The number of total degrees of freedom is the same as the total number of nodes because each node has one degree of freedom for the present problem. Rewriting Eq. (2.4.17) for the expanded matrix and vector gives

$$\begin{bmatrix} -3.111 & 2.9444 & 0 & 0 \\ 2.944 & -3.111 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix} + \begin{Bmatrix} 0.0185 \\ 0.0370 \\ 0 \\ 0 \end{Bmatrix} \qquad (2.4.20)$$

Similarly, Eq. (2.4.18) and Eq. (2.4.19) can be rewritten as

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -3.111 & 2.9444 & 0 \\ 0 & 2.9444 & -3.1111 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix} + \begin{Bmatrix} 0 \\ 0.0741 \\ 0.0926 \\ 0 \end{Bmatrix} \qquad (2.4.21)$$

$$
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & -3.1111 & 2.9444 \\
0 & 0 & 2.9444 & -3.1111
\end{bmatrix}
\begin{Bmatrix}
u_1 \\ u_2 \\ u_3 \\ u_4
\end{Bmatrix}
+
\begin{Bmatrix}
0 \\ 0 \\ 0.1296 \\ 0.1481
\end{Bmatrix}
\tag{2.4.22}
$$

Adding directly Eqs. (2.4.20) through (2.4.22) results in

$$
\begin{bmatrix}
-3.1111 & 2.9444 & 0 & 0 \\
2.9444 & -6.2222 & 2.9444 & 0 \\
0 & 2.9444 & -6.2222 & 2.9444 \\
0 & 0 & 2.9444 & -3.1111
\end{bmatrix}
\begin{Bmatrix}
u_1 \\ u_2 \\ u_3 \\ u_4
\end{Bmatrix}
$$
$$
+
\begin{Bmatrix}
0.0185 - u'(0) \\
0.1111 \\
0.2222 \\
0.1481 + u'(1)
\end{Bmatrix}
= 0
\tag{2.4.23}
$$

The Neuman boundary conditions are added to the right-hand side vector from Eq. (2.4.13). For the present problem, the Dirichlet boundary conditions are provided at both ends (i.e. $u_1 = 0$ and $u_4 = 0$). Therefore, the Neumann boundary conditions (i.e. $u'(0)$ and $u'(1)$) are not provided. Equation (2.4.23) can be solved with the given boundary conditions, $u_1 = 0$ and $u_4 = 0$, to find the rest of nodal variables and unknown Neumann boundary conditions. In actual finite element programming, Eqs (2.4.17) through (2.4.19) are directly summed into Eq. (2.4.23) without using Eqs (2.4.20) through (2.4.22). Equations (2.4.20) through (2.4.22) are used here only to help the conceptual understanding of the assembly process. Furthermore, in computer programming, unknown nodal values, called the primary variables, are solved first and then the unknown boundary conditions are solved later. To this end, Eq. (2.4.23) is modified with the known boundary conditions.

$$
\begin{bmatrix}
1 & 0 & 0 & 0 \\
2.9444 & -6.2222 & 2.9444 & 0 \\
0 & 2.9444 & -6.2222 & 2.9444 \\
0 & 0 & 0 & 1
\end{bmatrix}
\begin{Bmatrix}
u_1 \\ u_2 \\ u_3 \\ u_4
\end{Bmatrix}
=
\begin{Bmatrix}
0 \\ -0.1111 \\ -0.2222 \\ 0
\end{Bmatrix}
\tag{2.4.24}
$$

The first and last equations in Eq. (2.4.23) are replaced by the Dirichlet boundary conditions. From Eq. (2.4.24), the solution gives $u_1 = 0$, $u_2 = 0.0448$, $u_3 = 0.0569$, and $u_4 = 0$. These nodal solutions can be substituted into Eq. (2.4.23) to find $u'(0)$ and $u'(1)$. Once the nodal variables are determined, the solution within each element can be obtained from corresponding nodal variables and shape functions. For example, the solution within the first element $(0 \leq x \leq \frac{1}{3})$ is $u = H_1(x)u_1 + H_2(x)u_2 = 0.1344x$. ‡

2.5 Variational Method

The variational method is also commonly used to derive the finite element matrix equation. We want to derive the functional for the sample problem

$$\begin{cases} \frac{d^2u}{dx^2} - u = -x, & 0 < x < 1 \\ u(0) = 0, \text{ and } u(1) = 0 \end{cases} \tag{2.1.1}$$

The variational expression for Eq. (2.1.1) is

$$\delta J = \int_0^1 \left(-\frac{d^2u}{dx^2} + u - x \right) \delta u \, dx + \left[\frac{du}{dx} \delta u \right]_0^1 \tag{2.5.1}$$

where δ is the *variational* operator. The first term in the above equation is the differential equation and the second term is the unknown *Neumann boundary condition* (or natural boundary condition). Applying integration by parts to the first term of Eq. (2.5.1) yields

$$\delta J = \int_0^1 \left(\frac{du}{dx} \frac{d(\delta u)}{dx} + u\delta u - x\delta u \right) dx \tag{2.5.2}$$

Since the *variational* operator is commutative with both differential and integral operators (i.e. $\frac{d(\delta u)}{dx} = \delta(\frac{du}{dx})$ and $\int \delta u \, dx = \delta \int u \, dx$), Eq. (2.5.2) can be written as

$$\delta J = \delta \int_0^1 \left\{ \frac{1}{2} \left(\frac{du}{dx} \right)^2 + \frac{1}{2} u^2 - xu \right\} dx \tag{2.5.3}$$

The functional is obtained from Eq. (2.5.3) as

$$J = \int_0^1 \left\{ \frac{1}{2} \left(\frac{du}{dx} \right)^2 + \frac{1}{2} u^2 - xu \right\} dx \tag{2.5.4}$$

Conversely, taking variation of Eq. (2.5.4) will result in the differential equation as given in Eq. (2.1.1). Functional represents energy in many engineering applications. For example, the total potential energy in solid mechanics is a functional. The solution to the governing equation is obtained by minimizing the functional. The *principle of minimum total potential energy* in solid mechanics is one example to determine the stable equilibrium solution [4,5]. Energy principles are discussed in later chapters. For more detailed information for *variational method*, readers may refer to Refs [6-8].

2.6 Rayleigh-Ritz Method

The *Rayleigh-Ritz* method obtains an approximate solution to a differential equation with given boundary conditions using the functional of the equation. The procedure of this technique can be summarized in two steps as given below:

1. Assume an admissible solution which satisfies the *Dirichlet* boundary condition (or essential boundary condition) and contains unknown coefficients.

2. Substitute the assumed solution into the functional and find the unknown coefficients to minimize the functional.

♣ **Example 2.6.1** In order to solve Eq. (2.1.1) using the *Rayleigh-Ritz* method, we assume the following function as an approximate solution.

$$u = ax(1 - x) \tag{2.6.1}$$

where a is an unknown coefficient. This function satisfies the essential boundary conditions. Substituting Eq. (2.6.1) into the functional, Eq. (2.5.4), yields

$$J = \frac{1}{2}a^2 \int_0^1 [(1 - 2x)^2 + x^2(1 - x)^2]dx - a \int_0^1 x^2(1 - x)dx \tag{2.6.2}$$

Minimizing the functional with respect to the unknown coefficient a, i.e. $\frac{dJ}{da}=0$, yields a=0.2272. Therefore, the approximate solution is $u = 0.2272x(1 - x)$ which is the same as that obtained in Sec. 2.1 using Galerkin's method. In order to improve the approximate solution, we need to add more terms. For example, we may assume

$$u = a_1 x(1 - x) + a_2 x^2(1 - x) \tag{2.6.3}$$

where a_1 and a_2 are two unknown coefficients. We substitute the expression into the functional and take derivatives with respect to a_1 and a_2 in order to minimize the functional.

$$\frac{\partial J}{\partial a_1} = 0 \quad \text{and} \quad \frac{\partial J}{\partial a_2} = 0 \tag{2.6.4}$$

This operation will give solutions for unknown coefficients a_1 and a_2. ‡

2.7 Rayleigh-Ritz Finite Element Method

The *Rayleigh-Ritz* method can be applied to a problem domain using continuous piecewise functions. As a result, the problem domain is divided into subdomains of finite elements. For elements with two nodes apiece, the linear shape functions as in Eqs (2.4.7) and (2.4.8) can be used for the *Rayleigh-Ritz* method. The following example explains the finite element procedure using the *Rayleigh-Ritz* method.

♣ **Example 2.7.1** We will solve Example 2.4.1 again using the Rayleigh-Ritz method. The problem domain and its discretization are shown in Fig. 2.4.3. The functional can be expressed for the discretized domain as

$$J = \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} \left\{ \frac{1}{2}\left(\frac{du}{dx}\right)^2 + \frac{1}{2}u^2 - xu \right\}dx \tag{2.7.1}$$

where $n = 3$, $x_1 = 0$, $x_2 = 1/3$, $x_3 = 2/3$ and $x_4 = 1$ as shown in Fig. 2.4.3. Using the linear shape functions, the solution u for the ith element is expressed as

$$u = H_1(x)\, u_i + H_2(x)\, u_{i+1} = [H]\{u^i\} \tag{2.7.2}$$

where

$$[H] = [H_1 \quad H_2] \tag{2.7.3}$$

$$\{u^i\} = \{u_i \quad u_{i+1}\}^T \tag{2.7.4}$$

and H_1 and H_2 are given in Eqs (2.4.7) and (2.4.8). Substituting Eq. (2.7.2) into the functional yields

$$\int_{x_i}^{x_{i+1}} \left\{ \frac{1}{2}\left(\frac{du}{dx}\right)^2 + \frac{1}{2}u^2 - xu \right\} dx = \int_{x_i}^{x_{i+1}} \left\{ \frac{1}{2}\{u^i\}^T \left[\frac{dH}{dx}\right]^T \left[\frac{dH}{dx}\right]\{u^i\} + \right.$$

$$\left. \frac{1}{2}\{u^i\}^T [H]^T [H]\{u^i\} - \{u^i\}^T [H]^T x \right\} dx \tag{2.7.5}$$

in which

$$\left[\frac{dH}{dx}\right] = \left[\frac{dH_1}{dx} \quad \frac{dH_2}{dx}\right] \tag{2.7.6}$$

Evaluation of the integral in Eq. (2.7.5) gives

$$\frac{1}{2}\{u_i \; u_{i+1}\} \begin{bmatrix} \frac{1}{h_i} + \frac{h_i}{3} & -\frac{1}{h_i} + \frac{h_i}{6} \\ -\frac{1}{h_i} + \frac{h_i}{6} & \frac{1}{h_i} + \frac{h_i}{3} \end{bmatrix} \begin{Bmatrix} u_i \\ u_{i+1} \end{Bmatrix}$$

$$- \{u_i \; u_{i+1}\} \begin{Bmatrix} \frac{h_i}{6}(x_{i+1} + 2x_i) \\ \frac{h_i}{6}(2x_{i+1} + x_i) \end{Bmatrix} \tag{2.7.7}$$

Here, the matrix expression in Eq. (2.7.7) came from the first and second terms of the right-hand side of Eq. (2.7.5) while the vector expression came from the last term. Summing Eq. (2.7.7) over the total number of elements and substituting proper values give the functional

$$J = \frac{1}{2}\{u_1 \; u_2 \; u_3 \; u_4\} \begin{bmatrix} 3.1111 & -2.9444 & 0 & 0 \\ -2.9444 & 6.2222 & -2.9444 & 0 \\ 0 & -2.9444 & 6.2222 & -2.9444 \\ 0 & 0 & -2.9444 & 3.1111 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix}$$

$$- \{u_1 \; u_2 \; u_3 \; u_4\} \begin{Bmatrix} 0.0185 \\ 0.1111 \\ 0.2222 \\ 0.1481 \end{Bmatrix} \tag{2.7.8}$$

The summation process for Eq. (2.7.8) is the same as explained in Example 2.4.1. In order to find the solution, we need to minimize the functional with respect to

the unknown nodal vector $\{u\} = \{u_1 \ u_2 \ u_3 \ u_4\}^T$. Invoking $\frac{dJ}{d\{u\}} = 0$ results

in

$$\begin{bmatrix} 3.1111 & -2.9444 & 0 & 0 \\ -2.9444 & 6.2222 & -2.9444 & 0 \\ 0 & -2.9444 & 6.2222 & -2.9444 \\ 0 & 0 & -2.9444 & 3.1111 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix}$$

$$- \begin{Bmatrix} 0.0185 \\ 0.1111 \\ 0.2222 \\ 0.1481 \end{Bmatrix} = 0 \qquad\qquad (2.7.9)$$

Applying the boundary conditions $u_1 = 0$ and $u_4 = 0$ to Eq. (2.7.9) yields
Eq. (2.4.24) in Example 2.4.1. The solutions for nodal variables are $u_1 = 0$,
$u_2 = 0.0448$, $u_3 = 0.0569$, and $u_4 = 0$ again as before.   ‡

Some Simple Matrix Operations Using MatLab:

Define a matrix, A

```
>> A = [ 1 2 3 ; 4 5 6]

A =

     1     2     3
     4     5     6
```

Multiply by its transpose A^T

```
>> B = A'*A

B =

    17    22    27
    22    29    36
    27    36    45
```

Add 5 to the 2nd row and 2nd column of B(row,column)

```
>> B(2,2) = B(2,2) + 5

B =

    17    22    27
    22    34    36
    27    36    45
```

Invert B as B^{-1}

```
>> inv(B)

ans =

    1.3000   -0.1000   -0.7000
   -0.1000    0.2000   -0.1000
   -0.7000   -0.1000    0.5222

>> Binverse = inv(B)

Binverse =

    1.3000   -0.1000   -0.7000
   -0.1000    0.2000   -0.1000
   -0.7000   -0.1000    0.5222
```

Evaluate matrix product $(B^{-1})^T B^{-1} = I$

```
>> Binverse' * B
```

```
ans =

    1.0000   -0.0000   -0.0000
    0.0000    1.0000    0.0000
   -0.0000   -0.0000    1.0000
```

Extract the second column of B

```
>> C=B(1:3,2)

C =

    22
    34
    36
```

Solve the equation A x = C by decomposition

```
>> x=B\C

x =

   -0.0000
    1.0000
   -0.0000
```

Solve the equation A x = C by inversion

```
>> x=inv(B)*C

x =

    0.0000
    1.0000
   -0.0000
```

Use Elementary Functions to Solve Simple FE-Type Equations

```
function y = SpringElementStiffness(k)
%SpringElementStiffness    This function returns the element stiffness
%                          matrix for a spring with stiffness k.
%                          The size of the element stiffness matrix
%                          is 2 x 2.
y = [k -k ; -k k];


-------------------------------------------------------


function y = SpringAssemble(K,k,i,j)
%SpringAssemble    This function assembles the element stiffness
%                  matrix k of the spring with nodes i and j into the
%                  global stiffness matrix K.
%                  This function returns the global stiffness matrix K
%                  after the element stiffness matrix k is assembled.
K(i,i) = K(i,i) + k(1,1);
K(i,j) = K(i,j) + k(1,2);
K(j,i) = K(j,i) + k(2,1);
K(j,j) = K(j,j) + k(2,2);
y = K;


-------------------------------------------------------


function y = SpringElementForces(k,u)
%SpringElementForces    This function returns the element nodal force
%                       vector given the element stiffness matrix k
%                       and the element nodal displacement vector u.
y = k * u;
```

Solve:

$$\begin{Bmatrix} q_1 \\ q_2 = 0 \\ q_3 \end{Bmatrix} = 10^{-6} \begin{bmatrix} 1 & -1 & \\ -1 & (1+2) & -2 \\ & -2 & 2 \end{bmatrix} \begin{Bmatrix} h_1 = 20 \\ h_2 \\ h_3 = 25 \end{Bmatrix}$$

```
>> k1=SpringElementStiffness(1e-06)

>> k2=SpringElementStiffness(2e-06)

>> K=zeros(3,3)

>> K=SpringAssemble(K,k1,1,2)

>> K=SpringAssemble(K,k2,2,3)

>> q=zeros(3,1)

>> h=[20; 0; 25]

>> q = q - SpringElementForces(K,h)
```

```
>> K

>> h

>> h(2,1) = q(2,1)/K(2,2)

>> q=K*h
```

Full Solution Including Output

```
>> k1=SpringElementStiffness(1e-06)

k1 =

  1.0e-006 *

    1.0000   -1.0000
   -1.0000    1.0000

>> k2=SpringElementStiffness(2e-06)

k2 =

  1.0e-005 *

    0.2000   -0.2000
   -0.2000    0.2000

>> K=zeros(3,3)

K =

     0     0     0
     0     0     0
     0     0     0

>> K=SpringAssemble(K,k1,1,2)

K =

  1.0e-006 *

    1.0000   -1.0000        0
   -1.0000    1.0000        0
         0        0         0

>> K=SpringAssemble(K,k2,2,3)

K =

  1.0e-005 *

    0.1000   -0.1000        0
   -0.1000    0.3000   -0.2000
         0   -0.2000    0.2000
```

```
>> q=zeros(3,1)

q =

     0
     0
     0

>> h=[20; 0; 25]

h =

    20
     0
    25

>> q = q - SpringElementForces(K,h)

q =

  1.0e-004 *

   -0.2000
    0.7000
   -0.5000

>> K

K =

  1.0e-005 *

    0.1000   -0.1000         0
   -0.1000    0.3000   -0.2000
         0   -0.2000    0.2000

>> h

h =

    20
     0
    25

>> h(2,1) = q(2,1)/K(2,2)

h =

   20.0000
   23.3333
   25.0000

>> q=K*h

q =

  1.0e-005 *
```

```
  -0.3333
   0.0000
   0.3333

>>
```

2
Fluid Flow

[2:1] Fluid Flow and Pressure Diffusion

Recap of FEM

Comsol Applied to Flow

1D Element

3. Hydraulic Behavior (H)   Flow

3.1. Conservation of mass and Darcy's law
3.2. Steady behavior
    3.2.1. 1-dimensional elements
    3.2.2. 2-dimensional behavior – 2-D triangular, and 2-D isoparametric elements
3.3. Transient behavior
    3.3.1. Time stepping methods
3.4. Dual porosity flows

PROCESS COUPLINGS [T-H-M-C]

$$
\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & R_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}
\begin{Bmatrix} \underline{u} \\ \underline{p} \\ \underline{T} \\ \underline{c} \end{Bmatrix}
+
\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & S_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}
\begin{Bmatrix} \underline{\dot{u}} \\ \underline{\dot{p}} \\ \underline{\dot{T}} \\ \underline{\dot{c}} \end{Bmatrix}
=
\begin{Bmatrix} \underline{\dot{f}}+\cdots \\ \underline{q}_F+\cdots \\ \underline{q}_T+\cdots \\ \underline{q}_M+\cdots \end{Bmatrix}
$$

Conductance $\uparrow$ Storage $\uparrow$

Need to Understand:

① R_{22} $\underline{q}_F = \underline{R}_{22}\,\underline{p}$ $\sim$ $\underline{q}_F = \overline{\underline{R}}_{22}\,\underline{h}$

Conductance matrix $\begin{cases} 1-D \\ 2-D \end{cases}$

$R_{22} = \int_V \underline{a}^T \underline{D}\, \underline{a}\; dV$

② S_{22} Form of S_{22} $-$ $S_{22} = S_s \int \underline{b}^T \underline{b}\; dV$

③ Transient behavior:

$$\underline{q}_F = R_{22}\,\underline{p} + S_{22}\,\underline{\dot{p}}$$

Time stepping:

Implicit.

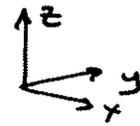Explicit.

EQUIVALENCE OF: HYDRAULIC CONDUCTIVITY (K) & PERMEABILITY (k)

and

HEAD (h) AND PRESSURE (P).

$$\frac{K}{\rho g} = \frac{k}{\mu} \qquad (1)$$

$$h = \frac{P}{\rho g} + z \qquad (2)$$

$$q_i = -\frac{k}{\mu}\left(\frac{\partial p}{\partial x_i} + \rho g \frac{\partial z}{\partial x_i}\right) \qquad (3)$$

Take derivative of (h) and multiply by ρg:

$$\rho g \frac{\partial}{\partial x}(h) = \rho g \frac{1}{\rho g}\frac{\partial p}{\partial x} + \rho g \frac{\partial z}{\partial x} \qquad (4)$$

Substitute (4) into (3)

$$q_i = -\underbrace{\frac{k}{\mu}(\rho g)}_{K}\frac{\partial h}{\partial x} = -K\frac{\partial h}{\partial x} \qquad (5)$$

SYSTEM TYPES

SOLID MECHANICS

o Conservation of momentum:
(Equilibrium), $VM_I = VW_E$

> Continuity (Compatibility):
$$\underline{\varepsilon} = \underline{a}\, u$$

Constitutive relation: $\underline{\sigma} = \underline{D}\,\underline{\varepsilon}$

> Initial Conditions

, Boundary Conditions

FLOW SYSTEM

o Conservation of mass:
$$\underline{\nabla}^T \underline{q} = 0$$

o Continuity: $\underline{h}_{,} = \underline{a}\,h$

o Constitutive reln. $\underline{v} = \underline{D}\,\underline{h}_{,}$

o ICs

o BCs

TRANSPORT

o Conservation of mass
$$\underline{\nabla}^T \underline{q} = 0$$

o Continuity: $\underline{c}_{,} = \underline{a}\,c$

o Constitutive:
diffusion - $\underline{v}_1 = \underline{D}\,\underline{c}_{,}$
advective - $\underline{v}_2 = \underline{A}\,\underline{c}$

o ICs

o BCs

— SOLVE SYSTEM EQUATIONS —

MASS BALANCE — FLOW
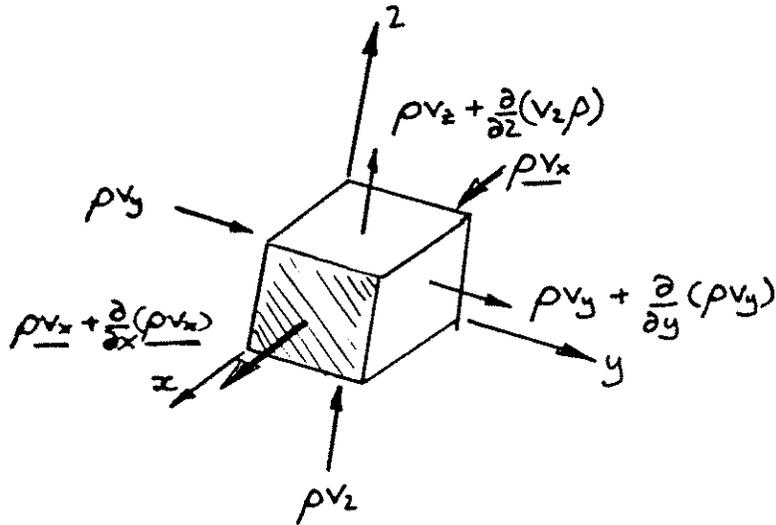
$$M_a = \text{MASS RATE IN} - \text{MASS RATE OUT}$$

Figure 2.2.1 Unit differential cube, dx = dy = dz = 1

For x-direction:
$$\rho v_x - \left[\rho v_x + \frac{\partial}{\partial x}(\rho v_x)\right] = M_a$$

gives
$$-\frac{\partial}{\partial x}(\rho v_x) - \frac{\partial}{\partial y}(\rho v_y) - \frac{\partial}{\partial z}(\rho v_z) = M_a$$

Assume ρ const. then
$$\rho\left[-\frac{\partial}{\partial x}(v_x) - \frac{\partial}{\partial y}(v_y) - \frac{\partial}{\partial z}(v_z)\right] = \rho \, S_s \frac{\partial h}{\partial t}$$
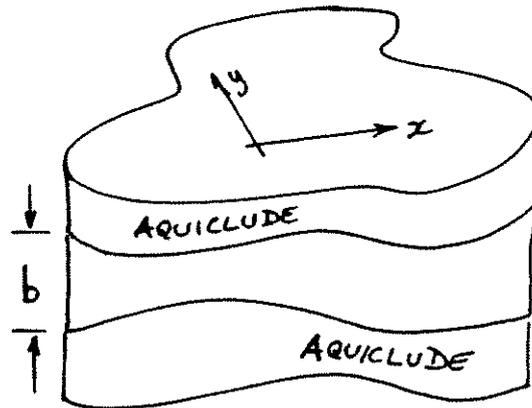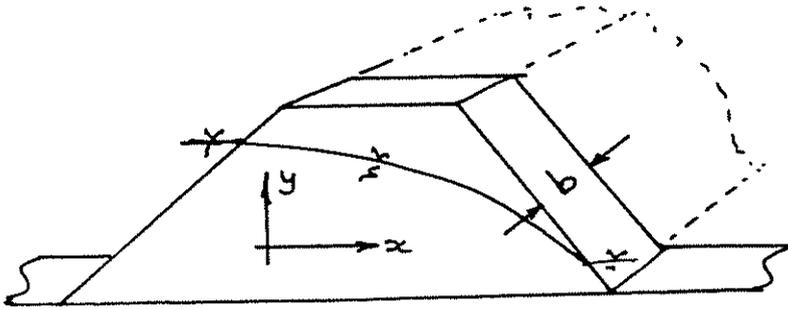
Continuity equation.

Special case

Figure 2.2.2. Two dimensional geometries (a) Confined vertical
section; (b) Confined flow in a horizontal aquifer.

Darcy's Law $\quad V_x = -K_x \dfrac{\partial h}{\partial x} \quad$ etc.

Substitute into continuity equation

$$\frac{\partial}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial h}{\partial z}\right) = S_s \frac{\partial h}{\partial t}$$
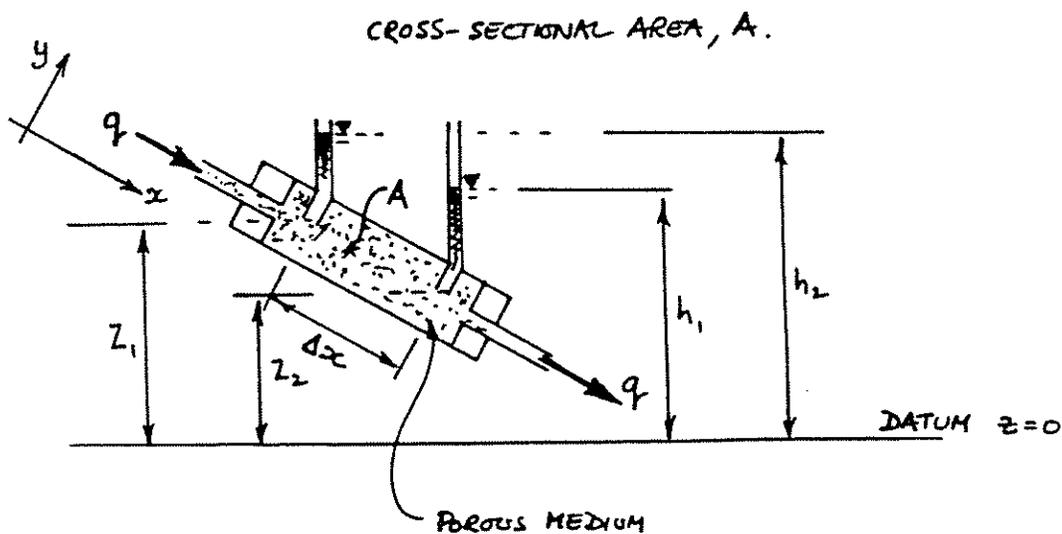
Figure 2.1.2.1 Constant head permeameter as defined by D'Arcy's experiment

Other variants of equation:

$$(2\text{-}D) \qquad K_x \frac{\partial^2 h}{\partial x^2} + K_y \frac{\partial^2 h}{\partial y^2} = S_s \frac{\partial h}{\partial t}$$

$$(\text{Aerial } 2\text{-}D) \qquad T = Kb \; ; \quad S = S_s b$$
$$(\text{Areal})$$
$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = \frac{S}{T} \frac{\partial h}{\partial t}$$
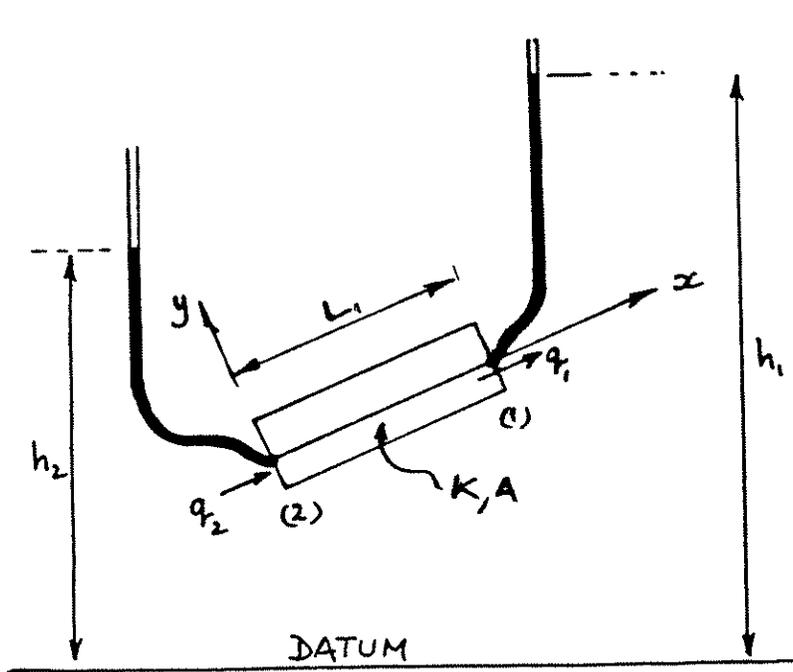
$$\underline{Darcy's\ Law}$$

$$q_1 = +A_1 K_1 \frac{(h_2 - h_1)}{L_1}$$

$$q_2 = -q_1$$

$$\begin{Bmatrix} q_1 \\ q_2 \end{Bmatrix} = -\frac{A_1 k_1}{L_1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} h_1 \\ h_2 \end{Bmatrix}$$

$$\underline{q} = \underbrace{\underline{K}}\qquad \underline{h}$$
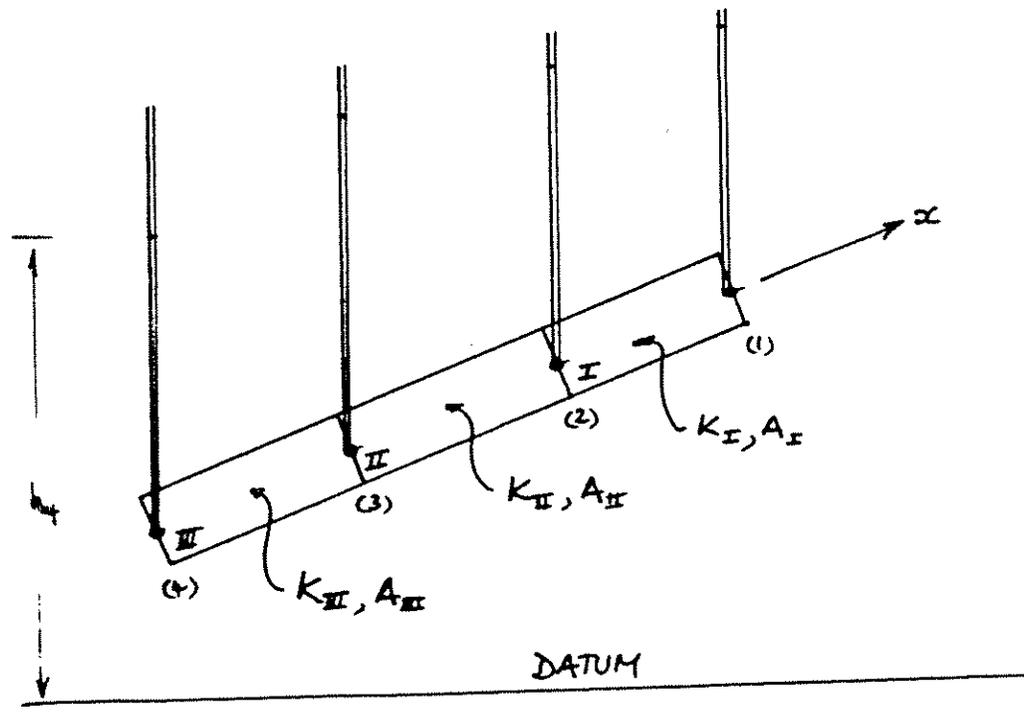
Figure 2.3.1 (a) Single element representing flow
in a pipe; (b) Multiple elements joined in series

(a) REAL

20 m

IMPERMEABLE

$K_I = 10^{-6}$

10m

$K_{II} = 2 \times 10^{-6}$ m/s

IMPERMEABLE

25m

20m

DATUM

(b) GLOBAL MESH

10m

10m

$h_1 = 20m$

$h_3 = 25m$

10m

(1) (2) (3)
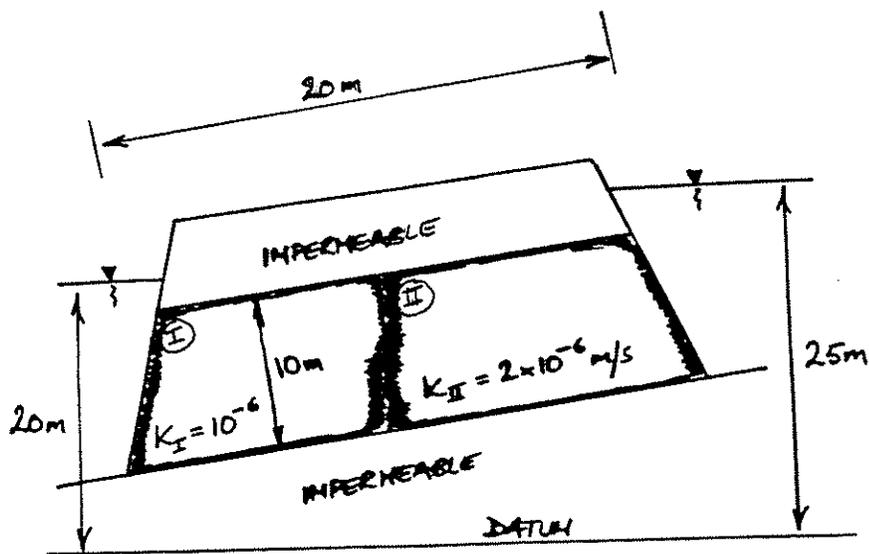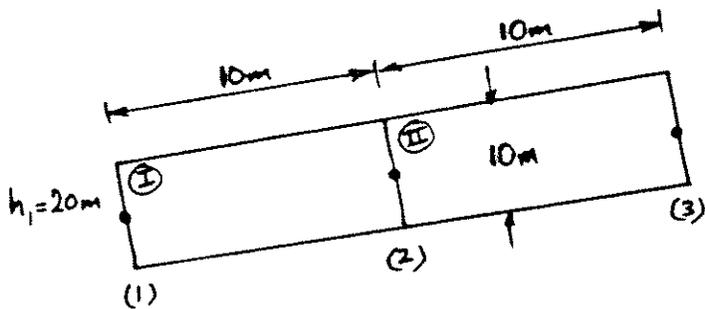
$A = 10m^2$

$K_I = 10^{-6} m/s$; $K_{II} = 2 \times 10^{-6} m/s$

$L = 10 m$

(c) LOCAL ELEMENT

CROSS-SECTIONAL AREA, A

L

(1) (2)
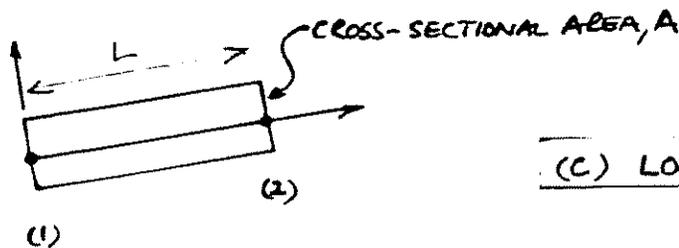
$$\underline{K} = \frac{AK}{L} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Figure 2.3.2 Illustrative example for one dimensional flow

ONE DIMENSIONAL EXAMPLE

$$\left.\begin{array}{l} A = 10\,m^2 \\ K = 10^{-6}\,m/s \\ L = 10\,m \end{array}\right\} \quad \frac{AK}{L} = 10^{-6}\,(m^2/s)$$

$$K_I = 10^{-6}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad\qquad K_{II} = 2\times10^{-6}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\begin{Bmatrix} q_1 \\ q_2 \\ q_3 \end{Bmatrix} = 10^{-6}\begin{bmatrix} 1 & -1 & 0 \\ -1 & (1+2) & -2 \\ 0 & -2 & 2 \end{bmatrix}\begin{Bmatrix} h_1 = 20 \\ h_2 = ? \\ h_3 = 25 \end{Bmatrix}$$

Solve: $\quad q_2 = 0 \quad$ since no prescribed flux

solve central equation only !!

$$\overset{0}{\cancel{q_2}} + 20\times10^{-6} + 2\times25\times10^{-6} = 3\times10^{-6}\times h_2$$

$$\frac{70}{3} = h_2$$

Resubstitute to determine flux magnitude:

$$q_1 = \left(20 - \frac{70}{3}\right)\times10^{-6} = -3.33\times10^{-6}\,m^3/s$$

$$q_2 = 0$$

$$q_3 = \left(-2\left(\frac{70}{3}\right) + 2(25)\right)\times10^{-6} = +3.33 \underset{\times10^{-6}}{}\,m^3/s$$

$$\left.\begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}\right\} \begin{array}{l} \text{Fluxes sum} \\ \text{to zero!} \end{array}$$

[2:2] Fluid Flow and Pressure Diffusion

Recap

Conservation of Mass

Galerkin Formulation

1D Element and Analysis

CONSERVATION OF MASS

Conservation Equation:
$$\frac{d\rho}{dt} + \frac{d}{dx}(\rho V_x) + \frac{d}{dy}(\rho V_y) = 0 \qquad (1)$$

Incompressible Fluid: $\qquad \rho = \text{constant} \qquad$ (in space and in time)

$$\therefore \quad d\rho/dt = 0 \qquad\qquad d/dx(\rho V_x) = \rho\, dV_x/dx + V_x\, \overset{0}{\cancel{d\rho/dx}}$$

$$\therefore \qquad \frac{dV_x}{dx} + \frac{dV_y}{dy} = 0 \qquad\qquad (2)$$

Slightly Compressible Fluid (Darcian): $\qquad (v^2/2g \rightarrow \text{small})$

$$d\rho/dt \quad = \quad \underbrace{d\rho/dP}\; dP/dt \qquad\qquad P = \text{reduced pressure}$$
$$P = p - \rho g z \qquad (3)$$

$$V_x \quad = \quad -\frac{k}{\mu}\frac{dP}{dx} \qquad\qquad\qquad (4)$$

Alternately: $\quad \dfrac{d\rho}{dt} = \dfrac{d}{dt}\left(\rho\dfrac{\Psi}{\Psi}\right) = \dfrac{1}{V}\dfrac{d}{dt}(\rho\Psi) = \dfrac{1}{\Psi}\rho\dfrac{d\Psi}{dt} + \dfrac{1}{\Psi}\Psi\,\overset{0}{\cancel{\dfrac{d\rho}{dt}}} = \rho\dfrac{1}{\Psi}\dfrac{d\Psi}{dP}\dfrac{dP}{dt}$

$$(5)$$

Combining (4) and (5) into (1)

$$\cancel{\rho}\,\frac{1}{\Psi}\underbrace{\frac{d\Psi}{dP}}\frac{dP}{dt}\cancel{\Psi} = \cancel{\rho}\,\frac{k}{\mu}\frac{d}{dx}\frac{dP}{dx} + \cancel{\rho}\,\frac{k}{\mu}\frac{d}{dy}\frac{dP}{dy} \qquad (6).$$

Compressibility, β
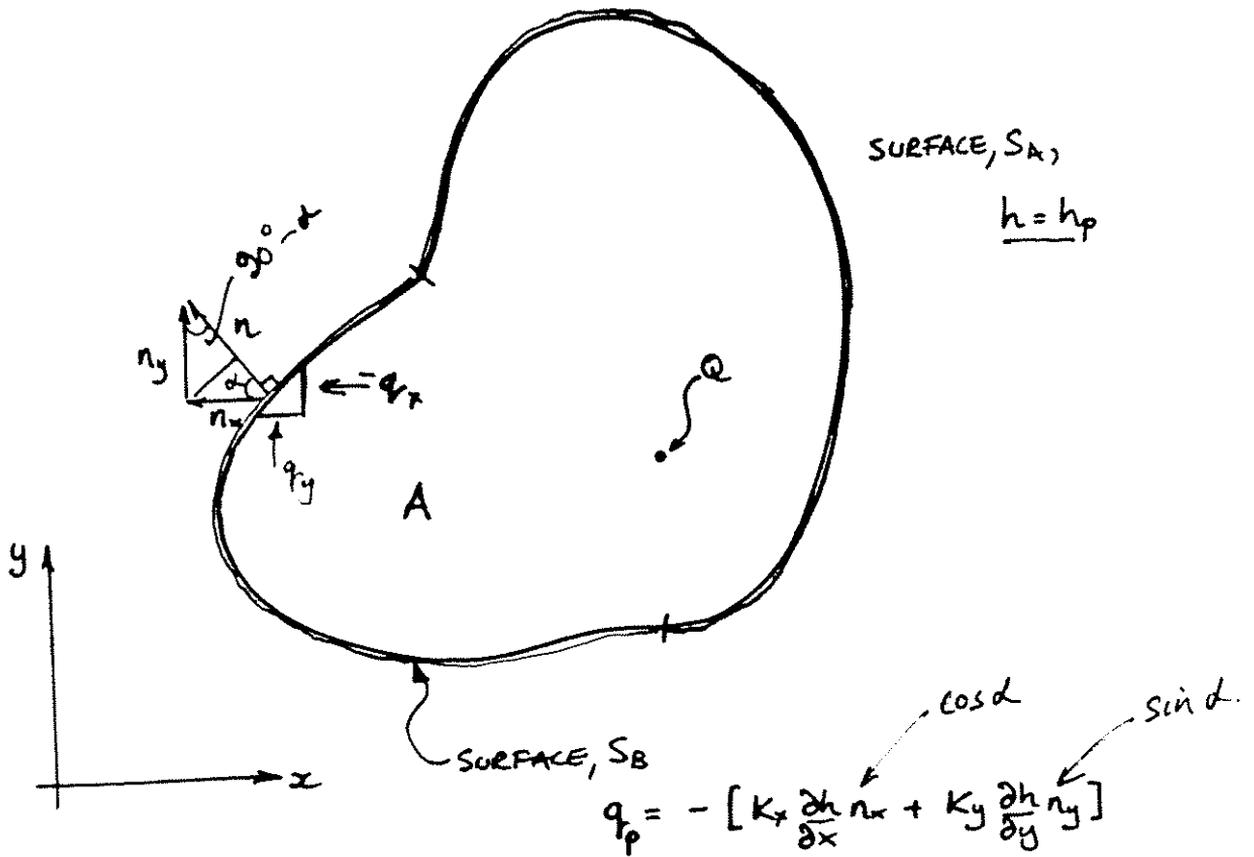
$$\beta\, dP/dt = k/\mu\, \nabla^2 P$$

Figure 2.3.1.1 Solution domain for physical problem

GALERKIN MODEL FOR FLOW

- No simple model available for general diffusion equation that is analogous to principle of virtual work. Use Galerkin approximation.

Sequence of steps

Equation
$$\frac{\partial}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) + Q = S_s \frac{\partial h}{\partial t} \qquad (1)$$

Boundary conditions

1. S_A; $\qquad h = h_P \qquad\qquad$ (2)

2. S_B; $\qquad q_P = -\left[K_x \frac{\partial h}{\partial x} n_x + K_y \frac{\partial h}{\partial y} n_y\right]$ (3)

Include boundary conditions in (1) to rewrite, weighted by a weighting function, W, as

$$\int_A W \left[\frac{\partial}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) + Q - S_s \frac{\partial h}{\partial t}\right] dx\, dy$$

$$- \int_{S_B} W \left[K_x \frac{\partial h}{\partial x} n_x + K_y \frac{\partial h}{\partial y} n_y + q_P\right] dS_B = 0 \qquad (4)$$

and W is an arbitrary scalar. If for any magnitude of W the differential equation is not satisfied then equation (4) cannot be true.

Wish to reduce the second order pde by one order. Use Green's theorem as:

$$\int_A w \frac{\partial}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right) dx\, dy = -\int_A \frac{\partial w}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right) dx\, dy$$

$$+ \int_S w\left(K_x \frac{\partial h}{\partial x}\right) n_x\, dS \qquad\qquad (5)$$

← volume integral

← surface integral

Similarly for $\int_A w \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) dx\, dy = \dots\dots$

Resubstituting (5) into (4) gives

$$-\int_A \left[\frac{\partial w}{\partial x} K_x \frac{\partial h}{\partial x} + \frac{\partial w}{\partial y} K_y \frac{\partial h}{\partial y} - wQ + wS_s \frac{\partial h}{\partial t}\right] dx\, dy$$

$$+ \int_S w\left[K_x \frac{\partial h}{\partial x} n_x + K_y \frac{\partial h}{\partial y} n_y\right] dS - \int_{S_B} w\left[K_x \frac{\partial h}{\partial x} n_x + K_y \frac{\partial h}{\partial y} n_y - q_p\right] dS_B$$

$$= 0 \qquad\qquad (6)$$

Rearrange terms of surface integral since $\int_S dS - \int_{S_B} dS_B = \int_{S_A} dS_A$

$$\underbrace{\int_A \left[\frac{\partial w}{\partial x} K_x \frac{\partial h}{\partial x} + \frac{\partial w}{\partial y} K_y \frac{\partial h}{\partial y}\right] dx\, dy}_{①} \qquad \underbrace{- \int_A wQ\, dx\, dy}_{②}$$

$$+ \int_A wS_s \frac{\partial h}{\partial t} dx\, dy$$

$$\underbrace{- \int_{S_A} w\left[K_x \frac{\partial h}{\partial x} n_x + K_y \frac{\partial h}{\partial y} n_y\right] dS_A}_{③} \quad \underbrace{- \int_{S_B} w\, q_p\, dS_B}_{④} = 0 \qquad (7)$$

Note: ① is first order in $\partial h/\partial x$ instead of $\partial^2 h/\partial x^2$

② Volumetric flux

③ Boundary flux defined on S_A only

④ Prescribed nodal flux conditions on S_B
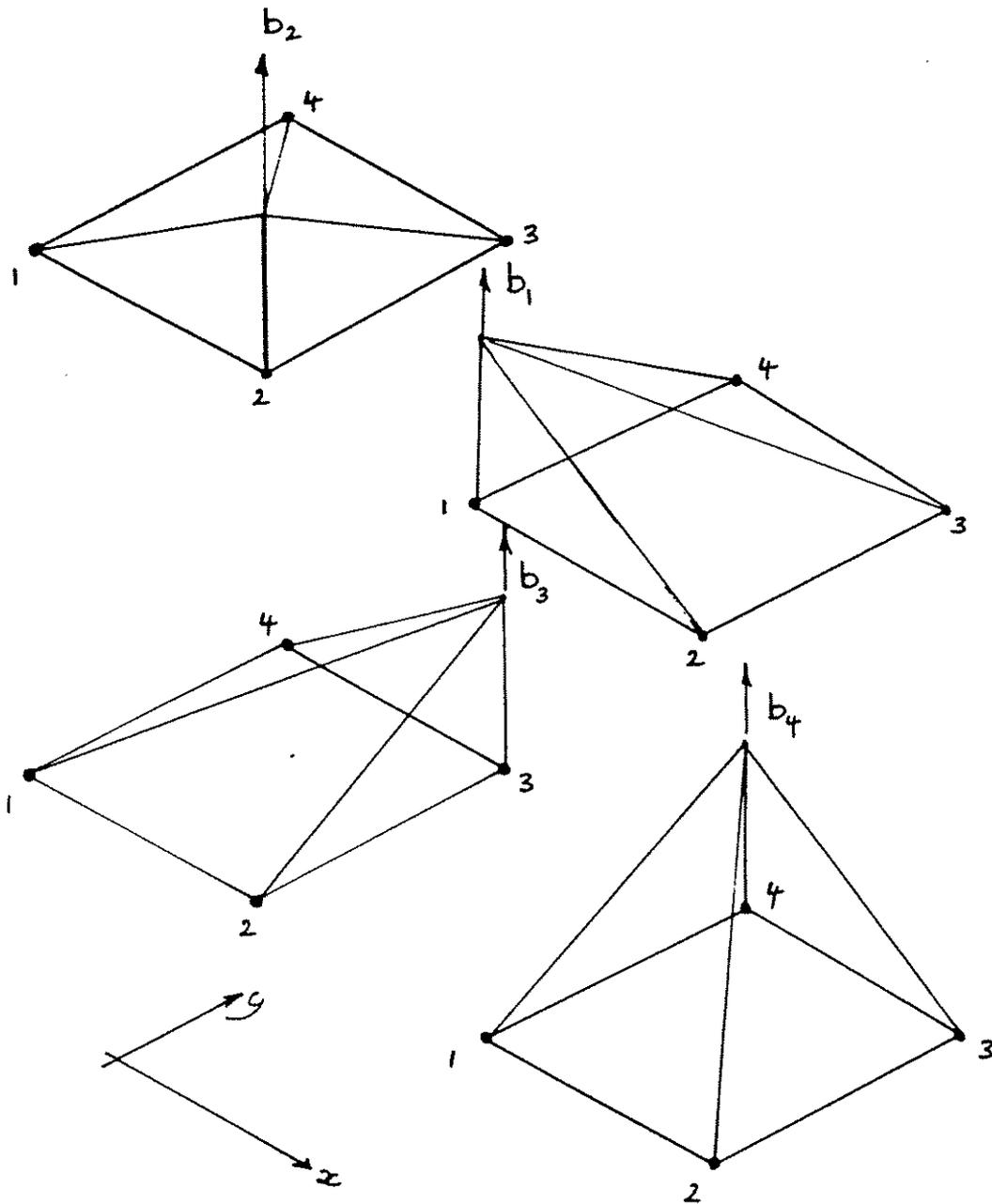
$$h = \sum_{i=1}^{4} b_i h_i = b_1 h_1 + b_2 h_2 + b_3 h_3 + b_4 h_4 = \underline{b}\,\underline{h}$$

$$\underline{b} = [\,b_1\,;\,b_2\,;\,b_3\,;\,b_4\,] \qquad \underline{h}^T = [\,h_1\,;\,h_2\,;\,h_3\,;\,h_4\,]$$

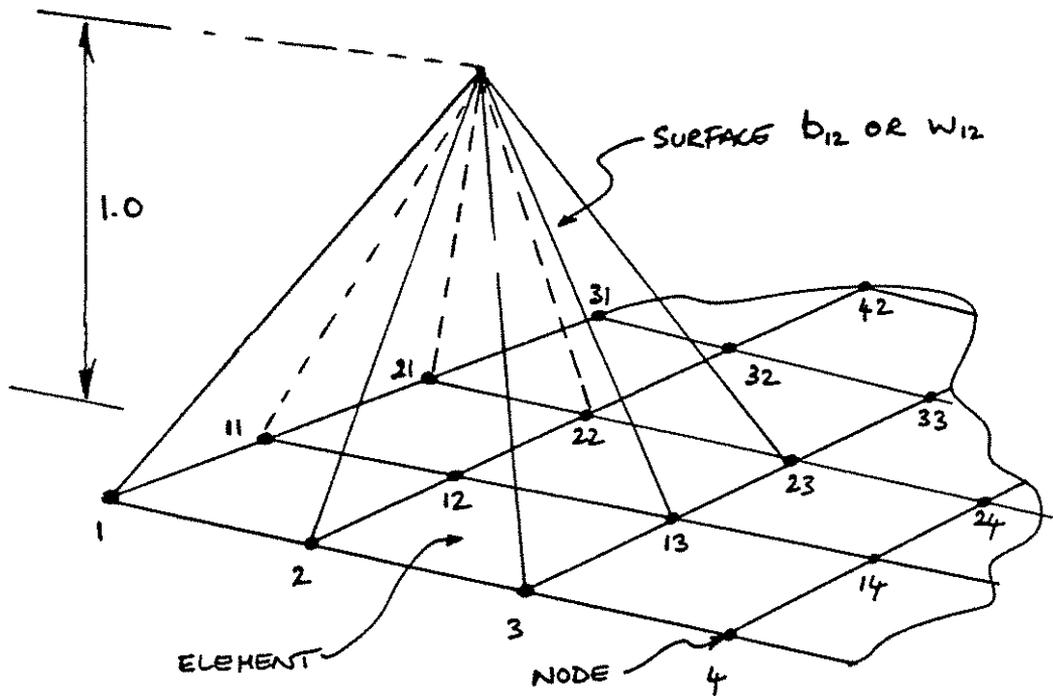Figure 2.3.1.3   Shape functions within a single four noded element

Figure 2.3.1.2 Form of global shape functions.
Note for node 12 the shape or basis function has
a magnitude of unity at node 12 and zero at all
other nodes in the grid.

In FE equation, we must define the form of variation over the element. Using shape function,

$$h = \sum_1^n b_i h_i$$

$$Q = \sum_1^n b_i Q_i \qquad\qquad (8)$$

$$q_p = \sum_1^n b_i q_{p_i}$$

Also define form of weighting, w, as

$$w_i = b_i \qquad\qquad (9)$$

as defined by Galerkin method.

Substitute (8) and (9) into (7) with summation, to give

$$\sum_{i=1}^n \sum_{j=1}^n \int_A \left[\frac{\partial b_i}{\partial x} K_x \frac{\partial b_j}{\partial x} + \frac{\partial b_i}{\partial y} K_y \frac{\partial b_j}{\partial y} \right] h_j \, dx \, dy$$

$$- \sum_{i=1}^n \sum_{j=1}^n \int_A b_i b_j \, dx \, dy \, Q_j$$

$$+ \sum_{i=1}^n \sum_{j=1}^n \int_A b_i S_s b_j \, dx \, dy \, \frac{\partial}{\partial t} h_j$$

$$- \sum_{i=1}^n \sum_{j=1}^n \int_{S_A} b_i \left[K_x \frac{\partial b_j}{\partial x} n_x + K_y \frac{\partial b_j}{\partial y} n_y \right] \partial S_A \, h_j$$

$$- \sum_{i=1}^n \sum_{j=1}^n \int_{S_B} b_i b_j \, dS_B \, q_{p_j} = 0 \qquad\qquad (10)$$

Note that from (3) $\quad q_p = -\left[K_x \frac{\partial h}{\partial x} n_x + K_y \frac{\partial h}{\partial y} n_y \right]$

$\therefore$ naturally satisfied on S_A since h is defined on S_A.

Writing in matrix form:

$$\int_A \underline{a}^T \underline{D} \, \underline{a} \, dx \, dy \, \underline{h} \quad - \quad \int_A \underline{b}^T \underline{b} \, dx \, dy \, \underline{Q}$$

$$+ \quad S_s \int_A \underline{b}^T \underline{b} \, dx \, dy \, \frac{\partial}{\partial t} \, \underline{h} \quad = \quad \int_{S_B} \underline{b}^T \underline{b} \, dS_B \, \underline{q}_P \qquad (11)$$

$$\underbrace{\phantom{\int_{S_B} \underline{b}^T \underline{b} \, dS_B \, \underline{q}_P}}_{\underline{q}_P}$$

and functions are prescribed

$$\underline{h}^T = [\, h_1, \, h_2 \, \cdots \, h_n \,]$$

$$\underline{b} = [\, b_1 \, ; \, b_2 \, ; \qquad b_n \,]$$

$$\underline{a} = \left\{ \begin{matrix} \partial/\partial x \\ \partial/\partial y \end{matrix} \right\} \underline{b}$$

$$\left. \begin{matrix} \\ \\ \\ \end{matrix} \right\}$$

$$h = \underline{b} \, \underline{h}$$

$$h_{,} = \frac{\partial}{\partial x}(\underline{b} \, \underline{h}) = \frac{\partial}{\partial x}(\underline{b}) \, \underline{h}$$

Solve (11) for various elements.

$$\underline{q} = \underline{D} \, \underline{h}_{,}$$

<u>**Summary of Notation – Diffusion Equation for Darcy Flow**</u>

Tensor:

$$A\frac{\partial p}{\partial t} + \nabla\cdot(-D\nabla p) = 0 \quad \text{with} \quad \nabla = \left\{ \begin{array}{c} \partial\!\!\big/\!\!\partial x \\ \partial\!\!\big/\!\!\partial y \\ \partial\!\!\big/\!\!\partial z \end{array} \right\} \quad and \quad \nabla\cdot\nabla = \partial^2\!\!\big/\!\!\partial x^2 + \partial^2\!\!\big/\!\!\partial y^2 + \partial^2\!\!\big/\!\!\partial z^2 \tag{1}$$

Matrix:

$$A\dot{p} - \underline{\nabla}^T D\,\underline{\nabla} p = 0 \quad \text{with} \quad \underline{\nabla} = \left\{ \begin{array}{c} \partial\!\!\big/\!\!\partial x \\ \partial\!\!\big/\!\!\partial y \\ \partial\!\!\big/\!\!\partial z \end{array} \right\}, \quad and \quad \nabla\cdot\nabla = \underline{\nabla}^T\underline{\nabla} = \nabla^2 \tag{2}$$

Parameters:

$$A = \beta \text{ (reservoir compressibility or storage); } D = \frac{k}{\mu} \text{ (permeability/dynamic viscosity)} \tag{3}$$

<u>**Finite Element Statement**</u>

Galerkin – Pre-weight by $\underline{b}^T$ and integrate over the volume of the domain:

$$\int_V \underline{b}^T[A\dot{p} - \underline{\nabla}^T D\,\underline{\nabla} p = 0]\,dV \tag{4}$$

Note that we can define pressures at a point, p, and pressure gradients, $\underline{\nabla} p$ or $\underline{p}$, in terms of nodal pressures, $\underline{p}$, as,

$$p = \underline{b}\,\underline{p} \tag{5}$$
$$p, = \underline{\nabla} p = \underline{\nabla}\,\underline{b}\,\underline{p} = \underline{a}\,\underline{p} \tag{6}$$

Substituting the nodal pressures of equation (5) and the gradient of pressure of equation (6) into equation (4) yields

$$\int_V \underline{b}^T[A\,\underline{b}\,\dot{\underline{p}} - \underline{\nabla}^T D\,\underline{a}\,\underline{p} = 0]\,dV \tag{7}$$

And noting the standard result for transposed matrices that $\underline{b}^T\,\underline{\nabla}^T = [\underline{\nabla}\,\underline{b}]^T = \underline{a}^T$ yields on substitution into equation (7).

$$\int_V [\underline{b}^T A\,\underline{b}\,\dot{\underline{p}} - \underline{b}^T\underline{\nabla}^T D\,\underline{a}\,\underline{p} = 0]\,dV \tag{8}$$

that results in

$$\int_V [\underbrace{\underline{b}^T A\,\underline{b}}_{\underline{S}}\,\dot{\underline{p}} - \underbrace{\underline{a}^T D\,\underline{a}}_{\underline{K}_d}\,\underline{p} = 0]\,dV \tag{9}$$

Yields

$$\underline{S}\dot{\underline{p}} + \underline{K}_d\underline{p} = \underline{q} \tag{10}$$

I-D ELEMENT $\underline{K} = \int_V \underline{a}^T \underline{D} \, \underline{a} \, dV = A \int_V \underline{a}^T \underline{D} \underline{a} \, dx$

$K_I = 10^{-6} \, m/s$

$K_{II} = 2 \times 10^6 \, m/s$

(a) GLOBAL

(b) LOCAL

$h = h_1 + \frac{x}{L}(h_2 - h_1) = $ SHAPE FUNCTION

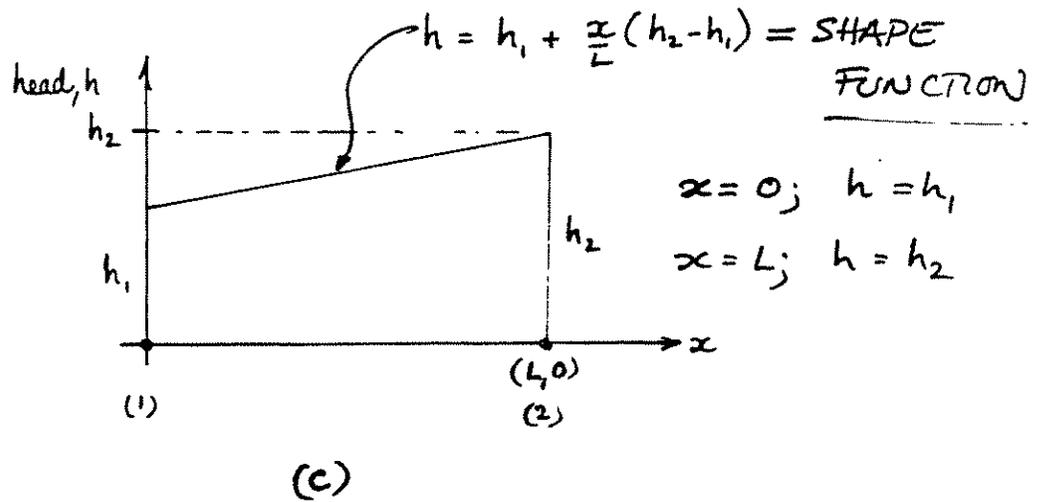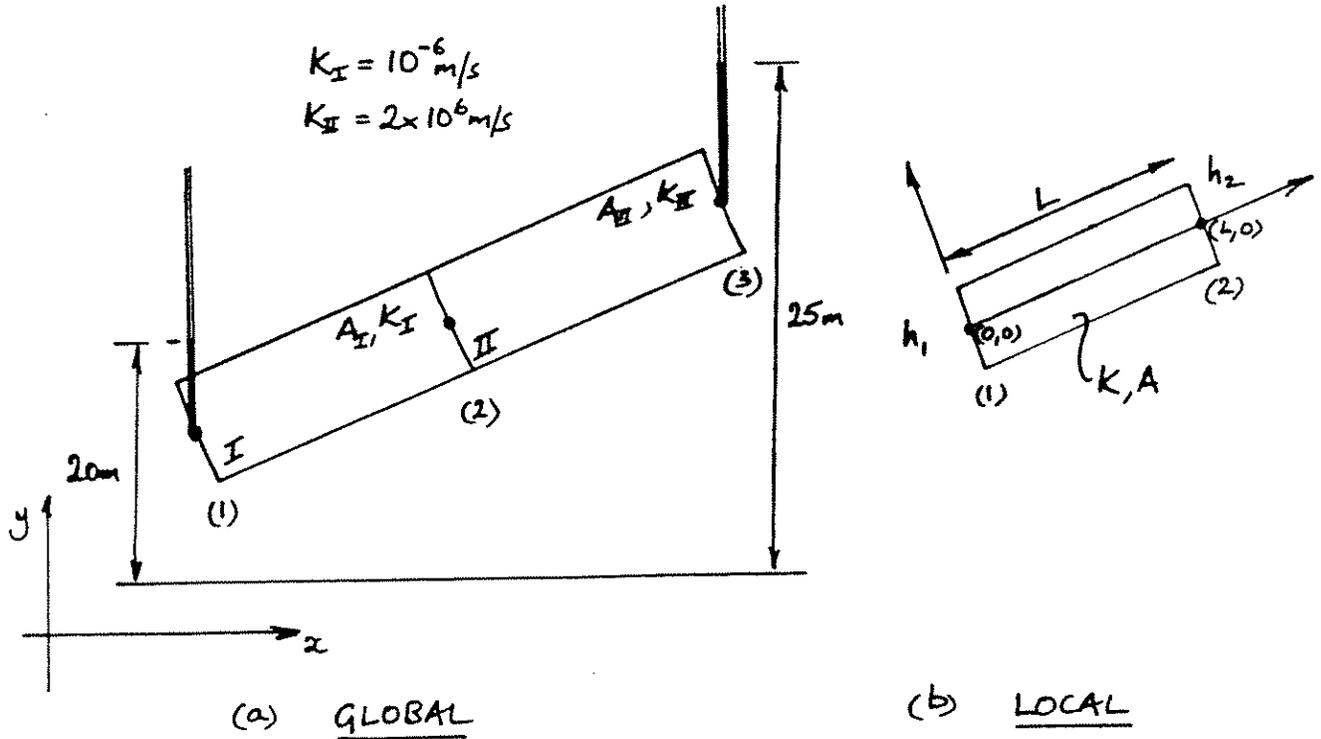$x = 0; \quad h = h_1$

$x = L; \quad h = h_2$

(c)

Figure 2.4.1.1 Global and local descriptions of
elements together with assumed linear head distribution

1-D ELEMENT (FLOWS)

$$K = \int_V \underline{a}^T \underline{D}\underline{a}\, dV = A \int_0^L \underline{a}^T \underline{D}\underline{a}\, dx$$

$\underline{D}$ Matrix : $\underline{D} = [K_x]$

$\underline{a}$ Matrix : $\underline{a} = \dfrac{\partial}{\partial x}\, \underline{b}$ since $h_, = \underline{a}\, \underline{h}$

or $h_, = \dfrac{\partial h}{\partial x} = \dfrac{\partial}{\partial x}(\underline{b}\underline{h}) = \dfrac{\partial}{\partial x}\underline{b}\underline{h}$

Choose linear distribution of h, as : $h = h_, + (h_2 - h_,)\dfrac{x}{L}$

or $h = [\,(1-\dfrac{x}{L})\; ;\; \dfrac{x}{L}\,]\begin{Bmatrix} h_1 \\ h_2 \end{Bmatrix}$

$$h = \qquad \underbrace{\phantom{[\,(1-\tfrac{x}{L})\; ;\; \tfrac{x}{L}\,]}}_{\underline{b}} \qquad \underline{h}$$

Then $\underline{a} = \dfrac{\partial}{\partial x}\, \underline{b}\; ;$ $\underline{a} = \dfrac{1}{L}[-1\; ;\; 1]$

Final $\underline{K}$ matrix as : $\underline{K} = \int_V \underline{a}^T \underline{D}\underline{a}\, dV$

$$\underline{K} = A\int_0^L \dfrac{1}{L}\begin{Bmatrix} -1 \\ 1 \end{Bmatrix} K_x\, \dfrac{1}{L}[-1\; ;\; 1]\, dx$$

$$\underline{K} = \dfrac{AK_x}{L^2}\int_0^L \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} dx = \pm\dfrac{AK_x}{L^2}[x]_0^L = \dfrac{AK_x}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

NOTE: SAME RESULT AS SIMPLE INTUITIVE CONSTRUCTION !!

[2:3] Fluid Flow and Pressure Diffusion

Recap

2D Triangular (Constant Gradient) Elements
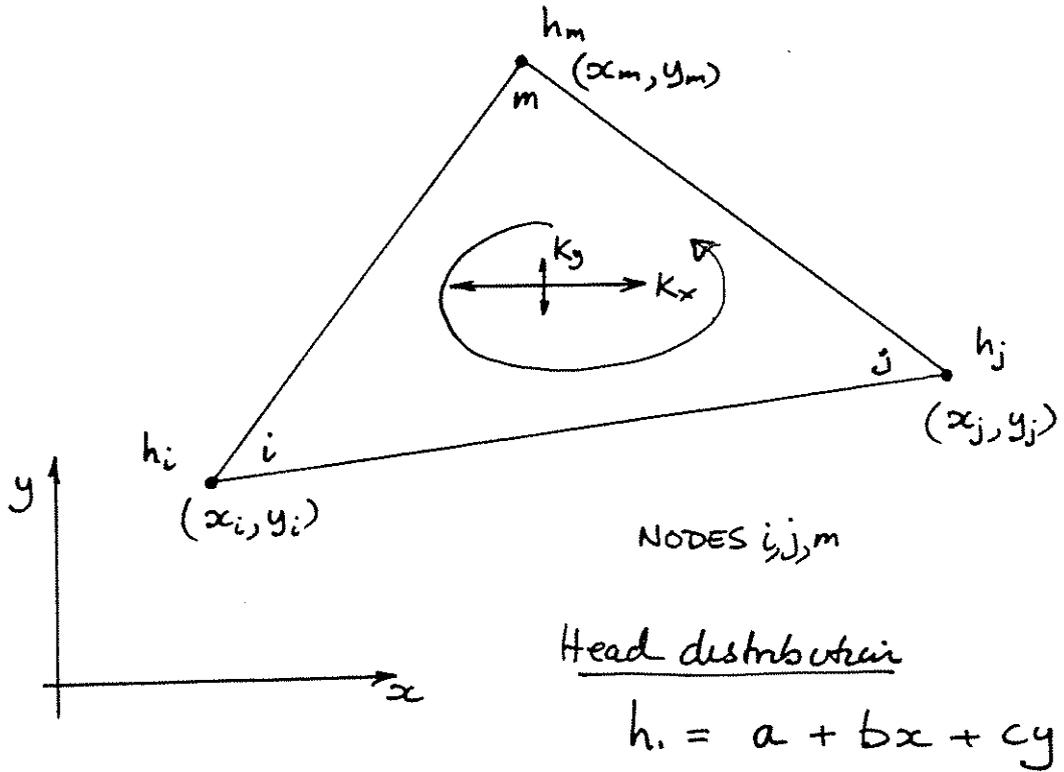
Derivation

Example

EGEEfem

Figure 2.4.2.1 Geometry of a triangular element

TRIANGULAR ELEMENT

$$\underline{K} = \int_v \underline{a}^T \underline{D} \underline{a} \, dv \quad (1)$$

$$\underline{v} = \underline{D} \underline{h},$$

$$\underline{D} = \begin{bmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy} \end{bmatrix} \cdot \qquad K_{xy} = K_{yx} \qquad or \qquad \begin{Bmatrix} v_x \\ v_y \end{Bmatrix} = \begin{bmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy} \end{bmatrix} \begin{Bmatrix} \partial h/\partial x \\ \partial h/\partial y \end{Bmatrix}$$

'a' Matrix $\qquad \underline{h}_j = \underline{a}\underline{h} \qquad or \qquad \begin{Bmatrix} \partial h/\partial x \\ \partial h/\partial y \end{Bmatrix} = \underline{a}\underline{h} \qquad (2)$

Choose shape functions: $\qquad h = a + bx + cy \qquad (3)$

$$\therefore \quad \frac{\partial h}{\partial x} = b \qquad \left.\begin{array}{c} \\ \\ \end{array}\right\} (4)$$
$$\frac{\partial h}{\partial y} = c$$

The head magnitudes are defined as h_i, h_j, h_m at the nodes.
$\therefore$ 3 equations may be determined as:

$$\left.\begin{array}{l} h_i = a + bx_i + cy_i \\ h_j = a + bx_j + cy_j \\ h_m = a + bx_m + cy_m \end{array}\right\} (5)$$

Writing in matrix form; the equations (5) may be presented as

$$\begin{Bmatrix} h_i \\ h_j \\ h_m \end{Bmatrix} = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_m & y_m \end{bmatrix} \begin{Bmatrix} a \\ b \\ c \end{Bmatrix} \qquad (6)$$

Inverting (6) to give the coefficients (b) and (c) of equation (4) yields:

$$\begin{Bmatrix} a \\ b \\ c \end{Bmatrix} = \frac{1}{2\Delta} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{Bmatrix} h_i \\ h_j \\ h_m \end{Bmatrix} \qquad (7)$$

2Δ = the determinant of the equation $\qquad (2\Delta = A_{11} + x_i A_{21} + y_i A_{31})$

Δ = area of triangle.

Returning to (2) and (4) and using (7), then

$$\begin{Bmatrix} \partial h/\partial x \\ \partial h/\partial y \end{Bmatrix} = \begin{Bmatrix} b \\ c \end{Bmatrix} = \frac{1}{2\Delta} \underbrace{\begin{bmatrix} A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{Bmatrix} h_i \\ h_j \\ h_m \end{Bmatrix}}_{\underline{a}} \qquad (8)$$

$\underline{K} = \int_V \underline{a}^T \underline{D} \underline{a} \, dV$

Since $\underline{a}$ and $\underline{D}$ are constant over the element, they are removed from the integral.

$$\underline{K} = \int_V \underline{a}^T \underline{D} \underline{a} \, dV = \underline{a}^T \underline{D} \underline{a} \int_V dV = \underline{a}^T \underline{D} \underline{a} \, \Delta \, \text{thickness}$$

where Δ is the scalar magnitude of element area.

STANDARD RESULT

$$\begin{Bmatrix} h_1 \\ h_2 \\ h_3 \end{Bmatrix} = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_m & y_m \end{bmatrix} \begin{Bmatrix} a \\ b \\ c \end{Bmatrix}$$

$$\underbrace{\phantom{\begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_m & y_m \end{bmatrix}}}_{B}$$

$$\begin{Bmatrix} a \\ b \\ c \end{Bmatrix} = \frac{1}{|B|} \begin{bmatrix} (x_j y_m - x_m y_j) & (x_m y_i - x_i y_m) & (x_i y_j - x_j y_i) \\ (y_j - y_m) & (y_m - y_i) & (y_i - y_j) \\ (x_m - x_j) & (x_i - x_m) & (x_j - x_i) \end{bmatrix} \begin{Bmatrix} h_1 \\ h_2 \\ h_3 \end{Bmatrix}$$

$$|B| = 2\Delta = 1 . \begin{vmatrix} x_j & y_j \\ x_m & y_m \end{vmatrix} - x_i . \begin{vmatrix} 1 & y_j \\ 1 & y_m \end{vmatrix} + y_i \begin{vmatrix} 1 & x_j \\ 1 & x_m \end{vmatrix}$$

$$2\Delta = (x_j y_m - x_m y_j) - x_i (y_m - y_j) + y_i (x_m - x_j)$$

II.5 Inversion (Adjoint Matrix)

It can be shown that

$$a \,(\text{adj } a) = |a|\, I \qquad (\text{II}.11)$$

where $|a|$ is the determinant of the matrix a and adj a, called the *adjoint* matrix, is the transpose of the matrix of cofactors of the determinant. Comparing (II.10) and (II.11) we see that

$$a^{-1} = \frac{\text{adj } a}{|a|} \qquad (\text{II}.12)$$

from which it is clear that the inverse does not exist when $|a|$ is zero, in which case a is said to be *singular*.

To illustrate the method we shall determine the inverse of the matrix

$$H = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_m & y_m \end{bmatrix} \qquad (\text{II}.13)$$

If we delete the pth row and qth column from the determinant of the matrix we obtain the minor H'_{pq}, e.g. deleting row 3 and column 1 we have

$$H'_{31} = \begin{vmatrix} x_i & y_i \\ x_j & y_j \end{vmatrix} \qquad (\text{II}.14)$$

The cofactor $\bar{H}_{pq}$ is the product of the minor and $(-1)^{(p+q)}$. When the cofactors are written as a matrix and then transposed we have the adjoint matrix
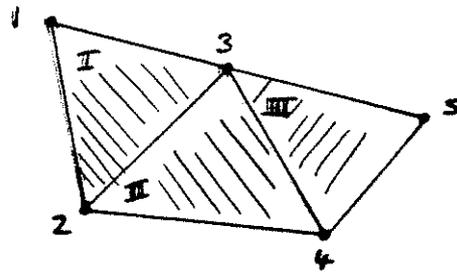
$$\text{adj } H = \begin{bmatrix} \begin{vmatrix} x_j & y_j \\ x_m & y_m \end{vmatrix} & -\begin{vmatrix} x_i & y_i \\ x_m & y_m \end{vmatrix} & \begin{vmatrix} x_i & y_i \\ x_j & y_j \end{vmatrix} \\ -\begin{vmatrix} 1 & y_j \\ 1 & y_m \end{vmatrix} & \begin{vmatrix} 1 & y_i \\ 1 & y_m \end{vmatrix} & -\begin{vmatrix} 1 & y_i \\ 1 & y_j \end{vmatrix} \\ \begin{vmatrix} 1 & x_j \\ 1 & x_m \end{vmatrix} & -\begin{vmatrix} 1 & x_i \\ 1 & x_m \end{vmatrix} & \begin{vmatrix} 1 & x_i \\ 1 & x_j \end{vmatrix} \end{bmatrix} \qquad (\text{II}.15)$$

For example H'_{31} of (II.14) is transposed to row 1 column 3. Expanding the

determinants we have

$$\text{adj } H = \begin{bmatrix} (x_j y_m - x_m y_j) & -(x_i y_m - x_m y_i) & (x_i y_j - x_j y_i) \\ -(y_m - y_j) & (y_m - y_i) & -(y_j - y_i) \\ (x_m - x_j) & -(x_m - x_i) & (x_j - x_i) \end{bmatrix} \qquad (\text{II}.16)$$

The inverse is obtained by dividing adj H by the determinant of H.

LOCAL ELEMENT MATRICES $\quad \underline{K}\,\underline{h} = \underline{q}$

$$\begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{21} & I_{32} & I_{33} \end{bmatrix} \begin{Bmatrix} h_1 \\ h_2 \\ h_3 \end{Bmatrix} = \begin{Bmatrix} q_1 \\ q_2 \\ q_3 \end{Bmatrix} \qquad \text{ELEMENT } I$$

$$\begin{matrix} & 2 & 4 & 3 & & & \\ 2 \\ 4 \\ 3 \end{matrix} \begin{bmatrix} II_{22} & II_{24} & II_{23} \\ II_{42} & II_{44} & II_{43} \\ II_{32} & II_{34} & II_{33} \end{bmatrix} \begin{Bmatrix} h_2 \\ h_4 \\ h_3 \end{Bmatrix} = \begin{Bmatrix} q_2 \\ q_4 \\ q_3 \end{Bmatrix} \qquad \text{ELEMENT } II$$

$$\begin{bmatrix} III_{44} & III_{45} & III_{43} \\ III_{54} & III_{55} & III_{53} \\ III_{34} & III_{35} & III_{33} \end{bmatrix} \begin{Bmatrix} h_4 \\ h_5 \\ h_3 \end{Bmatrix} = \begin{Bmatrix} q_4 \\ q_5 \\ q_3 \end{Bmatrix} \qquad \text{ELEMENT } III$$

NOTE : SUBSCRIPTS REPRESENT GLOBAL NODE NUMBERS

Figure 2.4.2.2 Local element conductance matrices
for a three element system

GLOBAL SYSTEM MATRIX $\underline{K}\,\underline{h} = \underline{q}$
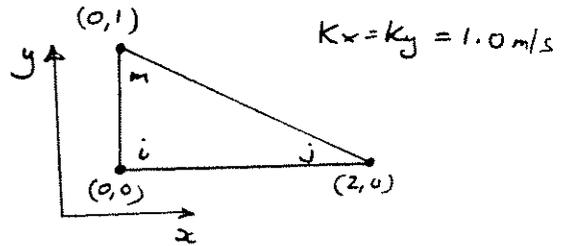
INDIVIDUAL TERMS ARE ADDED FROM THE LOCAL ELEMENT MATRICES.

Figure 2.4.2.3 Global assembly for three element example

(0,1)

$K_x = K_y = 1.0$ m/s

(0,0)   (2,0)

Example 2.4.2.1

    Evaluate the element conductance matrix for the triangular element shown.

From equation (2.4.2.9)

$$\underline{a} = \frac{1}{2\Delta}\begin{bmatrix} A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \frac{1}{2\Delta}\begin{bmatrix} -1 & 1 & 0 \\ -2 & 0 & 2 \end{bmatrix}$$

$$\underline{D} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

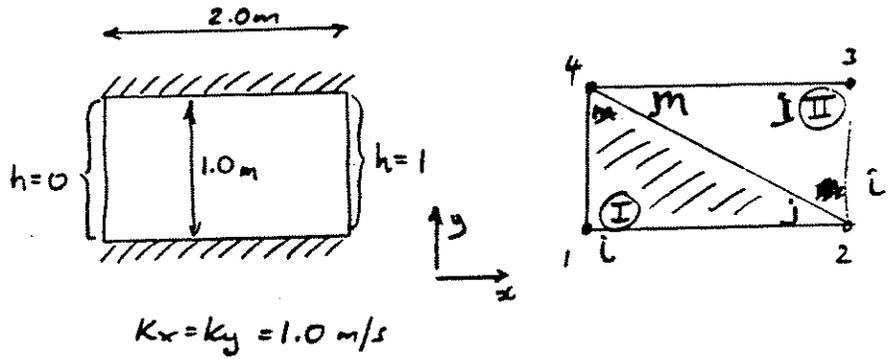$$\underline{K} = \underline{a}^T \underline{D}\, \underline{a}\, \Delta = \frac{1}{2\Delta}\begin{bmatrix} -1 & -2 \\ 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\frac{1}{2\Delta}\begin{bmatrix} -1 & 1 & 0 \\ -2 & 0 & 2 \end{bmatrix}\Delta$$

$$\underline{K} = \frac{1}{4\Delta}\begin{bmatrix} 5 & -1 & -4 \\ -1 & 1 & 0 \\ -4 & 0 & 4 \end{bmatrix} \qquad \text{where } \Delta = 1.0$$

Example 2.6.2.2

... the system shown,
evaluate flow rates using
the finite element method.

$K_x = k_y = 1.0$ m/s

Element Conductance Matrices $\underline{K}\, \underline{h} = \underline{q}$

$$\underline{K}_I = \frac{1}{4}\begin{bmatrix} 5 & -1 & \boxed{-4} \\ -1 & 1 & 0 \\ -4 & 0 & 4 \end{bmatrix}\begin{matrix}1\\2\\4\end{matrix} \quad ; \quad \underline{K}_{II} = \frac{1}{4}\begin{bmatrix} 4 & -4 & 0 \\ \boxed{-4} & 5 & -1 \\ 0 & -1 & 1 \end{bmatrix}\begin{matrix}2\\3\\4\end{matrix}$$

with column labels $\begin{matrix}1&2&4\end{matrix}$ for $\underline{K}_I$ and $\begin{matrix}2&3&4\end{matrix}$ for $\underline{K}_{II}$.

Note: $\underline{K}_{II}$ may be derived directly from $\underline{K}_I$ in this particular case since they possess identical geometries. Element II is merely rotated. The permutation $4 \to 2$; $1 \to 3$ and $2 \to 4$ may be used to exchange locations as, for example, $K_{I_{14}} = K_{II_{32}}$ as illustrated.

Global Conductance Matrix $\underline{K}\, \underline{h} = \underline{q}$

$$\frac{1}{4}\begin{bmatrix} 5 & -1 & & -4 \\ -1 & (1+4) & -4 & (0+0) \\ & -4 & 5 & -1 \\ -4 & (0+0) & -1 & (4+1) \end{bmatrix}\begin{Bmatrix} h_1 = 0 \\ h_2 = 1.0 \\ h_3 = 1.0 \\ h_4 = 0 \end{Bmatrix} = \begin{Bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{Bmatrix} = \frac{1}{4}\begin{Bmatrix} -1 \\ +1 \\ +1 \\ -1 \end{Bmatrix}$$

Note:
Discharge at left side $= q_1 + q_4 = -\frac{1}{2}$ m³/s
Discharge at right side $= q_2 + q_3 = +\frac{1}{2}$ m³/s

Total accumulation $\underline{ 0}$ m³/s

Check with D'arcy's law $q = AK\dfrac{dh}{d\ell} = (1.0)(1.0)\dfrac{1.0}{2.0} = \frac{1}{2}$ m³/s

Q.E.D.

[2:4] Fluid Flow and Pressure Diffusion

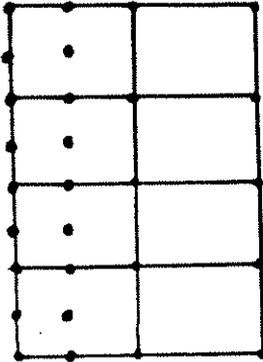Recap – with EGEEfem

2D Isoparametric Elements

Concept

1D example

Numerical integration

$$V = -K \frac{\partial h}{\partial x}$$
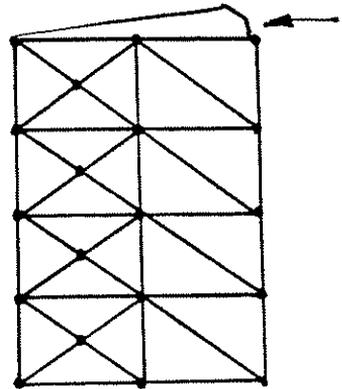
ISOPARAMETRIC ELEMENTS   TRIANGULAR ELEMENTS

Figure 2.4.3.2 Mesh scaling in areas of high hydraulic gradients
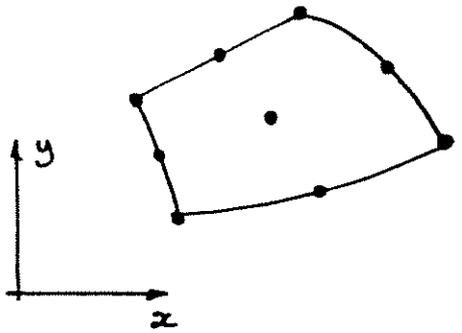
GLOBAL LOCATION

DESIRE

$$\underline{K} = \int_A \underline{a}^T \underline{D} \, \underline{a} \; \underline{dx \, dy}$$
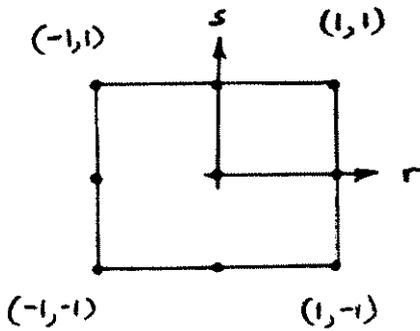
HAPPING

EVALUATE $\underline{K}$ MATRIX

$$\underline{K} = \int_{-1}^{+1} \int_{-1}^{+1} \underline{a}^T \underline{D} \, \underline{a} \, |\underline{J}| \, dr \, ds$$

REVERSE HAPPING
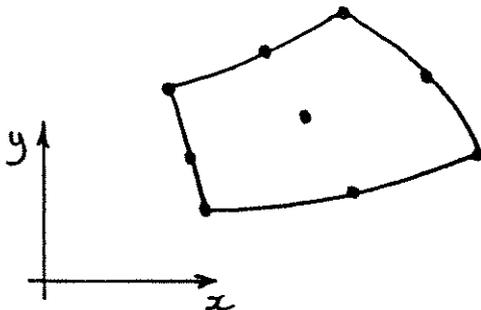
REMAP TO ORIGINAL

AND OBTAIN

$$\underline{K} = \int_A \underline{a}^T \underline{D} \, \underline{a} \, dx \, dy$$

Figure 2.4.3.3 Isoparametric concept

ISOPARAMETRIC
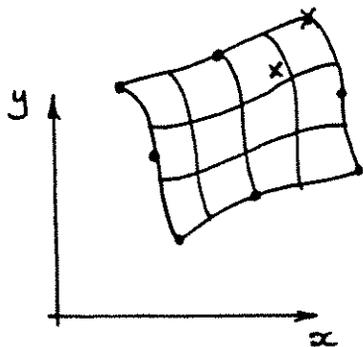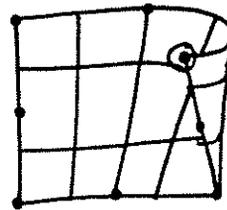
Figure 2.4.3.4 Concept of mapping through the transformation $|\underline{J}|$

GLOBAL   LOCAL

$$dx \, dy = |\underline{J}| \, dr \, ds$$

Figure 2.4.3.5 Mapping must be isoparametric, representing
no folding in the mapping process

FOR 1-D ELEMENT
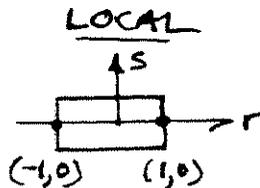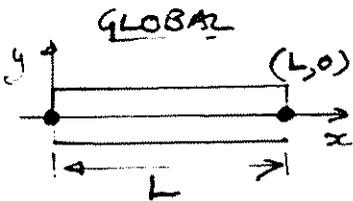
$$dx = |\underline{J}| \, dr$$

$$\frac{dx}{dr} = |\underline{J}| = \frac{L}{2}$$

Example 2.4.3.1 One-Dimensional Isoparametric Element

GLOBAL
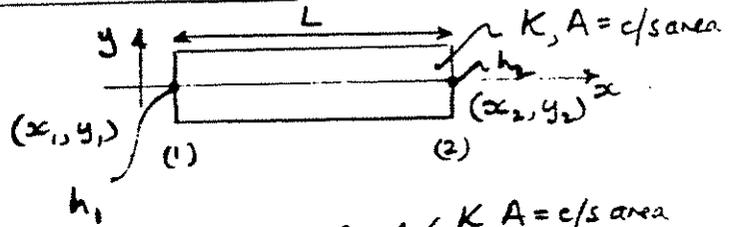
LOCAL
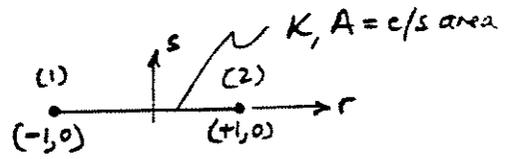
Global $\qquad \underline{K} = A \int_O^L \underline{a}^T \underline{D} \, \underline{a} \, dx \, dy$ $\qquad$ (1)

Transformed $\qquad \underline{K} = A \int_{-1}^{+1} \underline{a}^T \underline{D} \, \underline{a} \, |\underline{J}| \, dr$ $\qquad$ (2)

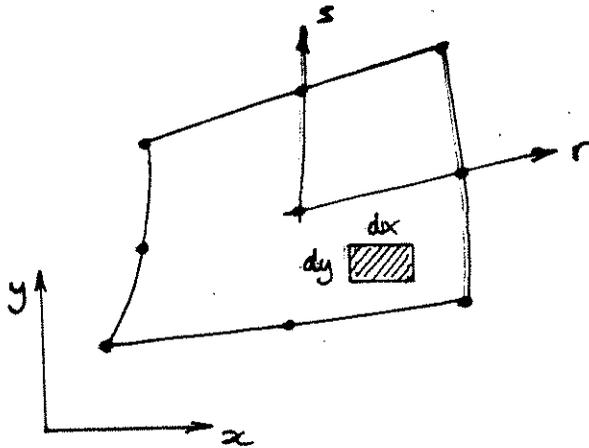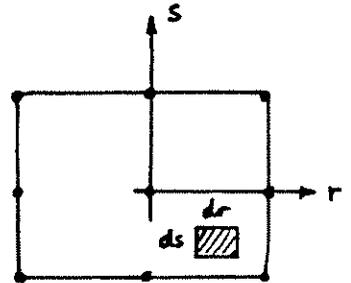'$\underline{D}$' Matrix $\qquad \underline{v} = \underline{D} \frac{\partial h}{\partial x}$ $\quad ; \quad \underline{D} = K$ $\qquad$ (3)

Shape Functions $\qquad \underline{b} = \frac{1}{2} [(1-r) \; ; \; (1+r)]$ $\qquad$ (4)

$$x = \underline{b} \, \underline{x} \qquad \text{for} \quad \underline{x}^T = [x_1 \; ; \; x_2] \qquad \left.\vphantom{\begin{array}{c} a \\ b \end{array}}\right\} (5)$$
$$h = \underline{b} \, \underline{h} \qquad \qquad \underline{h}^T = [h_1 \; ; \; h_2]$$

Note : Shape functions give 1.0 at node under consideration
$\qquad \qquad \qquad 0.0$ at all other nodes.
$\qquad \qquad$ Linear shape functions.

'$\underline{a}$' Matrix

From equation $(2.4.2.4)$, $\quad \frac{\partial h}{\partial x} = \underline{a} \, \underline{h}$ $\qquad$ (6)

From the chain rule $\qquad \frac{\partial h}{\partial x} = \frac{\partial r}{\partial x} \frac{\partial h}{\partial r}$ $\qquad$ (7)

Taking these components individually. The first is defined using equation (5b) as

$$\frac{\partial h}{\partial r} = \frac{\partial}{\partial r}(\underline{b} \, \underline{h}) \quad = \left[\frac{\partial}{\partial r}(\underline{b})\right] \underline{h} \qquad (8)$$

with $\underline{b}'$ labeled above

and substituting equation (4) for the shape function

$$\frac{\partial h}{\partial r} = \underline{b}' \underline{h} \qquad \underline{b}' = \frac{1}{2} [-1 \; ; \; 1] \qquad (9)$$

where $\underline{b}'$ is the derivative of the shape function matrix as

$$\underline{b}' = \frac{\partial}{\partial r}(b) = \frac{1}{2}[-1 \; ; \; 1] \qquad (10)$$

The magnitude of $\partial r / \partial x$ may similarly be determined from equation (7) as

$$\frac{\partial r}{\partial x} = \left[\frac{\partial x}{\partial r}\right]^{-1} = \left[\frac{\partial}{\partial r}(x)\right]^{-1} \underbrace{}_{x = \underline{b}x} \qquad (11)$$

Substituting equation (5a) into (11) gives

$$\frac{\partial r}{\partial x} = [\underline{b}'\underline{x}]^{-1} \qquad (12)$$

Equation (7) is recovered from backsubstituting equations (9) and (12) such that

$$\frac{\partial h}{\partial x} = \underbrace{[\underline{b}'\underline{x}]^{-1} \underline{b}' h}_{\underline{a}} \qquad (13)$$

and the $\underline{a}$ matrix may be evaluated as

$$\underline{a} = \frac{2}{2(x_2 - x_1)}[-1 \; ; \; 1] \qquad (14)$$

or

$$\underline{a} = \frac{1}{L}[-1 \; ; \; 1] \qquad (15)$$

Noting also that $dx = |\underline{J}| \, dr$

Then $\quad |\underline{J}| = \dfrac{dx}{dr} = \underline{b}'\underline{x} = \dfrac{1}{2}(x_2 - x_1) = \dfrac{L}{2} \qquad (16)$

allowing the conductance matrix $\underline{K}$ to be determined as

$$\underline{K} = A \int_{-1}^{+1} \underline{a}^T \underline{D}\, \underline{a}\, |\underline{J}|\, dr = A \int_{-1}^{+1} \frac{1}{L}\begin{bmatrix} -1 \\ 1 \end{bmatrix} K\, \frac{1}{L}[-1\, ;\, 1]\, \frac{L}{2}\, dr \qquad (17)$$

$$\underline{K} = \frac{AK}{2L} \int_{-1}^{+1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} dr \qquad (18)$$

with $\int_{-1}^{+1} dr = [r]_{-1}^{+1} = 2$

$$\underline{K} = \frac{AK}{L} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad (19)$$

where we note the similarity between the final result of equation (19) with that evaluated by the more direct method.

In this particular example there is no advantage in using the isoparametric technique since the integral of equation (18) is trivial to determine.

[2:5] Fluid Flow and Pressure Diffusion

Recap

Isoparametric Elements

Numerical integration

2D and 3D elements

NUMERICAL INTEGRATION

Evaluate integral as
$$\int_{-1}^{+1} f(r)\, dr \simeq \sum_{j=1}^{N} H_j\, f(a_j)$$

Weight factor ← (arrow pointing to H_j)

Function ← (arrow pointing to $f(a_j)$)

Quadrature gives the <u>exact</u> result if the degree of the function, k, is less than or equal to $2N-1$, where N is the order of integration.

$$f(r) = a_0 + a_1 r^1 + a_2 r^2 \dots + a_k r^k$$

$\therefore$ for $f(r) = r^2$ $\qquad k = 2$ $\quad \therefore$ $\qquad$ <u>$N = 2$ gives exact result</u>

EXAMPLE

$$I = \int_{-1}^{+1} (r^2)\, dr$$

(N=1) $\qquad I = 2 f(0) = 0$

(N=2) $\qquad I = (1) f(+.577\dots) + (1) f(-.577\dots)$
$\qquad\qquad I = \quad .333 \quad + \quad .333 \qquad = .66\dot{}$

(N=3) $\qquad\qquad\qquad\qquad\qquad\qquad\qquad I = .66\dot{}$

$\therefore$ no need to evaluate I by using $N > 2$.

TABLE 8.1
ABSCISSAE AND WEIGHT COEFFICIENTS OF THE GAUSSIAN QUADRATURE FORMULA

$$\int_{-1}^{1} f(x)\ dx = \sum_{j=1}^{n} H_i f(a_j).$$

$\pm a$	H
$n = 1$	
0	2·00000 00000 00000
$n = 2$	
0·57735 02691 89626	1·00000 00000 00000
$n = 3$	
0·77459 66692 41483	0·55555 55555 55556
0·00000 00000 00000	0·88888 88888 88889
$n = 4$	
0·86113 63115 94053	0·34785 48451 37454
0·33998 10435 84856	0·65214 51548 62546
$n = 5$	
0·90617 98459 38664	0·23692 68850 56189
0·53846 93101 05683	0·47862 86704 99366
0·00000 00000 00000	0·56888 88888 88889
$n = 6$	
0·93246 95142 03152	0·17132 44923 79170
0·66120 93864 66265	0·36076 15730 48139
0·23861 91860 83197	0·46791 39345 72691
$n = 7$	
0·94910 79123 42759	0·12948 49661 68870
0·74153 11855 99394	0·27970 53914 89277
0·40584 51513 77397	0·38183 00505 05119
0·00000 00000 00000	0·41795 91836 73469
$n = 8$	
0·96028 98564 97536	0·10122 85362 90376
0·79666 64774 13627	0·22238 10344 53374
0·52553 24099 16329	0·31370 66458 77887
0·18343 46424 95650	0·36268 37633 76362
$n = 9$	
0·96816 02395 07626	0·08127 43883 61574
0·83603 11073 26636	0·18064 81606 94857
0·61337 14327 00590	0·26061 06964 02935
0·32425 34234 03809	0·31234 70770 40003
0·00000 00000 00000	0·33023 93550 01260
$n = 10$	
0·97390 65285 17172	0·06667 13443 08688
0·86506 33666 88985	0·14945 13491 50581
0·67940 95682 99024	0·21908 63625 15982
0·43339 53941 29247	0·26926 67193 09996
0·14887 43389 81631	0·29552 42247 14753

2-D integration: $\int_{-1}^{+1}\int_{-1}^{+1} f(r,s)\, dr\, ds = \sum_{j=1}^{N} \sum_{i=1}^{N} H_i H_j f(r_i, s_j)$

Integration order	Degree of precision	Location of integration points
2 × 2	3	
3 × 3	5	
4 × 4	7	

[1] The location of any integration point in the x–y coordinate system is given by : $x_p = \Sigma h_i(r_p, s_p)x_i$ and $y_p = \Sigma h_i(r_p, s_p)y_i$

Figure 2.4.3.2.2 Integration in two dimensions using quadrature

Element	Reliable integration order
4-node	2 × 2
4-node distorted	2 × 2
8-node	3 × 3
8-node distorted	3 × 3
9-node	3 × 3
9-node distorted	3 × 3

Figure 2.4.3.2.1   Recommended orders of quadrature, n, in evaluating conductance matrices

Influence of Shape Function Form

$$q = Kh$$

After p 107

$Q \frac{1}{12}$

$\leftarrow$ .083

$\leftarrow$ .333

.1667 $\left\{ \begin{array}{l} \leftarrow .083 \\ \leftarrow .083 \end{array} \right.$ h=0

$\leftarrow$ .333

$\leftarrow$ .083

$\Sigma a = 1.0$

$\frac{1}{6} = \int b_1 \, dr$

$\frac{4}{6} = \int b_2 \, dr$

$\frac{1}{6} = \int b_3 \, dr$

$\frac{1}{6} = \int b_3 \, dr$

$\frac{4}{6} = \int b_4 \, dr$

$\frac{1}{6} = \int b_5 \, dr$

h=1.0

$$q = AK \frac{dh}{dr} = 2(1)\frac{(1)}{(2)} = 1.0$$
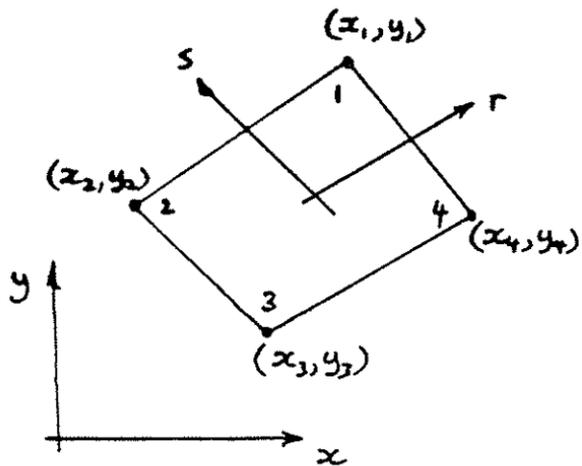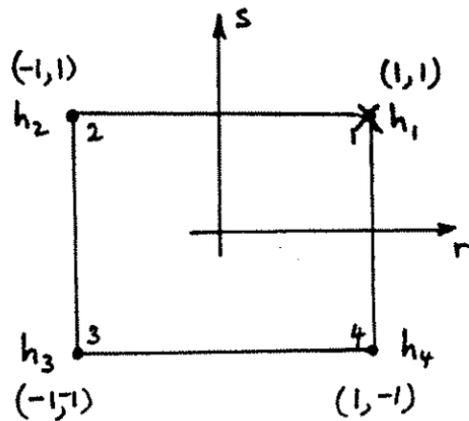
Figure 2.4.3.3.1 Global and local configurations
for a four noded quadrilateral element.

2-D ISOPARAMETRIC ELEMENT

$$\underline{K} = \int_V \underline{a}^T \underline{D} \, \underline{a} \, dx \, dy = \int_{-1}^{+1}\int_{-1}^{+1} \underline{a}^T \underline{D} \underline{a} |J| \, dr \, ds$$

In addition to the integral, we need to evaluate the $\underline{a}$ matrix with respect to local coordinates, r and s. To do this we define the shape functions within the element.

$$\underline{b} = \tfrac{1}{4}[(1+r)(1+s) \; ; \; (1-r)(1+s) \; ; \; (1-r)(1-s) \; ; \; (1+r)(1-s)] \quad (1)$$

Note that these are unity @ nodes under consideration and zero at other nodes.

Vectors of heads may be defined as can be coordinates:

$$\underline{h}^T = [\, h_1 \; ; \; h_2 \; ; \; h_3 \; ; \; h_4 \,]$$
$$\underline{x}^T = [\, x_1 \; ; \; x_2 \; ; \; x_3 \; ; \; x_4 \,] \qquad \Bigg\} \quad (2)$$
$$\underline{y}^T = [\, y_1 \; ; \; y_2 \; ; \; y_3 \; ; \; y_4 \,]$$

Enabling parameters to be defined at any point within the element as:

$$h = \underline{b}\,\underline{h}$$
$$x = \underline{b}\,\underline{x} \qquad \Bigg\} \quad (3)$$
$$y = \underline{b}\,\underline{y}$$

Head gradients in local coordinates are (we need global derivatives in '$\underline{a}$' matrix)

$$\left\{ \begin{array}{c} \dfrac{\partial h}{\partial r} \\[2mm] \dfrac{\partial h}{\partial s} \end{array} \right\} = \underbrace{\left\{ \begin{array}{c} \dfrac{\partial}{\partial r} \\[2mm] \dfrac{\partial}{\partial s} \end{array} \right\} \underline{b}}_{\underline{P}} \; h \qquad (4)$$

The individual derivatives may be defined as (with ref. to (1))

$$\underline{P} = \left\{ \begin{array}{c} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial s} \end{array} \right\} \underline{b} = \frac{1}{4} \left[\begin{array}{cccc} (1+s) & -(1+s) & -(1-s) & (1-s) \\ (1+r) & (1-r) & -(1-r) & -(1+r) \end{array} \right] \qquad (5)$$

and from (4), using (5)

$$\left\{ \begin{array}{c} \frac{\partial h}{\partial r} \\ \frac{\partial h}{\partial s} \end{array} \right\} = \underline{P}\underline{h} \qquad (6)$$

We have the gradients of head in <u>local</u> coordinates, therefore they must be transformed to enable us to determine the '$\underline{a}$' matrix, requiring $\underline{h}_{,} = \underline{a}\,\underline{h}$.

The chain rule is defined as

$$\begin{array}{ccc}
\frac{\partial h}{\partial r} & = & \frac{\partial x}{\partial r}\frac{\partial h}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial h}{\partial y} \\
\frac{\partial h}{\partial s} & = & \frac{\partial x}{\partial s}\frac{\partial h}{\partial x} + \frac{\partial y}{\partial s}\frac{\partial h}{\partial y}
\end{array} \qquad (7)$$

Or in matrix notation

$$\left\{ \begin{array}{c} \frac{\partial h}{\partial r} \\ \frac{\partial h}{\partial s} \end{array} \right\} = \left[\begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \end{array} \right] \left\{ \begin{array}{c} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial y} \end{array} \right\} \qquad (8)$$

In shorthand:

global derivatives

$$\left\{ \begin{array}{c} \frac{\partial h}{\partial r} \\ \frac{\partial h}{\partial s} \end{array} \right\} = \underline{J} \left\{ \begin{array}{c} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial y} \end{array} \right\} \qquad (9)$$

Inverting equation (9) to define the global gradients, gives,

$$\left\{ \begin{array}{c} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial y} \end{array} \right\} = \underline{J}^{-1} \left\{ \begin{array}{c} \frac{\partial h}{\partial r} \\ \frac{\partial h}{\partial s} \end{array} \right\} = \underbrace{\underline{J}^{-1} \underline{P} \underline{h}}_{\underline{a}} \quad (10)$$

and

$$\underline{J}^{-1} = \left[\begin{array}{cc} \frac{\partial y}{\partial s} & -\frac{\partial y}{\partial r} \\ -\frac{\partial x}{\partial s} & \frac{\partial x}{\partial r} \end{array} \right] \frac{1}{|J_m|} \quad (11)$$

$$|\underline{J_m}| = \frac{\partial x}{\partial r} \frac{\partial y}{\partial s} - \frac{\partial y}{\partial r} \frac{\partial x}{\partial s} \quad (12)$$

Note, that in equation (10) we have $(\partial h/\partial r, \partial h/\partial s)$ from equation (6) and the terms in (11) and (12) are purely related to element geometry and orientation. i.e.

$$\left\{ \begin{array}{c} \frac{\partial x}{\partial r} \\ \frac{\partial x}{\partial s} \end{array} \right\} = \underline{P} \underline{x} \quad (13)$$

and substituting (6) into (10) yields

$$\left\{ \begin{array}{c} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial y} \end{array} \right\} = \underbrace{\underline{J}^{-1} \underline{P}}_{\underline{a}} \underline{h} \quad (14)$$

∴ $\underline{a}$ matrix is determined.

$$\underline{D} = \left[\begin{array}{cc} K_x & K_{xy} \\ K_{yx} & K_y \end{array} \right]$$

And the full conductance matrix may be determined from:

$$\underline{K} = \int_{-1}^{+1} \int_{-1}^{+1} \underline{a}^T \underline{D} \, \underline{a} \, |\underline{J}| \, dr \, ds$$

$$\text{with} \qquad \underline{a} = \underline{J}^{-1} \underline{P}$$

Note the physical meaning of the terms in $\underline{J}^{-1}$ and $|\underline{J}|$

if the actual element in global space is a bi-unit square
then the terms of the matrices are
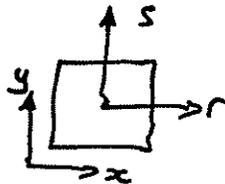
$$\partial x / \partial r = 1 \qquad \partial y / \partial s = 1$$
$$\partial x / \partial s = \partial y / \partial r = 0$$

$$\therefore |\underline{J}| = 1 \qquad \text{and} \qquad \underline{J} = \underline{J}^{-1} = \underline{I}$$

Physically suggest the transformation is 1 to 1.

(a) Two to 4 variable-number-nodes truss element

		Include only if node 3 is present	Include only if nodes 3 and 4 are present
$h_1 = \frac{1}{2}(1-r)$	....	$-\frac{1}{2}(1-r^2)$	$+\frac{1}{16}(-9r^3 + r^2 + 9r - 1)$
$h_2 = \frac{1}{2}(1+r)$	....	$-\frac{1}{2}(1-r^2)$	$+\frac{1}{16}(9r^3 + r^2 - 9r - 1)$
$h_3 = (1-r^2)$			$+\frac{1}{16}(27r^3 + 7r^2 - 27r - 7)$
$h_4 = \frac{1}{16}(-27r^3 - 9r^2 + 27r + 9)$			

(b) Interpolation functions

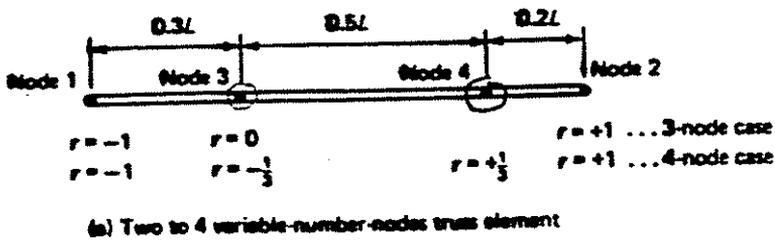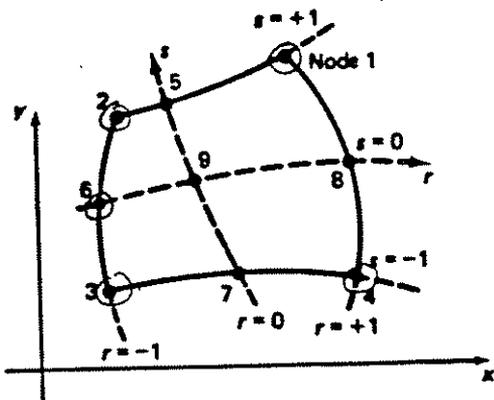*Interpolation functions of two to four variable-number-nodes one-dimensional element.*

(a) Four to 9 variable-number-nodes two-dimensional element

		Include only if node i is defined				
		$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
$h_1 =$	$\frac{1}{4}(1+r)(1+s)$	$-\frac{1}{2}h_5$	....	....	$-\frac{1}{2}h_8$	$-\frac{1}{4}h_9$
$h_2 =$	$\frac{1}{4}(1-r)(1+s)$	$-\frac{1}{2}h_5$	$-\frac{1}{2}h_6$			$-\frac{1}{4}h_9$
$h_3 =$	$\frac{1}{4}(1-r)(1-s)$	....	$-\frac{1}{2}h_6$	$-\frac{1}{2}h_7$		$-\frac{1}{4}h_9$
$h_4 =$	$\frac{1}{4}(1+r)(1-s)$	....	....	$-\frac{1}{2}h_7$	$-\frac{1}{2}h_8$	$-\frac{1}{4}h_9$
$h_5 =$	$\frac{1}{2}(1-r^2)(1+s)$	....	....	....	....	$-\frac{1}{2}h_9$
$h_6 =$	$\frac{1}{2}(1-s^2)(1-r)$	....	....	....	....	$-\frac{1}{2}h_9$
$h_7 =$	$\frac{1}{2}(1-r^2)(1-s)$	....	....	....	....	$-\frac{1}{2}h_9$
$h_8 =$	$\frac{1}{2}(1-s^2)(1+r)$	....	....	....	....	$-\frac{1}{2}h_9$
$h_9 =$	$(1-r^2)(1-s^2)$					

(b) Interpolation functions

Interpolation functions of four to nine variable-number-nodes two-dimensional element.

Figure 2.4.3.3.2   Interpolation functions for 1-D and 2-D

4 - 21 node element

(a) Eight to 20 variable number nodes three dimensional element

$h_1 = g_1 - (g_9 + g_{12} + g_{17})/2$ $h_6 = g_6 - (g_{13} + g_{14} + g_{18})/2$

$h_2 = g_2 - (g_9 + g_{10} + g_{18})/2$ $h_7 = g_7 - (g_{14} + g_{15} + g_{19})/2$

$h_3 = g_3 - (g_{10} + g_{11} + g_{19})/2$ $h_8 = g_8 - (g_{15} + g_{16} + g_{20})/2$

$h_4 = g_4 - (g_{11} + g_{12} + g_{20})/2$ $h_i = g_i$ for $i = 9, \ldots, 20$

$h_5 = g_5 - (g_{13} + g_{16} + g_{17})/2$

$g_i = 0$ if node i is not included; otherwise,

$g_i = G(r, r_i) G(s, s_i) G(t, t_i)$

$G(\beta, \beta_i) = \frac{1}{2}(1 + \beta_i \beta)$ for $\beta_i = \pm 1$

$G(\beta, \beta_i) = (1 - \beta^2)$ for $\beta_i = 0$; $\beta = r, s, t$

(b) Interpolation functions

Figure 2.4.3.4.1   Interpolation (shape) function for an
8-20 variable number node three dimensional element

[2:6] Fluid Flow and Pressure Diffusion

Recap

Transient Behavior $\quad \underline{K}\underline{h} + \underline{S}\dot{\underline{h}} = \underline{q}$

"Mass" matrices

TRANSIENT FLOW

General diffusion equation in FEM form is

$$\int_A \underline{a}^T \underline{D} \underline{a} \, dx \, dy \, \underline{h} \quad - \quad \int_A \underline{b}^T \underline{b} \, dx \, dy \, \underline{Q}$$

$$+ \underbrace{S_s \int_A \underline{b}^T \underline{b} \, dx \, dy}_{\underline{S}} \frac{\partial}{\partial t} \underline{h} \quad = \quad \underline{q} \qquad (1)$$
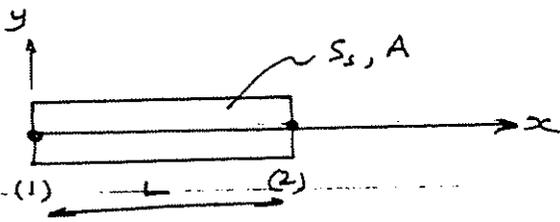
Use shorthand to represent equation as (neglecting $\underline{Q}$)

$$\underline{K} \underline{h}_z + \underline{S} \underline{\dot{h}}_z = \underline{q} \qquad (2)$$

Time steps in solution:

1) Determine $\underline{S}$ matrix
2) Perform integration of equations in time to remove $\partial/\partial t$.

Storage Matrix

$$\underline{S} = S_s \int_V b^T b \, dV$$

$$\underline{S} = A S_s \int_0^L \underline{b}^T \underline{b} \, dx \qquad \text{Shape functions:} \quad b = [\,(1-\tfrac{x}{L})\,; \; \tfrac{x}{L}\,]$$

Substituting into matrix relation:

$$\underline{S} = A S_s \int_0^L \begin{Bmatrix} 1-\tfrac{x}{L} \\ \tfrac{x}{L} \end{Bmatrix} \{1-\tfrac{x}{L} \; ; \; \tfrac{x}{L}\} \, dx = A S_s \int_0^L \begin{bmatrix} (1-\tfrac{x}{L})^2 & \tfrac{x}{L}(1-\tfrac{x}{L}) \\ \tfrac{x}{L}(1-\tfrac{x}{L}) & (1-\tfrac{x}{L})^2 \end{bmatrix} d$$

Evaluate component integrals:

$$\int_0^L (1-\tfrac{x}{L})^2 \, dx = \int_0^L (1 - \tfrac{2x}{L} + \tfrac{x^2}{L^2}) \, dx = [\, x - \tfrac{x^2}{L} + \tfrac{x^3}{3L^2} \,]_0^L$$

$$= L - L + \tfrac{L}{3} = \tfrac{1}{3}L$$

$$\int_0^L \tfrac{x}{L}(1-\tfrac{x}{L}) \, dx = \int_0^L (\tfrac{x}{L} - \tfrac{x^2}{L^2}) \, dx = [\, \tfrac{1}{2}\tfrac{x^2}{L} - \tfrac{1}{3}\tfrac{x^3}{L^2} \,]_0^L = \tfrac{1}{2}L - \tfrac{1}{3}L = \tfrac{1}{6}L$$
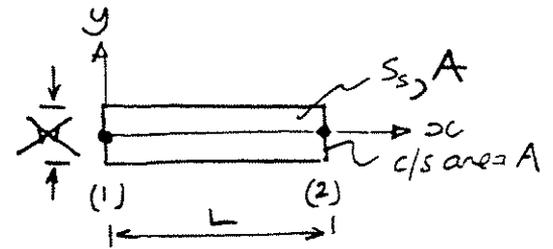
Resubstituting as Consistent Mass
$$\underline{S} = A S_s L \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{bmatrix}$$

Lumped Mass
$$\underline{S} = A S_s L \begin{bmatrix} \tfrac{1}{2} & 0 \\ 0 & \tfrac{1}{2} \end{bmatrix}$$
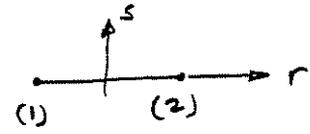
Note total mass of element is $\sum S_{ij} = 1 \underline{AL}_V \, S_s$

Example 2.4.4   Storage Matrix

GLOBAL

LOCAL

Consider a one-dimensional element of specific storage, S_s, and length, L.

Use the isoparametric concept

Then shape function, $\underline{b}$. $\underline{b} = \frac{1}{2}[(1-r) ; (1+r)]$ (1)

Storage matrix

$$\underline{S} = A S_s \int_0^L \underline{b}^T \underline{b} \, dx = S_s A \int_{-1}^{+1} \underline{b}^T \underline{b} |\underline{J}| \, dr \quad (2)$$

$$|\underline{J}| = \frac{\partial x}{\partial r} = \frac{L}{2} \quad (3)$$

Substituting (1) and (3) into (2) gives

$$\underline{S} = \frac{S_s A L}{8} \int_{-1}^{+1} \underbrace{\begin{bmatrix} (1-r)^2 & (1-r)(1+r) \\ (1-r)(1+r) & (1+r)^2 \end{bmatrix}}_{\underline{b}^T \underline{b}} dr \quad (4)$$

This may be evaluated either analytically or using 2 point quadrature $r = \frac{1}{\sqrt{3}}$.

$$\underline{S} = \frac{S_s A L}{8} \begin{bmatrix} 2.66 & 1.33 \\ 1.33 & 2.66 \end{bmatrix} \quad (\text{Consistent}) \quad (5)$$

This consistent matrix may be lumped at the nodes as

$$\underline{S} = \frac{S_s A L}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (\text{Lumped}) \quad (6)$$

Total amount of fluid in storage is given by $\sum_{i=1}^{2} \sum_{j=1}^{2} S_{ij} = S_s A L$

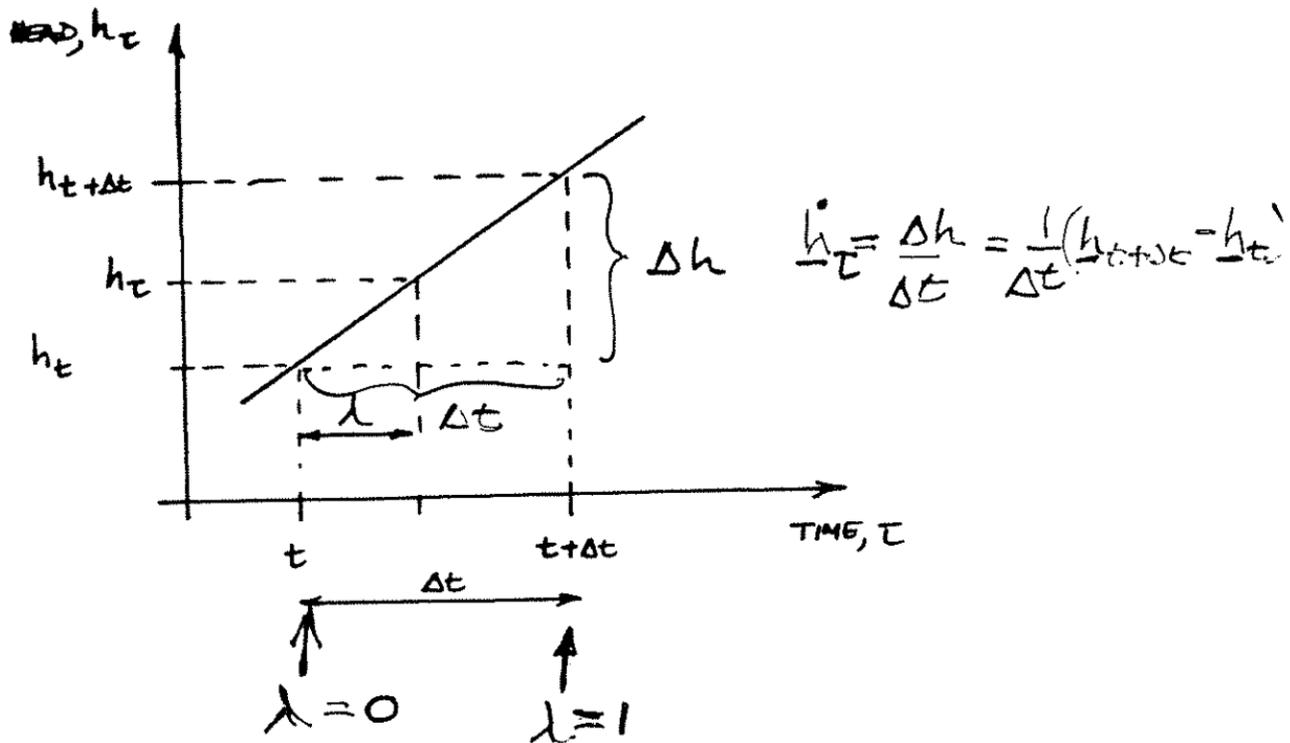or $S_s \times (\text{Element volume})$.

$$\dot{h_\tau} = \frac{\Delta h}{\Delta t} = \frac{1}{\Delta t}\left(h_{t+\Delta t} - h_t\right)$$

Figure 2.4.4.1.1 Variation in nodal head over the
time interval t to t + Δt

IMPLICIT TIME INTEGRATION

Exact statement @ time τ
$$\underline{K}\,\underline{h}_\tau + \underline{S}\,\underline{\dot{h}}_\tau = \underline{q}_\tau \qquad (1)$$

Write FE equations at time $\tau = t+\Delta t$ $\therefore$ using (1)

$$\underline{K}\,\underline{h}_{t+\Delta t} + \underline{S}\,\underline{\dot{h}}_{t+\Delta t} = \underline{q}_{t+\Delta t} \qquad (2)$$

Linear head change in time Δt, gives

$$\underline{\dot{h}}_\tau = \frac{1}{\Delta t}(\underline{h}_{t+\Delta t} - \underline{h}_t) \qquad (3)$$

and $\quad \underline{\dot{h}}_{t+\Delta t} \simeq \frac{1}{\Delta t}(\underline{h}_{t+\Delta t} - \underline{h}_t) \qquad (4)$

Substituting (4) into (2) and rearranging gives

$$\underline{K}\,\underline{h}_{t+\Delta t} + \underline{S}\frac{1}{\Delta t}(\underline{h}_{t+\Delta t} - \underline{h}_t) = \underline{q}_{t+\Delta t} \qquad (5)$$
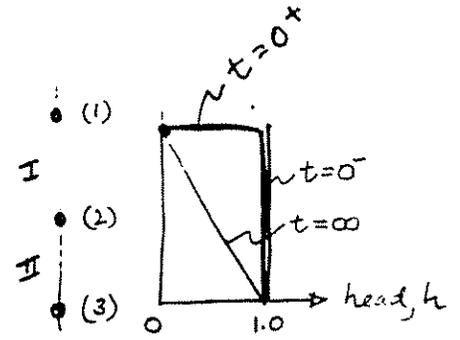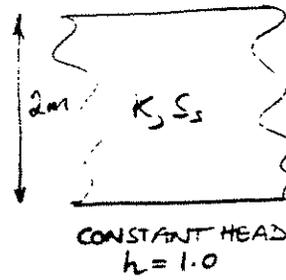
or. $\quad [\underline{K} + \frac{1}{\Delta t}\underline{S}]\,\underline{h}_{t+\Delta t} = \underline{q}_{t+\Delta t} + \frac{1}{\Delta t}\underline{S}\,\underline{h}_t \qquad (6)$

or $\quad \underline{K}^*\underline{h}_{t+\Delta t} = \underline{q}^*_{t+\Delta t}$

<u>Unconditionally stable method.</u> $\qquad (\lambda = 1.0)$

Example 2.4.4.1 One-Dimensional Process

One dimensional flow in a body, initially at $h = 1.0$. At time $t = 0^+$ the topsurface head is changed to $h = 0$ and held constant. The basal head is retained at $h = 1.0$.

CONSTANT HEAD
$h = 1.0$

$$K = 1.0 \text{ m/s} \quad ; \quad S_s = \underset{2.0}{\cancel{0.5}} \text{ m}^{-1} \quad ; \quad \Delta t = 0.1 \text{ s}$$

Using two node elements

$$\underline{K}_I = \underline{K}_{II} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{1}$$

$$\underline{S}_I = \underline{S}_{II} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

The system matrices are

$$\underline{K}\,\underline{h}_\tau + \underline{S}\,\underline{\dot{h}}_\tau = \underline{q}_\tau \tag{3}$$

$$\underbrace{\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}}_{\underline{K}} \begin{Bmatrix} h_1 \\ h_2 \\ h_3 \end{Bmatrix}_\tau + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\underline{S}} \begin{Bmatrix} \dot{h}_1 \\ \dot{h}_2 \\ \dot{h}_3 \end{Bmatrix}_\tau = \begin{Bmatrix} q_1 \\ q_2 \\ q_3 \end{Bmatrix}_\tau \tag{4}$$

Since the heads at nodes 1 and 3 are known at time $t = 0^+$; $h_1 = 0$; $h_3 = 1.0$. Rearranging equation (4) gives

$$\begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{Bmatrix} 0 \\ h_2 \\ 0 \end{Bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} \dot{h}_1 \\ \dot{h}_2 \\ \dot{h}_3 \end{Bmatrix}_{t+\Delta t} = \begin{Bmatrix} q_1 - h_1 \\ \cancel{q_2}^{\,0} + h_1 + h_3 \\ q_3 - h_3 \end{Bmatrix}_{t+\Delta t} \tag{5}$$

where only the second equation remains active. Rearranging to form of equation (2.6.4.1.8) the system matrices may be determined, in the central active equation, since h_1 and h_3 are known for all time. Correspondingly

$$\underline{K}^* = [\underline{K} + \tfrac{1}{\Delta t}\underline{S}] = 2 + \tfrac{1}{0.1}(2) = 22 \tag{6}$$

$$\underline{q}^*_{t+\Delta t} = \cancel{q_2}^{0} + h_1 + h_3 + \frac{1}{\Delta t} \underline{S} \, h_2 \tag{7}$$

$$= 0 + 0 + 1.0 + \frac{1}{0.1}(2) h_2 \qquad (h_2 = 1.0)$$

$$\underline{q}^*_{t=0.1} = 21.0 \tag{8}$$

The single equation is, therefore, from equations (6) and (8)

$$\underline{K}^* \, \underline{h}_{t+\Delta t} = \underline{q}^*_{t+\Delta t} \tag{9}$$

$$22 \, h_2 = 21 \qquad\qquad h_2 = 0.9545 @ \ t+\Delta t = 0.1$$

For the next time step, $\underline{K}^*$, is the same and only $\underline{q}^*_{t+\Delta t}$ changes. For time level $t+\Delta t = 0.2$

$$\underline{q}^*_{t+\Delta t} = \cancel{q_{t+\Delta t}}^{0} + \cancel{h_1}^{0} + \cancel{h_3}^{1.0} + \frac{1}{\Delta t} \underline{S} \cancel{h_t}^{(21/22)} \tag{10}$$

$$\underline{q}^*_{t+\Delta t} = 0 + 0 + 1.0 + \frac{1}{0.1}(2)\left(\frac{21}{22}\right) =$$

$$\underline{K}^* \, \underline{h}_{t+\Delta t} = \underline{q}^*_{t+\Delta t} \qquad\qquad h_2 = 0.9132 @ \ t+\Delta t = 0.2$$

The recurrence relationship is therefore

$$h_{t+\Delta t} = \frac{1}{K^*}\left(h_3 + \frac{1}{\Delta t} S \, h_t\right) = \frac{1}{22}\left(1 + 20 h_t\right)$$

enabling the history of heads to be evaluated as illustrated graphically for time steps varying between $0.01 < \Delta t < 5.0$ in Figure 26.4.1.1

Although unconditionally stable, the solution converges as Δt approaches 0.1 s. Thus, time discretization is important as an aspect separate from stability. The results illustrated in the figure are only approximate. The will be as adequate as the spatial discretization.

IMPLICIT SOLUTION - Unconditionally stable
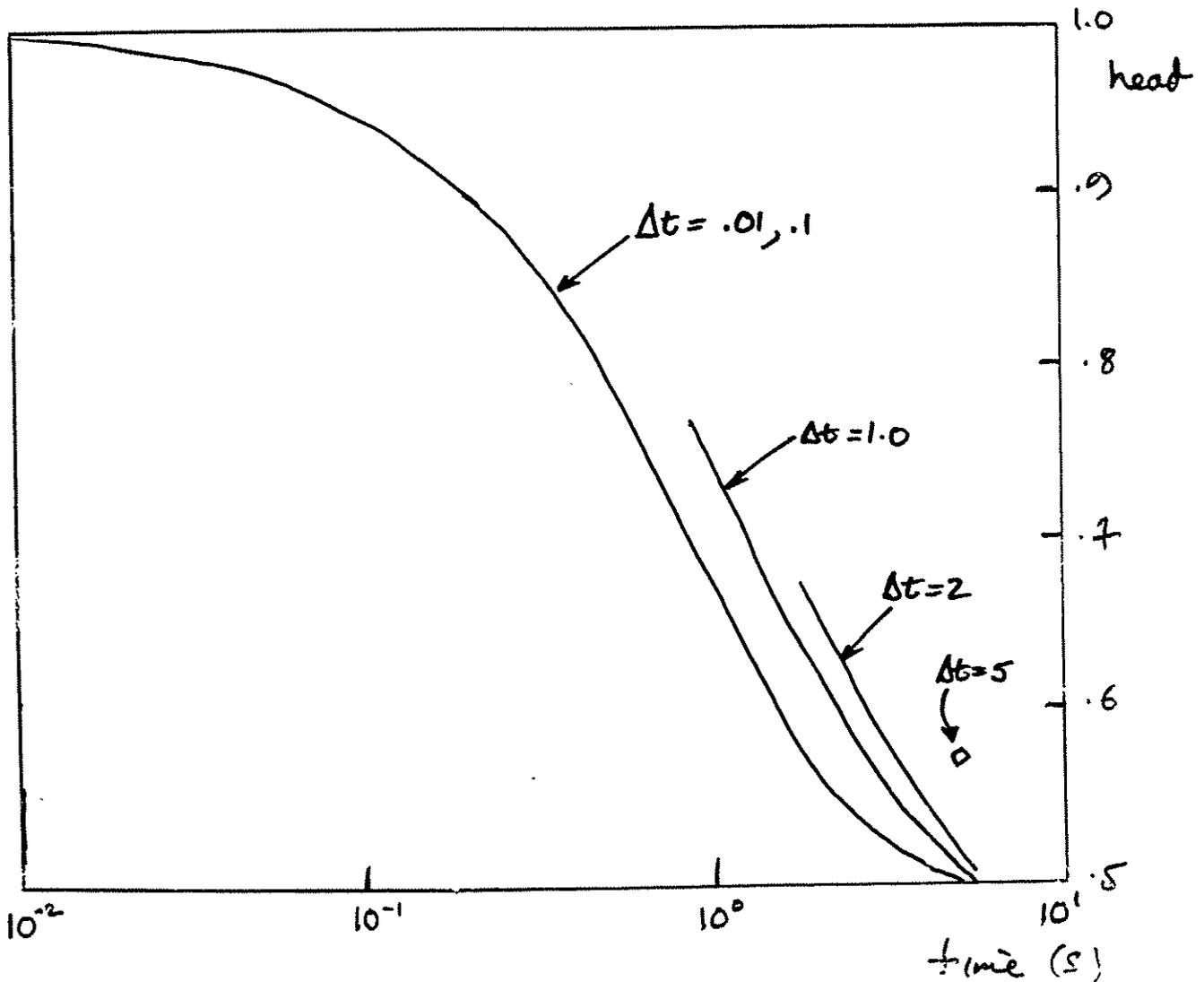
Solution converges for $\Delta t \leq .1$

$\therefore$ solution stable in time.

Not, however, exact solution since spatial discretization
is coarse.

[2:7] Fluid Flow and Pressure Diffusion

Recap

Transient Behavior $\quad \underline{K}\underline{h} + \underline{S}\underline{\dot{h}} = \underline{q}$

Time integration

EGEEfem

EXPLICIT TIME INTEGRATION

Instead, write equation at time $\tau = t$, then

$$\underline{K}\underline{h}_t + \underline{S}\dot{\underline{h}}_t = \underline{q}_t \qquad (1)$$

Form $\underline{S}$ as a lumped matrix (terms on diagonal only) then

$$\underline{S}^{-1} = \frac{1}{\underline{S}} \qquad (2)$$

and (1) may be rearranged as

$$\dot{\underline{h}}_t = \underline{S}^{-1}[\underline{q}_t - \underline{K}\underline{h}_t] \qquad (3)$$

and the time derivative of head may also be evaluated as

$$\underline{h}_{t+\Delta t} = \underline{h}_t + \Delta t\, \dot{\underline{h}}_t \qquad (4)$$
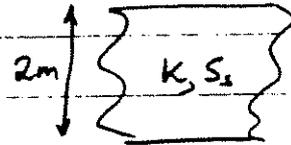
The magnitudes of heads may be evaluated as follows:

1. Evaluate $\underline{K}$ and $\underline{S}$
2. Evaluate $\dot{\underline{h}}_t$ from (3)
3. Evaluate $\underline{h}_{t+\Delta t}$ from (4)
4. Reevaluate $\dot{\underline{h}}_t$ with new heads where $\tau = t + \Delta t$

Conditionally stable.

Example 2.4.4.2   One -Dimensional Problem

Using the same physical parameters as
   Example 2.4.4.1.

$$2m \left\{ \begin{array}{c} K, S_s \end{array} \right\}$$

Assembling the system in rearranged form $\underline{K}\,\underline{h}_t + \underline{S}\,\dot{\underline{h}}_t = \underline{q}_t$

$$\begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{Bmatrix} 0 \\ h_2 \\ 0 \end{Bmatrix}_t + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} \dot{h}_1 \\ \dot{h}_2 \\ \dot{h}_3 \end{Bmatrix} = \begin{Bmatrix} q_1 - h_1 \\ q_2 + h_1 + h_3 \\ q_3 \quad - h_3 \end{Bmatrix} \quad (1)$$

The single active equation is $K\,h_2 + S\dot{h}_2 = q_2 + h_1 + h_3 \quad (2)$

Substituting into equation (2.4.4.2.3) gives

$$\dot{\underline{h}}_t = \underline{S}^{-1} [\,\underline{q}_{t+\Delta t} + h_1 + h_3 - \underline{K}\,\underline{h}_t\,] \qquad (3)$$

$$\dot{\underline{h}}_t = \tfrac{1}{2} [\,0 \quad + 0 \quad + \frac{1.0}{P} - (2)\underline{h}_t\,] \qquad (4)$$

$$\underline{h}_{t+\Delta t} = \underline{h}_t + \Delta t\,\dot{\underline{h}}_t \qquad (5)$$

For $\Delta t = 0.1s$ and other time step magnitudes the transient behavior of head at node 2 is given in Table 2.4.4.2.1 and shown graphically in Figure 2.4.4.2.1

The solution method is only conditionally stable. For $\Delta t \geq 1.0$ the solution oscillates.

The primary advantage is that a conductance matrix never has to be inverted or solved. This method is particularly suited to nonlinear problems where conductivity magnitudes change with time or head gradient.
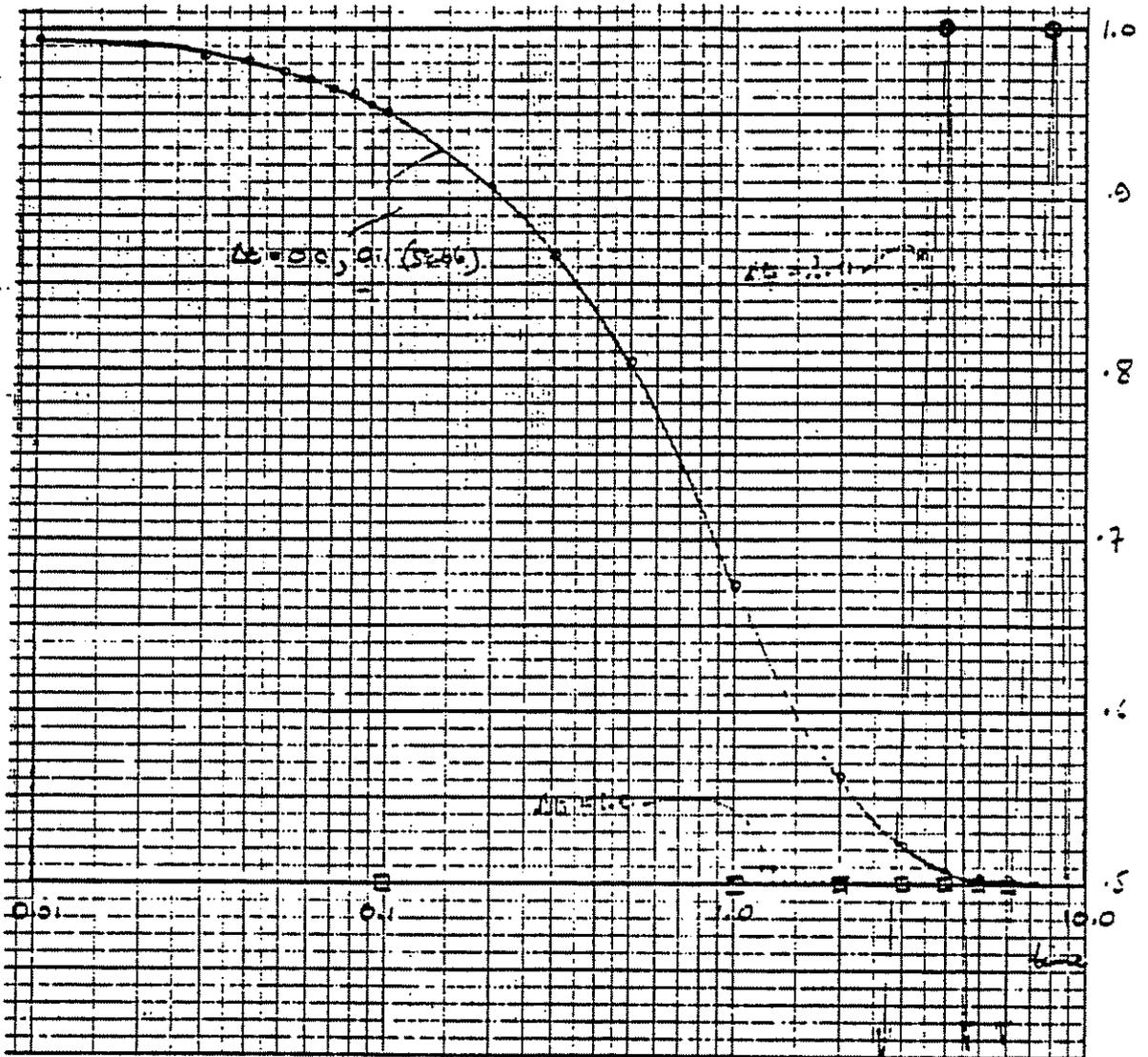
Note — unstable in this instance for $\Delta t > 1.0$

$\Delta t = 1.0 \rightarrow$ steady, incorrect solution for h

$\Delta t = 2.0 \rightarrow$ oscillating solution

NOTE: Very good correspondence with implicit solution

Figure 2.4.4.2.1   Transient response at node 2.

GENERAL TIME INTEGRATION

Define
$$h_\tau = h_t + \lambda(h_{t+\Delta t} - h_t) \qquad (1)$$

Linear gradient as
$$\underline{h}_\tau' = \tfrac{1}{\Delta t}(\underline{h}_{t+\Delta t} - \underline{h}_t) \qquad (2)$$

Substitute (1) and (2) into the following,

$$\underline{K}\,\underline{h}_\tau + \underline{S}\,\underline{h}_\tau' = \underline{q}_\tau \qquad (3)$$

To yield:
$$\underline{K}[(1-\lambda)\underline{h}_t + \lambda\underline{h}_{t+\Delta t}] + \tfrac{1}{\Delta t}\underline{S}(\underline{h}_{t+\Delta t} - \underline{h}_t) = \underline{q}_\tau \qquad (4)$$

This yields: Implicit for $\lambda = 1$ backward difference
Crank–Nicolson $\lambda = \tfrac{1}{2}$ central difference
Explicit for $\lambda = 0$ forward difference

$\lambda \geq \tfrac{1}{2}$ are unconditionally stable
$\lambda < \tfrac{1}{2}$ are conditionally stable.

```fortran
      SUBROUTINE ELMT04(D,UL,XL,IX,TL,S,P,NDF,NDM,NST,ISW)
      IMPLICIT REAL*8(A-H,O-Z)
C
C.... THREE NODED CONSTANT GRADIENT FLOW ELEMENT
C
C     USER INFORMATION
C
C     INPUT
C
C       VAR      FORMAT          DESCRIPTION
C       ------------------------------------------------------------
C
C       D(1)    F10.0           HYDRAULIC CONDUCTIVITY
C       D(2)    F10.0           SPECIFIC STORAGE
C
C-------------------------------------------------------------------
C
C     LOCAL NODAL NUMBERING MUST BE COUNTER-CLOCKWISE
C
C-------------------------------------------------------------------
C
C     VARIABLES
C
C     NEL    -    NUMBER OF NODES PER ELEMENT
C     NDF    -    NUMBER OF DEGREES OF FREEDOM PER NODE
C     NST    -    NUMBER OF DEGREES OF FREEDOM PER ELEMENT (NEN*NDF)
C     ISW    -    FUNCTION CALL NO.
C                 1 = READ ELEMENT SPECIFIC INPUT DATA
C                 2 = PERFORM MESH CHECK
C                 3 = FORM ELEMENT STIFFNESS MATRIX       - TANG
C                 4 = EVALUATE ELEMENT STRESSES           - STRE
C                 5 = FORM CONSISTENT/LUMPED MASS MATRIX  - CMAS/LMAS
C                 6 = FORM LOAD VECTOR                     - FORM
C                     OR EVALUATE NODAL FORCES            - REAC
C
C     ARRAYS - GIVEN
C
C     UL(1,J)    SPECIFIED HEAD BOUNDARY CONDITION FOR
C                DEGREE OF FREEDOM J (J=1,3)
C     XL(I,J)    COORDINATE IN THE I DIRECTION AT NODE J
C                EG. XL(1,3) IS X COORDINATE OF NODE K
C
C     ARRAYS - EVALUATED
C
C     A(  )      A MATRIX
C     C(  )      D MATRIX
C     S(I,J)     CONDUCTANCE MATRIX S = AT*D*A DV
C                FOR ROW (VERTICAL) I AND COLUMN (HORIZ.) J
C     P(I)       MODIFIED LOAD VECTOR FOR LOCAL DOF I (IGNORE)
C
C
C     FOR LMAS CALCULATION THE VECTOR LOCATIONS P(1), P(2), P(3)
C     ARE USED FOR THE STORAGE VECTOR
C
C
C
C-------------------------------------------------------------
```

```
       CHARACTER*4 O,HEAD
       COMMON  /CDATA/ O,HEAD(20),NUMNP,NUMEL,NUMMAT,NEN,NEQ,IPR
       COMMON  /ELDATA/ DM,N,MA,MCT,IEL,NEL
       DIMENSION D(2),UL(1,1),XL(NDM,1),IX(1),TL(1),S(NST,1),P(1)
      1      ,A(2,3),C(2,2)
C.... GO TO CORRECT ARRAY PROCESSOR
       GO TO(1,2,3,4,5,3),ISW
C.... INPUT MATERIAL PROPERTIES
1      READ (5,1000) D(1),D(2)
       WRITE(6,2000) D(1),D(2)
       RETURN
C.... MESH CHECKING FACILITY
2      RETURN
C.... CONDUCTANCE MATRIX COMPUTATION
3      CONTINUE
C      EVALUATE TERMS IN THE CONDUCTIVITY
C      TENSOR  ...I.E. THE D( ) MATRIX IN CLASS
C      AND PLACE THEM IN THE C(2,2) ARRAY
C
C
C.... EVALUATE COEFFICIENTS IN A( ) MATRIX
C
C
C.... COMPLETE TRIPLE MATRIX PRODUCT  AT*D*A
C      AND PLACE THE RESULT IN THE S(3,3) ARRAY
C
C
C
C.... PERFORM VOLUME INTEGRATION (*AREA)
C      BY EVALUATING THE DETERMINANT OF THE
C      COORDINATE MATRIX
C
C
C      END OF YOUR MODIFICATIONS
C
C.... MODIFY LOAD VECTOR FOR BOUNDARY CONDITIONS
       DO 320 I=1,3
       DO 320 J=1,3
320    P(I) = P(I) - S(I,J)*UL(1,J)
       RETURN
C.... END OF CONDUCTANCE MATRIX DETERMINATION
4      RETURN
C.... LUMPED MASS COMPUTATION
5      CONTINUE
C
C      EVALUATE DETERMINANT OF NODAL COORDINATE MATRIX
C      TO DEFINE AREA (VOLUME) OF ELEMENT.
C      APPLY PRODUCT OF VOLUME AND STORAGE EQUALLY
C      TO EACH OF THE NODES IN ARRAY P(3).
C
C
       RETURN
C.... FORMATS FOR INPUT AND OUTPUT
1000  FORMAT(2F10.0)
2000  FORMAT(/5X,'THREE NODED CONSTANT STRAIN ELEMENT',//
      1 10X,'HYDRAULIC CONDUCTIVITY   ',6X,E14.7,/
      2 10X,'SPECIFIC STORAGE         ',6X,E14.7,/)
```

END

```
FEAP    SIX TRIANGULAR ELEMENTS-FLOW-STEADY
    8   6   1   2   1   3
COORD
    1   2      0.0        0.0
    7   0      3.0        0.0
    2   2      0.0        1.0
    8   0      3.0        1.0

ELEM
    1   1   1   4   2
    2   1   1   3   4
    3   1   3   5   4
    4   1   5   6   4
    5   1   5   8   6
    6   1   5   7   8

MATE
    1   4          MATERIAL 1
      1.0       1.0

BOUN
    1       1
    2       1
    7       1
    8       1

FORC
    1       1.0
    2       1.0
    7       0.0
    8       0.0

END
MACR
TANG
FORM
SOLV
DISP
END
STOP
```

```
FEAP    SIX TRIANGULAR ELEMENTS-FLOW-TRANSIENT
    8   6   1   2   1   3
COORD
    1   2      0.0       0.0
    7   0      3.0       0.0
    2   2      0.0       1.0
    8   0      3.0       1.0

ELEM
    1   1   1   4   2
    2   1   1   3   4
    3   1   3   5   4
    4   1   5   6   4
    5   1   5   8   6
    6   1   5   7   8

MATE
    1   4           MATERIAL 1
       1.0      1.0

BOUN
    1       1
    2       1

FORC
    1       1.0
    2       1.0

END
MACR
DT          0.1
TANG
FORM
LMAS
LOOP        10
TIME
IMPL
SOLV
DISP
NEXT
END
STOP
```

3
Mass Transport

[3:1] Mass Transport

Introduction

Advection-Diffusion Equation

$$\underline{S}\,\underline{\dot{c}} + [\underline{K}_d + \underline{K}_a]\,\underline{c} = \underline{q} + \underline{R}$$

Galerkin method

1D Example – stability

Transient response

5. Mass (Chemical) Transport (C)   **Transport**
    5.1. Conservation of mass and Fick's law
    5.2. Steady behavior
    5.3. Transient behavior
    5.4. Considerations of local equilibrium

PROCESS COUPLINGS [T-H-M-C]

$$\begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & R_{44} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{p} \\ \underline{T} \\ \underline{c} \end{Bmatrix} + \begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & S_{44} \end{bmatrix} \begin{Bmatrix} \underline{\dot{u}} \\ \underline{\dot{p}} \\ \underline{\dot{T}} \\ \underline{\dot{c}} \end{Bmatrix} = \begin{Bmatrix} \underline{\dot{f}} + \dots \\ \underline{q}_F + \dots \\ \underline{q}_T + \dots \\ \underline{q}_M + \dots \end{Bmatrix}$$

R_{44} ↑

Diffusion/Dispersion
and Advection

S_{44} ↑

$\left(\dfrac{\partial c}{\partial t}\right)_{SPATIAL} + \left(\dfrac{\partial c}{\partial t}\right)_{REACTION}$

SOLID MECHANICS	FLOW SYSTEM	TRANSPORT

SOLID MECHANICS

- Conservation of momentum: (Equilibrium), $VW_I = VW_E$

- Continuity (Compatibility): $\underline{\varepsilon} = \underline{a}\,\underline{u}$

- Constitutive relation: $\underline{\sigma} = \underline{D}\,\underline{\varepsilon}$

- Initial Conditions

- Boundary Conditions

FLOW SYSTEM

- Conservation of mass: $\underline{\nabla}^T \underline{q} = 0$

- Continuity: $\underline{h}_{,} = \underline{a}\,h$

- Constitutive rel'n. $\underline{v} = \underline{D}\,\underline{h}_{,}$

- ICs

- BCs

TRANSPORT

- Conservation of mass $\underline{\nabla}^T \underline{q} = 0$

- Continuity: $\underline{c}_{,} = \underline{a}\,\underline{c}$

- Constitutive:
  diffusion - $\underline{v}_{,} = \underline{D}\,\underline{c}_{,}$
  advective - $\underline{v}_{,} = \underline{A}\,\underline{c}$

- ICs

- BCs

— SOLVE SYSTEM EQUATIONS —

TRANSPORT EQUATIONS

Two modes of transport: **Diffusion:** $\quad q_i^d = -D_{ij}\dfrac{\partial c}{\partial x_j}$ $\qquad$ (1)

→ flow, V
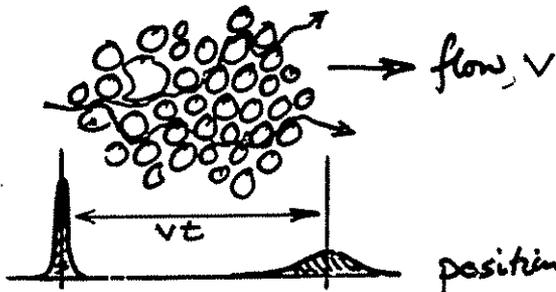
Fick's law and hydrodynamic dispersion.

c = mass per unit volume

increasing dispersion

(similar to a diffusive process)
(but as a result of flow tortuosity
and not diffusion gradient $\partial c/\partial x$)

Advection: $\qquad q_i^a = v_i c$ $\qquad\qquad v_i = \dfrac{V_{darcy}}{n}$

---

continuity: $\qquad\qquad \dfrac{\partial q_i}{\partial x_i} = \dfrac{\partial c}{\partial t}$ $\qquad$ (3)

substituting (1) and (2) into (3) gives (where $q_i = q_i^d + q_i^a$)

$$\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial c}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(v_i c) = \frac{\partial c}{\partial t} \qquad (4)$$

Or expanding to 2-D case, then,

$$\underbrace{D_x\frac{\partial^2 c}{\partial x^2} + D_y\frac{\partial^2 c}{\partial y^2}}_{\text{Diffusive/dispersive flux}} \;\underbrace{-\,v_x\frac{\partial c}{\partial x} - v_y\frac{\partial c}{\partial y}}_{\text{Advection flux}} = \frac{\partial c}{\partial t}$$

TRANSPORT EQUATIONS

Galerkin:
$$\int_A w\left[\underbrace{\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial c}{\partial x_j}\right)}_{\text{Diffusive}} - \underbrace{\frac{\partial}{\partial x_i}(v_i c)}_{\text{Advective}} - \frac{\partial c}{\partial t}\right]\,dx\,dy = 0 \qquad (1)$$

Form of equations with <u>advective</u> component are potentially unstable.

Use Galerkin technique and advective term is additional:

Define shape functions

$$c = \underline{b}\,\underline{c}$$
$$\underline{w} = \underline{b} \qquad\qquad (2)$$

Apply Green's Theorem to equations to give

$$\int_A \underline{a}^T \underline{D}\, \underline{a}\, dx\,dy\, \underline{c} \quad + \quad \int_A \underline{b}^T \underline{b}\, dx\,dy\, \underline{\dot{c}}$$

$$+ \int_A \underline{b}^T \underline{v}\, \underline{a}\, dx\,dy\, \underline{c} \quad = \quad \underline{q}$$

Where individual equations are defined as:

$$\underline{b} = [\,b_1,\, b_2 \;\ldots\; b_n\,]$$

$$\underline{a} = \left\{\begin{matrix} \partial/\partial x \\ \partial/\partial y \end{matrix}\right\}\, \underline{b}$$

$$\underline{D} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$$

$$\underline{c}^T = [\,c_1,\, c_2,\, c_3 \;\ldots\; c_n\,]$$

$$\underline{v} = [\,v_x\,;\, v_y\,]$$

$$\underline{q} = \text{fluxes @ boundary (prescribed)}$$

MATRIX FORM FOR EQUATIONS

$$[\underline{K}_d + \underline{K}_a]\, \underline{c}_\tau \;+\; \underline{S}\, \underline{\dot{c}}_\tau \;=\; \underline{q}_\tau$$

$$\underline{K}_d = \int_A \underline{a}^T \underline{D}\,\underline{a}\; dx\, dy$$

$$\underline{K}_a = \int_A \underline{b}^T \underline{v}\,\underline{a}\; dx\, dy$$

$$\underline{S} = \int_A \underline{b}^T \underline{b}\; dx\, dy \qquad = \text{volume of element.}$$

SOLVE AS LINEAR SYSTEM OF EQUATIONS

$$\underline{\bar{K}}\, \underline{c}_\tau \;+\; \underline{S}\, \underline{\dot{c}}_\tau \;=\; \underline{q}_\tau$$

Implicit: $\qquad \lambda = 1$

$\qquad\qquad\quad \tau = t + \Delta t$

$$\underline{\bar{K}}^* \underline{c}_{t+\Delta t} \; \underline{q}^*_{t+\Delta t}$$

$$\underline{\bar{K}}^* = [\underline{\bar{K}} + \frac{1}{\Delta t}\underline{S}]$$

$$\underline{q}^*_{t+\Delta t} = \underline{q}_{t+\Delta t} + \frac{1}{\Delta t}\underline{S}\underline{c}_t$$

Explicit: $\qquad \lambda = 0.0$

$\qquad\qquad\quad \tau = t$

$$\underline{\dot{c}}_t = \underline{S}^{-1}[\underline{q}_\tau - \underline{\bar{K}}\underline{c}_t]$$

$$\underline{c}_{t+\Delta t} = \underline{c}_t + \Delta t\, \underline{\dot{c}}_t$$

Summary of Notation – Advection-Diffusion Equation

Tensor:

$$A\frac{\partial c}{\partial t} + \nabla \cdot (-D\nabla c) = R \quad \text{with} \quad \nabla = \begin{Bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{Bmatrix} \quad and \quad \nabla \cdot \nabla = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2 \tag{1}$$

Matrix:

$$A\dot{c} - \underline{\nabla}^T D\ \underline{\nabla} c = R - \mathbf{v}^T \underline{\nabla} c \quad \text{with} \quad \underline{\nabla} = \begin{Bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{Bmatrix}, \quad \mathbf{v} = \begin{Bmatrix} v_x \\ v_y \\ v_z \end{Bmatrix}, \quad and \quad \nabla \cdot \nabla = \underline{\nabla}^T \underline{\nabla} = \nabla^2 \tag{2}$$

Finite Element Statement

Galerkin – Pre-weight by $\underline{b}^T$ and integrate over the volume of the domain**:**

$$\int_V \underline{b}^T [A\dot{c} - \underline{\nabla}^T D\ \underline{\nabla} c - R + \mathbf{v}^T \underline{\nabla} c = 0]\, dV \tag{3}$$

Note that we can define concentrations, c, and concentration gradients, $\underline{c}$, , in termos of nodal concentration, $\underline{c}$, as,

$$c = \underline{b}\ \underline{c} \tag{4}$$

$$c, = \nabla c = \nabla \underline{b}\ \underline{c} = \underline{a}\ \underline{c} \tag{5}$$

Substituting the nodal concentrations of equation **(4)** and the gradient of concentration of equation **(5)** into equation **(3)** yields

$$\int_V \underline{b}^T [A\underline{b}\ \dot{\underline{c}} - \underline{\nabla}^T D\ \underline{a}\ \underline{c} - R + \mathbf{v}^T \underline{a}\ \underline{c} = 0]\, dV \tag{6}$$

And noting the standard result for transposed matrices that $\underline{b}^T\ \underline{\nabla}^T = [\underline{\nabla}\ \underline{b}]^T = \underline{a}^T$ yields on substitution in equation **(6)**.

$$\int_V [\underline{b}^T A\ \underline{b}\ \dot{\underline{c}} - \underline{b}^T \underline{\nabla}^T D\ \underline{a}\ \underline{c} - \underline{b}^T R + \underline{b}^T \mathbf{v}^T \underline{a}\ \underline{c} = 0]\, dV \tag{7}$$
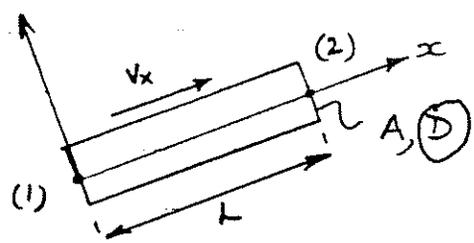
and noting that $R = \underline{b}\ \underline{R}$

$$\int_V [\underbrace{\underline{b}^T A\ \underline{b}}_{\underline{S}}\ \dot{\underline{c}} - \underbrace{\underline{a}^T D\ \underline{a}}_{\underline{K}_d}\ \underline{c} - \underbrace{\underline{b}^T \underline{b}}_{\underline{R}}\ \underline{R} + \underbrace{\underline{b}^T \mathbf{v}^T \underline{a}}_{\underline{K}_a}\ \underline{c} = 0]\, dV \tag{8}$$

Yields

$$\underline{S}\ \dot{\underline{c}} + [\underline{K}_d + \underline{K}_a]\ \underline{c} = \underline{q} + \underline{R} \tag{9}$$

Example 3.1.1   One Dimensional Element

$$\underline{b} = [\,(1 - \tfrac{x}{L})\,; \ \tfrac{x}{L}\,]$$

$$\underline{a} = \frac{\partial}{\partial x}\underline{b} = \frac{1}{L}[-1\,;\ 1]$$

$$\underline{c} = \begin{Bmatrix} c_1 \\ c_2 \end{Bmatrix}$$

From equation (3.1.3);

$$\underline{K_d} = \int_A \underline{a}^T \underline{D}\,\underline{a}\ dx\,dy \quad = \quad \frac{AD}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad \text{Diffusive}$$

$$\underline{K_a} = A\int_{x_o}^{L} \underline{b}^T \underline{v}\,\underline{a}\ dx\,dy \quad = \quad AV_x \int_{x_1}^{x_2} \begin{Bmatrix} (1 - \tfrac{x}{L}) \\ \tfrac{x}{L} \end{Bmatrix} \frac{1}{L}[-1\,;\ 1]\ dx$$

$$\underline{K_a} = \frac{AV_x}{L}\int_{x_1}^{x_2} \begin{bmatrix} -(1 - \tfrac{x}{L}) & (1 - \tfrac{x}{L}) \\ -\tfrac{x}{L} & \tfrac{x}{L} \end{bmatrix} dx$$

Integrals $\qquad \int_0^L (1 - \tfrac{x}{L})\,dx = [\,x - \tfrac{x^2}{2L}\,]_0^L = \tfrac{L}{2}$

$$\int_0^L \tfrac{x}{L}\,dx = [\,\tfrac{x^2}{2L}\,]_0^L = \tfrac{L}{2}$$
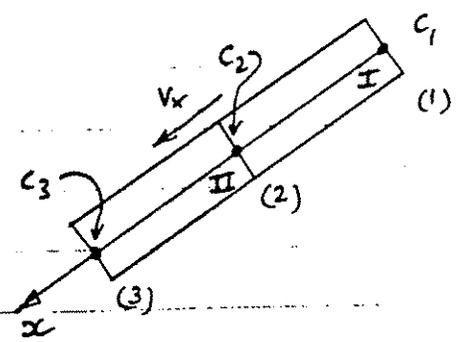
$$\underline{K_a} = \frac{AV_x}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \qquad \text{Advective}$$

3.1.2 Example — Steady Transport

Steady flow example
$A = 1.0$; $L = 1.0$
$C_1 = 1.0$; $C_3 = 0.0$

$$\underline{K_d} = D \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad ; \quad \underline{K_a} = \frac{V_x}{2} \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$\underline{K} = \{ \underline{K_d} + \underline{K_a} \} = \begin{bmatrix} (D - \frac{V}{2}) & -(D - \frac{V}{2}) \\ -(D + \frac{V}{2}) & (D + \frac{V}{2}) \end{bmatrix}$$
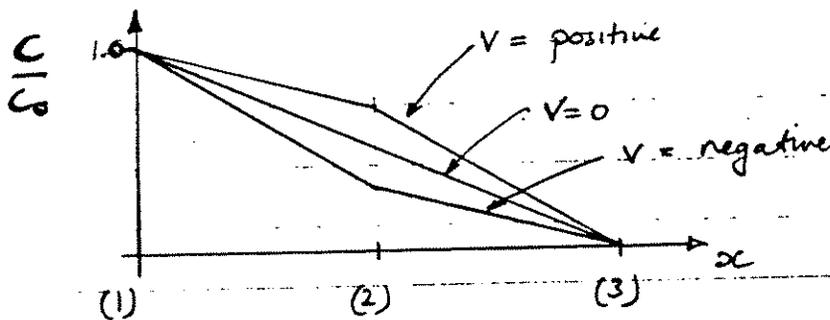
Only one unknown exists — $C_2 = ?$
A single equation results

$$[(D + \frac{V}{2}) + (D - \frac{V}{2})] \, C_2 = (D + \frac{V}{2}) C_1^{1.0}$$

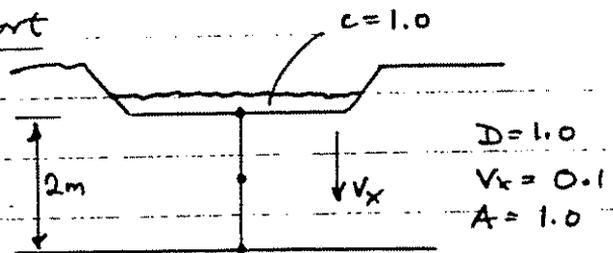$$C_2 = (\frac{1}{2} + \frac{V}{4D}) C_1^{1.0}$$

This distribution is illustrated below

The system is over constrained, as V increases, the solution is unbounded.

3.1.3 Example — Transient Transport

Leakage from a lagoon.

$c = 1.0$

$D = 1.0$
$V_x = 0.1$
$A = 1.0$

$\dfrac{AD}{L} = 1.0 \; ; \quad \dfrac{V_x A}{2} = +0.05$

$\dfrac{AL}{2} = 0.5$

$$\underline{K_d} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \; ; \qquad \underline{K_a} = 0.05 \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$[\underline{K_d} + \underline{K_a}] = \begin{bmatrix} 0.95 & -0.95 \\ -1.05 & 1.05 \end{bmatrix}$$

$$\underline{S} = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Global matrix at time τ

$$\begin{bmatrix} 0.95 & -0.95 & 0 \\ -1.05 & 2.0 & -0.95 \\ 0 & -1.05 & 1.05 \end{bmatrix} \begin{Bmatrix} c_1 \\ c_2 \\ c_3 \end{Bmatrix}_\tau + \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} \dot{c}_1 \\ \dot{c}_2 \\ \dot{c}_3 \end{Bmatrix}_\tau = \begin{Bmatrix} q_1 \\ q_2 = 0 \\ q_3 \end{Bmatrix}_\tau$$

Assume implicit time stepping with $\lambda = 1.0 \; ; \quad \tau = t + \Delta t$

$$\underline{K} \, \underline{c}_{t+\Delta t} + \underline{S} \, \underline{\dot{c}}_{t+\Delta t} = \underline{q}_{t+\Delta t}$$

$$\underline{K}^* \, \underline{c}_{t+\Delta t} = \underline{q}^*_{t+\Delta t}$$

$$\underline{K}^* = \underline{K} + \frac{1}{\Delta t} \underline{S}$$

$$\underline{q}^*_{t+\Delta t} = \underline{q}_{t+\Delta t} + \frac{1}{\Delta t} \underline{S} \, \underline{c}_t$$

Set boundary conditions $c_1 = 1.0$ ____ $t \geq 0$

   time step ____ $\Delta t = 1.0$

$$\begin{bmatrix} 1.45 & -0.95 & 0 \\ -1.05 & 3.0 & -0.95 \\ 0 & -1.05 & 1.55 \end{bmatrix} \begin{Bmatrix} c_1 = 1.0 \\ c_2 \\ c_3 \end{Bmatrix} = \begin{Bmatrix} q_1^{\,0} + \frac{1}{2}c_1 \\ q_2^{\,0} + c_2 \\ q_3^{\,0} + \frac{1}{2}c_3 \end{Bmatrix}$$

$\underbrace{}_{K^*}$ $\underbrace{}_{q^*}$

Rearrange for boundary conditions

$$\underbrace{\begin{bmatrix} 3.0 & -0.95 \\ -1.05 & 1.55 \end{bmatrix}}_{K^*} \begin{Bmatrix} c_2 \\ c_3 \end{Bmatrix}_{t+\Delta t} = \begin{Bmatrix} c_2 + 1.05 \\ \frac{1}{2}c_3 \end{Bmatrix}_{t}$$

Inverting

$$\underbrace{\begin{bmatrix} .4244 & .2061 \\ .2875 & .8214 \end{bmatrix}}_{K^{*-1}} \begin{Bmatrix} c_2 + 1.05 \\ \frac{1}{2}c_3 \end{Bmatrix}_{t} = \begin{Bmatrix} c_2 \\ c_3 \end{Bmatrix}_{t+\Delta t}$$

Results

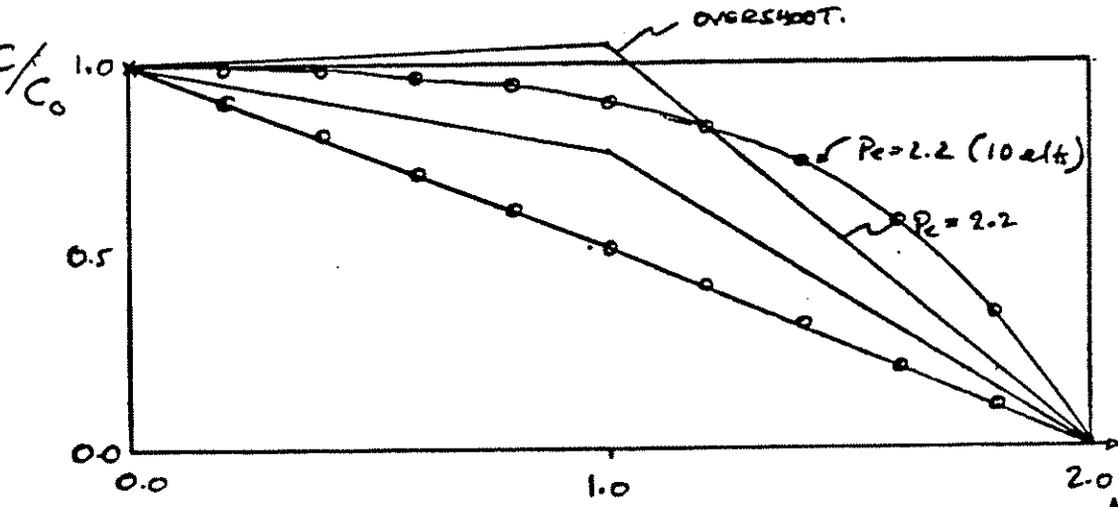| time, t	c_2	c_3
1.0	0.4456	0.3019
2.0	0.6740	0.5540
3.0	0.8037	0.7234
4.0	0.8808	0.8300

etc.

Note

The results converge to $c_2 = c_3 = 1.0$

Better results may be obtained by:
- Using more elements
- Using Crank-Nicolson, $\lambda = \frac{1}{2}$, time stepping and consistent, rather than lumped, approximations for, $\underline{S}$.

CONTAMINANT FLOW EXAMPLES

OVERSHOOT.

C/C_0

$Pe = 2.2 \ (10 \ elt)$

$P_e = 2.2$

STEADY STATE

$C_1 = 1.0$

$C_3 = 0.0$

—— 2 ELT SYSTEM

o 10 ELT SYSTEM

$D = 1.0$

V_z	0.1	1.0	2.2
P_e	0.1	1.0	2.2

NB. $\Delta L = 1.0$ for P_e

P_e & C_e Nos. W.R.T. $\frac{1}{2}$ FLOW DOMAIN LENGTH.

C/C_0

$t = 5.0$

$t = 20.0$

5.0

0.5

TRANSIENT

· $D = 1.0 ; V = 0.1 ; \Delta t =$

o $D = 1.0 ; V = 1.0 ; \Delta t =$

· $P_e = 0.1 \quad C_e = 0.1$

o $P_e = 1.0 \quad C_e = 0.1$

C/C_0

$t = 0.20 ; V = 50 ; P_e = 50 ; C_e = 0.2$

0.50

1.0
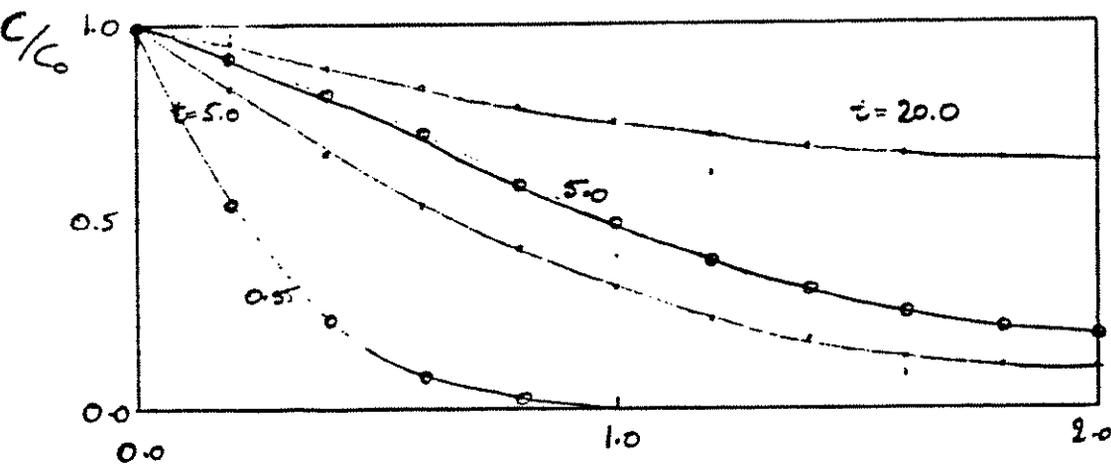
$t = 0.1$
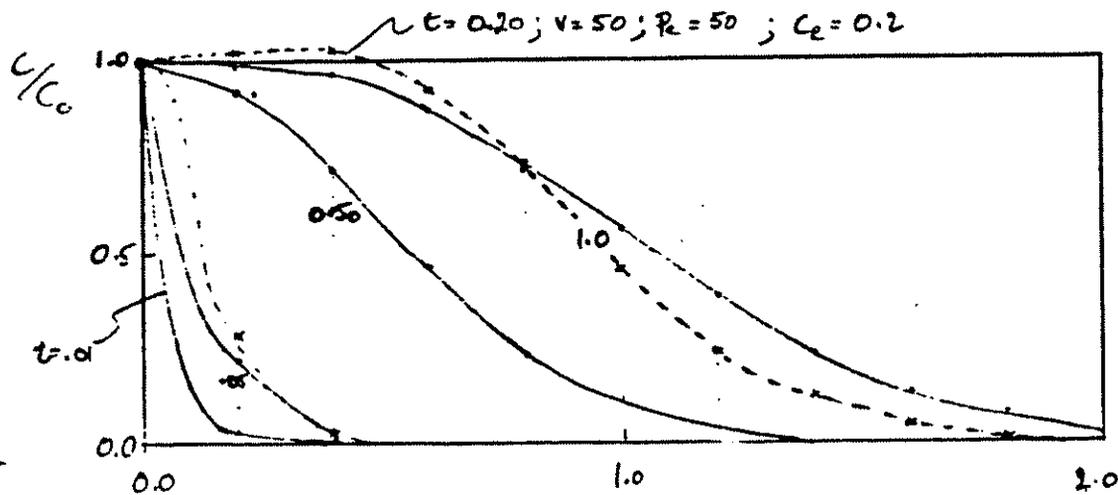
STRONGLY ADVECTIVE TRANSIENT.

$D = 1.0 ; V = 10.0 ; \Delta t = 0.$

$P_e = 10.0$

$C_R = 0.1$

[3:2] Mass Transport

Recap
$$\underline{S}\,\underline{\dot{c}} + [\underline{K}_d + \underline{K}_a]\,\underline{c} = \underline{q} + \underline{R}$$

2D Elements - heuristic

Stability

Upwind-weighting

Numerical dispersion

Example 3.1.4

Evaluate the advective
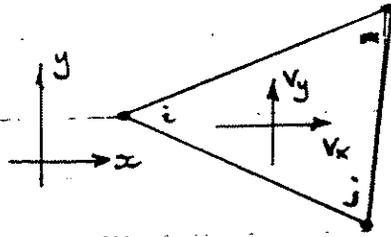matrix for a triangular
element

$$\left\{\begin{array}{l} \dfrac{\partial c}{\partial x} \\ \dfrac{\partial c}{\partial y} \end{array}\right\} \begin{array}{l} c_i = a\,\underline{c} \\ = \underline{a}\,\underline{c} \\ = \underline{a}\,\underline{c} \end{array}$$

$$\underline{K}_a = \int_V \underline{b}^T \underline{v}\,\underline{a}\ dx\,dy$$

$$\underline{v} = [\,v_x \quad v_y\,]$$

$$\underline{a} = \frac{1}{2\Delta}\begin{bmatrix} A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$
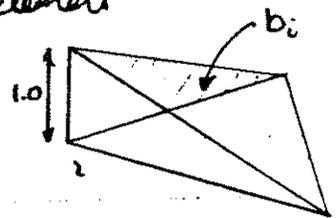
$$\underline{v}\,\underline{a} = \frac{1}{2\Delta}[\,(v_x A_{21} + v_y A_{31})\ ;\ (v_x A_{22} + v_y A_{32})\ ;\ (v_x A_{23} + v_y A_{33})\,]$$

Note $\underline{v}\,\underline{a}$ is constant over the element. The only remaining term
under the integral is

$$\underline{K}_a = \int \underline{b}^T\,dx\,dy\ \underline{v}\,\underline{a}$$

Integrals of the shape functions are merely the volume of a
pyramid of unit height over one vertex of the element

$$\text{Volume} = \frac{1}{3}\Delta \times \text{height}$$

$$\text{and} \quad \int \underline{b}\ dx\,dy = [\,\tfrac{1}{3}\ ;\ \tfrac{1}{3}\ ;\ \tfrac{1}{3}\,]\Delta$$

finally

$$\underline{K}_a = \int \underline{b}^T \underline{v}\,\underline{a}\ dx\,dy = \int \underline{b}^T\,dx\,dy\ \underline{v}\,\underline{a}$$

Note that $\underline{K}_a$ makes the element matrix non-symmetric.

STABILITY REQUIREMENTS - GALERKIN

$$P_e = \frac{\text{Advective flux}}{\text{Diffusive flux}} = \frac{V\ell}{D} \leq 2-10 \qquad (\text{Péclet})$$

$$C_R = \frac{V \Delta t}{\ell} \leq 1 \qquad (\text{Courant})$$

UPWIND WEIGHTING

- Reduced oscillation and overshoot.
- Numerical Dispersion

True solution

Upwind weighted solution
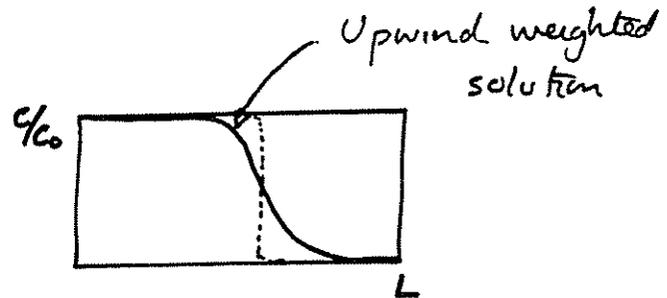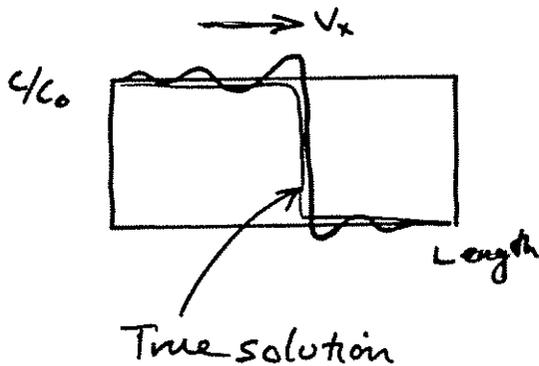
UPWIND WEIGHTING

NORMAL
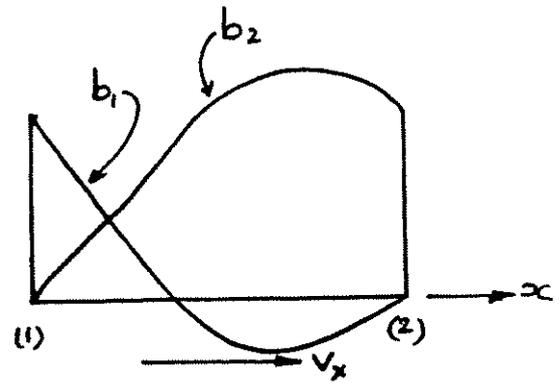
UPWIND WEIGHTED

$$b_1 = 1 - \zeta$$
$$b_2 = \zeta$$
$$\left.\right\} \quad \zeta = \frac{x}{L}$$

$$b_1 = W_1 = 1 - \zeta + 3\beta(\zeta^2 - \zeta)$$
$$b_2 = W_2 = \zeta - 3\beta(\zeta^2 - \zeta)$$

ELEMENT BASED SHAPE FUNCTIONS
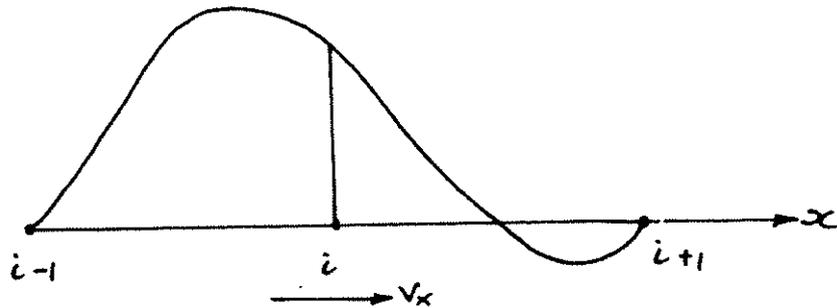
WEIGHTING FOR NODE i

Figure 3.1.1.1 Upwind weighted shape functions

UPWIND WEIGHTED EQUATIONS

$$\underline{W} = \left[\ 1 - \frac{x}{L} + 3\beta\left(\frac{x^2}{L^2} - \frac{x}{L}\right);\ \frac{x}{L} - 3\beta\left(\frac{x^2}{L^2} - \frac{x}{L}\right)\right] \qquad (1)$$

Apply only to Galerkin weighting component and not to spatial component of transport equation. Thus,

$$\int_A \underline{\bar{w}}^T \underline{D}\,\underline{a}\ dx\,dy\ \underline{c} \ + \ \int_A \underline{b}^T\underline{b}\ dx\,dy\ \underline{\dot{c}}$$

$$+ \int_A \underline{w}^T \underline{v}\ \underline{a}\ dx\,dy\ \underline{c}\ =\ \underline{q} \qquad (2)$$

$$\underline{\bar{w}} = \left\{\begin{array}{c} \partial/\partial x \\ \partial/\partial y \end{array}\right\}\ \underline{W} \qquad (3)$$

$$\beta \cong 1 - \frac{2}{Pe} \qquad\qquad Pe \geqslant 2 \qquad\qquad (4)$$
$$\qquad\qquad\qquad\qquad If\ Pe \leqslant 2\ \ \beta \to 0$$

$$Pe = \frac{V_i\,\ell}{D} \qquad\qquad\qquad\qquad (5)$$
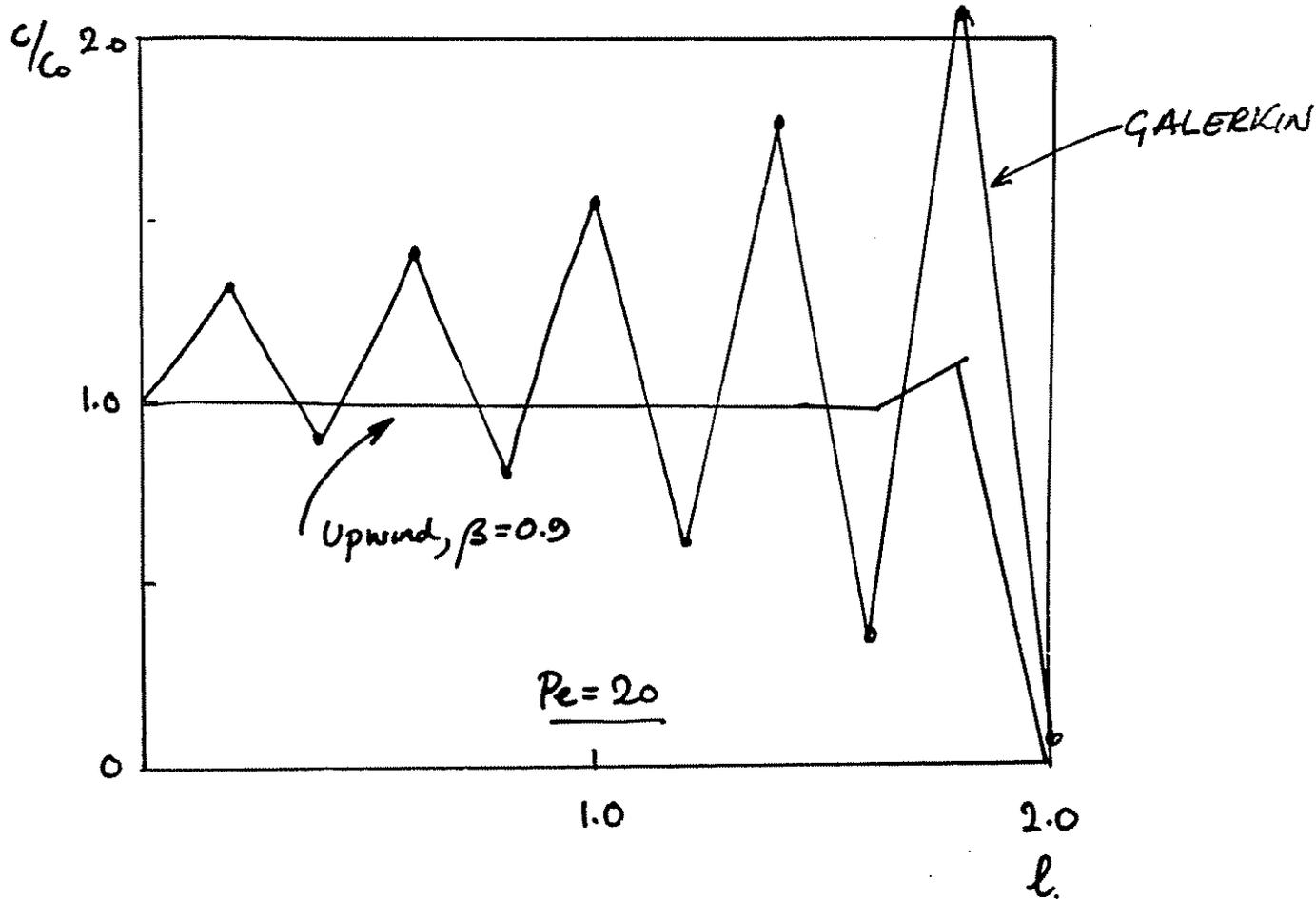
[3:3] Mass Transport

Recap
$$\underline{S}\,\underline{\dot{c}} + [\underline{K}_d + \underline{K}_a]\,\underline{c} = \underline{q} + \underline{R}$$

Reactive transport

   Sorption

   First-order reactions

   Multiple reactions

Reaction Rates

$$\frac{\partial c_i}{\partial t} + \nabla \cdot (c_i \mathbf{u}) - \nabla \cdot (D_i \nabla c_i) = R_i \tag{1}$$

For the reaction:

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C \tag{2}$$

$$
\begin{aligned}
\text{Forward rate} \quad &= \quad k_1[A][B] \\
\text{Reverse rate} \quad &= \quad k_2[C]
\end{aligned}
\tag{3}
$$

At equilibrium: $\qquad$ Forward rate = Reverse rate

$$k_1[A][B] = k_2[C] \tag{4}$$

$$\therefore \quad [A][B] = \frac{k_2}{k_1}[C] \tag{5}$$

For closed system and one mole each of [A] and of [B], with $k_1 = 10$ and $k_2 = 1$, then:

$$\frac{[A][B]}{[C]} = \frac{(1-X)^2}{X} = \frac{1}{10} \tag{6}$$

And $(1-X) = [A] = [B] = 0.916$ and X=$[C] = 0.0839$.

Implementation:

$$
\begin{aligned}
R_A \quad &= \quad -k_1[A][B] \quad + \quad k_2[C] \\
R_B \quad &= \quad -k_1[A][B] \quad + \quad k_2[C] \\
R_C \quad &= \quad +k_1[A][B] \quad - \quad k_2[C]
\end{aligned}
\tag{7}
$$

Generalized:

$$R_i = -k_i{}^f \prod_{j=1}^{N} [c_j{}^f]^{\alpha_j{}^r} + k_i{}^r \prod_{j=1}^{N} [c_j{}^r]^{\alpha_j{}^r} \tag{8}$$

Heats of reaction:

$$H_i = R_i \Delta H_i \tag{9}$$

And heat balance requires:

$$\rho c \frac{\partial T}{\partial t} + \nabla \cdot (T\mathbf{u}) - \nabla \cdot (\lambda \nabla T) = H_i \tag{10}$$

APPROACHES TO REACTIVE TRANSPORT MODELING

1. <u>Pseudo-reactive using linear isotherms</u>
   - Retardation approach → gives chromatographic effect
   - Adsorption of monolayer of components at sorption sites
   - No pore clogging or dissolution
   - Multi-component retardation possible, but no reaction/interaction

2. <u>First order reaction with a single component</u>
   - Reaction rate law defines rate of precipitation/dissolution
   - Dissolution/precipitation may be used to update
     pore volume/surface area/permeability

3. <u>Reaction with multiple reacting components</u>

   - Complex interacting solutes
   - Precipitation/dissolution

General approaches of 2. & 3.
   1. Determine $\partial c/\partial t$ in transport (spatial)
   2. Determine $\partial c/\partial t$ in species due to reaction.
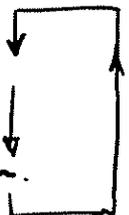
PSEUDO-REACTIVE SYSTEM — "RETARDATION" APPROACH

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2} - \frac{V_x}{n} \frac{\partial c}{\partial x} - \frac{\rho_b}{n} \frac{\partial c^*}{\partial t}$$

concentration on mineral surfaces

Linear isotherm:

(Aqueous) c

K_d

c^* (solid)

$$\frac{\partial c^*}{\partial t} = \frac{\partial c^*}{\partial c} \frac{\partial c}{\partial t} \implies K_d \frac{\partial c}{\partial t}$$

Replacing $\frac{\rho_b}{n} \frac{\partial c^*}{\partial t}$ yields a modified transport equation

$$\underbrace{\left(1 + \frac{\rho_b}{n} K_d\right)}_{R = \text{retardation coefficient}} \frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2} - \frac{V_x}{n} \frac{\partial c}{\partial x}$$

$$\frac{\partial c}{\partial t} = \frac{D}{R} \frac{\partial^2 c}{\partial x^2} - \frac{V_x}{Rn} \frac{\partial c}{\partial x}$$

Retarded at time, t.

Unretarded $R = 1$ at time, t

c/c_0

$l = \frac{V}{n} \cdot t$

$l = \frac{V}{Rn} t$
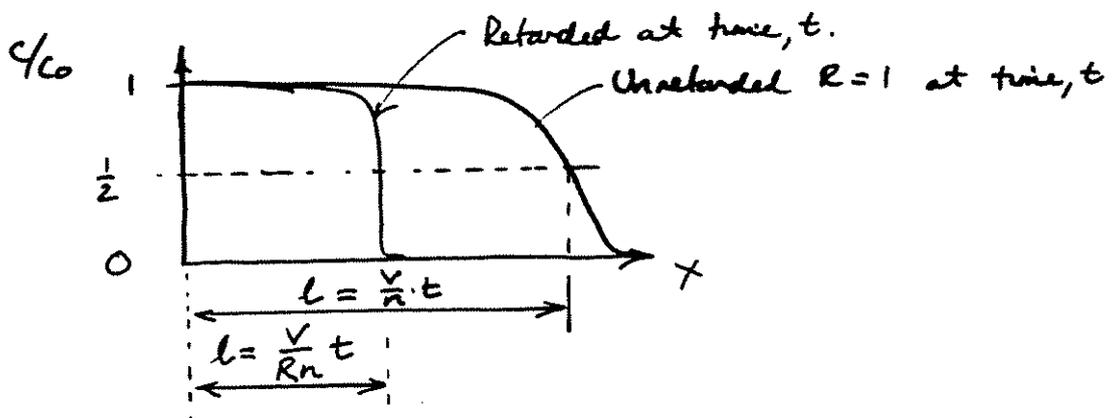
FIRST ORDER REACTION — SINGLE COMPONENT

Dissolution / Precipitation of Quartz

$$H_4 Si O_4 \,(aq) = SiO_2 \,(q) + 2H_2O \,(l) \qquad (1) \qquad \text{Stochiometric reaction}$$
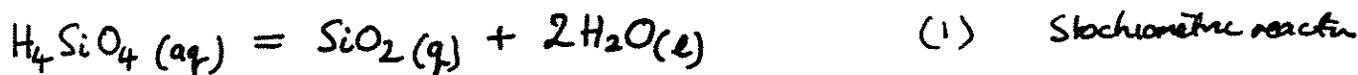
$$\left(\frac{\partial c}{\partial t}\right)_{Rxn} = -k(c - c_{eq}) \qquad (2) \qquad \text{Rate law}$$

equilibrium concentration — no precip./dissol.

Transport Equation (one component) for fluid phase.
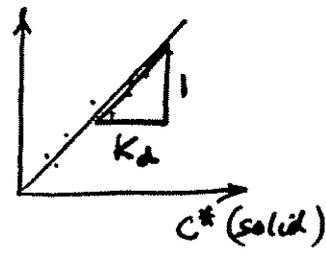
$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2} - \frac{V_x}{n}\frac{\partial c}{\partial x} + \left(\frac{\partial c}{\partial t}\right)_{Rxn}$$

Translate to FE statement:

$$\int_V \underline{b}^T \underline{b}\, d\forall\, \underline{\dot{c}} \;+\; D \int_V \underline{a}^T \underline{a}\, d\forall\, \underline{c} \;+\; \frac{V_x}{n}\int_V \underline{b}^T \underline{a}\, d\forall\, \underline{c} \;-\; k_1 \int \underline{b}^T \underline{b}\, d\forall (\underline{c} - \underline{c}_{eq}) = \underline{q}$$

$$\underline{K}_0\, \underline{\dot{c}} \;+\; D\, \underline{K}_{I\!I}\, \underline{c} \;+\; \frac{V_x}{n}\underline{K}_{I}\, \underline{c} \;-\; k_1\, \underline{K}_0 (\underline{\dot{c}} - \underline{c}_{eq}) = \underline{q}$$

Separate terms for solution (implicit time stepping ($\lambda = 1$)):

$$\left[\frac{1}{\Delta t}\underline{K}_0 + D\underline{K}_{I\!I} + \frac{V_x}{n}\underline{K}_I - k_1\underline{K}_0\right]\underline{c}^{t+\Delta t} = \underline{q}^{t+\Delta t} + \frac{1}{\Delta t}\underline{K}_0\underline{c}^t$$

Mass   + Diffusion + Advection + Dissol./Precip $-k_1\underline{K}_0\,\underline{c}^t$
Accumulation

Solve for $\underline{c}$ with time
These are fluid concentrations of any species (quartz)
Similar to retardation approach.

2.4. THE RATE LAW

We can show experimentally that for the general irreversible reaction

$$A + 2B + \ldots \rightarrow P + 2Q \ldots$$
$$\underbrace{}_{\text{reactants}} \quad \underbrace{}_{\text{products}}$$

we can write the rate law,

$$\frac{d[A]}{dt} = -k[A]^a[B]^b[P]^p[Q]^q \ldots \tag{2-1}$$

where

$\dfrac{d[A]}{dt}$ = time rate of change in molar concentration of species A,

k = reaction rate constant, and

$a, b, p, q, \ldots$ = constants

In this book, [ ] is used to signify concentration in moles/liter. We may use concentration units other than moles/liter in the rate law but in doing so we should use the same concentration unit for each species and realize that both the numerical value and units of the reaction rate constant will differ from those found when molecular concentrations are used.

Using our knowledge of the stoichiometry of the reaction, that is, the relative number of moles of species reacting and the relative number of moles of products being formed as the reaction proceeds, we can state that

$$\frac{d[A]}{dt} = \frac{1}{2}\frac{d[B]}{dt} \ldots = \frac{-d[P]}{dt} = \frac{-1}{2}\frac{d[Q]}{dt} \ldots \tag{2-2}$$

because 1 mole of A reacts for every 2 moles of B that react, and so forth, and 1 mole of P is formed for every mole of A that reacts, and so forth. We can determine the *reaction order* from the rate law. The *overall* reaction order is

$$a + b + p + q \ldots \tag{2-3}$$

while the order *with respect to* A is a, the order with respect to B is b, and so forth. If the reaction is irreversible, then $p, q, \ldots$, the exponents of the product concentration, are usually zero. For example, if

$$\frac{d[A]}{dt} = -k[A][B]^2$$

then we would say that the reaction was first order with respect to A, second order with respect to B, and third order overall. It is important to note that reaction order is generally not determined by the stoichiometry of the overall reaction. Laboratory experimentation is necessary to determine the order.

The following example illustrates several points that are important for a good understanding of the rate law.

Integrated forms of the rate law are very useful for analyzing rate data to determine reaction rate constants and reaction order. Let us first consider the irreversible reaction

$$A \rightarrow \text{products}$$

which has the rate law

$$\frac{d[A]}{dt} = -k[A]^n$$

To determine the behavior of [A] as a function of time, we must integrate the rate expression with respect to time. We will do this for several values of the reaction order, n. When $n = 0$, the reaction is zero order, and

$$\frac{d[A]}{dt} = -k[A]^0 = -k \tag{2-4}$$

Upon integrating, we obtain

$$[A] = [A]_0 - kt \tag{2-5}$$

where $[A]_0$ = the concentration of A at $t = 0$, that is, the initial concentration of A. The *half-life*, $t_{1/2}$, or time for 50 percent of the initial concentration to react can be obtained from Eq. 2-5 by setting $[A] = 0.5 [A]_0$ when $t = t_{1/2}$. Then

$$t_{1/2} = \frac{0.5[A]_0}{k}$$

When $n = 1$, the reaction is *first order*, both with respect to A and overall, and we can write,

$$\frac{d[A]}{dt} = -k[A] \tag{2-6}$$

Rearranging Eq. 2-6 and solving the integral,

$$\int_{[A]_0}^{[A]} \frac{d[A]}{[A]} = -\int_0^t k \, dt$$

we find

$$\ln [A] = \ln [A]_0 - kt \tag{2-7}$$

or

$$[A] = [A]_0 e^{-kt} \tag{2-8}$$

Examination of Eq. 2-7 suggests that the rate constant k may be determined experimentally from a plot of ln [A] versus t, which has a slope of $-k$. Also, from Eq. 2-8, when $[A] = 0.5 [A]_0$, we find the half-life to be

$$t_{1/2} = \frac{0.693}{k}$$

If the reaction is greater than first order, then we can write

$$\frac{d[A]}{dt} = -k[A]^n \tag{2-9}$$

PROCESS COUPLINGS [T-H-M-C]

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & R_{42} & \cdots & R_{44} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{p} \\ \underline{T} \\ \underline{c} \end{Bmatrix} + \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & S_{44} \end{bmatrix} \begin{Bmatrix} \underline{\dot{u}} \\ \underline{\dot{p}} \\ \underline{\dot{T}} \\ \underline{\dot{c}} \end{Bmatrix} = \begin{Bmatrix} \underline{\dot{f}} + \cdots \\ \underline{q}_F + \cdots \\ \underline{q}_T + \cdots \\ \underline{q}_M + \cdots \end{Bmatrix}$$

or

or

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c}{\partial x^2} - \frac{v_x}{n} \frac{\partial c}{\partial x} + \left(\frac{\partial c_i}{\partial t} \right)_{RxN}$$

Reaction with Multiple Reacting Components

See for example: Steefel & MacQuarrie in "Reactive Transport in Porous Media"
Ed. Lichtner, P.C., Steefel, C.I. and Oelkers, E.H.
Mineralogical Soc. of Amer. 1996.

Basic Approach

1. Write a single equation for each species considered, as:

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2} - \frac{v_x}{n} \frac{\partial c_i}{\partial x} + R_i c_i$$

Reaction rate: $R_i c_i = \frac{\partial c_i}{\partial t} = -k_i (c - c_{eq})$

$$\sim \quad \frac{\partial c_i}{\partial t} = -k_i [c]$$

Components are not necessarily independent

2. Solve sequentially as:

→ 1. Solve transport equation → Evaluate $\underline{c}_i$ @ all nodes
with $R_i = 0$

2. Evaluate reaction term

$$\frac{\partial \underline{c}_i}{\partial t} = -k_i (\underline{c}_i - \underline{c}_{eq_i})$$

$$\underline{c}_i^{n+1} \quad \underline{c}_i^n + \Delta t \left(\frac{\partial \underline{c}_i}{\partial t} \right)_{R \times N}$$

Repeat Step ①
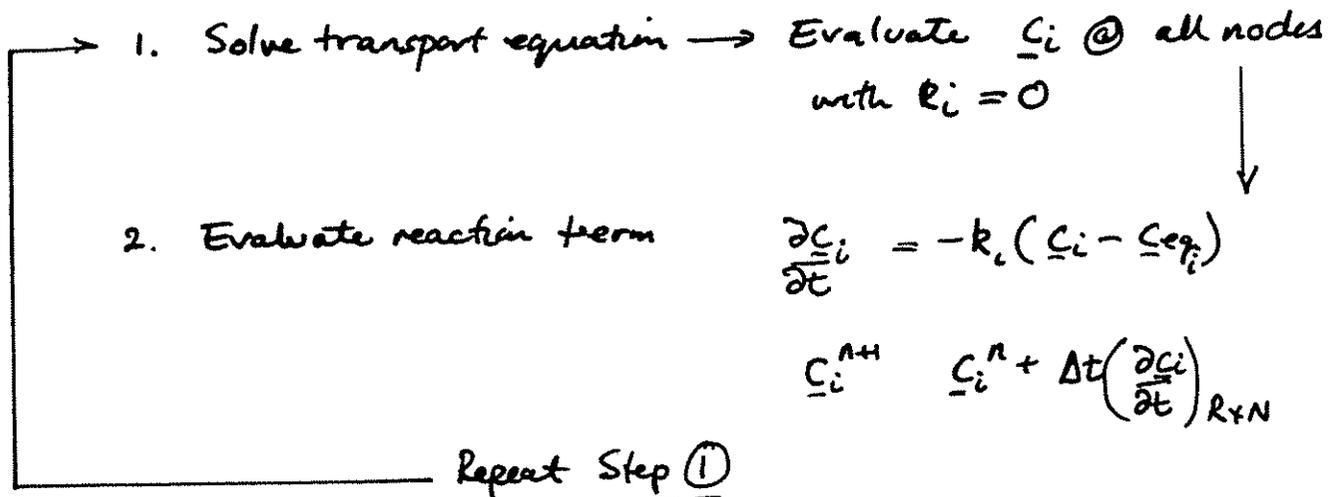
REACTION ALGORITHMS FOR MULTICOMPONENT SYSTEMS

Mathematical descriptions of reaction systems

The multicomponent, multi-species systems typical of those which occur in porous media require some special treatment, both because they involve multiple unknowns and because they are usually nonlinear. The mathematical description used, however, will depend on what form the reactions in the system are assumed to take. It is instructive to derive a general approach to handle multicomponent, multi-species reactive systems. Formulations for arbitrarily complex reaction systems characterized by both equilibrium and non-equilibrium reactions have been presented by Lichtner (1985), Lichtner (this volume), Friedly and Rubin (1992), Sevougian et al. (1993), and Chilakapati (1995). A clear discussion of one possible way of doing so is given by Chilakapati (1995). His approach begins with the most general case, a set of ordinary differential equations for each species in the system and reactions between the species described by kinetic rate laws. A system containing N_{tot} species and N_r reactions can be expressed as
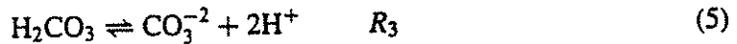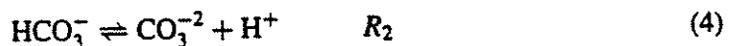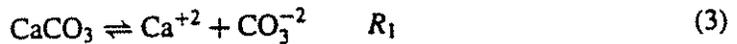
$$\mathbf{I} \cdot \frac{d\mathbf{C}}{dt} = \nu \cdot \mathbf{R}. \tag{1}$$

The raised dot indicates matrix multiplication. $\mathbf{I}$ is the identity matrix of dimension $N_{tot} \times N_{tot}$, $\mathbf{C}$ is the vector of solute concentrations of length N_{tot}, ν is a matrix of dimension $N_{tot} \times N_r$, and $\mathbf{R}$ is a vector of length N_r. For example, the matrix ν and the vector $\mathbf{R}$ have the form

$$\nu = \begin{bmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,N_r} \\ \nu_{2,1} & \nu_{2,2} & \cdots & \nu_{2,N_r} \\ \vdots & \cdots & \cdots & \vdots \\ \nu_{N_{tot},1} & \nu_{N_{tot},2} & \cdots & \nu_{N_{tot},N_r} \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} R_1 \\ R_2 \\ \cdots \\ R_{N_r} \end{bmatrix}. \tag{2}$$

The multiplication of the identity matrix by the derivatives of the individual species concentrations results in an ODE of similar form for each of the species in the system.

As an example, consider an aqueous system consisting of $Ca^{+2}, H^+, OH^-, CO_3^{-2}, HCO_3^-$, H_2CO_3, and $CaCO_3(s)$ (calcite). We ignore H_2O for the sake of conciseness. In addition, we assume that the following reactions occur, without yet specifying whether they are to be considered equilibrium or kinetically-controlled reactions,

$$CaCO_3 \rightleftharpoons Ca^{+2} + CO_3^{-2} \qquad R_1 \tag{3}$$

$$HCO_3^- \rightleftharpoons CO_3^{-2} + H^+ \qquad R_2 \tag{4}$$

$$H_2CO_3 \rightleftharpoons CO_3^{-2} + 2H^+ \qquad R_3 \tag{5}$$

$$H^+ + OH^- \rightleftharpoons H_2O \qquad R_4. \tag{6}$$

In the above equations R_i symbolizes the rate expression for reaction i. We also make no assumptions at this stage about whether the set of reactions included are linearly independent (although the reactions listed above are). We have shown the reactions to be reversible here (thus the symbol $\rightleftharpoons$) but the results below apply whether the reactions are irreversible or reversible since at this stage, one can think of the reaction rates as simply time-dependent expressions of the mole balances inherent in a balanced chemical reaction. The reversibility or lack thereof only determines whether the sign of the reaction rate can change. The term *reversible* is generally used by thermodynamicists to refer to equilibrium reactions (Lichtner, this volume), although we prefer to use it to refer to reactions which are sufficiently close to equilibrium that the backward reaction is important. It is quite possible in a steady-state flow system, for example, for backward reactions to be important and yet not to be at equilibrium (e.g. Nagy et

al., 1991; Nagy and Lasaga, 1992; Burch et al., 1993). According to this definition, the term *irreversible* is used for those reactions which proceed in only one direction (i.e. those that can be represented with a unidirectional arrow, $\longrightarrow$).

For our example aqueous system, the rates for the individual species can be written

$$\frac{d[H_2CO_3]}{dt} = -R_3 \tag{7}$$

$$\frac{d[HCO_3^-]}{dt} = -R_2 \tag{8}$$

$$\frac{d[CaCO_3]}{dt} = -R_1 \tag{9}$$

$$\frac{d[OH^-]}{dt} = -R_4 \tag{10}$$

$$\frac{d[H^+]}{dt} = R_2 + 2R_3 - R_4 \tag{11}$$

$$\frac{d[Ca^{+2}]}{dt} = R_1 \tag{12}$$

$$\frac{d[CO_3^{-2}]}{dt} = R_1 + R_2 + R_3. \tag{13}$$

In matrix form the system of equations becomes

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\frac{d[H_2CO_3]}{dt} \\
\frac{d[HCO_3^-]}{dt} \\
\frac{d[CaCO_3]}{dt} \\
\frac{d[OH^-]}{dt} \\
\frac{d[H^+]}{dt} \\
\frac{d[Ca^{+2}]}{dt} \\
\frac{d[CO_3^{-2}]}{dt}
\end{bmatrix}
=
\begin{bmatrix}
0 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 \\
0 & 1 & 2 & -1 \\
1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
R_1 \\
R_2 \\
R_3 \\
R_4
\end{bmatrix}
\tag{14}
$$

As written in Equation (14), the stoichiometric reaction matrix, ν, is referred to as being in *canonical* form (Smith and Missen, 1982; Lichtner, 1985; Lichtner, this volume). The system of equations is partitioned into the first four rows where the associated species (H_2CO_3, HCO_3^-, $CaCO_3(s)$, and OH^-) are involved in only one reaction while in the remaining three rows the species are involved in multiple reactions. The first four species are referred to as secondary or non-component species, while the last three are the primary or component species (Lichtner, this volume). These are also referred to as *basis* species because they form a basis which spans the concentration space. In this example, we have written all of the carbonate reactions using the species CO_3^{-2} precisely so as to restrict all of the other carbonate species to involvement in a single reaction. This is an essential first step in obtaining either the canonical formulation (Lichtner, 1985; Lichtner, this volume) or to writing the reactions in *tableaux* form (Morel and Hering, 1993), both of which assume that one is dealing with a set of linearly independent reactions, but it is not essential for what follows below. The procedure will also work if, for example, the formation of H_2CO_3 involved H^+ and HCO_3^- rather than 2 H^+ and CO_3^{-2}, although we will not obtain the conserved quantities (total H^-, total CO_3^{-2}, etc.) found in the tableaux method without additional manipulations.

The system of ODEs could be solved directly in the form of Equation (14) if the reactions are all described with kinetic rate laws. Alternatively, one can apply a Gauss-Jordan elimination

process to the matrix v and simultaneously to the identity matrix I until there are no pivots left (Chilakapati, 1995). The resulting transformed set of ODEs is now

$$M \cdot \frac{dC}{dt} = v^* \cdot R \tag{15}$$

which partitions the system of equations into N_r ODEs associated with reactions and N_c conservation laws with zero right-hand sides (i.e. no associated reactions). The number of conservation laws or mole balance equations is equal to

$$N_c = N_{tot} - \text{rank of } v = N_{tot} - N_r. \tag{16}$$

N_r, therefore, refers to the number of *linearly independent* reactions between the species in the system. For the sake of clarity, we make the first N_r rows of the matrix M the ODEs with associated reactions and the next N_c rows the conservation equations, so that the left hand of Equation (15) takes the form

$$
\begin{bmatrix}
M_{1.1} & \cdots & M_{1,N_r+N_c} \\
\vdots & \cdots & \vdots \\
M_{N_r.1} & \cdots & M_{N_r,N_r+N_c} \\
M_{N_r+1.1} & \cdots & M_{N_r+1,N_r} \\
\vdots & \cdots & \vdots \\
M_{N_r+N_c.1} & \cdots & M_{N_r+N_c,N_r+N_c}
\end{bmatrix}
\cdot
\begin{bmatrix}
\frac{dC_1}{dt} \\
\vdots \\
\frac{dC_{N_r}}{dt} \\
\vdots \\
\frac{dC_{N_r+N_c}}{dt}
\end{bmatrix}
\tag{17}
$$

In our example, the Gauss-Jordan elimination is carried out on the the matrix v on the right hand side of Equation (14) and the same row transformations are applied to the identity matrix, I, yielding

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
2 & 1 & 0 & -1 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\frac{d[H_2CO_3]}{dt} \\
\frac{d[HCO_3^-]}{dt} \\
\frac{d[CaCO_3]}{dt} \\
\frac{d[OH^-]}{dt} \\
\frac{d[H^+]}{dt} \\
\frac{d[Ca^{+2}]}{dt} \\
\frac{d[CO_3^{-2}]}{dt}
\end{bmatrix}
=
\begin{bmatrix}
0 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
R_1 \\
R_2 \\
R_3 \\
R_4
\end{bmatrix}
\tag{18}
$$

The stoichiometric reaction matrix, v^*, now consists of a nonsingular 4 by 4 matrix (N_r by N_r) and three rows of zeros corresponding to the N_c conservation equations. Writing out the ODEs in Equation (18), we find

$$\frac{d[H_2CO_3]}{dt} = -R_3 \tag{19}$$

$$\frac{d[HCO_3^-]}{dt} = -R_2 \tag{20}$$

$$\frac{d[CaCO_3]}{dt} = -R_1 \tag{21}$$

$$\frac{d[OH^-]}{dt} = -R_4 \tag{22}$$

Table 1. Tableaux for carbonate system, neglecting H_2O as a species and component.

		Components		
		H^+	Ca^{+2}	CO_3^{-2}
Species	H_2CO_3	2		1
	HCO_3^-	1		1
	$CaCO_3$		1	1
	OH^-	-1		
	H^+	1		
	Ca^{+2}		1	
	CO_3^{-2}			1

and

$$\frac{d}{dt}\left([H^+] + 2[H_2CO_3] + [HCO_3^-] - [OH^-]\right) = 0 \tag{23}$$

$$\frac{d}{dt}\left([Ca^{+2}] + [CaCO_3]\right) = 0 \tag{24}$$

$$\frac{d}{dt}\left([CO_3^{-2}] + [H_2CO_3] + [HCO_3^-] + [CaCO_3]\right) = 0. \tag{25}$$

From the example, it is apparent that we have eliminated the reactions in the equations originally corresponding to the species H^+, Ca^{+2}, and CO_3^{-2} by making use of the relations in the first four equations. The last three equations are mole balances for *total* H^+, Ca^{+2}, and CO_3^{-2}

$$TOTH^+ = [H^+] - [OH^-] + [HCO_3^-] + 2[H_2CO_3] \tag{26}$$

$$TOTCa^{+2} = [Ca^{+2}] + [CaCO_3] \tag{27}$$

$$TOTCO_3^{-2} = [CO_3^{-2}] + [H_2CO_3] + [HCO_3^-] + [CaCO_3]. \tag{28}$$

Note that the canonical form of the stoichiometric reaction matrix is identical to the *tableaux* form popularized by Morel and coworkers (Morel and Hering, 1993; Dzombak and Morel, 1990). By transposing the last three rows of the matrix **M** in Equation (17), we can write the matrix in tableaux form (Table 1).

The procedure has yielded expressions for the total concentrations of the N_c *primary* or *component* species. A more general form is given by

$$TOT_j = C_j + \sum_{i=1}^{N_r} v_{i,j} X_i \tag{29}$$

where C_j and X_i refer to the concentration of the primary and secondary species respectively. Note that the number of secondary species is equal to N_r, the number of linearly independent reactions in the system (i.e. the rank of the matrix v). Equation (27) and Equation (28) are recognizable as the total concentrations of calcium and carbonate respectively. The total concentration of H^+ is written in exactly the same form as the other equations, although its physical meaning is less clear because it may take on negative values due to the negative stoichiometric coefficients in the expression. The mole balance equation for total H^+ is just the *proton condition* equation referred to in many aquatic chemistry textbooks. Oxidation-reduction reactions are also easily handled with this method. If the redox reactions are written as whole cell reactions, there is no need in any application not involving an electrical current

(see Lichtner, this volume) to introduce the electron as an unknown. Writing the reactions as whole cell reactions allows redox reactions to be treated exactly like any other reaction.

CONVECTIVE HEAT FLOW

Assume — Thermal equilibrium of fluid and rock ∴ $T_R = T_F$

and only 1 variable

— Not necessary to make this assumption (convenient).

$$D^* \frac{\partial^2 T}{\partial x^2} - \rho_f c_f q_D \frac{\partial c}{\partial x} = (\overline{\rho c}) \frac{\partial T}{\partial t} \qquad \overset{\rho_f c_f n + (1-n)\rho_r c_r}{}$$

$D^* = D_r(1-n) + D_f(n)$

Darcy flux $\quad q_D = -\frac{k}{\mu} \frac{dp}{dx}$

FE Equations:

$$D^* \frac{\partial^2 T}{\partial x^2} \quad = \quad D^* \int_V \underline{a}^T \underline{a} \, dV \, \underline{T} \qquad = \quad \underline{K}_2 \, \underline{T}$$

$$\rho_f c_f q_D \frac{\partial T}{\partial x} \quad = \quad \rho_f c_f q_D \int_V \underline{b}^T \underline{a} \, dV \, \underline{T} \qquad = \quad \underline{K}_1 \, \underline{T}$$

$$\overline{\rho c} \frac{\partial T}{\partial t} \quad = \quad \overline{\rho c} \int_V \underline{b}^T \underline{b} \, dV \, \underline{\dot{T}} \qquad = \quad \underline{K}_0 \, \underline{\dot{T}}$$

ASSEMBLED EQUATIONS: (Implicit, $\lambda = 1$)

$$[\underline{K}_2 + \underline{K}_1] \underline{T}^{t+\Delta t} + \underline{K}_0 \underline{\dot{T}}^{t+\Delta t} = \underline{q}^{t+\Delta t}$$

$$\underline{\dot{T}} = \frac{1}{\Delta t} [\underline{T}^{t+\Delta t} - \underline{T}^t]$$

GATHERING TERMS:

$$\left[\underline{K}_2 + \underline{K}_1 + \frac{1}{\Delta t} \underline{K}_0\right] \underline{T}^{t+\Delta t} = \underline{q}^{t+\Delta t} + \frac{1}{\Delta t} \underline{K}_0 \underline{T}^t$$

— may change if q_D changes in <u>time</u>.

SYSTEM TYPES

SOLID MECHANICS	FLOW SYSTEM	TRANSPORT

- Conservation of momentum:
  (Equilibrium), $VW_I = VW_E$

- Conservation of mass:
  $$\underline{\nabla}^T \underline{q} = 0$$

- Conservation of mass
  $$\underline{\nabla}^T \underline{q} = 0$$

- Continuity (Compatibility):
  $$\underline{\varepsilon} = \underline{a}\,\underline{u}$$

- Continuity: $\underline{h}_, = \underline{a}\,\underline{h}$

- Continuity: $\underline{c}_, = \underline{a}\,\underline{c}$

- Constitutive relation: $\underline{\sigma} = \underline{D}\,\underline{\varepsilon}$

- Constitutive reln. $\underline{v} = \underline{D}\underline{h}_,$

- Constitutive:
  diffusion — $\underline{v}_1 = \underline{D}\underline{c}_,$
  advective — $\underline{v}_2 = \underline{A}\underline{c}$

- Initial Conditions

- ICs

- ICs

- Boundary Conditions

- BCs

- BCs

— SOLVE SYSTEM EQUATIONS —

```
          SUBROUTINE ELMT05(D,UL,XL,IX,TL,S,P,NDF,NDM,NST,ISW)
          IMPLICIT REAL*8(A-H,O-Z)
C
C.... THREE NODED CONSTANT GRADIENT TRANSPORT ELEMENT
C
C       USER INFORMATION
C
C       INPUT
C
C         VAR       FORMAT            DESCRIPTION
C         ------------------------------------------------------------
C
C         D(1)    F10.0             DIFFUSIVITY OR DISPERSION
C         D(2)    F10.0             HYDRAULIC COND/POROSITY
C----------------------------------------------------------------------
C
C       LOCAL NODAL NUMBERING MUST BE COUNTER-CLOCKWISE
C
C----------------------------------------------------------------------
C
C       VARIABLES
C
C       NEL    -     NUMBER OF NODES PER ELEMENT
C       NDF    -     NUMBER OF DEGREES OF FREEDOM PER NODE
C       NST    -     NUMBER OF DEGREES OF FREEDOM PER ELEMENT (NEN*NDF)
C       ISW    -     FUNCTION CALL NO.
C                    1 = READ ELEMENT SPECIFIC INPUT DATA
C                    2 = PERFORM MESH CHECK
C                    3 = FORM ELEMENT STIFFNESS MATRIX      - TANG
C                    4 = EVALUATE ELEMENT STRESSES          - STRE
C                    5 = FORM CONSISTENT/LUMPED MASS MATRIX - CMAS/LMAS
C                    6 = FORM LOAD VECTOR                   - FORM
C                        OR EVALUATE NODAL FORCES           - REAC
C
C       ARRAYS - GIVEN
C
C       UL(1,J)    SPECIFIED HEAD BOUNDARY CONDITION FOR
C                  DEGREE OF FREEDOM J (J=1,3)
C       XL(I,J)    COORDINATE IN THE I DIRECTION AT NODE J
C                  EG. XL(1,3) IS X COORDINATE OF NODE K
C       TL(J)      TEMPERATURE OR HEAD AT NODE J
C
C
C       ARRAYS - EVALUATED
C
C       A(  )      A MATRIX
C       C(  )      D MATRIX
C       S(I,J)     CONDUCTANCE MATRIX S = AT*D*A DV
C                                        + BT*V*A DV
C                  FOR ROW (VERTICAL) I AND COLUMN (HORIZ.) J
C       P(I)       MODIFIED LOAD VECTOR FOR LOCAL DOF I (IGNORE)
C
C
C       FOR LMAS CALCULATION THE VECTOR LOCATIONS P(1), P(2), P(3)
C       ARE USED FOR THE STORAGE VECTOR
C
C
```

```
C
C----------------------------------------------------------------
      CHARACTER*4 O,HEAD
      COMMON  /CDATA/ O,HEAD(20),NUMNP,NUMEL,NUMMAT,NEN,NEQ,IPR
      COMMON  /ELDATA/ DM,N,MA,MCT,IEL,NEL
      DIMENSION D(2),UL(1,1),XL(NDM,1),IX(1),TL(1),S(NST,1),P(1)
     1      ,A(2,3),C(2,2)
C.... GO TO CORRECT ARRAY PROCESSOR
      GO TO(1,2,3,4,5,3),ISW
C.... INPUT MATERIAL PROPERTIES
1     READ (5,1000) D(1),D(2)
      WRITE(6,2000) D(1),D(2)
      RETURN
C.... MESH CHECKING FACILITY
2     RETURN
C.... DIFFUSIVE-ADVECTIVE MATRIX COMPUTATION
C
C     DIFFUSIVE MATRIX COMPONENTS
C.... EVALUATE COEFFICIENTS IN A( ) MATRIX
C     AND PLACE IN A(2,3) ARRAY
3     CONTINUE
C
C
C.... EVALUATE VX AND VY FROM NODAL HEADS TL(I)
C
C
C.... EVALUATE CONSTITUTIVE MATRIX. THIS IS THE
C     D( ) MATRIX IN YOUR NOTES AND THE C(2,2) ARRAY.
C
C
C.... COMPLETE TRIPLE MATRIX PRODUCT  AT*D*A AND STORE
C     THE PRODUCT IN THE S(3,3) ARRAY.
C
C
C.... PERFORM VOLUME INTEGRATION (*AREA)
C     AND MULTIPLY TERMS OF THE S(3,3) MATRIX BY
C     AREA.
C
C
C.... EVALUATE
C.... ADVECTIVE MATRIX COMPONENTS
C.... SUBSTITUTE TERMS FOR ADVECTIVE FLUX
C
C.... CALCULATE ADVECTIVE VELOCITIES (VX AND VY) FROM
C     THE NODAL HEADS (STORED IN THE TL( ) ARRAY) AND
C     THE HYDRAULIC CONDUCTIVITY
C
C.... EVALUATE ADDITIONAL TERMS OF THE S(3,3) ARRAY
C     DUE TO ADVECTION.
C
C     THIS IS THE END OF YOUR ADDITIONS. RELAX.
C
C.... MODIFY LOAD VECTOR FOR BOUNDARY CONDITIONS
      DO 325 I=1,3
      DO 325 J=1,3
325   P(I) = P(I) - S(I,J)*UL(1,J)
      RETURN
```

```
C.... NOT USED
4      RETURN
C.... LUMPED AND CONSISTENT MASS COMPUTATION
5      B11 = XL(1,2)*XL(2,3) - XL(1,3)*XL(2,2)
       B21 = XL(2,2) - XL(2,3)
       B31 = XL(1,3) - XL(1,2)
       D2  = XL(1,1)*B21 + XL(2,1)*B31 + B11
       D2 = D2/2.
       DO 500 I=1,3
       DO 510 J=1,3
 510   S(I,J) = D2/12.
       S(I,I) = S(I,I) + D2/12.
 500   P(I) = D2/3.
       RETURN
C.... FORMATS FOR INPUT AND OUTPUT
1000   FORMAT(2F10.0)
2000   FORMAT(/5X,'THREE NODED TRANSPORT ELEMENT       ',//
      1 10X,'DIFFUSIVITY OR DISPERSION ',6X,E14.7,/
      2 10X,'HYDRAULIC COND/POROSITY   ',6X,E14.7,/)
       END
```

```
C
C..............................................................ELMT06
C
      SUBROUTINE ELMT06(D,UL,XL,IX,TL,S,P,NDF,NDM,NST,ISW)
      IMPLICIT REAL*8(A-H,O-Z)
C----------------------------------------------------------------------
C     TWO DIMENSIONAL MASS TRANSPORT ELEMENT
C
C     WITH UPWIND WEIGHTING
C
C
C----------------------------------------------------------------------
      CHARACTER*4 O,HEAD
      COMMON  /CDATA/ O,HEAD(20),NUMNP,NUMEL,NUMMAT,NEN,NEQ,IPR
      COMMON  /ELDATA/ DM,N,MA,MCT,IEL,NEL
      DIMENSION D(10),UL(1,1),XL(NDM,1),IX(4),TL(4),S(NST,1),P(8)
     1          ,SS(4,4),SC(4,4),PS(4)
C.... SET INITIAL PARAMETERS
      PI = 3.141592654
      XBAR = DABS(0.25*(XL(1,1)+XL(1,2)+XL(1,3)+XL(1,4)))
      RAD  = XBAR*2.*PI
C.... IF NOT AXISYMMETRIC
      RAD = 1.0
C
C.... CHECK DIMENSION OF INTERNAL ARRAYS
C
C.... GO TO CORRECT ARRAY PROCESSOR
      GO TO(1,2,3,2,5,3),ISW
C.... INPUT MATERIAL PROPERTIES
1     READ(5,1000) D(1),D(2)
      WRITE(6,2000) D(1),D(2)
      RETURN
2     RETURN
C.... FORM CONDUCTANCE MATRICES
3     CONTINUE
C.... EVALUATE VELOCITIES VX AND VY FROM NODAL HEADS TL(I)
      X11 = XL(1,2) - XL(1,1)
      X12 = XL(2,2) - XL(2,1)
      DL  = DSQRT(X11*X11+X12*X12)
      X21 = XL(1,3) - XL(1,2)
      X22 = XL(2,3) - XL(2,2)
      DB  = DSQRT(X21*X21+X22*X22)
      DH11 = TL(2) - TL(1)
      DH12 = TL(3) - TL(4)
      DH21 = TL(3) - TL(2)
      DH22 = TL(4) - TL(1)
      V11  =-D(2)*DH11/DL
      V12  =-D(2)*DH12/DL
      V21  =-D(2)*DH21/DB
      V22  =-D(2)*DH22/DB
      VX   = (V11+V12)/2.
      VY   = (V21+V22)/2.
C**************
      WRITE(6,2222) VX,VY
2222  FORMAT( 'VX,VY ',2E16.6 ,/ )
C***************8
C.... ZERO MATRIX
```

```
      DO 300 I=1,4
      DO 300 J=1,4
 300  S(I,J) = 0.0
C.... SET OPTIMUM (CRITICAL) PECLET NUMBER
      COEF  = 1./D(1)
      PEC11 = COEF*DL*DABS(V11)
      PEC12 = COEF*DL*DABS(V12)
      PEC21 = COEF*DB*DABS(V21)
      PEC22 = COEF*DB*DABS(V22)
C
      IF(PEC11.LE.2.) GO TO 301
      A11 = DABS(1.-2./PEC11)*V11/DABS(V11)
      GO TO 302
 301  A11 = 0.0
C
 302  IF(PEC12.LE.2.) GO TO 303
      A12 = DABS(1.-2./PEC12)*V12/DABS(V12)
      GO TO 304
 303  A12 = 0.0
C
 304  IF(PEC21.LE.2.) GO TO 305
      A21 = DABS(1.-2./PEC21)*V21/DABS(V21)
      GO TO 306
 305  A21 = 0.0
C
 306  IF(PEC22.LE.2.) GO TO 307
      A22 = DABS(1.-2./PEC22)*V22/DABS(V22)
      GO TO 308
 307  A22 = 0.0
 308  CONTINUE
C*****************
      WRITE(6,444) PEC11,PEC22,PEC12,PEC21
  444 FORMAT ( '11,22,12,21 ',4E12.2 )
C*******************
C.... EVALUATE SYSTEM MATRICES
C.... FORM DIFFUSION MATRIX
      CALL MAT4(SS,DL,DB,0.,0.,0.,0.,D(1),0.,1)
      DO 310 I=1,4
      DO 310 J=1,4
 310  S(I,J)  =  S(I,J)  + SS(I,J)
C.... FORM ADVECTIVE MATRIX
      CALL MAT4(SS,DL,DB,A11,A12,A21,A22,VX,VY,2)
      DO 311 I=1,4
      DO 311 J=1,4
 311  S(I,J)  =  S(I,J)  + SS(I,J)
C.... FOR RADIAL FLOW
      DO 330 I=1,4
      DO 330 J=1,4
 330  S(I,J) = S(I,J)*RAD
C.... REARRANGE FOR BOUNDARY CONDITIONS
      DO 400 I=1,4
      DO 400 J=1,4
 400  P(I) = P(I) - S(I,J)*UL(1,J)
      RETURN
C.... EVALUATE CONSISTENT MASS APPROXIMATIONS
5     X11 = XL(1,2) - XL(1,1)
      X12 = XL(2,2) - XL(2,1)
```

```
      DL  = DSQRT(X11*X11+X12*X12)
      X21 = XL(1,3) - XL(1,2)
      X22 = XL(2,3) - XL(2,2)
      DB  = DSQRT(X21*X21+X22*X22)
      CALL MAT4(SS,DL,DB,0.,0.,0.,0.,0.,0.,3)
C.... FOR RADIAL FLOW
      DO 510 I=1,4
      DO 510 J=1,4
 510  SS(I,J) = SS(I,J)*RAD
C.... LUMP CONSISTENT MATRICES
      DO 556 I=1,4
      SUM1 = 0.0
      DO 555 J=1,4
      SUM1 = SUM1 + SS(I,J)
 555  SS(I,J) = 0.0
      SS(I,I) = SUM1
 556  P(I) = SUM1
      RETURN
C.... FORMAT STATEMENTS
 1000 FORMAT( 2F10.0   )
 2000 FORMAT( 'HYDRAULIC DISPERSIVITY/DIFFUSION     ',E18.5,/
     .          'HYDRAULIC CONDUCTIVITY/POROSITY      ',E18.5,/ )
      END
C
C----------------------------------------------- MAT4
C
      SUBROUTINE MAT4(SS,DL,DB,ALPHA1,ALPHA2,BETA1,BETA2,D1,D2,ISW)
      IMPLICIT REAL*8(A-H,O-Z)
C-------------------------------------------------------
C     TO EVALUATE CLOSED FORM COEFFICIENT MATRICES FOR
C     AN UPWIND WEIGHTED FOUR-NODED RECTANGULAR ELEMENT
C
C     SWITCH PARAMETERS
C                     ISW = 1   FORM DIFFUSION MATRIX
C                     ISW = 2   FORM ADVECTION MATRIX
C                     ISW = 3   FORM CONSISTENT MASS
C
C-------------------------------------------------------
      DIMENSION SS(4,4),SC(4,4)
      DATA SC/4.,2.,1.,2.,2.,4.,2.,1.,
     .        1.,2.,4.,2.,2.,1.,2.,4./
C.... GO TO CORRECT PROCESSOR
      GO TO(1,2,3),ISW
C.... FORM DIFFUSION MATRIX
 1    DB2 = DB*DB
      DL2 = DL*DL
      E1 = D1/(6.*DB*DL)
      A1 =  (DL2-2.*DB2)*E1
      A2 = -(DB2+DL2)*E1
      A3 =  (DB2-2.*DL2)*E1
      SS(1,1) = -2.*A2
      SS(1,2) = A1
      SS(1,3) = A2
      SS(1,4) = A3
      SS(2,2) = -2.*A2
      SS(2,3) = A3
      SS(2,4) = A2
```

```
      SS(3,3) = -2.*A2
      SS(3,4) = A1
      SS(4,4) = -2.*A2
      DO 100 J=1,4
      DO 100 I=J,4
 100  SS(I,J) = SS(J,I)
      RETURN
C.... FORM ADVECTION MATRIX
 2    CONTINUE
      C1 = -D2/DB
      C2 = -D1/DL
      DO 200 I=1,4
      GO TO(21,22,23,24),I
C.... PARAMETERS FOR FIRST ROW
 21   AA =  3.*ALPHA1*(-DL/12.)
      BB =  3.*BETA2 *(-DB/12.)
      GO TO 25
C.... PARAMETERS FOR SECOND ROW
 22   AA = -3.*ALPHA1*(-DL/12.)
      BB =  3.*BETA1 *(-DB/12.)
      GO TO 25
C.... PARAMETERS FOR THIRD ROW
 23   AA = -3.*ALPHA2*(-DL/12.)
      BB = -3.*BETA1 *(-DB/12.)
      GO TO 25
C.... PARAMETERS FOR FOURTH ROW
 24   AA =  3.*ALPHA2*(-DL/12.)
      BB = -3.*BETA2 *(-DB/12.)
 25   CONTINUE
C.... EVALUATE OVERALL PARAMETERS
      A1 = DL/3. + AA
      A2 = DL/6. + AA
      A3 = DB/2. + 2.*BB
      B1 = DB/3. + BB
      B2 = DB/6. + BB
      B3 = DL/2. + 2.*AA
C.... FORM MATRIX BY ROW
      GO TO(26,27,28,29),I
C.... FIRST ROW
 26   SS(1,1) =  C1*A1*A3 + C2*B1*B3
      SS(1,2) =  C1*A2*A3 - C2*B1*B3
      SS(1,3) = -C1*A2*A3 - C2*B2*B3
      SS(1,4) = -C1*A1*A3 + C2*B2*B3
      GO TO 200
C.... SECOND ROW
 27   SS(2,1) =  C1*A2*A3 + C2*B1*B3
      SS(2,2) =  C1*A1*A3 - C2*B1*B3
      SS(2,3) = -C1*A1*A3 - C2*B2*B3
      SS(2,4) = -C1*A2*A3 + C2*B2*B3
      GO TO 200
C.... THIRD ROW
 28   SS(3,1) =  C1*A2*A3 + C2*B2*B3
      SS(3,2) =  C1*A1*A3 - C2*B2*B3
      SS(3,3) = -C1*A1*A3 - C2*B1*B3
      SS(3,4) = -C1*A2*A3 + C2*B1*B3
      GO TO 200
C.... FOURTH ROW
```

```
 29    SS(4,1) =  C1*A1*A3 + C2*B2*B3
       SS(4,2) =  C1*A2*A3 - C2*B2*B3
       SS(4,3) = -C1*A2*A3 - C2*B1*B3
       SS(4,4) = -C1*A1*A3 + C2*B1*B3
200    CONTINUE
       RETURN
C.... FORM CONSISTENT MASS MATRIX
  3    COEF = DL*DB/36.
       DO 300 I=1,4
       DO 300 J=1,4
300    SS(I,J) = COEF*SC(I,J)
       RETURN
       END
```

```
FEAP    SIX TRIANGULAR ELEMENTS-TRANSPORT-STEADY
    8   6   1   2   1   3
COORD
    1   2     0.0       0.0
    7   0     3.0       0.0
    2   2     0.0       1.0
    8   0     3.0       1.0

ELEM
    1   1   1   4   2
    2   1   1   3   4
    3   1   3   5   4
    4   1   5   6   4
    5   1   5   8   6
    6   1   5   7   8

MATE
    1   5           MATERIAL 1
      1.0       1.0

TEMP
    1           0.0
    2           0.0
    3           0.1
    4           0.1
    5           0.2
    6           0.2
    7           0.3
    8           0.3

BOUN
    1           1
    2           1
    7           1
    8           1

FORC
    1           1.0
    2           1.0
    7           0.0
    8           0.0

END
MACR
UTAN
FORM
SOLV
DISP
END
STOP
```

```
FEAP    SIX TRIANGULAR ELEMENTS-TRANSPORT-TRANSIENT
    8   6   1   2   1   3
COORD
    1   2      0.0         0.0
    7   0      3.0         0.0
    2   2      0.0         1.0
    8   0      3.0         1.0

ELEM
    1   1   1   4   2
    2   1   1   3   4
    3   1   3   5   4
    4   1   5   6   4
    5   1   5   8   6
    6   1   5   7   8

MATE
    1   5           MATERIAL 1
       1.0        1.0        0.0

TEMP
    1           0.0
    2           0.0
    3           0.1
    4           0.1
    5           0.2
    6           0.2
    7           0.3
    8           0.3

BOUN
    1           1
    2           1

FORC
    1           1.0
    2           1.0

END
MACR
DT          0.1
UTAN
FORM
LMAS
LOOP        10
TIME
IMPL
SOLV
DISP
NEXT
END
STOP
```

4

Momentum Transport - Fluids

Navier-Stokes

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho (\mathbf{v} \cdot \nabla) \mathbf{v} = \mathbf{F} - \nabla P + \mu \nabla^2 \mathbf{v}$$

$$\nabla \cdot \mathbf{v} = 0$$

2D element – triangular

22. Flow of Viscous Fluids; Some Special Problems of Convective Transport

22.1 Introduction

Throughout this book we have endeavoured to present the reader with a systematic approach to a variety of problems of the physical world which, once posed in mathematical terms, could be discretized and hence solved numerically. Problems of solid mechanics have, however, been predominant and to redress the balance this chapter is devoted to fluid mechanics.

Although we could start by writing the appropriate governing differential equations—and then solve these by applying the general principles of Chapter 3—we prefer to approach the mechanics of viscous fluid flow via their analogy to solid mechanics, and the first sections of this chapter are devoted to such an approach. This will permit the reader to utilize directly, or with minor modification, some of the programs developed for solids to solve certain fluid problems.

The major difference from the formulations already encountered lies in the convective terms which enter the equations of fluid mechanics problems. These lead to non-symmetric matrices if the conventional, Galerkin, approach is used in their discretization. Further, instability of computation can occur and this necessitates special discretization procedures so far not encountered in this text. We shall outline these in section 22.8.

Space limitations will not allow an exhaustive treatment to be presented here. In particular high-speed compressible (trans- or super-sonic) flow will not be considered. For supplementation the reader is referred to a series of conference proceedings and texts.[1-6] However, it is hoped that the contents of this chapter, together with those of Chapters 17, 20, and 23 in which some special fluid flow cases are treated, will give the reader a reasonably full picture of the possibilities open in this field. Some prior knowledge of fluid mechanics is naturally assumed—and for more detailed

treatment of the essentials the reader should consult some of the well-known texts.[7,8]

22.2 Basic Concepts of Viscous, Slightly Compressible Flow

22.2.1 *Equilibrium.* If an isolated volume of fluid is considered at some instance of time (Fig. 22.1) then, just as in a solid, in its interior the stresses σ must be in equilibrium with the body forces $\mathbf{b}$ which include the appropriate acceleration forces. Further, on its external surfaces the stresses σ must be in balance with the applied traction $\bar{\mathbf{t}}$. Thus, both the internal equilibrium equations and those on the boundary are *identical* to those pertaining to the solid. Using the nomenclature of Chapter 3, Eq. 3.40, and of Chapter 12, Eq. 12.14, we can write

$$\mathbf{L}^{\mathrm{T}}\sigma + \mathbf{b} = 0 \quad \text{in } \Omega \tag{22.1a}$$

and

$$\mathbf{G}\sigma = \bar{\mathbf{t}} \quad \text{on } \Gamma_t \tag{22.1b}$$

where Ω is the problem domain and Γ_t its boundary on which tractions are prescribed.
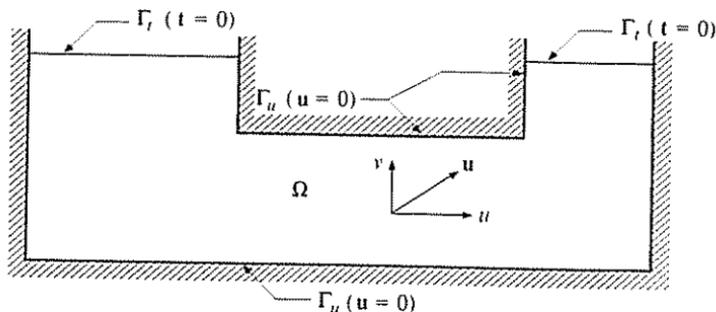
Fig. 22.1   A two-dimensional fluid flow domain

Thus the virtual work relationships used in Chapter 2 and discussed in Chapter 3 can once again be invoked. It is convenient now to apply virtual velocities $\delta\mathbf{u}$ in place of virtual displacements and we can write in place of Eqs. 22.1 the equivalent statement

$$\int_{\Omega} \delta\varepsilon^{\mathrm{T}}\sigma \, d\Omega - \int_{\Omega} \delta\mathbf{u}^{\mathrm{T}}\mathbf{b} \, d\Omega - \int_{\Gamma_t} \delta\mathbf{u}^{\mathrm{T}}\bar{\mathbf{t}} \, d\Gamma = 0 \tag{22.2}$$

where Γ_t stands for the part of the boundary on which tractions are specified and $\delta\mathbf{u} \neq 0$ there.

In the above,

$$\delta \dot{\varepsilon} = \mathbf{L}\, \delta \mathbf{u} \quad \text{and} \quad \dot{\varepsilon} = \mathbf{L}\mathbf{u} \tag{22.3}$$

defines the virtual *strain rate* by an identical expression to that used previously to define virtual strains (viz. Eq. 6.9. Chapter 6 for such a definition in three dimensions).

In fluid mechanics, due to the continually changing *displacements*, it is natural that we concentrate our attention on *velocities* and these at a fixed point of space will be denoted by $\mathbf{u}$—an identical symbol to that previously used for displacements. The body forces $\mathbf{b}$ per unit volume can be written, as in solid mechanics (*vide* Chapter 20), invoking d'Alembert's principle, as

$$\mathbf{b} = \mathbf{b}_0 - \rho \mathbf{c} \tag{22.4}$$

where $\mathbf{c}$ is the acceleration vector acting on each particle and ρ is the density. As we have defined velocity $\mathbf{u}$ at a point in space rather than with reference to a particle, the simple differentiation of the latter with respect to time does not suffice to define the acceleration. This is now given by the *total* (or particle) derivative of $\mathbf{u}$, e.g. for the x component

$$c_x = \frac{\mathrm{D}\mathbf{u}}{\mathrm{d}t} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial t}, \text{ etc.} \tag{22.5a}$$

As $\partial x/\partial t \equiv u$, etc., we can write the total acceleration vector as

$$\mathbf{c} = \frac{\partial \mathbf{u}}{\partial t} + (\nabla \cdot \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}\mathbf{u} \tag{22.5b}$$

where $\nabla^{\mathrm{T}} = [\partial/\partial x,\ \partial/\partial y,\ \partial/\partial z]$ and $(\nabla \cdot \mathbf{u}^{\mathrm{T}}) \equiv \mathbf{J}(\mathbf{u})$ is a Jacobian matrix.

Now even if the flow is steady, i.e., $\partial \mathbf{u}/\partial t = 0$, acceleration exists and here lies the principal difference from the solid mechanics formulation. Further, the expression for acceleration is non-linear in $\mathbf{u}$ and the problem is immediately of a non-linear nature.

22.2.2 *Constitutive relations.* In a fluid, by definition, no deviatoric stresses can be supported unless motion occurs. We can thus state quite generally that the deviatoric stresses are a function of the *strain rates* $\dot{\varepsilon}$.

If we define the pressure p as

$$p = -\sigma_m = (\sigma_x + \sigma_y + \sigma_z)/3 \tag{22.6}$$

we can write a very general linear relationship between the deviatoric stress σ' and strain rate as

$$\sigma' \equiv \sigma + \mathbf{m}p = \mathbf{D}'\dot{\varepsilon} \tag{22.7}$$

with

$$\mathbf{m}^{\mathrm{T}} = [1, 1, 1, 0, 0, 0]$$

For an isotropic incompressible fluid, by analogy with solid mechanic one constant μ, known as viscosity, defines completely the $\mathbf{D}'$ matrix.

$$\mathbf{D}' = \mu \begin{bmatrix} 2 & & & & & \\ & 2 & & & 0 & \\ & & 2 & & & \\ & & & 1 & & \\ & 0 & & & 1 & \\ & & & & & 1 \end{bmatrix} \qquad (22.8$$

Clearly μ plays here an identical role to that of the shear modulus G i elasticity (*vide* Chapter 11, Eq. (11.22)).

The constitutive relationship (22.7) is thus of an identical form to tha pertaining to incompressible solid mechanics with the strain rates no playing the role of strains, and additional constraint is thus necessar before the solution can be attempted.

22.2.3 *Continuity equation.* If an infinitesimal volume of space is con sidered then, quite generally, we can state that the nett rate of mas inflow is equal to the rate of mass accumulation. Thus, if ρ is the densit we can write

$$\frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) + \frac{\partial}{\partial z}(\rho w) - \frac{\partial \rho}{\partial t} \equiv \mathbf{\nabla}^{\mathrm{T}}(\rho \mathbf{u}) - \frac{\partial \rho}{\partial t} = 0. \qquad (22.9$$

Quite generally the pressure p and the density ρ are related by a suitab state relation

$$\rho = \rho(p). \qquad (22.10$$

If, however, the density changes are small, the continuity relationshi can be simplified to

$$\mathbf{\nabla}^{\mathrm{T}}\mathbf{u} \equiv \dot{\varepsilon}_v = 0 \qquad (22.11$$

stating simply that the rate of volumetric straining is identically zero. Th is analogous to the constraint used in incompressible solid mechanic (Chapters 11 and 12) and we shall in the main be concerned only wit problems where this incompressibility is enforced.

22.2.4 *Summary.* We have noted that a completely formal analog exists between the elasticity and viscous fluid mechanical problems.

Indeed, if we disregard the difference which occurs in the acceleratio forces and consider a purely incompressible flow the analogy is exac *Thus all the methodology developed for the solution of incompressible elasti solids is immediately available for the solution of viscous incompressibi flow under steady-state conditions, omitting acceleration terms.*

Following identification of terms is necessary

Elasticity		$\longleftrightarrow$	*Viscous flow*	
displacement	**u**	$\longleftrightarrow$	velocity	**u**
strain	ε	$\longleftrightarrow$	strain rate	$\dot\varepsilon$
stress	σ	$\longleftrightarrow$	stress	σ
shear modulus	G	$\longleftrightarrow$	viscosity	μ

The flow in which acceleration effects are negligible is generally known as creeping—and clearly for its solution any of the techniques already described for the solution of incompressible elasticity are immediately available.

Amongst these we have already encountered (and obviously more alternatives are possible):

1. The use of **u** and p as variables—with p entering the variational form as a Lagrangian multiplier (Chapter 12, p. 323).
2. The use of **u** as the only variable with the incompressibility constraint entering by use of a penalty function (Chapter 11, p. 286).
3. The use of equilibrating formulations (Chapter 12, p. 306).
4. The use of stream functions (Chapter 12, p. 324).

If acceleration effects are not negligible—their insertion into the discretization process (if this is achieved by Galerkin procedures) is simple and follows the lines used in structural dynamic effects in Chapter 20. However, the use of variational principles even in steady-state cases is no longer possible as true variational principles no longer exist.[9] In the next section we shall perform the discretization of the various types explicitly.

When formulating elasticity problems in Chapter 2 and elsewhere, we started from the virtual work principle as a basis and did not state the full governing equations explicitly. Such equations could be readily derived for elasticity if displacement formulation were to be used by eliminating the stresses and strains from the equilibrium equations and bear the name of Navier. Indeed these equations could (less conveniently) have been used for obtaining the first finite element discretization.

In fluid mechanics the conventional starting point of discretization is often based on similar equations.[10–21] Although we shall not pursue this line, which obviously leads to the same results as those obtainable by direct use of virtual work,[19] it is of interest to state explicitly the governing equations, which are known as those of Navier–Stokes.

Thus, if we eliminate σ from Eq. (22.1a) using relationships (22.3–5) and (22.7) we obtain a general Navier–Stokes equation

$$\rho\left[\frac{\partial \mathbf{u}}{\partial t} + (\nabla \mathbf{u}^T)^T \mathbf{u}\right] = -\mathbf{L}^T \mathbf{m} p + \mathbf{L}^T \mathbf{D}' \mathbf{L} \mathbf{u} + \mathbf{b}_0. \qquad (22.12)$$

With the form of $\mathbf{D}'$ given by Eq. (22.8) and $\mathbf{L}$ as defined previously, the above can be simplified to a more standard form. Thus in x direction

$$\rho\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z}\right) = -\frac{\partial p}{\partial x} + 2\frac{\partial}{\partial x}\left(\mu\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\mu\frac{\partial u}{\partial y} + \mu\frac{\partial v}{\partial x}\right)$$
$$+ \frac{\partial}{\partial z}\left(\mu\frac{\partial u}{\partial z} + \mu\frac{\partial w}{\partial x}\right) \qquad (22.13)$$

with similar equations in the y and z directions. For solution this equation has to be combined with that of continuity (Eq. (22.9)).

Alternative formulations can be devised, however, in which the stresses σ may be used as explicit variables.

22.3 Discretization of Viscous Flow Equations

Obviously with the analogy stated in the previous section, the discretization details could be omitted as these follow precisely the lines used in equivalent solid mechanics sections. For completion, however, we still give here brief details of the three useful forms.

22.3.1 *Velocity and pressure as variables.* In this we shall discretize the velocity and pressure in terms of independent parameters

$$\mathbf{u} = \mathbf{N}^u \mathbf{a}^u; \qquad p = \mathbf{N}^p \mathbf{a}^p \qquad (22.14)$$

Using the virtual work statement of Eq. (22.2), with

$$\delta\mathbf{u} = \mathbf{N}^u\,\delta\mathbf{a}^u \qquad \text{and} \qquad \delta\dot{\varepsilon} = (\mathbf{L}\mathbf{N}^u)\,\delta\mathbf{a}^u \equiv \mathbf{B}\,\delta\mathbf{a}^u \qquad (22.15)$$

we can write

$$\delta\mathbf{a}^{u\mathsf{T}}\left[\int_\Omega \mathbf{B}^\mathsf{T}\sigma\,d\Omega - \int_\Omega \mathbf{N}^{u\mathsf{T}}\mathbf{b}\,d\Omega - \int_{\Gamma_t} \mathbf{N}^{u\mathsf{T}}\bar{\mathbf{t}}\,d\Gamma\right] = 0 \qquad (22.16)$$

Noting that this is true for all variations $\delta\mathbf{a}^u$ we have, on inserting Eqs. (22.3–5) and (22.7)

$$\mathbf{K}\mathbf{a}^u + \bar{\mathbf{K}}\mathbf{a}^u + \mathbf{K}^p\mathbf{a}^p + \mathbf{M}\frac{d\mathbf{a}^u}{dt} + \mathbf{f}^u = 0 \qquad (22.17)$$

with coefficients given by

$$\mathbf{K}_{ij} = \int_\Omega \mathbf{B}_i^\mathsf{T}\mathbf{D}'\mathbf{B}_j\,d\Omega \qquad (22.18a)$$

$$\bar{\mathbf{K}}_{ij} = \int_\Omega \rho(\mathbf{N}_i^u)^\mathsf{T}\,(\nabla.(\mathbf{N}\mathbf{a}^u)^\mathsf{T})^\mathsf{T}\mathbf{N}_j\,d\Omega \qquad (22.18b)$$

$$\mathbf{K}_{ij}^p = -\int_\Omega \mathbf{B}_i^u\mathbf{m}\mathbf{N}_j^p\,d\Omega \qquad (22.18c)$$

$$M = \int_\Omega (N_i^u)^T \rho N_j^u \, d\Omega \qquad (22.18d)$$

$$f = -\int_\Omega (N_i^u)^T b_0 \, d\Omega - \int_{\Gamma_t} (N_i^u)^T \bar{t} \, d\Gamma. \qquad (22.18e)$$

We note immediately that, with the exception of the second matrix, as expected, the standard forms of elastic analysis are rediscovered with appropriate stiffness, mass, and force matrices.

To obtain the second equation necessary in view of the constraint equations we shall use the Galerkin process and simply pre-multiply the continuity equation (22.11) by $(\delta p)^T$ and integrate for the case of complete incompressibility.

Thus we have

$$(\delta a^p)^T \int_\Omega (N^p)^T \varepsilon_v \, d\Omega = 0 \qquad (22.19)$$

or, noting that this is true for all δa^p and writing

$$\varepsilon_v \equiv m^T L u \equiv m^T L N^u a^u \equiv m^T B a^u \qquad (22.20)$$

this results in an equation

$$(K^p)^T a^u = 0 \qquad (22.21)$$

with K^p taking on the form already given in Eq. (22.18c).

Equation systems (22.17) and (22.21) can be written as

$$\begin{bmatrix} K + \bar{K}, & K^p \\ K^{pT}, & 0 \end{bmatrix} \begin{Bmatrix} a^u \\ a^p \end{Bmatrix} + \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{Bmatrix} a^u \\ a^p \end{Bmatrix} + \begin{Bmatrix} f \\ 0 \end{Bmatrix} = 0 \qquad (22.22)$$

and can be used for the solution of transient viscous flow problems.

The reader will note that

(a) the equations are non-symmetric and non-linear if the velocities are large enough for $\bar{K}$ to be significant;

(b) when the problem is one of steady state and $\bar{K}$ is neglected, the Lagrangian form of incompressible elasticity equations of Chapter 12, p. 323, has been re-derived (now without use of a variational statement).

The formulation just presented is one of the most popular in the context of fluid mechanics and has been used frequently.[11,14,21] All the remarks made previously about over-constraint are once again applicable and practitioners find that, generally, a lower order of interpolation of p compared with that of u is desirable only to avoid over-constraints. Arguments of 'consistency' have, however, been used in the above

context but we believe that these are not the correct reasons for improved performance with such mixed interpolations.[20, 21]

22.3.2 *Stream function formulation.* It is possible to define the velocity field **u** by means of an auxiliary set of functions so that the continuity (incompressibility) condition is automatically satisfied.

Although vector stream functions can be obtained in three dimensions these have not proved successful and the approach is restricted generally to two dimensions. Writing thus

$$\mathbf{u} = \hat{\mathbf{L}}\psi: \qquad \hat{\mathbf{L}}^{\mathrm{T}} = \left[\frac{\partial}{\partial y}, -\frac{\partial}{\partial x}\right] \tag{22.23}$$

the velocity field automatically satisfies Eq. (22.11) written in two space dimensions.

If now we discretize the stream functions by writing

$$\psi = \hat{\mathbf{N}}\mathbf{a} \tag{22.24}$$

we note that the virtual work equation, (22.2), can be written again as

$$\delta\mathbf{a}^{\mathrm{T}}\left[\int_{\Omega} \hat{\mathbf{B}}^{\mathrm{T}}\sigma \, d\Omega - \int_{\Omega} (\hat{\mathbf{L}}\hat{\mathbf{N}})^{\mathrm{T}}\mathbf{b} \, d\Omega - \int_{\Gamma_t} (\hat{\mathbf{L}}\hat{\mathbf{N}})^{\mathrm{T}}\bar{\mathbf{t}} \, d\Gamma\right] = 0 \tag{22.25}$$

with

$$\hat{\mathbf{B}} = \mathbf{L}\hat{\mathbf{L}}\hat{\mathbf{N}}. \tag{22.26}$$

On insertion of Eqs. (22.5) and (22.7) we shall find that the coefficients of p disappear and a set of equations of the standard form

$$(\hat{\mathbf{K}}+\hat{\bar{\mathbf{K}}})\mathbf{a}+\hat{\mathbf{M}}\frac{d\mathbf{a}}{dt}+\mathbf{f} = 0 \tag{22.27}$$

can be written with

$$\hat{\mathbf{K}}_{ij} = \int_{\Omega} \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{D}'\hat{\mathbf{B}} \, d\Omega \qquad \hat{\mathbf{M}}_{ij} = \int_{\Omega} (\hat{\mathbf{L}}\hat{\mathbf{N}}_i)^{\mathrm{T}}\rho\hat{\mathbf{L}}\hat{\mathbf{N}}_j \, d\Omega$$

$$\hat{\bar{\mathbf{K}}}_{ij} = \int_{\Omega} (\hat{\mathbf{L}}\hat{\mathbf{N}}_i)^{\mathrm{T}}\rho(\nabla(\hat{\mathbf{L}}\hat{\mathbf{N}}\mathbf{a})^{\mathrm{T}})^{\mathrm{T}}\hat{\mathbf{L}}\hat{\mathbf{N}} \, d\Omega \tag{22.28}$$

$$\mathbf{f}_i = -\int_{\Omega} (\hat{\mathbf{L}}\hat{\mathbf{N}}_i)\mathbf{b}_0 \, d\Omega - \int_{\Gamma_i} (\hat{\mathbf{L}}\hat{\mathbf{N}}_i)^{\mathrm{T}}\bar{\mathbf{t}} \, d\Gamma.$$

Two points are worth noting beyond the existence of the non-linear and non-symmetric matrix $\hat{\bar{\mathbf{K}}}$ as in the previous formulation. These are, first, that the shape function $\hat{\mathbf{N}}$ now needs to possess C_1 continuity as second order derivatives exist in $\hat{\mathbf{B}}$ and, second, that the formulation is (almost) identical to that of plate bending problems. Indeed, solutions with this procedure have invariably utilized this analogy using many of the elements formerly noted in Chapter 10.[18, 22]

22.3.3 *'Penalty' function formulation*. This has been introduced as an effective procedure for incompressible elasticity in Chapter 11 and, hence, in the present context the method should be applicable. We shall, however, approach it without stating a variational principle.

To eliminate the variable p let us write in the constitutive relation (22.7)

$$p = \alpha \dot{\varepsilon}_v, \qquad (22.27)$$

where α is a large number. As $\dot{\varepsilon}_v \to 0$ by the constraint equation, Eq. (22.11), p will thus be a finite quantity.

With this substitution the need for discretizing p is eliminated, and for a discretized velocity $\mathbf{u} = \mathbf{N}\mathbf{a}$ we have

$$\dot{\varepsilon}_v = \mathbf{m}^T \mathbf{L}\mathbf{N}\mathbf{a}. \qquad (22.30)$$

Pursuing the discretization of Eqs. (22.16) and (22.17) we arrive now at

$$\mathbf{K}\mathbf{a} + \bar{\mathbf{K}}\mathbf{a} + \bar{\bar{\mathbf{K}}}\mathbf{a} + \mathbf{M}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a} + \mathbf{f} = 0 \qquad (22.31)$$

where all, but one, of the matrices are defined in Eq. (22.18) and

$$\bar{\bar{\mathbf{K}}}_{ij} = \int_\Omega (\mathbf{m}^T \mathbf{B}_i)^T \alpha(\mathbf{m}^T \mathbf{B}_j) \, \mathrm{d}\Omega. \qquad (23.32)$$

Again α can be recognized as analogous to the bulk modulus of elasticity, and indeed the standard form of nearly incompressible elasticity used in Chapter 11 is obtained for slow flow.

The procedure was first formulated in Reference 19 and used subsequently for creeping flow with 'reduced' integration elements. The first effective use of solutions for the full equations of viscous flow was made by Hughes *et al.*[23] using a bi-linear quadrilateral with a single point integration for the volumetric strain rate terms.

22.4 Some Applications of Viscous Flow Forms and Solution Techniques

22.4.1 *Steady-state creeping Newtonian flow*. By 'Newtonian' we mean that the problem is linear with a constant viscosity. With all acceleration terms rejected the formulation gives linear equations and little has to be said about this solution.

Entry flow. In this first example[19] of Fig. 22.2 a solution of entry flow in an axi-symmetric case is obtained

(*a*) by a stream function form in which the Hermitian rectangles of Chapter 10 are used;

(*b*) by a standard elasticity program with isoparametric, 8-node elements utilizing 2×2 Gauss point, 'reduced' integration (near-

Fig. 22.2   Velocity profile development in an axially symmetric flow. (*a*) by steam function approach; (*b*) **u** — penalty function approach ($v = 0.49995$); (*c*) $u - p$ formulation

incompressibility is here achieved by setting the equivalent of Poisson's ratio as 0·49995); and

(c) by velocity-pressure formulation using a parabolic interpolation for velocities and linear for pressures.[24]

All solutions give almost identical results at a comparable cost and compare well with a finite difference solution of the same problem carried out with a very fine subdivision.[25]

Although there is no apparent difference in the solution technique the last two procedures are easily generalized to three dimensions.

Solution for such three-dimensional flows have been presented in references 19 and 26.

Flow past an obstacle. This example, illustrated in Fig. 22.3 in which solution by both stream function and by penalty-velocity forms were obtained, brings out a further point of difficulty encountered with the former type of discretization. As the distribution of flow is not known initially, the value of the stream function on the obstacle is not known *a priori*—except that it is constant.

An additional requirement has now to be introduced.[27] This states that the rate of work done by boundary tractions on the stationary object must be zero, i.e., that

$$\int_\Gamma \delta \mathbf{u}^\mathsf{T} \bar{\mathbf{t}} \; d\Gamma \equiv \int_\Gamma (\hat{\mathbf{L}} \, \delta \psi)^\mathsf{T} \bar{\mathbf{t}} \; d\Gamma = 0 \qquad (22.33)$$

where Γ is the surface of the obstacle.

Imposition of this condition on the stream function parameters can be made if two independent solutions are carried out—a procedure which is clearly inconvenient and computationally expensive.

22.4.2 *Steady-state, creeping non-Newtonian flow. Visco-plastic metal flow.* In many fluids for which slow rates of flow are of interest the viscosity is a function of the strain rate $\dot{\varepsilon}$. Such fluids comprise many oils, chemicals, and indeed metals.

If the constitutive relation of Eqs. (22.7) and (22.8) is examined we find that it is convenient, in isotropic materials, to write it in terms of second stress and strain rate invariants, $\bar{\sigma}$ and $\dot{\bar{\varepsilon}}$ (for definitions see Chapter 18).

The relation can be written simply as

$$\bar{\sigma} = \mu \dot{\bar{\varepsilon}} \qquad (22.34)$$

which, for a Newtonian fluid, is of a form shown in Fig. 22.4. For non-Newtonian fluids the stress strain–rate invariant relationship may take various forms—a typical one being written as

$$\bar{\sigma} = \beta \dot{\bar{\varepsilon}}^n. \qquad (22.35)$$
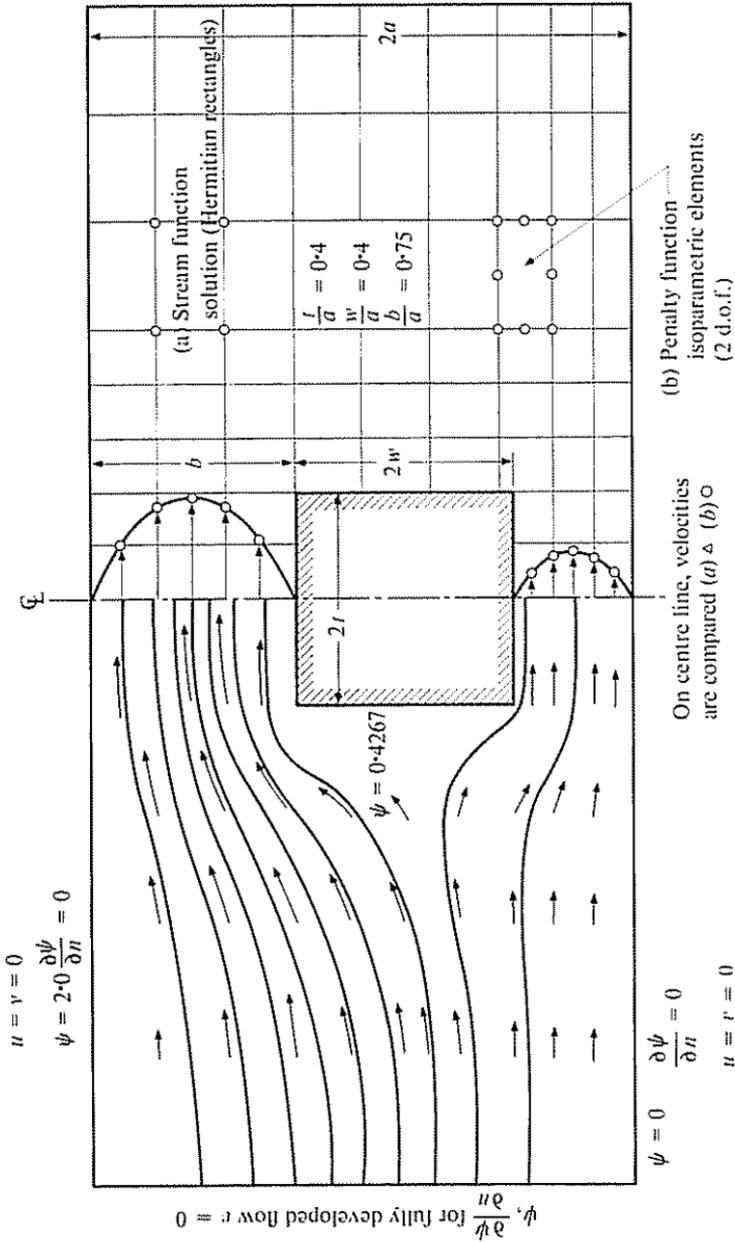
Fig. 22.3   Two-dimensional flow with a square obstruction placed asymmetrically

Clearly this can be represented as a behaviour of the standard form with

$$\mu \equiv \beta\dot{\bar{\varepsilon}}^{n-1}. \tag{22.36}$$

A very characteristic form of behaviour is known as that of Bingham fluid in which a yield stress is exhibited.

This can be written as a particular case of viscoplasticity, giving (see Fig. 22.4)

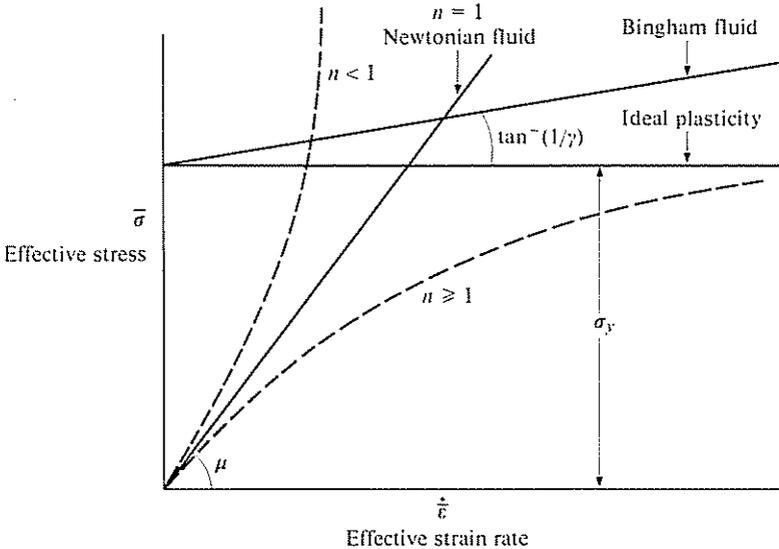$$\dot{\bar{\varepsilon}} = \gamma(\bar{\sigma} - \bar{\sigma}_y). \tag{22.37}$$

Fig. 22.4   Newtonian and non-Newtonian effective stress–strain relations

Again we can interpret this in terms of variable viscosity of expression (22.34), with

$$\mu = \frac{\dot{\bar{\varepsilon}}/\gamma + \bar{\sigma}_y}{\dot{\bar{\varepsilon}}} \tag{22.38}$$

reducing for an ideally plastic behaviour to

$$\mu = \frac{\bar{\sigma}_y}{\dot{\bar{\varepsilon}}}. \tag{22.39}$$

In Chapter 18 we have discussed the elasto-plastic and elasto-viscoplastic behaviour of many materials. If the deformations are such that elastic strains can be neglected, all such solids behave in effect as non-Newtonian fluids and solutions for their behaviour are easily attainable.

The creeping flow formulations of all types discussed in previous sections have resulted in a general form

$$Ka + f = 0 \tag{22.40}$$

Now $K = K(a)$ and is a symmetric matrix dependent on the viscosity and hence on the velocity parameters which define the effective strain rate $\dot{\varepsilon}$.

Various non-linear solution techniques can be adopted for this problem but the simplest in which the matrix K is recalculated iteratively with

$$a^{m+1} = -K^{m-1}f \tag{22.41}$$

gives very rapid convergence, even if quite severe non-linearity of the type given by Eq. (22.31) is encountered, providing the forcing function is one of specified boundary velocities.

Many solutions of non-Newtonian flow are available in the literature[28,29] but the plastic flow situations are of greatest interest.[19,24,27,30]

In Fig. 22.5 we show, for instance, a problem of steady-state extrusion solved[30] as a case of non-Newtonian flow. Comparison with solutions available for the same problems by classical slip line solution confirm the accuracy attainable in modelling such a flow.

22.4.3 Steady-state viscous flow with inclusion of convective acceleration terms. We have already remarked that, in the case of steady-state viscous flow in which the convective acceleration terms are retained, all formulations give non-linear equation systems (even if viscosity is Newtonian) of the form

$$(K + \bar{K}(a))a + f = 0 \tag{22.42}$$

in which $\bar{K}(a)$ is a non-symmetric matrix dependent on the solution parameters (velocities).

In fluid mechanics it is usual to characterize the flow by a non-dimensional parameter known as the Reynolds number R_n and defined as

$$R_n = \frac{\rho U d}{\mu} \tag{22.43}$$

where U and d are a characteristic velocity and dimension, respectively. The creeping form which we have previously discussed is thus the limiting case of $R_n \to 0$. Now we shall consider increasing the Reynolds number.

The solution techniques for the non-linear equation system depends evidently on the value of R_n.

For small values of R_n a modified Newton–Raphson technique is effective, using only the constant and symmetric matrix in solution.

At higher Reynolds numbers a full Newton–Raphson iteration is necessary and this involves a repeated solution of a *non-symmetric* equa-

K-Yield in pure shear

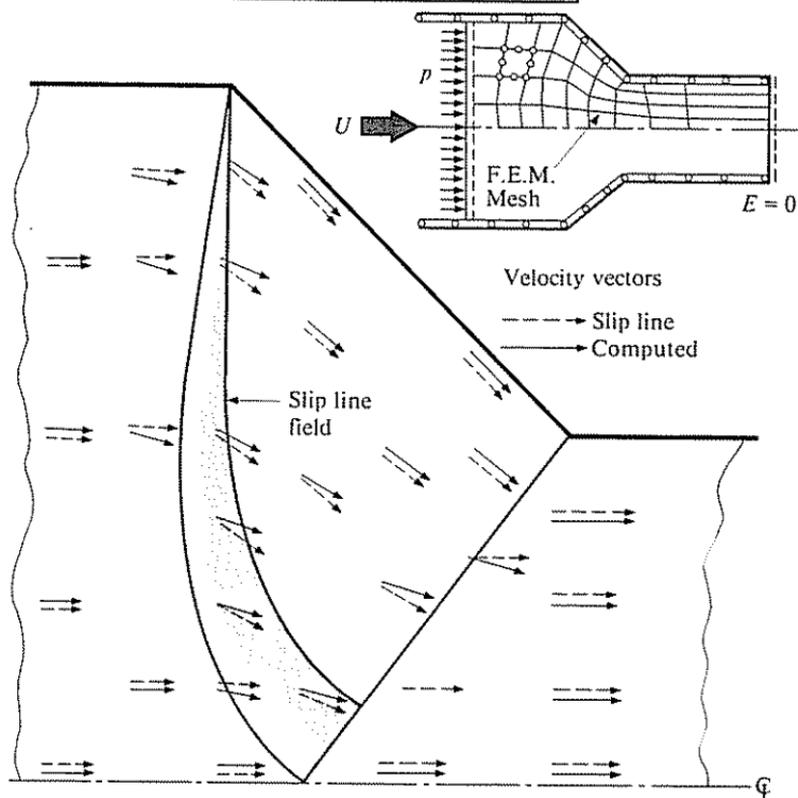| Extrusion pressure - $p/2\,K$	
Slip line	0·90
Penalty function	0·94
Stream Function	0·92
Elasto-plastic solution	0·93

Fig. 22.5 Plane strain extrusion; ideal plasticity; frictionless walls; penalty function solution

tions system. Other techniques, such as perturbation methods, etc., have also been used with success.[31]

When the Reynolds number becomes very large, and the convective terms predominate, convergence generally ceases. This occurs due to two causes: first, at some value of R_n the flow becomes physically highly unstable and turbulence sets in; second, an instability may be induced by the special character of the approximation to the convective term when the standard Galerkin form is used. This numerical instability will be discussed in depth later.

Purely as an example of such higher R_n computation, we show in Fig. 22.6 some solutions obtained using the velocity–pressure techniques for a flow around a two-dimensional obstacle.[20]
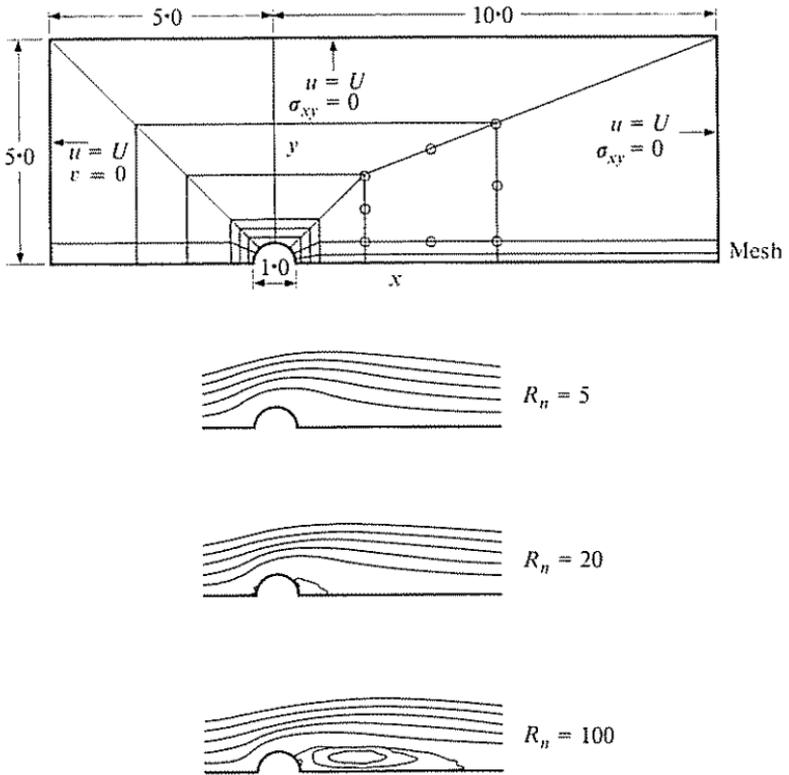
Fig. 22.6   Flow round a cylinder

22.5 Turbulent Flow

As the value of R_n increases, the turbulence which starts with large isolated eddies increases until it becomes widely distributed through the fluid. If *average velocities* are considered then the effect of the turbulence is analogous to that of viscosity and the flow can be represented by the standard viscous equations with the viscosity coefficient now replaced by an *eddy viscosity*, $\bar{\mu}$, which is dependent on the whole velocity field and its gradients. Indeed this behaviour may well be anisotropic, i.e., specified by several such coefficients. In principle, thus, turbulent flow approached in this way presents no more difficulties than those associated with non-Newtonian situations. In practice, unfortunately, no general explicit expressions for determining the eddy viscosity coefficients exist and, at

best, very rough solutions are attainable. We shall, however, make use of such turbulence concepts in some aspect of shallow water flow.

22.6 Transient, Time-dependent Flow and Free Surface Problems

In principle the transient flow Equations 22.22, 22.27 or 22.31 can be integrated using one or other of the time-stepping processes discussed in
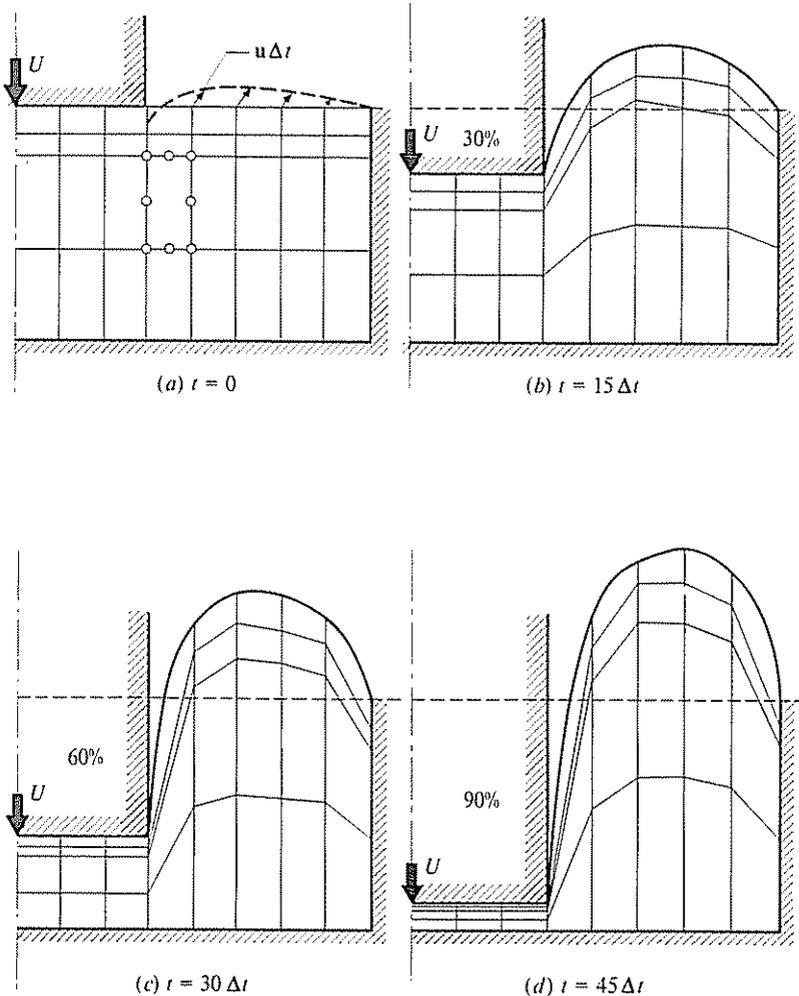

Fig. 22.7   Punch indentation problem (penalty function approach). Updated mesh and surface profile. 24 isoparametric elements. Ideally plastic material; (a), (b), (c) and (d) show various depths of indentation.

the previous chapter. Indeed it appears that such a procedure may present a useful technique for obtaining steady-state solutions if these exist. Little work has as yet been done on this aspect, but some solutions to simple problems have been obtained.[32,33]

Time-stepping techniques can readily be adopted to follow the development of the form of the free surface. With the initial position of this known, the velocities at the start of a time step determine the position of particles on the free surface at a later time. Iteration can be used here, but if the time interval is not large, a single forward integration giving a change of position, x, as

$$\Delta x^{m+1} = u^m \Delta t \qquad (22.44)$$

is effective.

This technique, when used in the context of slow (creeping) flow in which all acceleration effects are ignored, necessitates simply an updating of the free surface and a successive resolution of the problem with a new configuration. Indeed the whole mesh can be updated in this way, but if this is found to produce badly shaped elements a new mesh can be generated to fit the new surface at each stage.

Techniques of this kind are extremely useful in a variety of metal forming and rolling processes.[24,27,34,35,36] In Fig. 22.7 we show successive stages of deformation caused in an ideally plastic metal by a punch.[27]

Steady-state free surface problems are in a sense more difficult. Here we have to ensure that the traction-free surface develops velocities which are *strictly tangential to this surface*,[34,37] In such cases it is convenient to specify the original surface, obtain a velocity solution, and recompute a new surface by integrating from a known point, noting that the slope is given by the direction of the velocity vector at all points. Three or four repetitions of this process frequently suffice. In Fig. 22.8 we show an axi-symmetric drawing problem of a creeping Newtonian fluid emerging from a tube. The problem is of some importance in glass fibre drawing.

22.7 Shallow Water Flow: Estuaries and Lakes

22.7.1 *General equations.* In many problems of practical engineering importance the concern is with flow in bodies of water whose plan dimension is much larger than the depth. Lakes, estuaries, and indeed the oceans provide such examples for which a study of currents caused by wind action, periodic tidal forces, or wave drag is of interest. In contrast to the corresponding plane stress problems, the distribution of velocities across the depth is not uniform and often the changes of depth
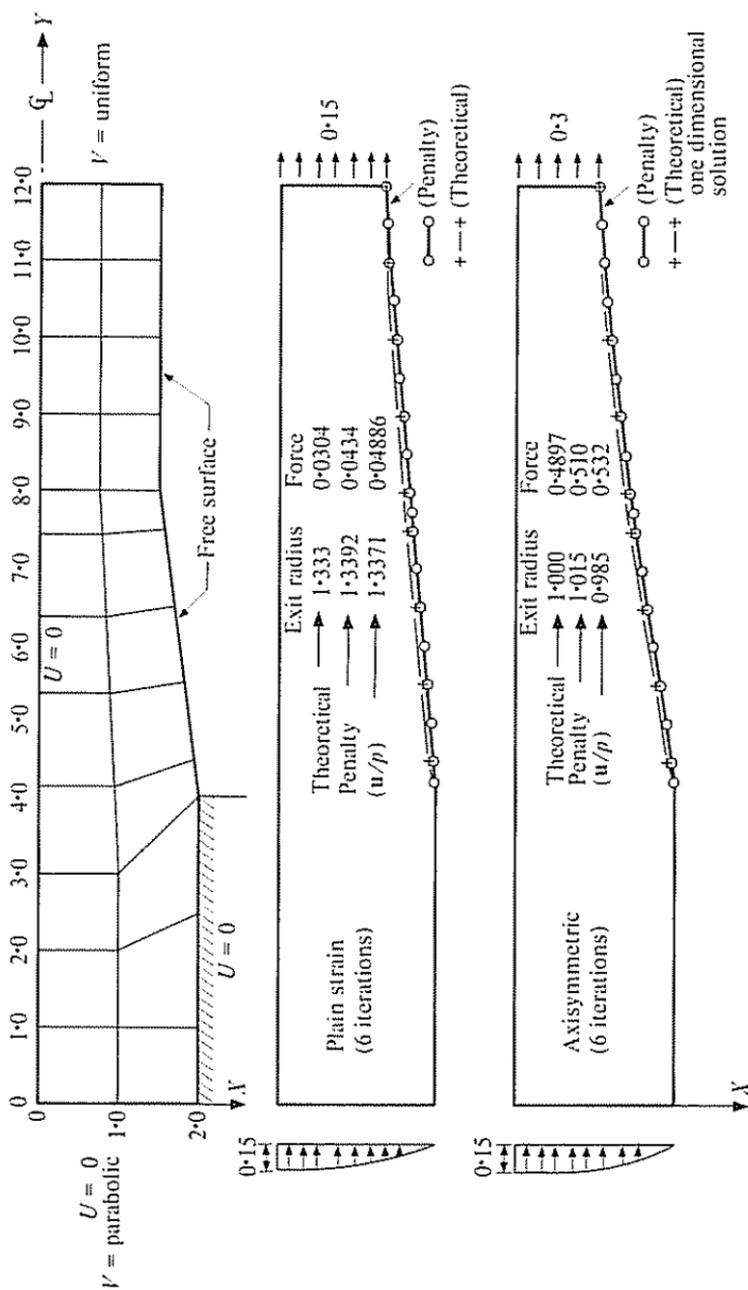
Fig. 22.8 Viscous incompressible jet—drawing problem

provide the main driving forces. Nevertheless, complete three-dimensional analysis of such flows is not practicable and two-dimensional approximations have to be made. Various forms of such approximations are available,[38,39,40] and here we shall present a derivation of a set of quite general equations which are of a form not dissimilar to those of Navier–Stokes' equations already derived, but now involve some additional terms.

In subsequent parts of this section we shall show particular simplified forms of the equations which are of some practical interest.

The derivation of basic shallow water equations uses the assumptions that the vertical accelerations are negligible and that the pressure distribution in the vertical directions is hydrostatic, i.e. (see Fig. 22.9),

$$p = \rho g(\eta - z) + p_a \tag{22.45}$$
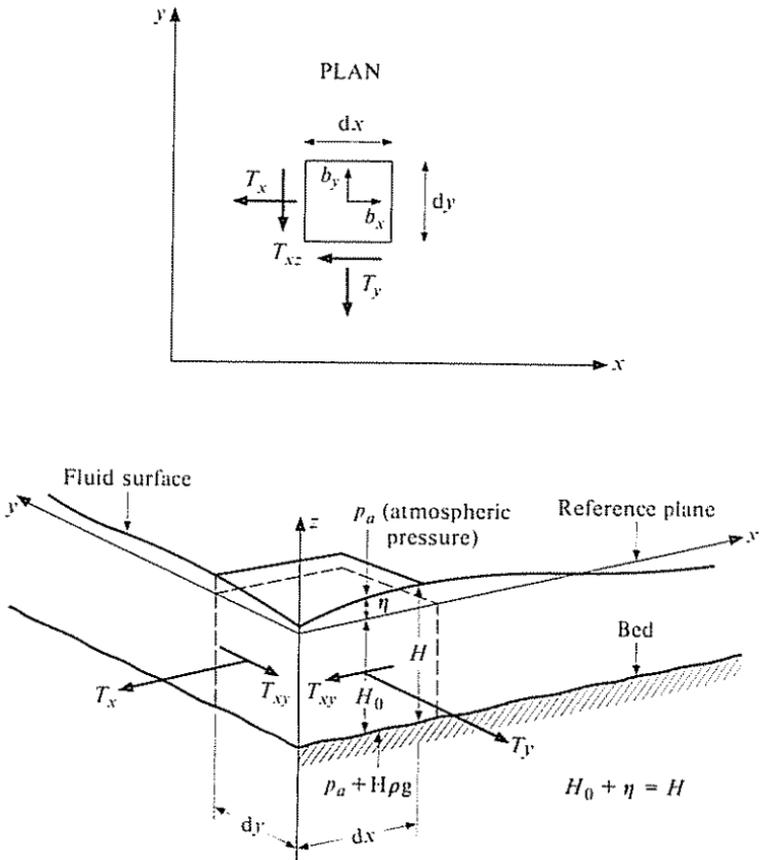
Fig. 22.9 Definitions for shallow water flow problem

where p_a is the atmospheric pressure. Further, we shall be concerned only with the average velocities in the plan direction, i.e., U or V.

$$U = \frac{1}{H} \int_{-H_0 + \eta}^{\eta} u \, dz; \qquad V = \frac{1}{H} \int_{-H_0 + \eta}^{\eta} v \, dz. \qquad (22.46)$$

With these two assumptions the overall continuity and equilibrium relations can be written.

For complete generality we assume that the density ρ can vary with position in the plan (of importance when density currents are considered) and we can write the continuity condition for a prism of unit plan area shown in Fig. 22.9 as

$$\frac{\partial}{\partial x}(\rho H U) + \frac{\partial}{\partial y}(\rho H V) - \rho \frac{\partial \eta}{\partial t} = 0 \qquad (22.47)$$

where the last term gives the rate of fluid accumulation due to the rising surface.

To examine the equilibrium in the plan directions we shall proceed in a manner completely analogous to that used in deriving the general viscous equilibrium equations.

First we observe that on the faces of an elementary prism we have tractions T_x, T_y, and T_{xy} which are due partly to the pressures and partly to turbulent mass transfer in which $\bar{\mu}$ is the eddy viscosity. Thus

$$T_x = -\int_{-H_0 + \eta}^{\eta} p \, dz + 2\bar{\mu}H \frac{\partial U}{\partial x} = -\rho g \frac{H^2}{2} - p_a H + 2\bar{\mu}H \frac{\partial U}{\partial x} \qquad (22.48a)$$

Similarly,

$$T_y = -\rho g \frac{H^2}{2} - p_a H + 2\bar{\mu}H \frac{\partial V}{\partial y} \qquad (22.48b)$$

and

$$T_{xy} = \bar{\mu}H \left(\frac{\partial U}{\partial y} + \frac{\partial V}{\partial x} \right). \qquad (22.48c)$$

We note that the tractions due to turbulent mass transfer are of precisely the same form as those associated with the viscosity coefficients in Eq. (22.7) but that the 'pressures' are now defined in terms of the depth H which plays, in shallow water equations, the same role as the density in compressible flow (in fact we shall find that the equations of shallow water flow bear a striking resemblance to compressible flow equations).

The tractions given by Eq. (22.48) must be in equilibrium with the appropriate body force vector **b** and the equilibrium equations (Eqs.

(22.1a) and (22.1b)) can again be written with the operator $\mathbf{L}$ appropriate to the two-dimensional problems as

$$\mathbf{L}^{\mathsf{T}}\mathbf{T}+\mathbf{b} = 0 \qquad \mathbf{T}^{\mathsf{T}} = [T_x, T_y, T_{xy}]. \qquad (22.49)$$

For discretization it is convenient to use the virtual work equation corresponding to Eq. (22.2), i.e.,

$$\int_\Omega \delta\dot{\varepsilon}^{\mathsf{T}}\mathbf{T}\,d\Omega - \int_\Omega \delta\mathbf{U}^{\mathsf{T}}\mathbf{b} - \int_\Gamma \delta\mathbf{U}^{\mathsf{T}}\overline{\mathbf{T}}\,d\Gamma = 0 \qquad (22.50)$$

where $\overline{\mathbf{T}}$ stands for prescribed boundary 'tractions' (which depend on the depths H) on suitable boundaries.

Once the body force vector is available the discretization can be written in the standard manner which we shall not pursue here in detail. However, it is essential to define the body force vector, as several terms not previously encountered now enter the problem. As before we can write (remembering that a depth of fluid H is considered)

$$\mathbf{b} = \mathbf{b}_0 - \rho\mathbf{c}H \qquad (22.51)$$

where the acceleration $\mathbf{c}$ is

$$\mathbf{c} = \frac{\partial\mathbf{U}}{\partial t} + (\nabla\mathbf{U}^{\mathsf{T}})^{\mathsf{T}}U$$

$$\mathbf{U}^{\mathsf{T}} = [U, V]. \qquad (22.52)$$

Further, the vector $\mathbf{b}_0$ is now specifically divided into several causes— and can be written as the sum of the following:

(a) Coriolis effects: if rotation of the earth is important due to extent of problems $-\rho f \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{U}$, where f is the Coriolis parameter ($f = 2 \times$ angular velocity of frame of reference rotation).

(b) Surface traction due to wind (or waves), τ.

(c) Bottom traction resisting motion, $-\beta\mathbf{U}$, with β a coefficient dependent on the absolute value $|\mathbf{U}|$ if turbulent conditions exist.

(d) Horizontal component of surface pressure, $p_a\nabla\eta$.

(e) Horizontal component of bottom pressure, $(p_a+\rho gH)\nabla H_0$.

Noting that the essential variables of the problem are the velocity (mean) vector $\mathbf{U}$ and the surface elevation η as

$$H = \eta + H_0 \qquad (22.53)$$

where H_0 is a known depth of the mean water surface, the full discretized equations can be written in a manner analogous to that described in section 22.3.1 by using the virtual work statement and a weighted form

of the continuity equation, Eq. (22.47). We shall spare the reader the details which he can readily fill in, but before proceeding further we shall give an explicit form of the equilibrium equation (in a manner equivalent to that of (Eq. 22.13)) as such equations have been used as the starting point of the discretization by many investigators. With an explicit form of the operator **L**, which we remind the reader is given by

$$\mathbf{L}^T = \begin{bmatrix} \dfrac{\partial}{\partial x} & 0 & \dfrac{\partial}{\partial y} \\[2mm] 0 & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial x} \end{bmatrix} \qquad (22.54)$$

the substitution of **b** and **T** into Eq. (22.49) results in two differential equations, i.e.,

$$H\rho\left(\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y} - fV\right)$$

$$= -\rho g H\frac{\partial \eta}{\partial x} - g\frac{H^2}{2}\frac{\partial \rho}{\partial x} - \beta U + H\frac{\partial p_a}{\partial x} + 2\frac{\partial}{\partial x}\left(\bar{\mu}\frac{\partial}{\partial x}(HU)\right)$$

$$+ \frac{\partial}{\partial y}\left(\bar{\mu}\frac{\partial}{\partial y}(HU) + \bar{\mu}\frac{\partial}{\partial x}(HV)\right) + \tau_x = 0 \qquad (22.55)$$

with a similar equation for the y direction.

As the surface elevation η is generally small compared with the depth H it simplifies matters to assume that

$$H \approx H_0.$$

With this simplification the reader will note that Eq. (22.55) together with the Eq. (22.47) are essentially similar to the Navier–Stokes equation in two dimension Eq. (22.13) and Eq. (22.9), but now contain some additional terms. Techniques of solution for both steady-state and transient situations for both problems will be essentially identical and therefore it is convenient to write programs capable of incorporating both classes of problems simultaneously.

In Fig. 22.10 we show some results of computations carried out for determination of tidal currents in Tokyo harbour.[41] Here the periodic nature of the tides was used to simplify the time response by a harmonic analysis applied despite the inherent non-linearity of the equations.

22.7.2 *Simplified shallow water flow equations.* Very few investigators have so far used the full set of equations of shallow flow even if certain forces or effects are absent.[42,43]

First, the horizontal eddy viscosity terms are frequently dropped.[44,45] When this is done it should be remembered that only one velocity component can be prescribed as the boundaries (the second order of the

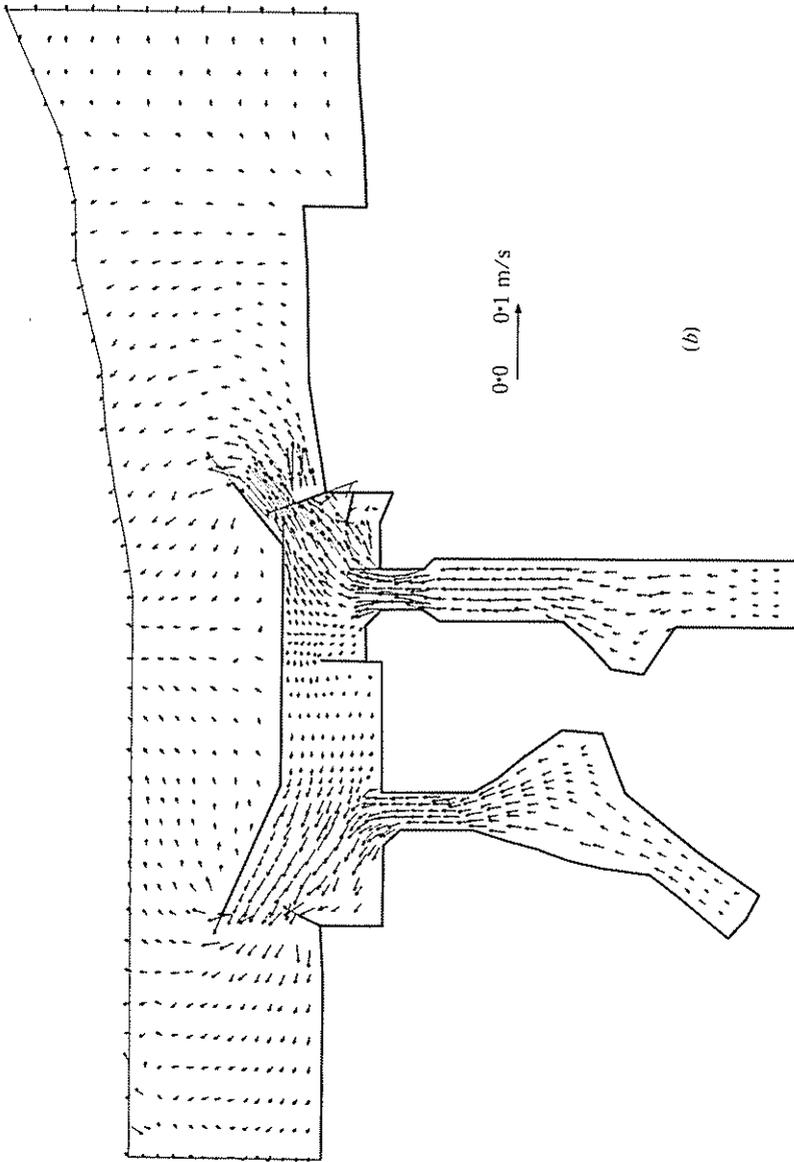(a)

0 1 km

0·0   0·1 m/s

(b)
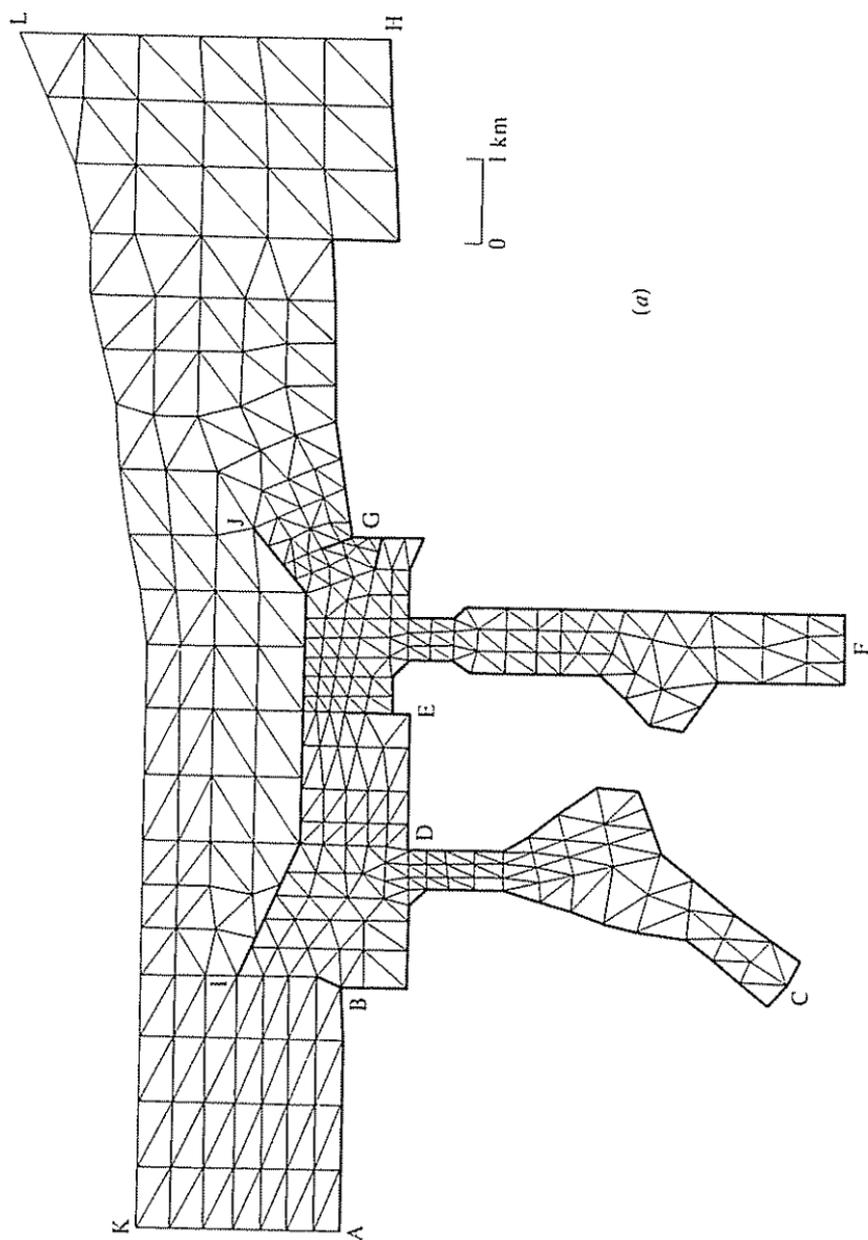
Fig. 22.10   Finite element mesh (a) and velocity distribution (b) due to tidal action in Tokyo harbour[1]

equations being now reduced to first). Indeed the 'no slip' condition on boundaries can no longer be imposed.

Second, the convective terms are frequently omitted thus linearizing the equation system (22.47) and (22.55) to a much simpler form (for constant density ρ and taking $H = H_0$).[46]

$$\frac{\partial}{\partial x}(H_0 U) + \frac{\partial}{\partial y}(H_0 V) - \frac{\partial \eta}{\partial t} = 0$$

$$\rho H_0\left(\frac{\partial U}{\partial t} - fV\right) = -\rho g H_0\frac{\partial \eta}{\partial x} - H_0\frac{\partial p_a}{\partial x} - \beta U + \tau_x \qquad (22.56)$$

$$\rho H_0\left(\frac{\partial V}{\partial t} + fU\right) = -\rho g H_0\frac{\partial \eta}{\partial y} - H_0\frac{\partial p_a}{\partial x} - \beta V + \tau_y.$$

For steady-state conditions it is convenient to introduce once again the notion of a stream function. Defining

$$H_0 U = \frac{\partial \psi}{\partial y}; \qquad H_0 V = -\frac{\partial \psi}{\partial x} \qquad (22.57)$$

the first of Eqs. (22.56) (with time differentiation omitted) is identically satisfied.

On elimination of η between the second pair we find that the governing equation reduces simply to the quasi-harmonic form (discussed in Chapter 17):

$$\frac{\partial}{\partial x}\left(\frac{\beta}{H_0}\frac{\partial \psi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\beta}{H_0}\frac{\partial \psi}{\partial y}\right) = \frac{\partial \tau_x}{\partial y} - \frac{\partial \tau_y}{\partial x} \qquad (22.58)$$

It is of interest to see that the Coriolis and pressure gradient forces do not now affect the solution. Despite this drastic simplification apparently reasonable predictions of wind (or wave) induced velocities can be made.[47,48] It seems, however, desirable to ascertain in all cases the errors due to the approximation; this can always be done by computing the contributions due to the omitted terms.

22.7.3 *Long wave equations.* If the time derivative terms of Eqs. (22.56) are not omitted we find that these equations are typical of wave problems.

For instance, if the drag, Coriolis, and pressure force terms are omitted we can eliminate U and V (by differentiation of the first equation with respect of time and the second and third by x and y, respectively) and obtain the classical wave equation of the form

$$\frac{\partial}{\partial x}\left(H_0\frac{\partial \eta}{\partial x}\right) + \frac{\partial}{\partial y}\left(H_0\frac{\partial \eta}{\partial y}\right) - \frac{1}{\rho g}\frac{\partial^2 \eta}{\partial t^2} = 0. \qquad (22.59)$$

We have discussed this type of equation in Chapter 20 and some aspects of it are dealt with in Chapter 23.

With the drag terms included it is possible, by the introduction of some further mathematical approximations, to derive an equation of type (22.59) with a damping term included.

22.8 Convective Transport Equation and some Special Finite Element Problems. 'Upwind' Weighting

22.8.1 *The convective transport problem.* In the basic fluid problem discussed in this chapter we have encountered a new type of term, i.e., that of convective acceleration (*vide* Eq. (22.5))

$$\rho(\nabla \mathbf{u}^T)^T \mathbf{u} \tag{22.60}$$

which has caused a major difficulty by introducing a *non-symmetric matrix* into the final equation.

Terms of this kind arise in an Eulerian formulation (i.e., one in which a fixed space element is considered) when a certain quantity is *transported* by a velocity field **u**. In Eq. (22.60) the quantity is the *momentum* but in many problems it may be, say, the amount of a chemical dissolved in the fluid or, of heat carried by it, etc.

If, for instance, we consider the heat transfer in a moving fluid in which the velocity **u** is known, then the heat balance equation derived by precisely the same reasoning as that concerning heat diffusion in Chapter 17 and Chapter 20 (Eq. 17.6 and Eq. 20.1) now contains an additional term. The balance equation now takes the following form:

$$\nabla^T(k \, \nabla\phi) + Q - c\frac{\partial\phi}{\partial t} - \nabla^T(c\phi\mathbf{u}) = 0 \tag{22.61}$$

where the last term is due to the transport of the heat content $c\phi$ by the moving fluid.

In steady flow, if the velocities obey the incompressibility condition, i.e. if

$$\nabla^T\mathbf{u} = 0 \tag{22.62}$$

this equation can be written as

$$\nabla^T(k \, \nabla\phi) - \nabla^T(c\phi)\mathbf{u} + Q = 0 \tag{22.63}$$

or if c is independent of position and the diffusivity is isotropic we have

$$\nabla^T(k' \, \nabla\phi) - (\nabla^T\phi)\mathbf{u} + Q' = 0 \tag{22.64}$$

with $k' = k/c$ and $Q' = Q/c$.

This type of convective problem is of extreme importance in all branches of physics and engineering. Heat transfer in fluid machinery, dispersion of pollutant in shallow water, etc., are but a few examples.

In Chapter 3 (section 3.5, p. 56) we have already treated this specific example by the Galerkin process and, beyond remarking that now non-symmetric matrices arise (just as in the corresponding fluid mechanics case), the impression was given that no special difficulties arise. However, in the context of high velocity fluid flows we have already remarked on page 621 that numerical instability has sometimes been noted. We shall now investigate the problem further—in the context of the simple form of Eq. (22.64). All the remarks which will be made are applicable to the more complex forms and, indeed, to the basic fluid flow problem itself.

In passing it should be noted that in equations of the type (22.64) the relative importance of the two terms will be obviously of crucial import-ance on the very nature of the problem. If, for instance, the conductivity (diffusivity) terms were to become zero, then we would have only a first order equation which clearly would not allow the specification of the same number of boundary conditions and would be one of initial value-propagation type.[49] We would indeed find that the temperature conditions at entry would govern entirely the solution and *downstream* conditions could not be imposed. Clearly, if k' has a small, but non-zero, value the solution will still have to be of the same nature, and the downstream effects will be highly localized. It is here that the essentials problems lie.

22.8.1 *General discretized form and the Galerkin approximation.* In section 3.5, p. 56 of Chapter 3 we have discretized the problem of Eq. (22.64) *a priori*, specifying the Galerkin form of weighting. Proceeding with an arbitrary weighting function set W_i we can similarly derive a more general discretization (which the reader can check as an exercise).

Writing

$$\phi = \sum N_i a_i = \mathbf{N}\mathbf{a} \qquad (22.65)$$

a system of linear equations of standard form

$$\mathbf{K}\mathbf{a} + \mathbf{f} = 0 \qquad (22.66)$$

is obtained, in which (omitting all boundary contributions for simplicity)

$$\mathbf{K}_{ij} = \int_\Omega (\nabla W_i)^\mathrm{T} k' (\nabla N_j)\, \mathrm{d}\Omega + \int_\Omega W_i \mathbf{u}^\mathrm{T} \nabla N_j\, \mathrm{d}\Omega \qquad (22.67)$$

$$\mathbf{f}_i = \int_\Omega W_i Q\, \mathrm{d}\Omega$$

Now the formulations and solutions can be obtained using any desired element form or weighting function.

To illustrate the difficulty which is encountered we shall consider a simple entry flow region, Fig. 22.11, where the velocity solution is assumed known (or derived as it was in this example by processes of viscous flow solution). With the diffusivity k' assumed constant, the solutions will be characterized by a non-dimensional (Péclet) number

$$P_e = Ud/k' \tag{22.68}$$

where U is the entry velocity and d a typical problem dimension (in our case half the duct width).

Using $W_i = N_i$, i.e., the standard Galerkin procedure, and a bilinear isoparametric element mesh shown in Fig. 22.11, we find that reasonable results are obtained for $P_e = 3 \cdot 75$ but that on increasing this the results deteriorate until at $P_e = 37 \cdot 5$ a meaningless oscillation is obtained. Clearly this situation is not acceptable and some corrective measure has to be taken.

22.8.2 *One- and two-dimensional problems—'upwind' weighting functions.* The difficulty just mentioned has been noted repeatedly in finite difference context[50–52] and in the finite element field the remedy was derived very recently.[53–56]

To appreciate the problem it is convenient to consider it first in one dimension, using standard linear interpolation functions shown in Fig. 22.12 and elements of a constant size h with a constant velocity u throughout. Without loss of generality the homogenous problem ($Q = 0$) is used and the weighting function will be assumed within each element to have a form

$$W_i = N_i(x) + \alpha F(x) \tag{22.68}$$

where α is positive when u is directed towards node; and

$$F(x) = -3x(x-h)/h^2 \tag{22.69}$$

is chosen so as to have a positive value in each element and zero values at the nodes to preserve C_0 continuity. Clearly $\alpha = 0$ will reproduce once again the classic Galerkin process.

On discretization and assembly of equations for a typical node the following (difference) equation is found

$$\left[1 + \frac{\gamma}{2}(|\alpha|+1)\right]\phi_{i-1} - (2+\gamma|\alpha|)\phi_i + \left[1 + \frac{\gamma}{2}(|\alpha|-1)\phi_{i+1}\right] = 0 \tag{22.70}$$

where the parameter γ is defined as

$$\gamma = uh/k' \tag{22.71}$$

(a) Problem statement
(velocity distribution determined independently)

(b) Mesh of linear elements

(c) Temperature distribution along $y = 0$ for $P_e = 3.75$ and 12.5

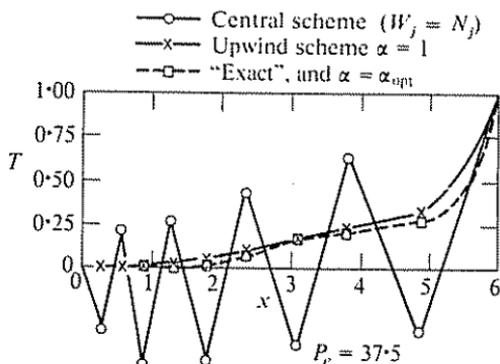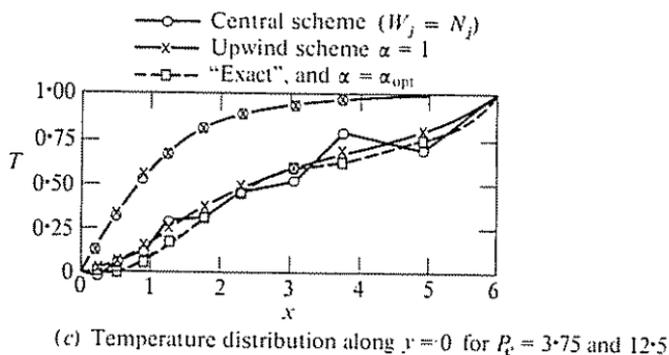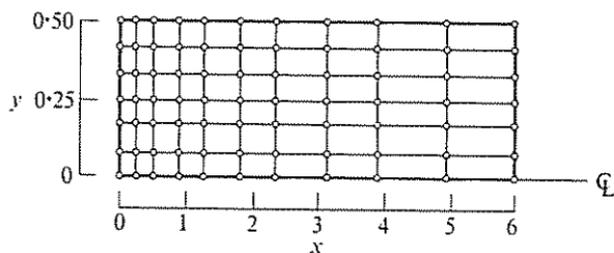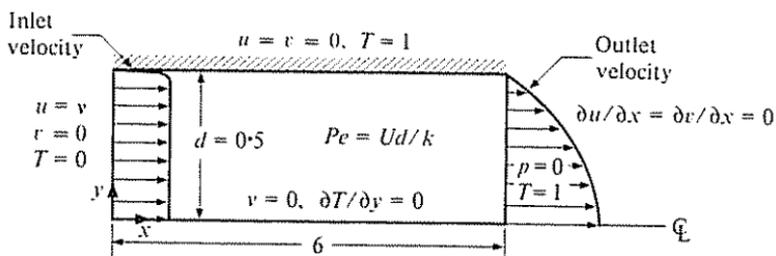(d) Temperature distribution along $y = 0$ for $P_e = 37.5$

Fig. 22.11   Thermal convection—diffusion in entry flow. (a) Problem statement;
(b) Mesh 2—linear elements; (c); and (d) temperature distribution for various P_e
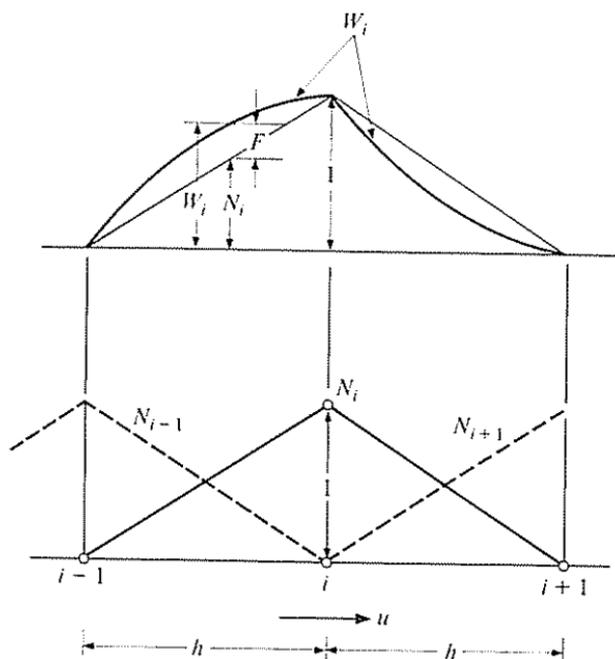numbers and discretization procedures.

Fig. 22.12   One-dimensional problem. Shape functions (N_i) and weighting functions (W_i). Constant velocity u

The 'exact' solution of this difference equation is obtained in reference 54 and is given by

$$\phi_i = A + B \left[\frac{1 + (|\alpha| + 1) \gamma/2}{1 + (|\alpha| - 1) \gamma/2} \right]^i \qquad (22.72)$$

where A and B are constants determined from the boundary conditions. The solution will be oscillatory unless

$$|\alpha| > \alpha_c = 1 - 2/\gamma: \qquad (\text{or } \gamma \leqslant 2). \qquad (22.73)$$

Further it can be shown[54] that the *exact* solution to the original differential equation will be obtained at nodal points if

$$|\alpha| = \alpha_0 = \coth \gamma/2 - 2/\gamma. \qquad (22.74)$$

In Chapter 3 we have indeed noted that in a simple diffusion problem of Fig. 3.4, in which $\gamma = 0$, such exact solutions were obtained with all meshes at the nodal points. The above result generalizes this observation, indicating the *best choice of a weighting function*.

In Fig. 22.13 we show that the stable and optimal values of α differ but little for higher values of γ and due to simplicity of computation the simpler expression (22.73) may be preferred. The process gives results very similar to those achieved in finite difference approaches by using 'upwind differences'. In a recent publication by Barrett[57] the possibility of using such optimal values of α is indeed also suggested.
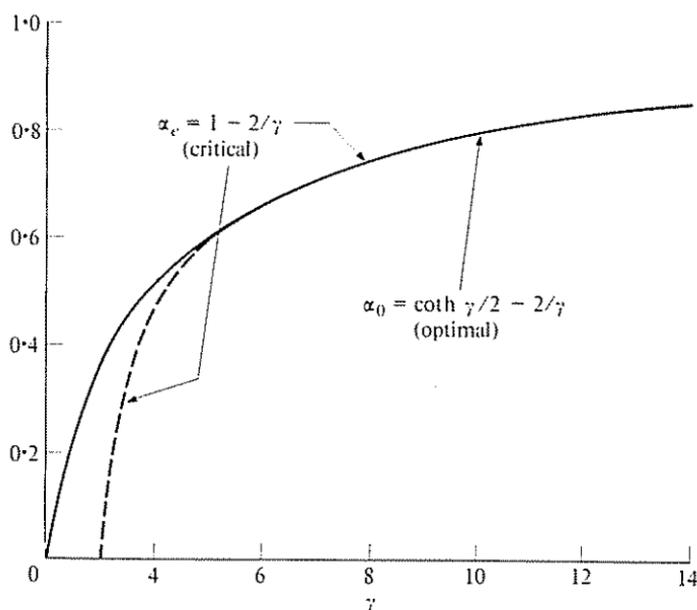

Fig. 22.13   Critical (stable) and optimal values of the 'upwind parameter' α for different values of $\gamma = uh/k'$

To generalize the result to a two- (or three-) dimensional field appears at first sight to be difficult, as simple 'exact' solutions of difference equations are not available. However, proceeding pragmatically the two-dimensional weighting functions can be derived by simply using the appropriate products of such one-dimensional functions as shown in Fig. 22.14. After all this was the basic process of deriving the two-dimensional shape functions discussed in Chapter 7.

To take account of the generally varying velocity field, the optimal α value is chosen in accordance with expression (22.74), depending on the flow *velocity component along the side* (e.g., u_{ij} from node i to node j). Figure 22.14 illustrates how such weighting functions may vary from node to node of the bilinear element. The success and stability of using $\alpha = \alpha_0$ is shown in example of Fig. 22.12.

Processes similar to the one described above can be developed, albeit
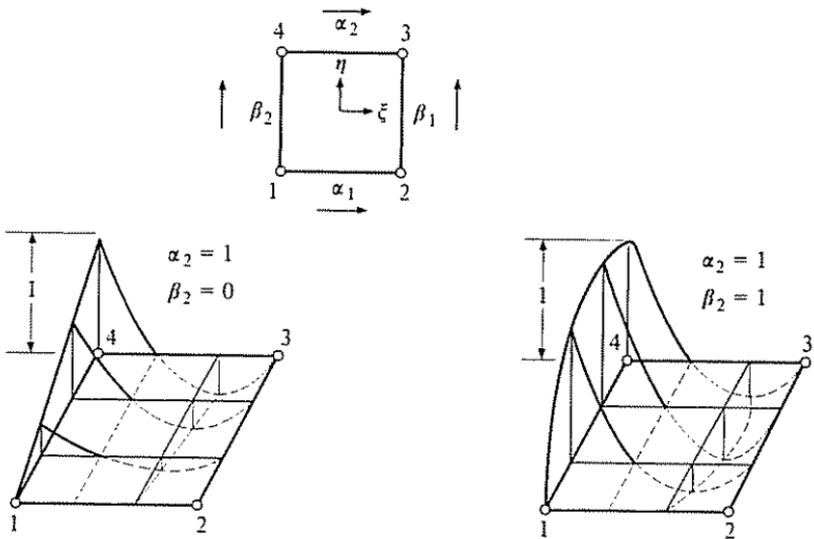
Fig. 22.14 Typical weighting functions for a bi-linear two-dimensional element (parent co-ordinates). Velocity sign convention

with more difficult mathematics, for higher order elements. This has been done in reference 56 for quadratic elements where now two parameters have to be determined along each side. The necessity for doing this is still there as the oscillations develop even to a greater extent with such higher order elements. In Fig. 22.15 we show, for instance, the meaningless results obtained with curved isoparametric quadratic elements for a diffusion problem to a hypothetic estuary and the improvement consequent on the use of 'upwinding'.
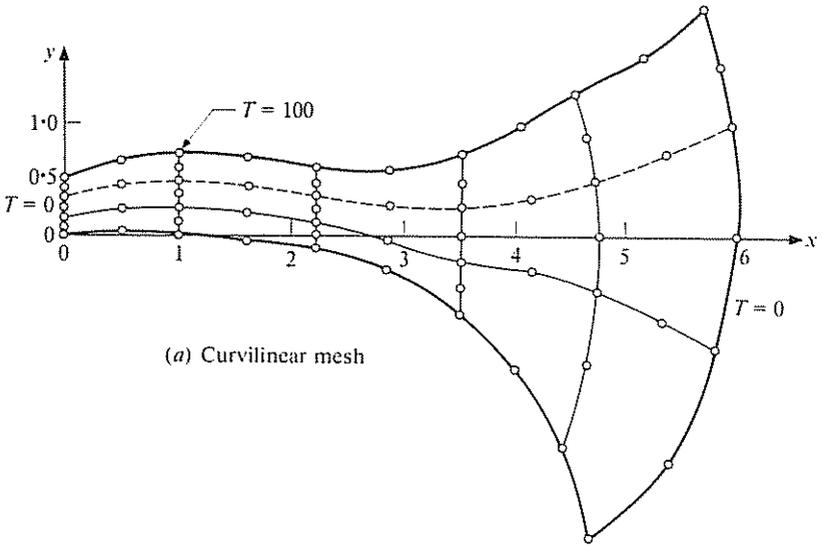
22.9 Some Further Problems of Fluid Mechanics and Concluding Remarks

The scope of this chapter has permitted only a brief mention of some typical fluid flow problems. Compressible subsonic and supersonic flow have not been touched upon despite quite an extensive literature already appearing on finite element approximations in those areas.[58-60]
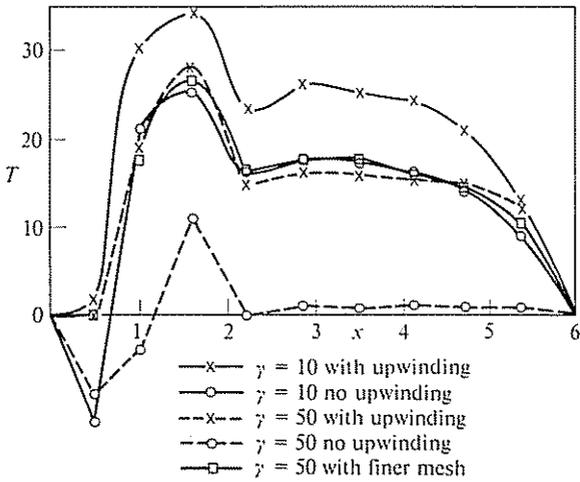
Even with incompressible situations many interesting coupled flow problems could not be discussed. For instance, density changes owing to temperature variation are frequently important contributors to the velocity field development. As the temperature field is in turn influenced by such velocity variations, an iterative complex formulation is necessary. Figure 22.16 from reference 53 shows such a coupled solution in a ventilating duct. Other similar coupled problems of heat generation, and

the consequent temperature changes (which in turn affect the viscosities), are discussed in the context of plastic non-Newtonian flows of metals in references 24, 34, and 56—the latter showing the importance of considering the convective terms by processes discussed in the last section.

While the applications of the finite element method in the field of classical fluid mechanics is fairly obvious, the flow of rocks on a geological
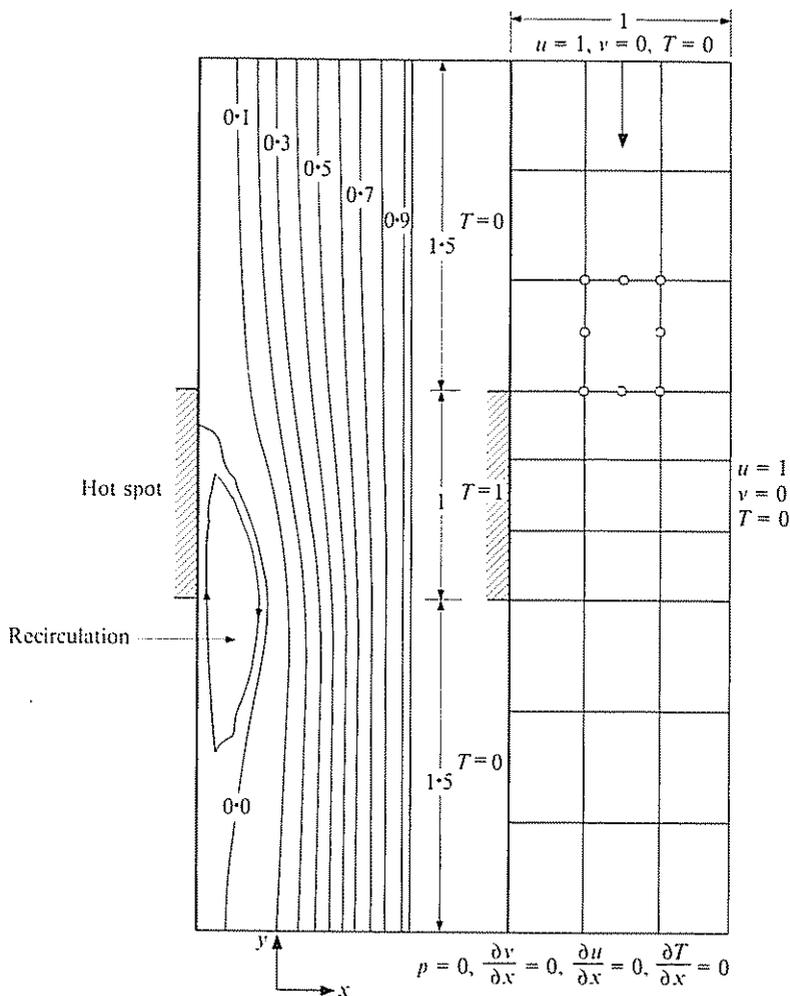

(a) Curvilinear mesh

—×— $\gamma = 10$ with upwinding
—○— $\gamma = 10$ no upwinding
---×-- $\gamma = 50$ with upwinding
--○-- $\gamma = 50$ no upwinding
—□— $\gamma = 50$ with finer mesh

(b) Temperature profile along dotted mesh line in (a)

Fig. 22.15   Steady state convective diffusion of a pollutant in an estuary

Combined convection streamlines $Re = 100$, $Gr = 20,000$, $Pr = 1\cdot0$ heated duct

Fig. 22.16   Recirculation caused by a hot spot in a duct with uniform velocity

time scale is but one of the new areas in which the principles of this chapter are directly applicable. A survey of possible application is given in reference 61.

References

1. J. T. ODEN, O. C. ZIENKIEWICZ, R. H. GALLAGHER, and C. TAYLOR (eds.), *Finite Element Methods in Flow Problems* (Proceedings 1st Symp., Swansea, 1974), Univ. of Alabama Press, 1974.

2. J. T. ODEN, O. C. ZIENKIEWICZ, R. H. GALLAGHER, and C. TAYLOR (eds.), *Finite Elements in Fluids*, Vols. I and II, J. Wiley and Sons, 1975.

3. *Proc. 2nd Int. Symp. on Finite Elements in Fluid Problems ICCAD*, St. Margharita Ligure, Italy, 1976.

4. *Finite Elements in Fluids*, Vol. III, Survey lectures presented at 2nd Int. Symp. at St. Margharita Ligure. To be published by J. Wiley & Sons, 1977.

5. B. L. HEWITT, C. R. ILLINGWORTH, G. C. LOCK, K. W. MANGLER, T. H. McDONELL, C. RICHARDSON and F. WALKDEN (eds.), *Computational Methods and Problems on Aeronautical Fluid Dynamics*, Proc. of Conf. at Univ. of Manchester, Academic Press, 1976.

6. J. J. CONNOR and C. A. BREBBIA, *Finite Element Techniques for Fluid Flow*, Newnes–Butterworths, 1976.

7. H. LAMB, *Hydrodynamics*, 6th ed., Cambridge Univ. Press, 1932.

8. G. K. BATCHELOR, *An Introduction to Fluid Dynamics*, Cambridge Univ. Press, 1967.

9. B. A. FINLAYSON, *The Method of Weighted Residuals and Variational Principles*, Academic Press, 1972.

10. J. T. ODEN and D. SOMOGYI, 'Finite element applications in fluid dynamics', *J. Eng. Mech. Div., Proc. Am. Soc. Civ. Eng.*, **95**, EM4, 821–6, 1969.

11. O. C. ZIENKIEWICZ and C. TAYLOR, 'Weighted residual processes in finite elements with particular reference to some transient and coupled problems', pp. 415–58, *Lectures on Finite Element Method in Continuum Mechanics*, 1970, Lisbon (eds. J. T. Oden and E. R. A. Oliveira), Univ. Alabama Press, Huntsville, 1973.

12. J. H. ARGYRIS and G. MARECZEK, 'Finite element analysis of slow incompressible viscous fluid motion', *Ingenieur Archiv.*, **43**, 92–109, 1974.

13. (a) J. T. ODEN, 'A finite element analog of the Navier–Stokes equations', *Proc. Am. Soc. Civ. Eng.*, **96**, EM4, 529–34, 1970.

13. (b) J. T. ODEN, 'The finite element in fluid mechanics', pp. 151–86, *Lectures on Finite Element Method in Continuum Mechanics*, 1970, Lisbon (eds. J. T. Oden and E. R. A. Oliveira), Univ. Alabama Press, Hunstville, 1973.

14. C. TAYLOR and P. HOOD, 'A numerical solution of the Navier–Stokes equations using the finite element techniques', *Comp. Fluids*, **1**, 73–100, 1973.

15. M. KAWAHARA, N. YOSHIMURA, and K. NAKAGAWA, 'Analysis of steady incompressible viscous flow', *Finite Element Methods in Flow Problems*, pp. 107–20, (eds. J. T. Oden, O. C. Zienkiewicz, R. H. Gallagher, and C. Taylor), Univ. Albama Press, Huntsville, 1974.

16. J. T. ODEN and L. C. WELLFORD Jr., 'Analysis of viscous flow by the finite element method', *J.A.I.A.A.*, **10**, 1590–9, 1972.

17. A. J. BAKER, 'Finite element solution algorithm for viscous incompressible fluid dynamics', *Int. J. Num. Meth. Eng.*, **6**, 89–101, 1973.

18. M. D. OLSON, 'Variational finite element methods for two dimensional and Navier–Stokes equations', Ref. 2, Vol. 1, pp. 57–72, J. Wiley and Sons, 1974.

19. O. C. ZIENKIEWICZ and P. N. GODBOLE, 'Viscous incompressible flow with special reference to non-Newtonian (plastic) flow', Ref. 2, Vol. 1, Ch. 2, pp. 25–71, J. Wiley and Sons, 1975.

20. P. HOOD and C. TAYLOR, 'Navier–Stokes equations using mixed inter-

polation', *Finite Element Method in Flow Problems*, pp. 121–32 (eds. J. T. Oden, O. C. Zienkiewicz, R. H. Gallagher, and C. Taylor), Univ. Alabama Press, Huntsville, 1974.

21. M. D. OLSON and S. Y. TUENN, 'Primitive variables versus stream function finite element solutions of the Navier–Stokes equation', pp. 55–68 of ref. 3.

22. B. ATKINSON, C. C. M. CARD, and B. M. IRONS, 'Application of the finite element method to creeping flow problems', *Trans. Inst. Chem. Eng.*, **48**, 276–84, 1970.

23. T. J. R. HUGHES, R. L. TAYLOR, and J. F. LEVY, 'A finite element method for incompressible viscous flows', pp. 1–16 of ref. 3.

24. P. C. JAIN, *Plastic flow in solids. Static, quasistatic and dynamic situations including temperature effects*, Ph.D. Thesis, Univ. of Wales, Swansea, to be submitted, 1976.

25. H. S. LEW and Y. C. FUNG, 'On low Reynolds number entry flow into a circular tube', *J. Bio-mech.*, **2**, 105–19, 1969.

26. P. N. GODBOLE, 'Creeping flow in rectangular ducts by the finite element method', *Int. J. Num. Meth. Eng.*, **9**, 727–30, 1975.

27. O. C. ZIENKIEWICZ and P. N. GODBOLE, 'Flow of plastic and visco-plastic solids with special reference to extrusion and forming processes', *Int. J. Num. Meth. Eng.*, **8**, 3–16, 1974.

28. K. PALIT and R. T. FENNER, 'Finite element analysis of two dimensional slow non-Newtonian flows', *A.I.Ch.E. Jl*, **18**, 1163–9, 1972.

29. K. PALIT and R. T. FENNER, 'Finite element analysis of slow non-Newtonian channel flow', *A.I.Ch.E. Jl*, **18**, 628–33, 1972.

30. O. C. ZIENKIEWICZ and P. N. GODBOLE, 'Penalty function approach to problems of plastic flow of metals with large surface deformations', *J. Strain Analysis*, **10**, 180–3, 1975.

31. M. KAWAHARA, N. YOSHIMURA, K. NAKAGAWA and H. OHSAKA, 'Steady and unsteady finite element analysis of incompressible viscous flow', *Int. J. Num. Meth. Eng.*, **10**, 437–56, 1976.

32. S. L. SMITH and C. A. BREBBIA, 'Finite element solution of Navier–Stokes equations for transient 2-dimensional incompressible flow', *J. Comp. Phys.*, **17**, 235–45, 1975.

33. C. H. LEE, 'Finite element method for transient linear viscous flow problems', *Proc. Int. Conf. on Numerical Methods in Fluid Dynamics*, 1973.

34. O. C. ZIENKIEWICZ, P. C. JAIN and E. OÑATE, *Flow of solids during forming and extrusion. Some aspects of numerical solutions*. Univ. College of Swansea, Report No. C/R/283/76.

35. G. C. CORNFIELD and R. H. JOHNSON, 'Theoretical prediction of plastic flow in hot rolling including the effect of various temperature distributions', *J. Iron Steel Inst.*, **211**, 567–73, 1973.

36. J. W. H. PRICE and J. M. ALEXANDER, 'The finite element analysis of two high temperature metal deformation processes', pp. 715–20 of ref. 3.

37. R. E. NICKELL, R. I. TANNER, and B. CASWELL, 'The solution of viscous incompressible jet and free surface flows using finite element methods', *J. Fluid Mech.*, **65**, Part 1, 189–206, 1974.

38. J. J. DRONKERS, 'Tidal computation for rivers, coastal waters and seas', *Proc. Am. Soc. Civ. Eng.*, **95**, H71, 29–77, 1969.

39. P. WELANDER, 'Wind action on shallow sea, some generalisations of Eckmann's theory', *Tellus*, **9**, 47–52, 1957.

40. R. T. CHENG, T. M. POWELL and T. M. DILLON, 'Numerical models of wind driven circulation in lakes'. *Appl. Math. Modelling*, **1**, 141–59, 1976.

41. M. Kawahara and K. Hasegawa, 'Periodic Galerkin finite element method of tidal flow', *Int. J. Num. Meth. Eng.*, **12**, 115–27, 1978.

42. T. Tenaka, T. Hirai and T. Katayama, 'Finite element applications to lake circulations on diffusion problems in Lake Drive', *Proc. 2nd Int. Symp. on Finite Elements in Fluid Problems*, St. Margharita, Italy, 1976.

43. J. Connor and J. Wang, 'Finite element modelling of hydrodynamic circulation' from *Numerical Methods in Fluid Dynamics* (eds. C. Brebbia and J. Connor), Pentech Press, 1974.

44. C. Taylor and J. M. Davis, 'Tidal propagation and dispersion in estuaries', Ref. 2, Vol. 1, Ch. 5, pp. 95–118, J. Wiley, 1975.

45. P. F. Hamblin, 'Finite element methods approach to the modelling of circulation, seiches, tides and storm surges in large lakes', (see ref. 3).

46. R. T. Cheng, 'Numerical investigation of Lake circulation around islands by the finite element method', *Int. J. Num. Meth. Eng.*, **5**, 103–12, 1972.

47. F. Arrizebalaya, G. M. Kovadi, and R. J. Krizek, 'Variational model for lake circulation' (see ref. 3).

48. R. T. Cheng and C. Tung, 'Wind driven lake circulation by the finite element method', *Proc. 13th Conf. on Great Lakes Research*, 1970.

49. S. Crandall, *Engineering Analysis*, McGraw-Hill, 1956.

50. R. Courant, E. Isaacson, and M. Rees, 'On the solution of non-linear hyperbolic differential equations by finite differences', *Comm. Pure Appl. Math.*, **V**, 243–55, 1952.

51. A. K. Runchal and M. Wolfstein, 'Numerical integration procedure for the steady state Navier–Stokes equations', *J. Mech. Eng. Sci.*, **11**, 445–53, 1969.

52. D. B. Spalding, 'A novel finite difference formulation for differential equations involving both first and second derivatives', *Int. J. Num. Meth. Eng.*, **4**, 551–9, 1972.

53. O. C. Zienkiewicz, R. H. Gallagher, and P. Hood, 'Newtonian and non-Newtonian viscous incompressible flow. Temperature induced flows. Finite element solutions', *The Mathematics of finite elements and applications II*, ed. J. Whiteman, Academic Press, 1977.

54. I. Christie, D. F. Griffiths, A. R. Mitchell and O. C. Zienkiewicz, 'Finite element methods for second order differential equations with significant first derivatives', *Int. J. Num. Meth. Eng.*, **10**, 1389–96, 1976.

55. O. C. Zienkiewicz, J. C. Heinrich, P. S. Huyakorn and A. R. Mitchell, 'An upwind finite element scheme for two dimensional convective transport equations', *Int. J. Num. Meth. Eng.*, **11**, 131–44, 1977.

56. J. C. Heinrich and O. C. Zienkiewicz, 'Quadratic finite element schemes for two dimensional convective-transport problems', *Int. J. Num. Meth. Eng.*, **1**, 1831–44, 1977.

57. K. E. Barrett, 'The numerical solution of singular perturbation boundary value problems', *Q. J. Mech. Appl. Math.*, **27**, 57–68, 1974.

58. J. Periaux, 'Three dimensional analysis of compressible potential flow', *Int. J. Num. Mech. Eng.*, **9**, 775–83, 1975.

59. T. E. Laskaris, 'Finite element analysis of compressible and incompressible viscous flow and heat transfer problems', *Physics of Fluids*, **18**, 1639–48, 1975.

60. S. G. Margolis, 'Finite element methods for compressible gas dynamic steam review', **17**, 385, 1975.

61. O. C. Zienkiewicz, 'The finite element method and the solution of some geophysical problems', *Phil. Trans. R. Soc. Lond. A.*, **283**, 139–51, 1976.

Momentum Transfer - Fluid Mechanics – Navier-Stokes Equations (Incompressible)

Tensor:

$$\rho\frac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v}\cdot\nabla)\mathbf{v} = \mathbf{F} - \nabla P + \mu\nabla^2\mathbf{v} \quad \text{with} \quad \nabla^2 = \nabla\cdot\nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{1}$$

$$\nabla\cdot\mathbf{v} = 0 \tag{2}$$

Expanded:

$$\rho\frac{\partial}{\partial t}\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} + \rho[v_x\frac{\partial}{\partial x} + v_y\frac{\partial}{\partial y} + v_z\frac{\partial}{\partial z}]\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} - \begin{bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{bmatrix}P + \mu[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}]\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{3}$$

$$[\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}] = 0 \tag{4}$$

In 2-D:

$$\rho[v_x\frac{\partial v_x}{\partial x} + v_y\frac{\partial v_x}{\partial y}] = F_x - \frac{\partial P}{\partial x} + \mu[\frac{\partial^2 v_x}{\partial x^2} + \frac{\partial^2 v_x}{\partial y^2}] \tag{5}$$

$$\rho[v_x\frac{\partial v_y}{\partial x} + v_y\frac{\partial v_y}{\partial y}] = F_y - \frac{\partial P}{\partial y} + \mu[\frac{\partial^2 v_y}{\partial x^2} + \frac{\partial^2 v_y}{\partial y^2}] \tag{6}$$

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_x}{\partial y} = 0 \tag{7}$$

NAVIER-STOKES EQUATIONS

2-D Equations:

$$\rho \left[v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} \right] = F_x - \frac{\partial P}{\partial x} + \mu \left[\frac{\partial^2 v_x}{\partial x^2} + \frac{\partial^2 v_x}{\partial y^2} \right]$$

$$\rho \left[v_x \frac{\partial v_y}{\partial x} + v_y \frac{\partial v_y}{\partial y} \right] = F_y - \frac{\partial P}{\partial y} + \mu \left[\frac{\partial^2 v_y}{\partial x^2} + \frac{\partial^2 v_y}{\partial y^2} \right]$$

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = 0$$

Consider the operators we have:

$$A \frac{\partial c}{\partial t} \neq -B \left(\frac{\partial c}{\partial x} + \frac{\partial c}{\partial y} \right) + C \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} \right)$$

$$\underline{q} = K_0 \underline{c} \qquad \underline{q} = K_I \underline{c} \qquad \underline{q} = \frac{K}{II} \underline{c}$$

① Ignore convective acceleration (Stokes flow) $\nabla P = \mu \nabla^2 v$ and $\nabla \cdot V = 0$

$$-F_x = K_I^x P + \mu K_{II} v_x$$

$$-F_y = K_I^y P + \mu K_{II} v_y$$

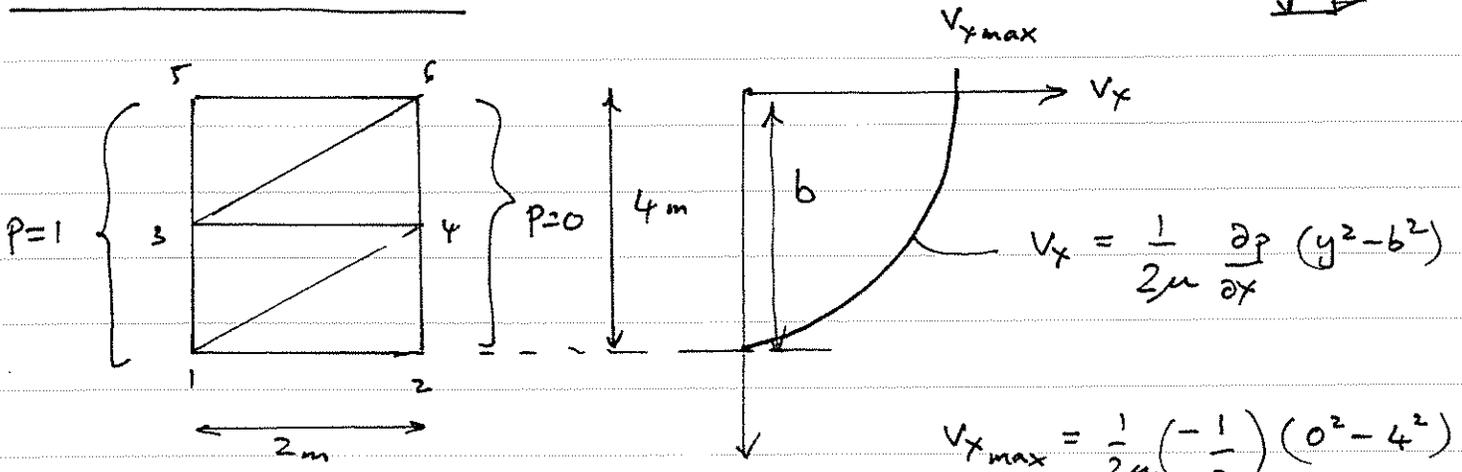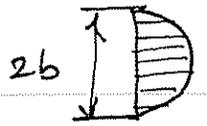$$0 = +K_I^{xT} v_x + K_I^{yT} v_y$$

$$F_x = \mu K_{II} v_x + K_I^x P$$

$$F_y = \mu K_{II} v_y + K_I^y P$$

$$0 = K_I^{xT} v_x + K_I^{yT} v_y$$

$$\begin{Bmatrix} F_x \\ F_y \\ 0 \end{Bmatrix} = \begin{bmatrix} \mu K_{II} & & K_I^x \\ & \mu K_{II} & K_I^y \\ K_I^{xT} & K_I^{yT} & \end{bmatrix} \begin{Bmatrix} v_x \\ v_y \\ P \end{Bmatrix}$$

EXAMPLE DATA FILE

$V_{x max}$

V_x
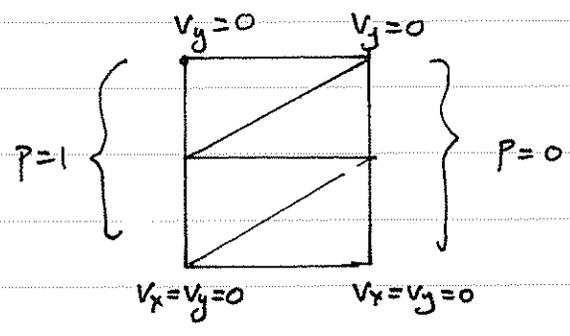
$P=1$ 5 ... 6 $P=0$ 4 m b

3 ... 4

1 ... 2

2 m

$$V_x = \frac{1}{2\mu} \frac{\partial p}{\partial x} (y^2 - b^2)$$

$$V_{x max} = \frac{1}{2\mu} \left(-\frac{1}{2}\right) (0^2 - 4^2)$$

$$= -\frac{1}{4\mu} (-4^2) = \frac{4}{\mu}$$

Boundary conditions

$V_y = 0$ $V_y = 0$

$P=1$ { } $P = 0$

$V_x = V_y = 0$ $V_x = V_y = 0$

```matlab
function [ Kl,Cl,Ml ] = elmt23( local_data )
% elmt23: 2-D three-node (triangular) element for Navier-Stokes flow
% (Constant gradient triangle, etc.)
%
%Element subroutine to construct local element matrix from local data
%
%   local_data.mater = [elt# Hyd.Cond Elt-thickness prop3 prop4 prop5 ] For material type
%
%   local_data.coords = [ x1 x2 of node 1;
%                         x1 x2 of node 2;
%                         x1 x2 of node 3]
%
%   local_data.dofs   = [ dof1 dof2 dof3] Global dof referenced to local dofs
%
   Kl = zeros(9,9);
   Cl = zeros(9,9);
   Ml = zeros(9,9);
   D  = zeros(2,2);
   A  = zeros(2,3);
%
%   Constitutive matrix - Darcy's Law, Fick's Law, Four
%
   density     = local_data.mater(1,2);
   viscosity   = local_data.mater(1,3);
   thickness   = 1

%----------------------------- Evaluate coefs and determinant for element area
    b = local_data.coords(2,1)*local_data.coords(3,2) - local_data.coords(3,1)*local_data.↵
coords(2,2);
    b21 = local_data.coords(2,2) - local_data.coords(3,2);
    b31 = local_data.coords(3,1) - local_data.coords(2,1);
    d2  = local_data.coords(1,1)*b21 + local_data.coords(1,2)*b31 + b;
    area = d2/2;
%----------------------------- Evaluate derivatives of shape functions
A = (1/d2) * [ b21  (local_data.coords(3,2)-local_data.coords(1,2))  (local_data.coords↵
(1,2)-local_data.coords(2,2));
               b31  (local_data.coords(1,1)-local_data.coords(3,1))  (local_data.coords↵
(2,1)-local_data.coords(1,1))];

%----------------------------- Evaluate second order matrices
    D = [ 1 0; 0 0 ];  secondorderinx = zeros(3,3); secondorderinx = A'*(D*A)↵
*area*thickness
    D = [ 0 0; 0 1 ];  secondorderiny = zeros(3,3); secondorderiny = A'*(D*A)↵
*area*thickness

%----------------------------- Evaluate first order matrices
    factor = ones(3,1)/3
    v = [ 1; 0 ];       firstorderinx = zeros(3,3);  firstorderinx = factor*(v'*A)↵
*area*thickness
    v = [ 0; 1 ];       firstorderiny = zeros(3,3);  firstorderiny = factor*(v'*A)↵
*area*thickness
```

```
%------------------------------ Assemble local stiffness matrix [ vx1 vx2 vx3 vy1 vy2↵
vy3 p1 p2 p3 ]
    null = zeros(3,3)

    Kl = [ viscosity*(secondorderinx+secondorderiny)    null↵
firstorderinx ;
           null                                         viscosity*↵
(secondorderinx+secondorderiny)      firstorderiny;
           firstorderinx'                               firstorderiny'↵
null ]

%------------------------------ Reorder stiffness matrix for dof
% [vx1 vx2 vx3 vy1 vy2 vy3 p1 p2 p3] ==> [vx1 vy1 p1 vx2 vy2 p2 vx3 vy3 p3]

% ----- Reorder columns
            tempcol = Kl(:,2); Kl(:,2) = Kl(:,4); Kl(:,4) = tempcol;
            tempcol = Kl(:,3); Kl(:,3) = Kl(:,7); Kl(:,7) = tempcol;
            tempcol = Kl(:,6); Kl(:,6) = Kl(:,8); Kl(:,8) = tempcol;
% ----- Reorder rows
            temprow = Kl(2,:); Kl(2,:) = Kl(4,:); Kl(4,:) = temprow;
            temprow = Kl(3,:); Kl(3,:) = Kl(7,:); Kl(7,:) = temprow;
            temprow = Kl(6,:); Kl(6,:) = Kl(8,:); Kl(8,:) = temprow;

%------------------------------ Evalaute consistent mass (storage) matrix
%     Cl = (area*thickness/3) * [ 1 0 0; 0 1 0; 0 0 1];
```

Note:

Pressure gradient: $\begin{Bmatrix} \partial p/\partial x \\ \partial p/\partial y \end{Bmatrix} = \begin{Bmatrix} \partial/\partial x \\ \partial/\partial y \end{Bmatrix} P = \begin{Bmatrix} \underline{K}_I^x \\ \underline{K}_I^y \end{Bmatrix} \underline{P}$

$\hookleftarrow \underline{K}_I$

Continuity term: $\begin{bmatrix} \dfrac{\partial v_x}{\partial x} + \dfrac{\partial v_y}{\partial y} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial}{\partial x} & ; & \dfrac{\partial}{\partial y} \end{bmatrix} \begin{Bmatrix} v_x \\ v_y \end{Bmatrix} = \underline{K}_I^T \begin{Bmatrix} \underline{v_x} \\ \underline{v_y} \end{Bmatrix}$

$\underline{K}_I^T \equiv \begin{bmatrix} \underline{K}_I^{xT} & \underline{K}_I^{yT} \end{bmatrix}$

ALTERNATIVE APPROACH — MATRICES FOR N-S.

$$\frac{\partial P}{\partial x} \equiv \Bigg\} \underline{a}^T \underline{m} P \qquad = \underline{a}^T \underline{m} \underline{N} \underline{P}$$

$$\frac{\partial P}{\partial y} \equiv \Bigg\} \qquad \underset{P = \underline{N}\underline{P}}{\overset{}{}}$$

$$\dot{\varepsilon}_v = \dot{\varepsilon}_x + \dot{\varepsilon}_y = \underline{m}^T \dot{\underline{\varepsilon}} = \overset{\underline{N}^T}{} \underline{m}^T \underline{a} \dot{\underline{v}} \qquad \equiv \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = 0$$

$$\dot{\varepsilon}_v = \underline{N}^T \underline{m}^T \underline{a} \; \dot{\underline{v}}$$

$$\therefore \quad \begin{bmatrix} \underline{K} & \underline{K}_b \\ \underline{K}_b^T & 0 \end{bmatrix} \begin{Bmatrix} \underline{v} \\ \underline{P} \end{Bmatrix} = \begin{Bmatrix} f \\ 0 \end{Bmatrix}$$

5

Momentum Transport - Solids

[5:1] Solid Mechanics

Principle of virtual work

1D element
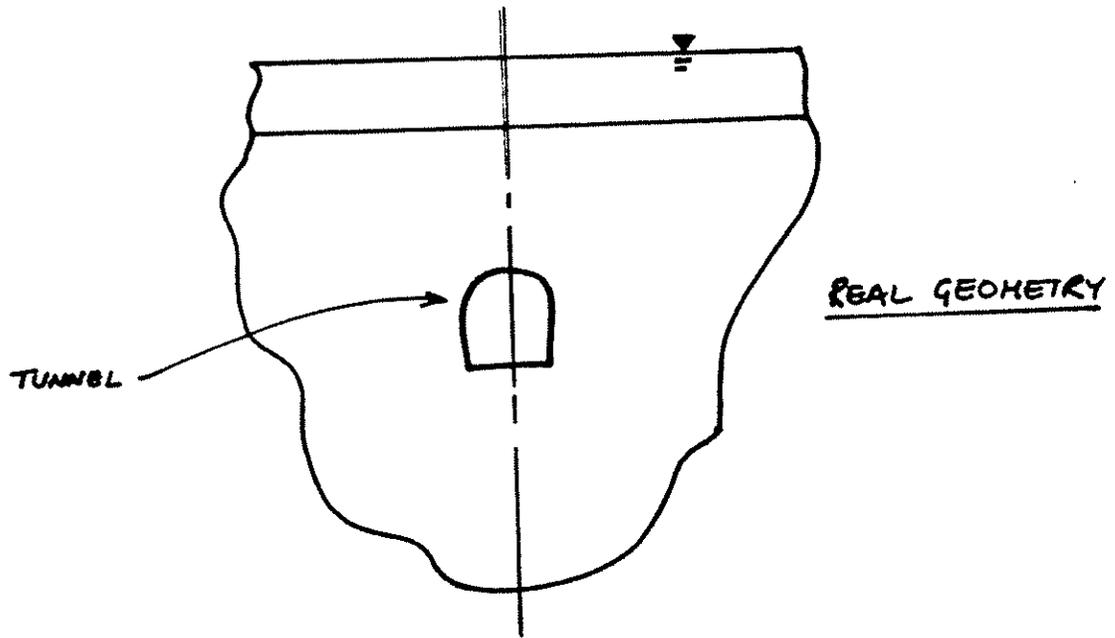
2D element

PROCESS COUPLINGS [T-H-M-C]

$$\begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{p} \\ \underline{T} \\ \underline{c} \end{Bmatrix} + \begin{bmatrix} S_{11} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \begin{Bmatrix} \dot{\underline{u}} \\ \dot{\underline{p}} \\ \dot{\underline{T}} \\ \dot{\underline{c}} \end{Bmatrix} = \begin{Bmatrix} \dot{\underline{f}} + \dots \\ \underline{q}_F + \dots \\ \underline{q}_T + \dots \\ \underline{q}_M + \dots \end{Bmatrix}$$

FINAL EQUATION

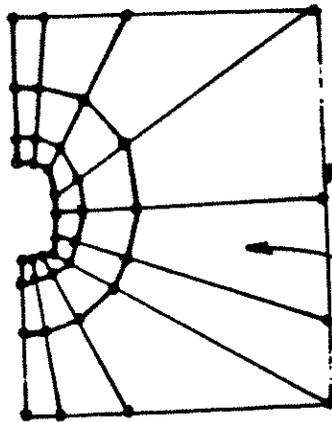$$\underline{f} = \underline{K}\,\underline{u} \qquad or \qquad \dot{\underline{f}} = \underline{K}\,\dot{\underline{u}}$$

$$K = \int_V \underline{a}^T \underline{D}\,\underline{a}\,dV$$

$$\begin{cases} \underline{\varepsilon} = \underline{a}\,\underline{u} \qquad or \qquad \dot{\underline{\varepsilon}} = \underline{a}\,\dot{\underline{u}} \\ \underline{u} = \underline{b}\,\underline{U} \\ \underline{\sigma} = \underline{D}\,\underline{\varepsilon} \end{cases}$$
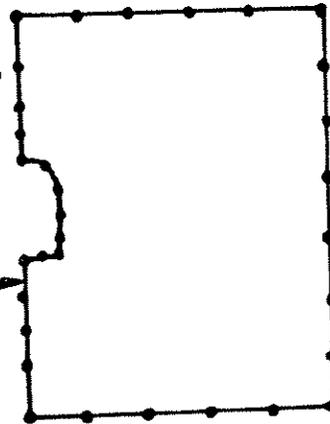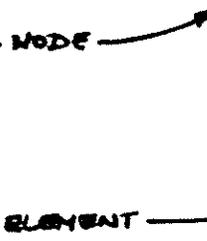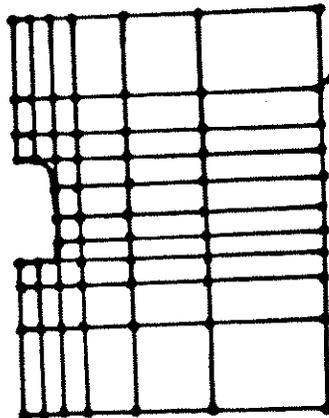
Figure 1.1 Domain and Integral Representations of a Flow Problem

SYSTEM TYPES

SOLID MECHANICS

- Conservation of momentum: (Equilibrium), $VW_I = VW_E$

- Continuity (Compatibility):
$$\varepsilon = a u$$

- Constitutive relation: $\sigma = D \varepsilon$

- Initial Conditions

- Boundary Conditions

FLOW SYSTEM

- Conservation of mass:
$$\nabla^T q = 0$$

- Continuity: $h_, = a h$

- Constitutive reln. $v = D h_,$

- ICs

- BCs

TRANSPORT

- Conservation of mass
$$\nabla^T q = 0$$

- Continuity: $c_, = a c$

- Constitutive:
   diffusion - $v_1 = D c_,$
   advective - $v_2 = A c$

- ICs

- BCs

— SOLVE SYSTEM EQUATIONS —

BASIC EQUATIONS — COMMONALITY IN SOLUTION

FLOW $\quad q = \underline{K}\underline{h}$

$$\underline{K} = \int_V \underline{a}^T \underline{D} \underline{a} \, dV$$
$$\underline{h}_, = \underline{a} \, \underline{h}$$
$$\underline{v} = \underline{D} \, \underline{h}_,$$

TRANSPORT $\quad \underline{q} = \underline{K} \underline{c}$

$$\underline{K} = \underline{k}_1 + \underline{k}_2$$
$$\underline{k}_1 = \int_V \underline{a}^T \underline{D} \underline{a} \, dV$$
$$\underline{k}_2 = \int_V \underline{b}^T \underline{v} \, \underline{a} \, dV$$

SOLID MECHANICS $\quad \underline{f} = \underline{K} \underline{u}$

$$\underline{K} = \int_V \underline{a}^T \underline{D} \underline{a} \, dV$$
$$\underline{E} = \underline{a} \, \underline{u}$$
$$\underline{\sigma} = \underline{D} \, \underline{E}$$

* ALL EQUATIONS REDUCE TO A SET OF SIMULTANEOUS
ALGEBRAIC EQUATIONS TO BE SOLVED FOR THE
DEPENDENT VARIABLES ONCE BOUNDARY CONDITIONS ARE
APPLIED.

* FEM CODES ARE STRUCTURED TO ALLOW:

    o VARIABLE ELEMENT TYPES

    o VARIABLE DIMENSIONALITY $\quad 1-D \longrightarrow 3-D$

    o SYMMETRIC/NONSYMMETRIC MATRICES

    o ITERATIVE SOLUTION

SOLID MECHANICS — PRINCIPLE OF VIRTUAL WORK

- Turner, Martin, Clough & Topp (1956) ; Argyris (1955)

PRINCIPLE OF VIRTUAL WORK (VW)

Definition: An elastic body is in equilibrium, if, for any
set of virtual displacements, $\bar{u}$, the virtual work of
the external forces is equal to the virtual strain energy
of the system (i.e. internal work)

$$\boxed{VW_E = VW_I} \qquad (1)$$

CONSIDER A DOMAIN

VOLUME, V
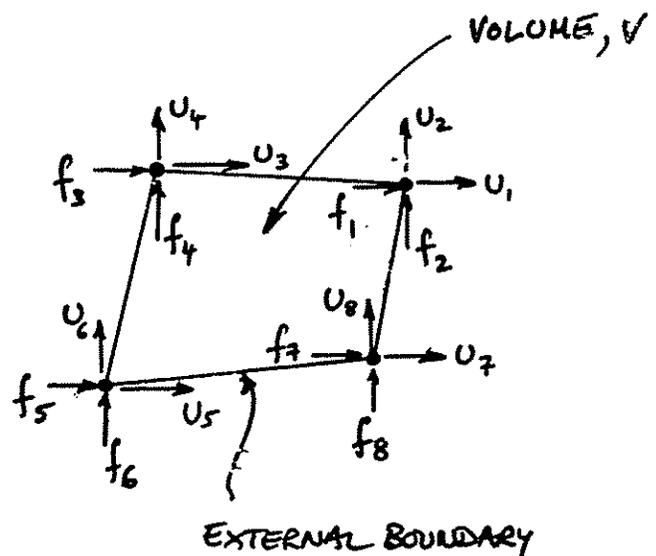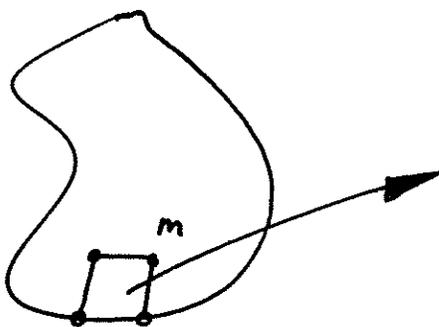
EXTERNAL BOUNDARY

Isolate element, m

$$\boxed{VW_E^{(m)} = VW_I^{(m)}} \qquad (2)$$

EXTERNAL VIRTUAL WORK, VW_E

$$VW_E = \text{Boundary forces} \times \text{Boundary displacements (virtual)}$$

For element with n degrees of freedom (example has $n = 8$):

$$\underline{u}^T = [u_1, u_2 \ldots u_n] \quad ; \quad \underline{f}^T = [f_1, f_2 \ldots f_n]$$

Sign convention: bar subscript $\equiv$ vector   eg. $\underline{u}$

bar superscript $\equiv$ virtual   e.g. $\bar{u}$

then

$$\boxed{VW_E = \sum_{i=1}^{n} \bar{u}_i f_i = \underline{\bar{u}}^T \underline{f}}$$

(1)

INTERNAL VIRTUAL WORK, VW_I

$$VW_I = \text{Body strain resulting from virtual displ.} \times \text{Stress change due to boundary conditions}$$

For 2-D system: $\underline{\bar{\varepsilon}}^T = [\varepsilon_x \ ; \ \varepsilon_y \ ; \ \gamma_{xy}] ; \quad \underline{\sigma} = \begin{Bmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{Bmatrix}$

$$\gamma_{xy} = (\alpha_1 + \alpha_2)$$
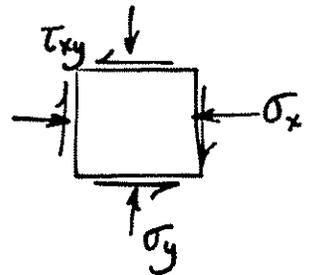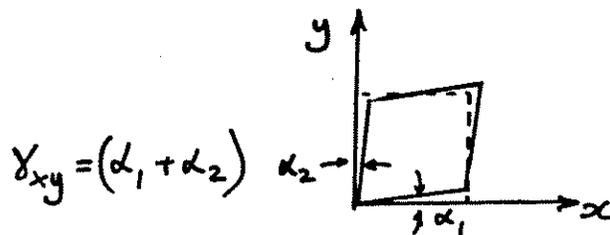
FOUR REQUIREMENTS TO SOLVE THE ELASTIC BOUNDARY VALUE PROBLEM

Equilibrium: Satisfied if $VW_E = VW_I$ (2)

Constitutive Law: $\underline{\sigma} = \underline{D}\,\underline{\varepsilon}$ (3)

$\underline{D}$ = Elasticity matrix (3×3)

Strain-displacement (compatability): $\underline{\varepsilon} = \underline{a}\,\underline{u}$ (4)

Boundary conditions :   supplied

Definition of $VW_I = \displaystyle\int_{Vol} \underline{\varepsilon}^T \underline{\sigma}\, dV$ (5)

Substituting (4) into (3) $\underline{\sigma} = \underline{D}\,\underline{a}\,\underline{u}$ (6)

Substituting (6) into (5) $VW_I = \displaystyle\int_V (\underline{a}\,\underline{\bar{u}})^T \underline{D}\,\underline{a}\,\underline{u}\, dV$ (7)

Note that $(\underline{x}\,\underline{y})^T = \underline{y}^T \underline{x}^T$ then; (7) becomes

$$\boxed{VW_I = \underline{\bar{u}}^T \int_V \underline{a}^T \underline{D}\,\underline{a}\, dV\, \underline{u}}$$ (8)

Recall that: $\boxed{VW_E = \underline{\bar{u}}^T \underline{f}}$ (9)

$\left.\vphantom{\begin{array}{c}a\\b\\c\end{array}}\right\}$ EQUATE

Equating (8) & (9) $\underline{\bar{u}}^T \underline{f} = \underline{\bar{u}}^T \underbrace{\int_V \underline{a}^T \underline{D}\,\underline{a}\, dV}_{\underline{K}}\, \underline{u}$ (10)

$$\boxed{\underline{f} = \underline{K}^{(m)} \underline{u}}$$
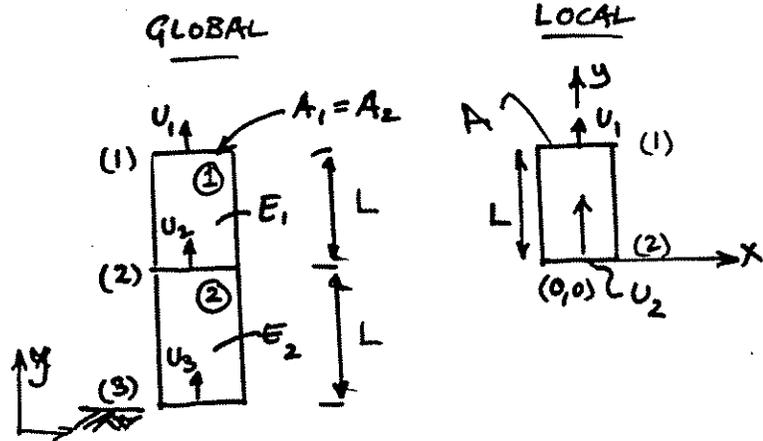
EXAMPLE 1-D PROBLEM

GLOBAL   LOCAL

For element ①;

$$\underline{f} = \underline{K}\,\underline{u} \qquad (1)$$

$$\underline{K} = \int_V \underline{a}^T \underline{D}\,\underline{a}\,dv \qquad (2) \quad \text{with} \qquad \underline{\sigma} = \underline{D}\,\underline{\varepsilon} \qquad (3)$$

$$\underline{\varepsilon} = \underline{a}\,\underline{u} \qquad (4)$$

$\underline{\underline{D}}$ matrix $\qquad \underline{D} = [E_1] \qquad \text{since} \quad E = \dfrac{\sigma}{\varepsilon} \qquad (5)$

$\underline{\underline{a}}$ matrix $\qquad \underline{\varepsilon} = \dfrac{\partial U_y}{\partial y} = \dfrac{\partial}{\partial y}(U_y) \qquad (6)$

$$\text{assume} \quad U_y = U_2 + (U_1 - U_2)\dfrac{y}{L} \qquad (7)$$

Then $\qquad U_y = U_1$ @ node ① when $y = L$

$\qquad U_y = U_2$ @ node ② when $y = 0 \qquad$ QED.

Substituting (7) into (6) then:

$$\underline{\varepsilon} = \dfrac{\partial}{\partial y}(U_y) = (U_1 - U_2)\dfrac{1}{L} = \dfrac{1}{L}\underbrace{[1 \; ; \; -1]}_{\underline{a}}\underbrace{\begin{Bmatrix} U_1 \\ U_2 \end{Bmatrix}}_{\underline{u}} \qquad (8)$$

Substituting (5) and (8) into (2)

$$\underline{K} = \int_V \underline{a}^T \underline{D}\,\underline{a}\,dv = A\int_0^L \dfrac{1}{L}\begin{bmatrix} 1 \\ -1 \end{bmatrix}[E_1]\dfrac{1}{L}[1 \; ; \; -1]\,dy \qquad (9)$$

Rearranging (9): $\quad K = A \int_0^L \frac{E}{L^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} dy \qquad$ (10)

Taking individual terms from (10), then $\quad \frac{AE}{L^2} \int_0^L 1 \, dy = \frac{AE}{L}$

and resubstituting into (10) then:

$$\underline{K}^{(1)} = \frac{A_1 E_1}{L_1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad (1)$$

and similarly for (2)

$$\underline{K}^{(2)} = \frac{A_2 E_2}{L_2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad (2)$$

MOVING TO THE GLOBAL LEVEL $\quad \sum_1^m VW_E^{(m)} = \sum_1^n VW_{\not{g}_I}^{(m)}$

$$\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix} = \begin{matrix} \text{①} \\ \text{②} \\ \text{③} \end{matrix} \begin{bmatrix} \frac{A E_1}{L_1} & -\frac{A E_1}{L_1} & 0 \\ -\frac{A E_1}{L_1} & \left(\frac{A E_1}{L_1} + \frac{A E_2}{L_2}\right) & -\frac{A E_2}{L_2} \\ 0 & -\frac{A E_2}{L_2} & \frac{A E_2}{L_2} \end{bmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \end{Bmatrix} \quad (11)$$
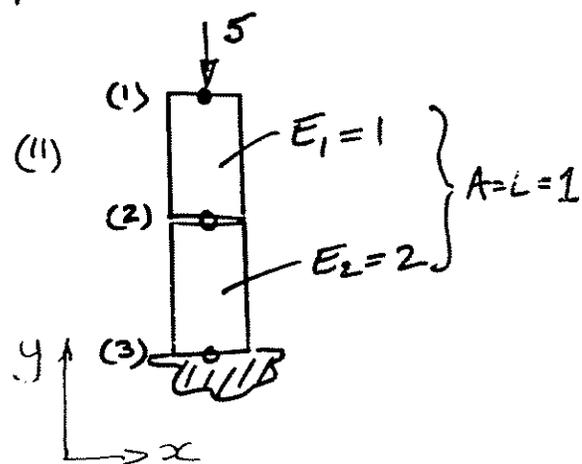
Then the system of equations becomes:

$$\begin{Bmatrix} -5 \\ f_2 \\ f_3 \end{Bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & (1+2) & -2 \\ 0 & -2 & 2 \end{bmatrix} \begin{Bmatrix} \underline{u}_1 \\ \underline{U}_2 \\ 0 \end{Bmatrix} \quad (12)$$

$U_1 \, ; \, U_2 \, ; \, f_3$ are unknowns

$f_2 = 0$ since internal node with no <u>prescribed</u> force.

Rearrange (12) for boundary conditions:

$$-\left\{\begin{array}{c} -5 - (0)(0) \\ 0 - (-2)(0) \\ \hline f_3 - (2)(0) \end{array}\right\} = \left[\begin{array}{cc} 1 & -1 \\ -1 & 3 \\ \hline 0 & -2 \end{array}\right] \left\{\begin{array}{c} u_1 \\ u_2 \end{array}\right\} \qquad (13)$$

Solve as: $\qquad u_1 = -15/2 \qquad ; \qquad u_2 = -5/2$

Resubstitute as: $\qquad f_3 = +5 \quad \checkmark \quad QED$

Alternatively: Shortening of element $(2) = u_2 - u_3 = -5/2$

$\qquad\qquad\qquad\qquad\qquad\qquad (1) = u_1 - u_2 = -10/2 \quad \checkmark$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad QED$

THREE-NODED CONSTANT STRAIN ELEMENT

$i, j, k = $ nodes 1, 2, 3.

$$\underline{K} = \int_V \underline{a}^T \underline{D} \, \underline{a} \, dV$$

- Six degrees of freedom.
- Assume linear variation in displ. for both U_x and U_y since 3 values of each are required to define each of two planar surfaces
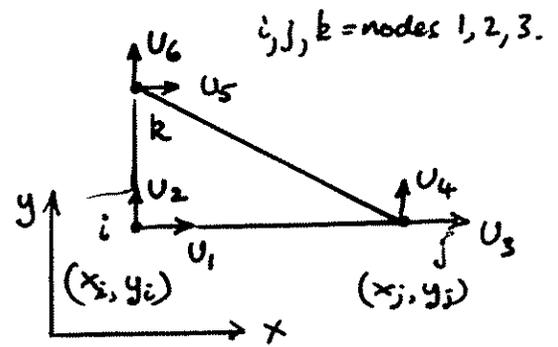
$$U_x = a + bx + cy \qquad \doteq \text{equation of a plane} \qquad (1)$$
$$U_y = d + ex + fy \qquad\qquad\qquad\qquad (2)$$

Strain components

$$\varepsilon_{xx} = \partial U_x / \partial x = b \qquad (3)$$
$$\varepsilon_{yy} = \partial U_y / \partial y = f \qquad (4)$$
$$\gamma_{xy} = \partial U_x / \partial y + \partial U_y / \partial x = c + e \qquad (5)$$

$\left.\begin{array}{l}\end{array}\right\}$ hence strain constant.

Require to obtain $(\underline{\varepsilon} = \underline{a}\,\underline{\upsilon})$ — The '$\underline{a}$' matrix.

Considering only displacements in the U_x direction first, we know that displacements $U_1, U_3,$ and U_5 occur at their respective coordinates (x, y). Substitution of coordinate values into equation (1) gives

$$U_1 = a + bx_i + c y_i \qquad (6)$$
$$U_3 = a + bx_j + c y_j \qquad (7)$$
$$U_5 = a + bx_k + c y_k \qquad (8)$$

or

$$\begin{Bmatrix} U_1 \\ U_3 \\ U_5 \end{Bmatrix} = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{bmatrix} \begin{Bmatrix} a \\ b \\ c \end{Bmatrix} \qquad (9)$$

This may be inverted to obtain (by any convenient method).

$$\begin{Bmatrix} a \\ b \\ c \end{Bmatrix} = \frac{1}{2\Delta} \begin{bmatrix} (x_j y_k - x_k y_j) & (x_k y_i - x_i y_k) & (x_i y_j - x_j y_i) \\ (y_j - y_k) & (y_k - y_i) & (y_i - y_j) \\ (x_k - x_j) & (x_i - x_k) & (x_j - x_i) \end{bmatrix} \begin{Bmatrix} U_1 \\ U_3 \\ U_5 \end{Bmatrix} \qquad (10)$$

Δ = the area of the triangular element.

Note permutation of subscripts — in rows $ijkijk \dots$

Shorthand notation:

$$\begin{Bmatrix} a \\ b \\ c \end{Bmatrix} = \frac{1}{2\Delta} \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \begin{Bmatrix} U_1 \\ U_3 \\ U_5 \end{Bmatrix} \qquad (11)$$

Resubstituting equation (11) into (1) gives $\qquad U_x = a + bx + cy$

$$U_x = \frac{1}{2\Delta} \Big[\overbrace{(\beta_{11} U_1 + \beta_{12} U_3 + \beta_{13} U_5)}^{a} + \overbrace{(\beta_{21} U_1 + \beta_{22} U_3 + \beta_{23} U_5)}^{b} x + \overbrace{(\beta_{31} U_1 + \beta_{32} U_3 + \beta_{33} U_5)}^{c} y \Big]$$
$$(12)$$

Gives U_x at any point in element $\underline{U}_x(x,y)$ defined by:

- nodal displacements, U_1, U_3, U_5
- geometry of element.

Referring to equation (3) $\qquad \varepsilon_{xx} = \partial U_x / \partial x = b \qquad (3)$

from equation (11)
$$\boxed{b = \varepsilon_{xx} = \frac{1}{2\Delta} (\beta_{21} U_1 + \beta_{22} U_3 + \beta_{23} U_5)} \qquad (13)$$

A similar procedure may be completed for displacements only in the y direction, U_y. First substitute the nodal coords. into equation (2)

$$\begin{Bmatrix} U_2 \\ U_4 \\ U_6 \end{Bmatrix} = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{bmatrix} \begin{Bmatrix} d \\ e \\ f \end{Bmatrix} \qquad (14)$$

Inverting (14) yields the same coefficient matrix of equation (10). Thus referring to the shorthand adopted in equation (11).

$$
\begin{Bmatrix} d \\ e \\ f \end{Bmatrix} = \frac{1}{2\Delta} \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \begin{Bmatrix} U_2 \\ U_4 \\ U_6 \end{Bmatrix}
\tag{15}
$$

Substituting into eqn (4)

$$
\varepsilon_{yy} = \frac{1}{2\Delta} \left(\beta_{31} U_2 + \beta_{32} U_4 + \beta_{33} U_6 \right)
\tag{16}
$$

and eqn (5)

$$
\gamma_{xy} = \frac{1}{2\Delta} \left(\beta_{31} U_1 + \beta_{32} U_3 + \beta_{33} U_5 + \beta_{21} U_2 + \beta_{22} U_4 + \beta_{23} U_6 \right)
\tag{17}
$$

Writing ($\underline{\varepsilon} = \underline{a}\,\underline{U}$) in matrix form

$$
\begin{Bmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \gamma_{xy} \end{Bmatrix} = \frac{1}{2\Delta} \begin{bmatrix} \beta_{21} & 0 & \beta_{22} & 0 & \beta_{23} & 0 \\ 0 & \beta_{31} & 0 & \beta_{32} & 0 & \beta_{33} \\ \beta_{31} & \beta_{21} & \beta_{32} & \beta_{22} & \beta_{33} & \beta_{23} \end{bmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \\ U_6 \end{Bmatrix}
\tag{18}
$$

And the $\underline{D}'$ matrix ($\underline{\sigma} = \underline{D}\,\underline{\varepsilon}$) for plane strain is;

$$
\begin{Bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{Bmatrix} = \frac{E}{(1+v)(1-2v)} \begin{bmatrix} (1-v) & v & 0 \\ v & (1-v) & 0 \\ 0 & 0 & \frac{1}{2}(1-2v) \end{bmatrix} \begin{Bmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \gamma_{xy} \end{Bmatrix}
\tag{19}
$$

Recap:

- Require to obtain stiffness matrix $\underline{K} = \int_V \underline{a}^T \underline{D}\, \underline{a}\, dV$ $\qquad \underline{f} = \underline{K}\underline{U}$
- $\underline{D}$ is known since $f(E, v)$ only
- $\underline{a}$ may be obtained for the element as $f(\text{geometry})$ only.

II.5 Inversion (Adjoint Matrix)

It can be shown that

$$a \, (\text{adj } a) = |a| I \tag{II.11}$$

where $|a|$ is the determinant of the matrix a and adj a, called the *adjoint matrix*, is the transpose of the matrix of cofactors of the determinant. Comparing (II.10) and (II.11) we see that

$$a^{-1} = \frac{\text{adj } a}{|a|} \tag{II.12}$$

from which it is clear that the inverse does not exist when $|a|$ is zero, in which case a is said to be *singular*.

To illustrate the method we shall determine the inverse of the matrix

$$\mathbf{H} = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_m & y_m \end{bmatrix} \tag{II.13}$$

If we delete the pth row and qth column from the determinant of the matrix we obtain the minor H'_{pq}, e.g. deleting row 3 and column 1 we have

$$H'_{31} = \begin{vmatrix} x_i & y_i \\ x_j & y_j \end{vmatrix} \tag{II.14}$$

The cofactor $\bar{H}_{pq}$ is the product of the minor and $(-1)^{(p+q)}$. When the cofactors are written as a matrix and then transposed we have the adjoint matrix

$$\text{adj } \mathbf{H} = \begin{bmatrix} \begin{vmatrix} x_j & y_j \\ x_m & y_m \end{vmatrix} & -\begin{vmatrix} x_i & y_i \\ x_m & y_m \end{vmatrix} & \begin{vmatrix} x_i & y_i \\ x_j & y_j \end{vmatrix} \\[2mm] -\begin{vmatrix} 1 & y_j \\ 1 & y_m \end{vmatrix} & \begin{vmatrix} 1 & y_i \\ 1 & y_m \end{vmatrix} & -\begin{vmatrix} 1 & y_i \\ 1 & y_j \end{vmatrix} \\[2mm] \begin{vmatrix} 1 & x_j \\ 1 & x_m \end{vmatrix} & -\begin{vmatrix} 1 & x_i \\ 1 & x_m \end{vmatrix} & \begin{vmatrix} 1 & x_i \\ 1 & x_j \end{vmatrix} \end{bmatrix} \tag{II.15}$$

For example H'_{31} of (II.14) is transposed to row 1 column 3. Expanding the determinants we have

$$\text{adj } \mathbf{H} = \begin{bmatrix} (x_j y_m - x_m y_j) & -(x_i y_m - x_m y_i) & (x_i y_j - x_j y_i) \\ -(y_m - y_j) & (y_m - y_i) & -(y_j - y_i) \\ (x_m - x_j) & -(x_m - x_i) & (x_j - x_i) \end{bmatrix} \tag{II.16}$$

The inverse is obtained by dividing adj H by the determinant of H.

[5:2] Solid Mechanics

Constitutive Relations

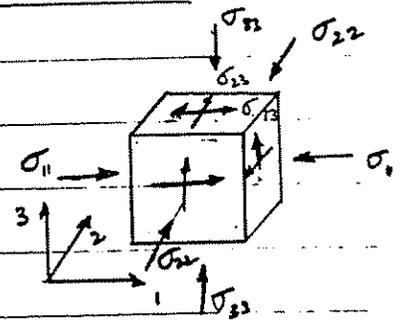STRESS/STRAIN RELATIONSHIPS FOR 3-D ISOTROPIC, LINEAR ELASTICITY

General equations

$$\varepsilon_{11} = \frac{1}{E}\left[\sigma_{11} - \nu(\sigma_{22} + \sigma_{33})\right] \quad (1)$$

$$\varepsilon_{22} = \frac{1}{E}\left[\sigma_{22} - \nu(\sigma_{11} + \sigma_{33})\right] \quad (2)$$

$$\varepsilon_{33} = \frac{1}{E}\left[\sigma_{33} - \nu(\sigma_{11} + \sigma_{22})\right] \quad (3)$$

$$\gamma_{12} = \sigma_{12}/G \quad ; \quad \gamma_{13} = \sigma_{13}/G \quad ; \quad \gamma_{23} = \sigma_{23}/G \quad (4) \qquad G = \frac{E}{2(1+\nu)}$$

σ_{13} ← orthogonal plane to x_3 axis

↳ direction of stress

For 2-D representation, let (1,2) be the plane of interest with the 3 axis perpendicular to this plane, e.g. Tunnel

Plane strain : By definition; • The (1,2) plane is a 'principal' plane on which no shear stresses act

$$\therefore \quad \sigma_{13} = \sigma_{23} = 0$$

$$\longrightarrow \quad \gamma_{13} = \gamma_{23} = 0$$

• No displacement (strain) is allowed perpendicular to the (1,2) plane

$$\therefore \quad \varepsilon_{33} = 0$$

Setting $\varepsilon_{33} = 0$ in equation (3)

$$\sigma_{33} = \nu(\sigma_{11} + \sigma_{22})$$

Substituting into (1) and (2)

$$\varepsilon_{11} = \frac{1}{E}\left[(1-\nu^2)\sigma_{11} - \nu(1+\nu)\sigma_{22}\right]$$

$$\varepsilon_{22} = \frac{1}{E}\left[(1-\nu^2)\sigma_{22} - \nu(1+\nu)\sigma_{11}\right]$$

or $\quad \underline{\varepsilon} = \underline{A}\,\underline{\sigma}$

$$\begin{Bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \gamma_{12} \end{Bmatrix} = \frac{1}{E}\begin{bmatrix} (1-\nu^2) & -\nu(1+\nu) & 0 \\ -\nu(1+\nu) & (1-\nu^2) & 0 \\ 0 & 0 & E/G \end{bmatrix} \begin{Bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{Bmatrix}$$

Since for $\underline{\sigma} = \underline{D}\,\underline{\varepsilon}$, $\underline{D} = \underline{A}^{-1}$

The third equation of the matrix identity is independent of the other 2 therefore $\longrightarrow$ $\sigma_{12} = G\,\gamma_{12}$. The remaining 2×2 matrix may be inverted to give.

$$
\begin{Bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{Bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} (1-\nu) & \nu & 0 \\ \nu & (1-\nu) & 0 \\ 0 & 0 & \frac{1}{2}(1-2\nu) \end{bmatrix} \begin{Bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \gamma_{12} \end{Bmatrix}
$$

N.B. since $\underline{D} = \underline{A}^{-1}$; $\underline{A}^{-1}\underline{D} = \underline{I}$ as a check. ✓

Plane stress

Definition ; • $(1,2)$ plane is principal plane ∴ $\sigma_{13} = \sigma_{23} = 0$

• No stress perpendicular to $(1,2)$ plane $\sigma_{33} = 0$

Substituting $\sigma_{13} = \sigma_{23} = 0$ into eqn(4) and $\sigma_{33} = 0$ into eqns (1, 2, 3) and rearranging terms:

$$
\underline{\varepsilon} = \underline{A}\,\underline{\sigma} \qquad \begin{Bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \gamma_{12} \end{Bmatrix} = \frac{1}{E} \begin{bmatrix} 1 & -\nu & 0 \\ -\nu & 1 & 0 \\ 0 & 0 & 2(1+\nu) \end{bmatrix} \begin{Bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{Bmatrix}
$$

$$
\underline{\sigma} = \underline{D}\,\underline{\varepsilon} \qquad \begin{Bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{Bmatrix} = \frac{E}{(1-\nu^2)} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1}{2}(1-\nu) \end{bmatrix} \begin{Bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \gamma_{12} \end{Bmatrix}
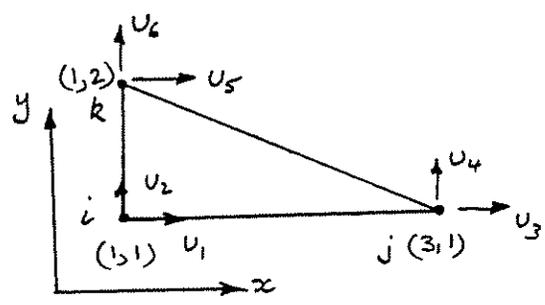$$

Plane strain — most useful in geotechnical, geological situations
eg - slice through a dam, a tunnel — confined problems

Plane stress — many uses in structural mechanics -ie plate bending
fracture mechanics re small specimens.

NUMERICAL EXAMPLE

Assume for simplicity that

$$E = 1 \quad ; \quad \nu = 0.0 \quad (\text{compressible, cork})$$

$$\Delta = 1.0$$

'D' matrix from eqn (19)

$$\underline{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \tag{1}$$

'a' matrix from eqn (18)

Area of triangular elt $\Delta = \cancel{2} \, 1.0$

$$\beta_{21} = y_j - y_k = -1$$
$$\beta_{22} = y_k - y_i = 1$$
$$\beta_{23} = y_i - y_j = 0$$
$$\beta_{31} = x_k - x_j = -2$$
$$\beta_{32} = x_i - x_k = 0$$
$$\beta_{33} = x_j - x_i = 2$$

$$\underline{a} = \frac{1}{2} \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 & 2 \\ -2 & -1 & 0 & 1 & 2 & 0 \end{bmatrix} \tag{2}$$

Note that in this case

$$K = \int_V \underline{a}^T \underline{D} \, \underline{a} \, dv = \underline{a}^T \underline{D} \, \underline{a} \int_V dv \tag{3}$$

$$K = \underline{a}^T \underline{D} \, \underline{a} \, t \, \Delta \tag{4}$$

since $\underline{a}$ and $\underline{D}$ matrices constant where t = element thickness
Δ = element area

$$
\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{Bmatrix}
=
\frac{t\Delta}{4}
\begin{bmatrix}
3 & 1 & -1 & -1 & -2 & 0 \\
  & 4\frac{1}{2} & 0 & -\frac{1}{2} & -1 & -4 \\
  & & 1 & 0 & 0 & 0 \\
  & & & \frac{1}{2} & 1 & 0 \\
  & \text{Symmetric} & & & 2 & 0 \\
  & & & & & 4
\end{bmatrix}
\begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{Bmatrix}
\qquad (5)
$$

Note the characteristic features:

- Positive definite (leading diagonal terms positive and non-zero)
- Symmetric about leading diagonal.

```
      SUBROUTINE ELMT03(D,UL,XL,IX,TL,S,P,NDF,NDM,NST,ISW)
      IMPLICIT REAL*8(A-H,O-Z)
C
C.... THREE NODED PLANE STRAIN ELEMENT FOR SOLID MECHANICS
C
C     USER INFORMATION
C
C     INPUT
C
C       VAR     FORMAT            DESCRIPTION
C       -------------------------------------------------------
C
C       D(1)    F10.0             MODULUS OF DEFORMATION
C       D(2)    F10.0             POISSON RATIO
C
C-------------------------------------------------------------------
C
C     LOCAL NODAL NUMBERING MUST BE COUNTER-CLOCKWISE
C
C-------------------------------------------------------------------
C
C     VARIABLES
C
C     NEL    -     NUMBER OF NODES PER ELEMENT
C     NDF    -     NUMBER OF DEGREES OF FREEDOM PER NODE
C     NST    -     NUMBER OF DEGREES OF FREEDOM PER ELEMENT (NEN*NDF)
C     ISW    -     FUNCTION CALL NO.
C                  1 = READ ELEMENT SPECIFIC INPUT DATA
C                  2 = PERFORM MESH CHECK
C                  3 = FORM ELEMENT STIFFNESS MATRIX       - TANG
C                  4 = EVALUATE ELEMENT STRESSES           - STRE
C                  5 = FORM CONSISTENT/LUMPED MASS MATRIX  - CMAS
C                  6 = FORM LOAD VECTOR                    - FORM
C                        OR EVALUATE NODAL FORCES          - REAC
C
C     ARRAYS - GIVEN
C
C     UL(1,J)    SPECIFIED DISPLACEMENT BOUNDARY CONDITION FOR
C                DEGREE OF FREEDOM J (J=1,6)
C     XL(I,J)    COORDINATE IN THE I DIRECTION AT NODE J
C                EG. XL(1,3) IS X COORDINATE OF NODE K
C
C     ARRAYS - EVALUATED
C
C     A(  )      A MATRIX
C     C(  )      D MATRIX
C     S(I,J)     STIFFNESS MATRIX  S = AT*D*A DV
C                FOR ROW (VERTICAL) I AND COLUMN (HORIZ.) J
C     P(I)       MODIFIED LOAD VECTOR FOR LOCAL DOF I (IGNORE)
C
C-------------------------------------------------------------
      CHARACTER*4 O,HEAD
      COMMON  /CDATA/ O,HEAD(20),NUMNP,NUMEL,NUMMAT,NEN,NEQ,IPR
      COMMON  /ELDATA/ DM,N,MA,MCT,IEL,NEL
      DIMENSION D(2),UL(1,1),XL(NDM,1),IX(1),TL(1),S(NST,1),P(1)
     1      ,A(3,6),C(3,3)
C.... GO TO CORRECT ARRAY PROCESSOR
```

```
       GO TO(1,2,3,4,5,3),ISW
C.... INPUT MATERIAL PROPERTIES
1      READ (5,1000) D(1),D(2)
       WRITE(6,2000) D(1),D(2)
       RETURN
C.... MESH CHECKING FACILITY
2      RETURN
C.... STIFFNESS MATRIX COMPUTATION
C..
3      CONTINUE
C..
C..    THE STIFFNESS MATRIX MUST BE COMPUTED IN THIS PORTION
C..    OF THE ELEMENT SUBROUTINE. BETWEEN STATEMENTS 3 AND 300
C..    THE STIFFNESS MATRIX MUST BE EVALUATED.
C..
C..    USE THE FOLLOWING STEPS TO EVALUATE THE MATRIX S(6,6)
C..
C..    S(6,6) MUST BE EVALUATED BEFORE STATEMENT 300
C..
C      EVALUATE C( ) MATRIX
C..
C.... EVALUATE COEFFICIENTS IN A( ) MATRIX
C..
C.... COMPLETE TRIPLE MATRIX PRODUCT  AT*D*A
C..
C.... PERFORM VOLUME INTEGRATION (*AREA)
C..
C.... MODIFY LOAD VECTOR FOR BOUNDARY CONDITIONS
C..
C..    THIS IS THE END OF YOUR ADDITIONS
C..
300    CONTINUE
       DO 320 I=1,6
       DO 320 J=1,6
320    P(I) = P(I) - S(I,J)*UL(1,J)
       RETURN
C.... EVALUATE ELEMENT STRESSES
4      RETURN
C.... LUMPED MASS COMPUTATION
5      RETURN
C.... FORMATS FOR INPUT AND OUTPUT
1000  FORMAT(2F10.0)
2000  FORMAT(/5X,'THREE NODED CONSTANT STRAIN ELEMENT',//
     1 10X,'DEFORMATION MODULUS       ',6X,E14.7,/
     2 10X,'POISSON RATIO             ',6X,E14.7,/)
       END
```

```
FEAP  1-D LOADING CASE FOR A COLUMN  (File asst2.d)
   11   10    2    1    1    2
COORD
    1    1      0.0
   11    0      2.0

ELEM
    1    1    1    2    1
    5    1    5    6
    6    2    6    7    1
   10    2   10   11

MATE
    1    8           BLOCK 1
      1.0  1000000.
    2    8           BLOCK 2
      1.0  1000000.

BOUN
    1         0
   11         1

FORC
    1        -5.0
   11         0.0

END
MACR
TANG
FORM
SOLV
DISP
REAC
END
STOP
```

```
FEAP   TWIN TRIANGULAR ELEMENTS - SOLID MECHANICS
    4    2    1    2    2    3
COORD
    1              0.0        0.0
    2              1.0        0.0
    3              1.0        1.0
    4              0.0        1.0

ELEM
    1    1    1    2    3
    2    1    1    3    4

MATE
    1    3            MATERIAL 1
       1.0        .25

BOUN
    1         1    1
    2         0    1
    3         0    0
    4         0    0

FORC
    1              0.0        0.0
    2              0.0        0.0
    3              0.0       -1.0
    4              0.0       -1.0

END
MACR
TANG
FORM
SOLV
DISP
REAC
END
STOP
```

6

Linked

Mechanisms

[6:1] Linked Mechanisms

Dual Porosity/Dual Permeability

Concept

Dual permeability

Heuristic derivation

Comsol implementation

DUAL PERMEABILITY MODELS

FRACTURES AND POROUS BLOCKS HAVE DIFFERENT HYDRAULIC PARAMETERS AND THEREFORE RESPONSE TIMES.

- FRACTURES   (high K, low S)
- MATRIX   (low K, high S)

Double diffusion equations:

leakage parameter

specific surface area

$$K_1 \left[\frac{\partial^2 h_1}{\partial x^2} + \frac{\partial^2 h_1}{\partial y^2} \right] = S_1 \frac{\partial h_1}{\partial t} + \pi a^2 (h_1 - h_2) \qquad (1)$$

$$K_2 \left[\frac{\partial^2 h_2}{\partial x^2} + \frac{\partial^2 h_2}{\partial y^2} \right] = S_2 \frac{\partial h_2}{\partial t} - \pi a^2 (h_1 - h_2) \qquad (2)$$

h_1 = porous medium ; h_2 = fracture.

FEM equations:

$$\underline{K}_1 \underline{h}_1^\tau + \underline{S}_1 \underline{\dot{h}}_1^\tau + \underline{B}(\underline{h}_1^\tau - \underline{h}_2^\tau) = \underline{q}_1^\tau$$

$$\underline{K}_2 \underline{h}_2^\tau + \underline{S}_2 \underline{\dot{h}}_2^\tau - \underline{B}(\underline{h}_1^\tau - \underline{h}_2^\tau) = \underline{q}_2^\tau$$

$$\underline{B} = \pi a^2 \int \underline{b}^T \underline{b} \, dv$$

Add time integration; Implicit:

$$\lambda = 1.0$$

$$\tau = t + \Delta t$$

$$\underline{\dot{h}}_1^\tau = \frac{1}{\Delta t} (\underline{h}_1^{t + \Delta t} - \underline{h}_1^t) \quad etc.$$

$$K_1 \underline{h}_1^{t+\Delta t} + \frac{1}{\Delta t} \underline{S}_1 (\underline{h}_1^{t+\Delta t} - \underline{h}_1^t) + \underline{B}(\underline{h}_1^{t+\Delta t} - \underline{h}_2^{t+\Delta t}) = \underline{q}_1^{t+\Delta t}$$

$$K_2 \underline{h}_2^{t+\Delta t} + \frac{1}{\Delta t} \underline{S}_2 (\underline{h}_2^{t+\Delta t} - \underline{h}_2^t) - \underline{B}(\underline{h}_1^{t+\Delta t} - \underline{h}_2^{t+\Delta t}) = \underline{q}_2^{t+\Delta t}$$

__Rearrange in matrix form as__

$$\begin{bmatrix} \underline{A}_1 & -\underline{A}_3 \\ -\underline{A}_3 & \underline{A}_2 \end{bmatrix} \begin{Bmatrix} \underline{h}_1 \\ \underline{h}_2 \end{Bmatrix}^{t+\Delta t} = \begin{Bmatrix} \underline{q}_1^{t+\Delta t} + \frac{1}{\Delta t} \underline{S}_1 \underline{h}_1^t \\ \underline{q}_2^{t+\Delta t} + \frac{1}{\Delta t} \underline{S}_2 \underline{h}_2^t \end{Bmatrix}$$

$$\underline{A}_1 = K_1 + \frac{1}{\Delta t} \underline{S}_1 + \underline{B}$$

$$\underline{A}_2 = K_2 + \frac{1}{\Delta t} \underline{S}_2 + \underline{B}$$

$$\underline{A}_3 = \underline{B}$$

$$K_1 = \int_V \underline{a}^T \underline{D} \underline{a} \, dv$$

$$\underline{S}_1 = S_1 \int_V \underline{b}^T \underline{b} \, dv$$

$$\underline{B} = \pi a^2 \int_V \underline{b}^T \underline{b} \, dv$$

[6:2] Linked Mechanisms

HM – Poromechanics

Effective stresses
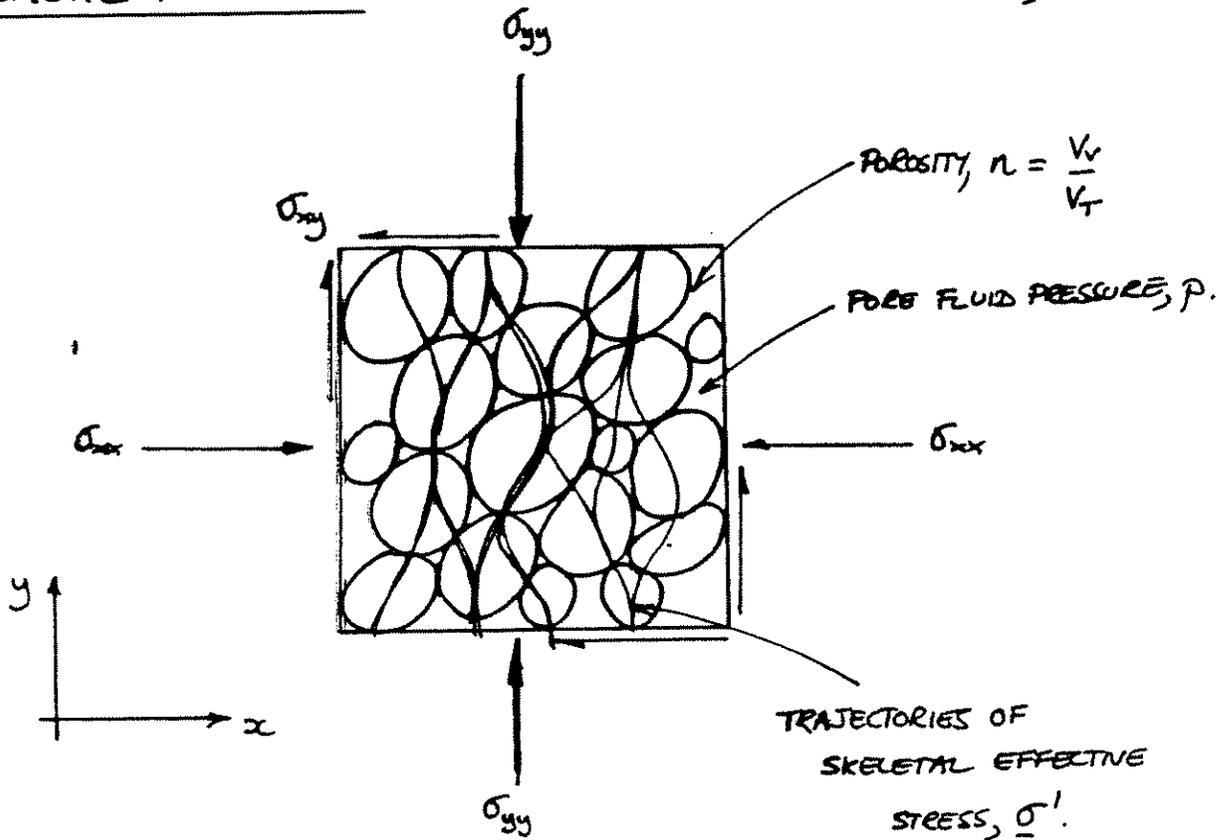
FE equations

Summary equations (Biot, 1941)

EGEEfem

Comsol

PROCESS COUPLINGS [T-<u>H</u>-<u>M</u>-C]

$$
\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & R_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{p} \\ \underline{T} \\ \underline{c} \end{Bmatrix} + \begin{bmatrix} \underline{S}_{11} & \underline{S}_{12} & \cdots & \cdots \\ \underline{S}_{21} & \underline{S}_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{Bmatrix} \underline{\dot{u}} \\ \underline{\dot{p}} \\ \underline{\dot{T}} \\ \underline{\dot{c}} \end{Bmatrix} = \begin{Bmatrix} \underline{\dot{f}} + \cdots \\ \underline{q}_F + \cdots \\ \underline{q}_T + \cdots \\ \underline{q}_M + \cdots \end{Bmatrix}
$$

POROELASTIC RESPONSE   M.A. BIOT (1941)

σ_{yy}

POROSITY, $n = \dfrac{V_v}{V_T}$

σ_{xy}

PORE FLUID PRESSURE, P.

σ_{xx} σ_{xx}

y

x

TRAJECTORIES OF
SKELETAL EFFECTIVE
STRESS, $\underline{\sigma}'$.

σ_{yy}

TOTAL STRESS = EFFECTIVE STRESS + PORE PRESSURE

Figure 4.1.1 Saturated porous medium

PORO ELASTIC EQUATIONS

EQUILIBRIUM (Virtual work)

Terzaghi's Law of effective stresses
$$\partial \underline{\sigma} = \partial \underline{\sigma}' + \underline{m} \partial p \qquad (1)$$

$$\partial \begin{Bmatrix} \sigma_x \\ \sigma_y \\ \sigma_{xy} \end{Bmatrix} = \partial \begin{Bmatrix} \sigma_x \\ \sigma_y \\ \sigma_{xy} \end{Bmatrix}' + \begin{Bmatrix} 1 \\ 1 \\ 0 \end{Bmatrix} \partial p \qquad (2)$$
$$\underset{\underline{m}}{}$$

Constitutive Law
$$\partial \underline{\sigma}' = \underline{E} \, \partial \underline{\varepsilon} \qquad (3) \qquad \left(\text{note } \underline{E} \equiv \underline{D} \right)$$

Substitute (3) into (1) gives

$$\partial \underline{\sigma} = E \partial \underline{\varepsilon} + \underline{m} \, \partial p \qquad (4)$$

Recall:
$VW_I = VW_E$
$\int \underline{\varepsilon}^T \underline{\sigma} \, dv = \underline{u}^T \underline{f}$
$\int (\partial \underline{u})^T \underline{\sigma} \, dv = \underline{u}^T \underline{f}$
$\int \underline{u}^T \underline{a}^T \underline{\sigma} \, dv = \underline{u}^T \underline{f}$

From previous, equilibrium may be stated as:

$$\int_A \underline{A}^T \partial \underline{\sigma} \; dx \, dy = \partial \underline{f} \qquad (5)$$
$$\uparrow$$
$$\underline{a}^T$$

Substituting (4) into (5) gives

$$\int_A \underline{A}^T \underline{E} \underline{A} \; dx \, dy \, \partial \underline{u} \quad + \quad \int_A \underline{A}^T \underline{m} \, \partial p \; dx \, dy \; = \; \partial \underline{f} \qquad (6)$$
$$\uparrow$$
$$\partial \underline{\varepsilon} = \underline{A} \, \partial \underline{u}$$

and if $p = \underline{b} \, \underline{p}$ then replacing in (6)

$$\boxed{\int_A \underline{A}^T \underline{E} \underline{A} \; dx \, dy \; \partial \underline{\dot{u}} \quad + \quad \int_A \underline{A}^T \underline{m} \, \underline{b} \; dx \, dy \; \partial \underline{\dot{p}} \; = \; \partial \underline{\dot{f}}} \qquad (7)$$

And divide through by time, Δt.

$$\underline{B}_{11} \, \underline{\dot{u}} \; + \; \underline{B}_{12} \, \underline{\dot{p}} \; = \; \underline{\dot{f}} \qquad (8)$$

CONSERVATION OF MASS

$\uparrow y$
$\llcorner \rightarrow x$
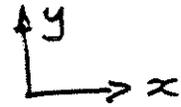
Darcy's Law: $\qquad \begin{Bmatrix} v_x \\ v_y \end{Bmatrix} = -\dfrac{k}{\mu} \begin{Bmatrix} \partial/\partial x \\ \partial/\partial y \end{Bmatrix} \underbrace{(p + \gamma y)}_{h\gamma_w}$ $\qquad$ (9)

$\underline{v} \quad = -\dfrac{k}{\mu} \, \underline{\nabla}(p + \gamma y)$ $\qquad$ (10)

Conservation of mass:

$$\underline{\nabla}^T \underline{v} = \underline{m}^T \underline{\dot{\varepsilon}} - \dfrac{n}{k_f} \dot{p}$$ $\qquad$ (11)

$$\underline{m}^T \underline{\dot{\varepsilon}} = (\dot{\varepsilon}_x + \dot{\varepsilon}_y)$$

Substituting (10) into (11)

$$-\underline{\nabla}^T \dfrac{k}{\mu} \underline{\nabla}(p + \gamma y) = \underline{m}^T \underline{\dot{\varepsilon}} - \dfrac{n}{k_f} \dot{p}$$ $\qquad$ (12)

or, more familiarly

$$-\dfrac{k}{\mu} \underbrace{\left[\dfrac{\partial^2 p}{\partial x^2} + \dfrac{\partial^2 p}{\partial y^2} \right]}_{\text{diffusion}} = \underbrace{\dfrac{\partial}{\partial t}(\dot{\varepsilon}_x + \dot{\varepsilon}_y)}_{\text{volume strain}} - \underbrace{\dfrac{n}{k_f} \dfrac{\partial p}{\partial t}}_{\text{fluid strain}}$$ $\qquad$ (13)

Apply Galerkin weighting:

$$\int_A w \left[-\underline{\nabla}^T \dfrac{k}{\mu} \underline{\nabla} \underbrace{\underline{b}\,\underline{p}}_{p = \underline{b}p} - \underline{m}^T \underbrace{\underline{A}\,\underline{\dot{u}}}_{\dot{\varepsilon} = \underline{A}\dot{u}} + \dfrac{n}{k_f} \underline{b}\,\underline{\dot{p}} \right] dx\,dy = 0$$ $\qquad$ (14)

Integrate in usual manner with, $w = b$, to give

$$\boxed{\begin{array}{l} -\displaystyle\int_A \underline{a}^T \underline{k}\, \underline{a} \; dx\,dy \; \underline{p} \;+\; \int_A \underline{b}^T \underline{m}^T \underline{A} \; dx\,dy \; \underline{\dot{u}} \\[2mm] \qquad -\dfrac{n}{k_f} \displaystyle\int_A \underline{b}^T \underline{b} \; dx\,dy \; \underline{\dot{p}} = \underline{q} \end{array}}$$ $\qquad$ (15)

with
$$\left\{ \begin{array}{c} \partial b/\partial x \\ \partial b/\partial y \end{array} \right\} = \underline{a} \tag{16}$$

and
$$\underline{K} = \frac{k}{\mu} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{17}$$

Symbolically this may be written (15) as

$$\underline{C}_{22} \, \underline{p} - \underline{B}_{22} \, \dot{\underline{p}} + \underline{B}_{21} \, \dot{\underline{u}} = \underline{q} \tag{18}$$

COMBINING MATRIX EQUATIONS — CONSERVATION OF MOMENTUM
— CONSERVATION OF MASS

$$\begin{bmatrix} 0 & 0 \\ 0 & -\underline{C}_{22} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{P} \end{Bmatrix}_\tau + \begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix} \begin{Bmatrix} \underline{\dot{u}} \\ \underline{\dot{P}} \end{Bmatrix}_\tau = \begin{Bmatrix} \underline{\dot{f}} \\ \underline{q} \end{Bmatrix}_\tau \qquad (19)$$

note $\underline{B}_{12} = \underline{B}_{21}^T$ $\therefore$ symmetric.

Defining time derivatives as:
$$\left. \begin{aligned} \underline{\dot{u}}_\tau &= \tfrac{1}{\Delta t}(\underline{u}_{t+\Delta t} - \underline{u}_t) \\ \underline{\dot{P}}_\tau &= \tfrac{1}{\Delta t}(\underline{P}_{t+\Delta t} - \underline{P}_t) \end{aligned} \right\} \quad (20)$$

and assuming, for simplicity, $\ell = 1.0$
$$\therefore \tau = t + \Delta t$$

Then substituting (20) into (19) gives

$$\boxed{\ \tfrac{1}{\Delta t}\begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & (\underline{B}_{22}+\Delta t\,\underline{C}_{22}) \end{bmatrix}\begin{Bmatrix} \underline{u} \\ P \end{Bmatrix}_{t+\Delta t} = \begin{Bmatrix} \underline{\dot{f}} \\ \underline{q} \end{Bmatrix}_{t+\Delta t} + \tfrac{1}{\Delta t}\begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix}\begin{Bmatrix} \underline{u} \\ \underline{P} \end{Bmatrix}_t\ } \quad (21)$$

or $\quad \underline{K}^* \underline{h}_{t+\Delta t} = \underline{q}^*_{t+\Delta t}$

EQUILIBRIUM

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} = b_x - \rho \frac{\partial^2 u_x}{\partial t^2}$$

(1.2) 3 eqns

$$\varepsilon_x = \frac{\partial u_x}{\partial x} \quad ; \quad \gamma_{xy} = \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \quad ; \quad \theta = \Delta V_f / V$$

CONSTITUTIVE

$$\sigma_x = 2G \left(\varepsilon_x + \nu \frac{\varepsilon_v}{1-2\nu} \right) - \alpha p$$

(2.11) (3 eq.)

$$\varepsilon_x = \frac{1}{E}\left[\sigma_x - \nu(\sigma_y + \sigma_z) \right] + \frac{P}{3H}$$

(2.4) (3 eq.)

$$\tau_{xy} = G\gamma_{xy} \qquad (3 eq.)$$

$$\gamma_{xy} = \tau_{xy}/G \qquad (3 eq.)$$

$$p = (\theta - \alpha \varepsilon_v) Q \qquad (2.12)\ (1 eq.)$$

$$\theta = \frac{1}{3H}(\sigma_x + \sigma_y + \sigma_z) + \frac{P}{R} \quad (1 eq.)$$

$$\theta = \alpha \varepsilon_v + \frac{P}{Q} \qquad (2.10)\ 1$$

(2.12)

Evaluate parameters: (E, ν, H, R)

$E \ \& \ \nu$ for (2.4) $\longrightarrow G$

$H \ \& \ R$ from (2.10)

Then $\quad \alpha = \frac{2(1+\nu)}{3(1-2\nu)} \frac{G}{H} \quad ; \quad \frac{1}{Q} = \frac{1}{R} - \frac{\alpha}{H}$

FLOW

$$\theta = \alpha \varepsilon_v + \frac{P}{Q} \qquad\qquad v_x = -k \frac{dp}{dx}$$

$$\frac{\partial \theta}{\partial t} = -\frac{\partial v_x}{\partial x} - \frac{\partial v_y}{\partial y} - \frac{\partial v_z}{\partial z}$$

Substitute (2.11) into (1.2) (3 eqns) $\quad G\nabla^2 u_x + \frac{G}{(1-2\nu)} \frac{\partial \varepsilon_v}{\partial x} - \alpha \frac{\partial p}{\partial x} = 0$

Substitute for flow: $\quad k \nabla^2 p = \frac{1}{Q} \frac{\partial p}{\partial t} + \alpha \frac{\partial \varepsilon_v}{\partial t}$

FINITE ELEMENT METHOD

Biot equations give • Constitutive law and equilibrium equations
 • Flow equation

Deformation

$$G \nabla^2 u_{x_i} + \frac{G}{(1-2v)} \frac{\partial \varepsilon_v}{\partial x} - \alpha \frac{\partial p}{\partial x} = b \qquad \text{3 equns.}$$

Divide through by $\partial/\partial t$

$$G \nabla^2 \frac{\partial u}{\partial t} \nearrow^{\dot{u}} + \frac{G}{(1-2v)} \frac{\partial^2 u}{\partial x^2} \frac{\partial u}{\partial t} \nearrow^{\dot{u}} - \alpha \frac{\partial}{\partial x} \frac{\partial p}{\partial t} \nearrow^{\dot{p}} = b$$
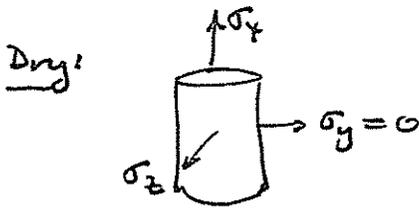
Flow equation

$$K \nabla^2 p = \frac{\partial}{\partial x} \alpha \frac{\partial u}{\partial t} \nearrow^{\dot{u}} + \frac{1}{Q} \frac{\partial p}{\partial t} \nearrow^{\dot{p}}$$

Matrix equations

$$\begin{bmatrix} 0 & 0 \\ 0 & E \end{bmatrix} \left\{ \begin{matrix} u \\ p \end{matrix} \right\}^t + \begin{bmatrix} A & B \\ C & D \end{bmatrix} \left\{ \begin{matrix} \dot{u} \\ \dot{p} \end{matrix} \right\}^t = \left\{ \begin{matrix} b \\ q \end{matrix} \right\}^t$$

PHYSICAL INTERPRETATION OF PARAMETERS $\qquad$ $E, \nu, H, R \longrightarrow G, Q, \alpha$

Dry:

$\sigma_y = 0$

$$E = \frac{\sigma_x}{\varepsilon_x} \quad ; \quad \varepsilon_z = \varepsilon_y = -\nu \varepsilon_x = -\nu \frac{\sigma_x}{E}$$

Fluid filled + jacketed:

$\Delta V_f \, ; \, -p$

— zero volume tube

V

— membrane

No change in total stress
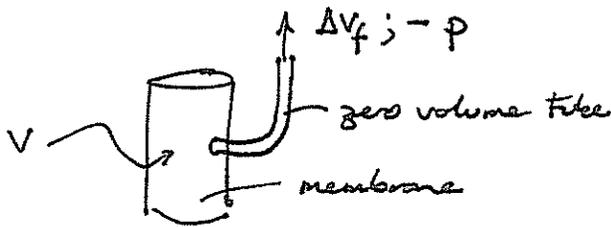
1. Apply $-p$ and remove ΔV_f

$$\theta = \frac{\Delta V_f}{V} \; ; \; \theta = \frac{1}{3H}(\sigma_x + \sigma_y + \sigma_z) + \frac{P}{R}$$

$$\therefore \quad R = P/\theta$$

(fluid strain with change in effective stress)

2. Measure ΔV_s (volume change of soil) for applied $-p$ and ΔV_f

$$\varepsilon_v = \varepsilon_x + \varepsilon_y + \varepsilon_z = \frac{3\sigma_m}{E}(1-2\nu) + \frac{P}{H} \qquad \therefore \quad H = \frac{P}{\varepsilon_v}$$

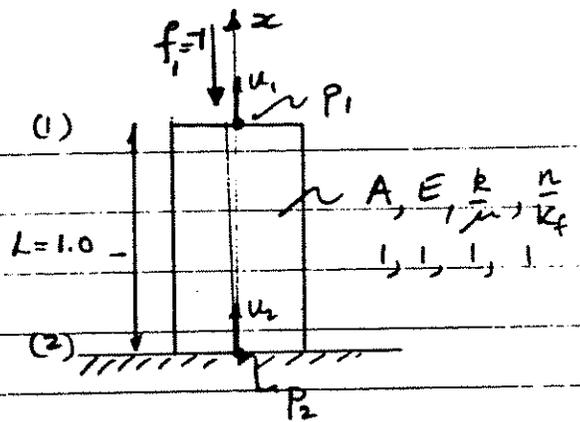(solid strain with effective stress)

From: $\quad E, \nu, H, R$

$$G = \frac{E}{2(1+\nu)}$$

$$\alpha = \frac{2(1+\nu)}{3(1-2\nu)} \frac{G}{H}$$

$$\frac{1}{Q} = \frac{1}{R} - \frac{\alpha}{H}$$

EXAMPLE 4.1

Consider a two noded one-
dimensional element — as shown

Using similar terminology

$$\varepsilon_x = \frac{1}{L} [1 \; ; \; -1] \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} \qquad \underline{\varepsilon} = \underline{A} \cdot u \qquad (1)$$

$$P = [\frac{x}{L} \; ; \; (1-\frac{x}{L})] \begin{Bmatrix} P_1 \\ P_2 \end{Bmatrix} \qquad P = \underline{b} \, \underline{p} \qquad (2)$$

$$\frac{\partial P}{\partial x} = \frac{1}{L} [1 \; ; \; -1] \begin{Bmatrix} P_1 \\ P_2 \end{Bmatrix} \qquad \frac{\partial P}{\partial x} = \underline{a} \, \underline{p} \qquad (3)$$

$$\underline{m} = [1] \qquad ; \qquad \underline{E} = E \qquad (4)$$

Matrices

$$\underline{B}_{11} = \int_A \underline{A}^T \underline{E} \, \underline{A} \, dx \, dy \qquad = \frac{AE}{L} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad (5)$$

$$\underline{B}_{22} = \frac{n}{k_f} \int_A \underline{b}^T \underline{b} \, dx \, dy \qquad = \frac{n}{k_f} \frac{AL}{8} \begin{bmatrix} 8/3 & 4/3 \\ 4/3 & 8/3 \end{bmatrix} \qquad (6)$$

$$\text{or lumped} \qquad = \frac{n}{k_f} \frac{AL}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad (7)$$

$$\underline{C}_{22} = \frac{k}{\mu} \int_A \underline{a}^T \underline{a} \, dx \, dy \qquad = \frac{k}{\mu} \frac{A}{L} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad (8)$$

$$(\quad \underline{B}_{12} = \underline{B}_{21}^T = \int_A \underline{A}^T \underline{m} \, \underline{b} \, dx \, dy$$

$$\underline{B}_{12} = A \int_A \frac{1}{L} \begin{Bmatrix} 1 \\ -1 \end{Bmatrix} [1] \left\{ \frac{x}{L} \; ; \; (1-\frac{x}{L}) \right\} dx$$

$$\int_0^L \frac{x}{L} dx = \left[\frac{x^2}{2L} \right]_0^L = \frac{\cancel{L}}{\cancel{L}} \left(\frac{L}{2} \right)$$

$$\underline{B}_{12} = \frac{A}{L} \int_0^L \begin{bmatrix} \frac{x}{L} & (1-\frac{x}{L}) \\ -\frac{x}{L} & -(1-\frac{x}{L}) \end{bmatrix} dx$$

$$\int_0^L (1-\frac{x}{L}) dx = \left[x - \frac{x^2}{2L} \right]_0^L = \cancel{(1-\frac{1}{2L})}$$

$$\left(L - \frac{L}{2} \right)$$

$L = 1.0$

$$\underline{B}_{12} = \frac{A}{2\cancel{X}} \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \qquad \frac{\cancel{L}}{2} \frac{A}{\cancel{L}} = \frac{A}{2} \qquad \qquad (9)$$

$$\text{and} \quad \underline{B}_{21} = \underline{B}_{12}^T$$

$$\frac{A}{L} \left(\cancel{L} - \frac{L}{2} \right) = A$$

$$A \left(1 - \frac{1}{2} \right) = \left(\frac{1}{2} \right) A$$

Transient Solution

$$(\quad \frac{1}{\Delta t} \begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} + \Delta t \underline{C}_{22} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{p} \end{Bmatrix}_{t+\Delta t} = \begin{Bmatrix} \dot{\underline{f}} \\ \underline{q} \end{Bmatrix} + \frac{1}{\Delta t} \begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{p} \end{Bmatrix}_t \quad (10)$$

Setting $\Delta t = 1.0$, for one element then

$$\begin{bmatrix} 1 & -1 & \frac{1}{2} & \frac{1}{2} \\ -1 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & a_{11} & \frac{k}{\mu} \\ \frac{1}{2} & -\frac{1}{2} & \frac{k}{\mu} & a_{11} \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2=0 \\ p_1 \\ p_2 \end{Bmatrix}_{t+\Delta t} = \begin{bmatrix} 1 & -1 & \frac{1}{2} & \frac{1}{2} \\ -1 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & a_{12} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & a_{12} \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ p_1 \\ p_2 \end{Bmatrix}_t + \begin{Bmatrix} \dot{f}_1 \\ \dot{f}_2 \\ q_1 \\ q_2 \end{Bmatrix}$$

$$a_{11} = -\frac{k}{\mu} \Delta t - \frac{n}{k_f} \frac{AL}{2} \qquad\qquad a_{12} = \frac{-n}{k_f} \frac{AL}{2}$$

$($

The boundary conditions may be prescribed as changes in total stresses by the application of a vertical force, f_1, at time $0 + \frac{1}{2}\Delta t$. This must be distributed as shown right.
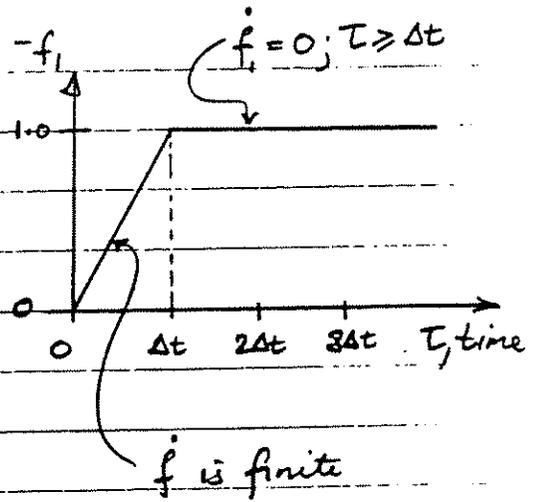
Rate of force application is then

$$\underline{\dot{f}} = \frac{1}{\Delta t}\left[\underline{f}_{t+\Delta t} - \underline{f}_t\right]$$

To solve equation system

1. The magnitudes of $\underline{f}$ are known at all times (prescribed)

2. Prescribe other boundary conditions

$\quad f_1 = 0$ @ $\tau = 0$ $\qquad\qquad f_1 = -1.0$ @ $\tau = \Delta t$

$\quad u_2 = 0 \qquad \tau \geqslant 0$

$\quad p_1 = 0 \qquad \tau \geqslant 0$

Recall basic forms of equations:

$$\frac{\partial^2 H}{\partial x^2} \quad \longrightarrow \quad \int \underline{a}^T \underline{D} \underline{a} \ dV \ \underline{H}$$

$$v \frac{\partial H}{\partial x} \quad \longrightarrow \quad \int \underline{b}^T v \ \underline{a} \ \partial V \ \underline{H}$$

$$\frac{\partial H}{\partial t} \equiv \dot{H} \quad \longrightarrow \quad \int \underline{b}^T \underline{b} \ dV \ \dot{\underline{H}}$$

1-D Biot eqn

$$G \frac{\partial^2 u}{\partial x^2} + \frac{G}{(1-2v)} \frac{\partial \varepsilon}{\partial x} - \alpha \frac{\partial p}{\partial x} = 0$$

$$k \frac{\partial^2 p}{\partial x^2} = \alpha \frac{\partial \varepsilon}{\partial t} \overset{\frac{\partial u}{\partial x}}{\diagup} + \frac{1}{Q} \frac{\partial p}{\partial t}$$

Second equation.

$$k \frac{\partial^2 p}{\partial x^2} = \alpha \frac{\partial}{\partial t} \frac{\partial u}{\partial x} + \frac{1}{Q} \frac{\partial p}{\partial t}$$

$$k \frac{\partial^2 p}{\partial x^2} = \alpha \frac{\partial \dot{u}}{\partial x} + \frac{1}{Q} \dot{p}$$

$$\int \underline{a}^T \underline{D} \underline{a} \, dV \, \underline{p} \ \pm \ \int \underline{b}^T \alpha \, \underline{a} \ dV \ \underline{\dot{u}} \ \pm \ \frac{1}{Q} \int \underline{b}^T \underline{b} \, dV \, \underline{\dot{p}} \ = \ q$$

$$\int \underline{b}^T \underline{m} \, \underline{a} \ dV \ \underline{\dot{u}}$$

GENERAL THEORY OF THREE-DIMENSIONAL CONSOLIDATION

BY

MAURICE A. BIOT

General Theory of Three-Dimensional Consolidation*

Maurice A. Biot
Columbia University, New York, New York

The settlement of soils under load is caused by a phenomenon called consolidation, whose mechanism is known to be in many cases identical with the process of squeezing water out of an elastic porous medium. The mathematical physical consequences of this viewpoint are established in the present paper. The number of physical constants necessary to determine the properties of the soil is derived along with the general equations for the prediction of settlements and stresses in three-dimensional problems. Simple applications are treated as examples. The operational calculus is shown to be a powerful method of solution of consolidation problems.

Introduction

IT is well known to engineering practice that a soil under load does not assume an instantaneous deflection under that load, but settles gradually at a variable rate. Such settlement is very apparent in clays and sands saturated with water. The settlement is caused by a gradual adaptation of the soil to the load variation. This process is known as *soil consolidation*. A simple mechanism to explain this phenomenon was first proposed by K. Terzaghi.[1] He assumes that the grains or particles constituting the soil are more or less bound together by certain molecular forces and constitute a porous material with elastic properties. The voids of the elastic skeleton are filled with water. A good example of such a model is a rubber sponge saturated with water. A load applied to this system will produce a gradual settlement, depending on the rate at which the water is being squeezed out of the voids. Terzaghi applied these concepts to the analysis of the settlement of a column of soil under a constant load and prevented from lateral expansion. The remarkable success of this theory in predicting the settlement for many types of soils has been one of the strongest incentives in the creation of a science of soil mechanics.

Terzaghi's treatment, however, is restricted to the one-dimensional problem of a column under a constant load. From the viewpoint of mathematical physics two generalizations of this are

possible: the extension to the three-dimensional case, and the establishment of equations valid for any arbitrary load variable with time. The theory was first presented by the author in rather abstract form in a previous publication.[2] The present paper gives a more rigorous and complete treatment of the theory which leads to results more general than those obtained in the previous paper.

The following basic properties of the soil are assumed: (1) isotropy of the material, (2) reversibility of stress-strain relations under final equilibrium conditions, (3) linearity of stress-strain relations, (4) small strains, (5) the water contained in the pores is incompressible, (6) the water may contain air bubbles, (7) the water flows through the porous skeleton according to Darcy's law.

Of these basic assumptions (2) and (3) are most subject to criticism. However, we should keep in mind that they also constitute the basis of Terzaghi's theory, which has been found quite satisfactory for the practical requirements of engineering. In fact it can be imagined that the grains composing the soil are held together in a certain pattern by surface tension forces and tend to assume a configuration of minimum potential energy. This would especially be true for the colloidal particles constituting clay. It seems reasonable to assume that for small strains, when the grain pattern is not too much disturbed, the assumption of reversibility will be applicable.

The assumption of isotropy is not essential and

[1] K. Terzaghi, *Erdbaumechanik auf Bodenphysikalischer Grundlage* (Leipzig F. Deuticke, 1925); "Principle of soil mechanics," Eng. News Record (1925), a series of articles.

[2] M. A. Biot, "Le problème de la Consolidation des Matières argileuses sous une charge," Ann. Soc. Sci. Bruxelles **B55**, 110–113 (1935).

anisotropy can easily be introduced as a refinement. Another refinement which might be of practical importance is the influence, upon the stress distribution and the settlement, of the state of initial stress in the soil before application of the load. It was shown by the present author[3] that this influence is greater for materials of low elastic modulus. Both refinements will be left out of the present theory in order to avoid undue heaviness of presentation.

The first and second sections deal mainly with the mathematical formulation of the physical properties of the soil and the number of constants necessary to describe these properties. The number of these constants including Darcy's permeability coefficient is found equal to five in the most general case. Section 3 gives a discussion of the physical interpretation of these various constants. In Sections 4 and 5 are established the fundamental equations for the consolidation and an application is made to the one-dimensional problem corresponding to a standard soil test. Section 6 gives the simplified theory for the case most important in practice of a soil completely saturated with water. The equations for this case coincide with those of the previous publication.[2] In the last section is shown how the mathematical tool known as the *operational calculus* can be applied most conveniently for the calculation of the settlement without having to calculate any stress or water pressure distribution inside the soil. This method of attack constitutes a major simplification and proves to be of high value in the solution of the more complex two- and three-dimensional problems. In the present paper applications are restricted to one-dimensional examples. A series of applications to practical cases of two-dimensional consolidation will be the object of subsequent papers.

1. SOIL STRESSES

Consider a small cubic element of the consolidating soil, its sides being parallel with the coordinate axes. This element is taken to be large enough compared to the size of the pores so that it may be treated as homogeneous, and at the

[3] M. A. Biot, "Nonlinear theory of elasticity and the linearized case for a body under initial stress." .

same time small enough, compared to the scale of the macroscopic phenomena in which we are interested, so that it may be considered as infinitesimal in the mathematical treatment.

The average stress condition in the soil is then represented by forces distributed uniformly on the faces of this cubic element. The corresponding stress components are denoted by

$$\begin{matrix} \sigma_x & \tau_z & \tau_y \\ \tau_z & \sigma_y & \tau_x \\ \tau_y & \tau_x & \sigma_z. \end{matrix} \qquad (1.1)$$

They must satisfy the well-known equilibrium conditions of a stress field.

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_z}{\partial y} + \frac{\partial \tau_y}{\partial z} = 0,$$

$$\frac{\partial \tau_z}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_x}{\partial z} = 0, \qquad (1.2)$$

$$\frac{\partial \tau_y}{\partial x} + \frac{\partial \tau_x}{\partial y} + \frac{\partial \sigma_z}{\partial z} = 0.$$

Physically we may think of these stresses as composed of two parts; one which is caused by the hydrostatic pressure of the water filling the pores, the other caused by the average stress in the skeleton. In this sense the stresses in the soil are said to be carried partly by the water and partly by the solid constituent.

2. STRAIN RELATED TO STRESS AND WATER PRESSURE

We now call our attention to the strain in the soil. Denoting by u, v, w the components of the displacement of the soil and assuming the strain to be small, the values of the strain components are

$$e_x = \frac{\partial u}{\partial x}, \quad \gamma_x = \frac{\partial w}{\partial y} + \frac{\partial v}{\partial z},$$

$$e_y = \frac{\partial v}{\partial y}, \quad \gamma_y = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}, \qquad (2.1)$$

$$e_z = \frac{\partial w}{\partial z}, \quad \gamma_z = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}.$$

In order to describe completely the macroscopic condition of the soil we must consider an addi-

tional variable giving the amount of water in the pores. We therefore denote by θ the increment of water volume per unit volume of soil and call this quantity the *variation in water content*. The *increment of water pressure* will be denoted by σ.

Let us consider a cubic element of soil. The water pressure in the pores may be considered as uniform throughout, provided either the size of the element is small enough or, if this is not the case, provided the changes occur at sufficiently slow rate to render the pressure differences negligible.

It is clear that if we assume the changes in the soil to occur by reversible processes the macroscopic condition of the soil must be a definite function of the stresses and the water pressure i.e., the seven variables

$$e_x \quad e_y \quad e_z \qquad \gamma_x \quad \gamma_y \quad \gamma_z \quad \theta$$

must be definite functions of the variables:

$$\sigma_x \quad \sigma_y \quad \sigma_z \qquad \tau_x \quad \tau_y \quad \tau_z \quad \sigma.$$

Furthermore if we assume the strains and the variations in water content to be small quantities, the relation between these two sets of variables may be taken as linear in first approximation. We first consider these functional relations for the particular case where $\sigma = 0$. The six components of strain are then functions only of the six stress components $\sigma_x \, \sigma_y \, \sigma_z \, \tau_x \, \tau_y \, \tau_z$. Assuming the soil to have isotropic properties these relations must reduce to the well-known expressions of Hooke's law for an isotropic elastic body in the theory of elasticity; we have

$$e_x = \frac{\sigma_x}{E} - \frac{\nu}{E}(\sigma_y + \sigma_z),$$

$$e_y = \frac{\sigma_y}{E} - \frac{\nu}{E}(\sigma_z + \sigma_x),$$

$$e_z = \frac{\sigma_z}{E} - \frac{\nu}{E}(\sigma_x + \sigma_y), \qquad (2.2)$$

$$\gamma_x = \tau_x/G,$$

$$\gamma_y = \tau_y/G,$$

$$\gamma_z = \tau_z/G.$$

In these relations the constants E, G, ν may be interpreted, respectively, as Young's modulus,

the shear modulus and Poisson's ratio for the solid skeleton. There are only two distinct constants because of the relation

$$G = \frac{E}{2(1+\nu)}. \qquad (2.3)$$

Suppose now that the effect of the water pressure σ is introduced. First it cannot produce any shearing strain by reason of the assumed isotropy of the soil; second for the same reason its effect must be the same on all three components of strain $e_x \, e_y \, e_z$. Hence taking into account the influence of σ relations (2.2) become

$$e_x = \frac{\sigma_x}{E} - \frac{\nu}{E}(\sigma_y + \sigma_z) + \frac{\sigma}{3H},$$

$$e_y = \frac{\sigma_y}{E} - \frac{\nu}{E}(\sigma_z + \sigma_x) + \frac{\sigma}{3H}, \qquad (2.4)$$

$$e_z = \frac{\sigma_z}{E} - \frac{\nu}{E}(\sigma_x + \sigma_y) + \frac{\sigma}{3H},$$

$$\gamma_x = \tau_x/G,$$

$$\gamma_y = \tau_y/G,$$

$$\gamma_z = \tau_z/G,$$

where H is an additional physical constant. These relations express the six strain components of the soil as a function of the stresses in the soil and the pressure of the water in the pores. We still have to consider the dependence of the increment of water content θ on these same variables. The most general relation is

$$\theta = a_1\sigma_x + a_2\sigma_y + a_3\sigma_z + a_4\tau_x$$
$$+ a_5\tau_y + a_6\tau_z + a_7\sigma. \qquad (2.5)$$

Now because of the isotropy of the material a change in sign of $\tau_x \, \tau_y \, \tau_z$ cannot affect the water content; therefore $a_4 = a_5 = a_6 = 0$ and the effect of the shear stress components on θ vanishes. Furthermore all three directions x, y, z must have equivalent properties $a_1 = a_2 = a_3$. Therefore relation (2.5) may be written in the form

$$\theta = \frac{1}{3H_1}(\sigma_x + \sigma_y + \sigma_z) + \frac{\sigma}{R}, \qquad (2.6)$$

where H_1 and R are two physical constants.

Relations (2.4) and (2.6) contain five distinct physical constants. We are now going to prove that this number may be reduced to four; in fact that $H = H_1$ if we introduce the assumption of the existence of a potential energy of the soil. This assumption means that if the changes occur at an infinitely slow rate, the work done to bring the soil from the initial condition to its final state of strain and water content, is independent of the way by which the final state is reached and is a definite function of the six strain components and the water content. This assumption follows quite naturally from that of reversibility introduced above, since the absence of a potential energy would then imply that an indefinite amount of energy could be drawn out of the soil by loading and unloading along a closed cycle.

The potential energy of the soil per unit volume is

$$U = \tfrac{1}{2}(\sigma_x e_x + \sigma_y e_y + \sigma_z e_z + \tau_x \gamma_x + \tau_y \gamma_y + \tau_z \gamma_z + \sigma\theta). \quad (2.7)$$

In order to prove that $H = H_1$ let us consider a particular condition of stress such that

$$\sigma_x = \sigma_y = \sigma_z = \sigma_1,$$
$$\tau_x = \tau_y = \tau_z = 0.$$

Then the potential energy becomes

$$U = \tfrac{1}{2}(\sigma_1 \epsilon + \sigma\theta) \quad \text{with} \quad \epsilon = e_x + e_y + e_z$$

and Eqs. (2.4) and (2.6)

$$\epsilon = \frac{3(1-2\nu)}{E}\sigma_1 + \frac{\sigma}{H}, \quad \theta = \sigma_1/H_1 + \sigma/R. \quad (2.8)$$

The quantity ϵ represents the volume increase of the soil per unit initial volume. Solving for σ_1 and σ

$$\sigma_1 = \frac{\epsilon}{R\Delta} - \frac{\theta}{H\Delta},$$

$$\sigma = \frac{-\epsilon}{H_1\Delta} + \frac{3(1-2\nu)\theta}{E\Delta}, \quad (2.9)$$

$$\Delta = \frac{3(1-2\nu)}{ER} - \frac{1}{HH_1}.$$

The potential energy in this case may be con-

sidered as a function of the two variables ϵ, θ. Now we must have

$$\frac{\partial U}{\partial \epsilon} = \sigma_1, \quad \frac{\partial U}{\partial \theta} = \sigma.$$

Hence

$$\frac{\partial \sigma_1}{\partial \theta} = \frac{\partial \sigma}{\partial \epsilon}$$

or

$$\frac{1}{H\Delta} = \frac{1}{H_1\Delta}.$$

We have thus proved that $H = H_1$ and we may write

$$\theta = \frac{1}{3H}(\sigma_x + \sigma_y + \sigma_z) + \frac{\sigma}{R}. \quad (2.10)$$

Relations (2.4) and (2.10) are the fundamental relations describing completely in first approximation the properties of the soil, for strain and water content, under equilibrium conditions. They contain four distinct physical constants G, ν, H and R. For further use it is convenient to express the stresses as functions of the strain and the water pressure σ. Solving Eq. (2.4) with respect to the stresses we find

$$\sigma_x = 2G\left(e_x + \frac{\nu\epsilon}{1-2\nu}\right) - \alpha\sigma,$$

$$\sigma_y = 2G\left(e_y + \frac{\nu\epsilon}{1-2\nu}\right) - \alpha\sigma,$$

$$\sigma_z = 2G\left(e_z + \frac{\nu\epsilon}{1-2\nu}\right) - \alpha\sigma, \quad (2.11)$$

$$\tau_x = G\gamma_x,$$
$$\tau_y = G\gamma_y,$$
$$\tau_z = G\gamma_z$$

with

$$\alpha = \frac{2(1+\nu)}{3(1-2\nu)}\frac{G}{H}.$$

In the same way we may express the variation in water content as

$$\theta = \alpha\epsilon + \sigma/Q, \quad (2.12)$$

where

$$\frac{1}{Q} = \frac{1}{R} - \frac{\alpha}{H}.$$

3. Physical Interpretation of the Soil Constants

The constants E, G and ν have the same meaning as Young's modulus the shear modulus and the Poisson ratio in the theory of elasticity provided time has been allowed for the excess water to squeeze out. These quantities may be considered as the average elastic constants of the solid skeleton. There are only two distinct such constants since they must satisfy relation (2.3). Assume, for example, that a column of soil supports an axial load $p_0 = -\sigma_z$ while allowed to expand freely laterally. If the load has been applied long enough so that a final state of settlement is reached, i.e., all the excess water has been squeezed out and $\sigma = 0$ then the axial strain is, according to (2.4),

$$e_z = -\frac{p_0}{E} \qquad (3.1)$$

and the lateral strain

$$e_x = e_y = \frac{\nu p_0}{E} = -\nu e_z. \qquad (3.2)$$

The coefficient ν measures the ratio of the lateral bulging to the vertical strain under final equilibrium conditions.

To interpret the constants H and R consider a sample of soil enclosed in a thin rubber bag so that the stresses applied to the soil be zero. Let us drain the water from this soil through a thin tube passing through the walls of the bag. If a negative pressure $-\sigma$ is applied to the tube a certain amount of water will be sucked out. This amount is given by (2.10)

$$\theta = -\frac{\sigma}{R}. \qquad (3.3)$$

The corresponding volume change of the soil is given by (2.4)

$$\epsilon = -\frac{\sigma}{H}. \qquad (3.4)$$

The coefficient $1/H$ is a measure of the compressibility of the soil for a change in water pressure, while $1/R$ measures the change in water content for a given change in water pres-

sure. The two elastic constants and the constants H and R are the four distinct constants which under our assumption define completely the physical proportions of an isotropic soil in the equilibrium conditions.

Other constants have been derived from these four. For instance α is a coefficient defined as

$$\alpha = \frac{2(1+\nu)}{3(1-2\nu)} \frac{G}{H}. \qquad (3.5)$$

According to (2.12) it measures the ratio of the water volume squeezed out to the volume change of the soil if the latter is compressed while allowing the water to escape ($\sigma = 0$). The coefficient $1/Q$ defined as

$$\frac{1}{Q} = \frac{1}{R} - \frac{\alpha}{H} \qquad (3.6)$$

is a measure of the amount of water which can be forced into the soil under pressure while the volume of the soil is kept constant. It is quite obvious that the constants α and Q will be of significance for a soil not completely saturated with water and containing air bubbles. In that case the constants α and Q can take values depending on the degree of saturation of the soil.

The standard soil test suggests the derivation of additional constants. A column of soil supports a load $p_0 = -\sigma_z$ and is confined laterally in a rigid sheath so that no lateral expansion can occur. The water is allowed to escape for instance by applying the load through a porous slab. When all the excess water has been squeezed out the axial strain is given by relations (2.11) in which we put $\sigma = 0$. We write

$$e_z = -p_0 a. \qquad (3.7)$$

The coefficient

$$a = \frac{1-2\nu}{2G(1-\nu)} \qquad (3.8)$$

will be called the *final compressibility*.

If we measure the axial strain just after the load has been applied so that the water has not had time to flow out, we must put $\theta = 0$ in relation (2.12). We deduce the value of the water pressure

$$\sigma = -\alpha Q e_z. \qquad (3.9)$$

substituting this value in (2.11) we write

$$e_z = -p_0 a_i. \quad (3.10)$$

The coefficient

$$a_i = \frac{a}{1+\alpha^2 aQ} \quad (3.11)$$

will be called the *instantaneous compressibility*.

The physical constants considered above refer to the properties of the soil for the state of equilibrium when the water pressure is uniform throughout. We shall see hereafter that in order to study the transient state we must add to the four distinct constants above the so-called *coefficient of permeability* of the soil.

4. General Equations Governing Consolidation

We now proceed to establish the differential equations for the transient phenomenon of consolidation, i.e., those equations governing the distribution of stress, water content, and settlement as a function of time in a soil under given loads.

Substituting expression (2.11) for the stresses into the equilibrium conditions (1.2) we find

$$G\nabla^2 u + \frac{G}{1-2\nu}\frac{\partial \epsilon}{\partial x} - \alpha\frac{\partial \sigma}{\partial x} = 0,$$

$$G\nabla^2 v + \frac{G}{1-2\nu}\frac{\partial \epsilon}{\partial y} - \alpha\frac{\partial \sigma}{\partial y} = 0,$$

$$G\nabla^2 w + \frac{G}{1-2\nu}\frac{\partial \epsilon}{\partial z} - \alpha\frac{\partial \sigma}{\partial z} = 0,$$

$$\quad (4.1)$$

$$\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2.$$

There are three equations with four unknowns u, v, w, σ. In order to have a complete system we need one more equation. This is done by introducing Darcy's law governing the flow of water in a porous medium. We consider again an elementary cube of soil and call V_x the volume of water flowing per second and unit area through the face of this cube perpendicular to the x axis. In the same way we define V_y and V_z. According to Darcy's law these three components of the rate of flow are related to the water pressure by the relations

$$V_x = -k\frac{\partial \sigma}{\partial x}, \quad V_y = -k\frac{\partial \sigma}{\partial y}, \quad V_z = -k\frac{\partial \sigma}{\partial z}. \quad (4.2)$$

The physical constant k is called the *coefficient of permeability* of the soil. On the other hand, if we assume the water to be incompressible the rate of water content of an element of soil must be equal to the volume of water entering per second through the surface of the element, hence

$$\frac{\partial \theta}{\partial t} = -\frac{\partial V_x}{\partial x} - \frac{\partial V_y}{\partial y} - \frac{\partial V_z}{\partial z}. \quad (4.3)$$

Combining Eqs. (2.2) (4.2) and (4.3) we obtain

$$k\nabla^2\sigma = \alpha\frac{\partial \epsilon}{\partial t} + \frac{1}{Q}\frac{\partial \sigma}{\partial t}. \quad (4.4)$$

The four differential Eqs. (4.1) and (4.4) are the basic equations satisfied by the four unknowns u, v, w, σ.

5. Application to a Standard Soil Test

Let us examine the particular case of a column of soil supporting a load $p_0 = -\sigma_z$ and confined laterally in a rigid sheath so that no lateral expansion can occur. It is assumed also that no water can escape laterally or through the bottom while it is free to escape at the upper surface by applying the load through a very porous slab.

Take the z axis positive downward; the only component of displacement in this case will be w. Both w and the water pressure σ will depend only on the coordinate z and the time t. The differential Eqs. (4.1) and (4.4) become

$$\frac{1}{a}\frac{\partial^2 w}{\partial z^2} - \alpha\frac{\partial w}{\partial z} = 0, \quad (5.1)$$

$$k\frac{\partial^2 \sigma}{\partial z^2} = \alpha\frac{\partial^2 w}{\partial z\partial t} + \frac{1}{Q}\frac{\partial \sigma}{\partial t}, \quad (5.2)$$

where a is the final compressibility defined by (3.8). The stress σ_z throughout the loaded column is a constant. From (2.11) we have

$$p_0 = -\sigma_z = -\frac{1}{a}\frac{\partial w}{\partial z} + \alpha\sigma \tag{5.3}$$

and from (2.12)

$$\theta = \alpha\frac{\partial w}{\partial z} + \frac{\sigma}{Q}.$$

Note that Eq. (5.3) implies (5.1) and that

$$\frac{1}{a}\frac{\partial^2 w}{\partial z\partial t} = \alpha\frac{\partial\sigma}{\partial t}.$$

This relation carried into (5.2) gives

$$\frac{\partial^2\sigma}{\partial z^2} = \frac{1}{c}\frac{\partial\sigma}{\partial t}, \tag{5.4}$$

with

$$\frac{1}{c} = \alpha^2\frac{a}{k} + \frac{1}{Qk}. \tag{5.5}$$

The constant c is called the *consolidation constant*. Equation (5.4) shows the important result that the water pressure satisfies the well-known equation of heat conduction. This equation along with the boundary and the initial conditions leads to a complete solution of the problem of consolidation.

Taking the height of the soil column to be h and $z=0$ at the top we have the boundary conditions

$$\sigma = 0 \quad \text{for } z=0,$$

$$\frac{\partial\sigma}{\partial z} = 0 \quad \text{for } z=h. \tag{5.6}$$

The first condition expresses that the pressure of the water under the load is zero because the permeability of the slab through which the load is applied is assumed to be large with respect to that of the soil. The second condition expresses that no water escapes through the bottom.

The initial condition is that the change of water content is zero when the load is applied because the water must escape with a finite velocity. Hence from (2.12)

$$\theta = \alpha\frac{\partial w}{\partial z} + \frac{\sigma}{Q} = 0 \quad \text{for } t=0.$$

Carrying this into (5.3) we derive the initial value of the water pressure

$$\sigma = p_0 \Big/ \left(\frac{1}{\alpha a Q} + \alpha\right) \quad \text{for } t=0 \quad \text{or} \quad \sigma = \frac{a-a_i}{\alpha a}p_0, \tag{5.7}$$

where a_i and a are the instantaneous and final compressibility coefficients defined by (3.8) and (3.11).

The solution of the differential equation (5.4) with the boundary conditions (5.6) and the initial condition (5.7) may be written in the form of a series

$$\sigma = \frac{4}{\pi}\frac{a-a_i}{\alpha a}p_0\left\{\exp\left[-\left(\frac{\pi}{2h}\right)^2 ct\right]\sin\frac{\pi z}{2h} + \frac{1}{3}\exp\left[-\left(\frac{3\pi}{2h}\right)^2 ct\right]\sin\frac{3\pi z}{2h} + \cdots\right\}. \tag{5.8}$$

The settlement may be found from relation (5.3). We have

$$\frac{\partial w}{\partial z} = \alpha a\sigma - a p_0. \tag{5.9}$$

The total settlement is

$$w_0 = -\int_0^h \frac{\partial w}{\partial z} dz = -\frac{8}{\pi^2}(a-a_i)hp_0 \sum_0^\infty \frac{1}{(2n+1)^2} \exp\left\{-\left[\frac{(2n+1)\pi}{2h}\right]^2 ct\right\} + ahp_0. \tag{5.10}$$

Immediately after loading ($t=0$), the deflection is

$$w_i = -\frac{8}{\pi^2}(a-a_i)hp_0 \sum_0^\infty \frac{1}{(2n+1)^2} + ahp.$$

Taking into account that

$$\sum_0^\infty \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}, \quad w_i = a_i hp_0, \tag{5.11}$$

which checks with the result (3.10) above. The final deflection for $t = \infty$ is

$$w_\infty = ahp_0. \tag{5.12}$$

It is of interest to find a simplified expression for the law of settlement in the period of time immediately after loading. To do this we first eliminate the initial deflection w_i by considering

$$w_s = w_0 - w_i = \frac{8}{\pi^2}(a-a_i)hp_0 \sum_0^\infty \frac{1}{(2n+1)^2}\left\{1 - \exp\left[-\left(\frac{(2n+1)\pi}{2h}\right)^2 ct\right]\right\}. \tag{5.13}$$

This expresses that part of the deflection which is caused by consolidation. We then consider the rate of settlement.

$$\frac{dw_s}{dt} = \frac{2c(a-a_i)}{h}p_0 \sum_0^\infty \exp\left\{-\left[\frac{(2n+1)\pi}{2h}\right]^2 ct\right\}. \tag{5.14}$$

For $t=0$ this series does not converge; which means that at the first instant of loading the rate of settlement is infinite. Hence the curve representing the settlement w_s as a function of time starts with a vertical slope and tends asymptotically toward the value $(a-a_i)hp_0$ as shown in Fig. 1 (curve 1). It is obvious that during the initial period of settlement the height h of the column cannot have any influence on the phenomenon because the water pressure at the depth $z=h$ has not yet had time to change. Therefore in order to find the nature of the settlement curve in the vicinity of $t=0$ it is enough to consider the case where $h = \infty$. In this case we put

$$n/h = \xi, \quad 1/h = \Delta\xi$$

and write (5.14) as

$$\frac{dw_s}{dt} = 2c(a-a_i)p_0 \sum_0^\infty \exp\left[-\pi^2(\xi + \tfrac{1}{2}\Delta\xi)^2 ct\right]\Delta\xi$$

for $h = \infty$. The rate of settlement becomes the integral

$$\frac{dw_s}{dt} = 2c(a-a_i)p_0 \int_0^\infty \exp(-\pi^2\xi^2 ct)d\xi = \frac{c(a-a_i)p_0}{(\pi ct)^{\frac{1}{2}}}. \tag{5.15}$$

The value of the settlement is obtained by integration

$$w_s = \int_0^t \frac{dw_s}{dt} dt = 2(a-a_i)p_0\left(\frac{ct}{\pi}\right)^{\frac{1}{2}}. \tag{5.16}$$

It follows a parabolic curve as a function of time (curve 2 in Fig. 1).

6. SIMPLIFIED THEORY FOR A SATURATED CLAY

For a completely saturated clay the standard test shows that the initial compressibility a_i may be taken equal to zero compared to the final compressibility a, and that the volume change of the soil is equal to the amount of water squeezed out. According to (2.12) and (3.11) this implies

$$Q = \infty, \quad \alpha = 1. \tag{6.1}$$

This reduces the number of physical constants of the soil to the two elastic constants and the permeability. From relations (3.5) and (3.6) we deduce

$$H = R = \frac{2G(1+\nu)}{3(1-2\nu)} \tag{6.2}$$

and from (5.5) the value of the consolidation constant takes the simple form

$$c = k/a. \tag{6.3}$$

Relation (2.12) becomes

$$\theta = \epsilon. \tag{6.4}$$

The general differential equations (4.1) and (4.4) are simplified,

$$G\nabla^2 u + \frac{G}{1-2\nu}\frac{\partial \epsilon}{\partial x} - \frac{\partial \sigma}{\partial x} = 0,$$

$$G\nabla^2 v + \frac{G}{1-2\nu}\frac{\partial \epsilon}{\partial y} - \frac{\partial \sigma}{\partial y} = 0, \tag{6.5}$$

$$G\nabla^2 w + \frac{G}{1-2\nu}\frac{\partial \epsilon}{\partial z} - \frac{\partial \sigma}{\partial z} = 0,$$

$$k\nabla\sigma^2 = \frac{\partial \epsilon}{\partial t}. \tag{6.6}$$

By adding the derivatives with respect to x, y, z of Eqs. (6.5), respectively, we find

$$\nabla\epsilon^2 = a\nabla\sigma^2, \tag{6.7}$$

where a is the final compressibility given by (3.8). From (6.6) and (6.7) we derive

$$\nabla\epsilon^2 = \frac{1}{c}\frac{\partial \epsilon}{\partial t}. \tag{6.8}$$

Hence the volume change of the soil satisfies the equation of heat conduction.

Equations (6.5) and (6.8) are the fundamental equations governing the consolidation of a completely saturated clay. Because of (6.4) the initial condition $\theta = 0$ becomes $\epsilon = 0$, i.e., at the instant of loading no volume change of the soil occurs. This condition introduced in Eq. (6.7) shows that at the instant of loading the water pressure in the pores also satisfies Laplace's equation.

$$\nabla\sigma^2 = 0. \tag{6.9}$$

The settlement for the standard test of a column of clay of height h under the load p_0 is given by (5.13) by putting $a_i = 0$.

$$w_s = \frac{8}{\pi^2}ahp_0 \sum_0^\infty \frac{1}{(2n+1)^2}$$

$$\times \left\{1 - \exp\left[-\left(\frac{(2n+1)\pi}{2h}\right)^2 ct\right]\right\}. \tag{6.10}$$

From (5.16) the settlement for an infinitely high column is

$$w_s = 2ap_0 \left(\frac{ct}{\pi}\right)^{\frac{1}{2}}. \tag{6.11}$$

It is easy to imagine a mechanical model having the properties implied in these equations. Consider a system made of a great number of small rigid particles held together by tiny helical springs. This system will be elastically deformable and will possess average elastic constants. If we fill completely with water the voids between the
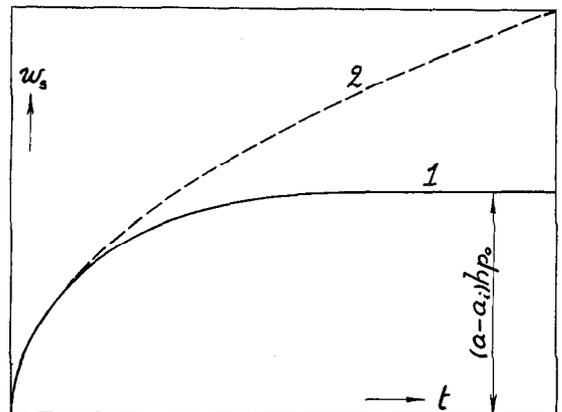

FIG. 1. Settlement caused by consolidation as a function of time. Curve 1 represents the settlement of a column of height h under a load p_0. Curve 2 represents the settlement for an infinitely high column.

particles, we shall have a model of a completely saturated clay.

Obviously such a system is incompressible if no water is allowed to be squeezed out (this corresponds to the condition $Q = \infty$) and the change of volume is equal to the volume of water squeezed out (this corresponds to the condition $\alpha = 1$). If the systems contained air bubbles this would not be the case and we would have to consider the general case where Q is finite and $\alpha \neq 1$.

Whether this model represents schematically the actual constitution of soils is uncertain. It is quite possible, however, that the soil particles are held together by capillary forces which behave in pretty much the same way as the springs of the model.

7. Operational Calculus Applied to Consolidation

The calculation of settlement under a suddenly applied load leads naturally to the application of operational methods, developed by Heaviside for the analysis of transients in electric circuits. As an illustration of the power and simplicity introduced by the operational calculus in the treatment of consolidation problem we shall derive by this procedure the settlement of a completely saturated clay column already calculated in the previous section. In subsequent articles the operational method will be used extensively for the solution of various consolidation problems. We consider the case of a clay column infinitely high and take as before the top to be the origin of the vertical coordinate z. For a completely saturated clay $\alpha = 1$, $Q = \infty$ and with the operational notations, replacing $\partial/\partial t$ by p,

Eqs. (5.1) become

$$\frac{1}{a}\frac{\partial^2 w}{\partial z^2} = \frac{\partial \sigma}{\partial z}, \quad k\frac{\partial^2 \sigma}{\partial z^2} = p\frac{\partial w}{\partial z}. \tag{7.1}$$

A solution of these equations which vanishes at infinity is

$$w = C_1 e^{-z(p/c)^{\frac{1}{2}}},$$
$$\sigma = C_2 - \frac{1}{a}\left(\frac{p}{c}\right)^{\frac{1}{2}} C_1 e^{-z(p/c)^{\frac{1}{2}}}. \tag{7.2}$$

The boundary conditions are for $z = 0$

$$\sigma_z = -1 = \frac{1}{a}\frac{\partial w}{\partial z}, \quad \sigma = 0.$$

Hence

$$C_1 = a\left(\frac{c}{p}\right)^{\frac{1}{2}}, \quad C_2 = 1.$$

The settlement w_s at the top ($z = 0$) caused by the sudden application of a unit load is

$$w_s = a\left(\frac{c}{p}\right)^{\frac{1}{2}} \cdot 1(t).$$

The meaning of this symbolic expression is derived from the operational equation[4]

$$\frac{1}{p^{\frac{1}{2}}}1(t) = 2\left(\frac{t}{\pi}\right)^{\frac{1}{2}}. \tag{7.3}$$

The settlement as a function of time under the load p_0 is therefore

$$w_s = 2ap_0\left(\frac{ct}{\pi}\right)^{\frac{1}{2}}. \tag{7.4}$$

This coincides with the value (6.11) above.

---

[4] V. Bush, *Operational Circuit Analysis* (John Wiley, New York, 1929), p. 192.

[6:3] Linked Mechanisms

HM – Poromechanics

    Comsol-based

        Implementation

        Validation

    EGEEfem-based

HM – Dual Porosity/Permeability Poromechs

THM – Thermomechanics

    Comsol-based

Validation, Verification, & Certification   and related QA/QC
[IEEE Standard Glossary of Software Engineering Terminology]

Verification is "The process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase."

*i.e. Whether the model re*presents the physics and chemistry you have programmed into it.

Validation is "The process of evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements."

i.e. Whether the physics and chemistry represent the real world.

Certification is "A written guarantee that a system or component complies with its specified requirements and is acceptable for operational use."

i.e. What would you bet on it?

SUMMARY

EQUILIBRIUM

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} = b_x - \rho \frac{\partial^2 u_x}{\partial t^2}$$

(1.2)   3 equs

$$\varepsilon_x = \frac{\partial u_x}{\partial x} \quad ; \quad \gamma_{xy} = \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \quad ; \quad \theta = \Delta_f / V$$

CONSTITUTIVE

$$\sigma_x = 2G\left(\varepsilon_x + \frac{\nu \, \varepsilon_v}{1-2\nu}\right) - \alpha p$$

(2.11) (3 eq)

$$\tau_{xy} = G\gamma_{xy} \qquad\qquad (3 eq)$$

$$p = (\theta - \alpha \varepsilon_v)Q \qquad (2.12)\ (1 eq)$$

$$\varepsilon_x = \frac{1}{E}\left[\sigma_x - \nu(\sigma_y + \sigma_z)\right] + \frac{P}{3H}$$

(2.4) (3 eq)

$$\gamma_{xy} = \tau_{xy}/G \qquad\qquad (3 eq)$$

$$\theta = \frac{1}{3H}(\sigma_x + \sigma_y + \sigma_z) + \frac{P}{R} \quad (1 eq)$$

$$\theta = \alpha \varepsilon_v + \frac{P}{Q}$$

(2.10) 1

(2.12)

Evaluate parameters: (E, ν, H, R)

$E \ \& \ \nu$ from (2.4) $\longrightarrow G$

$H \ \& \ R$ from (2.10)

Then $\alpha = \frac{2(1+\nu)}{3(1-2\nu)} \frac{G}{H}$; $\frac{1}{Q} = \frac{1}{R} - \frac{\alpha}{H}$

Flow

$$\theta = \alpha \varepsilon_v + \frac{P}{Q} \qquad\qquad v_x = -k\frac{dp}{dx}$$
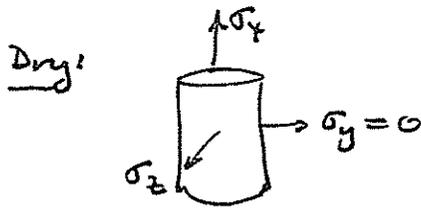
$$\frac{\partial \theta}{\partial t} = -\frac{\partial v_x}{\partial x} - \frac{\partial v_y}{\partial y} - \frac{\partial v_z}{\partial z}$$

Substitute (2.11) into (1.2) (3 equs) $G\nabla^2 u_x + \frac{G}{(1-2\nu)}\frac{\partial \varepsilon_v}{\partial x} - \alpha\frac{\partial P}{\partial x} = 0$

Substitute for flow: $k\nabla^2 P = \frac{1}{Q}\frac{\partial P}{\partial t} + \alpha\frac{\partial \varepsilon_v}{\partial t}$
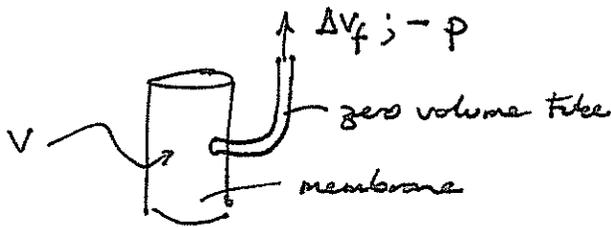
PHYSICAL INTERPRETATION OF PARAMETERS $E, \upsilon, H, R \longrightarrow G, Q, \alpha$

Dry:

$$E = \frac{\sigma_x}{\varepsilon_x} \quad ; \quad \varepsilon_z = \varepsilon_y = -\upsilon \varepsilon_x = -\upsilon \frac{\sigma_x}{E}$$

Fluid filled + jacketed:

No change in total stress

1. Apply $-p$ and remove ΔV_f

$$\theta = \frac{\Delta V_f}{V} \quad ; \quad \theta = \frac{1}{3H}(\sigma_x + \sigma_y + \sigma_z)^{\nearrow 0} + \frac{P}{R}$$

$$\therefore \quad R = P/\theta$$

(fluid strain with change in effective stress).

2. Measure ΔV_s (volume change of soil) for applied $-p$ and ΔV_f

$$\varepsilon_V = \varepsilon_x + \varepsilon_y + \varepsilon_z = \frac{3\sigma_m^{\nearrow 0}}{E}(1-2\upsilon) + \frac{P}{H} \qquad \therefore \quad H = \frac{P}{\varepsilon_V}$$

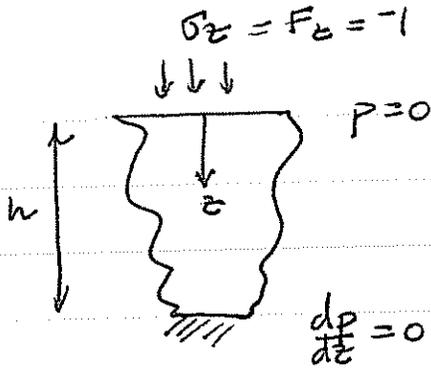(solid strain with effective stress)

From : E, υ, H, R

$$G = \frac{E}{2(1+\upsilon)}$$

$$\alpha = \frac{2(1+\upsilon)}{3(1-2\upsilon)} \frac{G}{H}$$

$$\frac{1}{Q} = \frac{1}{R} - \frac{\alpha}{H}$$

$$\sigma_z = F_z = -1$$

$$\downarrow\downarrow\downarrow$$

$$P = 0$$

$$h$$

$$\frac{dp}{dz} = 0$$

$$\left.\begin{array}{l} E = 1 \ (Pa) \\ \nu = 0 \end{array}\right\} \ G = 1/2$$

$$D = k/\mu = 10^{-3} \ m^2/Pa \cdot s$$

$$\alpha = 1$$

$$1/Q = 10^{-5} \ Pa^{-1}$$

Initial Pressure, p_i:

$$\Theta = \alpha \, \varepsilon_v + \frac{P}{Q} \qquad \therefore \quad \varepsilon_v = \frac{\partial v_z}{\partial z} = -\alpha \frac{1}{\alpha} \frac{P}{Q}$$

$$\sigma_z = 2G\left(\varepsilon_z + \frac{\nu}{(1-2\nu)}\left(\cancel{\varepsilon_x} + \cancel{\varepsilon_y} + \varepsilon_z\right)\right) - \alpha P$$

$$\sigma_t = \frac{2G\,(1-\nu)}{(1-2\nu)}\,\varepsilon_z - \alpha P$$

$$-\frac{\sigma_z}{P} = \frac{2G\,(1-\nu)}{(1-2\nu)}\frac{1}{\alpha Q} + \alpha \longrightarrow 0 + 1$$

$$\therefore \quad P = -\sigma_z$$

Time to 50% consolidation, t_D^{50}:

$$\frac{k}{\mu}\nabla^2 P = \frac{1}{Q}\frac{\partial P}{\partial t} + \alpha \frac{\partial \varepsilon_v}{\partial t}$$

For $\frac{\partial \sigma}{\partial t} \equiv 0$:

$$\varepsilon_v = \frac{\dot{P}}{3H} \quad \text{and} \quad H = \frac{2(1+\nu)}{3(1-2\nu)}\frac{G}{\alpha}$$

$$\frac{k}{\mu}\left(\frac{1}{Q} + \frac{\alpha^2}{G}\frac{(1-2\nu)}{2(1+\nu)}\right)^{-1}\nabla^2 P = \frac{\partial P}{\partial t} \quad \Rightarrow \quad c\frac{\partial^2 P}{\partial z^2} = \frac{\partial P}{\partial t}$$

$$\underbrace{\phantom{\frac{k}{\mu}\left(\frac{1}{Q} + \frac{\alpha^2}{G}\frac{(1-2\nu)}{2(1+\nu)}\right)^{-1}}}_{c}$$

$$c = \left(10^{-5} + \frac{2}{1}\cdot\frac{1}{2}\right)10^{-3} = \frac{1.0}{0.5}\times 10^{-3}$$

$$125s \ 256s$$

50% complete $t_D^{50} = 0.2$; $\qquad 0.2 = \frac{ct}{h^2} \qquad \therefore \quad \frac{0.2\,h^2}{c} = t = \frac{0.2(0.8)^2}{100.5\times 10^{-3}} =$

[6:4] Linked Mechanisms

HM - EGEEfem implementation

HM – Dual porosity/permeability models

THM – Implicit coupling

Explicitly coupled codes

COMBINING MATRIX EQUATIONS —   CONSERVATION OF MOMENTUM

—   CONSERVATION OF MASS

$$
\begin{bmatrix} 0 & 0 \\ 0 & -\underline{C}_{22} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{P} \end{Bmatrix}_{\tau} + \begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix} \begin{Bmatrix} \underline{\dot{u}} \\ \underline{\dot{P}} \end{Bmatrix}_{\tau} = \begin{Bmatrix} \underline{\dot{f}} \\ \underline{q} \end{Bmatrix}_{\tau} \qquad (19)
$$

note $\underline{B}_{12} = \underline{B}_{21}^{T}$ $\therefore$ symmetric.

Defining time derivatives as:
$$
\begin{aligned}
\underline{\dot{u}}_{\tau} &= \frac{1}{\Delta t}\left(\underline{u}_{t+\Delta t} - \underline{u}_{t}\right) \\
\underline{\dot{P}}_{\tau} &= \frac{1}{\Delta t}\left(\underline{P}_{t+\Delta t} - \underline{P}_{t}\right)
\end{aligned} \right\} \qquad (20)
$$

and assuming, for simplicity, $\mathcal{L} = 1.0$

$$\therefore \ \tau = t + \Delta t$$

Then substituting (20) into (19) gives

$$
\boxed{\frac{1}{\Delta t}\begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & (\underline{B}_{22}+\Delta t\,\underline{C}_{22}) \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{P} \end{Bmatrix}_{t+\Delta t} = \begin{Bmatrix} \underline{\dot{f}} \\ \underline{q} \end{Bmatrix}_{t+\Delta t} + \frac{1}{\Delta t}\begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix} \begin{Bmatrix} \underline{u} \\ \underline{P} \end{Bmatrix}_{t}} \qquad (21)
$$

or $\underline{K}^{*}\,\underline{h}_{t+\Delta t} = \underline{q}^{*}_{t+\Delta t}$

EQUILIBRIUM EQUATION

<u>Terzaghi effective stress</u>

$$\partial\underline{\sigma}_1 = \partial\underline{\sigma}_1' + \underline{m}\,\partial p_1$$
$$\partial\underline{\sigma}_2 = \partial\underline{\sigma}_2' + \underline{m}\,\partial p_2 \tag{1}$$

<u>Equilibrium (local)</u> $\partial\underline{\sigma}_1 = \partial\underline{\sigma}_2 = \partial\underline{\sigma}$ (2)

<u>Constitutive</u> $\partial\underline{\sigma}_1' = \underline{D}_1\,\partial\underline{\varepsilon}_1$ (3)
$$\partial\underline{\sigma}_2' = \underline{D}_2\,\partial\underline{\varepsilon}_2$$

<u>Inverse constitutive</u> $\partial\underline{\varepsilon}_1 = \underline{C}_1\,\partial\underline{\sigma}_1'$ $\underline{C}_1 = \underline{D}_1^{-1}$ (4)
$$\partial\underline{\varepsilon}_2 = \underline{C}_2\,\partial\underline{\sigma}_2' \qquad \underline{C}_2 = \underline{D}_2^{-1}$$

<u>Total strain</u> $\partial\underline{\varepsilon} = \partial\underline{\varepsilon}_1 + \partial\underline{\varepsilon}_2$ (5)

Providing reference length includes planes (1) and (2)

<u>Substituting (1) into (4) and then into (5)</u>

$$\partial\underline{\varepsilon} = (\underline{C}_1 + \underline{C}_2)\,\partial\underline{\sigma} - \underline{C}_1\,\underline{m}\,\partial p_1 - \underline{C}_2\,\underline{m}\,\partial p_2$$

$$\partial\underline{\sigma} = \underline{D}_{12}\left(\partial\underline{\varepsilon} + \underline{C}_1\,\underline{m}\,\partial p_1 + \underline{C}_2\,\underline{m}\,\partial p_2\right) \tag{6}$$

$$\text{where } \underline{D}_{12} = (\underline{C}_1 + \underline{C}_2)^{-1}$$

<u>Global Equilibrium (FE)</u> $\int \underline{B}^T\,\partial\underline{\sigma}\,dV - \partial\underline{f} = 0$ (7)

Where $\partial\underline{\varepsilon} = \underline{B}\,\partial\underline{u}$ (8)

$$\partial p_n = \underline{N}\,\partial p_n$$
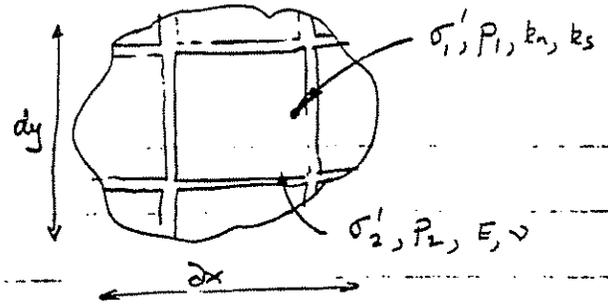
<u>Substituting (6) and (8) into (7) and divide by Δt</u>

$$\int \underline{B}^T\underline{D}_{12}\,\underline{B}\,dV\,\dot{\underline{u}} + \int \underline{B}^T\underline{D}_{12}\,\underline{C}_1\,\underline{m}\,\underline{N}\,dV\,\dot{p}_1$$
$$+ \int \underline{B}^T\underline{D}_{12}\,\underline{C}_2\,\underline{m}\,\underline{N}\,dV\,\dot{p}_2 = \dot{\underline{f}} \tag{9}$$

CONSERVATION OF MASS

<u>Phase 1 - Constitutive</u> $\underline{v} = -\dfrac{k_1}{\mu}\,\underline{\nabla}(p_1 + \gamma z)$ (10)

<u>Continuity</u> $\underline{\nabla}^T\underline{v} = \underline{m}^T\dot{\underline{\varepsilon}} - \dfrac{n_1}{K_f}\dot{p}_1 + q_{12} -$

DUAL POROSITY POROELASTICITY

σ_1', P_1, k_n, k_s
σ_2', P_2, E, ν
dy dx

<u>Integrate using Green's Identity.</u>

$$\frac{1}{\mu}\int \underline{A}^T\underline{k}_1\,\underline{A}\,dV\,\dot{p}_1 + \int \underline{N}^T\underline{m}^T\underline{C}_1\,\underline{D}_{12}\,\underline{B}\,dV$$

$$+ \frac{n_1}{K_f}\int \underline{N}^T\underline{N}\,dV\,\dot{p}_1 - \int \underline{N}^T\underline{N}\,dV\,q_{12} = \frac{\gamma}{\mu}\int \underline{A}^T\underline{k}_1\,\underline{A}\,d \tag{1}$$

<u>Similarly for phase 2.</u>

$$\frac{1}{\mu}\int \underline{A}^T\underline{k}_2\,\underline{A}\,dV\,\dot{p}_2 + \int \underline{N}^T\underline{m}^T\underline{C}_2\,\underline{D}_{12}\,\underline{B}\,dV\,\dot{u}$$

$$+ \frac{n_2}{K_f}\int \underline{N}^T\underline{N}\,dV\,\dot{p}_2 - \int \underline{N}^T\underline{N}\,dV\,q_{12} = \frac{\gamma}{\mu}\int \underline{A}^T\underline{k}_2\,\underline{A}\,dV \tag{1}$$

MATRIX EQUATIONS (COMBINED)

$$\begin{Bmatrix} 0 \\ \underline{K}_1 p_1 \\ \underline{K}_2 p_2 \end{Bmatrix}_{l+1} + \begin{bmatrix} F & G_1 & G_2 \\ E_1 & S_1 & 0 \\ E_2 & 0 & S_2 \end{bmatrix} \begin{Bmatrix} \dot{u} \\ \dot{p}_1 \\ \dot{p}_2 \end{Bmatrix}_{l+1} = \begin{Bmatrix} \dot{f} \\ \underline{H}q_{12} + \underline{K}_1 YZ \\ \underline{H}q_{21} + \underline{K}_2 YZ \end{Bmatrix}_{l+1}$$

Fully implicit $\dot{u}^{l+1} = \dfrac{1}{\Delta t}(u^{l+1} - u^l)$
etc. for $\dot{p}_1, \dot{p}_2$

gives $\dfrac{1}{\Delta t}[\quad]\{\ \}^{l+1} = \dfrac{1}{\Delta t}[\quad]\{\ \}^l + \{\ \}$
$\underbrace{\qquad}_{\text{solve}}$

COMMENTS.

1. Form of H must be defined (Dual porosity)

2. Displacements are lumped $\underline{u}$ not u_1, u_2

Thermo – Mechanical – Hydraulic System:

Mechanical:

$$G \nabla^2 u_x + \frac{G}{(1-2v)} \frac{\partial \varepsilon_v}{\partial x} - \alpha_b \frac{\partial p}{\partial x} - \alpha_t \frac{\partial (T-T_0)}{\partial x} = \rho \frac{\partial^2 u_y}{\partial t^2}$$

$$f(\alpha, E, v)$$

(3 equations)

Fluid – Flow:

$$\frac{k}{\mu} \nabla^2 p = \frac{1}{Q} \frac{\partial p}{\partial t} + \alpha_b \frac{\partial \varepsilon_v}{\partial t} - \alpha_f \frac{\partial T}{\partial t}$$
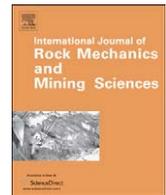
Thermal – Flow:

$$\lambda \nabla^2 T - v \rho c \, \partial T = \rho c \frac{\partial T}{\partial t}$$

Numerical simulation of thermal-hydrologic-mechanical-chemical processes in deformable, fractured porous media

Joshua Taron [a,*], Derek Elsworth [a], Ki-Bok Min [b]

[a] Department of Energy and Mineral Engineering and Center for Geomechanics, Geofluids, and Geohazards (G3), Pennsylvania State University, University Park, Pennsylvania, USA
[b] School of Civil, Environmental, and Mining Engineering, The University of Adelaide, SA, Australia

ABSTRACT

A method is introduced to couple the thermal (T), hydrologic (H), and chemical precipitation/dissolution (C) capabilities of TOUGHREACT with the mechanical (M) framework of FLAC^{3D} to examine THMC processes in deformable, fractured porous media. The combined influence of stress-driven asperity dissolution, thermal-hydro-mechanical asperity compaction/dilation, and mineral precipitation/dissolution alter the permeability of fractures during thermal, hydraulic, and chemical stimulation. Fracture and matrix are mechanically linked through linear, dual-porosity poroelasticity. Stress-dissolution effects are driven by augmented effective stresses incrementally defined at steady state with feedbacks to the transport system as a mass source, and to the mechanical system as an equivalent chemical strain. Porosity, permeability, stiffness, and chemical composition may be spatially heterogeneous and evolve with local temperature, effective stress and chemical potential. Changes in total stress generate undrained fluid pressure increments which are passed from the mechanical analysis to the transport logic with a correction to enforce conservation of fluid mass. Analytical comparisons confirm the ability of the model to represent the rapid, undrained response of the fluid-mechanical system to mechanical loading. We then focus on a full thermal loading/unloading cycle of a constrained fractured mass and follow irreversible alteration in *in-situ* stress and permeability resulting from both mechanical and chemical effects. A subsequent paper [Taron J, Elsworth D. Thermal-hydrologic-mechanical-chemical processes in the evolution of engineered geothermal reservoirs. Int J Rock Mech Min Sci 2009; this issue, doi:10.1016/j.ijrmms.2009.01.007] follows the evolution of mechanical and transport properties in an EGS reservoir, and outlines in greater detail the strength of coupling between THMC mechanisms.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

It is well known that fractured rocks exhibit changes in mechanical compliance and hydraulic conductivity when subjected to thermal, hydraulic, mechanical, and chemical forces. In many engineering applications it is important to be able to predict the direction and magnitude of these changes. However, the interplay between temperature, effective stress, chemical potential, and fracture response is complex: it is not only influenced by anisotropic and spatially varying fracture properties, but also by fracture properties that are dynamic, and evolve with the dynamic nature of the applied forces.

The gaping or sealing of natural fractures has clear implications in reservoirs for the sequestration of CO_2 [2] and radioactive waste repositories [3], where the release of CO_2 or the redistribution of pore fluids around contained radioactive waste is a primary concern. Volcanic environments are also impacted, as in the case of failing volcanic domes [4], where elevated fluid pressures may destabilize an existing volcanic pile. In other cases, such as petroleum or gas reservoirs, hot dry rock [5] or enhanced geothermal systems [6] (HDR/EGS), engineered stimulations may beneficially improve fluid circulation; a topic of significant interest since the majority of worldwide geothermal capacity is contained within low permeability rock masses [6,7].

Despite their importance, the competing influence of processes that degrade fluid conductivity in dominant fractures, such as thin-film pressure solution [8–10] and mineral precipitation, and those that enhance it, such as shear dilation [11,12], mineral dissolution [13–15], and strain energy driven free-face dissolution [9,16] have yet to be addressed at geologic scale. To examine these processes together, a link between chemical and mechanical behavior that maintains dependence on thermal and hydrologic changes is required, i.e., THMC coupling. And while several THM [3,17–20] and THC [14] coupling methodologies have been suggested, to the authors' knowledge no single numerical simulator has been introduced to examine THMC processes in a

* Corresponding author. Tel.: +1 814 863 9733; fax: +1 814 865 3248.
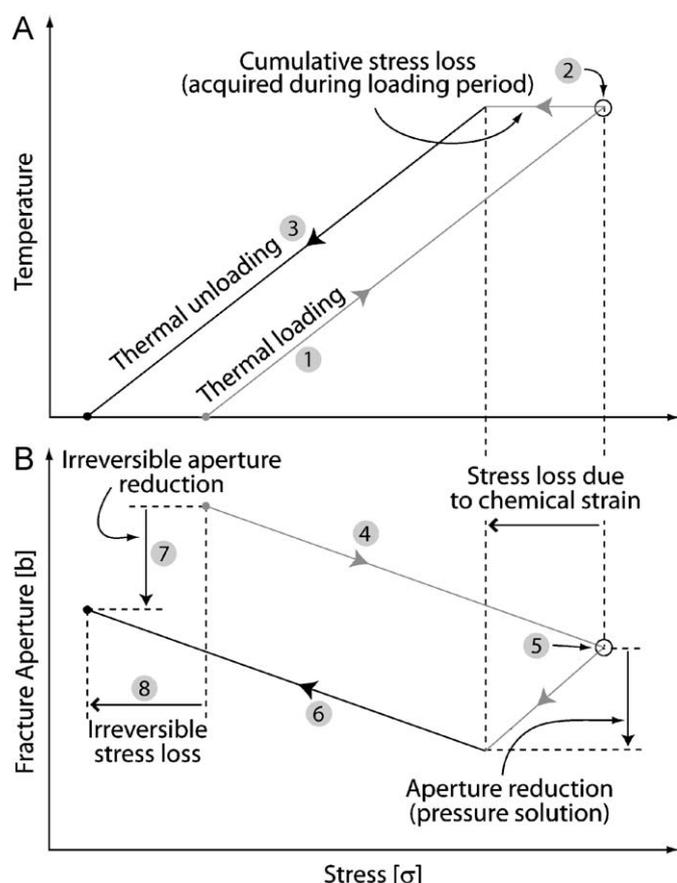 E-mail address: jmt269@psu.edu (J. Taron).

Fig. 1. Conceptual, behavioral trend of thermally loaded and fractured rock: (A) follow light gray line as (1) increasing temperature builds stress (partially reduced by elastic fracture strain). (2) Irreversible fracture strains reduce stress, which, for illustrative purpose, is applied at the end of loading (3). Thermal unloading follows the black line. (B) Follow gray temperature (stress) loading line (4) elastic reduction in fracture aperture (idealized as linear). Loading reaches maximum value (5). Aperture irreversibly closes (chemical strain) and causes corresponding drop in stress. Black (6) unloading line returns the system to its resting state for an (7) irreversible aperture reduction and (8) corresponding irreversible stress loss.

construct that is applicable to the broad variety of above-mentioned engineering applications.

Fig. 1 illustrates the potential error in excluding the chemical–mechanical link from numerical modeling. In the figure, we follow a complete cycle of thermal/stress loading in a chemically active fractured rock. During the loading/unloading cycle, reversible (elastic) and irreversible (chemical–mechanical: pressure solution or other) changes in aperture occur, with the ultimate result that after unloading, once the system has been returned to its initial background state, we see an irreversible aperture reduction, and a corresponding irreversible loss in the state of stress. These two occurrences (7 and 8 in Fig. 1) are the behaviors of primary interest, as they indicate a complete and potentially significant alteration of the resting system that cannot be represented without the inclusion of THMC processes.

2. Model capabilities

We now introduce and implement a method for coupling the multiphase, multi-component, non-isothermal thermodynamics, reactive transport, and chemical precipitation/dissolution capabilities of TOUGHREACT [14] with the mechanical framework of FLAC³D [21] to generate a coupled THMC simulator. This "modular" approach, first proposed by Settari [22] to couple

geomechanics with reservoir flow simulation, has some advantages over the development of a single coupled program. Modular approaches will typically be more rapid and less expensive to develop, although working within the framework of an existing code can sometimes lack the freedom that is inherent in "from scratch" code development. Additionally, as pointed out by Settari and Mourits [23], the modular construction allows for easier implementation of future advances in constitutive relationships or modeling structures (rather than modifying an entire coding structure), and the system can utilize highly sophisticated, rigorously validated existing codes developed at high cost. It can take many years for a new modeling structure to be validated by the research community, but in the case of TOUGHREACT and FLAC³D, each has been extensively scrutinized and each code is "qualified" for regulated programs, such as the US radioactive waste program.

Furthermore, single codes often simplify behavior beyond the principal scope of the analysis. For example, complex geomechanical codes may represent the flow system as only single phase, and complex reactive transport codes often incorporate mechanical response as invariant total stresses. Appropriate coupling enables the important subtleties of geomechanical response to be followed while maintaining complex fluid thermodynamics and reactive processes. Although development time is shortened in this modular approach, execution times are commonly extended, as neither code is optimized for the couplings, and data transfer must occur between the concurrently or sequentially executing codes. As suggested by Settari and Mourits [23] and Minkoff et al. [24], however, this may not always be the case, because in systems where geomechanics may be loosely coupled (not changing at a rapid pace) the geomechanics simulation may not need to be conducted very often, thus improving computational efficiency over fully coupled codes where mechanics are equilibrated at every fluid flow time step.

The coupled analysis that we present incorporates features unique to engineered geosystems, particularly those under elevated temperature and chemical potential, involving the undrained pressure response in a dual-porosity medium and stress-chemistry effects including the role of mechanically mediated chemical dissolution of bridging fracture asperities. FLAC³D is exercised purely in mechanical mode, where undrained fluid pressures may be evaluated (externally) from local total stresses. This undrained methodology allows calculation of the short-term build-up in fluid pressures that result from an instantaneous change in stress, provided we have knowledge of the compressibility of the pore fluids and the solid matrix. In this way, the complex thermodynamics of phase equilibria of multi-phase water mixtures, and even multi-component mixtures (such as CO_2 and water), can be tracked in the pre-existing framework of TOUGHREACT. As TOUGHREACT has no use for compressibility, however, it is necessary to code this capability into the program or, as we have done, to insert a thermodynamic calculation into the external linking module (discussed later). For water mixtures, we utilize the 1997 International Association for the Properties of Water and Steam (IAPWS) steam table equations [25]. For CO_2 mixtures, an appropriate equation of state would be required, and we have not yet added this capability. If a system is unsaturated (such as in HDR/EGS), fluid compressibility is very large, and the undrained poroelastic equations approach their drained counterparts. Therefore, while our construct is tailored to saturated systems, drained systems are automatically accommodated.

FLAC³D is applied independent of time to accommodate the incremental equilibration of stresses for various mechanical constitutive relationships. TOUGHREACT performs time-dependent transport calculations, tracking thermodynamic relationships for temperature, phase equilibria, and pore pressure dissipation

together with aqueous chemical equilibrium and kinetic precipitation/dissolution in a dual-porosity medium. Under large thermal stresses, shear failure may be expected, and FLAC^{3D} is capable of handling this with the constitutive theories of Mohr–Coulomb or Hoek–Brown. Plastic flow is also possible, although this would require consideration of permeability changes that occur during fracture shear and also fracture compression. This complexity is not addressed here, and will be the topic of a future paper.

3. Simulation logic

Simulation is executed within FLAC^{3D}'s FISH programming language [21], where external operations by TOUGHREACT and the linking module are controlled. TOUGHREACT, an integral finite difference code [26], calculates all properties at the central coordinate of element volumes. In contrast, the first order finite difference program FLAC^{3D}, with explicit temporal derivatives and a mixed discretization method that overlays constant strain-rate tetrahedral elements with the final zone elements (adding greater freedom in methods of plastic flow), utilizes properties of state (p,T) at corner nodes and mechanical variables (σ,u) at central coordinates. Correspondingly, state properties from central TOUGHREACT nodes are interpolated to connecting corner nodes of FLAC^{3D}. Stress (not displacement) outputs from FLAC^{3D} are used as the independent variable in constitutive relationships. The parsing of stresses to TOUGHREACT is direct, as they are calculated centrally within the node-centered blocks of FLAC^{3D} (in spatial agreement with TOUGHREACT).

In its current construction, the codes iterate upon the same numerical grid. This structure, however, is not required. As pointed out by Minkoff et al. [24], un-matched meshes are one benefit to a modular code. For example [24], it may be desirable to conduct flow simulations upon a reservoir area impacted by fluid injection and withdrawal only, while the mechanical grid may include the reservoir area in addition to all overburden up to the ground surface. Neither must the overlapping simulation areas utilize identical grid spacing, such that it may be desirable to refine the fluid flow mesh to capture some complex physics in a specific area, without adapting the mechanical mesh to agree. It is only required that interpolation of data accommodate the differences in mesh extent and geometry.

Sequential execution of the two programs is linked by a separate code capable of parsing data outputs from each primary simulator as input to the companion. This separate code is referred to as the "interpolation module". The module is a Fortran 90 executable, and maintains access to data outputs from TOUGHREACT and FLAC^{3D}. In addition to data interpolation, this module executes constitutive relationships including permeability evolution, dual-porosity poroelastic response to stress, and thermodynamically controlled fluid compressibility.

All transient calculations take place within TOUGHREACT, and it is here that the time step is controlled for conditions of fluid velocity, grid size, and reaction rates. Additionally, there is a secondary (explicit) time step that controls how often stress is corrected to changes in fluid pressure (for what length of time TOUGHREACT conducts a flow simulation before allowing stress equilibration in FLAC^{3D}). This frequency is controlled in the interpolation module. If the magnitude of stress change in the system over one time step is beneath a pre-determined tolerance, the frequency is decreased (if stress is not changing, mechanical re-equilibration is unnecessary), and vice versa for an upper tolerance. Coupling is explicit and constitutive calculations are performed once per iteration (assuming constant constitutive values throughout a fluid flow time step), requiring sufficiently
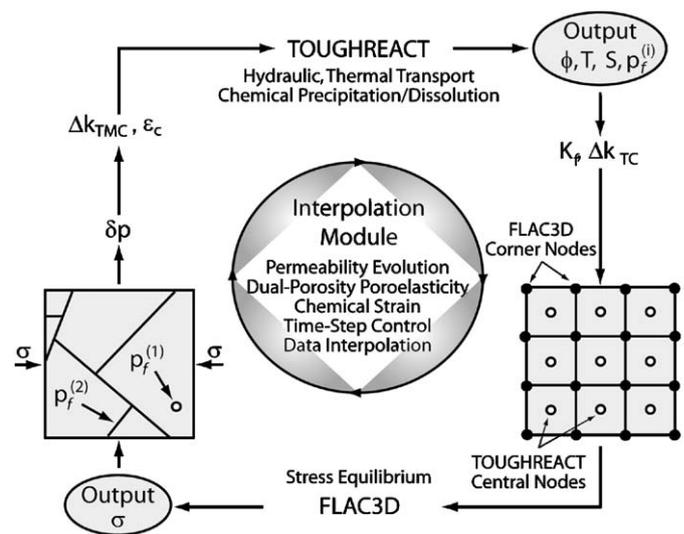

Fig. 2. Coupling relationship between TOUGHREACT, FLAC^{3D}, and the interpolation module.

small time steps relative to the rapidity of change in the system. The validity of utilizing such a methodology is discussed in later sections to provide insight into this explicit time step.

The coupling cycle is shown in Fig. 2, and is comparable to the loose coupling, modular structure of Minkoff et al. [24] and Rutqvist et al. [3]. Simulation begins with equilibration of temperature (T) and pore fluid pressure (p_f) in TOUGHREACT, where porosity (ϕ) changes due to mineral precipitation/dissolution and liquid saturation (S) are also obtained. Constitutive relationships in the interpolation module transform these outputs into fluid bulk modulus (K_f), as obtained from IAPWS steam table equations, and permeability change due to mineral behavior (Δk_{TC}). The TOUGHREACT central node data (p_f,T) are then interpolated to corner node information as input to FLAC^{3D}. After stress equilibration in FLAC3D, the interpolation module uses stress outputs within a dual-porosity framework, consisting of matrix ($p_f^{(1)}$) and fracture ($p_f^{(2)}$) pore fluid pressures, to obtain the pressure response to the new stress field via domain (*matrix, fracture*) and state (p,T) specific Skempton coefficients. Effective stress is then used to obtain the permeability change due to pressure solution type behavior (Δk_{TMC}) while chemical strain (ε_C) is accommodated in the stress field (discussed later). Parameters then re-enter TOUGHREACT for the next time step.

4. Governing equations

The physical system of interest is modeled herein as a multi-continuum, fully or partially saturated fracture/matrix system with direct communication between the domains. Local thermal equilibrium is assumed between the fluid and solid (at a single point in continuum space, the fluid and solid exhibit the same temperature), but not between separate fracture and matrix domains. From this framework, a differential of pressure and temperature may develop between the fracture and matrix, with properties of pressure and temperature dissipation influencing the rapidity of transfer from local changes in the fracture system into the surrounding matrix blocks, and vice versa. As such, the multi-continuum distinction is fully maintained within the numerically represented THC system, while local continuity of stress requires equilibrium of stresses between fracture and matrix, which is then represented within the single continuum framework of FLAC^{3D}. For this transition, physical characteristics

are delegated based upon dual-porosity poroelastic theory [27–31]. The governing balance equations and their constitutive counterparts are discussed below.

4.1. Conservation of momentum—solid

Mechanical equilibrium of the solid phase is governed by the balance of linear momentum,

$$\sigma_{ij,j} + b_i = \rho \dot{v}_i, \tag{1}$$

where b_i are the body forces per unit volume, $\dot{v}_i$ are the material time derivatives of velocities, and $\sigma_{ij,j}$ represents the divergence of the transpose of the Cauchy stress tensor. In an iterative formulation, for static equilibrium of the medium, the momentum balance becomes the common force equilibrium relation

$$\sigma_{ij,j} = -b_i. \tag{2}$$

The resulting unknowns can be related to each other through any of several elastic or plastic constitutive relationships. We begin with the case of an isotropic, elastic solid, thus introducing the stress/strain constitutive relationship for a medium with two distinct porosities (see dual-porosity discussion below), including the effects of pore fluid pressure, p, and temperature, T (a combined equation utilizing constitutive poroelasticity (e.g. [32], Eq. 7.42), with thermoelastic response, and utilizing two distinct pore fluid pressures as in Wilson and Aifantis [27]),

$$\sigma_{ij} = 2G\varepsilon_{ij} + \frac{2Gv}{1-2v}\varepsilon_{kk}\delta_{ij} - (\alpha_p^{(1)}p + \alpha_p^{(2)}p)\delta_{ij} - \alpha_T T\delta_{ij}, \tag{3}$$

where G is the shear modulus, v is the Poisson ratio, $\alpha_p^{(i)}$ and α_T are the coupling coefficients for fluid and thermal effects for the $^{(1)}$ fracture and $^{(2)}$ matrix, δ_{ij} is the Kronecker delta, and the linearized ("small") strains are defined as the symmetric part of the displacement gradient $u_{i,j}$, i.e.,

$$\varepsilon_{ij} = \tfrac{1}{2}(u_{i,j} + u_{j,i}). \tag{4}$$

Inserting Eq. (4) into Eq. (3) and the result into the equilibrium equation, Eq. (2), yields the Navier equation for the displacements, u

$$G\nabla u_i + \frac{G}{1-2v}u_{k,ki} = (\alpha_p^{(1)}p_{,i} + \alpha_p^{(2)}p_{,i}) + \alpha_T T_{,i} - b_i, \tag{5}$$

4.2. Conservation of momentum, mass, and energy—fluid

Fluid, aqueous species, and energy are transported through the system as defined by their respective mass and energy balances. The master equation for these processes is given in integral form as

$$\frac{d}{dt}\int_V M_\kappa \, dV = \int_\Gamma \mathbf{F}_\kappa \cdot \mathbf{n} + \int_V q_\kappa \, dV, \tag{6}$$

where the left-hand side represents the rate of accumulation of the conserved quantity (M_κ is mass of fluid, mineral mass, or energy density) resulting from the arrival of the fluxes $\mathbf{F}_\kappa$, (of fluid, mass, or energy) across the boundary, Γ, and complemented by volume sources, q_κ, distributed over the nominal element volume, V, for each component, κ (gas, liquid, advected species, or heat). In this discussion we have adopted (for clarity of coefficients) standard tensor notation, where bold values represent first or second order tensors. Eq. (6) may be transformed into its common PDE counterpart through use of the divergence theorem

$$\frac{\partial M_\kappa}{\partial t} = -\nabla \cdot \mathbf{F}_\kappa + q_\kappa, \tag{7}$$

where the mass, flux, and source terms must then be independently determined for a given system.

Mass, or energy density, M_κ, in Eq. (7) is defined for each component, κ, as the summation of the various contributions to the component across all phases (subscripted l, g, s for liquid, gas, or solid) as

$$M_\kappa = \phi S_l \rho_l X_l + \phi S_g \rho_g X_g + (1-\phi)\rho_s X_s, \tag{8}$$

where S is phase saturation, ρ is density (or species concentration), ϕ is porosity, and $X_{s,l,g}$ is mass fraction (or internal energy). Simplification then occurs for each calculation. The third term disappears for fluid mass calculations (no solid phase present), while the second and third terms are excluded from aqueous species mass (species may be present within the liquid medium, but not solid or gaseous).

Fluxes, $\mathbf{F}$, in Eq. (7) are given by the summation across phases ($\beta = l,g$) of the advective and diffusive terms as

$$\mathbf{F} = \sum_{\beta=l,g} \left(-X_\beta \rho_\beta \frac{\mathbf{k}k_\beta^r}{\mu_\beta}(\nabla p_\beta - \rho_\beta \mathbf{g}) \right) - \lambda_\beta \nabla C, \tag{9}$$

where the first term represents the contribution of advection through consideration of the multiphase extension of Darcy's law for relative permeability, k^r, intrinsic permeability vector, $\mathbf{k}$, dynamic viscosity, μ, and, as before, ρ is density of fluid (or concentration of species) and X is mass fraction for fluid transport, specific enthalpy for heat flow, or unity for chemical calculations. The second term represents diffusive transport as governed by the laws of Fick and Fourier, and introduces conductivity, λ_β, and gradient (of temperature or concentration), ∇C. This last diffusive term is only present when calculating the flux of temperature or concentration, and therefore disappears when calculating pure liquid flux. For heat flow calculations, λ_β is thermal conductivity, while for chemical flux $\lambda_\beta = \rho_\beta \tau \phi S_\beta D_\beta$ with tortuosity, τ, and diffusion coefficient, D_β. Note that a hydrodynamic dispersion concept is not utilized in the classic Fickian sense. Instead, TOUGHREACT utilizes the interaction of regions with differing velocities (fracture and matrix in a dual-porosity construct) to induce solute mixing [33]. In the case of mineral mass, the flux term disappears (colloid transport is not considered).

The source term, q_κ, in Eq. (7) may be comprised of an injection or withdrawal source or as an increase in species concentration (or mineral mass) due to dissolution (or precipitation). A thermal source may also arise due to a release of energy during chemical reactions. This last case is not currently considered. Sources of aqueous species and/or mineral mass are discussed in the following.

4.3. Chemical precipitation/dissolution

A generalized rate law for precipitation/dissolution of a mineral, m, is [34,35],

$$r_m = \text{sgn}(\log(Q_m/K_m^e))k_m^c A_m f(a_i)\left|1 - \left(\frac{Q_m}{K_m^e}\right)^\varphi\right|^n, \tag{10}$$

where k^c is the rate constant, A is the specific mineral reactive surface area per kg of H_2O, K^e is the mineral/water equilibrium constant, and Q is the ion activity product. The function $f(a_i)$ represents some (inhibiting or catalyzing) dependence on the activities of individual ions in solution such as H^+ and OH^- [36], and $\text{sgn}(Q_m/K_m^e)$ provides a direction of reaction: positive for supersaturated precipitation. The exponential parameters, φ and n, indicate an experimental order of reaction, commonly assumed to be unity. An additional term (multiplied by Eq. (10)) may also be introduced to represent the dependency of reactive surface area on liquid saturation [33]. Dependency of the rate constant

may be handled, to a reasonable approximation [37], via the Arrhenius expression,

$$k^c = k_{25}^c \exp\left(-\frac{E_a}{R_u}\left(\frac{1}{T} - \frac{1}{298.15}\right)\right), \tag{11}$$

for the rate constant at 25 °C, k_{25}, activation energy, E_a, and gas constant, R_u.

In the case of amorphous silica an alternate expression may be used following [38], where the precipitation rates reported in [39] were observed to underestimate behavior in geothermal systems. This new rate law, based upon experimental data for more complex geothermal fluids, becomes, in a form modified in [40] to approach zero as Q/K approaches one (i.e., as the system approaches equilibrium)

$$r_m = \text{sgn}(\log(Q_m/K_m^e))k_m A_m f(a_i)\left|\left(\frac{Q_m}{K_m^e}\right)^\mu - \left(\frac{K_m^e}{Q_m}\right)^{2\mu}\right|^n. \tag{12}$$

These are the formulations utilized in TOUGHREACT. Reactions between aqueous species (homogeneous reactions) are assumed to be at local equilibrium, and therefore governed by the relationship between the concentrations of basis (primary) species and their activities, partitioned by the stoichiometric coefficients. This relationship is termed the law of mass action (e.g. [34]). The assumption of local equilibrium greatly reduces the number of chemical unknowns and ODEs (between primary and secondary species), and is accurate to the extent that the true reaction rates outpace the rate of fluid transport in a given system. This is a correct assumption for most aqueous species [34] (and flow systems), but less so for slower redox reactions [33,34]. In TOUGHREACT, species activities are obtained from an extended Debye–Hückel equation with parameters taken from [41].

5. Deformable dual-porosity material

To represent the pressure loading of a fully or nearly liquid saturated system (particularly at high temperature and pressure and with multi-component liquids) coupling of the above formulation requires the undrained (instantaneous) response of pore fluid pressure to mechanical loading in both the fracture and matrix domains. Hydrologic considerations allow a timed pressure-dissipation response throughout the fracture dominated fluid system and between the fracture/matrix companionship following undrained loading.

Classically, a dual-porosity material is represented as a porous matrix partitioned into blocks by a mutually orthogonal fracture network [42,43]. In this scenario, permeability is much higher within the fracture network, thus allowing global flow to occur primarily through the fractures, while the vast majority of storage occurs within the higher porosity matrix (due to its larger global fraction of the medium). Interchange of fluid and heat between fractures and matrix, so-called "interporosity flow", is driven by pressure or temperature gradients between the two domains.

Expansion of this classic two-domain interaction into "multiple interacting continua" [44,45] allows the gradual evolution of gradients between fracture and matrix through the existence of one or more intermediate continua placed, mathematically, some linear distance from the fracture domain. This development has allowed for numerical approximations to more accurately represent the slow invasion of locally (to the fracture) altered pressures and temperatures deeply into the matrix blocks, and introduced dispersive mixing that arises at the interface of zones with differing fluid velocities. While this multi-continuum methodology may be adopted in TOUGHREACT to represent dual-permeability fluid transport with uniformly constant stress fields in

time, we do not seek such an expansion with respect to a flow-deformation response [46]. As such, a dual-porosity framework with two interacting continua (fracture and matrix) is utilized in this study, while a compatible poroelastic theory carries this behavior into the mechanical domain.

5.1. Fluid pressure response

Extension of Biot's poroelastic theory [47–50] to a dual-porosity framework has been previously addressed [27–31, 46,51]. The methodologies presented in these works provide an adequate framework for the phenomenological representation of poroelastic coefficients capable of describing flow-deformation response in such a medium.

Continuity of fluid mass is represented in a compressible media as,

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot \mathbf{F} = 0, \tag{13}$$

where ζ is the increment of fluid content as in [52], and comprises the relative motion between fluid and solid. Inserting Darcy's law for the flux term yields

$$\frac{\partial \zeta}{\partial t} - \frac{k}{\mu}\nabla^2 p = q. \tag{14}$$

Biot's [48] linear-poroelastic constitutive equivalence, for volumetric strain, e, is

$$\begin{pmatrix} e \\ \zeta \end{pmatrix} = \begin{pmatrix} \dfrac{1}{K} & \dfrac{1}{H} \\ \dfrac{1}{H} & \dfrac{1}{R} \end{pmatrix}\begin{pmatrix} \sigma \\ p \end{pmatrix}, \tag{15}$$

where the coefficients $1/K$, $1/H$, and $1/R$ are the bulk drained compressibility, poroelastic expansion, and specific storage, respectively. Substituting

$$B \equiv -\left.\frac{\delta p}{\delta \sigma}\right|_{\zeta=0} = \frac{R}{H}, \tag{16}$$

for the Skempton coefficient, $\alpha \equiv K/H$, for the Biot–Willis coefficient and

$$\frac{1}{R} \equiv \left.\frac{\delta \zeta}{\delta p}\right|_{\sigma=0} = \frac{\alpha}{KB}, \tag{17}$$

for the specific storage, condensing Eq. (15) to relate fluid content to strain, and substituting its time derivative in Eq. (14), establishes the flow condition for a single-porosity medium with no fluid sources

$$\frac{\alpha}{BK_u}\dot{p} + \alpha\dot{e} = \frac{k}{\mu}\nabla^2 p, \tag{18}$$

where we have utilized the relationship for undrained bulk modulus,

$$K_u \equiv \left.\frac{\delta \sigma}{\delta e}\right|_{\zeta=0} = \frac{K}{1-\alpha B}. \tag{19}$$

Extending to a dual-porosity medium, we follow the same procedure leading to the dual-porosity form of Eq. (5), where Eq. (18) is modified to exhibit two separate fluid pressures (for fracture and matrix) with flow between them governed by, in its simplest form, an instantaneous pressure differential, $\Delta p = (p_1-p_2)$ [42], to obtain two continuity relationships [28],

$$\frac{k^{(i)}}{\mu}\nabla^2 p^{(i)} = \frac{\alpha^{(i)}}{K_u^{(i)}B^{(i)}}\dot{p}^{(i)} + \alpha^{(i)}\dot{e} + (-1)^i\gamma\Delta p, \tag{20}$$

where i is not a repetitive index, but represents the existence of two separate equations for the matrix ($i = 1$) and fracture ($i = 2$),

and γ is the cross coupling coefficient for flow exchange between the two domains [53]. Eq. (20) states that the divergence of fluid flux for a given control volume must equal the rate of accumulation within that volume, and is thus a statement of mass conservation.

5.2. Dual-porosity load response

The general linear relation between strain, increment of fluid content, total stress (σ), and pore fluid pressure (p), simply extends Eq. (15) to allow, again, for two separate fluid pressures [31]

$$\begin{pmatrix} \delta e \\ -\delta\zeta^{(1)} \\ -\delta\zeta^{(2)} \end{pmatrix} = \begin{pmatrix} c_{11}c_{12}c_{13} \\ c_{21}c_{22}c_{23} \\ c_{31}c_{32}c_{33} \end{pmatrix} \begin{pmatrix} -\delta\sigma \\ -\delta p^{(1)} \\ -\delta p^{(2)} \end{pmatrix},$$

(21)

where the superscripts refer to the (1) matrix and (2) fracture domains. The single porosity coefficients of Biot are no longer applicable, and are replaced by the uknown coupling coefficients, c_{ij}, that may be designated via a phenomenological deconstruction similar to that of Biot and Willis [52]. The coefficient matrix can be shown to be symmetric [31] by the Betti reciprocal theorem. Performing manipulations of the above equation through isolation of independent components (i.e. long-time versus short-time limits) allows determination of the central coefficients (see detailed procedure in [31,30]).

Herein we assume that $c_{23} = c_{32} = 0$ [31], which differs slightly from the procedure of [27,28,30]. Examination of Eq. (21) shows that this assumption implies the following: an undrained application of stress that influences a change in fluid content for the fracture domain does so through modification of fracture fluid pressure, and does not influence that of the matrix. The reverse is also true, with the overall implication being, see discussion in Berryman and Wang [31], that in the undrained limit the matrix and fracture domains are completely separate. This can be considered a justification for a dual-porosity approach [31].

In our analysis, the purpose of dual-porosity elasticity is to attain Skempton coefficients representing both the fracture and matrix domains

$$\delta p^{(1)} = B^{(1)}\delta\sigma = -\frac{c_{12}}{c_{22}}\delta\sigma$$

$$\delta p^{(2)} = B^{(2)}\delta\sigma = -\frac{c_{13}}{c_{33}}\delta\sigma,$$

(22)

which represent the undrained ($\delta\zeta = 0$) build in pore fluid pressure in each domain for a given change in stress as provided by FLAC^{3D}. Relationships to calculate these two Skempton coefficients are provided in Table 2 of [31]. For this procedure, we choose as the known coefficients $K^{(1)}$, K, $K_s^{(1)}$, and K_f, where K_s is the solid grain modulus (in a microhomogeneous medium [54]) and the fluid bulk modulus,

$$\frac{1}{K_f} \equiv \frac{1}{V}\frac{\delta V}{\delta p}\bigg|_T,$$

(23)

is calculated in the interpolation module as a function of position, temperature, and pressure utilizing the IAPWS steam table equations [25]. For a complete reconstruction of the individual relations required to represent the dual-porosity poroelastic response, refer to [31,51].

5.3. Effect on the global mass balance

Injection of fluid mass into TOUGHREACT in the form of fluid pressure violates conservation of mass by an amount proportional to the compressibility of the local fluids. A change in pressure by

this procedure necessitates a change in local fluid volume, and therefore appearing or disappearing mass. However, when the local element is fully saturated, a stiff fluid will not significantly respond (volumetrically) to stress induced pressure changes, while for unsaturated media even a significant volumetric response will not in general dictate a noticeable change in mass. Nonetheless, we err on the side of safety and correct for this discrepancy with a recast of Eq. (23),

$$dV = \frac{1}{K_f}Vdp\bigg|_T,$$

(24)

which indicates the volume (or mass) error due to an increase in pressure, dp (at a given temperature). To correct for potential mass loss, we alter elemental volumes (physically reduce the volume of the mesh element) within TOUGHREACT by this amount (in an integral finite difference formulation, this does not require the alteration of geometric coordinates). In our simulations, including both single and multiphase flow with water/steam phase changes occurring, we have not detected total system mass losses greater than ~0.01% of total system mass.

6. Undrained fluid/mechanical response

We now examine the error that our formulation introduces to the fluid-mechanical coupling. Excluding constitutive approximations, error may be introduced into the coupling procedure as it has been described up to this point in two primary ways: explicit time step size, and the equilibration step between a stress change and its undrained pressure response (Fig. 3).

The first is a direct byproduct of explicit coupling, insomuch as an increase in time step (length of the TOUGHREACT fluid step between each mechanical equilibration), allows a greater amount of fluid pressure to diffuse between each mechanical equilibration, introducing error proportional to the fluid diffusivity and inversely proportional to the rate of mechanical change (not the amount of mechanical change per timestep, $d\sigma$, which implies proportionality to error, but the rate of change per unit time ($d\sigma/dt$), implying inverse proportionality).

The second form of error, shown in Fig. 3, is due to the nature of the undrained pressure response, which may not be fully accommodated by a single stress equilibrium step. In other words, at a given time step a fixed pressure field enters FLAC^{3D} and is accommodated by a calculated stress distribution. This stress distribution induces a modification of the previously fixed pressure field, and this new pressure field may, in turn, produce a redistribution of the stress field whether or not any fluid is allowed to diffuse (within TOUGHREACT). A number of steps may
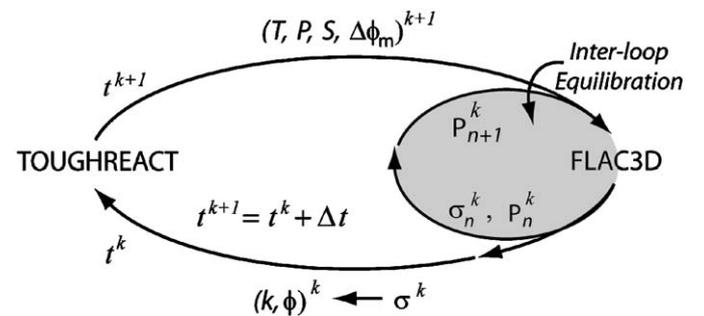

Fig. 3. Relationship between coupling methodologies. Interior looping may occur over n steps (at fixed time, $t = t^k$) to equilibrate the response of stress to an undrained increase in pressure. Alternatively, this inter-looping may be excluded in favor of a "leapfrog" method, where a single stress equilibration (run of FLAC^{3D}) is conducted per time step.

be required to find the true equilibrium magnitude of stress and pressure, which tends to asymptote at a value higher than is suggested by a single equilibration step. This is not necessarily a Mandel-Cryer type effect [55,56], which is a real occurrence and would require the action of a diffusing fluid pressure and redistribution of stresses around the diffusing magnitudes (although the behavior is comparable). The case where FLAC³ᴰ is run once per explicit time step (single equilibration step) is referred to herein as the "leapfrog method" (Fig. 3). Each of these possible error sources ([1]explicit time step and [2]leapfrog versus $p \rightleftharpoons \sigma$ iteration) requires further examination, which consequently leads to validation of the undrained fluid-mechanical coupling.

6.1. Fluid-mechanical couple: instantaneous loading

In one dimension, we may examine the accuracy of the fluid-mechanical coupling in comparison to the classical fluid diffusion equation of hydrogeology (e.g. [57]),

$$\frac{\partial p}{\partial t} - c_f \frac{\partial^2 p}{\partial z^2} = 0, \tag{25}$$

which is a specific poroelastic result of Eqs. (14) and (15) restricted to a one-dimensional column of soil (or rock) under constant applied vertical stress [58], and gives its form to the analytical solution for heat flow [59, p. 96]

$$p(z,t) = \frac{4p_0}{\pi} \sum_{m=0}^{\infty} \frac{1}{2m+1} \exp(-\psi^2 c_f t) \sin(\psi z), \tag{26}$$

where $\Psi = (2m+1)\pi/2L$, and $p_0 = B^{(v)}\sigma_0$ is the initial undrained pressure response to the applied vertical stress (σ_0). The one-dimensional Skempton coefficient ("loading efficiency" in [58]) is given by

$$B^{(v)} = -\frac{B(1 + v_u)}{3(1 - v_u)}, \tag{27}$$

for the Skempton coefficient, B, and undrained Poisson ratio, v_u. This is the canonical consolidation problem of a one-dimensional column of soil subjected to a constant vertical stress applied at $t = 0^+$ to the top of the column, with fluid pressure allowed to drain freely from the point of applied stress. A similar solution is available for column displacement u (e.g. [48,58]),

$$\frac{\partial^2 u}{\partial z^2} = c_m \frac{\partial p}{\partial z}, \tag{28}$$

for Geertsma's [60] uniaxial expansion coefficient (consolidation coefficient), $c_m \equiv \alpha/K^{(v)}$, with uniaxial bulk modulus, $K^{(v)} = K+4G/3$. Under the same boundary conditions as above, the analytical solution is [58]

$$\Delta u(z,t) = c_m p_0 \left[(L-z) - \frac{8L}{\pi^2} \sum_{m=0}^{\infty} \frac{1}{(2m+1)^2} \exp(-\Psi^2 c_f t) \cos(\Psi z) \right], \tag{29}$$

with definitions the same as for Eq. (26), and the instantaneous displacement at the time of stress application $u(z,0^+) = \sigma_0(L-z)/K_u^{(v)}$, for the undrained unaxial bulk modulus,

$$K_u^{(v)} = \frac{K_u(1 + v_u)}{3(1 - v_u)}. \tag{30}$$

All undrained parameters approach their drained counterparts as fluid compressibility becomes large, or fluid saturation approaches zero.

Results of a TOUGHREACT-FLAC³ᴰ simulation mimicking these boundary conditions are presented against these analytical solutions in Fig. 4. A column of porous rock ($E = 13$ GPa,
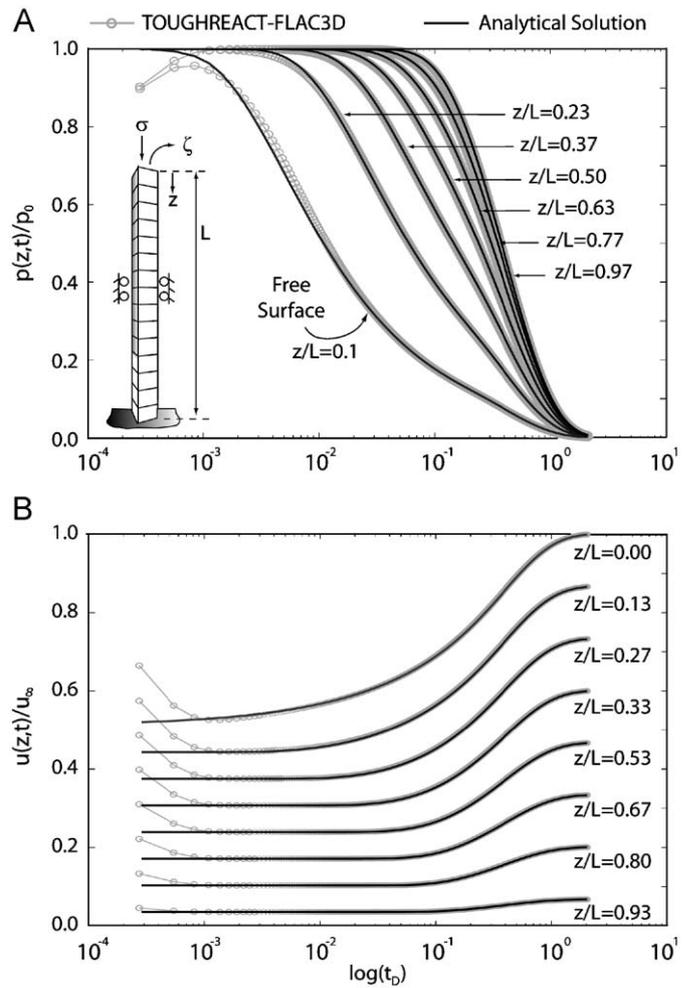

Fig. 4. Comparison of TOUGHREACT-FLAC³ᴰ fluid-mechanical coupling simulation versus analytical results in one dimension: (A) normalized ($p_0 = p(z,0) = B^{(v)}\sigma_0$) pressure diffusion response versus diffusive time ($t_D = ct/L^2$); (B) normalized ($u_\infty = u(0,\infty) = \sigma_0 L/K^{(v)}$) displacement response versus diffusive time.

$v = 0.22$) with displacements constrained laterally and pore pressure initially zero, is subjected to an applied vertical load, $\sigma_0 = 50$ MPa, at $t = 0^+$, and pressure is allowed to drain freely from the top of the column only. Time step was chosen large enough to illustrate the error incorporated in very early times (near the time of undrained loading) due to the leapfrog method of simulation (Fig. 3).

Pressure builds up (and elastic displacement decreases) in the early stages as the model cycles between stress equilibration and undrained pressure response (*leapfrog* artifact). Following the instantaneous loading period (50 MPa applied over one time step) numerical results overlay nearly identically the analytical solution as pressure diffuses and stress accommodates the pressure reduction. A slightly greater error occurs at points nearest the free draining surface (left-most curve in Fig. 4A) due to the explicit time step size, where a greater rate of fluid diffusion allows the fluid to move greater distances before being accommodated by a mechanical response.

6.2. Fluid-mechanical couple: constant loading rate

In light of Fig. 4, it is of interest to examine more precisely the error that arises while the sample is being loaded. To do so, we wish to utilize the same geometry, but apply the load gradually over a finite loading period at a given loading rate, $d\sigma_0/dt$ (rate of

increase of applied load at the top of the column per unit time). Here, we maintain the one-dimensional form, but alter the governing diffusion equation (25) to accommodate a constant loading rate [58]:

$$\frac{1}{c_f}\frac{\partial p}{\partial t} - \frac{\partial^2 p}{\partial z^2} = \frac{c_m \mu}{k}\frac{d\sigma_0}{dt}, \tag{31}$$

with the series solution adjusted so that, as above, the free draining boundary is at $z = 0$ [59, p. 130],

$$p(z,t) = \frac{c_m d\sigma_0}{dt}\frac{L^2\mu}{2k}\left(1 - \frac{(L-z)^2}{L^2} - \frac{32}{\pi^3}\sum_{m=0}^{\infty}\frac{(-1)^m}{(2m+1)^3}\right.$$
$$\left. \times \exp(-\Psi^2 c_f t)\cos(\Psi(L-z))\right). \tag{32}$$

Results of the gradual loading analysis are presented in Fig. 5. Loading rate refers to the rate of increase of applied load at the top of the column per unit time. The amount of load change per iteration ($d\sigma_0$) is a function of the time step (dt), so that a smaller load change is experienced per iteration as the time step is decreased. Time steps were chosen for A and B such that $d\sigma_0 = (d\sigma_0/dt)\times dt$ is the same magnitude in each case. From the figure, two primary conclusions are apparent. Firstly, at the slowest loading rate (Fig. 5A) and smallest time step (and correspondingly smallest value of $d\sigma_0$) there is no difference between the leapfrog approach and a simulation with additional $p \rightleftharpoons \sigma$ iteration (inter-looping), proving the intuitive result that small explicit time steps remove the need for inter-looping. In this

case, if the time step is too large to capture the fluid-mechanical coupling, then inter-looping has little effect because more error is introduced by the fluid-mechanical couple than by the leapfrog method (evidenced by the fact that the dashed lines do not improve in accuracy over their corresponding solid lines). Secondly, a faster loading rate (Fig. 5B) results in greater error due to the leapfrog method, but lesser error due to the explicit time step size (evidenced by the relative accuracy of all three dashed lines). In other words, mechanical change (loading) is faster relative to fluid diffusion, and so the explicit time step size may be larger and still accommodate the fluid-mechanical coupling because less frequent mechanical equilibration is required to keep up with the relatively slower fluid diffusion. However, precisely because the loading rate is faster, greater error will result due to the non-iterative equilibration of stress and pressure. Therefore, a larger time step is viable, but only with inter-looping. In any case, the system may be accurately represented with the proper selection of time step and iterative method for a given rate of mechanical change, and at the slower loading rate (likely closer to those that might be seen in natural systems) the leapfrog method is sufficient provided that the explicit time step is reasonably small. For now, experimentation is required to guarantee accurate coupling.

7. THMC mediated aperture/permeability change

Having now examined the fluid-mechanical mechanism, we proceed to introduce further complexities that surround chemical behavior. And, because constitutive behavior in a geological system is generally non-linear, responses mediated by stress, fluid pressure, temperature, and chemical potential often require empirical examination. Notably, permeability of the system may change by orders of magnitude in response to changes in effective stress. In the following, we describe changes in permeability resulting from both stress and chemical effects, utilizing the empirical relationship proposed in [61]. That relationship is further developed herein to accommodate unloading of fracture asperities in a manner that suggests fracture gaping may occur only through mechanical means (or by thermal contribution to the stress field). Section 7.1 presents the governing loading equations as found in [61], whereas Section 7.2 illustrates an unloading construct similar to that used in [61], but where unloading is allowed to occur only through mechanical means.

7.1. Loading behavior

Hydraulic aperture of a fracture under an applied effective stress, σ', may be defined empirically as [3]

$$b_m = b_m^r + (b^0 - b_m^r)\exp(-\omega\sigma'), \tag{33}$$

where b_m is the hydraulic aperture (subscripted m indicating changes due solely to mechanical effects), b^0 is the aperture under no mechanical stress, b_m^r is the residual aperture at maximum mechanical loading and ω is a constant that defines the non-linear stiffness of the fracture.

The dissolution of bridging asperities may also reduce the effective aperture of the fracture. These "chemical" effects may be accommodated in the relationship for fracture aperture in a form that includes the mechanical compaction process of Eq. (33) and pressure solution-type dissolution of contacting asperities, where we have substituted $b_m^{max} = b^0 - b_m^r$ as the maximum possible mechanical closure [61]

$$b_{mc} = b_m^r + \{b_m^r - b_c^r + b_m^{max}\exp(-\omega\sigma')\}\cdot\exp(-\sigma'(\beta - \chi/T)), \tag{34}$$
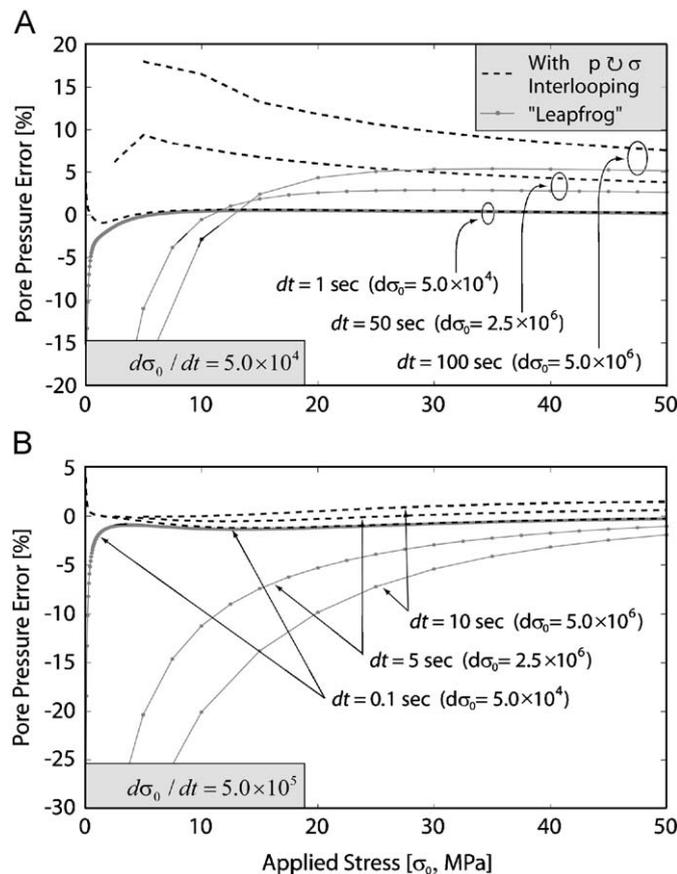

Fig. 5. Error (compared to analytical solution) in undrained pore pressure response for constant loading rate of one-dimensional vertical column for (A) slower loading rate ($d\sigma_0/dt = 5.0 \times 10^4$) and (B) faster loading rate ($d\sigma_0/dt = 5.0 \times 10^5$). "Leapfrog" method of simulation is solid gray line with data points. Additional inter-looping method is dashed black line.

where T is temperature and the empirical constants β and χ define the chemical compaction process. The subscripted c represents changes due to chemical effects and b_c^r is the residual aperture at maximum chemical loading. Buried within these two constants (see [61]) is the critical stress [61, modified from 62,63],

$$\sigma_c = \frac{E_m(1 - T/T_m)}{4V_m},\tag{35}$$

where E_m is the heat of fusion, T_m is the temperature of fusion, and V_m is the molar volume of the mineral comprising the fracture asperity. Dissolution of the contacting asperity will progress where the local asperity stress exceeds this critical stress that represents both the chemical and mechanical potential of the contact.

Permeability is evaluated for an orthogonal set of persistent fractures of spacing s, from the cubic law, [64,65] $k = b^3/12\,s$. Note that Eq. (34) represents equilibrium behavior, where chemically mediated changes have run to completion (i.e., it is a thermodynamic, not a kinetic relationship).

7.2. Unloading behavior

The above constitutive relationship governs aperture closure under conditions of thermal/mechanical loading due to the effects of mechanical deformation (Eq. (33)) and chemical alteration including mechanical deformation (Eq. (34)). If utilized in its entirety and without memory of any previous mechanical/thermal state, this represents the case of complete reversibility. However, aperture closure should not be viewed as completely reversible or irreversible, but as a mechanism that is dependent on the initial stress state and subsequent loading, as well as one that maintains memory of some attained stress magnitude and a subsequent unloading period.

For instance, subsurface storage of radioactive waste is characterized by a loading period during which temperature steadily increases and fracture apertures correspondingly decrease, followed by a period of sustained cooling towards the background state, implying a reversal of this process (fracture gaping). Alternatively, geothermal reservoirs are largely characterized by unloading behavior, where the maximum stress/temperature condition is the in-situ state of the fractured mass, and the injection of cooler circulation fluids causes unloading from this in-situ state. It is of some interest to determine the precise behavior of such an unloading period and its beginning transition.

The mechanical component of fracture closure is not a completely reversible process, but exhibits hysteresis as governed by both the elastic and plastic properties of the contacting asperities. Furthermore, while chemical behaviors may contribute to permeability increases through the action of thermodynamically governed dissolution, pressure solution type mechanisms as discussed above are incapable of inducing gaping of the fracture during an unloading stage (barring the inclusion of "force of crystallization" processes, pressure solution is irreversible). Therefore, it is apparent that an additional term is needed to describe the reversible portion of mechanical closure, while excluding the possibility of chemical reversibility. In this aim, we follow a procedure similar to that of Min et al. [61] to develop an unloading relationship, but maintain a reversibility that is due purely to mechanical effects.

In the simplest formulation, this need may be addressed through a mechanical recovery ratio, R_m, that governs the degree of elastic reversibility, and is defined as the ratio of the potential unloading mechanical aperture change, $b_{m(u)}^{\max}$, to the maximum potential loading mechanical aperture change, $b_m^{\max}$, as

$$R_m = \frac{b_{m(u)}^{\max}}{b_m^{\max}}.\tag{36}$$

It is first necessary to examine the case of a mass unloaded from a state of infinite stress with the unloading version of Eq. (33)

$$b_{m(u)} = b_m^r + b_{m(u)}^{\max}\exp(-\omega\sigma'),\tag{37}$$

or, from the definition of recovery ratio

$$b_{m(u)} = b_m^r + R_m b_m^{\max}\exp(-\omega\sigma').\tag{38}$$

However, the unloading process is dependent on the maximum loading stress (initial unloading stress). The difference in aperture between this maximum loading stress and some unloaded state is, utilizing Eq. (38),

$$\Delta b_{m(u)} = R_m b_m^{\max}\exp(-\omega\sigma_{(u)}') - R_m b_m^{\max}\exp(-\omega\sigma_{\max}'),\tag{39}$$

with the maximum (prior to unloading) effective stress $\sigma_{\max}' > \sigma_{(u)}'$, for any subsequent unloading effective stress, $\sigma_{(u)}'$. This inequality states that load cycling is not considered. The unloaded aperture is then comprised of the difference between this change and the fully loaded aperture, b^f:

$$b_{m(u)} = b^f + \Delta b_{m(u)}.\tag{40}$$

In the case of mechanical loading and unloading, the aperture at maximum loading stress, b^f, is equivalent to the final loaded aperture, $b_m(\sigma_{\max}')$, and so the unloading aperture is obtained by substituting Eq. (33) in Eq. (40). However, we are seeking the relationship for a fracture that has been chemically and mechanically loaded, and then unloaded along a path defined by the recoverable portion of mechanical loading. Therefore, substituting $b^f = b_{mc}(\sigma_{\max}')$ and inserting Eq. (39) into Eq. (40) and simplifying yields

$$b_{m(u)} = b_{mc}(\sigma_{\max}') + R_m b_m^{\max}\{\exp(-\omega\sigma_{(u)}') - \exp(-\omega\sigma_{\max}')\},\tag{41}$$

where $b_{mc}(\sigma_{\max}')$ is Eq. (34) evaluated at $\sigma' = \sigma_{\max}'$. This relationship defines the aperture at a stress magnitude lower than and obtained a posteriori the fully loaded state. Eqs. (34) and (41) then fully define the loading and unloading cycle, respectively, of a fractured mass. The required empirical parameters are shown in Table 1. Parameters were obtained through a comparison with experimental results introduced in the heated block test of Terra Tek [66], where the aperture was monitored during a complete loading and unloading cycle in-situ, on a 2×2 m cube of granitic gneiss subjected to stresses supplied by flatjacks with temperature alteration via borehole heaters. The original experimental results of Hardin et al. [66] are shown in Fig. 6, alongside theoretical reproduction of this behavior calculated with Eqs. (34) and (41). In the figure, loading begins at point 9 (and is isothermal for the first three data points) and continues until point 16 (non-isothermally), before being unloaded to the initial state at point 21. Hardin et al. also performed two intermediate load/unload cycles at points 13 and 16. These two intermediate cycles are not considered here, and the analytical solution is incapable of

Table 1
Parameters of the permeability constitutive relationship as utilized in Fig. 6.

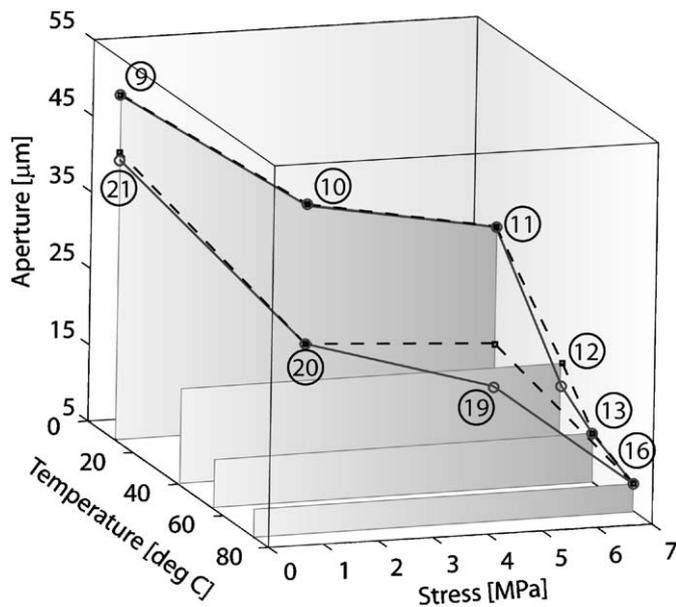| Parameter	Fitted value
Residual mechanical aperture, b_m^r (μm)	6.0
Residual chemical aperture, b_c^r (μm)	3.0
Constant in aperture relationship, β	1.00
Constant in aperture relationship, χ	345
Stiffness coefficient (1/MPa)	0.375
Mechanical recovery ratio, R_m	0.8

Fig. 6. Comparison of the analytical results of Eqs. (34) and (41) against experimental results of [66]. Experimental results are shown as black dashed line with solid data points. Gray solid line with hollow data points is the analytical solution. Each data point is numbered to correspond with the original data points of [66].

representing them. Agreement between the two data sets is satisfactory for the primary points of interest (intermediate loading/unloading cycling is not considered), excluding point 19, where unloading aperture cannot be reproduced with the given analytical model (which is purely mechanical and does not undergo unloading with decreasing temperature unless the stress field is altered).

7.3. Force of crystallization

Chemical precipitation is commonly assumed to cause a reduction in fracture aperture due to a buildup of deposited species along the fracture face. Contrary to this assumption is the concept of "force of crystallization", dating back to 1896 with the work of Dunn [67] and 1920 with Tabor [68], with a phenomenological model presented by Weyl [8]. Force of crystallization operates analogously and inversely to pressure solution where, instead of relieving fracture stress through dissolution at asperity contacts, if the fluid is sufficiently super saturated mineral precipitation and crystal growth may exert pressure at contact points and lead to physical gaping of the fracture. Further discussion of the mechanism is available in the literature (e.g. [69–71]). While we do not, in a fundamental sense, implicitly consider the impact of this process in our model, the current logic is capable of accommodating this effect in a straightforward manner—should solution concentrations be sufficiently super saturated. The phenomenological relationship for pressure solution that we utilize is able to adequately match the laboratory studies on which it is based, all of which involve significantly under saturated fluids only.

8. THC mediated porosity/permeability change

Thermo–chemical induced changes in permeability may be referenced to precipitation/dissolution behaviors along the continuum fracture and matrix domains. Here, aperture changes are caused by the addition or removal of mineral components from the walls of (at the scale of these investigations) an assumed uniform fracture face, or an isotropic porous volume fraction. This is not precisely "free face dissolution" (which implies contribution of strain energy to thermodynamic dissolution), but a purely chemically driven process governed by the rates of reaction as previously discussed. In the following, we assume that processes of this type may act independently from pressure solution over a single time step, thus enabling them to be additive over that time step. This does not indicate process independence, which would allow chemical analyses to be conducted separately of TM or of TMC without loss of accuracy. These processes are still strongly dependent on one another outside of a time step. For example, changes in permeability from pressure solution (or chemical precipitation/dissolution) will alter the flow characteristics and residence times of circulating fluids, thus modifying thermal transport. Changes in local temperature in this manner alter the stress field and modify chemical reaction rates. Modified reaction rates and residence times influence the characteristics of chemical reaction, while modified temperature and stress influence pressure solution and thermal gaping.

Changes in fracture aperture due to THC behavior are accommodated via the chemical precipitation behavior incorporated in TOUGHREACT. Addition or removal of mineral mass from the continuum system results in a change in fracture or matrix porosity within a nominal element volume, as given by the overall change in the volume of minerals present by [40,72],

$$\phi = 1 - \sum_{m=1}^{N} f_m^{rx} - f^u, \tag{42}$$

where f^{rx} is the volume fraction of mineral m in the surrounding rock $v_{mineral}/v_{medium}$, and f^u is the volume fraction of the non-reactive surrounding rock. Relations between fracture porosity and permeability are provided in the literature. One such possibility is a simple cubic relationship [34]:

$$k = k_i \left(\frac{\phi}{\phi_i} \right)^3, \tag{43}$$

where the subscript, i, refers to an initial property and k, and ϕ are permeability and porosity, respectively. While several such relations may be implemented from within TOUGHREACT, it is necessary in our case to calculate permeability changes externally in order to operate multiple mechanisms simultaneously. Compatibility between the permeability change due to this behavior and that of pressure solution can be indexed to the change in fracture aperture by, as before, $b = \sqrt[3]{12ks}$. Aperture change via this mechanism is then assumed additive to the THMC aperture reductions associated with pressure solution driven compaction.

Several options also exist for the relationship between matrix porosity and permeability. One such possibility is the Carman–Kozeny equation [73],

$$k = k_i \frac{(1 - \phi_i)^2}{(1 - \phi)^2} \left(\frac{\phi}{\phi_i} \right)^3, \tag{44}$$

where all parameters are as previously defined, although matrix permeability is likely an insignificant contributor (in many cases) to overall system behavior.

9. Chemical strain and stress

Modifications in fracture aperture necessarily lead to changes in the local stress field. However, because FLAC³D uses grid point displacements to calculate strains, see Eq. (4), and does not store values of strain, no provision is available to input strains due to
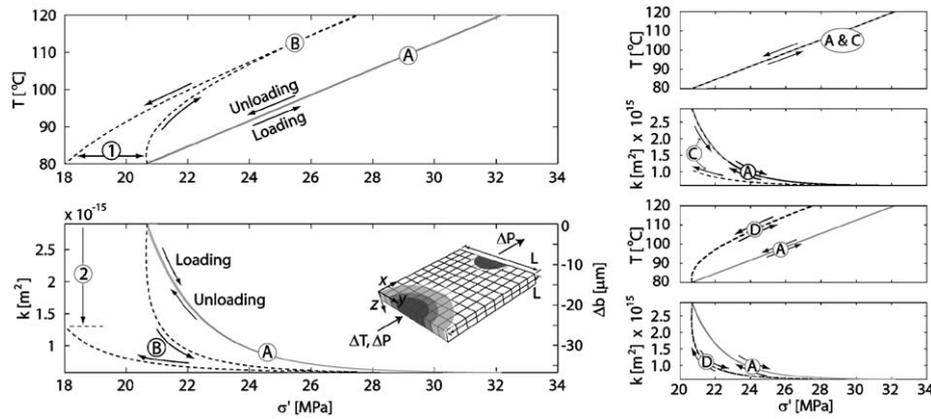
Fig. 7. Thermal loading/unloading cycle examining the effects of chemical strain. Parameters: $E = 13\,GPa$, $v = 0.22$, $\alpha_T = 12 \times 10^{-6}/K$.

aperture change and subsequently convert them into gridpoint displacements. An alternative method is needed.

For equally spaced orthogonal fractures, the impact of a change in aperture on local linear strain (unidirectional from a single fracture) is represented by

$$\varepsilon_{CH} = \frac{\Delta b}{s}, \tag{45}$$

where s is fracture spacing and the subscript CH refers to the "chemical strain" component of total strain owing to aperture change. In the usual manner, total strain, ε, can be spectrally decomposed into components due to mechanical, M, chemical, CH, and thermal, T, behaviors as, $\varepsilon = \varepsilon_M + \varepsilon_T + \varepsilon_{CH}$. Considering only thermal and chemical effects, the thermal/chemical strain is $\varepsilon_{TC} = \varepsilon_T + A$, where A is some constant representing the chemical portion. At incremental equilibrium we have $\varepsilon_{TC} = \alpha_T \Delta T + A$ which, upon rearranging, becomes

$$\varepsilon_{TC} = \Delta T \left(\alpha_T + \frac{A}{\Delta T} \right), \tag{46}$$

where $A = \Delta b/s$. This relationship provides a method to accommodate chemical strain by altering the coefficient of thermal expansion, α_T, in FLAC^{3D} at all nominal element volumes for respective aperture changes, and, as desired, maintains a non-linear dependence on temperature. However, because the function is undefined for temperature changes approaching zero, care should be taken in its application. In physical systems where aperture change, which is a strong function of the effective stress field, is dominated by thermal stress, such as geothermal systems, such strains will be tracked appropriately, but in systems that are nearly isothermal this method will be ineffective in transferring information to the mechanical system (which may or may not be necessary, as isothermal systems are unlikely to experience chemical strain to the same degree).

9.1. Chemical strain in cyclic loading

Chemical strain is defined here as thermo-chemo-mechanically irreversible reduction in fracture aperture that results in a relaxation of stress in the surrounding rock. As illustrated in Fig. 1, this process is proposed to be of significant importance in fractured reservoirs and replicating it one of the primary goals of THMC modeling. To examine this process, we consider the case of a liquid saturated, high temperature and pressure fractured mass subjected to a complete cycle of thermal loading and unloading (Fig. 7). The model is a pseudo three-dimensional mass (unit width in the z-direction, discretized in x and y) with zero-displacement boundaries and initially at a uniform temperature of

$80\,°C$, and $\sigma' = 20.8\,MPa$. A high temperature ($120\,°C$) and pressure ($2\,MPa$ above *in-situ*) source is placed at one end of the geometry ($x = 0$, $y = L/2$) with a low pressure source ($2\,MPa$ below *in-situ*) at the opposing end ($x = L$, $y = L/2$), allowing the thermal source to translate across the geometry with the fluid pressure gradient. After thermal breakthrough to the injection temperature, the temperature source is reversed to $80\,°C$, so that the mass then gradually declines to its initial temperature state. Progress is monitored at the central coordinate ($x = L/2$, $y = L/2$), and the results of temperature and aperture change versus stress at this location are displayed.

In the figure we present four cases incorporating different assumptions of response, which may be compared to the conceptual representation of Fig. 1. Fig. 7A is the baseline case, with completely reversible permeability change (Eq. (34) only), and no feedback of this chemical strain on the stress field (Eq. (46) not used). Fig. 7B represents the case of complete permeability constitutive treatment (Eqs. (34) and (41)) and includes feedback on stress field (Eq. (46)). Fig. 7C maintains full permeability constitutive treatment (as in 7b), but this time does not include feedback on stress (Eq. (47) not used). Finally, Fig. 7D considers complete reversibility (as in 8a), but this time includes feedback on the stress field (Eq. (47)).

The non-linear dependence of aperture on the temperature/stress field is evident, as is the non-linear dependence of stress on temperature that results from the feedback of chemical strain on the stress field. Two-dominant impacts on the system, hysteretic in nature, are visible by comparing the initial, ambient system with the final, ambient system. Importantly, when the system returns to its initial state, there has been an irreversible reduction in the stress field as well as an irreversible decrease in permeability. Neither of these occurrences, intuitively operative and significant in natural systems, may be represented without the inclusion of thermal, hydrologic, mechanical, and chemical processes.

10. Conclusions

A coupled THMC simulator has been developed with the capability of reproducing the undrained loading behavior of a fractured rock mass. Reactive transport has been included in the model via the equilibrium behavior of aqueous species (homogeneous reactions) and through kinetic considerations of mineral precipitation and dissolution. From multi-continuum hydrogeologic analysis, multi-phase fluid behavior is coupled to the mechanical response in one continuum via dual-porosity poroelasticity and thermodynamically controlled fluid

compressibility. Permeability of the mass is followed with a new constitutive relationship representing thermal loading and unloading behavior: Closure of the fracture is controlled by thermal-elastic compaction and the dissolution of stress-concentrated asperities, while dilation occurs via thermal-hydraulic stress relaxation. Bulk permeability is also modified by the precipitation/dissolution kinetics of mineral species. The explicit coupling between THC and M behaviors is shown to reproduce the rapid response of a loaded mass. Additional couplings have also been explored, and a subsequent paper [1] examines the strength of coupling between THMC mechanisms as well as the application of this model to an EGS scenario.

Chemical strain is accommodated by the permeability constitutive relationship, and its impact on the stress field of a geologic environment is illustrated. For the first time, we present geologic scale numerical results illustrating the conceptual model that thermal loading may lead to an irreversible reduction in aperture and stress, so that the *in-situ* system may be completely altered by a cycle of loading.

Acknowledgment

References

[1] Taron J, Elsworth D. Thermal-hydrologic-mechanical-chemical processes in the evolution of engineered geothermal reservoirs. Int J Rock Mech Min Sci 2009; this issue, doi:10.1016/j.ijrmms.2009.01.007.

[2] Wawersik WR, Rudnicki JW. Terrestrial sequestration of CO_2—an assessment of research needs. Report from envited panelist workshop, May 1997, US Dept Energy Geosci Res Prog, 1998.

[3] Rutqvist J, Wu Y-S, Tsang C-F, et al. A modeling approach for analysis of coupled multiphase fluid flow, heat transfer, and deformation in fractured porous rock. Int J Rock Mech Min Sci 2002;39:429–42.

[4] Taron J, Elsworth D, Thompson G, et al. Mechanisms for rainfall-concurrent lava dome collapses at Soufriere Hills Volcano, 2000–2002. J Volcanol Geoth Res 2007;160:195–209.

[5] Nemat-Nasser S, Keer LM, Parihar KS. Unstable growth of thermally induced interacting cracks in brittle solids. Int J Solids Struct 1977;14:409–30.

[6] Hunsbedt A, Kruger P, London AL. Recovery of energy from fracture-stimulated geothermal reservoirs. J Petrol Tech 1977;29:940–6.

[7] Pruess K. Enhanced geothermal systems (EGS) using CO_2 as working fluid—a novel approach for generating renewable energy with simultaneous sequestration of carbon. Geothermics 2006;35(4):351–67.

[8] Weyl PK. Pressure solution and the force of crystallization—a phenomenological theory. J Geophys Res 1959;64:2001–25.

[9] Paterson MS. Nonhydrostatic thermodynamics and its geologic applications. Rev Geophys Space Phys 1973;11:355–89.

[10] Revil A. Pervasive pressure-solution transfer: a poro-visco-plastic model. Geophys Res Lett 1999;26(2):255–8.

[11] Elsworth D, Goodman RE. Characterization of rock fissure hydraulic conductivity using idealized wall roughness profiles. Int J Rock Mech Min Sci 1986;23(3):233–43.

[12] Barton N, Bandis S, Bakhtar K. Strength, deformation and conductivity coupling of rock joints. Int J Rock Mech Min Sci 1985;22(3):121–40.

[13] Rose P, Xu T, Kovac KM, et al. Chemical stimulation in near-wellbore geothermal formations: silica dissolution in the presence of calcite at high temperature and high pH. In: Proceedings of the 32nd workshop geothermal reservoir engineering, Stanford University; 2007.

[14] Xu T, Sonnenthal E, Spycher N, et al. TOUGHREACT—A simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media: applications to geothermal injectivity and CO_2 geological sequestration. Comp Geosci 2006;32:145–65.

[15] Nami P, Schellschmidt R, Schindler M, et al. Chemical stimulation operations for reservoir development of the deep crystalline HDR/EGS system at Soultz-Souz-Forêts (France). In: Proceedings of the 32nd workshop geothermal reservoir engineering, Stanford University; 2007.

[16] De Boer RB. On the thermodynamics of pressure solution-interaction between chemical and mechanical forces. Geochem Cosmochem Acta 1977;41:249–56.

[17] Bower KM, Zyvoloski G. A numerical model for thermo-hydro-mechanical coupling in fractured rock. Int J Rock Mech Min Sci 1997;34(8):1201–11.

[18] Gawin D, Schrefler BA. Thermo-hydro-mechanical analysis of partially saturated porous materials. Eng Comput 1996;13(7):113.

[19] Taron J, Min K-B, Yasuhara H, et al. Numerical simulation of coupled thermo-hydro-chemo-mechanical processes through the linking of hydrothermal and solid mechanics codes. In: Proceedings of the 41st US symposium on rock mechanics, Golden, Colo, June 17–21, 2006.

[20] Swenson D, Gosavi S, Hardeman B. Integration of poroelasticity into TOUGH2. In: Proceedings of the 29th international workshop geothermal reservoir engineering, Stanford University; 2004.

[21] Itasca Consulting Group Inc. FLAC3D Manual: fast Lagrangian analysis of continua in 3 dimensions—version 2.0. Itasca Consulting Group Inc, Minneapolis; 1997.

[22] Settari A. Modeling of fracture and deformation processes in oil sands. In: Proceedings of the fourth UNITAR/UNDP conference on heavy crude and tar sands, Edmonton; 1988. Paper no. 43.

[23] Settari A, Mourits FM. Coupling of geomechanics and reservoir simulation models. Comp Methods Adv Geomech 1994:2151–8.

[24] Minkoff SE, Stone CM, Bryant S, et al. Coupled fluid flow and geomechanical deformation modeling. J Pet Sci Eng 2003;38:37–56.

[25] IAPWS, Industrial formulation 1997 for the thermodynamic properties of water and steam. IAPWS Release, IAPWS Secretariat, 1997: International Association for the Properties of Water and Steam.

[26] Narasimhan TN, Witherspoon PA. An integrated finite difference method for analyzing fluid flow in porous media. Water Resour Res 1976;12(1): 57–64.

[27] Wilson RK, Aifantis EC. On the theory of consolidation with double porosity. Int J Eng Sci 1982;20(9):1009–35.

[28] Khaled MY, Beskos DE, Aifantis EC. On the theory of consolidation with double porosity, III, A finite element formulation. Int J Numer Anal Methods Geomech 1984;8(2):101–23.

[29] Cho TF, Plesha ME, Haimson BC. Continuum modelling of jounted porous rock. Int J Numer Anal Methods Geomech 1991;15:333–53.

[30] Elsworth D, Bai M. Flow-deformation response of dual-porosity media. J Geotech Eng 1992;118(1):107–24.

[31] Berryman JG, Wang HF. The elastic coefficients of double-porosity models for fluid transport in jointed rock. J Geophys Res 1995;100(12): 24,611–27.

[32] Jaeger JC, Cook NGW, Zimmerman RW. Fundamentals of rock mechanics. 4th ed. Oxford: Wiley-Blackwell; 2007.

[33] Xu T, Pruess K. Modeling multiphase non-isothermal fluid flow and reactive geochemical transport in variably saturated fractured rocks: 1. Methodology. Am J Sci 2001;301:16–33.

[34] Steefel CI, Lasaga AC. A coupled model for transport of multiple chemical species and kinetic precipitation/dissolution reactions with applications to reactive flow in single phase hydrothermal system. Am J Sci 1994;294: 529–92.

[35] Lasaga AC. Chemical kinetics of water-rock interactions. J Geophys Res 1984; 89(B6):4009–25.

[36] Lasaga AC, Soler JM, Ganor J, et al. Chemical weathering rate laws and global geochemical cycles. Geochem Cosmochem Acta 1994;58:2361–86.

[37] Lasaga AC. Rate laws of chemical reactions. In: Lasaga AC, Kirkpatrick RJ, editors. Kinetics of geochemical processes, reviews in mineralogy, vol. 8. Washington: Mineral Soc Amer; 1981.

[38] Carroll S, Mroczek E, Alai M, et al. Amorphous silica precipitation (60–120 °C): comparison of laboratory and field rates. Geochem Cosmochem Acta 1998;62:1379–96.

[39] Rimstidt JD, Barnes HL. The kinetics of silica-water reactions. Geochem Cosmochem Acta 1980;44:1683–99.

[40] Xu T, Sonnenthal E, Spycher N, et al. TOUGHREACT user's guide: a simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media. Report LBNL-55460, Lawrence Berkeley National Laboratory, 2004.

[41] Helgeson HC, Kirkham DH, Flowers DC. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures: IV. Calculation of activity coefficients, osmotic coefficients, and apparent molal and standard and relative partial molal properties to 600 °C and 5 kb. Am J Sci 1981;281:1249–516.

[42] Warren JE, Root PJ. The behavior of naturally fractured reservoirs. Soc Petrol Eng J 1963;3:245–55.

[43] Barenblatt GI, Zheltov YP, Kochina IN. Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. J Appl Math Mech 1960;24: 1286–303.

[44] Pruess K, Narasimhan TN. On fluid reserves and the production of superheated steam from fractured, vapor-dominated geothermal reservoirs. J Geophys Res 1982;87(B11):9329–39.

[45] Pruess K, Narasimhan TN. A practical method for modeling fluid and heat flow in fractured porous media. Soc Petrol Eng J 1985;25(1):14–26.

[46] Bai M, Elsworth D, Roegiers J-C. Modeling of naturally fractured reservoirs using deformation dependent flow mechanism. Int J Rock Mech Min Sci 1993;30(7):1185–91.

[47] Biot MA. Theory of propagation of elastic waves in a fluid-saturated porous solid, II. Higher frequency range. J Acoust Soc Am 1956;28(2):179–91.

[48] Biot MA. General theory of three-dimensional consolidation. J Appl Phys 1941;12:155–64.

[49] Biot MA. Theory of propagation of elastic waves in a fluid-saturated porous solid, I. Low-frequency range. J Acoust Soc Am 1956;28(2):168–78.

[50] Biot MA. Mechanics of deformation and acoustic propagation in porous media. J Appl Phys 1962;33(4):1482–98.

[51] Berryman JG, Pride SR. Models for computing geomechanical constants of double-porosity materials from the constituents' properties. J Geophys Res 2002;107(B3):2052.

[52] Biot MA, Willis DG. The elastic coefficients of the theory of consolidation. J Appl Mech 1957;24:594–601.

[53] Zimmerman RW, Chen G, Hadgu T, Bodvarsson GS. A numerical dual-porosity model with semi-analytical treatment of fracture/matrix flow. Water Resour Res 1993;29:2127–37.

[54] Detournay E, Cheng AH-D. Fundamentals of poroelasticity. In: Hudson JA, editor. Comprehensive rock engineering. New York: Pergamon; 1993. p. 113–71.

[55] Mandel J. Consolidation des sols (etude mathematique). Geotechnique 1953;3:287–99.

[56] Abousleiman Y, Cheng AH-D, Cui L, et al. Mandel's problem revisited. Geotechnique 1996;46(2):187–95.

[57] Rice JR, Cleary MP. Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. Rev Geophys Space Phys 1976;14(2):227–40.

[58] Wang HF. Theory of linear poroelasticity. Princeton: Princeton University Press; 2000.

[59] Carslaw HS, Jaeger JC. Conduction of heat in solids. 2nd ed. Oxford: Oxford University Press; 1959.

[60] Geertsma J. Problems of rock mechanics in petroleum production engineering. In: Proceedings of the first international congress rock mechanics, vol. 1, Lisbon, 1966. p. 585–94.

[61] Min K-B, Rutqvist J, Elsworth D. Chemically and mechanically mediated influences on the transport and mechanical characteristics of rock fractures. Int J Rock Mech Min Sci 2009;46(1):80–9.

[62] Stephenson LP, Plumley WJ, Palciauskas VV. A model for sandstone compaction by grain interpenetration. J Sediment Petrol 1992;62:11–22.

[63] Yasuhara H, Elsworth D, Polak A. A mechanistic model for compaction of granular aggregates moderated by pressure solution. J Geophys Res 2003; 108(B11).

[64] Snow DT. Anisotropic permeability of fractured media. Water Resour Res 1969;5(6):1273–89.

[65] Witherspoon PA, Wang JSY, Iwai K, et al. Validity of cubic law for fluid flow in a deformable rock fracture. Water Resour Res 1980;16(6):1016–24.

[66] Hardin EL, Barton N, Lingle R, Board MP, Voegele MD. A heated flatjack test series to measure the thermomechanical and transport properties of in situ rock masses. Report ONWI-260, Columbus, Ohio: Office of Nuclear Waste Isolation; 1982.

[67] Dunn EJ. Reports on the Bendigo Goldfield. Special reports nos. 1 and 2, Department of Mines, Victoria, Australia, 1896. p. 16.

[68] Taber S. The mechanics of vein formation (with discussion). Am Inst Min Met Eng Trans 1920;61:3–41.

[69] Wiltschko DV, Morse JW. Crystallization pressure versus "crack seal" as the mechanism for banded veins. Geology 2001;29(1):79–82.

[70] Dewers T, Ortoleva P. Force of crystallization during the growth of siliceous concretions. Geology 1990;18:204–7.

[71] Maliva RG, Siever R. Diagenetic replacement controlled by force of crytallization. Geology 1988;16:688–91.

[72] Lichtner PC. Continuum formulation of multicomponent-multiphase reactive transport. In: Lichtner PC, Steefel CI, Oelkers EH, editors. Reactive transport in porous media. Washington, DC: Mineral Soc Amer; 1996.

[73] Bear J. Dynamics of fluids in porous media. New York: Elsevier; 1972.

7

Alternative Solution Methods

[7:1] Alternative Solution Models

Lagrangian-Eulerian Models

Lagrangian-Eulerian Methods

Lagrangian-Eulerian Approach

These methods use the idea of operator splitting to solve the diffusion component of flow using Eulerian methods, but treat the advective term using Lagrangian methods.

For example, the advection-dispersion equation is given as:

$$\frac{\partial M}{\partial t} + \mathbf{V} \cdot \nabla M = D \nabla^2 M, \tag{1}$$

where M denotes the mass of the solute, $\mathbf{V}$ is the velocity, and D is the diffusion coefficient.

Applying the Galerkin finite element method, Eq. (1) may be written in discrete form for the Lagrangian-Eulerian approach, as [Yeh, 1990],

$$\left[\int_R N_i N_i dR \right] \frac{DM_i}{Dt} + \left[\int_R (\nabla N_i) D (\nabla N_i) dR \right] M_i = \int_B D \cdot (\nabla M) \cdot n N_i dB \quad (i = 1, 2, \cdots, N), \tag{2}$$

Where the substantial derivative is defined as $\dfrac{D}{Dt} = \dfrac{\partial}{\partial t} + \mathbf{V} \cdot \nabla$, and where N_i is the shape function at the ith node, DM_i/Dt is the Lagrangian derivative of M_i with respect to time, R is the region of interest, B is the global boundary, and N is the total number of the nodes in the system. Integrating Eq. (2) using explicit time stepping, linear interpolation, and a fixed time step Δt, yields,

$$\left([W]/\Delta t + [K] \right) \{ M^{n+1} \} = \left([W]/\Delta t \right) \{ M^* \} + \{ B \}, \tag{3}$$

where

$$[W] = \sum \left(\int_R N_i N_i dR \right), \tag{4}$$

$$[K] = \sum \left(\int_R (\nabla N_i) D (\nabla N_i) dR \right), \tag{5}$$

$$\{ B \} = \sum \left(\int_B D \cdot (\nabla M) \cdot n N_i dB \right), \tag{6}$$

And where $\{M^{n+1}\}$ is the mass at the new time and $\{M^*\}$ is the Lagrangian mass. To obtain the Lagrangian mass at t^{n+1}, a forward-particle-tracked mass M_j^p, is first computed during time step Δt, given as,

$$M_j^p = M\left(\mathbf{x}_j^*, t^{n+1}\right) = M_j^n \quad \left(j = 1, 2, \cdots, N\right), \tag{7}$$

where

$$\mathbf{x}_j^* = \mathbf{x}_j^n + \mathbf{V}_j \Delta t \quad \left(j = 1, 2, \cdots, N\right), \tag{8}$$

in which $\mathbf{x}_j^*$ is the fictitious particle position at t^{n+1} when traveling from nodal location $\mathbf{x}_j^n$ at t^n. Subsequently, applying finite element interpolation with the shape functions, the Lagrangian mass M_i^* at each node is evaluated as (see **Figure 1**),

$$M_i^* = \sum_{j=1}^{N} M_j^p N_j\left(\mathbf{x}_i\right) \quad \left(i = 1, 2, \cdots, N\right). \tag{9}$$

Once the Lagrangian mass M_i^* is obtained, the final mass M_i^{n+1} at t^{n+1} is computed using Eq. (3).

Careful treatment is required for no-flow boundaries. Solutes are not allowed to cross no-flow boundaries, but for certain choices of large time steps, particles may be inadvertently ejected, as illustrated in **Figure 2**. This condition may be corrected by relocating the escaping particle back into the flow-field by using its closest projection to the boundary, as illustrated in **Figure 2**. This correction will thus tend to return the particle close to its true flow trajectory.
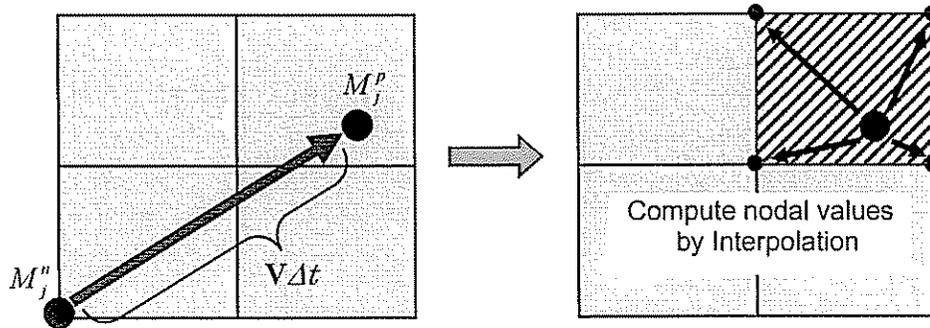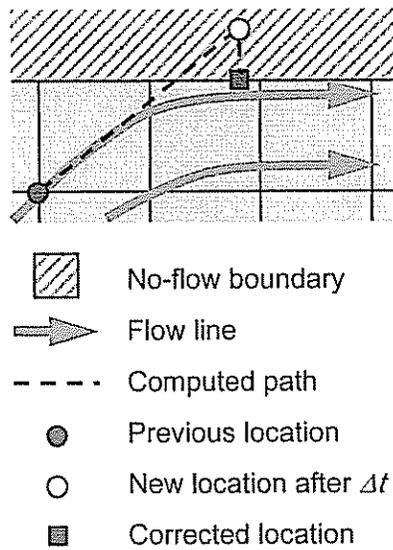
Figure 1. Trajectory of Lagrangian mass.

	No-flow boundary
	Flow line
	Computed path
	Previous location
	New location after Δt
	Corrected location

Figure 2. Schematic of correction applied to the computed particle location near no-flow boundary.

A numerical model simulating reactive transport and evolution of fracture permeability

Hideaki Yasuhara[1,*,†] and Derek Elsworth[2]

[1] *Department of Civil and Environmental Engineering, Ehime University, 3 Bunkyo-cho,
Matsuyama 790-8577, Japan*
[2] *Department of Energy and Geo-Environmental Engineering, Penn State University,
University Park, PA 16802, U.S.A.*

SUMMARY

A numerical model is presented to describe the evolution of fracture aperture (and related permeability) mediated by the competing chemical processes of pressure solution and free-face dissolution/precipitation; pressure (dis)solution and precipitation effect net-reduction in aperture and free-face dissolution effects net-increase. These processes are incorporated to examine coupled thermo-hydro-mechano-chemo responses during a flow-through experiment, and applied to reckon the effect of forced fluid injection within rock fractures at geothermal and petroleum sites. The model accommodates advection-dominant transport systems by employing the Lagrangian–Eulerian method. This enables changes in aperture and solute concentration within a fracture to be followed with time for arbitrary driving effective stresses, fluid and rock temperatures, and fluid flow rates. This allows a systematic evaluation of evolving linked mechanical and chemical processes. Changes in fracture aperture and solute concentration tracked within a well-constrained flow-through test completed on a natural fracture in novaculite (*Earth Planet. Sci. Lett.* 2006, in press) are compared with the distributed parameter model. These results show relatively good agreement, excepting an enigmatic abrupt reduction in fracture aperture in the early experimental period, suggesting that other mechanisms such as mechanical creep and clogging induced by unanticipated local precipitation need to be quantified and incorporated. The model is applied to examine the evolution in fracture permeability for different inlet conditions, including localized (rather than distributed) injection. Predictions show the evolution of preferential flow paths driven by dissolution, and also define the sense of permeability evolution at field scale. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: fracture permeability; Lagrangian–Eulerian method; dissolution

1. INTRODUCTION

Coupled thermal–hydraulic–mechanical–chemical (THMC) processes exert significant influence in controlling the evolution of the mechanical and transport properties for fractured rocks. The

---

*Correspondence to: Hideaki Yasuhara, Department of Civil and Environmental Engineering, Ehime University,
3 Bunkyo-cho, Matsuyama 790-8577, Japan.
†E-mail: hide@dpc.ehime-u.ac.jp

Copyright © 2006 John Wiley & Sons, Ltd.

competition between agents that reduce porosity (grain interpenetration, compaction, pressure solution, and precipitation) and those that generate porosity (dilation and free-face dissolution) control the rates, magnitudes, and sense of permeability modification, strength gain, and change in stiffness. In turn, these processes are important in defining the evolution of porosity and permeability in subsiding basins, in geothermal and petroleum reservoirs, and around repositories for the entombment of radioactive wastes, and in defining rates and magnitudes of strength gain that impact recurrence times and magnitudes of earthquakes.

To better understand the effects of temperature, stress, and fluid chemistry on the evolution of fracture permeability, only a limited number of experiments have been conducted under hydrothermal conditions, indicating the conflicting predictions on evolution in fracture permeability; sealing, gaping, or spontaneous switching between sealing and gaping is observed to result from net dissolution or precipitation within a fracture. Dissolution-driven sealing, likely resulting from dissolution beneath propping asperities in contact, is reported for natural and artificial fractures at elevated temperatures ($> 300°C$) in sandstone [1, 2], in granite [3], and in quartz [4], and at modest temperatures ($50–150°C$) in tuff [5] and in novaculite [6]. These are supplemented by results at both high confining stress (> 150 MPa) in granite [7] and at low stress (0.2 MPa) in marble where an acidic permeant is circulated [8]. Conversely, precipitation-driven sealing is observed in tuff at a range of temperatures [9]. Gaping is observed in hydrocarbon reservoir rocks [10, 11], and spontaneous or induced switching from sealing to gaping is reported at ambient temperatures ($20°C$) in limestone [12] and at modest temperatures ($20–120°C$) in novaculite [13]. These limited studies on fractures provide no conclusive view of the effects controlling the evolution of the transport properties, and the evolving rates and magnitudes of fracture permeability driven by interaction between the mechanical and chemical processes remain poorly constrained.

Modelling studies are an important supplement to experimental observations of the evolution in the transport properties of fractures under hydrothermal conditions. Such studies allow complexly interacting processes to be unraveled, to explain counter-intuitive results. These models must incorporate the interactions of reactive mass transport and mechanical effects, with these approaches complicated where flows are advection dominant—as they may be for flows in fractures. Difficulties result in accurately solving using numerical methods where flows are dominated by advective transport since numerical oscillation may result for Eulerian approaches where local Peclet numbers are large. To circumvent this problem, a variety of numerical treatments may be employed; the easiest involving a reduction in the spatial element (or grid) size. However, for very large velocities, this treatment is not always practical, due to the requisite large number of elements. For purely advective flows, Eulerian methods are intrinsically unstable, and erroneous oscillations may not be removed. Upstream-weighted finite element (FEM) and finite difference (FDM) methods may enable oscillations to be eliminated for high Peclet numbers, but these methods may generate artificial or numerical dispersion, resulting from their incapacity to preserve the sharpness of the front (or steep concentration gradients). An alternative to these flawed methods is the mixed Lagrangian–Eulerian approach [14–16] that overcomes many of the innate problems in high velocity flows. This method accommodates the advection term through a Lagrangian approach—the advective component is solved by tracking particles along characteristic pathlines, with all other terms in the solute-transport equation solved from an Eulerian viewpoint on a grid fixed in space. This method has the advantage that numerical oscillations and artificial dispersion are automatically

damped, and the procedure may continuously handle problems with mesh Courant numbers in excess of unity.

In this study, a Lagrangian–Eulerian algorithm is presented to follow the progress of evolution in permeability when a fracture is subjected to chemical dissolution by circulating hydrothermal fluids. Notably, our focuses are in examining the chemical processes that significantly influence the evolution in fracture permeability and in accommodating evolving advection-dominant transport problems. To demonstrate capability and validity of the model, predictions are compared with a companion flow-through experiment conducted on a stressed natural fracture in novaculite (> 99.5% quartz) [13].

2. MODEL DESCRIPTION

A numerical model is developed to describe the stress- and temperature-dependent evolution in aperture (permeability) within a single fracture mediated by chemical dissolution. This model accommodates solution for fluid flow and solute transport processes under advection-dominant conditions. The virtual fracture is constructed using the exact topography of two rough surfaces in contact, which have been previously profiled in 3-D [17]. From this prescribed initial aperture distribution within the fracture, and assuming steady conditions, the fluid velocity field is calculated from the Reynolds approximation. The local rate of dissolution/precipitation throughout the whole fracture domain is then determined, and the updated concentration distribution is obtained. Subsequently, the new aperture distribution resulting from chemical reaction is updated, and the final concentration distribution is obtained by solving the advection–diffusion equation. Each calculation process is explained in detail in the following.

2.1. FE mesh

The rectilinear two-dimensional mesh occupies the mean plane of the fracture and uses data from the measured topography of two rough surfaces in contact—the profile measured by 3-D roughness profiling (for details, see References [13, 17]). Each node in the fracture mesh has a local aperture datum that may be determined simply by point-by-point subtraction of the two digitized surfaces. However, careful positioning and orientation of the two surfaces is required before the subtraction since the profiles are initially unmated—the upper and lower rough surfaces are measured in an open-book format. To limit skewing of the aperture data, the mean planes of both surfaces are calculated and are made parallel to each other [18].

Figure 1 shows the parallel digitized rough surfaces of a natural fracture in novaculite. The differenced surface (i.e. the point-by-point subtraction of two surfaces) represents the distribution of the mechanical aperture, rather than the hydraulic aperture recovered from the flow-through experimental results. However, as a first-order estimation, the arithmetic mean aperture of the initial differenced surface is used for model prediction. This is calibrated by adjusting the separation between the two virtual parallel surfaces, and setting the initial hydraulic aperture to that obtained from the experiment [13]—mechanical and hydraulic apertures are assumed approximately equivalent [18].

Note that at contact points between the rough surfaces, a finite thickness water-film is assumed. This allows diffusive transport of mineral mass dissolved and then mobilized at these contacts by the elevated chemical potential beneath the stressed surface. Such a thin water-film
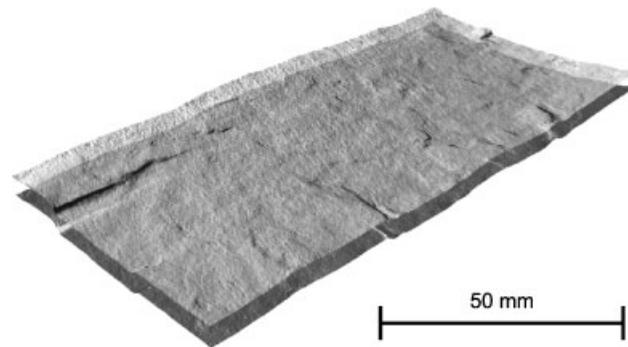
Figure 1. Oblique view of the parallel rough surfaces digitized by the 3-D laser profilometer system. The digitized surface measures $50 \times 89.5 \, \text{mm}^2$.

at the contact may be a function of applied stress and may range from less than 1 nm to a few hundreds nanometers [19, 20]. In this study, we presume a constant water-film thickness, ω, of 4.0 nm as this thickness remains ill-constrained.

2.2. Fluid flow distribution

The FE mesh of the fracture aperture distribution is utilized for fluid flow simulations. The flow simulation is conducted using the steady-state approximation of the Navier–Stokes equation for incompressible laminar flow (the Reynolds approximation) [21, 22] as

$$\nabla \cdot \left(\frac{b^3}{12\mu} \nabla p \right) = 0 \tag{1}$$

where b is the local aperture, μ is the fluid viscosity, and p is the fluid pressure driving flow. This returns a steady distribution of velocity, which is updated as the aperture distribution changes. Although the Reynolds equation is known to overestimate fluid velocity when fracture aperture is small relative to surface roughness [23], this error is small in relation to the other uncertainties within the analysis.

2.3. Pressure solution and free-face dissolution

Dissolution-dependent evolution of the fracture aperture is controlled by the competing influences of pressure (dis)solution and free-face dissolution. Fracture aperture (or related permeability) may decrease if pressure solution dominates, or may increase if free-face dissolution prevails. Pressure solution within a fracture incorporates three serial processes; dissolution at asperity contacts, diffusion along the interfacial water-film, and precipitation at the pore (fracture) wall, and may result in net reduction of fracture aperture. Conversely, if the mass rate of supply to the fluid occupying the fracture void is sufficiently low, or the flow-system sufficiently open, then the solute concentration in the pore fluid will be below the equilibrium concentration, net dissolution at free walls may dominate, and the fracture will widen. The competition between pressure solution and precipitation in the fracture void, that together contribute to a net reduction in permeability, and dissolution from the wall of the fracture void, that increases permeability, will prescribe the dominant effect; either net sealing or gaping.

Importantly, the dominant mechanism may change with stress and chemical condition of the solvent, or as a result of the evolution of fracture topography, and flow topology.

Here, stress- and temperature-dependent dissolution at contacting asperities and free-face dissolution/precipitation are systematically defined. First, dissolution at the asperity contacts provides a source of mass into the fracture cavity. Applying nonhydrostatic and nonequilibrium thermodynamics and then considering the chemical potential difference between the compressive site of contact and the less-stressed site of the pore wall, that is the motive force driving pressure solution, the source of mass injected into the fracture void space is most conveniently defined in terms of a dissolution mass flux, dM_{diss}^{PS}/dt, given as (for details, see Reference [24]),

$$\frac{dM_{diss}^{PS}}{dt} = \frac{3V_m^2(\sigma_a - \sigma_c)k_+\rho_g A_c}{RT} \tag{2}$$

where V_m is molar volume of the solid ($2.27 \times 10^{-5}\,\text{m}^3\,\text{mol}^{-1}$ for quartz), σ_a is the disjoining pressure [25] equal to the amount by which the pressure acting at a contact area exceeds the hydrostatic pore pressure, k_+ is the dissolution rate constant of the solid, ρ_g is the solid density ($2650\,\text{kg}\,\text{m}^{-3}$ for quartz), A_c is the size of the local contact area, R is the gas constant, and T is the temperature of the system. σ_c is the critical stress that defines stress state where the compaction of indenting asperity contacts will effectively halt. Where confining stress is applied to a rock fracture, asperity indentation will occur as a result of high localized contact stresses. Transient interpenetration may develop by plastic creep as the contact stress remains in excess of a critical stress, σ_c. Where stresses remain in excess of the critical interpenetration stress, dissolution will proceed in the water-film enveloping the interface, and mass will be removed by dissolution and transported by diffusion. This process will continue until the applied contact stress is sufficiently reduced by the growth of the contact area that compaction essentially ceases. The limiting stress may be defined by considering the energy balance under applied stress and temperature conditions, given by Revil [26] modified from Reference [27],

$$\sigma_c = \frac{E_m(1 - T/T_m)}{4V_m} \tag{3}$$

where E_m and T_m are the heat and temperature of fusion, respectively ($E_m = 8.57\,\text{kJ}\,\text{mol}^{-1}$, $T_m = 1883\,\text{K}$ for quartz).

Next, free-face dissolution and precipitation components are quantified as mass fluxes, dM_{diss}^{FF}/dt and dM_{prec}/dt, defined by the dissolution/precipitation rate constants and the difference between the fluid mass concentration in the pore space and the equilibrium concentration, defined as (modified from Reference [28]),

$$\frac{dM_{diss}^{FF}}{dt} = k_+ A_{pore}\rho_g V_m\left(1 - \left(\frac{C_{pore}}{C_{eq}}\right)^m\right)^n \tag{4}$$

$$\frac{dM_{prec}}{dt} = k_- A_{pore}\rho_g V_m\left(\left(\frac{C_{pore}}{C_{eq}}\right)^m - 1\right)^n \tag{5}$$

where A_{pore} is the area of the fracture void, k_- is the precipitation rate constant of the dissolved mineral, C_{pore} is the concentration in the pore space, and C_{eq} is the equilibrium solubility of the dissolved mineral. m and n are two positive numbers normally constrained by experiment; for quartz–water reaction, the reaction kinetics is likely first order [29], and in the model m and n are set to unity. Note that the free-face dissolution/precipitation mass fluxes will be zero as the mass

concentration in the pore fluid is either greater or smaller than the equilibrium solubility, respectively.

Dissolution/precipitation rate constant $k_{+/-}$, equilibrium solubility C_{eq}, and diffusion coefficient D of quartz have all Arrhenius-type dependence with temperature, given by

$$k_+ = k_+^0 \exp(-E_{k+}/RT) \tag{6}$$

$$k_- = k_-^0 \exp(-E_{k-}/RT) \tag{7}$$

$$C_{eq} = C_{eq}^0 \exp(-E_C/RT) \tag{8}$$

$$D = D_0 \exp(-E_D/RT) \tag{9}$$

Appropriate magnitudes are selected for these constants defining the temperature dependence as, $k_{+/-}^0 = 1.59/1.27 \, \text{mol m}^2 \, \text{s}^{-1}$ and $E_{k+/-} = 71.3/48.9 \, \text{kJ mol}^{-1}$ [30], $C_{eq}^0 = 274.9 \, \text{kg m}^{-3}$ and $E_C = 26.4 \, \text{kJ mol}^{-1}$ [31], and $D_0 = 5.2 \times 10^{-8} \, \text{m}^2 \, \text{s}^{-1}$ and $E_D = 13.5 \, \text{kJ mol}^{-1}$ [26].

2.4. Lagrangian–Eulerian approach

The solute transport in a fracture is modelled by the mixed Lagrangian–Eulerian approach. An advection–diffusion equation is given as

$$\frac{\partial M}{\partial t} + \mathbf{V} \cdot \nabla M = D \nabla^2 M \tag{10}$$

where M denotes the mass of the solute, $\mathbf{V}$ is the velocity, and D is the diffusion coefficient. The diffusion coefficient of the solute may be different between contact and void nodes. The diffusivity inside contacts may be a few orders of magnitudes smaller than that in the bulk pore fluid due to electro-viscous effects [32, 33], although others [34] justify that this has minor influence. Correspondingly, we use the same value of the diffusion coefficient for both contact and void points.

Applying the Galerkin FEM Equation (10) may be written in discrete form for the Lagrangian–Eulerian approach, as [15]

$$\left[\int_R N_i N_i \, dR\right]\frac{DM_i}{Dt} + \left[\int_R (\nabla N_i)D(\nabla N_i) \, dR\right]M_i = \int_B D \cdot (\nabla M) \cdot n N_i \, dB \quad (i = 1, 2, \ldots, N) \tag{11}$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla \tag{12}$$

in which N_i is the shape function at the ith node, DM_i/Dt is the Lagrangian derivative of M_i with respect to time, R is the region of interest, B is the global boundary, and N is the total number of the nodes in the system. Integrating Equation (11) using explicit time stepping, linear interpolation, and a fixed time step Δt, yields,

$$([W]/\Delta t + [K])\{M^{n+1}\} = ([W]/\Delta t)\{M^*\} + \{B\} \tag{13}$$

where

$$[W] = \sum \left(\int_R N_i N_i \, dR\right) \tag{14}$$

$$[K] = \sum \left(\int_R (\nabla N_i) D (\nabla N_i) \, dR \right) \tag{15}$$

$$\{B\} = \sum \left(\int_B D \cdot (\nabla M) \cdot n N_i \, dB \right) \tag{16}$$

in which $\{M^{n+1}\}$ is the mass at the new time and $\{M^*\}$ is the Lagrangian mass. To obtain the Lagrangian mass at t^{n+1}, a forward-particle-tracked mass M_j^p, is first computed during time step Δt, given as

$$M_j^p = M(\mathbf{x}_j^*, t^{n+1}) = M_j^n \quad (j = 1, 2, \ldots, N) \tag{17}$$

where

$$\mathbf{x}_j^* = \mathbf{x}_j^n + \mathbf{V}_j \Delta t \quad (j = 1, 2, \ldots, N) \tag{18}$$

in which $\mathbf{x}_j^*$ is the fictitious particle position at t^{n+1} when travelling from nodal location $\mathbf{x}_j^n$ at t^n. Subsequently, applying FE interpolation with the shape functions, the Lagrangian mass M_i^* at each node is evaluated as (see Figure 2),

$$M_i^* = \sum_{j=1}^{N} M_j^p N_j(\mathbf{x}_i) \quad (i = 1, 2, \ldots, N) \tag{19}$$

Once the Lagrangian mass M_i^* is obtained, the final mass M_i^{n+1} at t^{n+1} is computed using Equation (13).

Careful treatment is required for no-flow boundaries. Solutes are not allowed to cross no-flow boundaries, but for certain choices of large time steps, particles may be inadvertently ejected, as illustrated in Figure 3. This condition may be corrected by relocating the escaping particle back into the flow-field by using its closest projection to the boundary, as illustrated in Figure 3. This correction will thus tend to return the particle close to its true flow trajectory.

2.5. Overall computational procedure

With the fracture topography digitized, the flow simulation defines the initial flow velocity field. Mineral mass is either injected-into, or removed-from, the flow-field, depending on the relative dominance of processes of pressure solution and free-face dissolution/precipitation. These components are then transported within the fluid phase, until conditions dictate their removal to
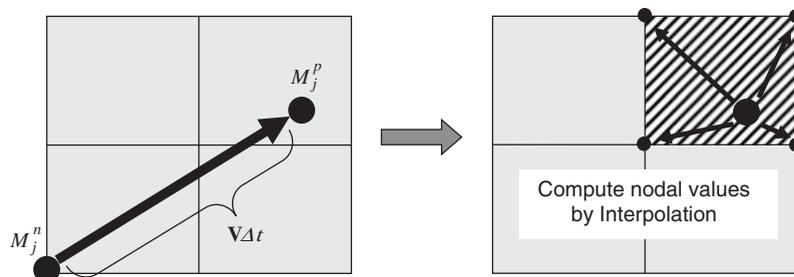

Figure 2. Schematic illustration of computing the Lagrangian mass at the nodal locations. During time step Δt, the nodal mass M_j^n travels to M_j^p, and the new nodal values are interpolated using shape functions.
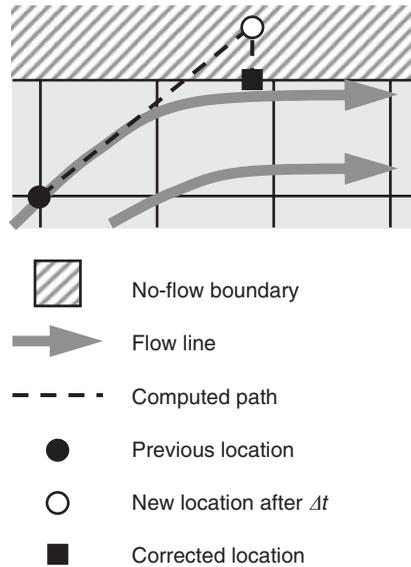
Figure 3. Schematic of correction applied to the computed particle location near no-flow boundary.

the fracture walls. Importantly, prior characterizations of an interface region as a separate diffusive domain [35], are unnecessary, as mass diffusion within the water-film separating contacts is automatically accommodated by the macro-scale FE mesh. Note that the flow simulation, the chemical processes, and the solute transport equation are solved sequentially, rather than simultaneously. The main points of the computational procedure are as follows.

First, the initial fracture topography is set to generate the FE mesh—each node has a local aperture datum. Second, the initial and boundary conditions (i.e. temperatures, stresses, flow rates, and flow or no-flow boundaries) are applied, and flow simulation (Equation (1)) is conducted using the aperture mesh to obtain the distribution of flow velocities within the fracture. The simulation retrieves elemental velocities at Gauss points, and nodal velocities are interpolated using shape functions to accommodate solving the solute transport equation (i.e. obtaining the fictitious particle position using nodal velocity as shown in Equation (18)).

Third, the dissolution/precipitation processes at contact points and void wall are evaluated using Equations (2)–(5) at every single node. In Equation (2), the stress acting at contact points σ_a is simply defined by,

$$\sigma_a = \sigma_{\mathrm{eff}}\frac{n_{\mathrm{c}}}{n_t} \tag{20}$$

where σ_{eff} is the confining effective stress prescribed. n_{c} and n_t is the numbers of contact points and total nodes, respectively. This assumes that the contacting stresses are equivalent at all contact areas distributed within the fracture. The effective area for pressure dissolution (i.e. A_{c}) is assumed equal to the area of a single element, and that for free-face dissolution/precipitation (i.e. A_{pore}) is assumed equal to twice the elemental area (both upper and lower void walls
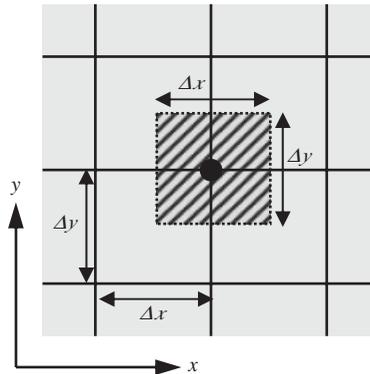
Figure 4. Representation of contact or void area (shaded area) at each node for chemical calculations. The effective area for pressure dissolution and free-face dissolution/precipitation is equivalent to either the area of a single element, or twice that area, respectively.

contribute) (Figure 4). Consequently, the nodal aperture b_i at t^{n+1} is calculated as

$$b_i^{n+1} = b_i^n - \frac{dM_{\text{diss}}^{\text{PS}}}{dt} \frac{\Delta t}{\rho_g A_e} + \left(\frac{dM_{\text{diss}}^{\text{FF}}}{dt} \bigg|_i - \frac{dM_{\text{prec}}}{dt} \bigg|_i \right) \frac{\Delta t}{\rho_g A_e} \quad (i = 1, 2, \ldots, N) \tag{21}$$

where A_e is the area of an element. If b_i^{n+1} becomes smaller than the water-film thickness of 4 nm, it is then indexed as a contacting node and $b_i^{n+1} = 4.0$ nm. Simultaneously, the nodal mass dissolved is also updated as

$$M_i^{c,n+1} = M_i^n - \frac{dM_{\text{diss}}^{\text{PS}}}{dt} \Delta t \quad (\text{if } b_i^{n+1} = 4.0 \text{ nm}) \tag{22}$$

$$M_i^{c,n+1} = M_i^n - \left(\frac{dM_{\text{diss}}^{\text{FF}}}{dt} \bigg|_i - \frac{dM_{\text{prec}}}{dt} \bigg|_i \right) \Delta t \quad (\text{if } b_i^{n+1} > 4.0 \text{ nm}) \tag{23}$$

where $M_i^{c,n+1}(i = 1, 2, \ldots, N)$ is the updated mass at each node after the incremented time, but is not the final one at t^{n+1} since it is subsequently modified by solute transport.

Finally, the contribution of solute transport (Equation (10)) is computed using the Lagrangian–Eulerian approach. The final mass at t^{n+1} is calculated as schematically shown in Figure 5. Then, the concentration at each node is evaluated using the updated mass at t^{n+1}, given as

$$C_i^{n+1} = \frac{M_i^{n+1}}{b_i^{n+1} A_e} \quad (i = 1, 2, \ldots, N) \tag{24}$$

The updated concentrations are used to calculate free-face dissolution/precipitation (Equations (4) and (5)) at the next time step. The lumped concentration travelling out of the domain C_{out}^{n+1}, which is directly analogous to mineral efflux measurements made during the experiments, is also calculated by

$$C_{\text{out}}^{n+1} = \frac{\sum M_{\text{out}}^{n+1}}{Q \Delta t} \tag{25}$$

where $\sum M_{\text{out}}^{n+1}$ is the summation of the solute mass exiting the outflow boundary during one time increment, and Q is the flow rate.

In summary, the combined algorithm incorporates an initial evaluation of aperture distribution from profile data, fluid transport simulation, and subsequent evaluation of mineral mass transport and redistribution, as outlined in Figure 6. This procedure allows the evolution
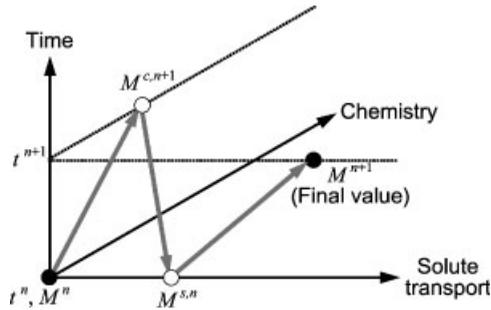

Figure 5. Schematic of calculation sequence during incremental time to obtain nodal mass, involving chemical processes and solute transport. Note that $M^{c,n+1} = M^{s,n}$.
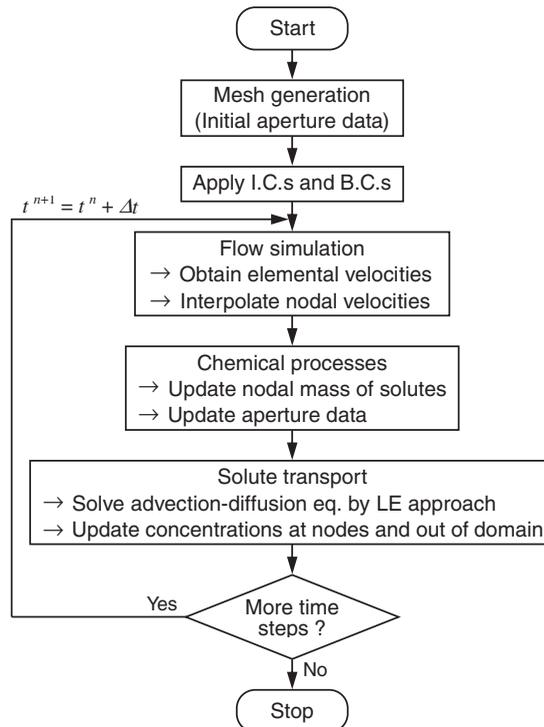
Figure 6. Flow chart for the overall computational procedure.

of fracture aperture to be followed where stress, temperature, and fluid flow conditions mediate behaviour.

3. COMPARISON WITH EXPERIMENTAL MEASUREMENTS

This numerical model is applied to describe the time-dependent evolution in aperture obtained from a companion experiment in a stressed natural fracture of novaculite [13]. The flow-through experiment is conducted on a natural fracture of Arkannsas novaculite, which has a uniform grain size in the range 1–6 μm and high quartz content of > 99.5% [36], at a constant effective stress of 1.38 MPa (200 psi) and at elevated temperatures in the range 20–120°C. Distilled water is used as a permeant and thus, the chemical system is relatively simple (i.e. $SiO_2 + 2H_2O \leftrightarrow Si(OH)_4$). The experimental conditions during the entire length of experiments (3150 h) are listed in Table I.

Prior to applying the current model, the lumped parameter model previously developed [35, 37] is first adopted to predict the progress of mean aperture closure and evolution of Si efflux, for the same experiment. Then, the current model is applied to quantify the experimental observations, and predictions given by the two different models are compared and examined, relative to the experimental measurements.

3.1. Lumped parameter model comparison

Lumped parameter models [35, 37] are capable of approximately representing the principal chemical processes of pressure solution at mineral contacts, solute diffusion along these contacts, and precipitation on the void wall of a fracture at a single representative contact. These solutions may also represent free-face dissolution, together with changes in fracture aperture and mineral mass concentration in the effluent fluid that result. These solutions are approximate in that they require a single representative contact to be defined—all processes at the contacting walls, and in the void, are averages of the entire contact area and void volume, respectively. Importantly, characteristic differences of the lumped parameter model from the numerical model developed in Section 2 are that fracture topography is simplified by a representative contact surrounded by an appropriate tributary area (see Figure 7) and that a

Table I. Experimental conditions [37].

Time (h)	Temperature (°C)	Flow rate (mL min^{-1})	Flow direction
0–121	20	1.0	Original
121–380	20	0.5	Original
380–858	20	0.25	Original
858–930	20	0.0	—
930–1266	20	0.25	Original
1266–1292	20	0.125	Original
1292–1494	20	0.125	Reversed
1494–1869	20	0.0625	Original
1869–2255	40	0.0625	Original
2255–2875	80	0.0625	Original
2875–3150	120	0.125	Original

simple, but physically plausible, relation between fracture aperture and fracture contact-area ratio is defined to represent the irreversible alteration in fracture geometry caused by pressure solution and free-face dissolution. Correspondingly, processes are innately averaged, and no account is made for the spatial structure. Such models typically make adequate predictions of homogeneously distributed behaviours, but not of localized effects, such as the evolution of a through-going dissolution conduit (wormhole) [12, 37].

The digitized fracture obtained through the profiling data constrains the relation between the fracture aperture and the contact-area ratio as shown in Figure 8. This relation is approximated by the regression curve, given as [13]

$$\langle b \rangle = 2.5 + 16.0 \exp(-(R_c - R_{c0})/20.0) \tag{26}$$

where $\langle b \rangle$ is the mean mechanical aperture, and R_c is the contact-area ratio. The initial aperture is set $18.5\,\mu m$ because the hydraulic aperture evaluated from the companion flow-through experiment [13] started initially with this value.
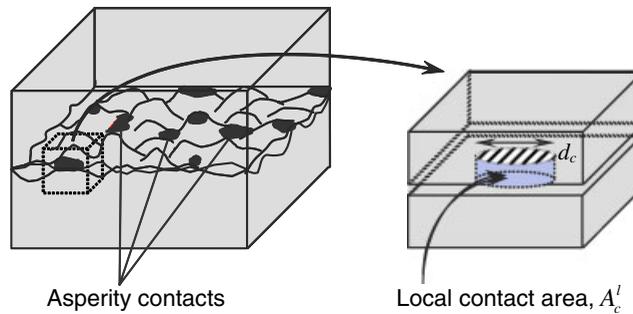

Figure 7. Idealized representation of asperity contact condition for lumped parameter model. A representative contact area A_c^l (right) represents the assumed average area of each contact (left), and is considered circular in shape of diameter d_c.
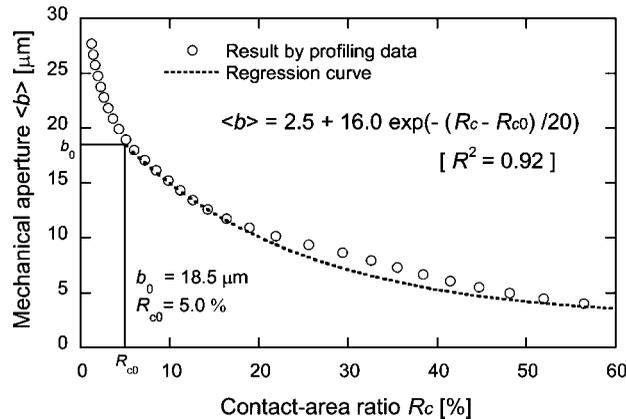

Figure 8. Relation between mean aperture and contact-area ratio. Circles are evaluated from point-by-point subtraction using the profiling data, and the dotted line is the regression curve of $\langle b \rangle = 2.5 + 16.0 \exp(-(R_c - R_{c0})/20)$ with the correlation coefficient, $R^2 = 0.92$.

A detailed description for the calculation procedure is reviewed in References [35, 37] and is also summarized in Appendix A. Parameters utilized in the predictions are listed in Table II. Predicted changes in fracture aperture and Si concentration are shown in Figure 9 together with the data measured through the experiments. Note that we omit the predictions during the reversed flow experiment (stage II) due to the unanticipated sharp reductions in aperture resulting from changes in fluid pressure distribution within the fracture by the switching of flow direction, which is not able to be predicted by the model. Thus, the unaccountable reduction is followed by resetting the aperture according to that recorded in the experiment and the contact-area ratio is updated using Equation (26). To closely match the evolution in fracture aperture in the experiment, the significant parameters of reaction rate constants $k_{+/-}$ for pressure dissolution (Equation (2)) and free-face dissolution (Equations (4) and (5)) are separately increased by factors shown in Table III. Also, the critical stress σ_c defined by Equation (3) is reduced by a factor of one-tenth to follow the large aperture reduction (~ 18.5 to $\sim 10.0\,\mu m$) in the early experiment (0 to ~ 800 h); if the unmodified σ_c is used for the predictions, such a large decrease in aperture is not predicted because stress acting on contacts (i.e. σ_a) becomes equal to σ_c and then no further compaction proceeds (see Equation (2)). This indicates that critical stress σ_c may be smaller than that defined by Equation (3) and more data are needed to quantify this process. However, in this work we merely select a value of one-tenth σ_c in an attempt to replicate the experimental results.

As shown in Figure 9, the predictions of fracture aperture and Si concentrations using the augmented $k_{+/-}$ are in good agreement with the actual data although the applied multipliers are relatively large. Note that precipitation, which may reduce fracture aperture, exerts little influence on the change in aperture—solute concentration is much lower than equilibrium solubility as a result of the dominant effect of strongly advective transport and short residence time in the relatively short core (~ 10 cm). The multipliers applied to follow the experimental measurements are large, specifically for those in stages I–IV, implicating that other mechanisms may dominate over pressure solution and/or free-face dissolution, or the model may be incapable of representing the overall processes since a detailed topology for a fracture is not involved. This concern is further examined in the following section by accommodating a spatial distribution of contacts and apertures, using the FEM developed in this work.

3.2. Distributed parameter (numerical) model comparison

The numerical model developed in Section 2 is applied here to follow experimental observations of changes in fracture aperture and Si concentration. The latter are measured directly, and the former are inferred from measurements of flow rate and differential fluid pressure. The flow is along the long dimension of the image (Figure 1), from left to right, with no-flow boundaries applied along the two long sides, parallel to the flow direction of the fracture. Reasonable computational limits are placed on both memory and runtime—calculations are conducted using a constant element size of $2 \times 2\,mm^2$. Note that even for this fine grid, local element Peclet numbers are of the order of 10^5 and result in a fully advection-dominant system.

Flow rates and parameters utilized in the predictions are summarized in Tables I and II. Predicted rates of aperture evolution and Si concentration history are matched with the actual measurements, as shown in Figure 10. Required modifications in parameters, necessary to replicate the experiments are listed in Table IV. Both the evolutions in fracture aperture and in Si concentration are in fairly good agreement with those observed. The predictions of Si

H. YASUHARA AND D. ELSWORTH

Table II. Parameters used to describe the evolution of fracture aperture.

Effective stress, σ_{eff} (MPa)	Diffusion path width, ω (nm)	Temperature, T (°C)	Dissolution rate constant, k_+ (mol m^{-2} s^{-1})	Precipitation rate constant, k_- (mol m^{-2} s^{-1})	Equilibrium solubility, C_{eq} (kg m^{-3})	Diffusion coefficient, D (m^2 s^{-1})	Critical stress, σ_{c} (MPa)	Dynamic viscosity, μ (Pa s)
1.38	4.0	20	3.14×10^{-13}	2.46×10^{-9}	6.12×10^{-3}	2.04×10^{-10}	79.7	1.00×10^{-3}
		40	2.03×10^{-12}	8.85×10^{-9}	1.09×10^{-2}	2.91×10^{-10}	78.7	6.53×10^{-4}
		80	4.52×10^{-11}	7.43×10^{-8}	3.25×10^{-2}	5.24×10^{-10}	76.7	3.54×10^{-4}
		120	5.35×10^{-10}	4.04×10^{-7}	8.38×10^{-2}	8.36×10^{-10}	74.7	2.32×10^{-4}
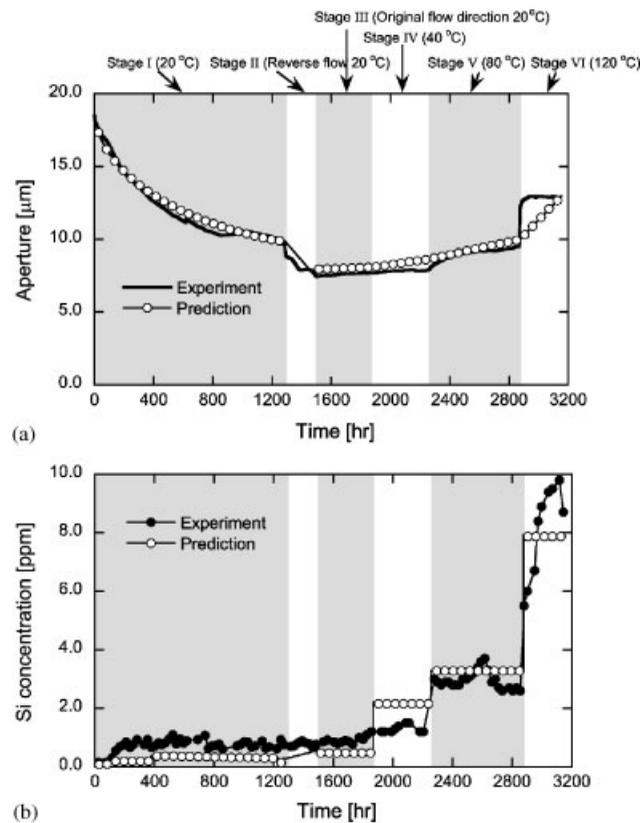
Figure 9. Comparisons of changes in: (a) aperture; and (b) Si concentration between the experimental results [13] and the predictions by the lumped parameter model. Open circles represent the predictions using modified values of reaction rate constants $k_{+/-}$ shown in Table III.

Table III. Experimental conditions and modification of parameters used in the analysis by the lumped parameter model.

	Test stages					
Parameters	I	II	III	IV	V	VI
Temperature (°C)	20	20	20	40	80	120
Flow direction	Original	Reversed	Original	Original	Original	Original
σ_c (Equations (2) and (3))	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$
k_+ (Equation (2))	$k_+ \times 10^6$	—	$k_+ \times 10^4$	$k_+ \times 10^4$	$k_+ \times 500$	$k_+ \times 200$
$k_{+/-}$ (Equations (4) and (5))	$K_{+/-} \times 10^4$	—	$k_{+/-} \times 10^4$	$k_{+/-} \times 10^4$	$k_{+/-} \times 500$	$k_{+/-} \times 200$

concentrations, especially those during stages III and IV underestimate the real data; the predictions in stages III and IV are ~ 0.1 and ~ 0.2 ppm, relative to the measurements of ~ 0.9 and ~ 1.3 ppm, respectively. A systematic improvement in predictions between the lumped
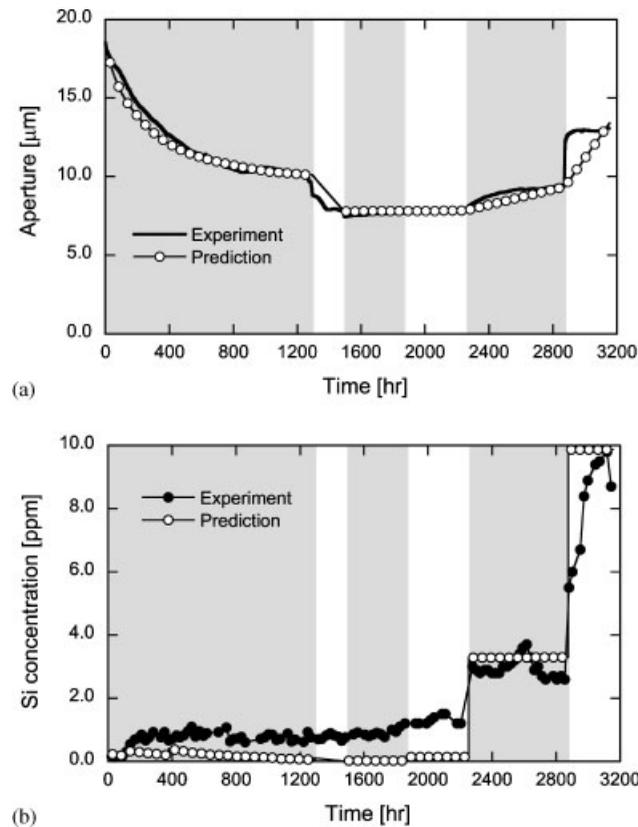
Figure 10. Comparisons of changes in: (a) aperture; and (b) Si concentration between the experimental results [13] and the predictions by the distributed parameter model developed in this work. Open circles represent the predictions using modified values of reaction rate constants $k_{+/-}$ shown in Table IV.

Table IV. Experimental conditions and modification of parameters used in the analysis by the distributed parameter (current) model.

Parameters	Test stages					
	I	II	III	IV	V	VI
Temperature (°)	20	20	20	40	80	120
Flow direction	Original	Reversed	Original	Original	Original	Original
σ_c (Equations (2) and (3))	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$	$\sigma_c \times 0.1$
k_+ (Equation (2))	$k_+ \times 5.0 \times 10^6$	—	$k_+ \times 30$	$k_+ \times 30$	$k_+ \times 30$	$k_+ \times 30$
$k_{+/-}$ (Equations (4) and (5))	$k_{+/-} \times 30$	—	$k_{+/-} \times 30$	$k_{+/-} \times 30$	$k_{+/-} \times 30$	$k_{+/-} \times 15$

parameter model and the numerical model is apparent in the applied multipliers for reaction rate constants to replicate the experiments (see Tables III and IV); the modifiers are much smaller for the numerical model—the small and constant magnitude multiplier of 30 is applied

throughout the experimental period except for pressure dissolution in stage I (i.e. 5.0×10^6) and for free-face dissolution/precipitation in stage VI (i.e. 15). However, the large multipliers required to replicate pressure dissolution during stage I (i.e. 10^6 for the lumped model and 5.0×10^6 for the numerical model) including the relatively abrupt and large aperture reduction, remain enigmatic. This implicates other mechanisms, such as mechanical creep and clogging resulting from locally high and unanticipated precipitation rates, of which neither are accommodated in the current description.

An important component of the model is the ability to follow the evolution in local aperture with time. Comparison between fracture apertures measured at the close of the experiment (3150 h) by X-ray CT [13], and those independently predicted by the model are shown in Figure 11. The white shaded area in the CT image represents apertures greater than the CT resolution threshold of 60 μm. The scanning resolution for the X-ray CT is insufficient for a rigorous quantitative comparison between the CT image and the prediction. However, the model prediction is in qualitatively good agreement at several regions with large aperture (or void), with the CT image.

Flow patterns within the fracture are predicted with time and are shown at the beginning (0 h) and end (3150 h) of the experimental period (Figure 12). Flow velocities within the fracture at 0 h are entirely faster than those at 3150 h because of the larger flow rate prescribed (i.e. 1.0 vs 0.125 mL min^{-1}, see Table I). As apparent in Figure 12, flow at both times is tortuous due to the effects of surface roughness and contact area, and in particular the flow at the end is randomly
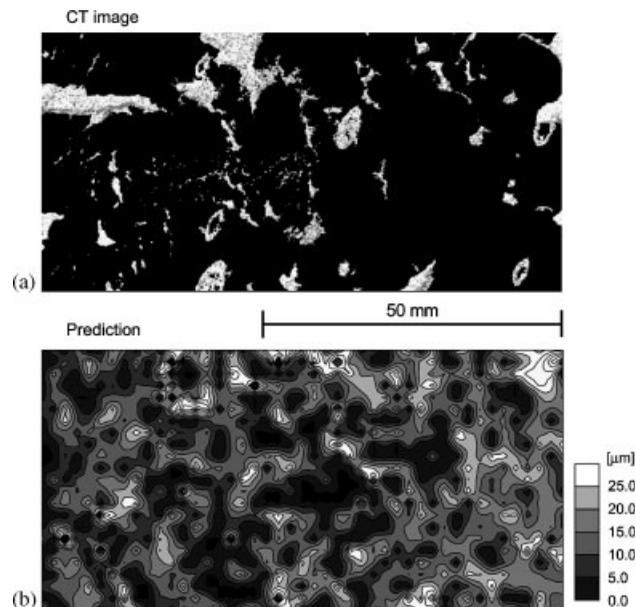

Figure 11. Qualitative comparison in aperture distribution between: (a) the X-ray CT image post-experiment (after Reference [13]); and (b) the predicted response at the end of simulation. (a) White coloured area represents aperture greater than the threshold of 60 μm and the black area shows aperture smaller than the threshold or contact area, while (b) white area is aperture greater than 25 μm, with contact area shown in black.
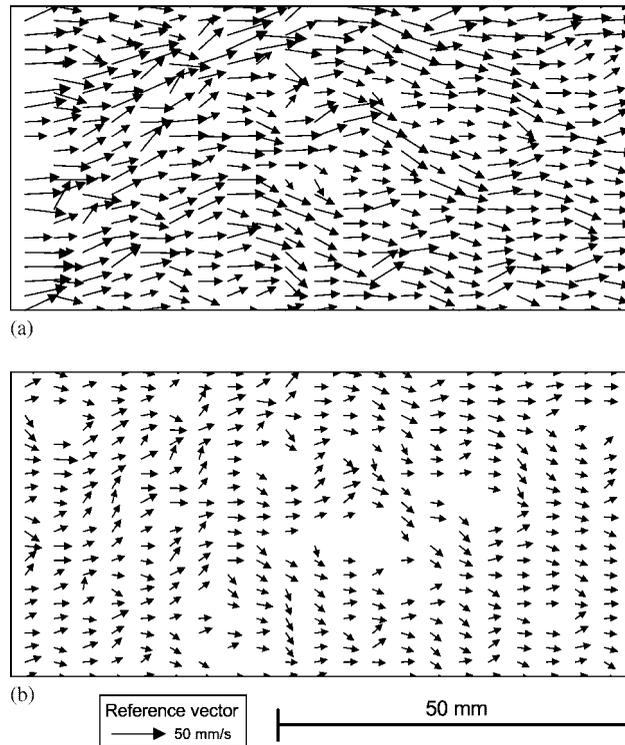
(a)

(b)

| Reference vector |
| ---→ 50 mm/s |

50 mm

Figure 12. Flow field resulting from the FE solution of the Reynolds equation at: (a) 0 h; and (b) 3150 h. Vectors represent the relative magnitude and direction of local flow. Note that the flow is distributed randomly with flow-excluded zones growing with increased time, and related contact area.

distributed throughout the fracture without clear channelling although circumventing the regions of sufficiently small aperture and/or contact. Correspondingly, no preferential flow paths are generated during net-dissolution (or erosion) processes, which is congruent with the experimental measurements constrained by X-ray CT images and later Wood's metal injection [13]. This contrasts with other experiments where the evolution of flow channels formed through net dissolution [8, 12] are evident.

4. IMPLICATIONS OF EVOLUTION IN PERMEABILITY

Both the companion flow-through experiment [13] and the model predictions confirm that no preferential flow paths evolve within the fracture (i.e. pressure solution and free-face dissolution). This is likely due to the prescribed boundary conditions; the flow is injected throughout the fracture inlet with relatively high flow rates. In contrast, at the anticipated larger *in situ* scales of geothermal and petroleum reservoirs dissolution to enhance fracture permeability may not occur in the broad area throughout fractures of interest, but proceed

within the limited regions of fractures with preferential flow paths because flow is spatially restricted through limited numbers of injection and recovery wells.

To examine the effect of narrowed fluid injection relative to fracture length, which simply simulates fluid injection in geothermal and petroleum reservoirs, a simple numerical experiment is conducted. Injection into the same fracture is applied at a point ($Q = 1.0 \, \text{mL} \, \text{min}^{-1}$ at the central node on the inlet boundary). The applied temperature is 120°C and the corresponding parameters for the prediction are listed in Table II. Predicted flow patterns within the fracture after 100 h, overlaying the distribution of fracture apertures are shown in Figure 13(a), together with the difference in apertures between 0 and 100 h, shown in Figure 13(b). Employing no modifications in reaction rate constants for this prediction, free-face dissolution dominates over the effect of pressure dissolution, resulting in the mean aperture consistently increasing at a rate of gaping of $3.7 \times 10^{-13} \, \text{m} \, \text{s}^{-1}$ throughout the prediction. As apparent in Figure 13 several flow paths are generated during the 100 h virtual experiment, with net dissolution and erosion concentrated within the upper half of the plan-view of the fracture, generating a broad flow channel. This is likely due to the combined effect of the narrowed flow injection port and is sensitive to the initial conditions of the local aperture distribution (or roughness). Notably, the restricted flow presents a positive feedback that favours the development of localized flow conduits (worm-hole-like flow channels). Clearly, this effect is influenced by geometric factors relating to the scale of the sample—larger samples may develop multiple distributed flow channels with the ephemeral dominance of these channels switching with the progress of the transport network.
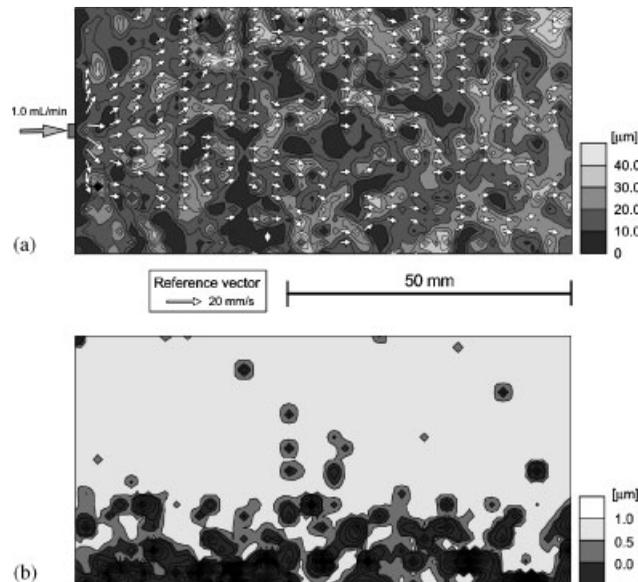

Figure 13. Results of numerical flow-through simulation at 100 h after flow started: (a) overlay of flow field on aperture contour. Note that several flow paths are formed; and (b) contour of aperture difference between 0 and 100 h. Lighter shading represent dissolution (erosion) regions.

5. CONCLUSIONS

A numerical model is developed to represent the evolution in fracture aperture mediated by the significant processes of pressure solution, and incorporating the serial processes of dissolution, diffusion, precipitation, and free-face dissolution. The model defines an initial distribution of fracture apertures, and supplements this with a Reynolds equation solution for the evolving flow-field—this flow-field is used to calculate the influence of chemical processes of pressure solution and free-face dissolution/precipitation in sequentially modifying the initial aperture distribution. Significantly, where advective flows dominate, as is anticipated to be the norm, the transport equation is solved via Lagrangian–Eulerian algorithm. The model is capable of predicting the evolution in aperture and solute concentration for a single fracture under arbitrary prescribed stress, temperature, and flow rate conditions. Predictions of both lumped parameter [35, 37] and distributed parameter representation of a single well-constrained experiment [13] are examined to test the adequacy of each model in representing experimentally observed behaviour. Notably, if the controlling parameters of reaction rate constants are increased by a factor of 30, the numerical model show excellent agreement with the experimental observations although a sharp reduction of the fracture aperture during the early experiment period is unable to be followed. This mismatch is likely attributed to processes of mechanical creep, that are not represented in the model. Both observations by X-ray CT, and model predictions, conclude that fluid flow at the conclusion of the experiment (3150 h) is broadly distributed throughout the fracture—no preferential flow paths are generated.

The model is applied to examine the evolution of fracture aperture (or permeability) where injection is concentrated at a point, as an approximation of injection into a reservoir. The predictions show the propensity to develop channelized dissolution features that concentrate flow. This exercise portends the potential to determine the form, plausible rates, magnitudes, and senses of permeability enhancement at field scale.

APPENDIX A: LUMPED PARAMETER MODEL

The evolution of fracture aperture is controlled by the competing influences of pressure solution, which incorporates interfacial dissolution, diffusion, and precipitation, and free-face dissolution. Interfacial dissolution at contacting asperities (Equation (2)) and free-face dissolution/precipitation (Equations (4) and (5)) are defined in the main text, and interfacial diffusion is defined herein in terms of the diffusive mass flux, dM_{diff}/dt, as [24]

$$\frac{dM_{diff}}{dt} = \frac{2\pi\omega D}{\ln(d_c/2a)}(C_{int} - C_{pore}) \tag{A1}$$

where ω is the thickness of the water-film trapped at the interface, D is the diffusion coefficient, and $(C_{int})_{x=a}$ and $(C_{pore})_{x=d_c/2}$ are mineral concentrations in the interface fluid and pore space, respectively.

A single fracture is idealized as two rough surfaces held apart by bridging asperities, as illustrated in Figure 7 (left). The average contact-area ratio, R_c, may be determined by defining a representative contact area, A_c^l, surrounded by an appropriate tributary area, A_t^l, (Figure 7,

right), and is assumed equivalent to the ratio of the summed local contact areas, A_c^t, to the total fracture area, A_t^t, given as [35]

$$R_c = \frac{A_c^l}{A_t^l} = \frac{A_c^t}{A_t^t} \tag{A2}$$

Within this tributary area, the contact diameter, d_c, of the local contact area, A_c^l, is defined as

$$d_c = \sqrt{\frac{4A_c^l}{\pi}} \tag{A3}$$

For uniaxial compaction, the normal forces acting on the tributary area and the contacting asperity balance, yielding the stress applied at the contact area, σ_a, as

$$\sigma_{\text{eff}} \cdot A_t^l = \sigma_a \cdot A_c^l \Rightarrow \sigma_a = \frac{\sigma_{\text{eff}}}{R_c} \tag{A4}$$

where σ_{eff} is the average macroscopic effective stress.

Interactive processes of pressure solution and free-face dissolution irreversibly alter the geometry of the fracture surfaces, and the relation between fracture aperture and contact area may be defined to follow this modification of the fracture aperture and contact-area ratio within the tributary domain. A simple, but physically viable, relation between them is defined as [35]

$$\langle b \rangle = a_1 + a_2 \exp(-(R_c - R_{c0})/a_3) \tag{A5}$$

where $\langle b \rangle$ is the mean aperture, R_c is the contact-area ratio, and a_i ($i = 1, 2, 3$) is a constant. This curve is adopted as a straightforward and representative relation between fracture contact area and aperture, to define the phenomenology of fracture sealing/gaping by pressure solution/ free-face dissolution.

A.1. Computational procedure

The individual processes of dissolution at asperity contacts, diffusion along interfacial water-film, and free-face dissolution/precipitation are combined to define the progress of aperture reduction of the fracture with time. In the initial condition, a small representative contact area is set with the initial aperture of the fracture. An effective stress is applied, as amplified by the tributary geometry, and during time step Δt, appropriate magnitudes of mass dissolution at the representative contact area, diffusion, and free-face dissolution/precipitation are simultaneously evaluated from Equations (2), (A1), (4), and (5), respectively. Physically, the dissolved mass evaluated from Equation (2) is supplied to the interface, and domain shortening (i.e. aperture reduction) proceeds as this mass passes along the interface by diffusion, as defined by Equation (A1). From the known magnitude of the diffusing mass, the updated contact area and aperture are calculated using the relation of Equation (A5) (the integration of Equation (A5) represents the volume that is removed, and its volume is matched by the diffused volume). A portion of the mass that diffuses to the pore fluid may deposit to the free surface of the fracture (Equation (5)), resulting in an additive reduction in fracture void volume. Alternately, net dissolution from the fracture wall (Equation (4)) and resulting enlargement of the void cavity will compete with the closure occasioned by the shortening of the bridging asperity. The dominant process will prescribe whether the fracture gapes or seals. This deposition or dissolution on the free surface is controlled by the relative concentration differential between the pore fluid solution and the equilibrium concentration of that fluid (Equations (4) and (5)). Concurrently, mineral

concentrations in the immobile fluid layer beneath the asperity contact, and the mobile fluid in the fracture void fluid are updated, as [35]

$$C_{int}|_{t+\Delta t} = \frac{(D_1 + V_p/2\Delta t) \cdot (dM_{diss}/dt + V_p/4\Delta t \cdot C_{int}|_t) + D_1 V_p/2\Delta t \cdot C_{pore}|_t}{(D_1 + V_p/4\Delta t) \cdot (D_1 + V_p/2\Delta t) - D_1^2} \quad (A6)$$

where

$$D_1 = \frac{2\pi\varpi D_b}{\ln(d_c/2a)} \quad (A7)$$

$$C_{pore} = \frac{1}{Q}\left(\frac{dM_{diff}}{dt} + \frac{dM_{diss}^{FF}}{dt}\right) \quad (\text{if } C_{pore} < C_{eq}) \quad (A8)$$

$$C_{pore} = \frac{1}{Q}\left(\frac{dM_{diff}}{dt} - \frac{dM_{prec}}{dt}\right) \quad (\text{if } C_{pore} > C_{eq}) \quad (A9)$$

and Q denotes the flow rate.

These relations are used iteratively to follow the evolution of dissolved concentrations in the fracture void, and resulting closure-history of the fracture.

NOMENCLATURE

A_c	area of local contact (m^2)
A_e	area of one element (m^2)
A_{pore}	area of fracture void (m^2)
b	local aperture (m)
$\langle b \rangle$	mean mechanical aperture (m)
C_{eq}	equilibrium solubility (kg m^{-3})
C_{out}	lumped concentration travelling out of domain (kg m^{-3})
C_{pore}	concentration in pore space (kg m^{-3})
D	diffusion coefficient (m^2 s^{-1})
$E_{k_{+/-}}$	activation energy for dissolution/precipitation (J mol^{-1})
E_m	heat of fusion (J mol^{-1})
E_C	activation energy for solubility (J mol^{-1})
E_D	activation energy for diffusion (J mol^{-1})
$k_{+/-}$	dissolution/precipitation rate constant (mol m^{-2} s^{-1})
M^p	forward-particle-tracked mass (kg)
M_{prec}	precipitation mass (kg)
M_{diss}^{FF}	dissolution mass at pore space (kg)
M_{diss}^{PS}	dissolution mass at contact area (kg)
M^*	Lagrangian mass (kg)
n	number of time steps (dimensionless)
n_c	number of contact nodes (dimensionless)
n_t	number of total nodes (dimensionless)
N_i	shape function at ith node (dimensionless)
p	fluid pressure (Pa)

Q	flow rate ($m^3\,s^{-1}$)
R	gas constant ($J\,mol^{-1}\,K^{-1}$)
R_c	contact-area ratio (dimensionless)
T	temperature (K)
T_m	temperature of fusion (K)
$\mathbf{V}$	fluid velocity ($m\,s^{-1}$)
V_m	molar volume (m^3)
$\mathbf{x}$	particle position (m)
$\mathbf{x}^*$	fictitious particle position (m)

Greek letters

μ	fluid viscosity (Pa s)
ρ_g	density ($kg\,m^{-3}$)
σ_a	contact stress (Pa)
σ_c	critical stress (Pa)
σ_{eff}	effective stress (Pa)

REFERENCES

1. Morrow CA, Moore DE, Lockner DA. Permeability reduction in granite under hydrothermal conditions. *Journal of Geophysical Research* 2001; **106**:30 551–30 560.
2. Tenthorey E, Cox SF, Todd HF. Evolution of strength recovery and permeability during fluid–rock reaction in experimental fault zones. *Earth and Planetary Science Letters* 2003; **206**:161–172.
3. Moore DE, Lockner DA, Byerlee JD. Reduction of permeability in granite at elevated temperatures. *Science* 1994; **265**:1558–1561.
4. Beer NM, Hickman SH. Stress-induced, time-dependent fracture closure at hydrothermal conditions. *Journal of Geophysical Research* 2004; **109**:B02211, doi:10.1029/2002JB001782.
5. Lin W, Roberts J, Glassley W, Ruddle D. Fracture and matrix permeability at elevated temperatures. *Workshop on Significant Issues and Available Data*. Near-field/Altered-zone coupled effects expert elicitation project, San Francisco, November 1997.
6. Polak A, Elsworth D, Yasuhara H, Grader A, Halleck P. Permeability reduction of a natural fracture under net dissolution by hydrothermal fluids. *Geophysical Research Letters* 2003; **30**(20):2020, doi:10.1029/2003GL017575.
7. Durham WB, Bonner BP. Self-propping and fluid flow in slightly offset joints at high effective pressures. *Journal of Geophysical Research* 1994; **99**:9391–9399.
8. Durham WB, Bourcier WL, Burton EA. Direct observation of reactive flow in a single fracture. *Water Resources Research* 2001; **37**:1–12.
9. Dobson PF, Kneafsey TJ, Sonnenthal EL, Spycher N, Apps JA. Experimental and numerical simulation of dissolution and precipitation: implications for fracture sealing at Yucca Mountain, Nevada. *Journal of Contaminant Hydrology* 2003; **62–63**:459–476.
10. Bryant SL, Schechter RS, Lake LW. Interactions of precipitation/dissolution waves in ion exchange in flow through permeable media. *AIChE Journal* 1986; **32**:751–764.
11. Liu X, Ormond A, Bartko K, Li Y, Ortoleva P. A geochemical reaction-transport simulator for matrix acidizing analysis and design. *Journal of Petroleum Science and Engineering* 1997; **17**:181–196.
12. Polak A, Elsworth D, Liu J, Grader AS. Spontaneous switching of permeability changes in a limestone fracture with net dissolution. *Water Resources Research* 2004; **40**:W03502, doi:10.1029/2003WR002717.

13. Yasuhara H, Polak A, Mitani Y, Grader A, Haleck P, Elsworth D. Evolution of fracture permeability through fluid–rock reaction under hydrothermal conditions. *Earth and Planetary Science Letters* 2006, in press.
14. Neuman SP. Adaptive Eulerian–Lagrangian finite element method for advection–dispersion. *International Journal for Numerical Methods in Engineering* 1984; **20**:321–337.
15. Yeh GT. A Lagrangian–Eulerian method with zoomable hidden fine-mesh approach to solving advection–dispersion equations. *Water Resource Research* 1990; **26**(6):1133–1144.
16. Konikow LF, Goode DJ, Hornberger GZ. A three-dimensional method-of-characteristics-transport model (MOC3D). *Water-Resources Investigations Report 96-4267*, U.S. Geological Survey, Reston, VA, 1996.
17. Mitani Y, Esaki T, Zhou G, Nakashima Y. Experiments and simulation of shear-flow coupling properties of a rock joint. In *Proceedings of the 39th U.S. Rock Mechanics Symposium*, Culligan PJ, Einstein HH, Whittle AJ (eds). Cambridge, 2003; 1459–1464.
18. Piggott AR, Elsworth D. Laboratory assessment of the equivalent apertures of a rock fracture. *Geophysical Research Letters* 1993; **30**(13):1387–1390.
19. Pashley RM. Hydration forces between mica surfaces in electrolyte solutions. *Advances in Colloid and Interface Science* 1982; **16**:57–62.
20. Horn RG, Smith DT, Haller W. Surface forces and viscosity of water measured between silica sheets. *Chemical Physics Letters* 1989; **162**:404–408.
21. Brown SR. Fluid flow through rock joints: the effect of surface roughness. *Journal of Geophysical Research* 1987; **92**:1337–1347.
22. Zimmerman RW, Kumar S, Bodvarsson GS. Lubrication theory analysis of the permeability of rough-walled fractures. *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts* 1991; **28**:325–331.
23. Brown SR, Stockman HW, Reeves SJ. Applicability of the Reynolds equation for modeling fluid flow between rough surfaces. *Geophysical Research Letters* 1995; **22**(18):2537–2540.
24. Yasuhara H, Elsworth D, Polak A. A mechanistic model for compaction of granular aggregates moderated by pressure solution. *Journal of Geophysical Research* 2003; **108**:B11, 2530, doi:10.1029/2003JB002536.
25. Heidug WK. Intergranular solid–fluid phase transformations under stress: the effect of surface forces. *Journal of Geophysical Research* 1995; **100**:5931–5940.
26. Revil A. Pervasive pressure-solution transfer: a poro-visco-plastic model. *Geophysical Research Letters* 1999; **26**:255–258.
27. Stephenson LP, Plumley WJ, Palciauskas VV. A model for sandstone compaction by grain interpenetration. *Journal of Sedimentary Petrology* 1992; **62**:11–22.
28. Steefel CI, Lasaga AC. A coupled model for transport of multiple chemical species and kinetic precipitation/dissolution reactions with application to reactive flow in single phase hydrothermal systems. *American Journal of Science* 1994; **294**(5):529–592.
29. Rimstidt JD, Barnes HL. The kinetics of silica–water reactions. *Geochimica et Cosmochimica Acta* 1980; **44**:1683–1699.
30. Dove PM, Crerar DA. Kinetics of quartz dissolution in electrolyte solutions using a hydrothermal mixed flow reactor. *Geochimica et Cosmochimica Acta* 1990; **54**:955–969.
31. Fournier RO, Potter RWII. An equation correlating the solubility of quartz in water from 25°C to 900°C at pressure up to 10,000 bars. *Geochimica et Cosmochimica Acta* 1982; **46**:1969–1973.
32. Rutter EH. The kinetics of rock deformation by pressure solution. *Philosophical Transactions of the Royal Society of London*, Series A 1976; **283**:203–219.
33. Tada R, Maliva R, Siever R. Rate laws for water-assisted compaction and stress-induced water–rock interaction in sandstones. *Geochimica et Cosmochimica Acta* 1987; **51**:2295–2301.
34. Revil A. Pervasive pressure solution transfer in a quartz sand. *Journal of Geophysical Research* 2001; **106**:8665–8686.
35. Yasuhara H, Elsworth D, Polak A. Evolution of permeability in a natural fracture: significant role of pressure solution. *Journal of Geophysical Research* 2004; **109**(B3):B03204, doi:10.1029/2003JB002663.
36. Lee VW, Mackwell SJ, Brantley SL. The effect of fluid chemistry on wetting textures in novaculite. *Journal of Geophysical Research* 1991; **96**:10 023–10 037.
37. Yasuhara H, Elsworth D, Polak A, Liu J, Grader A, Halleck P. Spontaneous permeability switching in fractures in carbonate: lumped parameter representation of mechanically- and chemically-mediated dissolution. *Transport in Porous Media* 2006, in press.

[7:2] Alternative Solution Models

Level Set Methods

Application of the level set method to the finite element solution of two-phase flows

M. Quecedo[1] and M. Pastor[2,*,†]

[1] *ENUSA and E.T.S. de Caminos, Madrid, Spain*
[2] *Centro de Estudios y Experimentación de OBRAS Públicas (CEDEX) and E.T.S. de Caminos, Madrid, Spain*

SUMMARY

This paper presents a method to solve two-phase flows using the finite element method. On one hand, the algorithm used to solve the Navier–Stokes equations provides the neccessary stabilization for using the efficient and accurate three-node triangles for both the velocity and pressure fields. On the other hand, the interface position is described by the zero-level set of an indicator function. To maintain accuracy, even for large-density ratios, the pseudoconcentration function is corrected at the end of each time step using an algorithm successfully used in the finite difference context. Coupling of both problems is solved in a staggered way. As demonstrated by the solution of a number of numerical tests, the procedure allows dealing with problems involving two interacting fluids with a large-density ratio. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: level set; Navier–Stokes; free surface; characteristics Galerkin; uncompressible flows; fractional step

1. INTRODUCTION

There is a large variety of problems, such as the motion of droplets and bubbles, free surface flows, mould filling, debris flow, etc. involving two fluids interacting across a gas–liquid interface. In these problems, a jump in the fluid characteristics, viscosity and notably density, exists across the interface and the evolution of the system strongly depends on the fluid-to-fluid interaction. Therefore, accurate calculation of the interface evolution is critical for the problem solution.

To describe the evolution of the interface, finite difference practitioners have already developed a number of methods. These methods can be classified as front-tracking and volume-tracking algorithms.

Front-tracking methods use, apart form the stationary mesh used to discretize the overall domain, an additional set of computational elements to describe the interface. In 2D cases, the interface is

---

Copyright © 2001 John Wiley & Sons, Ltd.

formed by line segments connecting chains of points that are advected by the flow. To maintain the front resolution during the advection, extra points are incorporated or deleted from the moving interface as it expands or contracts.

Although this method results in a sharp definition of the front, is expensive and complex to implement even for 2D problems solution and its extension to 3D cases is non-trivial. Furthermore, there is a serious difficulty for this method to handle merging interfaces. To solve the latter issue, merging algorithms and other alternatives have been developed [1].

In the volume-tracking methods, a marker is advected by the flow and used to define the regions occupied by the different fluids. Marker particles were earlier used by Harlow and Welch [2, 3] in the marker-and-cell (MAC) method. This procedure was later extended by the marker function concept, the volume-of-fluid (VOF), due to Hirt and Nichols [4]. Further developments on this approach include the simple line interface calculation (SLIC) [5, 6]. However, the VOF method uses a mesh of rectangular donor–acceptor cells difficult to use by the finite element method.

Mixed approaches of cells and finite elements have been successfully applied [7], but the pseudoconcentration function (PCM) of Thompson [8] overcomes the above difficulty within the FEM context and it has been successfully used by Lewis [9]. However, the existence of high gradients in both the velocity and the pseudoconcentration, can result in non-physical oscillations of the PCM leading to the inception of false interfaces.

This problem has been addressed by Dhatt [10] and, later, by Medale [11] for 2D problems. As their proposals require the solution of a number of topological issues, Riemann problems, etc. they seem to be complicated to be implemented for general type of problems and their extension to 3D situations is non-straightforward.

As an alternative, level set methods [12] have been successfully applied within the finite difference context to solve flow problems involving two gases of similar densities [13] and fluids with much larger-density ratios.

Sussman [14] proposed an efficient algorithm to maintain the function indicating the different fluid regions as a distance function. In this way, solution accuracy was remarkably improved. Furthermore, (1) there is no obvious restriction to extent Sussman's method to the solution of 3D problems and (2) the algorithm used for this purpose is very close to that used to advect the indicator function by the fluid flow. For these reasons, Sussman's proposal is followed in this paper to calculate the interface position using the FEM.

As regards the basic flow calculation, to fulfil the Babuška–Brezzi condition Dahtt, Medale and other authors have used mixed elements with different order of interpolation for velocities and pressures.

These high-order elements are not specially well suited to capture steep gradients in the field variables as those occurring across the interface and for adaptive remeshing [15]. For these purposes, linear triangles in both fields are preferred. Besides, in this way, both the Navier–Stokes equations and the advection of the indicator function are solved using the same mesh. Unfortunately, triangles with linear interpolation of velocity and pressure, do not fulfil the Babuška–Brezzi condition and require stabilization [16] to avoid spurious oscillations in the pressure field.

However, some fractional step algorithms, based either on the Taylor–Galerkin [17] or on the characteristic Galerkin methods [18, 19], provide the required stabilization. The method followed in this paper uses the characteristic Galerkin procedure for compressible/incompressible flow proposed by Zienkiewicz et al. [18, 19] to solve the basic flow problem.

The paper is organized as follows: Section 2 presents the mass balance and momentum equations describing the motion of the two fluids together with the general strategy used to advect

the interface. Section 3 describes the discretization of the governing equations. A characteristic Galerkin, fractional step method is used for the hydrodynamics of the two phases. The advection of the interface is performed in two steps: (1) transport of the interface using a characteristic Galerkin algorithm and (2) iterative correction of the pseudoconcentration. Finally, the performance of the proposed method is assessed in Section 4 using classical tests involving fluids with large differences in density and interface merging.

2. GOVERNING EQUATIONS

2.1. Equations of motion

Determination of the unsteady, viscid, incompressible, flow of two interacting fluids requires the solution of the Navier–Stokes equations

$$\rho \frac{\partial \bar{u}}{\partial t} + \rho \operatorname{div}(\bar{u} \otimes \bar{u}) = \operatorname{div} \tau - \operatorname{grad} p + \rho \bar{b} \tag{1}$$

$$\operatorname{div} \bar{u} = 0 \tag{2}$$

supplemented by the state equations for the density and viscosity, which for immiscible fluids are

$$\frac{\partial \rho}{\partial t} + \operatorname{grad} \rho \cdot \bar{u} = 0 \tag{3}$$

$$\frac{\partial \mu}{\partial t} + \operatorname{grad} \mu \cdot \bar{u} = 0 \tag{4}$$

In the above equations, $\rho(\bar{x}, t)$ and $\mu(\bar{x}, t)$ are the discontinuous density and viscosity fields, respectively, $\bar{u}(\bar{x}, t)$ is the fluid velocity, $p(\bar{x}, t)$ is the pressure, $\tau(\bar{x}, t)$ is the deviatoric stress tensor which, for an incompressible, simple viscid fluid is related to the symmetric strain rate tensor through the viscosity coefficient, μ,

$$\tau = 2\mu \dot{\varepsilon} = \mu(\operatorname{grad} \bar{u} + \operatorname{grad}^{\mathrm{T}} \bar{u})$$

and, $\bar{b}(\bar{x}, t)$ is a body force, typically, gravity. Another forces, as surface tension at the fluids interface, are ignored in this paper.

Equations (3) and (4) state that for immiscible fluids, the density and viscosity are constant along particle paths, i.e. the material derivative of both variables is zero.

From the spatial point of view, the density and viscosity are constant inside each fluid, experiencing a jump only at the fluids interface, i.e. the front. Therefore, ρ and μ can be written as a function of a smooth variable, ϕ, whose zero-level set indicates the fluid-to-fluid interface

$$\{\rho, \mu\} = \begin{cases} \{\rho, \mu\}_1 & \text{if } \phi \geqslant 0 \\ \{\rho, \mu\}_2 & \text{if } \phi < 0 \end{cases} \tag{5}$$

In this way, the solution of the state equations (3) and (4) reduces to the simple advection by the fluid flow of the indicator function ϕ

$$\frac{\partial \phi}{\partial t} + \operatorname{grad} \phi \cdot \bar{u} = 0 \tag{6}$$

and the determination of its zero-level set at each time step.

Besides, the representation of the interface using a smooth function instead of the step one used in the equation (5) levels out most of the numerical oscillations caused by the shortest wavelengths present in the mesh, while maintaining the accuracy in the interface position.

This approach to the solution of the problem defined by Equations (1) (2) (5) and (6) works well for small-density ratios, as verified in Reference [13] in the finite difference context, but it results in unwanted instabilities in the pressure field for larger ratios as those found in gas–liquid systems.

Additionally, the advection of the indicator by a non-uniform velocity field can result in non-physical, spurious, interfaces appearing in the problem solution. It is important to notice that this effect is not caused by inaccuracies or instabilities of the numerical scheme but by high-velocity gradients present in the fluid field [11]. Clearly, poor numerical schemes will make this issue to become worse.

Finally, the numerical scheme contaminates the numerical solution either by diffusion, in the case of low-order schemes, or destroying the front sharpness causing oscillations, in the high-order case.

2.2. Front capturing

When using the FEM to analyse mould-filling problems, the above issues have been successfully solved by (1) front tracking and ϕ function reconstruction from the new front position as in Reference [10] or (2) by reducing the front advection calculations to a limited number of elements located at the sides of the front: the cursor, which is updated as the front propagates [11].

As these front-tracking methods require too much book-keeping effort and the solution of several topological issues, among other difficulties, the approach adopted in this paper is based on

1. The diffuse representation of the front.
2. Front capturing using the level set approach.

As regards the first point, to avoid the abrupt changes in the density and viscosity fields implied by (5) when crossing the front, these properties are interpolated through a constant thickness tube of width 2δ surrounding the front [1]. The zero-level set of ϕ indicates the front position and δ is taken of the order of the mesh size [10].

Front smoothing further prevents from oscillations of length scale of the order of the mesh size [1] and, finally, maintaining the tube thickness constant through the advection eliminates the diffusion issues. Front smoothing has been previously used in Reference [11], where the transition region is formed by the elements crossed by the interface. However, using the cursor concept in Reference [11], the diversity of element sizes in the mesh results in a changing front thickness as the front is advected through the mesh.

Concerning the interpolation of the fluid properties, different alternatives exist. This paper considers the simple linear interpolation

$$\rho = \begin{cases} \rho_1 & \text{for } \phi \leqslant -\delta \\ \rho_1 + \dfrac{\rho_2 - \rho_1}{2\delta}(\phi + \delta) & \text{for } -\delta < \phi < \delta \\ \rho_2 & \text{for } \phi \geqslant \delta \end{cases} \tag{7}$$

although extra smoothing can be gained considering other functions, as the sine function used in Reference [14].

However, the interpolation of the fluid properties as a function of ϕ, requires the value of ϕ indicating the signed distance to the interface, i.e. ϕ being a distance function. Besides, requiring $|\text{grad}\,\phi| = 1$ also avoids spurious interfaces appearing and contaminating the problem solution.

To keep ϕ as a distance, this paper follows the approach proposed by Sussman *et al.* [14] within the finite difference context: once ϕ has been advected up to a certain time t^n, to correct $\phi(\bar{x}, t^n)$ the following problem is evolved to steady state:

$$\frac{\partial \psi}{\partial t} + S(\phi^n)|\text{grad}\,\psi| = S(\phi^n) \tag{8}$$

with initial conditions

$$\psi(\bar{x}, t) = \phi^n(\bar{x})$$

where

- $S(\ )$ is the sign function and
- ϕ^n is the solution of Equation (6) at time t^n, $\phi(\bar{x}, t^n)$.

Clearly, the zero-level set of ψ, indicating the front position, matches that of ϕ^n, thus the term front capturing, and, when reaching the steady state, $|\text{grad}\,\psi| = 1$. Thus, while ϕ^n is not a distance function, the steady state solution of Equation (8) will be.

Equation (8) can also be written as

$$\frac{\partial \psi}{\partial t} + S(\phi^n)\frac{\text{grad}\,\psi}{|\text{grad}\,\psi|}\,\text{grad}\,\psi = S(\phi^n) \tag{9}$$

showing that this problem consists in advecting ψ by a velocity field

$$\bar{v} = S(\phi_n)\,\text{grad}\,\psi/|\text{grad}\,\psi|$$

Therefore, the same algorithms used for advecting the indicator function can be used with advantage to evolve Equation (8) to steady state. Besides, specific numerical schemes for this type of equation are also available [20].

It is also noted that the velocity field is the unit normal pointing outward from the zero-level set. Therefore, the zero-level set is the appropriate inlet boundary to prescribe the boundary conditions for this problem.

2.3. Summary

The method proposed in this paper to solve the two fluids flow problem consists, of the following for each time step:

1. Solving the Navier–Stokes equations of motion (1) and (2) with the appropriate initial and boundary conditions.
2. Solving the advection of the indicator function (6) initialized as a signed distance to the interface and considering the appropriate boundary conditions.
3. Capturing the interface and keeping the indicator function as a signed distance to the interface by evolving Equation (9) to steady state.
4. Interpolating the density and viscosity (7) for the next calculations.

The next section describes the discretization procedure for these problems.

3. DISCRETIZATION

There is a large variety of FEM-based procedures available for the solution of advection-dominated problems (see, for instance, References [21, 22]). Among them, the characteristics Galerkin method has demonstrated an excellent performance in general type of problems [23], specially in situations involving steep gradients in the field variables [24]. Thus, this procedure has been chosen to solve the Navier–Stokes and the two advection problems.

The next section presents the details of the discretization.

3.1. Navier–Stokes

Time discretization of Equation (1) along the characteristics [18] results in the following semi-implicit equation for time increment n

$$\rho \frac{\Delta \bar{u}^n}{\Delta t} + \rho \operatorname{div}(\bar{u}^n \otimes \bar{u}^n) - \operatorname{div} \tau^n - \frac{\Delta t}{2} \bar{u}^n \cdot \operatorname{grad}[\rho \operatorname{div}(\bar{u}^n \otimes \bar{u}^n) - \operatorname{div} \tau^n]^n + \operatorname{grad} p^{n+1} = \bar{0}$$

Body forces, as gravity, have been ignored in the above equation derivation.

Following now a fractional step procedure as proposed by Chorin [25], the velocity can be decomposed into two parts

$$\Delta \bar{u}^n = \Delta \bar{u}^{*,n} + \Delta \bar{u}^{**,n} \tag{10}$$

such as

$$\rho \frac{\Delta \bar{u}^{*,n}}{\Delta t} + \rho \operatorname{div}(\bar{u}^n \otimes \bar{u}^n) - \operatorname{div} \tau^n - \frac{\Delta t}{2} \bar{u}^n \cdot \operatorname{grad}[\rho \operatorname{div}(\bar{u}^n \otimes \bar{u}^n) - \operatorname{div} \tau^n]^n = \bar{0} \tag{11}$$

$$\rho \frac{\Delta \bar{u}^{**,n}}{\Delta t} + \operatorname{grad} p^{n+1} = \bar{0}$$

which are complemented by the continuity equation

$$\operatorname{div} \bar{u}^{n+1} = 0 \tag{12}$$

For the spatial discretization, the computationally efficient three-nodes triangles are preferred. Furthermore, it is well known that the excellent performance of the low-order elements in the solution of problems involving sharp fronts [26]. However, the Babuška–Brezzi condition [27, 28] precludes using equal order of interpolation for both the velocity and pressure fields unless some stabilization is provided.

The selected scheme provides such stabilization [17, 21, 24, 29, 30] and therefore it allows using the linear interpolation of both the velocity and pressures

$$\bar{u} = \mathbf{N}\bar{u}$$

$$p = \bar{N} \cdot \bar{p}$$

where $\mathbf{N}/\bar{N}$ are the shape functions matrix/vector for the three-node triangle.

3.1.1. Step 1 : *fractional velocity discretization.* Therefore, the discretization of the fractional momentum equation starts from

$$\rho \frac{\Delta \bar{u}^*}{\Delta t} + \rho \operatorname{div}(\bar{u} \otimes \bar{u}) - \operatorname{div} \tau - \frac{\Delta t}{2} \bar{u} \cdot \operatorname{grad}[\rho \operatorname{div}(\bar{u} \otimes \bar{u}) - \operatorname{div} \tau] = \bar{0}$$

where the superscript n has been dropped for convenience. Using the standard Galerkin weighting it results in

$$\int_{\Omega} \frac{\rho}{\Delta t} \mathbf{N}^{\mathrm{T}} \mathbf{N} \, \mathrm{d}\Omega \, \Delta \bar{\mathbf{u}}^{**} + \int_{\Omega} \rho \left(\mathbf{N}^{\mathrm{T}} + \frac{\Delta t}{2} \mathbf{C}^{\mathrm{T}} \right) \operatorname{div}(\bar{u} \otimes \bar{u}) \, \mathrm{d}\Omega - \int_{\Gamma - \Gamma_{\sigma}} \mathbf{N}^{\mathrm{T}} \tau \bar{n} \, \mathrm{d}\gamma$$

$$- \frac{\Delta t}{2} \int_{\Gamma} \rho \, \bar{u} \cdot \bar{n} \, \mathbf{N}^{\mathrm{T}} \operatorname{div}(\bar{u} \otimes \bar{u}) \, \mathrm{d}\gamma - \int_{\Gamma_{\sigma}} \mathbf{N}^{\mathrm{T}} \bar{t}_s \, \mathrm{d}\gamma + \int_{\Omega} \mathbf{B}^{\mathrm{T}} \tau \, \mathrm{d}\Omega = \bar{0} \qquad (13)$$

where, considering Cartesian co-ordinates,

$$\mathbf{C}^I = \begin{pmatrix} N^I \operatorname{div} \bar{u} + \dfrac{\partial N^I}{\partial x} u_x + \dfrac{\partial N^I}{\partial y} u_y & 0 \\[2ex] 0 & N^I \operatorname{div} \bar{u} + \dfrac{\partial N^I}{\partial x} u_x + \dfrac{\partial N^I}{\partial y} u_y \end{pmatrix} \quad \text{for node } I$$

$$\mathbf{B}^I = \begin{pmatrix} \dfrac{\partial N^I}{\partial x} & 0 \\[2ex] 0 & \dfrac{\partial N^I}{\partial y} \\[2ex] \dfrac{\partial N^I}{\partial y} & \dfrac{\partial N^I}{\partial x} \end{pmatrix}$$

$$\tau = \begin{pmatrix} \tau_x \\ \tau_y \\ \tau_{xy} \end{pmatrix}$$

Besides, by using linear elements, the contribution of the higher-order spatial derivatives of the velocity field, $\bar{u} \cdot \operatorname{grad}(\operatorname{div} \tau)$, has been neglected [18].

The boundary conditions are:

- On Γ_{σ} a shear traction $\bar{t}_s$ is prescribed. As $\bar{t}_s = \tau \bar{n}$, care should be taken to properly deal with the pressure in the shear traction prescription: $\bar{t}_s = \tau \bar{n} = \sigma \bar{n} + p\bar{n} = \bar{t} + p\bar{n}$.
- Fluid velocities are prescribed on $\Gamma - \Gamma_{\sigma}$. However, conditions on $\Delta \bar{u}^*$ are unknown. Therefore, the integral $\int_{\Gamma - \Gamma_{\sigma}} \mathbf{N}^{\mathrm{T}} \tau \bar{n} \, \mathrm{d}\gamma$ cannot be omitted by assuming $\Delta \bar{u}^*$ is known at the boundary. Calculating this integral, as proposed by Codina [19], eliminates additional assumptions, as in Reference [17], and possible inaccuracies. In this way, $\Delta \bar{u}^*$ is calculated at each node in the mesh and used later.

  It is pointed out that in the case of linear triangles, the first derivatives, i.e. gradient and divergence, of the different fields involved in the calculations are constant within the element. Therefore, they will always be calculated at one Gauss point location.

However, $\rho(\mathbf{N}^{\mathrm{T}} + \Delta t/2\mathbf{C}^{\mathrm{T}}) \operatorname{div}(\bar{u} \otimes \bar{u})$ is cubic at the front, where the density varies linearly, and quadratic everywhere outside the front. Therefore, density, velocities and shape functions involved in the integral calculation are determined at three Gauss points. Although slightly inaccurate at the front, it provides an adequate compromise between computational effort and the required accuracy.

Finally, constant source terms, such as gravity, can be directly added to the right-hand side of Equation (13) as

$$\int_{\Omega} \rho \mathbf{N}^{\mathrm{T}} \bar{b} \, \mathrm{d}\Omega$$

This integral is calculated using one Gauss point.

3.1.2. Step 2 : *continuity equation discretization.* Taking into account the continuity equation at time t^{n+1}

$$\operatorname{div} \bar{u}^{n+1} = 0$$

and using the incremental momentum split,

$$\rho \frac{\Delta \bar{u}^{**,n}}{\Delta t} + \operatorname{grad} p^{n+1} = \bar{0} \tag{14}$$

the time discretized continuity equation results in

$$\operatorname{div} \bar{u}^{n,*} - \Delta t \operatorname{div} \left(\frac{1}{\rho} \operatorname{grad} p^{n+1} \right) = 0$$

were $\bar{u}^{n,*} = \bar{u}^n + \Delta \bar{u}^{n,*}$. Note that to improve the accuracy, as explained in the previous section, the term $\operatorname{div} \bar{u}^n$ is kept in the calculations.

Following now the standard Galerkin discretization, this equation becomes

$$\int_{\Omega} \frac{1}{\rho} \operatorname{grad} \bar{N} \operatorname{grad}^{\mathrm{T}} \bar{N} \, \mathrm{d}\Omega \, \Delta \bar{p}^n$$

$$= -\frac{1}{\Delta t} \int_{\Omega} \operatorname{div} \bar{u}^{n,*} \bar{N} \, \mathrm{d}\Omega - \int_{\Omega} \frac{1}{\rho} \operatorname{grad} \bar{N} \operatorname{grad} p^n \, \mathrm{d}\Omega + \int_{\Gamma - \Gamma_p} \frac{1}{\rho} \operatorname{grad} p^{n+1} \cdot \bar{n} \, \bar{N} \, \mathrm{d}\gamma$$

where it has been assumed that the pressure is prescribed on Γ_p.

To calculate the boundary integral, Equation (14) is projected along the normal, $\bar{n}$,

$$\frac{1}{\rho} \operatorname{grad} p^{n+1} \cdot \bar{n} = -\frac{1}{\Delta t} [\bar{u}^{n+1} - (\bar{u}^n + \Delta \bar{u}^{*,n})] \cdot \bar{n}$$

resulting in

$$\int_{\Gamma - \Gamma_p} \frac{1}{\rho} \operatorname{grad} p^{n+1} \cdot \bar{n} \, \bar{N} \, \mathrm{d}\Omega = -\frac{1}{\Delta t} \int_{\Gamma - \Gamma_p} [\bar{u}^{n+1} - (\bar{u}^n + \Delta \bar{u}^{*,n})] \bar{n} \, \bar{N} \, \mathrm{d}\gamma$$

This procedure avoids neglecting grad p^{n+1} as in References [19, 21] and, thus the boundary integral on $\Gamma - \Gamma_p$.

Another possibility, which has been successfully used in the context of Solid Dynamics [31], is approximating grad p^{n+1} by its value at the previous time step, grad p^n.

The former alternative provides better accuracy and therefore it is preferred. In this way, the system of equations to solve and calculate the pressure increment is

$$\int_\Omega \frac{1}{\rho} \text{grad}\, \bar{N}\, \text{grad}^T\, \bar{N}\, \mathrm{d}\Omega\, \Delta \bar{p}^n = -\frac{1}{\Delta t} \int_\Omega \text{div}\, \bar{u}^{n,*} \bar{N}^T\, \mathrm{d}\Omega$$

$$-\int_\Omega \frac{1}{\rho} \text{grad}\, \bar{N}\, \text{grad}\, p^n\, \mathrm{d}\Omega - \frac{1}{\Delta t} \int_{\Gamma - \Gamma_p} [\bar{u}^{n+1} - (\bar{u}^n + \Delta \bar{u}^{*,n})]\, \bar{n} \bar{N}^T\, \delta\Omega$$

3.1.3. Step 3 : *velocity correction.* Once the pressure increment has been determined in the previous step, the velocity increment $\Delta \bar{u}^*$ should be corrected for the effects of pressure. This is done by solving the spatial discretization of (14) to calculate $\Delta \bar{u}^{**}$:

$$\int_\Omega \frac{\rho}{\Delta t} \mathbf{N}^T \mathbf{N}\, \mathrm{d}\Omega\, \Delta \bar{u}^{**,n} + \int_\Omega \mathbf{N}^T \text{grad}\, p^{n+1}\, \mathrm{d}\Omega = \bar{0} \tag{15}$$

which, according to (10) is added to $\Delta \bar{u}^{*,n}$ to obtain $\Delta \bar{u}^n$.

Finally, $\bar{u}^{n+1}$ is obtained taking into account the corresponding boundary conditions on Γ_u.

3.2. Indicator advection

3.2.1. Step 1: *pure advection.* The indicator function ϕ is advected by the fluid velocity according to

$$\frac{\partial \phi}{\partial t} + \text{grad}\, \phi \cdot \bar{u} = 0 \tag{16}$$

and initialized as the signed distance to the initial front position

$$\phi(\bar{x}, t = 0) = \pm\, \text{distance to the front} \tag{17}$$

The time discretization of (16) along a characteristic results in

$$\frac{\Delta \phi^n}{\Delta t} + \text{grad}\, \phi^n \cdot \bar{u}^n - \frac{\Delta t}{2} \text{grad}(\text{grad}\, \phi \cdot \bar{u})^n \cdot \bar{u}^n = 0$$

and the Galerkin spatial discretization in

$$\int_\Omega \frac{1}{\Delta t} \bar{N} \otimes \bar{N}\, \mathrm{d}\Omega\, \Delta \bar{\phi} + \int_\Omega \text{grad}\, \phi \cdot \bar{u} \left(\bar{N} + \frac{\Delta t}{2} (\bar{u} \cdot \text{grad}\, \bar{N} + \text{div}\, \bar{u}\, \bar{N}) \right)\, \mathrm{d}\Omega$$

$$-\frac{\Delta t}{2} \int_{\Gamma - \Gamma_\phi} \text{grad}\, \phi \cdot \bar{u}(\bar{u} \cdot \bar{n}) \bar{N}\, \mathrm{d}\gamma = \bar{0} \tag{18}$$

where the superscript, n, has been dropped for convenience.

The procedure followed to initialize ϕ in the general case as the signed distance to the initial front position, Equation (17), consists of:

- making first $\phi(\bar{x}, t)$ to be a signed step function of amplitude $2a$, $\phi(\bar{x}, t) = a\,S(\bar{x} - \bar{\gamma})$, where
  — $2a$ is the value assigned for the jump across the front. In the examples presented in this paper, $a = \delta$, as defined in Equation (13).
  — $\bar{\gamma}$ indicates the initial front position and
  — $S()$ is the sign function.
- evolving this step function of amplitude $2a$ using the method described in Section 3.2.2 to convert it to a distance function.

1. To get a well-posed initial boundary value problem, in addition to the initial conditions (17), the value of ϕ should be prescribed at any time along that part of the boundary, Γ_ϕ, which corresponds to the incoming characteristics [21, 26, 32]. As in the current problem the characteristics are particle paths, ϕ should be prescribed at the fluid inlet boundaries, where $\bar{u} \cdot \bar{n} < 0$.

The right value of ϕ on Γ_ϕ at any time t is the distance to the front

$$\phi(\bar{x}, t) = \pm \text{ distance to the front at time } t; \quad \bar{x} \in \Gamma_\phi$$

but, as the front position is unknown, $\phi(t)\,|_{\Gamma_\phi}$ is also unknown.

The method used here consists of approximating the value at t^{n+1} by the value of t^n

$$\phi(\bar{x}, t^{n+1}) = \phi(\bar{x}, t^n) \quad \bar{x} \in \Gamma_\phi$$

leaving the calculation of the right value of $\phi(\bar{x}, t^{n+1})$ for the global correction step performed next to keep ϕ as a distance function.

Finally, as most of the integrals in (18) involve second-order functions, the examples presented in this paper are solved using three Gauss points. However, it is pointed out that only one Gauss point could be sufficiently accurate and slightly faster [26].

3.2.2. Step 2 : correction. As described in Section 2.2, once ϕ has been advected up to time t^n using the algorithm described in the previous section, it is transformed into a distance function, denoted as ψ, by evolving

$$\frac{\partial \psi}{\partial t} + S(\phi^n)\frac{\operatorname{grad} \psi}{|\operatorname{grad} \psi|} \operatorname{grad} \psi = S(\phi^n) \tag{19}$$

to steady state, with initial conditions

$$\psi(\bar{x}, 0) = \phi(\bar{x}, t^n)$$

As already pointed out, $\phi(\bar{x}, t^n)$ and the steady-state solution of (19), ψ, share the same zero-level set. Thus, (19) preserves the fluid-to-fluid interface.

Taking into account that the velocity field in this problem is

$$\bar{v} \equiv S(\phi^n)\frac{\operatorname{grad} \psi}{|\operatorname{grad} \psi|} \tag{20}$$

and, thus,

$$\bar{v} \cdot \operatorname{grad}(\psi) = S(\phi^n)|\operatorname{grad}\psi|$$

the discretized form of Equation (19) is

$$\int_\Omega \frac{1}{\Delta t} \bar{N} \otimes \bar{N} \, \mathrm{d}\Omega \, \Delta \bar{\psi}^k = \int_\Omega \bar{N} \, S(\phi^n) \left[1 - |\operatorname{grad}\psi^k| \left(1 + \frac{\Delta t}{2} \operatorname{div} \bar{v}^k \right) \right] \mathrm{d}\Omega$$

$$- \int_\Omega \frac{\Delta t}{2} \operatorname{grad}\psi \operatorname{grad}\bar{N} \, \mathrm{d}\Omega + \frac{\Delta t}{2} \int_\Gamma \operatorname{grad}\psi \cdot \bar{n} \, \bar{N} \, \mathrm{d}\gamma = \bar{0}$$

(21)

where the superscript k stands for the iteration number towards the steady state.

As only the steady state is of interest, use of the local time-stepping procedure [21, 26] is suggested for convergence acceleration.

It is pointed out that as the velocity field in this problem

$$\bar{v} \equiv S(\phi^n) \frac{\operatorname{grad}\psi}{|\operatorname{grad}\psi|}$$

is not divergence free an stabilization term $\operatorname{div}\bar{v}$ appears in Equation (21). This term involves the calculation of second-order derivatives of ψ. In the authors experience, neglecting it could slow down convergence to steady state and it could result in errors. Therefore, nodal recovery techniques are used to calculate the nodal values of $\operatorname{grad}\psi$ and then, using its definition, the nodal velocities. From the nodal velocities, calculation of $\operatorname{div}\bar{v}$ is straightforward.

For the simple linear triangle used in this paper, the variational recovery is sufficient. Furthermore, Equation (21) uses the $\operatorname{div}\bar{v}^k$ term only as a correction to the first-order advection term. Thus, small errors in the recovery are irrelevant and, consequently, the recovery calculations are carried out simply using the lumped mass matrix.

Finally, to get a well-posed initial boundary value problem, the boundary conditions should be stated. As for any hyperbolic problem, the value of ψ should be prescribed at the inlet boundaries. In the current problem, the characteristics are given by the vector $\bar{v}$ (20), the unit normal pointing outward from the front, i.e. the inlet boundary for this problem is just the front.

Therefore (1) the boundary integral in (21) should be extended to the whole body boundary and (2) ψ should be prescribed only at the front, where it should be zero.

Regarding the second point, the front thickness is 2δ. Therefore, all nodes located inside the front, i.e. those nodes fulfilling $|\psi(\bar{x})^k| \leqslant \delta$, are left unchanged during the iteration k of the correction step.

The value of the front thickness, 2δ, as stated in Section 2.2, is of the order of the mesh size. In the examples presented in this paper

$$\delta = \tfrac{3}{2} h_{\max}$$

(22)

where $h_{\max}$ is the size of the largest element in the mesh.

Another possibility to address the issue of prescribing the boundary values of ψ consists in fixing only those nodes adjacent to the zero-level set of ψ. This purpose can be achieved by detecting the elements crossed by the front and leaving unchanged their nodal values during the correction steps

Both methods leave unchanged the position of the zero-level set of ψ during the correction steps. The results of the numerical tests presented in this paper do not favour one specific procedure either in terms of accuracy or efficiency.

The stopping criterion for the iteration used in this work is

$$\sum_{1}^{\text{total nodes}} |\psi^{k+1}(\bar{x}, t^n) - \psi^k(\bar{x}, t^n)| < \delta \tag{23}$$

In case the error calculation (23) extends over the whole mesh, much of the time is expended fulfilling the criterion in zones located far from the front. As the critical issue in the far field is avoiding appearance of false interfaces, small errors in the far field could be accepted. In this case, nodes located far from the front are removed from the error calculation in (23) and convergence accelerated.

Comparison of the results for a number of tests cases showed that extending the calculations in (14) to a tube thickness

$$|\psi^{k+1}(\bar{x}, t^n)| < 10 * \delta$$

is sufficient for most applications. However, care should be taken to avoid the front crossing the tube at the end of a time step.

3.3. Stability requirements on the time step increment

Coupling of the two problems is solved in each time step in a staggered way. First, the Navier–Stokes equations are solved and then, the new front position is determined as the solution of the transport problem. In this way, the density and viscosity to be used for the next time step can be calculated.

The feedback from the advection step, i.e. the new front position, is extremely important in the case of high-density ratios between the two fluids while for low-density ratios, a number of advection steps can be advanced within the same Navier–Stokes step.

The maximum time step increment allowed for the solution of the Navier–Stokes equation is calculated from the condition [21]

$$C \leqslant \sqrt{\frac{1}{P_e^2} + \alpha} - \frac{1}{P_e}$$

where C is the Courant number; $C = |\bar{u}|/h/\Delta t$, P_e is the Peclet number, $P_e = |\bar{u}|\mu/\rho * h/2$ and h is the element size, which has been considered here as the minimum triangle height and, $\alpha = 1$ when using the lumped mass matrix and $\alpha = \frac{1}{3}$ when using the consistent mass matrix.

The corresponding restriction in the time step increment for the advection equation is

$$C \leqslant \alpha \tag{24}$$

Therefore, for one of the two problems, Navier–Stokes or the indicator advection, the time step will not be the optimum one. This will result in unwanted oscillations and lack of accuracy for that problem [21]. Furthermore, even for the problem setting the time step increment and, thus, being the optimum for that problem, the time step increment will not be optimum for all the elements in the mesh due to their different sizes.
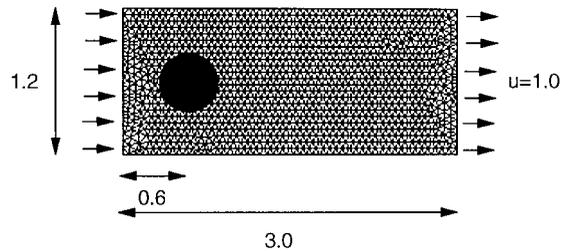
Figure 1. A colour drop advected by a uniform flow. Problem lay-out.

To solve this issue, this paper follows the approach presented in Reference [21]. This approach consists in calculating the time step increment used in the stabilization terms as

$$\Delta t = \frac{\alpha_{\text{opt}}\, h}{|\bar{v}|}$$

where the value of α_{opt} can also be found in Reference [21].

The test cases presented in the next section have been solved using the lumped mass matrix for the Navier–Stokes problem and the consistent matrix for the advection one. However, in the general case, using the consistent mass matrix also for the solution of the Navier–Stokes problem will result in a more accurate solution of the basic flow problem [21, 26] and will help in eliminating spurious oscillations. To save computer effort, the consistent matrix can be approximated by the iterative correction described in Reference [33].

Finally, it is pointed out that the indicator function correction places no restriction in the critical time step for the overall problem solution. However, the internal time step used to solve it, should fulfil with the same conditions as those for the advection problem solution described above, Equation (15).

4. NUMERICAL EXAMPLES

This section shows the capabilities of the methods presented in this paper when solving a number of test problems.

4.1. A colour drop advected by a uniform flow

This test involves a drop of a colour, with radius 0.3, advected by a uniform flow in a 3.0×1.2 rectangular domain. The flow velocity is 1.0, constant everywhere in the domain and it is assumed that the presence of the drop does not affect the flow.

Figure 1 presents the domain, discretized using 1213 nodes in 2280 triangular elements. The drop advection is solved using the characteristics Galerkin method described in Section 3.2.1. The indicator function ϕ is taken as a cone with unit height.

Figure 2 depicts the contours of ϕ at different instants. It is observed that the method is diffusing the drop contour. Thus, an interpolation of the fluid properties across the interface as that in Equation (7) would result in oscillations.

Figure 3 presents the drop contour, i.e. the zero-level set of ϕ at the same instants as above, calculated using the proposed reinitialization. It is observed that the contour remains sharp during the transport and maintains the initial circular shape.
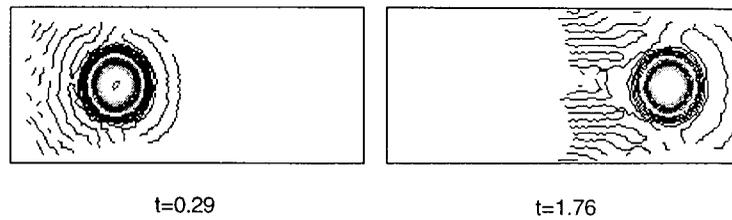
Figure 2. Pure advection of a colour drop by a uniform flow. Contours of the indicator function.
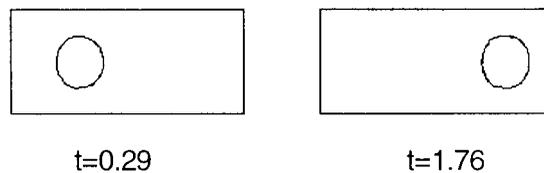
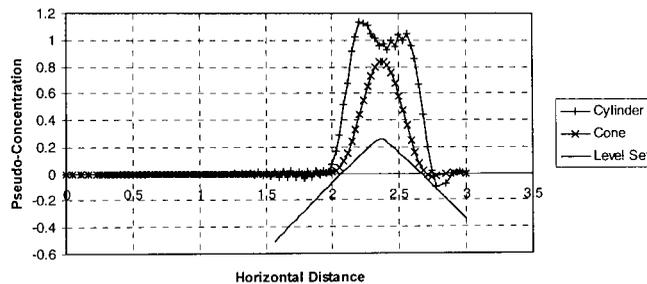Figure 3. Drop contour calculated using the proposed reinitialization.

Figure 4. Advection of a colour drop. Pseudo-concentration function profiles calculated along the central horizontal line assuming ϕ is initialized as: (a) a cylinder; (b) a cone and (c) using the level set with reinitialization.

To further substantiate these observations, Figure 4 depicts the pseudo-concentration function profiles along a central horizontal line at $t = 1.76$ for these two cases. Also included in this figure is the profile calculated when using a steep ϕ function as it is a cylinder. These results make clear that the reinitialization levels out the oscillations existing even using a smooth definitions for ϕ, as it is the cone, and it maintains the drop radius through the advection.

4.2. Flow in a T-branch

This example tests the capability of the proposed advection technique to transport and to merge two interfaces. For this purpose, this test analyses a colour coming into a T-branch from two different inlet boundaries and advected by a steady-state flow.

Figure 5 presents the problem lay-out. The mesh consists of 552 nodes and 972 linear nodes triangles.
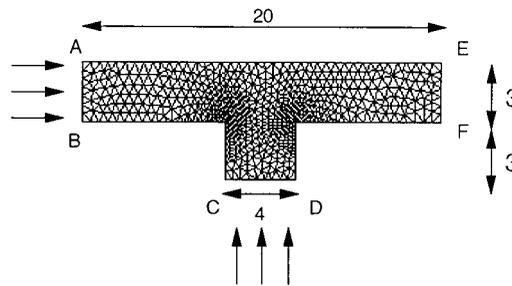
Figure 5. Flow in a T-branch: mesh and problem lay-out.
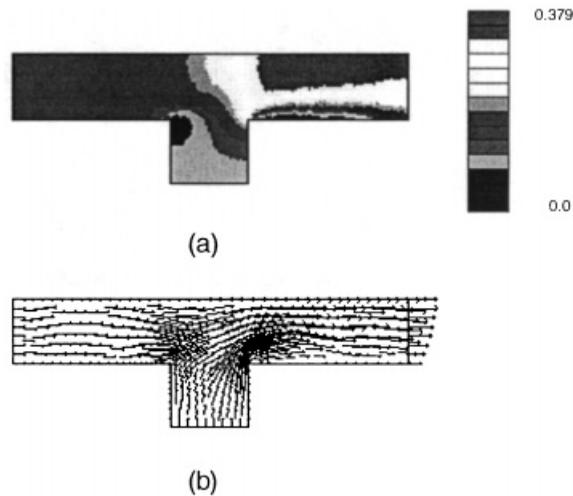
Figure 6. Flow in a T-branch: (a) velocity contours and (b) velocity vectors.

The colour comes into the domain with a velocity $u = 0.15$ m/s at section AB and with velocity $u = 0.10$ m/s at section CD. The pressure is prescribed as zero at section EF and perfect slippage between the fluid and the walls is assumed at the remaining walls of the domain. Normal velocity is set to zero along AE, BC and DF.

Considering $Re = 25$ calculated at section AB, the steady-state flow has been determined independently from the colour advection using the algorithm described in Section 3.1. Figure 6 presents the calculated velocity vectors and contours for steady-state conditions.

Figure 7 presents the evolution of the interfaces through its advection. Clearly, the coalescence of the two interfaces is held without any difficulty.

To illustrate the need of the indicator function reinitialization to accurately calculate the interface position, Figure 8 presents the results obtained now without the proposed reinitialization. It is clear that in this case as the front meets areas where velocity gradients exists, the solution accuracy greatly deteriorates. Therefore, it is important performing the propose reinitialization when the solution of the flow problem is strongly coupled to the interface position, as in the examples presented next.
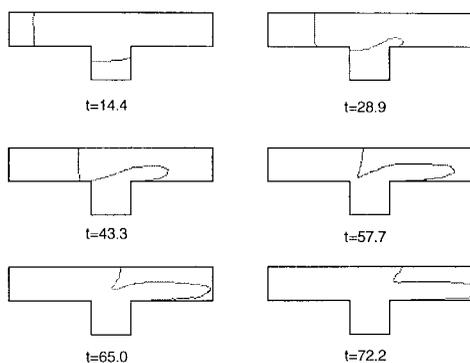
Figure 7. Evolution of the colour interface as advected by the flow.
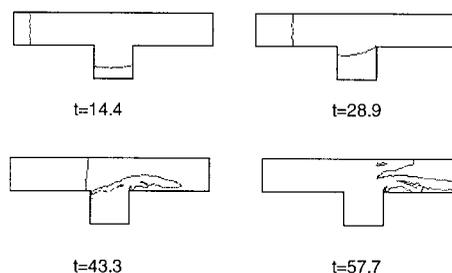
Figure 8. Evolution of the colour interface calculated without reinitialization.
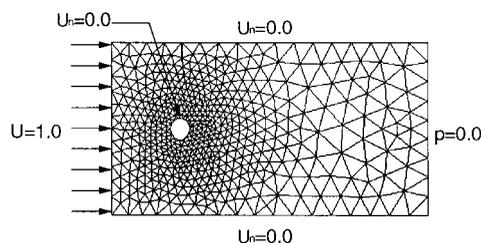
Figure 9. Flow around a cylinder: problem lay-out and mesh.

4.3. Flow around a cylinder

The next example shows the performance of the method when solving the flow around a cylinder. Initially, a fluid with viscosity $\mu = 2.0 \times 10^{-5}$ Pa s and density $\rho = 1.0$ kg/m^3 is at rest filling a rectangular domain which contains a circular cylinder.

At time zero, another fluid, characterized by $\mu = 2.0 \times 10^{-5}$ Pa s and $\rho = 10^3$ kg/m^3, starts entering into the domain with horizontal uniform unit velocity, interacting with the existing one. Figure 9 depicts the problem lay-out and the mesh of 529 nodes in 988 three-node triangles used in the problem discretization.

Figure 10 presents the calculated pressure and velocity contours along with the interface position at different instants. It is pointed out that while using the same order of interpolation for velocities and pressures, oscillations in the pressure field are not present.

4.4. Step cavity

Dhatt [10] and Medale [11] used this test to check the performance of their methods in complex flow problems. The test consists in filling a cavity with a step and initially filled with air at rest, using a molten metal. Therefore, the example includes again the effect of each fluid flow on the other fluid and involves large-density ratios between both fluids.

Figure 11 presents the problem lay-out and the discretization by a mesh of 557 nodes and 989 three-node triangles.
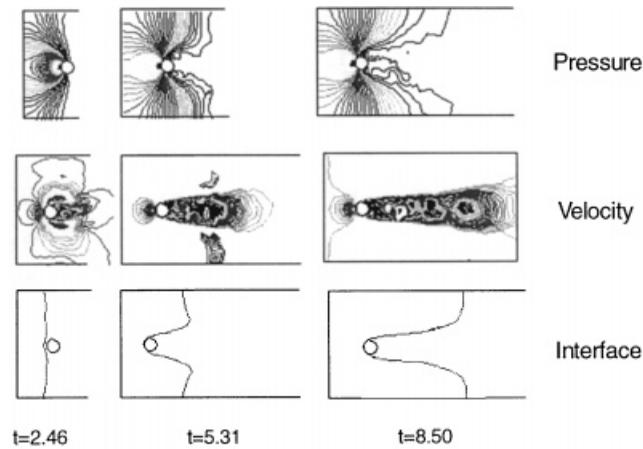
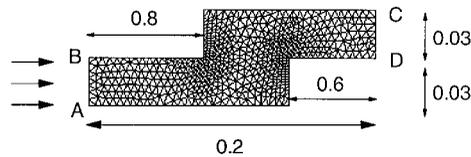Figure 10. Flow around a cylinder: pressure and velocity contours and interface postion evolution.

Figure 11. Step cavity: problem lay-out and mesh of three-node triangles.

A fluid with density 1.0×10^3 and viscosity $\mu = 2.0 \times 10^{-2}$ starts coming into the domain through section AB, with a horizontal velocity of 0.1. The cavity is initially filled with air, with density $\rho = 1.0$ and viscosity $\mu = 2.0 \times 10^{-5}$. A zero pressure datum is set at C and perfect slippage is allowed on the walls, i.e. there is no friction of the fluid with the walls. Normal velocities along BC and A are set to zero. Finally, the test includes the effect of a gravity force with value 9.81.

Figure 12 presents the calculated interface position evolution and the velocity vectors.

5. CONCLUSIONS

This paper presents a method to solve flow problems involving two immiscible fluids within the FEM context.

The proposed procedure discretizes the Navier–Stokes equations and the pseudoconcentration function advection along the characteristics, the characteristics Galerkin method. Additionally, the solution of the Navier–Stokes problem uses a fractional-step method that allows equal order of interpolation for both the velocity and pressure fields. In this way the same mesh of three-node triangles can be used for both problems solution.

Interface position is determined using the level set method and a fast algorithm to preserve accuracy. This procedure shows no problem in handling different interfaces merging and, contrary to existing approaches, it can be easily extended to the solution of 3D problems.
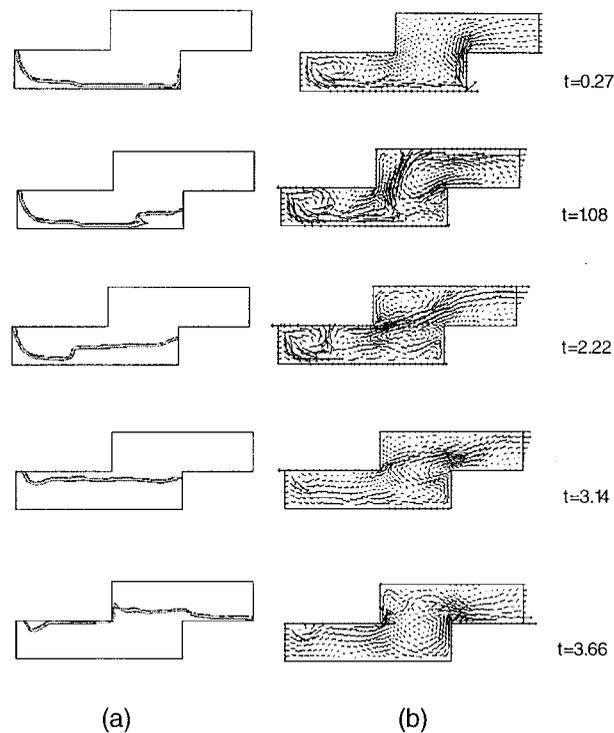
t=0.27

t=1.08

t=2.22

t=3.14

t=3.66

(a)   (b)

Figure 12. Step cavity: (a) interface position evolution and (b) velocity vectors.

REFERENCES

 1. Unverdi S, Tryggvason G. A Front-Tracking method for viscous, incompressible, multi-fluid flows. *Journal of Computational Physics* 1992; **100**:25–37.
 2. Harlow JH, Welch JE. *Physics of Fluids* 1965; **8**:2182.
 3. Harlow JH, Welch JE. Numerical study of large amplitude free surface motion. *Physics of Fluids* 1966; **9**:842–851.
 4. Hirt CW, Nichols BD. Volume of Fluid (VOF) method for the dynamics of free boundaries. *Journal of Computational Physics* 1981; **39**:201–225.
 5. Noh W, Woodward P. Simple line interface calculation. In *Proceedings of the 5th International Conference on Numerical Methods in Fluid Dynamics*, Vooren AI, Zanbergen PJ (eds). Springer: Berlin, 1976; 330.
 6. Chorin A. Flame advection and propagation algorithms. *Journal of Computational Physics* 1980; **35**:1–11.
 7. Dieterlen MR, Maronnier V, Rappaz J. Numerical simulation of free surface flows. In *Computational Mechanics. New Trends and Application*, Idelsohn ES, Dvorkin E (eds). CIMNE: Barcelona, 1998.
 8. Thompson E. Use of pseudo-concentrations to follow creeping viscous flows during transient analysis. *International Journal for Numerical Methods in Fluids* 1986; **6**:749–761.
 9. Lewis RW, Usmani AS, Cross JT. Efficient mould filling simulation in castings by an explicit finite element method. *International Journal for Numerical Methods in Fluids* 1995; **20**:493–506.
10. Dhatt G, Gao DM, Ben Cheikh A. A finite element simulation of metal flow in moulds. *International Journal for Numerical Methods in Engineering* 1990; **30**:821–831.

11. Medale M, Jaeger M. Numerical simulation of incompressible flows with moving interfaces. *International Journal for Numerical Methods in Fluids* 1997; **24**:615–638.
12. Osher S, Sethian J. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulation. *Journal of Computational Physics* 1988; **79**:12–49.
13. Mulder W, Osher S, Sethian J. Computing interface motion in compressible gas dynamics. *Journal of Computational Physics* 1992; **100**:209–228.
14. Sussman M, Smereka P, Osher S. A level set approach for computing solutions to incompressible two-phase flow. *Journal of Computational Physics* 1994; **114**:146–159.
15. Peraire J, Vahdati M, Morgan K, Zienkiewicz O. Adaptive remeshing for compressible flow computations. *Journal of Computational Physics* 1987; **72**:449–466.
16. Hughes TJR, Franca LP, Balestra M. A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska–Brezzi condition: a stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Computer Methods in Applied Mechanics and Engineering* 1986; **59**:85–99.
17. Akin J. *Finite Elements for Analysis and Design*. Academic Press Inc: New York, 1994.
18. Zienkiewicz OC, Codina R. A general algorithm for compressible and incompressible flow—Part I. The split, characteristic-based scheme. *International Journal for Numerical Methods in Fluids* 1995; **20**:869–885.
19. Codina R, Vasquez M, Zienkiewicz OC. A fractional step method for compressible flows: boundary conditions and incompressible limit. In *Proceedings of International Conference On Finite Elements in Fluids—New Trends and Applications*, Venezia, 1995; 409–418.
20. Barth T, Sethian J. Numerical schemes for the Hamilton–Jacobi and level set equations on triangulated domains. *Journal of Computational Physics* 1998; **145**:1–40.
21. Zienkiewicz OC, Taylor RL. *The Finite Element Method* (4th edn), vol. 2. McGraw-Hill: New York, 1991.
22. Codina R. Comparison of some finite element methods for solving the diffusion–convection–reaction equation. *Computer Methods in Applied Mechanics and Engineering* 1998; **156**:185–210.
23. Zienkiewicz OC, Morgan K, Datya Sai BVK, Codina R, Vazquez M. A general algorithm for compressible and incompressible flow—Part II. Tests on the explicit form. *International Journal for Numerical Methods in Fluids* 1995; **20**:887–913.
24. Zienkiewicz OC, Ortiz P. A Split-Characteristic based finite element model for the shallow water equations. *International Journal for Numerical Methods in Fluids* 1995; **20**:1061–1080.
25. Chorin A. Numerical solution of incompressible flow problems. *Studies in Numerical Analysis* 1968; **2**:64–71.
26. Peraire J. A finite element method for convection dominated flows. *Ph.D. Thesis*, University of Wales, Swansea, 1986.
27. Babuska I. The finite element method with Lagrange multipliers. *Numerical Mathematics* 1973; **20**:179–192.
28. Brezzi F. On the existence, uniqueness and approximations of saddle point problems arising from Lagrange multipliers. *RAIRO 8-R2*, 1974; 129–151.
29. Kawahara M, Ohmiya K. Finite element analysis of density flow using velocity correction method. *International Journal for Numerical Methods in Fluids* 1985; **5**:981–993.
30. Schneider GE, Raithby GD, Yovanovich MM. Finite element analysis of incompressible flow incorporating equal order pressure and velocity interpolation. In *Numerical Methods for Laminar and Turbulent Flow*, Taylor C, Morgan K, Brebbia C (eds). Pentech Press: Plymouth, 1978.
31. Quecedo M, Pastor M, Zienkiewicz OC. A fractional step method for non-linear solid dynamics. *Computers and Structures* 2000; **74**:535–545.
32. Hirsch C. *Numerical Computation of Internal and External Flows*, vol. 2. Wiley: New York, 1990.
33. Zienkiewicz OC, Taylor RL. *The Finite Element Method* (4th edn), vol. 1. McGraw-Hill: New York, 1989.

A combined molecular dynamics and finite element method technique applied to laser induced pressure wave propagation

Julia A. Smirnova, Leonid V. Zhigilei, Barbara J. Garrison

Department of Chemistry, 152 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Abstract

Analysis of a variety of dynamic phenomena requires simultaneous resolution at both atomistic and continuum length scales. A combined molecular dynamics and finite element method approach, which we discuss in this paper, allows us to find the balance between the necessary level of detail and computational cost. The combined method is applied to the propagation of a laser-induced pressure wave in a solid. We find good agreement of the wave profile in the molecular dynamics and finite element regions. This computational approach can be useful in cases where a detailed atomic-level analysis is necessary in localized spatially separated regions whereas continuum mechanics and thermodynamics is sufficient in the remainder of the system. © 1999 Elsevier Science B.V. All rights reserved.

PACS: 02.70.Ns; 02.70.Dh; 61.80.Az
Keywords: Multiscale simulation; Molecular dynamics; Finite-element method; Laser ablation

1. Introduction

Over the years a number of approaches have been developed to simulate dynamics in condensed phases. In the atomistic regime, the molecular dynamics (MD) simulation technique has been successfully applied to a variety of phenomena including structures of liquids [1], energetic particle bombardment of solids [2], reactions at surfaces [3], and crack propagation [4,5]. On the other hand, the finite element (FE) method is successful in modeling propagation of elastic waves and heat transfer through material at macroscopic length scales [6,7]. A schism arises, however, when one wants to examine phenomena that occur at an intermediate length regime and yet still retain an atomic-level resolution in some regions of interest. In principle, one could make MD simulations larger but even simulations with $\sim 10^6$–10^8 particles are not sufficiently large to deal efficiently with situations such as propagation of the pressure waves developed in simulations of laser ablation [8], energetic cluster impact [9] or crack propagation [4,5,10].

In particular, the generation of strong pressure waves is a natural result of the fast energy deposition in short pulse laser ablation [8,11,12]. The development and propagation of these waves occur at length scales that are beyond the capability of the MD simulation technique [8,11]. One problem is that a pressure wave reflected from the boundaries of the MD computational cell can interfere with the processes in the ablation region and hinder interpretation of the results of the simulation. Recently we developed a simple and computationally efficient approach for simulating the non-reflecting propagation of a pres-

sure wave out from the MD computational cell [11]. While providing an efficient way to avoid artifacts due to pressure wave reflection, the non-reflecting boundary conditions are not sufficient for more challenging scenarios in which there are several regions where molecular-level analysis is needed. For example, in addition to ablation at the front side of the irradiated sample, the interaction of the laser induced compressive pressure pulse with the back surface of the target can cause the desorption of the molecules adsorbed at the back surface, an effect known as acoustic desorption [13], or lead to the failure process known as back spallation [12]. Moreover, in many heterogeneous systems such as pigmented biological tissues [11,14] or polymer films containing graphitic nanoparticle sensitizers [15], the ablation or damage mechanisms are defined by intensive processes occurring in the immediate vicinity of the spatially localized absorbers embedded in a transparent medium. While the sizes of the systems of interest in these cases are far beyond the capabilities of the MD method, the areas where an atomic or molecular level analysis is necessary can be small enough to be amenable to a treatment by the MD method. The rest of the system, where relative displacements of the atoms are small, can be treated by the use of continuum mechanics and thermodynamics.

A natural approach to the simulation of multiscale processes, thus, is to combine a MD simulation for the critical regions within the system with a FE method for a continuum description of the remainder of the system. There have been a number of works where the FE method is used to simulate an adequate static [10,16,17], and dynamic [10] response of surrounding material to the processes in the MD computational cell. In the present work we demonstrate that application of a combined MD–FE technique can be extended to a multiscale simulation of a system with multiple interacting MD and FE regions.

Here we test this approach on the propagation of a laser induced pressure wave from the ablation region through a micrometer-sized sample using a two-dimensional (2D) model. The extension to three dimensions is straightforward and is currently under development. The computational method is described in Section 2, and the application of the method to the propagation of a pressure wave through the successively arranged MD, FE, and another MD region of

the model is given in Section 3.

2. Computational method

A computational approach for multiscale dynamic simulations that combines molecular and continuum descriptions of different parts of the system is outlined in this section. We give the essence of the MD and FE techniques, show the computational similarities, and discuss a simple prescription for combining the two methods.

In MD simulations a computational cell is represented by a set of N particles with coordinates $\{r_i\}$ and momenta $\{p_i\}$. The time evolution of the system is governed by Newton's second law,

$$m_i \, d^2 r_i / dt^2 = -\nabla_i U(r_1, r_2, \ldots, r_N) \,, \qquad (1)$$

where m_i is the mass of the ith particle, $F_i = -\nabla_i U(r_1, r_2, \ldots, r_N)$ is the force acting on the ith particle due to interaction with other particles in the system, and $U(r_1, r_2, \ldots, r_N)$ is the interaction potential. The initial positions and velocities of the particles together with the interaction potential define the whole set of thermodynamic, elastic and mechanical properties of the model material.

The set of $3N$ second-order differential equations, Eq. (1), is often solved by recasting it as a set of $6N$ first-order Hamilton's equations of motion,

$$dp_i / dt = -\nabla_i U(r_1, r_2, \ldots, r_N) \,,$$
$$dr_i / dt = p_i / m_i \,. \qquad (2)$$

Given the initial positions and momenta of the system, integration of Eq. (2) yields the total trajectory of the system. With a knowledge of the trajectories of all the particles, one can calculate spatial and temporal distributions of energy, temperature and pressure, as well as monitor the structural and phase changes in the system.

In the FE method the continuum system is divided into a finite number of elements which are usually much larger than an individual particle in a MD simulation. Each element is characterized by its geometry, a sequence of points or nodes on its periphery, and by a set of properties of the material. In particular, triangular plane-stress elements, used in 2D simulations discussed in the next section, are shown in Fig. 1. The
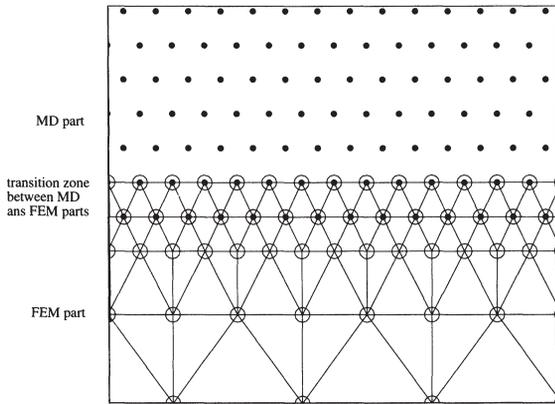
Fig. 1. Transition zone between the MD and FE regions of the system. Black dots represent MD particles. Open circles represent FE nodes. Open circles with black dots inside represent node-particles in the transition layer.

dynamics of a non-dissipative medium is defined, in the linear approximation, by the following equation for the displacements, $\boldsymbol{a}$, of a set of n nodes,

$$\mathbf{M}\,d^2\boldsymbol{a}/dt^2 = -\mathbf{K}\boldsymbol{a} + \boldsymbol{F}_{\mathrm{ex}}\,. \qquad (3)$$

In this system of n coupled second-order differential equations, $d^2\boldsymbol{a}/dt^2$, is a vector of nodal accelerations, $\mathbf{M}$ is a mass matrix defined by the geometry of the elements and the density of material, $\mathbf{K}$ is the stiffness matrix that is defined by the geometry of the elements and elastic moduli of material, and $\boldsymbol{F}_{\mathrm{ex}}$ is the external force applied to the nodes. The techniques for construction of a FE mesh and calculation of mass and stiffness matrices appropriate for the simulation of a particular system are covered in an extensive literature, see for example Ref. [18].

A dynamic simulation with both MD and FE techniques involves integration of the equations of motion, Eqs. (1) and (3). An apparent similarity of Eqs. (1) and (3) suggests the possibility for coupling of the two dynamic simulation techniques. From one perspective, if the interaction potential used in the MD method is assumed to be harmonic, then $\nabla_i U$ becomes a linear function of displacement as in Eq. (3). From the other side, the stiffness matrix, $\mathbf{K}$, is defined by the elastic constants of the system [18] and calculation of the elastic constants from a given functional form of the interaction potential is straightforward [10]. Thermal effects, damping, and nonlinearity can be introduced into the finite-element analysis [6,7,10,18] providing

a more accurate match with the properties of MD system defined by a realistic interaction potential.

In order to effectively combine the regions described by the MD and FE methods into a single model, one not only has to ensure consistency between the properties of the discrete and continuum media, but also to provide a smooth transition between the two media. In the present work the coupling of the two descriptions of the media is brought about by a transition zone in which the FE nodes coincide with the positions of the particles in the MD region, Fig. 1. The width of the transition zone is equal to the cutoff distance of the interaction potential used in the MD region, two layers of particles in this case. This provides a complete set of neighbors within the interaction range for all particles in the MD region. Particles that belong to the transition zone interact via the interaction potential with the MD region. At the same time the transition zone constitutes a part of the FE grid, where the nodes coincide with the MD particles, and experience the nodal forces due to the FE grid. The forces exerted on the particle-nodes in the transition zone due to the interaction with the MD region make up the external forces $\boldsymbol{F}_{\mathrm{ex}}$ in the equations of motion for the FE nodes, Eq. (3). $\boldsymbol{F}_{\mathrm{ex}}$ is nonzero only in the transition zone between the MD and continuum regions. In order to avoid a density mismatch at the boundary, the mass at each node in the transition zone is set equal to the mass of the MD particle. Both Eqs. (1) and (3) are solved using the same integration scheme, which increases the stability in the transition region.

3. Application to laser ablation

In this section we illustrate the computational efficiency and accuracy of the combined MD and FE approach. The method is applied to the multiscale simulation of laser ablation of a molecular solid and propagation of a laser induced pressure wave from the ablation region through a micrometer-sized sample. The schematic view of the model consisting of the successively arranged MD, FE, and another MD region is shown in Fig. 2.

A relatively small surface region, part A in Fig. 2, where complex processes of laser energy deposition, overheating, buildup of high pressure, disintegration
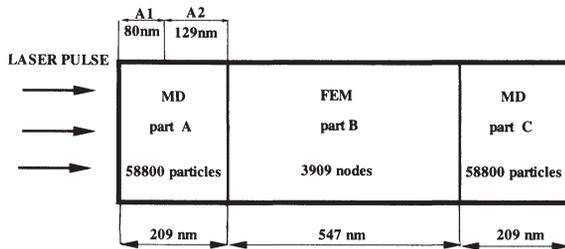
Fig. 2. Schematic picture of the model system.

and ejection of a significant amount of material are occurring, does require a molecular level analysis and is simulated by the MD method. A breathing sphere model for MD simulations of laser ablation [8] has significantly expanded the time and length scales accessible for this molecular level simulation and has provided insight into the microscopic mechanisms of laser ablation and damage of organic solids [8,11,19,20]. In this work we use a 2D version of the breathing sphere model, described in detail in Ref. [8], for simulation of the surface region of the irradiated sample.

The high pressure associated with the fast energy deposition and ablative recoil lead to the development of a pressure wave that propagates deeper into the sample [8]. The long-range propagation of the pressure wave is simulated using the FE method, part B in Fig. 2. In this work the FE part of the system is described within the linear approximation with triangular elements [18] under plane stress conditions [21]. Generally, complete compatibility between the MD and FE regions requires implementation of nonlinear elasticity and inclusion of anisotropy of the material in the FE method [10]. Introduction of nonlinearity into the FE method involves an adjustment of the stiffness matrix at each integration step whereas anisotropy increases significantly the number of nonzero elements in the stiffness matrix [6,7,22]. Although neither of these factors make an apparent problem, the stiffness matrix in this simple test case could, in principle, be ~ 60 million double precision numbers. We have thus made efforts to use only the nonzero elements to save computer memory. Thus for this first test case, we chose to use an isotropic and linear stiffness matrix. In order to partially account for nonlinearity within our linear isotropic FE continuum, we calculate the stiffness matrix based on an effective elastic modulus obtained from the average velocity of the pressure wave

propagation in the MD region. This approximation allows us to decrease energy reflection at the boundary between the MD and FE parts as compared to the simulation with an elastic modulus obtained from the interaction potential [10,21]. The value of the effective 2D Young's modulus used in the stiffness matrix is 0.27 eV/Å^2, 22% higher than the one obtained from the interaction potential. A value of $1/3$ is used for Poisson's ratio of a two-dimensional isotropic material [21]. A regular triangular mesh with 145 rows of elements along the direction of the pressure wave propagation is used to represent the continuum region. The internode distance varies from 0.58 nm in the transition layer, which corresponds to the interparticle distance in the MD region (Fig. 1), to 4 nm in the bulk. Of the total number of 3909 finite-element nodes about one third are in the vicinity of the MD regions in order to provide a smooth transition from the small elements in the transition region to the larger elements in the middle of the continuum region.

The third part of the system, marked as part C in Fig. 2, is a region at the back of the sample. The interaction of the laser induced pressure wave with the back surface can cause a mechanical damage in the surface region [12] or lead to the desorption of the adsorbed molecules [13]. In order to study the microscopic mechanisms of the damage and desorption in part C, we have to switch back from the continuum FE method in part B to a molecular level MD method.

Thus, the complete computational cell (Fig. 2) contains 117600 MD particles in two MD regions, a finite-element mesh with 3909 nodes, and 560 particle-nodes in two transition zones. The size of the system is 81 by 965 nm. Periodic boundary conditions in the direction parallel to the surface are imposed, thus the effects of the edges of the laser beam are neglected and a plane pressure wave propagates from the ablation region. The laser penetration depth is 32 nm and the laser pulse duration is 15 ps. The MD regions consume most ($\sim 98\%$) of the computer time in the simulation.

The computational setup described above is used to simulate the propagation of a compressive pressure wave within the irradiated sample. Fig. 3 shows the pressure profile at different times following irradiation with a 15 ps laser pulse. The pressure wave generated in the ablation region reaches the first transition zone approximately 60 ps after the start of the laser pulse. From 60 ps to 245 ps the wave moves through the FE
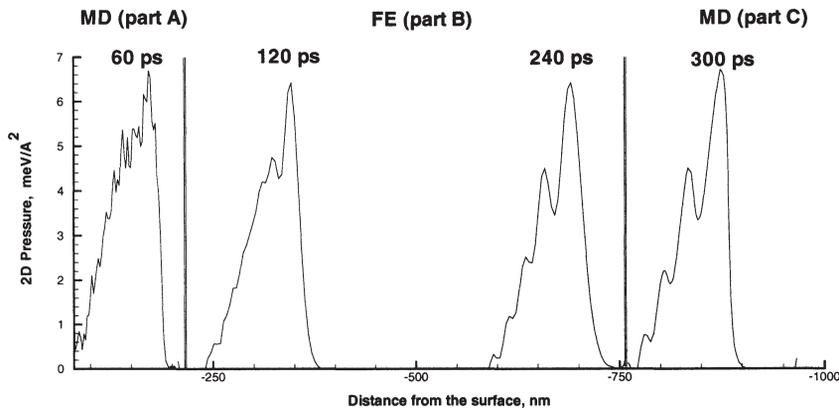
Fig. 3. The propagation of the pressure wave through the different regions of the model sample. Pressure profiles are shown for four different times after the start of the 15 ps laser pulse.

region and by 245 ps it reaches the second MD region. We observe a good agreement in the wave profile shape in all three regions of the model (parts A, B, and C). There are, however, a few changes in the wave profile. First, the wave front becomes less steep after passing through the first transition region. Second, there is an apparent smoothing of the wave profile as it has propagated to the FE region. Both of these differences are ramifications of having a molecular resolution in the MD region and a more course grid in the FE region. The high spatial resolution of the pressure wave simply cannot be described by the FE grid. Third, while the pressure wave propagates through the FE region, additional peaks spaced by 10–20 nm appear on the wave profile. The spacings between the peaks correspond to the characteristic frequencies of the FE grid. It appears that the sharp front of the pressure wave has caused a ringing in the FE grid. It is, of course, desirable to eliminate these differences. To accomplish this, however, would demand that the FE method resolution be the same as in the MD region. The wave propagates through the second transition zone between the FE and MD regions (parts B and C) without changes because in this case the wave is going from a course grid to a finer MD resolution.

In order to test additionally the combined MD–FE approach, we performed a large-scale MD simulation with a 310 nm long computational cell consisting of 86800 particles. Propagation of the laser induced pressure wave from the ablation region to the back side of this computational cell takes about 100 ps. The pressure distributions at 100 ps in the combined MD–FE model and in the pure MD model are shown in Fig. 4. Except for differences due to the smoothing in going from the fine MD resolution to the course FE grid, both profiles have the same amplitude and shape. Comparisons of the energy in these two models indicate that only $\sim 5\%$ of the pressure wave energy is reflected at the transition zone.

4. Conclusion

A computational technique based on a combination of MD and FE methods has been implemented and tested on the propagation of a pressure wave induced by laser irradiation of an organic solid. Good agreement of both the total energy of the wave and the wave profile in MD and FE parts is observed in the simulation. The real strength of this combined approach is that a pressure wave can be transported over micron dimensions without losing the essential characteristics of the wave profile. Certainly improvements can be made. For example, nonlinear elasticity, anisotropy, and heat transfer can be included in the FE method. For our application, however, it is doubtful that any of these changes would improve the agreement between the pressure profiles shown in Fig. 4. An improvement could be achieved by decreasing the grid spacing in the FE region. This would lead to an increased memory requirement for the multiplication of matrices in the FE calculation as well as an increase in computer time. Ultimately each specific application determines
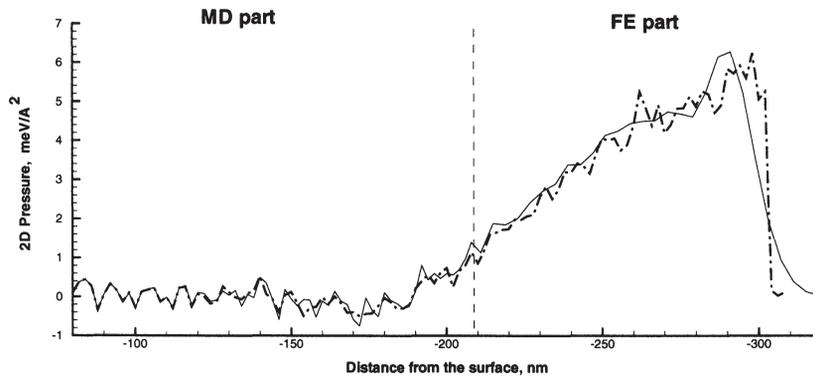
Fig. 4. The pressure wave profile in the MD–FE system and the reference pure MD system at 100 ps. The solid line is for the combined MD–FE system and the dash-dotted line is for the pure MD system.

its own tolerable accuracy and acceptable level of distortions caused by a more course space resolution in the FE method.

In summary, the combined MD–FE technique allows one to balance the level of details necessary to provide reasonable accuracy in some regions of the model with computational cost. In the field of laser ablation this approach can readily be applied to study back spallation, acoustic desorption, or laser ablation/damage of heterogeneous systems with spatially localized absorbers.

Acknowledgements

References

[1] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids (Clarendon Press, Oxford, 1987).

[2] R.S. Taylor, B.J. Garrison, Langmuir 11 (1995) 1220.

[3] B.J. Garrison, K.B.S. Prasad, D. Srivastava, Chem. Rev. 96 (1996) 1327.

[4] F.F. Abraham, Europhys. Lett. 38 (1997) 103.

[5] P.S. Lomdahl, R. Thompson, B.L. Holian, Phys. Rev. Lett. 76 (1996) 2318.

[6] J.H. Argyris, H.-P. Mlejnek, Dynamics of Structure

[7] T. Belytschko, T.J.R. Hughes, eds., Computational Methods for Transient Analysis (Elsevier, Amsterdam, 1983).

[8] L.V. Zhigilei, P.B.S. Kodali, B.J. Garrison, J. Phys. Chem. B 101 (1997) 2028; Chem. Phys. Lett. 276 (1997) 269; J. Phys. Chem. B 102 (1998) 2845.

[9] M. Moseler, J. Nordiek, H. Haberland, Phys. Rev. B 56 (1997) 15439.

[10] S. Kohlhoff, P. Gumbsh, H.F. Fischmeister, Phil. Mag. A 64 (1991) 851.

[11] L.V. Zhigilei, B.J. Garrison, in: Laser-Tissue Interaction IX, S.L. Jacques, ed., Proc. SPIE 3254 (1998) 135.

[12] I. Gilath, in: High-Pressure Shock Compression of Solids II, L. Davison, D.E. Grady, M. Shahinpoor, eds. (Springer, New York, 1996) p. 90.

[13] V.V. Golovlev, S.L. Allman, W.R. Garrett, N.I. Taranenko, C.H. Chen, Int. J. Mass Spec. Ion Process. 169/170 (1997) 69.

[14] S.L. Jacques, A.A. Oraevsky, R. Thompson, B.S. Gerstman, in: Laser-Tissue Interaction IX, S.L. Jacques, ed., Proc. SPIE 2134A (1994) 54.

[15] X. Wen, D.E. Hare, D.D. Dlott, Appl. Phys. Lett. 64 (1994) 184.

[16] M. Mullins, M.A. Dokainish, Phil. Mag. A 46 (1982) 771.

[17] E.B. Tadmor, M. Ortiz, R. Phillips, Phil. Mag. A 73 (1996) 1529.

[18] O.C. Zienkiewicz, R.L. Taylor, The Finite Element Method (McGraw-Hill, London, 1989).

[19] L.V. Zhigilei, B.J. Garrison, Appl. Surf. Sci. 127–129 (1998) 142.

[20] L.V. Zhigilei, B.J. Garrison, Appl. Phys. Lett. 71 (1997) 551; Rapid Commun. Mass Spectrom. 12 (1998) 1273.

[21] P.-O. Esbjørn, E.J. Jensen, J. Phys. Chem. Solids 37 (1976) 1081.

[22] J.M. McGlaun, P. Yarrington, in: High-Pressure Shock Compression of Solids, J.R. Asay, M. Shahinpoor, eds. (Springer, New York, 1993) p. 323.

(Elsevier, Amsterdam, 1991).

[7:3] Alternative Solution Models

Boundary Element Methods

11. Boundary Element Methods – Introduction

BOUNDARY ELEMENT METHOD - INDIRECT

Kernel functions:
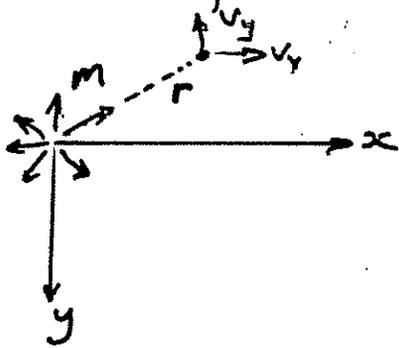
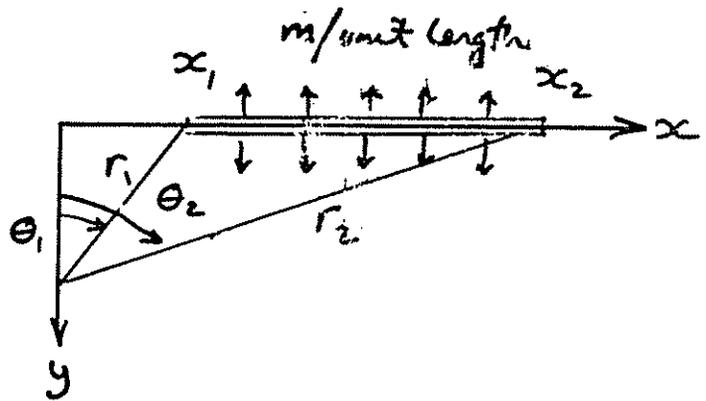$$h = \frac{m}{2\pi} \ln(r)$$

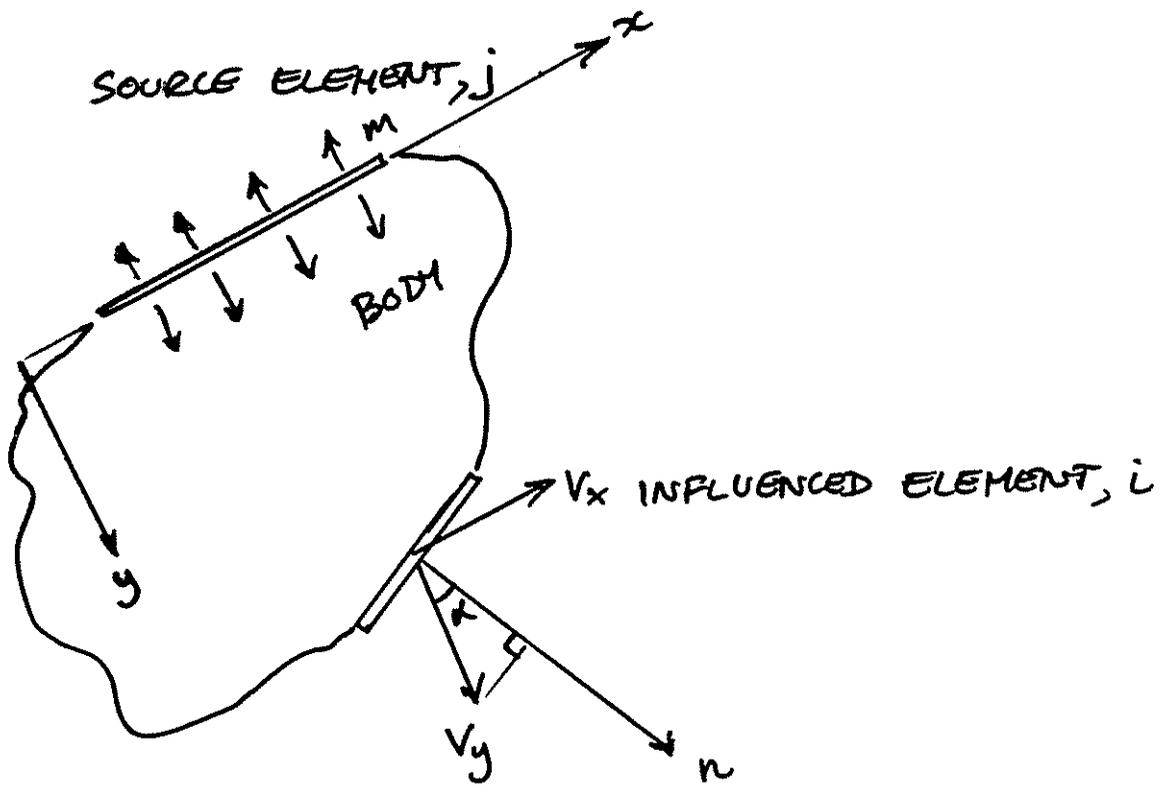$$V_x = -K\frac{\partial h}{\partial x} = -\frac{Km}{2\pi}\left(\frac{x}{r^2}\right)$$

$$V_y = -K\frac{\partial h}{\partial y} = -\frac{Km}{2\pi}\left(\frac{y}{r^2}\right)$$

Integrated over line segment:

$$h = \frac{m}{2\pi}\left[x\ln(r) - x + y\theta\right]_1^2 = B_{ij}$$

$$V_x = -K\frac{\partial h}{\partial x} = -\frac{Km}{2\pi}\left[\ln(r)\right]_1^2 \left.\vphantom{\frac{Km}{2\pi}}\right\}$$

$$V_y = -K\frac{\partial h}{\partial y} = -\frac{Km}{2\pi}\left[\theta\right]_1^2 \quad\Bigg\} \; A_{ij}$$

SOURCE ELEMENT, j

x

m

BODY

V_x INFLUENCED ELEMENT, i

y

V_y

n

$$V \cdot n = V_x \sin\alpha + V_y \cos\alpha$$

$$h_i = \sum_{j=1}^{M} B_{ij} \, m_j$$

$$(V \cdot n)_i = \sum_{j=1}^{M} A_{ij} \, m_j$$

$\left.\begin{array}{c} \\ \\ \end{array}\right\}$ System equations
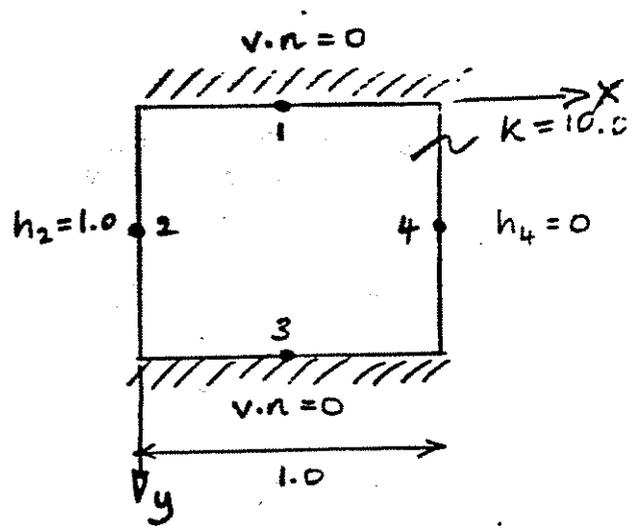
Note $B_{ij} =$ influence of source at
j on head at i

$A_{ij} =$ influence of source at
j on velocity at i

PROCEDURE

1. Write

$$h_i = \sum_{j=1}^{M} B_{ij} m_j \qquad \text{if head b.c. prescribed @ } i$$

$$(v \cdot n)_i = \sum_{j=1}^{M} A_{ij} m_j \qquad \text{if flux b.c. prescribed @ } i.$$

2. A total of M equations result where $(v \cdot n)_i$ and h_i are known as boundary conditions and A_{ij}, B_{ij} kernels are functions of geometry and K, only.

3. Solve for fictitious source strengths, $m_j \qquad j = 1, M$.

4. Determine complementary equations to determine complementary boundary condition

$$(v \cdot n)_i = \sum_{j=1}^{M} A_{ij} m_j \qquad \text{if head b.c. prescribed}$$

$$h_i = \sum_{j=1}^{M} B_{ij} m_j \qquad \text{if flux b.c. prescribed}$$

5. Determine heads and velocities if desired at internal points with new A_{ij}, B_{ij} matrices.

Example 6.1.1 Indirect Method

Consider the square body
shown, and discretized
crudely by only four
elements.

The induced to the boundary velocity at node 1 may be
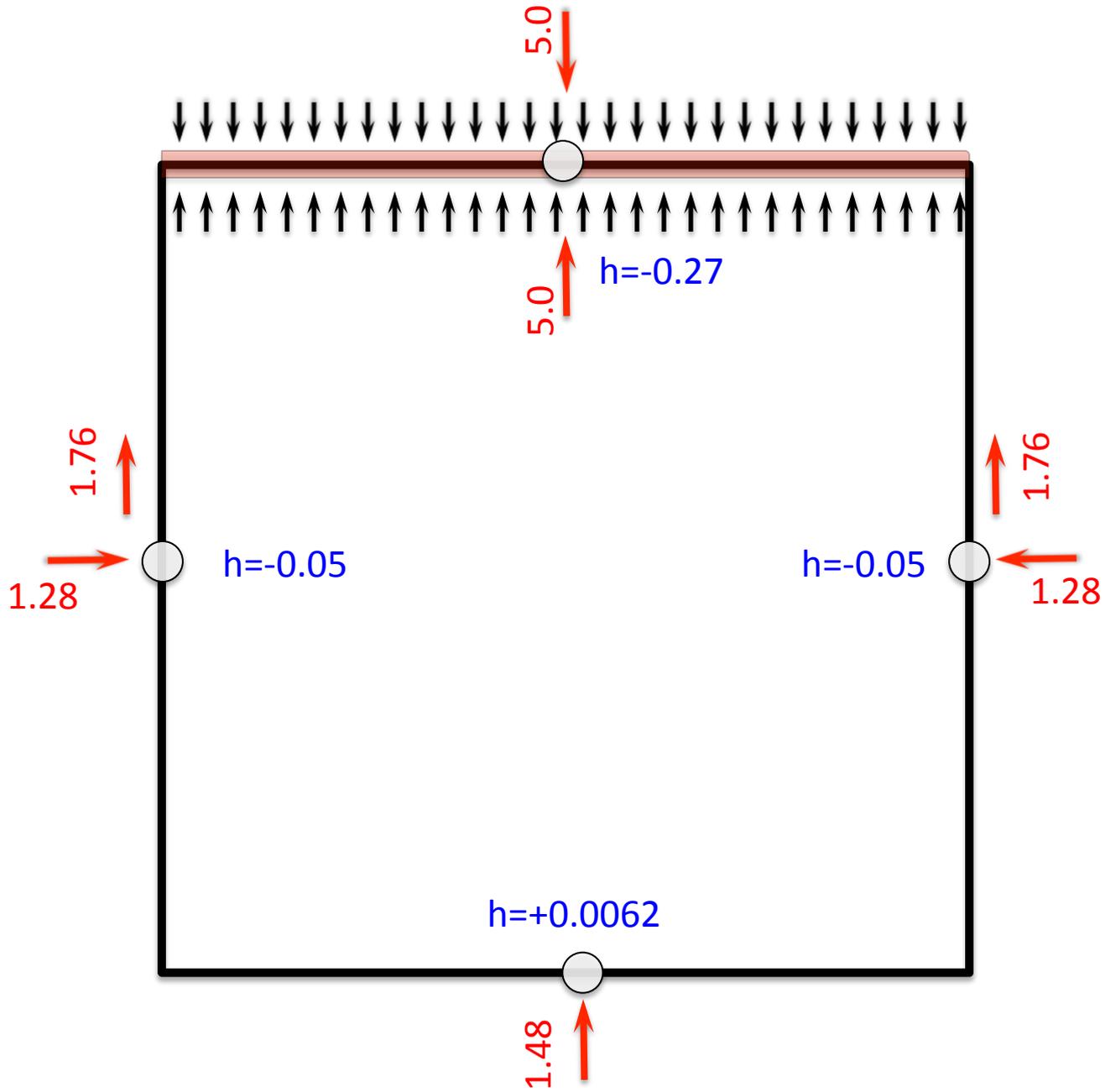written in terms of the source strengths at all other elements.

$$(v \cdot n)_1 = \sum_{j=1}^{4} A_{ij} \, m_j$$
$$h_2 = \sum B_{ij} \, m_j$$
$$(v \cdot n)_3 = \sum A_{ij} \, m_j$$
$$h_4 = \sum B_{ij} \, m_j$$

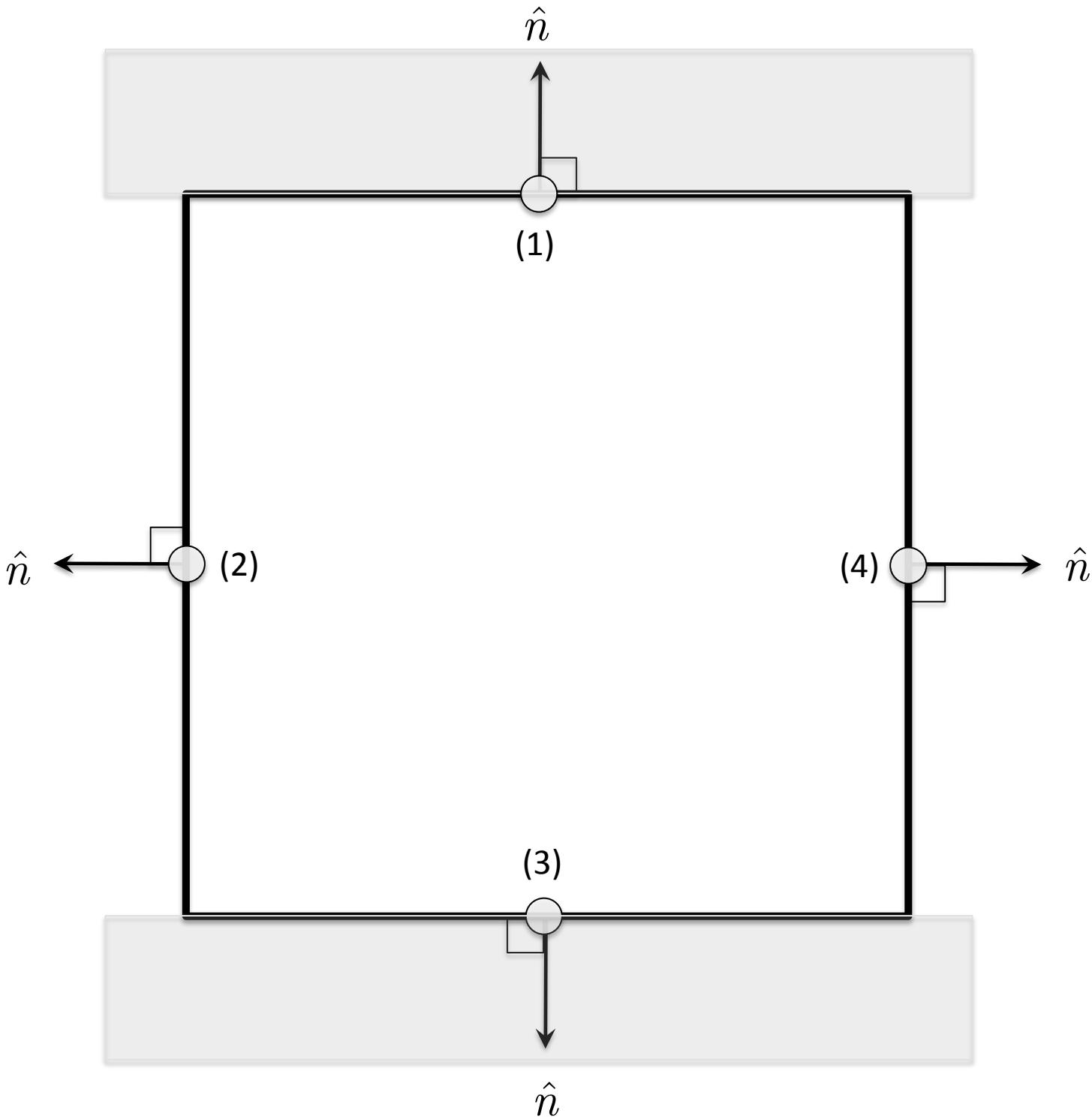From the symmetry of the problem, the influence of element
1 on all others fully defines the induced velocities and heads
for all other possible sources. These are tabulated below.

Source	Effect	h	v_x	v_y
1	1	−.2695	0	−5.0
	2	−.0533	+1.281	−1.762
	3	.0062	0	−1.476
	4	−.0533	−1.281	−1.762
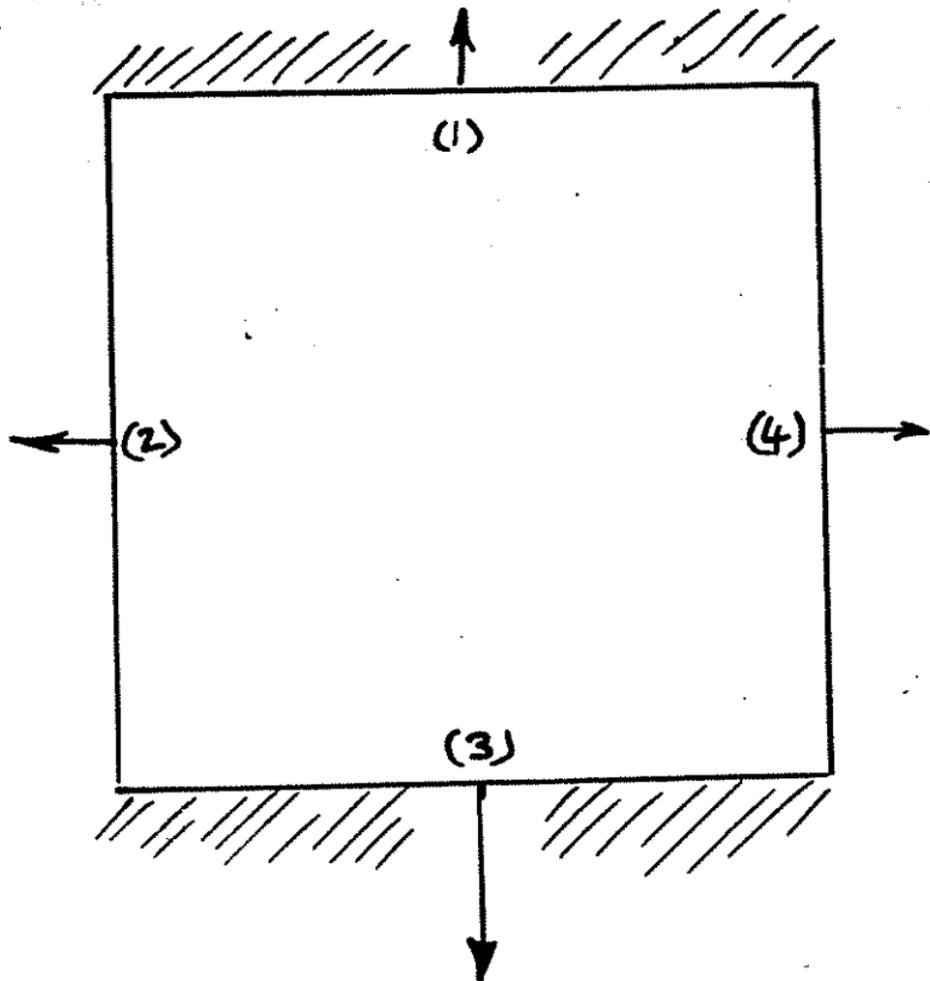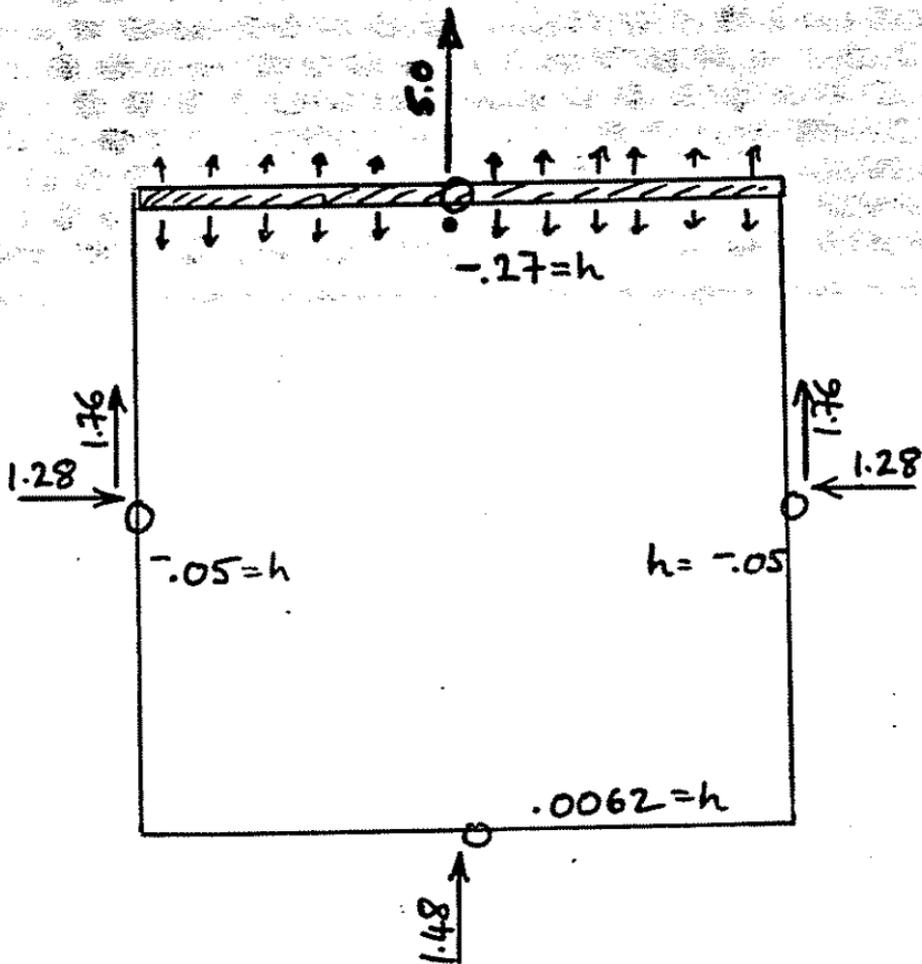
From this the full matrix of coefficients may be evaluated, in $A1$ and solve for the source strengths $M1$ required to give the prescribed boundary conditions BC

$$BC = A1 \, M1$$

With $M1$ evaluated, these magnitudes are substituted into $A2$ to give the element unknowns EU as

$$EU = A2 \, M1$$

and give the resulting

$$
\begin{Bmatrix}
h_1 \\
(v \cdot n)_2 \\
h_3 \\
(v \cdot n)_4
\end{Bmatrix}
=
\begin{Bmatrix}
-436 \\
-11.416 \\
.436 \\
-9.296
\end{Bmatrix}
\qquad
\begin{matrix}
-14.16 \\
\\
9.29
\end{matrix}
$$

representing the unknown magnitudes.

The crude elements chosen to represent the behavior of the block results in the poor evaluation of the complementary boundary conditions, above. These magnitudes are reasonable considering the crudity of the representation.

Source	Effect	h	V_r	V_z
1	1	−.2695	0	−5.0
	2	−.0533	+1.281	−1.762
	3	.0062	0	−1.476
	4	−.0533	−1.281	−1.762
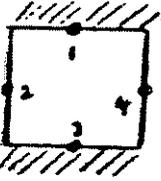
$K = 10.0$

Side dimension $= 1.0$

$$\underline{BC} = \begin{Bmatrix} V.n_1 \\ h_2 \\ V.n_3 \\ h_4 \end{Bmatrix}$$

MATRIX OF INFLUENCE COEFFICIENTS A1

	1	2	3	4
1	0.5000000E+01	−0.1281000E+01	−0.1476000E+01	−0.1281000E+01
2	−0.5300000E−01	−0.2700000E+00	−0.5300000E−01	0.6000001E−02
3	−0.1476000E+01	−0.1281000E+01	0.5000000E+01	−0.1281000E+01
4	−0.5300000E−01	0.6000001E−02	−0.5300000E−01	−0.2700000E+00

MATRIX OF INFLUENCE COEFFICIENTS A2

	1	2	3	4
1	−0.2700000E+00	−0.5300000E−01	0.6000001E−02	−0.5300000E−01
2	−0.1281000E+01	0.5000000E+01	−0.1281000E+01	−0.5300000E−01
3	0.6000001E−02	−0.5300000E−01	−0.2700000E+00	−0.1476000E+01
4	−0.1281000E+01	−0.1476000E+01	−0.1281000E+01	0.5000000E+01

VECTOR OF FICTICIOUS SINK STRENGTHS M1

	1
1	−0.1065805E+01
2	−0.3277597E+01
3	−0.1065805E+01
4	0.3455919E+00

VECTOR OF ELEMENT UNKNOWNS V*N

	1
1	0.4367689E+00
2	−0.1416748E+02
3	0.4367689E+00
4	0.9296286E+01

Solution process

1. $\{B.C.$ for all elements$\} = [A1] \{M1\}$

2. Solve for $[M1]$

3. $\{$Element unknowns$\} = [A2] \{M1\}$

4. Solve for $\{$Element unknowns$\}$

BOUNDARY ELEMENT METHOD — DIRECT
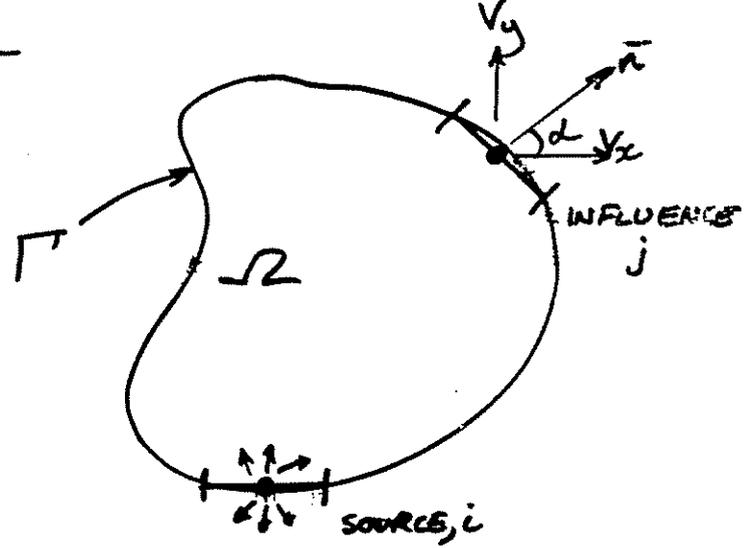
Boundary constraint equation:

$$c_{ij}(p)\, h_j(p) + \int_\Gamma V_{ij}(p,q)\, h_j(q)\, d\Gamma = \int_\Gamma H_{ij}(p,q)\, \underline{V}_j(q) \cdot \bar{n}\, d\Gamma \tag{7}$$

$\int_\Gamma V_{ij}\, d\Gamma$ is integrated effect of a unit source at element i
on the resulting normal flux at boundary element j.

$\int_\Gamma H_{ij}\, d\Gamma$ is integrated effect of a unit source at element i
on the resulting head at boundary element j.

$$(v \cdot n) = [V_x \,;\, V_y]_j \begin{Bmatrix} \cos\alpha \\ \sin\alpha \end{Bmatrix}$$

$c_{ij}(p)$ is the free term, caused by bringing the source
to the boundary

$$c_{ij}(p) = \delta_{ij} = 1$$

$$c_{ij}(p) = \tfrac{1}{2}\,\delta_{ij} = \tfrac{1}{2}$$

PROCEDURE

1. Apply unit source/sink at the centre of a single element. Evaluate the ordered heads at elements j and normal velocities at elements j.

$$h(q)_j = \frac{m}{2\pi} \ln(r) \qquad ; \qquad V(q)_{j_r} = \frac{-K}{2\pi r}$$

2. If h_j and $(V \cdot \bar{n})_j$ are evaluated for each element j then we have H_{ij} and V_{ij} for $j = 1, N$ for $i = $ fixed (at node where source is applied) One equation (1) will result.

3. Repeat (2.) by moving the source/sink to each element in turn and evaluating the sums V_{ij}, H_{ij} for all i, $i = 1, N$

   This results in N equations

4. All integrals may be evaluated easily by analytical or numerical means, except where $i = j$

   For $\int H_{ii} d\Gamma$ use logarithmic quadrature (Stroud and Secrest)

   For $\int V_{ii} d\Gamma$ set all $h_j = 1$ $j = 1, n$ then by definition $V_{jj} \cdot n = 0$ and equation (1) reduces to.

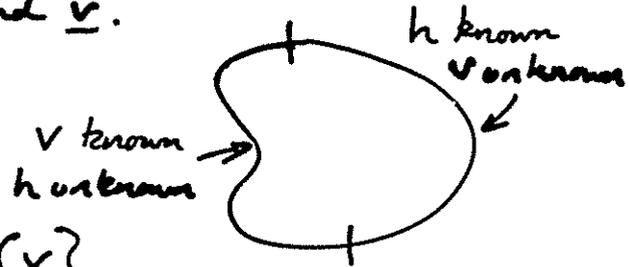$$C_{ij}(p) \; h_j(p) \Big|^{1.0} = -\sum_{\substack{j=1 \\ j \neq i}}^{N} \int V_{ij}(p, q) \; h_j(q) \Big|^{1.0} d\Gamma$$

5. Assemble matrix identity $\quad \underline{V}\,\underline{h} = \underline{H}\,\underline{v}$

$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \underset{n\times n}{} \underset{n\times 1}{} \quad \underset{n\times n}{} \underset{n\times 1}{}$

where only $\quad m$ values of head are known $\quad (m \leq n)$

$\quad\quad\quad\quad\quad n-m$ values of flux $(v-n)$ are known.

Partition matrices for known $\underline{h}$ and $\underline{v}$.

relative to unknown $\underline{h}$ and $\underline{v}$.

h known
v unknown

v known
h unknown

$$\begin{bmatrix} \underline{V}_{11} & \underline{V}_{12} \\ \underline{V}_{21} & \underline{V}_{22} \end{bmatrix} \begin{Bmatrix} \underline{h} \\ \underline{h} \end{Bmatrix} = \begin{bmatrix} \underline{H}_{11} & \underline{H}_{12} \\ \underline{H}_{21} & \underline{H}_{22} \end{bmatrix} \begin{Bmatrix} \underline{v} \\ \underline{v} \end{Bmatrix}$$

Rearrange as:

$$\begin{bmatrix} \underline{V}_{11} & -\underline{H}_{12} \\ \underline{V}_{21} & -\underline{H}_{22} \end{bmatrix} \begin{Bmatrix} \underline{h} \\ \underline{v} \end{Bmatrix} = \begin{bmatrix} \underline{H}_{11} & -\underline{V}_{12} \\ \underline{H}_{21} & -\underline{V}_{22} \end{bmatrix} \begin{Bmatrix} \underline{v} \\ \underline{h} \end{Bmatrix}$$

Solve as

$$\begin{Bmatrix} \underline{v} \\ \underline{h} \end{Bmatrix} = \begin{bmatrix} \underline{H}_{11} & -\underline{V}_{12} \\ \underline{H}_{21} & -\underline{V}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \underline{V}_{11} & -\underline{H}_{12} \\ \underline{V}_{21} & -\underline{H}_{21} \end{bmatrix} \begin{Bmatrix} \underline{h} \\ \underline{v} \end{Bmatrix}$$
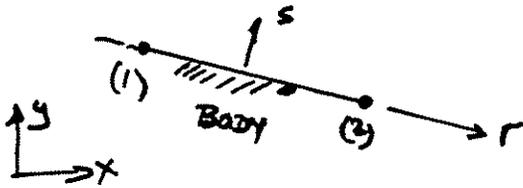
and determine $\underline{v}, \underline{h}$.

6. Evaluate internal conditions, h, by applying (1)
   at internal points.

BEM — ISOPARAMETRIC

- 1. Accurate representation of curved boundaries
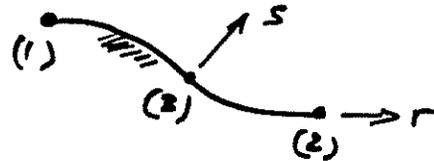- 2. Accurate representation of potential variation / velocity variation.

2 - Node

$$b_1 = \tfrac{1}{2}(1-r)$$
$$b_2 = \tfrac{1}{2}(1+r)$$

$$\underline{b} = [\, b_1 \,;\, b_2 \,]$$

3 - Node

$$b_1 = \tfrac{1}{2}(1-r) - \tfrac{1}{2}(1-r^2)$$
$$b_2 = \tfrac{1}{2}(1+r) - \tfrac{1}{2}(1-r^2)$$
$$b_3 = (1-r^2)$$

$$\underline{b} = [\, b_1 \,;\, b_2 \,;\, b_3 \,]$$

Parametric mapping

$$x = \underline{b}\,\underline{x}$$
$$y = \underline{b}\,\underline{y}$$
$$h = \underline{b}\,\underline{h}$$
$$(v \cdot n) = \underline{b}\,\underline{(v \cdot n)}$$

Mapping

$$d\Gamma = |\underline{J}|\, dr$$
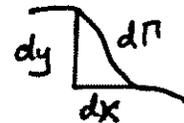
$$\frac{d\Gamma}{dr} = |\underline{J}| = \frac{L}{2}$$

$$\frac{d\Gamma}{dr} = \sqrt{\left(\frac{dx}{dr}\right)^2 + \left(\frac{dy}{dr}\right)^2}$$

Pythagoras.

and $\quad \dfrac{dx}{dr} = \dfrac{d}{dr}(x) = \dfrac{d}{dr}(\underline{b}\underline{x}) = \dfrac{d\underline{b}}{dr}\underline{z}$
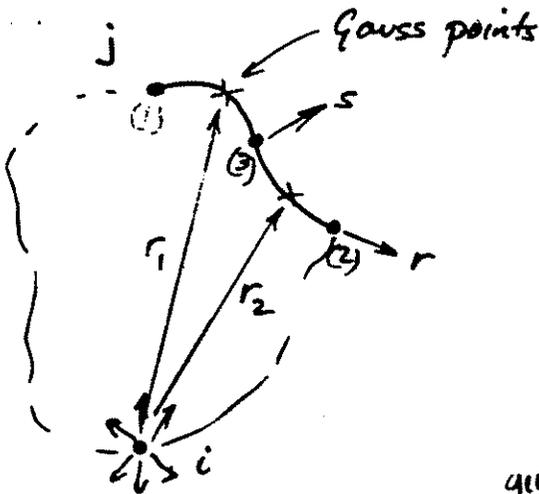
Substituting these concepts into the boundary constraint equation

$$c(p)\,h(p) + \sum_{j=1}^{N}\int_{-1}^{+1} V(p,q)\,h(q)\,\dfrac{d\Pi}{dr}\,dr = \sum_{j=1}^{N}\int_{-1}^{+1} H(p,q)\,(V(q)\cdot\hat{n})\,\dfrac{d\Pi}{dr}\,dr$$

$$h(q) = \underline{b}\,\underline{h} \qquad\qquad\qquad (V(q)\cdot n) = \underline{b}\,(V\cdot n)$$

Example evaluation of coefficient $H(p,q)$ for $\dfrac{\text{element}}{\text{node}}$ j.

$$H(p,q) = \left[\int_{-1}^{+1} \dfrac{1}{2\pi}\,\overset{h(q)}{\overbrace{\ln(r)}}\,\underline{b}\,\dfrac{d\Pi}{dr}\,dr \right]$$

Gauss points

gives $\qquad [\,H_{i1}\,;\,H_{i2}\,;\,H_{i3}\,]\begin{Bmatrix} V_1 \\ V_2 \\ V_3 \end{Bmatrix}$

<u>GLOBAL MATRIX</u> $\qquad\qquad j = 1.2.3$

$$i\left[\,\cdots\cdots\;\underline{H_{i_1}\,H_{i_2}\,H_{i_3}}\,\right]\begin{Bmatrix} \underset{=}{V_1} \\ V_2 \\ V_3 \\ = \end{Bmatrix}$$

COUPLED BEM - FEM

BEM System
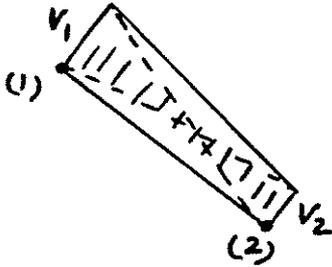
$$q = v \ (\text{length})$$

$$q_1 = v_1 \int b_1 \frac{d\Pi}{dr} dr \quad ; \quad q_2 = v_2 \int b_2 \frac{d\Pi}{dr} dr$$

$$\text{or} \quad \underline{q} = \underline{v}\,\underline{l}$$

BEM Equation

$$\underline{H}\,\underline{v} = \underline{V}\,\underline{h}$$

$$\underline{v} = \underline{H}^{-1}\underline{V}\,\underline{h}$$

$$\underline{l}\,\underline{v} = \underline{l}\,\underline{H}^{-1}\underline{V}\,\underline{h}$$

$$\underline{q} = \underbrace{\underline{l}\,\underline{H}^{-1}\underline{V}}_{\underline{K}}\,\underline{h}$$

FEM Equations

$$\underline{q} = \underline{K}\,\underline{h}$$

Coupled

$$\left\{ \frac{q}{q} \right\} = \begin{bmatrix} \underline{K} & 0 \\ 0 & \underline{K} \end{bmatrix} \left\{ \frac{\underline{h}}{\underline{h}} \right\}$$

TEACHBE3.DAT — BEM SUPERELEMENT

GLOBAL — LOCAL

COUPLED BE-FE GEOMETRY

(BRACKETED NOS. ARE LOCAL DEGREES OF FREEDOM.)

EXAMPLE

$V_2 = 0 \quad h_2 = ?$

h_1
$V_1 = ?$

$h_1 = $ known
$V_1 = ?$

$V_2 = 0$
$h_2 = ?$
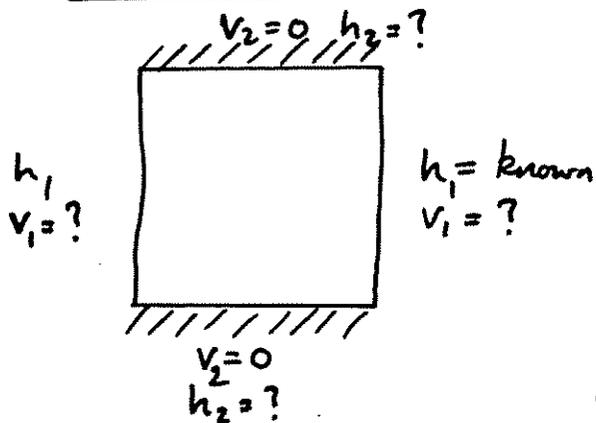
BEM equation

$$\begin{bmatrix} \underline{H}_{11} & \underline{H}_{12} \\ \underline{H}_{21} & \underline{H}_{22} \end{bmatrix} \begin{Bmatrix} \underline{V}_1 \\ \underline{V}_2 \end{Bmatrix} = \begin{bmatrix} \underline{V}_{11} & \underline{V}_{12} \\ \underline{V}_{21} & \underline{V}_{22} \end{bmatrix} \begin{Bmatrix} \underline{h}_1 \\ \underline{h}_2 \end{Bmatrix}$$

Rearrange as

$$\underbrace{\begin{bmatrix} \underline{H}_{11} & -\underline{V}_{12} \\ \underline{H}_{21} & -\underline{V}_{22} \end{bmatrix}}_{\underline{A}} \begin{Bmatrix} \underline{V}_1 \\ \underline{h}_2 \end{Bmatrix} = \underbrace{\begin{bmatrix} \underline{V}_{11} & -\underline{H}_{12} \\ \underline{V}_{21} & -\underline{H}_{22} \end{bmatrix}}_{\underline{B}} \begin{Bmatrix} \underline{h}_1 \\ \underline{V}_2 \end{Bmatrix} \to 0$$

$\underline{K}$ matrix $\searrow$

$$\begin{Bmatrix} \underline{V}_1 \\ \underline{h}_2 \end{Bmatrix} = \underline{A}^{-1}\underline{B} \begin{Bmatrix} \underline{h}_1 \\ \underline{V}_2 \end{Bmatrix} = \begin{bmatrix} \boxed{////} & \\ & \end{bmatrix} \begin{Bmatrix} \underline{h}_1 \\ \underline{V}_2 \end{Bmatrix}$$

Discard
remainder

For example — $\underline{K}$ matrix, BEM is 10×10
in size.

$$\underline{V}_1 = \underbrace{\underline{A}^{-1}\underline{B}}_{\underline{K}} \, \underline{h}_1$$

$$\underline{q} = \underline{K} \quad \underline{h}$$

A Boundary Element–Finite Element Procedure
For Porous and Fractured Media Flow

DEREK ELSWORTH

Department of Mineral Engineering, Pennsylvania State University, University Park

A coupled boundary element–finite element procedure is presented for linear and nonlinear fluid flow simulation in porous and fractured aquifers. Quadratic variation of both element geometry and fundamental singularity is used in the constitutively linear direct boundary element formulation. Compatible 3- to 9-noded Lagrangian finite elements are used to represent the plane flow domain for mixed linear and nonlinear flows, alike. Nodes on the external contour of the boundary element domain are only retained if flux boundary conditions are not prescribed, thus resulting in reduced matrix dimension. The geometric conductance of the linear boundary element region is evaluated only once. The resulting system matrices remain sparse, positive definite, and may be arranged for symmetry. Nonlinearity, in this context, is restricted to turbulent flow at high Reynolds numbers, although other nonlinearities may be easily accommodated using a similar procedure. A Missbach relationship is implemented to represent turbulent flow in rock fractures. Turbulent effects are confined to the finite element domain, and the resulting nonlinear equations are solved by direct iteration. Validation studies are completed against analytical solutions to linear and nonlinear flow problems. Excellent agreement is obtained with relatively sparing nodal coverage.

INTRODUCTION

Numerical models provide an effective means of rapidly evaluating a number of comparative scenarios in the quantification of groundwater flows. The heterogeneous and discontinuous nature of rock aquifers, combined with the limited access and penetration of standard site investigation procedures, makes the acquisition and interpretation of basic hydrologic data extremely difficult. High-quality numerical simulation techniques therefore provide an extremely important tool with which the impact of varied engineering or resource exploitation schemes may be readily evaluated. Sensitivity analyses of this nature provide a firm basis upon which subsequent judgemental decisions may be made [*Bachmat et al.*, 1980].

Of the powerful numerical techniques available, formulations may be divided between domain and boundary formulations. Associated with individual models are intrinsic merits and demerits which regulate their performance in any set engineering situation. Domain formulations encompass finite element and finite difference methods and require that the interior of the flow field is suitably discretized. Conversely, boundary solution procedures require only that the external edge contours of separate hydraulic zones be delimited as in the direct and indirect boundary element methods.

Domain methods offer powerful attributes in that complex nonlinear flow behavior, such as that evident in partially saturated [*Neuman*, 1973] or turbulent flow [*Elsworth*, 1985], may be easily accommodated. The system matrices are nonfully populated and in many instances are sparse, allowing considerable economy in storage requirements and execution time. Further computational savings may be realized with the finite element class of domain solutions where elemental and global system matrices are guaranteed symmetric and positive definite for linear and nonlinear potential flow problems alike. The extensive meshing within the domain, however, exacerbates data input requirements and introduces additional inter-

nal degrees of freedom for which results are sometimes not required. Thus although the matrix bandwidth may be small, the number of active equations comprising the system may be extremely large.

Boundary solution procedures are ideally suited to problem geometries of large volume to surface area ratio (equidimensional). Relatively trivial meshing is required, the discretization being limited to the edge contour of hydraulically homogeneous zones. System matrices are, however, asymmetric and fully populated within identified hydraulic subregions. Additionally, the virtue exalted in requiring discretization over the domain contour only is negated if nonlinear analysis of the interior is attempted. Primarily for this reason, boundary solution methods have not enjoyed popular application to nonlinear problems.

Coupled boundary element–finite element procedures offer the potential of using each of the different numerical procedures in the environment to which they are best suited. The innate strength of domain methods in dealing with constitutive nonlinearity, together with the relatively favorable structure of the system matrices make them ideal candidates to describe the behavior of nonlinear regions embedded within otherwise linear systems. The effectiveness with which boundary element procedures may accommodate volumetrically large but constitutively linear domains presents an ideal medium with which the far field may be adequately represented. Nonlinear effects discussed in the following sections are restricted to turbulent flows in fractured and porous-fractured media.

PREVIOUS APPLICATION

Previous applications of physical coupling between domain and integral methods are evident within the continuum mechanics literature. These applications span the fields of wave mechanics [*Chen and Mei*, 1974; *Shaw*, 1978], electrostatics [*Silvester and Hsieh*, 1971], and elastostatics [*Brady and Wassyng*, 1981], although this list is not exhaustive. A fine summary and critical commentary on many of these methods is given in the work by *Zienkiewicz et al.* [1977]. Application to problems of Darcy fluid flow have been investigated by

TABLE 1. Equivalent Fracture Hydraulic Conductivities

Hydraulic Zone	Equivalent Hydraulic Conductivity K_e	Exponent α
1	$\dfrac{gb^2}{12v}$	1.0
2	$\dfrac{1}{b}\left[\dfrac{g}{0.079}\left(\dfrac{2}{v}\right)^{1/4} b^3\right]^{4/7}$	4/7
3	$4g^{1/2} \log\left[\dfrac{3.7}{k/2b}\right] b^{1/2}$	1/2
4	$\dfrac{gb^2}{12v(1 + 8.8(k/2b)^{3/2})}$	1.0
5	$4g^{1/2} \log\left[\dfrac{1.9}{(k/2b)}\right] b^{1/2}$	1/2

Shapiro and Andersson [1983]. A coupled procedure to accommodate line finite elements representing fractures in two dimensional space was presented using constant singularity boundary elements and linear variation finite elements.

The following presents a coupled procedure using Lagrangian quadratic basis functions to represent element geometry and dependent variables at the interface between finite element and boundary element regions. Interelement compatability is therefore strictly enforced. A method of straightforward coupling is used to condense out unnecessary nodal equations and application is investigated to linear and nonlinear flow problems.

FLOW NONLINEARITY

A generalized constitutive relationship for flow in saturated porous and fractured media may be represented by Darcy's law

$$v = -K\left(\frac{\partial\phi}{\partial x}\right)\frac{\partial\phi}{\partial x} \qquad (1)$$

where v is the Darcy flow velocity, $\partial\phi/\partial x$ is the driving hydraulic gradient, and $K(\partial\phi/\partial x)$ is the gradient dependent hydraulic conductivity. The nonlinearity arises from mixed inertial and turbulent effects which operate simultaneously as flow velocities become significant. Both inertial and turbulent effects are manifest as increased flow impedence when Darcy velocities are increased. Inertial impedance results from spatial accelerations within the flow field that may commonly be attributed to converging flow. These effects have been observed experimentally and may be deduced based on consideration of mometum balance within the Navier-Stokes equations [*Irmay*, 1958]. Turbulent effects may be evident at the high-flow velocities possible within open voided or fractured rock masses. Fractures, especially, provide open conduits in which high velocity flows may be realized under relatively modest hydraulic gradients. For rock fractures, the transition to turbulent flow is most conveniently indexed by recourse to the Reynolds number Re such that

$$Re = 2bv/v \qquad (2)$$

where b is the nominal fracture aperture, and v is the fluid kinematic viscosity. The nondimensional Reynolds number is extremely useful in fracture flow applications in that it is possible to define the range over which certain hydraulic parameters are applicable. These hydraulic parameters are the constants in the commonly used Missbach and Forchheimer flow laws.

The Forchheimer law uses a polynomial expression to relate the Darcy velocity v to driving hydraulic gradient $\partial\phi/\partial x$ as

$$\partial\phi/\partial x = \bar{a}v + \bar{b}v^2 \qquad (3)$$

where $\bar{a}$ and $\bar{b}$ are experimentally determined parameters assumed constant over a given range of Reynolds numbers. The general correctness of this expression may be deduced from manipulation of the Navier-Stokes equations [*Irmay*, 1958] with the constants $\bar{a}$ and $\bar{b}$ being properties of both the fluid and transmitting medium. For low velocity flows, $\bar{a}$ is equivalent to the reciprocal of hydraulic conductivity and $\bar{b}$ is near zero.

Despite the analytical robustness of the Forchheimer relationship, the more compact Missbach law has found greater favor within groundwater applications related to fracture hydrology [*Louis*, 1969] and flow in open voided materials [*Leps*, 1973] with some exceptions [*Volker*, 1969; 1975]. The Missbach law links Darcy velocity v to driving hydraulic gradient through a power relationship of the form

$$\partial\phi/\partial x = cv^e \qquad (4)$$

where the proportionality constant c and the power exponent e are constant over given ranges of Reynolds number. The Missbach relationship of (4) may be inverted to yield

$$v = -K_e\left[\frac{\partial\phi}{\partial x}\right]^\alpha \qquad (5)$$

where $\alpha = 1/e$ and the equivalent hydraulic conductivity K_e is constant only over a given range of Reynolds numbers. For laminar flow, K_e is equivalent to the saturated hydraulic conductivity, and α is unity. For fully turbulent flow in a rough-walled fracture, the equivalent hydraulic conductivity K_e may be determined empirically, and α is equal to 1/2. Transition from laminar to turbulent flow is indexed by a critical Reynolds number Re_c. For rough-walled fractures, both the critical Reynolds number and the equivalent hydraulic conductivity are controlled by the ratio of mean fracture wall roughness to fracture double aperture $k/2b$. Experimentally derived suites of results are available [*Louis*, 1969] to quantify these parameters. Equivalent hydraulic conductivity magnitudes are given in Table 1 referring to the hydraulic zones, one through five, depicted in Figure 1. These results are germane to the following.

FINITE ELEMENT IMPLEMENTATION

The nonlinear hydraulic conductivity of (5) may be rearranged into a form directly analagous to Darcy's law for one dimensional flow as

$$v = -\left[K_e\left[\frac{\partial\phi}{\partial x}\right]^{\alpha-1}\right]\frac{\partial\phi}{\partial x} = -\bar{K}\frac{\partial\phi}{\partial x} \qquad (6)$$

where $\bar{K}$ is an equivalent scalar value of nonlinear hydraulic conductivity, and α is set equal to 1 or 1/2 for laminar or turbulent flow, respectively. For two dimensional flow, the appropriate hydraulic conductivity tensor relating cartesian Darcy velocities to cartesian gradients is given by $-\bar{K}\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. For multinoded plane elements, parametric representation of geometry (x, y) and total hydraulic
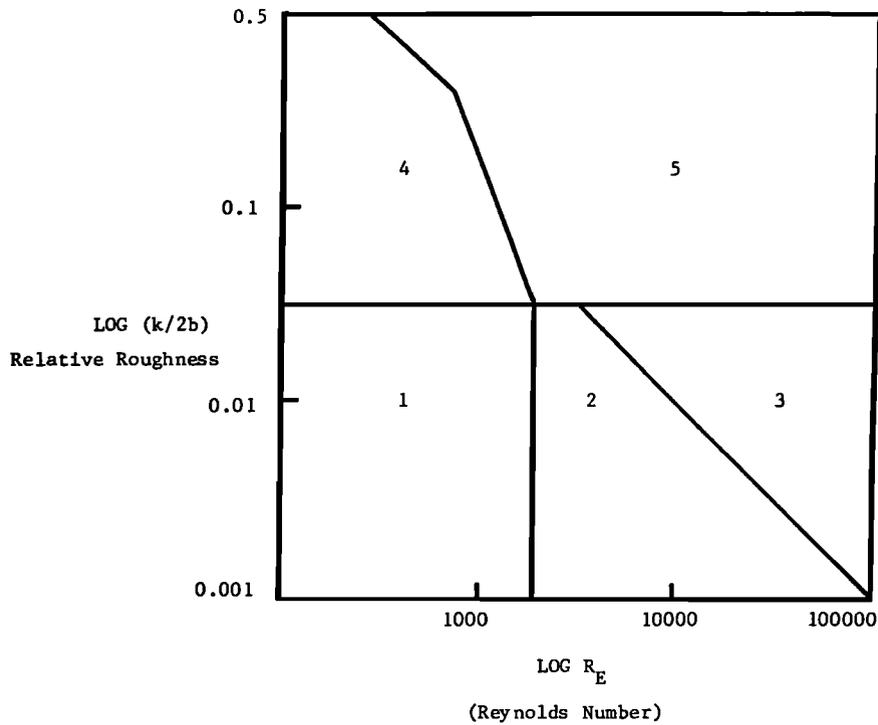
Fig. 1. Hydraulic zones for fracture flow [*Louis*, 1969].

head ϕ is appropriate. The point values of any of these parameters within the bounds of a single element is therefore

$$x = \mathbf{h}^T \mathbf{x} \tag{7a}$$

$$y = \mathbf{h}^T \mathbf{y} \tag{7b}$$

$$\phi = \mathbf{h}^T \boldsymbol{\phi} \tag{7c}$$

where $\mathbf{h}^T$ is a vector of basis functions, the vectoral (boldfaced) quantities are nodal values, and the left-hand sides represent interpolated values. Since equivalent nonlinear hydraulic conductivity $\bar{K}$ in the turbulent regime is a dependent function of hydraulic gradient, $\bar{K}$ will, in general, vary within individual elements. Substituting Darcy's law of the form given in (6) into the normal Galerkin method and subsequently applying Green's theorem yields the matrix equation

$$\mathbf{q} = \mathbf{K}\boldsymbol{\phi} \tag{8}$$

where $\mathbf{q}$ is a vector of prescribed nodal discharges defined per unit area, and $\mathbf{K}$ is a geometric conductance matrix. Equation (8) is equally valid at the elemental and global scales. For two dimensional analysis, the area integration required to evaluate the geometric conductance matrix $\mathbf{K}$ at the elemental level is given by

$$\mathbf{K} = b \int_\Omega \mathbf{a}^T \bar{\mathbf{K}} \mathbf{a} \, d\Omega \tag{9}$$

where $\mathbf{a}$ is a vector containing the derivatives of the shape functions $\mathbf{h}$ with respect to global coordinates; $\bar{\mathbf{K}}$ is a 2×2 diagonal matrix (i.e., $-\bar{K}\mathbf{I}$) containing the magnitude of the equivalent nonlinear hydraulic conductivity $\bar{K}$ at all nonzero entries; and Ω is the area of the element. For the two-dimensional case, the thickness b is considered constant over a

single element and Lagrangian basis functions $\mathbf{h}$ for a variable 3- to 9-noded element are used.

Rather than describe the variation of equivalent nonlinear hydraulic conductivity over the elemental domain using the nodal based shape functions of (7), the magnitude of $\bar{K}$ may be readily evaluated at the internal Gauss points. Dual or triple point quadrature may be used to evaluate all integrals of (9) with a dual-point scheme having proved sufficiently accurate for all examples completed to date. Since, for the turbulent case, $\bar{\mathbf{K}}$ is a function of the maximum in-fissure hydraulic gradient, the magnitude of the gradients with respect to global coordinates are given as

$$\left\{ \begin{array}{c} \dfrac{\partial \phi}{\partial x} \\[2mm] \dfrac{\partial \phi}{\partial y} \end{array} \right\} = \mathbf{a}\boldsymbol{\phi} \tag{10}$$

and the maximum hydraulic gradient is computed as the vector sum of the orthogonal components. Since the formulation is nonlinear with respect to nodal values of total head an iterative solution is implemented. For the global system, a laminar solution is first sought to provide initial nodal heads. This solution is used to evaluate hydraulic gradients and hence revise hydraulic conductivities. The direct iteration sequence employed is

$$\bar{\mathbf{K}}^l = f(\mathbf{a}\boldsymbol{\phi}^l) \tag{11}$$

$$\mathbf{q}^{l+1} = \mathbf{K}^l \boldsymbol{\phi}^{l+1} \tag{12}$$

where the superscripted l refers to the iteration cycle and $f(\ )$ refers to "a function of." Only those elements in which the
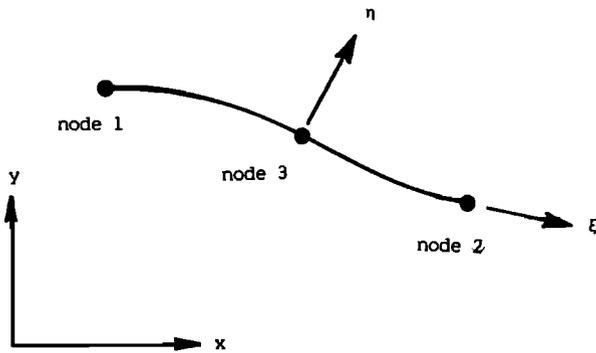
Fig. 2. Representation of a three-node isoparametric boundary element.

hydraulic conductivity K changes over a single-iteration cycle require to be reevaluated.

BOUNDARY ELEMENT ANALYSIS

To ensure effective coupling between the finite element and boundary element domains it is important that nonlinear effects propagating throughout the finite element region do not encroach into the boundary solution region. Under this proviso, the boundary domain is assumed to be constitutively linear and the formulation is able to operate in its most advantageous mode. In order that flow continuity between the domain and integral regions is maintained, the boundary element procedure must use basis functions compatible with those of the finite element region. For the boundary element procedure discussed in the following, isoparametric representation of both singularity and geometry is used. The element geometry is illustrated in Figure 2.

The boundary constraint equation corresponding to the direct formulation of the boundary element method may be stated as [*Jaswon and Symm*, 1977]

$$c(i)\phi(i) + \int_\Gamma V(i,j)\phi(j) \, d\Gamma = \int_\Gamma \Phi(i,j)v(j) \cdot n \, d\Gamma \quad (13)$$

where $V(i,j)$ and $\Phi(i,j)$ are kernel functions describing the influence effected at point j due to a unit source located at point i. The total hydraulic potential $\phi(j)$ and normal to the boundary velocity $v(j) \cdot n$ may be evaluated at any point on the boundary Γ from the kernel solutions. The free term $c(i)$ is a function of the domain geometry and is equal to δ_{ij} for an internal source and $\frac{1}{2}\delta_{ij}$ where point i is located on a smooth boundary with δ_{ij} being the Kronecker delta. For two dimensional porous media flow, the kernels for a line source are [*Kellog*, 1953]

$$\Phi(i,j) = \frac{M}{2\pi} \ln r \quad (14a)$$

$$V(i,j) = \frac{-KM}{2\pi r} \quad (14b)$$

where r is radius (i to j); K is the formation hydraulic conductivity; and M is the source strength. Lagrangian basis functions are used to define the geometry of an element where (13) may be rewritten in terms of local coordinates for a single biunit line element as

$$c(i)\phi(i) + \int_{-1}^{+1} V(i,j)\phi(j) \frac{d\Gamma}{d\xi} \, d\xi = \int_{-1}^{+1} \Phi(i,j)v(j)n \frac{d\Gamma}{d\xi} \, d\xi \quad (15)$$

and the Jacobian is identified as

$$\frac{d\Gamma}{d\xi} = \left[\left(\frac{dx}{d\xi}\right)^2 + \left(\frac{dy}{d\xi}\right)^2\right]^{1/2} \quad (16)$$

with

$$x = \mathbf{h}^T\mathbf{x} \quad (17a)$$

$$y = \mathbf{h}^T\mathbf{y} \quad (17b)$$

$$\phi = \mathbf{h}^T\boldsymbol{\phi} \quad (17c)$$

$$v \cdot n = \mathbf{h}^T(\mathbf{v} \cdot \mathbf{n}) \quad (17d)$$

where $\mathbf{h}^T$ contains a different family of basis functions from those identified in (7) previous. The Lagrangian basis functions are one dimensional in this case, varying only over the length of the element and are represented in local coordinates as

$$\mathbf{h}^T = \frac{1}{2}[(1-\xi) - (1-\xi^2); (1+\xi) - (1-\xi^2); 2(1-\xi^2)] \quad (18)$$

where ξ represents the natural coordinates of the biunit element with $-1 < \xi < 1$. Similar functional variation for both heads and boundary velocities are used, each being of quadratic form. Since velocities are related to the gradient of head, it may be desireable to use interpolation one degree lower for velocities than that for heads. The results of validation studies completed did not warrant implementation of this constraint. Under parametric representation, the integrals of (15) are evaluated by Gauss quadrature for all nodes comprising the boundary element system [*Stroud and Secrest*, 1966; *Elsworth*, 1986b]. Where a sharp corner is encountered at a node, the V kernel integrations are completed on adjacent segments where there is slope continuity on each element segment. These quantities are then summed to yield the nodal weighted flux out of the region rather than represent flux in any particular normal (to the boundary) direction. For a system of m nodes, each with a single degree of freedom, m simultaneous equations result. In matrix format these may be represented as

$$\mathbf{V}\boldsymbol{\phi} = \boldsymbol{\Phi}\mathbf{v} \cdot \mathbf{n} \quad (19)$$

which, for m known or prescribed nodal boundary conditions yields a solvable set. After performing appropriate column interchanges on (19) to rearrange all known boundary conditions to the right-hand side vector, the identity may be solved to yield a geometric conductance matrix such that

$$[\boldsymbol{\Phi}^{-1}\mathbf{V}]\boldsymbol{\phi} = \mathbf{v} \cdot \mathbf{n} \quad (20)$$

which is of similar form to the finite element statement of (8). Premultiplying (20) by the ranked cross-sectional area of flow will convert Darcy flow velocities directly to discharge quantities such that

$$\mathbf{q} = \mathbf{b} \int_\Gamma \mathbf{h}^T\mathbf{v} \cdot \mathbf{n} \, d\Gamma \quad (21)$$

where $\mathbf{q}$ is a vector of nodal discharges, and $\mathbf{h}^T$ is a vector of element by element defined basis functions. The constant out of plane thickness of the element is given by b, which is unity for plane flow or equal to fracture aperture for fracture flow applications. Identities (8) and (20) are fully compatible in a rigorous fashion. Interelement flow continuity is maintained

between boundary and domain formulations in a straightforward manner.

Boundary Conditions

Simultaneous solution of (19) is only possible if either head or velocity boundary conditions are prescribed at all nodes of the boundary solution domain. Since, in general, the boundary nodes that interface directly with the finite element mesh will have "a priori" undefined boundary conditions, it is necessary to prescribe artificial boundary conditions to aid the symbolic inversion of (20). Column substitution is first completed to move all nodal quantities corresponding to known total head and, as yet, unconstrained head boundary conditions to the right-hand side of (19). The right-hand side is completely defined if, for all the unconstrained nodes, the head at one node is held at unity and all others are set to zero. The system of equations may then be solved. When repeated for all unconstrained nodes this procedure results directly in a geometric conductance matrix linking nodal heads to nodal discharges. If l prescribed head nodes exist on a boundary domain of m nodes then the resulting geometric conductance matrix from the boundary solution procedure is fully populated and $l \times l$ in dimension. Thus the symbolic inversion of (20) is equivalent to solving a system of m equations for l different solution vectors.

All nodes corresponding to prescribed velocity boundary conditions are effectively condensed out and no equations require to be set up in the following coupled solution of the finite element and boundary element geometric conductance matrices [*Elsworth*, 1986b]. Equation (20) represents the geometric conductance matrix for a single multinoded element. The conductance matrix may be directly substituted into standard finite element matrix assembly routines as a single multinoded element with appropriate nodal connections. For linear flow, the matrix entries for the boundary element domain are invariant and require to be evaluated only once.

Matrix Symmetry

No particular problems arise in coupling boundary and domain methods if the boundary solution matrices are asymmetric, although the procedure may be expedited if both system matrices are symmetric. If a variational formulation is adopted the geometric conductance matrix (equation (20)) may be made symmetric after formation according to the method of *Zienkiewicz et al.* [1977]. In generality, different functional variation may be chosen for normal velocities $v \cdot n$ and heads ϕ along the boundary of the domain. If heads and velocities are defined by shape functions H^a and H^b relative to the entire boundary of the domain then

$$\phi = H^a \phi \tag{22}$$

$$v \cdot n = H^b v \cdot n \tag{23}$$

Nodal fluxes $v \cdot n$ at the boundary are related to heads by the geometric conductance relationship of (20) such that

$$v \cdot n = [\Phi^{-1}V]\phi \tag{20'}$$

The total potential π of the region may be given for the case where nodal heads only are prescribed as

$$\pi = \frac{1}{2} \int_\Gamma (v \cdot n)^T \phi \, d\Gamma \tag{24}$$

which on substitution of (20), (22) and (23) gives

$$\pi = \frac{1}{2}\phi^T \int_\Gamma [[\Phi^{-1}V]^T H^{bT} H^a \, d\Gamma]\phi \tag{25}$$

and may be minimized appropriately to give a revised geometric conductance matrix K

$$K = \frac{1}{2} \int_\Gamma [((\Phi^{-1}V)^T H^{bT} H^a)^T + ((\Phi^{-1}V)H^{bT}H^a)] \, d\Gamma \tag{26}$$

where symmetry is guaranteed. The functional variation over individual elements enforced in the current formulation is identical for head and normal velocity and therefore $H^a \equiv H^b$. To guarantee matrix symmetry in the boundary element formulation, a surrogate to (26) is invoked [*Banerjee and Butterfield*, 1981] such that

$$K = \frac{1}{2}[(\Phi^{-1}V)^T + (\Phi^{-1}V)] \tag{27}$$

to avoid the integration enforced within (26). This approach has been found to be entirely adequate as is illustrated in the following validation exercises.

VALIDATION

Analytical solutions for linear and nonlinear flow within simple domains are used to examine the accuracy, versatility, and utility of the proposed coupled formulation.

Linear Flow

The performance of the coupled procedure is first examined for the case of a concentrically holed, circular, porous disc containing both embedded and fully penetrating finite element domains. The ability to prescribe boundary conditions on a node by node basis for both the finite element and boundary element domains provides no particular differences in meshing and execution for embedded or penetrating domains. Disc geometries are illustrated in Figure 3 for the two individual cases with inner radius $r = a$. The variation in hydraulic potential with radius is shown in Figure 4. Excellent agreement is maintained between analytical and numerical solutions even for relatively modest nodal coverage. The presence of perpendicular corners at the interface between boundary element and finite element domains are shown not to adversely affect results.

In the case of a semi-infinite domain, the coupled solution procedure may similarly be shown to perform satisfactorily. The solution for a pressure tunnel within a saturated porous half space is used (J. W. Bray, personal communication, 1980). In this example, the direct boundary element procedure requires that the solution domain remains finite but may be expanded to considerable dimension without computational penalty. The expanded representation of the half space domain is illustrated in Figure 5. The problem geometry comprises a single circular tunnel of radius 5 m present at a depth of 40 m below the ground surface. The piezometric surface to the domain is coincident with the ground surface and unit head is applied in the tunnel annulus. The boundary element discretization comprises 48 interior and 32 exterior nodes divided between 40 three-noded elements. For the finite element domain, 8 nine-noded Lagrangian elements are used totaling 45 nodes. Zero flux boundary conditions are applied to the
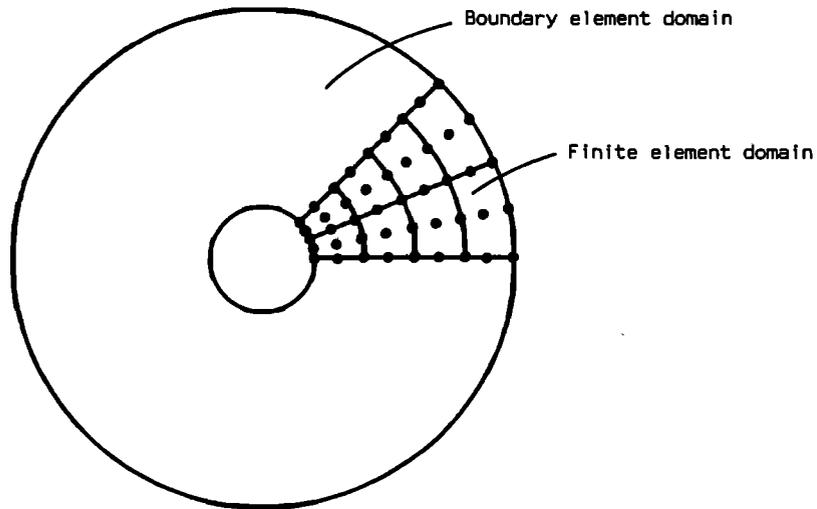
(b)
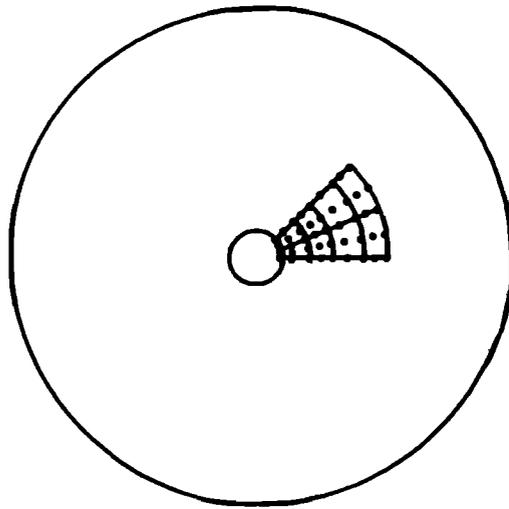
(a)

Fig. 3.   Circular concentrically perforated disc with (a) embedded and (b) fully penetrating finite element domains.

lower and side elements of the boundary element exterior with the result that the conductance matrix derived from the boundary element discretization retains only 57 degrees of freedom. The assembled finite element–boundary element procedure has a total of 84 degrees of freedom. The variation in normalized hydraulic potential for the solution geometry is illustrated in Figure 6. The nodal potentials along sections A-A' and B-B' are shown to yield excellent agreement wth the analytical solution. This excellent agreement is maintained despite use of relatively sparing nodal coverage in the finite element domain. Similarly, the large discrepancy in physical magnitude of the boundary element and finite element domains, as illustrated in Figure 5 has not affected solution accuracy.

In addition to being capable of representing conditions of porous media flow, the coupled model may be used in fracture flow applications. Analytical solution is vailable for the case of an infinite porous medium traversed by a single fracture of finite length and infinite hydraulic conductivity [*Gringarten*,
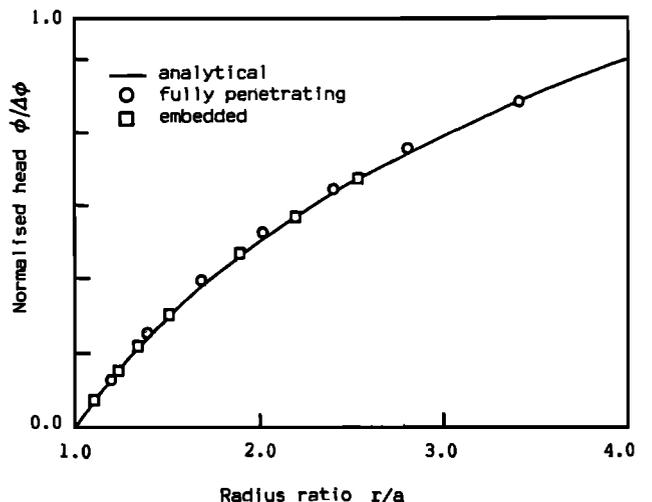

Fig. 4.   Variation in normalized hydraulic head with radius for perforated disc geometries.
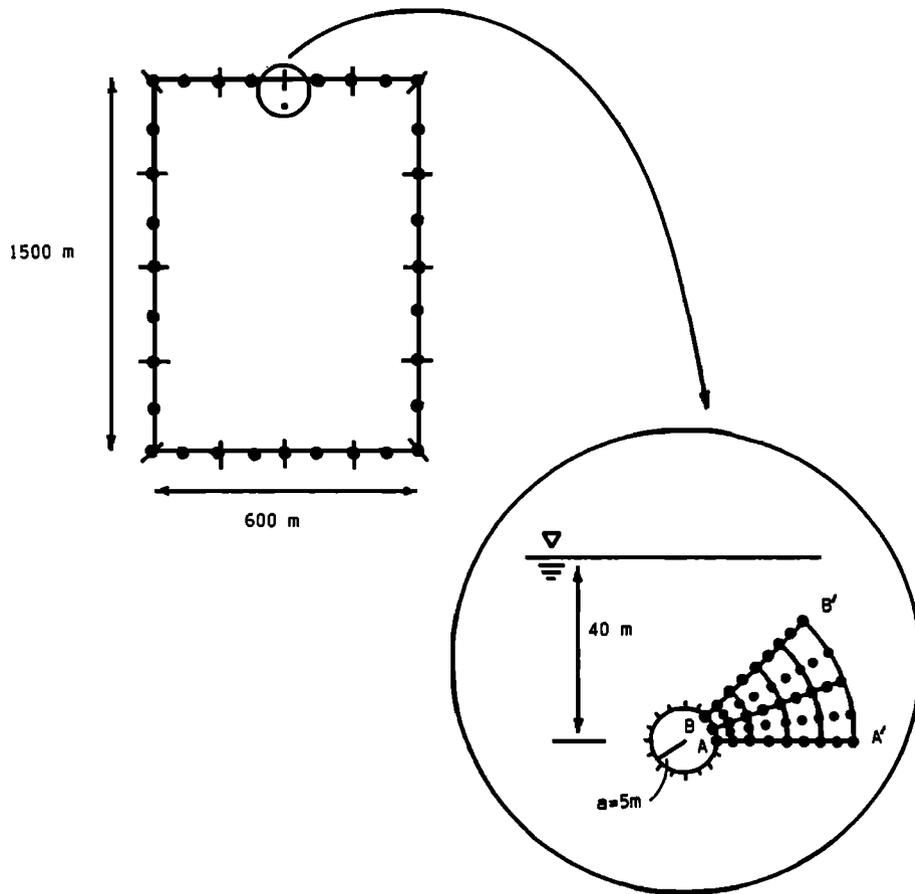
Fig. 5.   Discretization geometry used for a pressure tunnel problem (to scale).

1974]. The fracture is symmetrically disposed about a central wellbore from which an infinite reservoir is pumped. The boundary element discretization of the truncated infinite domain is illustrated in Figure 7a; the internal crack is divided into 21 nodes and 10 elements, and the external boundary is devided into 16 nodes and 8 elements. The external constant potential boundary is arbitrarily located at a distance of 50 m to simulate the infinite domain. The domain external nodes are all equally spaced on the circumference although the large hydraulic gradients and velocities manifest at the crack tip are best reproduced if nodal concentrations are located at the fracture tip. The discretization density is illustrated in Figure 7a where individual elements cover 0.56, 0.24, 0.12, 0.06, and 0.02 m of the fracture half length. In accordance with a validation example reported by Shapiro and Andersson [1983], the surrounding porous medium is represented by a formation conductivity of 1 m/d and central well discharge of 1 m²/d.

The boundary element model used in this procedure uses internal slit elements to represent the internal fracture. The essential component of this element is that it allows discharge into the element from the surrounding medium on either side. The formulation of the element has been adequately described elsewhere [Elsworth, 1986a] and no further explanation will be given here. With the slit element in place within the boundary solution domain, the relevant matrix identities may be assembled and inverted to yield the geometric conductance of the system. To this condensed system, fracture line elements representing the internal fracture are added and the system solved in finite element format using a central producing wellbore. In agreement with the example completed by Shapiro

and Andersson [1983], fracture conductivities of 10⁴ m/d are ascribed to the vertical fracture to simulate "infinite" conductivity. Using this conductivity contrast, excellent agreement between the analytical results of Gringarten [1974] and the
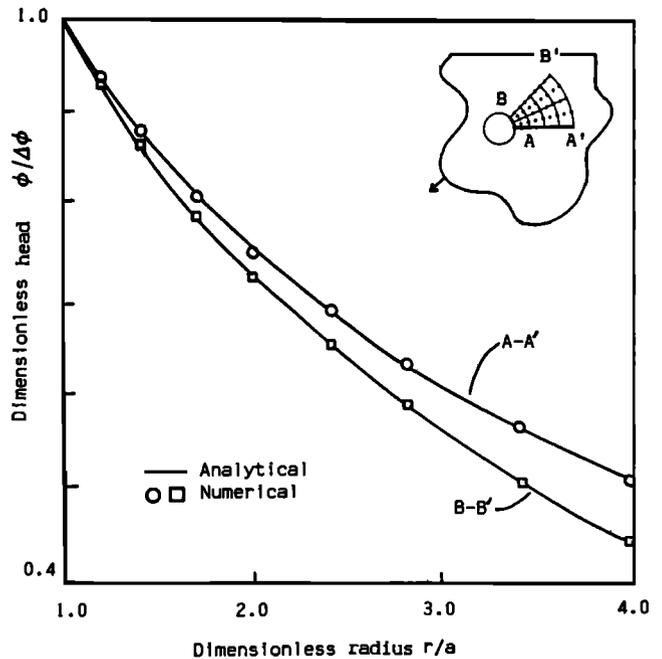

Fig. 6.   Variation in total hydraulic head with radius for pressure tunnel geometry.
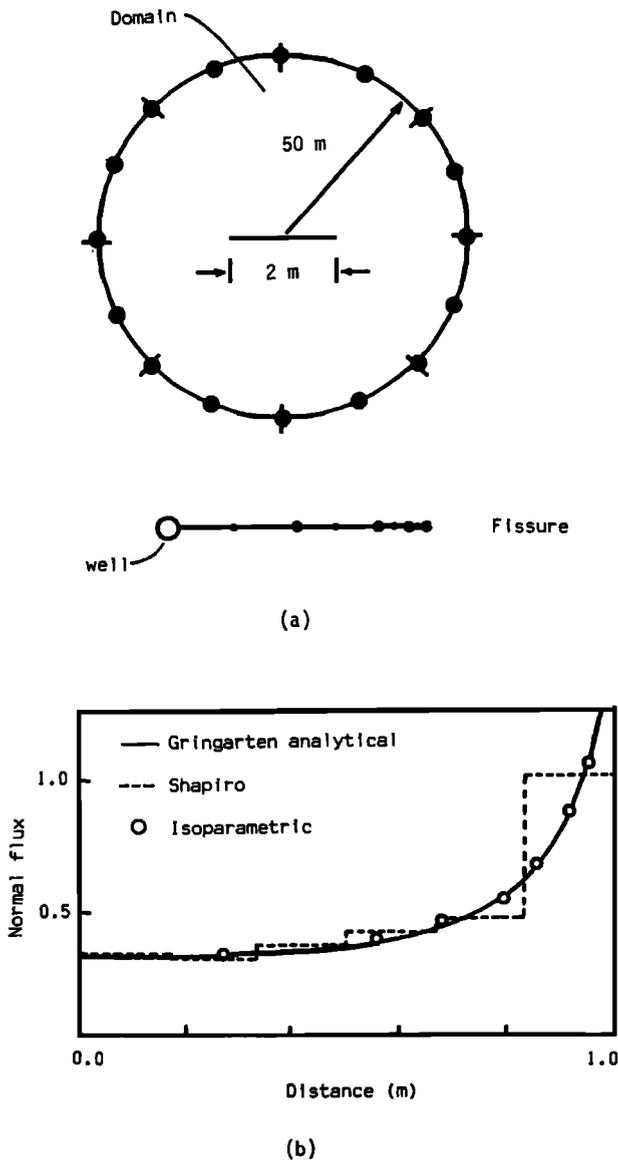
(a)

(b)

Fig. 7. Geometric representation of a fracture within a pseudo infinite porous medium (a) and results for the isoparametric model (b).

model formulated in this work are obtained. The results are illustrated in Figure 7b.

A second solution procedure may be applied to the problem whereby more note is made of the true character of the central fracture of infinite conductivity. The infinite conductivity of the fracture implies that head losses along the fracture length will be zero. Thus the only compatible solution to the problem is one in which nodal potentials along the internal slit are constant. Similarly, normal (to the fracture) mass fluxes must total 1 m²/d when integrated over the fracture length. Therefore the problem may be solved iteratively satisfying these two internal constraints of constant potential and prescribed total normal flux. Solution by this procedure yields identical results

to those previous. Clearly, the fracture to porous medium conductivity contrast of 4 orders of magnitude used in the alternate problem treatment is sufficient to represent the infinite conductivity of the fracture. Comparison of the results from this model with those of the analytical solution illustrates the ability of the formulation to faithfully represent the high flux gradients at the crack trip. The crack tip flux is recorded at 11.7 m/d representing an assymptote set at approximately ten times the height of the figure vertical axis.

Nonlinear Flow

Radial flow within a single circular fracture pierced centrally by a well bore is a problem for which an analyical solution is avalable (B. Amadei, personal communication, 1983). For validation, an axisymmetric geometry is chosen with domain external and internal radii of 6.0 and 0.25 m respectively. Finite element discretization reaches to a radius of 2.0 m. The combined boundary element-finite element mesh illustrated in Figure 8 is used. The domain comprises 51 finite element nodes and 20 boundary element nodes. The boundary conditions for the boundary element domain are such that only six active degrees of freedom are retained in the condensed geometric conductance matrix. For a nominal fracture aperture b of 1.0 cm, fracture relative roughness $k/2b$ of 0.5, fluid kinematic viscosity v of 1.8×10^{-6} m²/s, and a head differential across the system of 0.022 m, the nonlinear flow results are illustrated in Figure 9. Excellent agreement is obtained between the analytical and numerical results. The numerical results are completed using two point Gauss quadrature in evaluating the nonlinear conductance matrix integrals. For this particular example, the results following eight iteration cycles are graphically indistinguishable from those of over 20 iterations duration. Acceptable results are normally obtained after 10 iterations. It is apparent from these simple validation exercises that the proposed formulation is capable of returning satisfactory results to a variety of linear and nonlinear potential flow applications.

CONCLUSIONS

A coupled solution procedure is presented that is capable of representing linear and nonlinear flows in porous and fractured media. The coupling is performed in a straightforward manner through noting respective nodal conductance associations. This procedure allows arbitrarily embedded or located nonlinear zones to be easily analyzed. The boundary element domain may be simply considered as a single multinoded finite element and accommodated appropriately.

The boundary element procedure is particularly suited to representing volumetrically large or pseudo infinite domains where system matrix size or solution stability is, within reason, unaffected by domain dimension. Where prescribed flux nodes are included on the boundary element edge contour, the corresponding system equations are not retained at the global level. Depending on mesh specific details, this results in considerable computational saving both at the stage of reducing the bound-
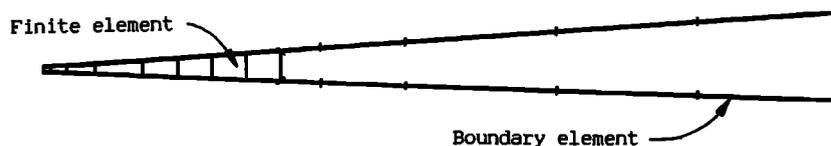

Fig. 8. Discretization of turbulent radial flow within a planar rock fracture.
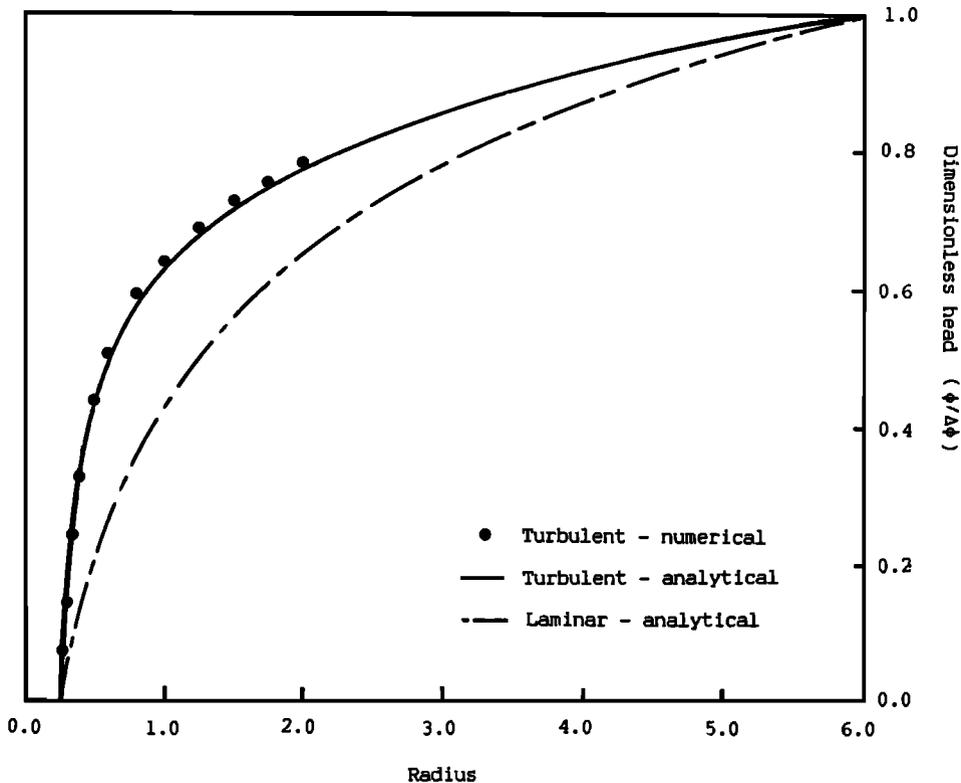
Fig. 9. Variation in normalized head with radius for turbulent radial flow within a planar rock fracture.

ary element domain and later in global system matrix assembly and solution.

The use of quadratic functional variation for both the finite element and boundary element domains ensures compatibility in the strictest sense. This facet appears especially useful in accurately representing regions of high gradient within the flow domain at crack tips and other singularities.

Nonlinear flow effects are accommodated most effectively where the nonlinearity is confined within the finite element domain. This allows the domain and integral methods to operate to maximum advantage. Since nonlinear effects, such as turbulence, are commonly of limited areal extent, coupled procedures provide a viable method of analysis. This is especially true where the zones of expected turbulence may be delimited a priori, say, by the known presence of highly conductive fractures. In other instances, where turbulent areas cannot be identified before analysis, some form of self-adaptive capability in the analysis would clearly be an advantage. Such concerns are not addressed herein. The proposed procedure is also applicable to other nonlinear flow problems.

The resulting matrix identities for the boundary element domain may be made positive definite and symmetric. This facet allows execution using readily available finite element coding arrangements accommodating storage of geometric conductance matrix terms above, and including, the leading diagonal only.

NOTATION

K hydraulic conductivity.

K_e equivalent hydraulic conductivity.

$\bar{K}$ equivalent nonlinear hydraulic conductivity.

$\mathbf{K}$ geometric conductance matrix (FEM).

$\bar{\mathbf{K}}$ medium hydraulic conductivity tensor.

M line source or sink strength.

Re, Re_c Reynolds number, critical Reynolds number.

$V(i, j), \Phi(i, j)$ kernel terms.

$\mathbf{V}, \mathbf{\Phi}$ matrices of integrated kernel terms.

$\mathbf{a}^T$ vector of basis function derivatives in global coordinates.

$\bar{a}, \bar{b}$ Forchheimer equation constants.

b fracture aperture.

c, e Missbach equation constant, Missbach equation exponent.

$c(i)$ free term.

g gravitational acceleration.

$\mathbf{h}^T$ vector of element basis functions.

k fracture absolute roughness.

l iteration count.

$\mathbf{n}$ domain unit outward normal.

$q, \mathbf{q}$ discharge, vector of nodal discharges.

r radius of separation of kernel functions.

$v, \mathbf{v}$ Darcy flow velocity, vector of nodal flow velocities.

x, y Cartesian coordinates.

α turbulent flow exponent.

δ_{ij} Kronecker delta.

ϕ total hydraulic head.

Ω domain area.

Γ domain external contour.

ξ, η local coordinates.

v fluid kinematic viscosity.

REFERENCES

Bachmat, Y., J. Bredehoeft, B. Andrews, D. Holtz, and S. Sebastian, *Groundwater Management: The Use of Numerical Models, Water Resour. Monogr. Ser.*, vol. 5, edited by P. van der Heijde et al., AGU, Washington, D. C., 1980.

Banerjee, P. K., and R. Butterfield, *Boundary Element Methods in Engineering Science*, McGraw-Hill, New York, 1981.

Brady, B. H. G., and A. Wassyng, A coupled boundary element-finite

element method of stress analysis, *Int. J. Rock Mech. Min. Sci.*, *16*, 235–244, 1981.

Chen, H. S., and C. C. Mei, Oscillations and wave forces in a man-made harbor in open sea, paper presented at Proceedings, 10th Symposium on Naval Hydrodynamics, MIT, Cambridge, Mass., 1974.

Elsworth, D., Coupled finite element/boundary element analysis for nonlinear flow in rock fractures and fracture networks, in *Proceedings of the 26th U.S. Symposium on Rock Mechanics*, pp. 633–641, Balkema Publishers, Rotterdam, 1985.

Elsworth, D., A hybrid boundary element-finite element analysis procedure for fluid flow simulation in fractured rock masses, *Int. J. Numer. Anal. Method Geomechan.*, *10*(6), 569–584, 1986a.

Elsworth, D., A model to evaluate the transient hydraulic response of three-dimensional sparsely fractured rock masses, *Water Resour. Res.*, *22*, 1809–1819, 1986b.

Gringarten, A. C., H. J., Ramey, and R. Raghavan, Unsteady state pressure distributions created by a well with a single infinite conductivity vertical fracture, *Soc. Petr. Eng. J.*, *14*, 347–360, 1974.

Irmay, S., On the theoretical derivation of Darcy and Forchheimer formulas, *Eos. Trans. AGU*, *30*, 702–707, 1958.

Jaswon, M. A., and G. T. Symm, *Integral Equation Methods in Potential Theory and Elastostatics*, Academic, Orlando, Fla., 1977.

Kellog, O. D., *Foundations of Potential Theory*, Dover, Mineola, N. Y., 1953.

Leps, T. M., Flow through rockfill, in *Embankment Dam Engineering*, edited by S. G. Poulos, pp. 87–107, John Wiley, New York, 1973.

Louis, C., A study of groundwater flow in rock and its influence on the stability of rock masses, *Rock Mech. Res. Rep.*, *10*, Imperial Coll., London, September 1969.

Neuman, S. P., Saturated-unsaturated seepage by finite elements, *J. Hydraul. Eng.*, *99*(HY12), 2233–2251, 1973.

Shapiro, A. M., and J. Andersson, Steady state fluid response of fractured rock: A boundary element solution for a coupled discrete fractured continuum model, *Water Res. Res.*, *19*(4), 959–969, 1983.

Shaw, R. P., Coupling boundary integral equation methods to other numerical techniques, in *Recent Developments in Boundary Element Methods*, Southhampton University, 1978.

Silvester, P., and M. S. Hsieh, Finite element solution of 2D exterior field problems, *Proc. Inst. Electr. Eng.*, *118*(12), 1943–1947, 1971.

Stroud, A. H., and D. Secrest, *Gaussian Quadrature Formulas*, Prentice-Hall, Englewood Cliffs, N. J., 1966.

Volker, R. E., Nonlinear flow in porous media by finite elements, *J. Hydraul. Eng.*, *95*(HY6), 2093–2114, 1969.

Volker, R. E., Solutions for unconfined non-Darcy seepage, *J. Irrig. Drain. Div. Am. Soc. Civ. Eng.*, *101*(IR1), 53–65, 1975.

Zienkiewicz, O. C., D. W. Kelly, and P. Bettes, The coupling of the finite element method and boundary element procedures, *Int. J. Numer. Meth. Eng.*, *11*, 355–375, 1977.

D. Elsworth, Department of Mineral Engineering, Pennsylvania State University, 104 Mineral Sciences Building, University Park, PA 16802.

8

Alternative Solution Methods [Cont'd]

[8:1] Alternative Solution Methods [Cont'd]

SPH – Smoothed Particle Hydrodynamics

LBM – Lattice Boltzmann Methods

DEM – Distinct Element Methods

XFEM – Extended FE Method

Smoothed particle hydrodynamics: theory and application to non-spherical stars

R. A. Gingold and J. J. Monaghan[*] *Institute of Astronomy,*
Madingley Road. Cambridge, CB3 0HA

Summary. A new hydrodynamic code applicable to a space of an arbitrary number of dimensions is discussed and applied to a variety of polytropic stellar models. The principal feature of the method is the use of statistical techniques to recover analytical expressions for the physical variables from a known distribution of fluid elements. The equations of motion take the form of Newtonian equations for particles. Starting with a non-axisymmetric distribution of approximately 80 particles in three dimensions, the method is found to reproduce the structure of uniformly rotating and magnetic polytropes to within a few per cent. The method may be easily extended to deal with more complicated physical models.

1 Introduction

Many of the most interesting problems in astrophysics involve systems with large departures from spherical symmetry. This may occur either because the initial state lacks spherical symmetry, as in the case of a protostar forming from a dense interstellar cloud, or because non-spherical forces arising from rotation or magnetic fields, as in the case of the fission of a rotating star, play an important part in the dynamics. Frequently these sources of non-spherical symmetry will be found combined.

Because of the complexity of these systems numerical methods are required to follow their evolution. However, the standard finite difference representations of the continuum equations are of limited use, because of the very large number of grid points required to treat each coordinate on an equal footing. If, for example, 20 points along the radial direction give adequate accuracy for a spherical polytrope, we may require $(20)^3$ such points to give the same accuracy for a highly distorted polytrope. This difficulty is mirrored in the evaluation of multiple integrals.

For the astrophysical problems a numerical method which allows reasonable accuracy for a small number of points is required. Ideally it should also be simple to program and robust. An early attempt to provide such an alternative to the standard finite difference method was made by Pasta & Ulam (1959). They replaced the continuous fluid by a fictitious set of

[*] Permanent address: Mathematics Department, Monash University, Clayton, Victoria 3168, Australia.

particles with inter-particle forces designed to mimic the true pressure and other body forces. The weakness in this method is that transport processes are difficult to include correctly.

A better method is to make use of the Lagrangian description of fluid flow which automatically focuses attention on fluid elements. In the discrete version, parcels of fluid move according to the Newtonian equations with forces due to the pressure gradient and other body forces: gravity, rotation and magnetic. The central feature of our analysis* is the method we use to determine the forces from the current positions of the fluid elements.

For fluid elements of equal mass, the number per unit volume must be proportional to the density. In addition, unless special symmetry is introduced from the start, the positions of the elements will be random because of the complicated motion which is inevitable for large N-body systems. We therefore make the assumption that, at any time, the positions of the fluid elements are randomly distributed according to the density. To recover the density from the known distribution of elements is then equivalent to recovering a probability distribution from a sample. Statisticians have given two methods for doing this which are well suited to the fluid problem. The first is the smoothing kernel method (Bartlett 1963; Parzen 1962), and the second is the delta spline technique (Boneva, Kendall & Stepanov 1971). Both methods may be thought of as an approximation to an integral determined according to the Monte Carlo procedure. Since the Monte Carlo method is known to give reasonable estimates of multiple integrals with fewer points than finite difference methods often require, it is plausible to expect a reduction in work if the statistical smoothing methods are used. We call this method smoothed particle hydrodynamics (SPH).

In this paper we first give a detailed description of the smoothing method and establish conditions which guide the choice of the smoothing kernel. Static spherical polytropes are then studied by relaxing from an initial non-spherical configuration with a damping term in the equations of motion. The free non-spherical oscillations of polytropes are then examined. Finally the departures from spherical symmetry produced by uniform rotation and magnetic fields in polytropes are determined and compared with results from perturbation theory.

2 Recovering distributions and body forces

2.1 THE DENSITY DISTRIBUTION

The equation of motion of the jth element of fluid with volume Δv_j, centre of mass $\mathbf{r}_j$ and density ρ_j is

$$\rho_j \Delta v_j \frac{d^2 \mathbf{r}_j}{dt^2} = -\Delta v_j \nabla P + \rho_j \Delta v_j \mathbf{F}_j, \tag{2.1}$$

or

$$\frac{d^2 \mathbf{r}_j}{dt^2} = -\frac{1}{\rho_j} \nabla P + \mathbf{F}_j, \tag{2.2}$$

where $\mathbf{F}_j$ is the body force acting on the element of fluid and ∇P is the pressure gradient at $\mathbf{r}_j$. Since in our approximation the element of fluid is described dynamically by a point, we shall call it a particle, and (2.2) the equation of motion of the jth particle.

It is convenient to begin our analysis by considering the calculation of a smoothed

density from a set of points, the various $\mathbf{r}_j$ distributed according to the density. Following Parzen (1962), we consider a smoothed density $\rho_s(\mathbf{r})$ defined by

$$\rho_s(\mathbf{r}) = \int W(\mathbf{r} - \mathbf{r}') \, \rho(\mathbf{r}') \, d\mathbf{r}', \tag{2.3}$$

where W is a function satisfying the condition

$$\int W(\mathbf{r}) \, d\mathbf{r} = 1, \tag{2.4}$$

where the integration is over all space.

If $\rho(\mathbf{r}')$ is unknown, (2.3) cannot be evaluated, but if we have a set of N points $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ distributed according to ρ, the integral can be evaluated by the Monte Carlo method (Hammersley & Handscomb 1964). Thus, defining $\rho_N(\mathbf{r})$ by

$$\rho_N(\mathbf{r}) = \frac{M}{N} \sum_{j=1}^{N} W(\mathbf{r} - \mathbf{r}_j), \tag{2.5}$$

where

$$M = \int \rho(\mathbf{r}) d\mathbf{r}, \tag{2.6}$$

we find, with E denoting the expectation

$$E\left[\rho_N(\mathbf{r})\right] \equiv \frac{1}{M^N} \int \ldots \int \rho_N(\mathbf{r}) \prod_{i=1}^{N} \rho(\mathbf{r}_i) d\mathbf{r}_i = \rho_s(\mathbf{r}). \tag{2.7}$$

In our numerical procedure only one sample distribution is produced each time. The equality (2.7) is therefore to be understood as implying that if we were to create an ensemble of models, each starting with a different array of points consistent with the initial conditions, then the ensemble average of $\rho_N(\mathbf{r})$ would be $\rho_s(\mathbf{r})$.

The error involved in replacing $\rho_s(\mathbf{r})$ by $\rho_N(\mathbf{r})$ is $\pm\sigma$, where σ is defined by

$$\sigma^2 = E\left[(\rho_N(\mathbf{r}) - \rho_s(\mathbf{r}))^2\right]$$

$$\doteq \frac{M^2}{N^2} \sum_j W^2(\mathbf{r} - \mathbf{r}_j) - \frac{1}{N}\left[\frac{M}{N} \sum_j W(\mathbf{r} - \mathbf{r}_j)\right]^2 \tag{2.8}$$

To complete the chain of analysis it is necessary to show that a $W(\mathbf{r})$ can always be chosen so that, as N increases, $\rho_s(\mathbf{r})$ becomes a better approximation to $\rho(\mathbf{r})$. We establish this result in the next section.

2.2 CHOOSING THE KERNEL $W(\mathbf{r})$

Intuitively it seems reasonable to expect that $W(\mathbf{r})$ can be made more like $\delta(\mathbf{r})$ as N becomes larger. If this is the case then

$$\rho_s(\mathbf{r}) \rightarrow \rho(\mathbf{r}) \text{ as } N \rightarrow \infty.$$

To make this result more precise it is convenient to write $W(\mathbf{r})$ in the form

$$W(\mathbf{r}) = \frac{1}{h^3} K(\mathbf{r}/h), \qquad (2.9)$$

where h is a parameter with the dimensions of length and the space is assumed to be three dimensional. By an easy generalization of a theorem due to Parzen (1962) we find that if

$$h \to 0 \text{ as } N \to \infty$$

and if $K(\mathbf{u})$ is a Borel function satisfying

$$\int K(\mathbf{u}) \, d\mathbf{u} = 1, \quad |\mathbf{u}^2 K(\mathbf{u})| \to 0 \quad \text{as} \quad |\mathbf{u}| \to \infty, \quad \int |K(\mathbf{u})| \, d\mathbf{u} < \infty$$

where the integrals are over all space, then

$$\rho_N(\mathbf{r}) \to \rho(\mathbf{r}) \text{ as } N \to \infty.$$

It proves convenient to choose $W(\mathbf{r})$ to be an even function. Typical examples of kernel functions in three dimensions are

$$\text{(i)} \left(\frac{1}{\pi h^2}\right)^{3/2} \exp\left(-r^2/h^2\right) \quad \text{(ii)} \frac{3H(1-|\mathbf{r}|/h)}{4\pi h^3} \quad \text{(iii)} \frac{S(|\mathbf{r}|/h)}{h^3} \qquad (2.10)$$

where H is the Heaviside step function and S is the spherical delta spline discussed in Appendix 1. Each of the functions in (2.10) is a member of a sequence of functions which represents the delta function.

In addition to requiring $\rho_N(\mathbf{r}) \to \rho(\mathbf{r})$ we require that σ should be as small as possible. To satisfy these conditions we choose h by minimizing the functional

$$L(\mathbf{r}) = \{E[\rho_N(\mathbf{r})] - \rho(\mathbf{r})\}^2 + E[(\rho_N(\mathbf{r}) - \rho_s(\mathbf{r}))^2]$$

$$= E[(\rho_N(\mathbf{r}) - \rho(\mathbf{r}))^2]. \qquad (2.11)$$

Using (2.5) we find

$$L(\mathbf{r}) = \frac{M}{N} \int W^2(\mathbf{r} - \mathbf{r}') \rho(\mathbf{r}') d\mathbf{r}' + \left(1 - \frac{1}{N}\right) \left[\int W(\mathbf{r} - \mathbf{r}') \rho(\mathbf{r}') d\mathbf{r}'\right]^2$$

$$+ \rho^2(\mathbf{r}) - 2\rho(\mathbf{r}) \int W(\mathbf{r} - \mathbf{r}') \rho(\mathbf{r}') d\mathbf{r}'. \qquad (2.12)$$

Since $W(\mathbf{r} - \mathbf{r}')$ is strongly peaked at $\mathbf{r} = \mathbf{r}'$, we can expand $\rho(\mathbf{r}')$ about $\mathbf{r}$. Keeping only the dominant terms, we find

$$L(\mathbf{r}) \sim \frac{M}{N} \rho(\mathbf{r}) \int W^2(\mathbf{r}') d\mathbf{r}' + \left\{\frac{\nabla^2 \rho}{6} \int W(\mathbf{r}') \, \mathbf{r}'^2 d\mathbf{r}'\right\}^2. \qquad (2.13)$$

Using (2.9) the minimum of $L(\mathbf{r})$ is found to occur at

$$h^7 = \frac{27}{N} \frac{M\rho}{(\nabla^2\rho)^2} \frac{\int K^2(\mathbf{u}) d\mathbf{u}}{\left\{\int K(\mathbf{u}) u^2 d\mathbf{u}\right\}^2}. \qquad (2.14)$$

Since ρ is unknown (2.14) cannot be used directly except to infer

$$h \propto 1/N^{1/7} \text{ and } L_{\min} \propto 1/N^{4/7}.$$

We find an appropriate choice of h to be given by

$$h = b \left(\langle \mathbf{r}^2 \rangle - \langle \mathbf{r} \rangle^2 \right)^{1/2}, \tag{2.15}$$

where b is adjustable and

$$\langle c(\mathbf{r}) \rangle = \frac{\int \rho c\, d\mathbf{r}}{M} \doteq \frac{1}{N} \sum_j c(\mathbf{r}_j).$$

In the problems we consider, the derivative of the smoothed function is required to be continuous. For this reason the second of the kernel functions in (2.10) is not useful. Of the other possible kernel functions we have concentrated on the Gaussian and spline functions.

To decide between the Gaussian and the spline kernels we took a known distribution and distributed a set of points. The goodness of fit $[\rho_N(\mathbf{r}) - \rho(\mathbf{r})]^2$ was then evaluated for various values of b. For forty points there is negligible difference between the two kernels, but for 80 points the Gaussian was much more accurate. For this reason we prefer the Gaussian kernel and use if for the results reported here.

For our stellar models we choose b by requiring the smoothed particle model to fit the known density of the spherically symmetric hydrostatic model. More elaborate procedures could be used but we have found those described to be successful for the models considered.

2.3 THE GRAVITATIONAL POTENTIAL

We use the gravitational potential ϕ defined by

$$\phi = -G \int \frac{\rho_N(\mathbf{r}')d\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.16}$$

Using (2.5)

$$\phi = -\frac{GM}{N} \sum_{j=1}^{N} \int \frac{W(\mathbf{r}' - \mathbf{r}_j)d\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.17}$$

$$I_j = \int \frac{W(\mathbf{r}' - \mathbf{r}_j)\, d\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|}, \tag{2.18}$$

can be evaluated easily noting that

$$\nabla^2 I_j = -4\pi W(\mathbf{r} - \mathbf{r}_j). \tag{2.19}$$

We find

$$\nabla\phi = -\frac{GM}{N} \sum_{j=1}^{N} \left\{ -\frac{4\pi}{u_j^2} \int_0^{u_j} W(u)\, u^2 du \right\} \nabla u_j, \tag{2.20}$$

where

$$\mathbf{u}_j = \mathbf{r} - \mathbf{r}_j.$$

For the Gaussian W defined by (2.10) (i) with $f = 1/h^2$

$$\nabla\phi = -\frac{GM}{N} \sum_{j=1}^{N} \frac{2}{u_j} \left(\frac{f}{\pi}\right)^{1/2} \left[\exp\left(-fu_j^2\right) - \frac{1}{u_j} \int_0^{u_j} \exp\left(-fu^2\right) du \right] \nabla u_j \tag{2.21}$$

and

$$I_j = \frac{2}{u_j} \left(\frac{f}{\pi}\right)^{1/2} \int_0^{u_j} \exp\left(-fu^2\right) du.$$

The equivalent formulae for the delta spline W involve polynomials, and are easier to evaluate.

2.4 GENERAL DISTRIBUTIONS

To find the smoothed version of any other scalar (or vector) field $A(\mathbf{r})$, we define the smoothed field $A_s(\mathbf{r})$ by

$$A_s(\mathbf{r}) = \int W(\mathbf{r}-\mathbf{r}') A(\mathbf{r}') d\mathbf{r}', \tag{2.22}$$

where, in general, the kernel differs from that in (2.3). Then an estimate of $A_s(\mathbf{r})$ is

$$A_N(\mathbf{r}) = \frac{M}{N} \sum_{j=1}^{N} W(\mathbf{r}-\mathbf{r}_j) \frac{A(r_j)}{\rho(\mathbf{r}_j)}. \tag{2.23}$$

The error in this estimate is $\pm\sigma$ where now

$$\sigma^2 = \frac{M^2}{N^2} \sum_j W^2(\mathbf{r}-\mathbf{r}_j) \frac{A^2(\mathbf{r}_j)}{\rho^2(\mathbf{r}_j)} - \frac{1}{N} A_N^2. \tag{2.24}$$

The approximations involved become better when $A(\mathbf{r})$ is distributed similarly to the density. This is the case for temperature and entropy, but for the magnetic field it is not in general true. To deal with this case importance sampling is useful and we discuss its application in the next subsection. Where the field has known symmetry properties antithetic variables can be used to improve the accuracy.

2.5 THE MAGNETIC FIELD AND CURRENT

According to the prescription given in Section 2.4 an estimate of the magnetic field is given by

$$\mathbf{B}_N(\mathbf{r}) = \frac{M}{N} \sum_{j=1}^{N} W(\mathbf{r}-\mathbf{r}_j) \frac{\mathbf{B}_j}{\rho_j}, \tag{2.25}$$

and an estimate of the current by

$$\mathbf{J}_N(\mathbf{r}) = \epsilon_0 c^2 \frac{M}{N} \sum_j \nabla W \times \frac{\mathbf{B}_j}{\rho_j}. \tag{2.26}$$

However it is usually the case that there is field inside and outside the star and (2.25) is then a poor approximation to $\mathbf{B}_s(\mathbf{r})$, and (2.26) is an even poorer approximation to $\mathbf{J}_s(\mathbf{r})$. To

improve the approximation we use importance sampling (Hammersley & Handscomb 1964) in the form

$$B_s(r) = B_0(r) + \int W(r - r') [B(r') - B_0(r')] \, dr', \tag{2.27}$$

where $B_0(r)$ is an approximation to $B(r)$. To obtain this approximation we solve

$$\epsilon_0 c^2 \nabla \times B_0 = J_N(r), \tag{2.28}$$

in the form

$$B_0 = B_{ext} + \frac{1}{4\pi\epsilon_0 c^2} \int \frac{J_N(r) \times (r - r') \, dr'}{|r - r'|^3}, \tag{2.29}$$

where B_{ext} is any superimposed external field. It could, for example, be the field permeating an interstellar cloud from which a star is forming. Substituting (2.26) into (2.29) we find

$$B_0(r) = B_{ext} - \frac{M}{4\pi N} \sum_{j=1}^{N} \left[\left(\frac{B_j}{\rho_j} \cdot \frac{\partial}{\partial r} \right) \frac{\partial I_j}{\partial r_j} - \frac{4\pi}{\rho_j} B_j W(r - r_j) \right], \tag{2.30}$$

where I_j is defined by (2.18).

For the Gaussian kernel (2.10)(i) with $f = 1/h^2$, the approximate field becomes

$$B_0(r) = B_{ext} + \frac{M}{N} \sum_j \left(\frac{f}{\pi} \right)^{3/2} \left\{ \frac{B_j}{\rho_j} \left[\exp(-fu^2) - \frac{1}{u^3} \int_0^u \exp(-fv^2) \, v^2 \, dv \right] \right. $$
$$\left. + u \left(\frac{B_j \cdot u}{\rho_j u} \right) \left[\frac{3}{u^4} \int_0^u \exp(-fv^2) v^2 \, dv - \frac{\exp(-fu^2)}{u} \right] \right\} \tag{2.31}$$

where $u = u_j = r - r_j$.

The field we use is

$$B_N(r) = B_0(r) + \frac{M}{N} \sum_j \frac{W(r - r_j)}{\rho_j} \{ B(r_j) - B_0(r_j) \} \tag{2.32}$$

and the current is obtained from the curl of (2.32). Thus

$$J_N(r) = \frac{\epsilon_0 c^2 M}{N} \sum_j \frac{\nabla W}{\rho_j} \times \{ 2 B(r_j) - B_0(r_j) \} + J_{ext}(r), \tag{2.33}$$

where J_{ext} is the current associated with B_{ext}. This procedure, as we show later, gives a good fit to the current and the field.

3 Equations of motion

The equations of motion of the fluid particles for a uniformly rotating polytrope of index n, with an internal magnetic field B are

$$\frac{d^2 r_j}{dt^2} = -\Gamma \frac{dr_j}{dt} - K \rho_j^{1/n - 1} \frac{(1 + n)}{n} \nabla\rho - \nabla\phi - \Omega \times (\Omega \times r) - 2\Omega \times \frac{dr_j}{dt} + \frac{J \times B}{\rho}, \quad j = 1, 2, \ldots, N \tag{3.1}$$

where the pressure $P = K\rho^{1 + 1/n}$, Ω is the angular velocity, **B** the magnetic field and **J** the current. The damping term $\Gamma d\mathbf{r}_j/dt$ has been introduced to allow static models to be calculated. For the rotating models considered here only the static structure is required, and the Coriolis term can be dropped. Using the dimensionless variables $\mathbf{x}_j$, D, τ, **b** defined by

$$\rho = \lambda D, \quad \mathbf{r}_j = \alpha \mathbf{x}_j, \quad t = \beta \tau, \quad \mathbf{B} = B\mathbf{b} \tag{3.2}$$

where

$$\alpha^2 = \frac{(n+1) K \lambda^{1/n - 1}}{4 \pi G}, \quad \beta^2 = \frac{\alpha^2 n}{K \lambda^{1/n}(1 + n)}, \tag{3.3}$$

and λ and B can be chosen for convenience. (3.1) becomes, on dropping the Coriolis term,

$$\frac{d^2 \mathbf{x}_j}{d\tau^2} = -\frac{\gamma d\mathbf{x}_j}{d\tau} - D_j^{1/n - 1} \nabla D_j - \frac{n}{4\pi} \nabla \Phi - \omega^2 \mathbf{i} \times (\mathbf{i} \times \mathbf{x}_j) + \eta \frac{(\nabla \times \mathbf{b}) \times \mathbf{b}}{D_j}, \tag{3.4}$$

where γ is a new damping constant

$$\mathbf{\Omega} = \left(\frac{4 \pi G \lambda}{n}\right)^{1/2} \omega \, \mathbf{i}, \quad \eta = \frac{\epsilon_0 c^2 n B^2}{4 \pi G \lambda^2 \alpha^2}, \tag{3.5}$$

and Φ is the scaled gravitational potential where

$$\Phi = \frac{\phi}{G \lambda \alpha^3} = \frac{\phi Q}{GM}$$

and we have chosen the value of $Q = \int D(\mathbf{x}) d\mathbf{x}$ such that were the representations of integrals by sums in Section 2 to be exact, then $D(0)$ would equal unity and λ would be the central density. Our scaled variables are therefore similar to the usual polytropic variables (Chandrasekhar 1939). However, since in our models $D(0) \neq 1$, our length scale is related to the polytropic variable ξ by $\xi = D(0)^{(n-1)/2n}|\mathbf{x}|$.

For the models considered here the magnetic field variation is calculated in the flux freezing approximation

$$\frac{d}{dt}\left(\frac{\mathbf{B}}{\rho}\right) = \left(\frac{\mathbf{B}}{\rho} \cdot \nabla\right)\mathbf{v}, \tag{3.6}$$

where d/dt is a derivative following the motion. To integrate this equation forward we replace $\mathbf{v}$ by the smoothed velocity field. Equation (3.6) has the advantage that it automatically generates the quantity $\mathbf{B}/\rho$ required at each fluid element to produce the smoothed field.

4 Numerical tests — spherical models

4.1 CONSTRUCTION OF STATIC MODELS

To construct a static model we follow the damped motion of a set of particles from some initial distribution of position and velocity until the system comes to rest. Typically the particles were initially at rest, distributed in space either according to a random Gaussian distribution or alternatively on a spherically symmetric cubic lattice. In the former case the
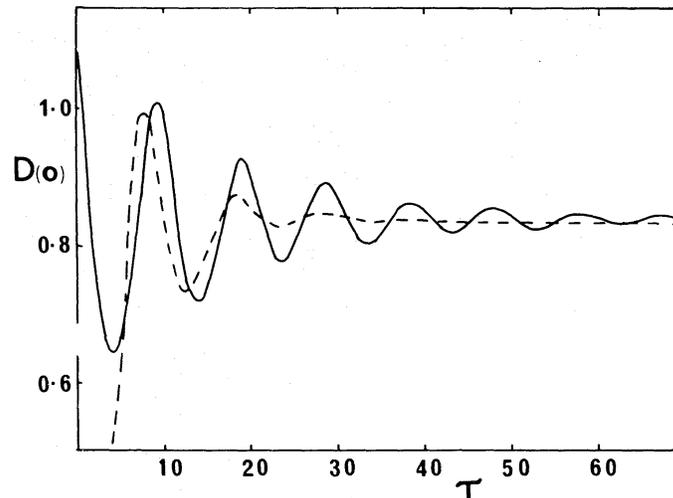
Figure 1. The central density $D(0)$ as a function of time τ for two damped hydrodynamic sequences. The initial configurations are given in the text.

initial coordinates of the particles were adjusted so that the centre of mass was at the centre of the coordinate system. As a check, the position and velocity of the centre of mass were monitored throughout the calculations.

The approach to equilibrium for two initial configurations with different degrees of damping is illustrated in Fig. 1 for a polytrope of index 1. The solid line represents the behaviour of $D(0)$ as a function of the scaled time τ, in a sequence that commences with 33 particles on a cubic lattice and $\gamma = 0.05$. The broken curve shows a sequence, with $\gamma = 0.15$, commencing with 40 particles distributed normally about the origin.

The models finally obtained are found to be nearly independent of both the damping, the initial configuration, and the number of particles. These model sequences commence, of course, with a good deal of spherical symmetry. Quite irregular initial distributions can also be successfully treated. Fig. 2(a) shows the density profile in the (X, Y) plane of an initially non-spherically symmetric distribution which leads to the symmetric distribution of Fig. 2(b) representing a polytrope of index 1.

4.2 STATIC STRUCTURE

Polytropes of index 1 and 1.5, constructed using about 40 particles and a wide range of smoothing constant b (defined by equation 2.15) were found to have density profiles which matched the true density to within a few per cent over the bulk of the star. The density profile in the outermost 10 per cent of the polytropic radius for the polytrope of index 1, however, reflects the nature of the smoothing function rather more than it does the actual distribution of matter. The size of this region can be decreased by employing a larger number of particles.

Sequences that commenced with a high degree of spherical symmetry yielded similarly highly symmetric polytropes. The departure from spherical symmetry in other cases was less than 2 per cent.

Polytropes of index 2.5 were also constructed employing both a range in the number of particles and in the value of the smoothing constant b. In Fig. 3 we display the appropriately scaled density profiles of two such models. The curve represents the true density for $n = 2.5$ while the filled circles show the density profile of a model constructed with $N = 80$. Since this model is highly symmetrical the density profile along only one axis is shown. Also

13

R. A. Gingold and J. J. Monaghan
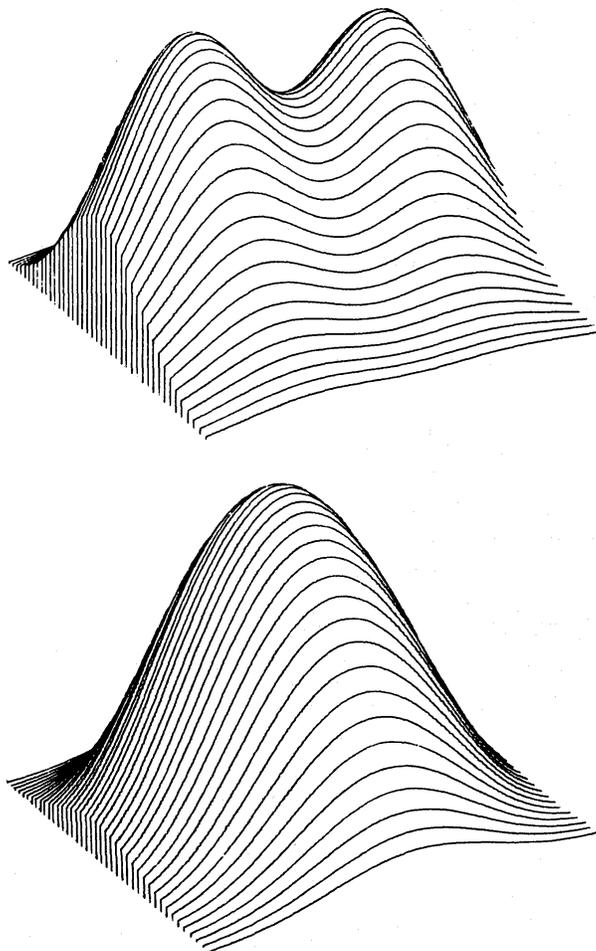
Figure 2. An example of initial and final smoothed density in the x,y plane for a polytrope of index 1. The initial state was a superposition of two Gaussian density distributions.
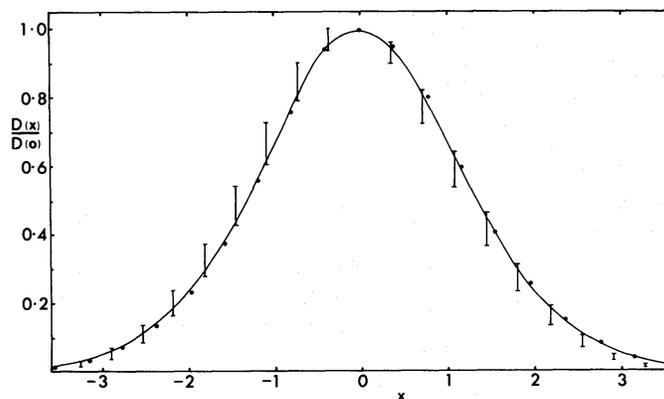
Figure 3. Density profiles for a polytrope of index 2.5. The Emden density is shown: ———; the 80-particle SPH is shown: • • • •. The variation in density for a given X along the x, y, z axes is indicated by the size of the filled circle. The analogous variation for 40 particles is shown by the bar.

indicated in Fig. 3 is the less symmetrical density profile of a model constructed with one half as many particles. The range in density at points on each of the three coordinate axes for this model is indicated by the vertical bars. The density profiles in this case fit the true density more closely than might appear from the figure. Along each of the coordinate axes

Table 1. Parameters of polytropes of index 2.5.

N	40	40	40	40	40	80	200	200	200
$1/b^2$	0.5	1.0	1.5	2.0	2.5	2.0	1.6	2.0	3.0
$1/h^2$	0.35	0.60	1.05	1.69	2.81	1.01	0.67	0.79	1.19
$D(0)$	0.72	1.22	1.97	2.97	4.92	1.52	1.06	1.08	1.34
$\langle \xi^2 \rangle$	1.17	1.88	2.15	2.27	2.31	2.55	2.47	2.65	3.01

the density profile is quite good, but the peak value is offset from the origin. Nevertheless, the improvement achieved by increasing the number of particles from 40 to 80 is remarkable and surpasses the $\sqrt{N}$ improvement we would expect in Monte Carlo integrations. This is probably due to 40 particles being intrinsically too few.

Some brief details of various models for $n = 2.5$ are given in Table 1. In each case the sequence commenced with a Gaussian distribution of particles about the origin. Although the final values of $D(0)$ vary with both the smoothing parameter and the number of particles, in each case the density profile after dividing by $D(0)$ is similar. Also displayed in Table 1 are the mean squared radial position of the particles given by

$$\langle \xi^2 \rangle = \frac{1}{N} \sum \xi_j^2 \cong \int \rho \xi^2 \, dv \Big/ \int \rho \, dv.$$

These values are smaller than the value of 4.8 obtained by performing the integrations using the density profile in the above expression. This discrepancy is not surprising since the integrand $\rho \xi^4$ has a sharp maximum beyond the position of the bulk of our particles.

4.3 UNDAMPED OSCILLATIONS

Several hydrodynamic sequences were followed with damping excluded. These were found to oscillate in a mixture of modes reflecting the initial state of the model. In each case a dominant period of oscillation of the central density was manifest. This matched the periods of oscillation of polytropes of index n in the range 1–2.5 given by Kopal (1938) to within 10 per cent for $N \sim 40$. This error can be reduced by using more particles. During extended runs over many cycles of large amplitude oscillations ($\delta D(0)/D(0) \sim 0.3$) the total energy E of systems with $N \sim 40$ was found to oscillate with $|\delta E/E| \lesssim 0.1$. This error is consistent with replacing integrals by sums according to the Monte Carlo procedure.

5 Numerical tests — non-spherical models

5.1 UNIFORMLY ROTATING POLYTROPES

Uniformly rotating polytropes were studied to determine the accuracy with which the technique reproduced a non-spherical structure.

In Fig. 4 the polar and equatorial density profiles are shown for a rapidly rotating polytrope of index 1.5 for which $\omega^2 = 0.024$. The figure also shows the density profiles for the same model obtained using the approximation technique of Monaghan & Roxburgh (1965). It is clear that the agreement is good. All models were found to be symmetric about the rotation axis and the equator to within 5 per cent.

Because our models do not have $D(0) = 1$, the parameter α of Monaghan & Roxburgh is related to our ω^2 by $\alpha = 2\omega^2/nD(0)$. The model shown is therefore on the verge of breakup. Since our method does not produce fluid particles near the edge, the critical ω corre-
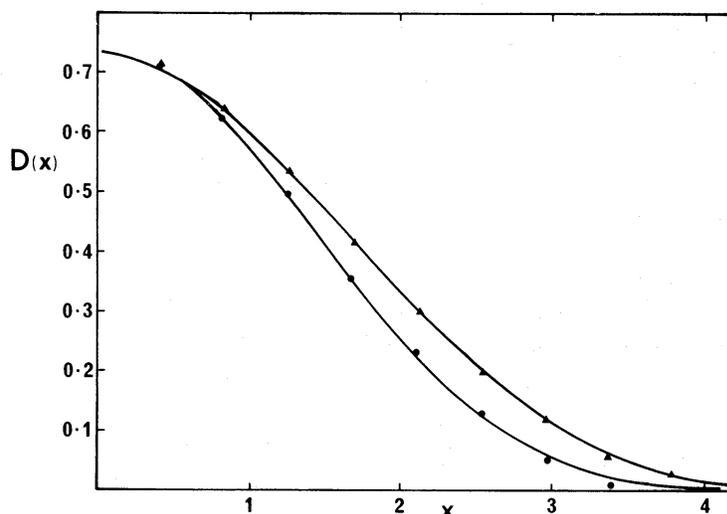
Figure 4. The density profiles for a uniformly rotating polytrope of index 1.5. The SPH results are shown thus: polar density: lower curve; equatorial density: upper curve. Perturbation analysis (Monaghan & Roxburgh 1965) shown by • • • • • and ▲ ▲ ▲ ▲ .

sponding to breakup cannot be determined accurately. Of course, with more particles, and therefore a smaller h, the critical ω can be calculated as accurately as desired. Alternatively test particles could be introduced.

5.2 MAGNETIC POLYTROPES

The static structure of polytropes with both poloidal and toroidal fields was studied by starting with a static, non-rotating, polytrope and then superimposing the field. The polytrope was then allowed to relax to a static structure. Because the main purpose of this study was to explore the numerical method we chose initial fields which were known solutions of
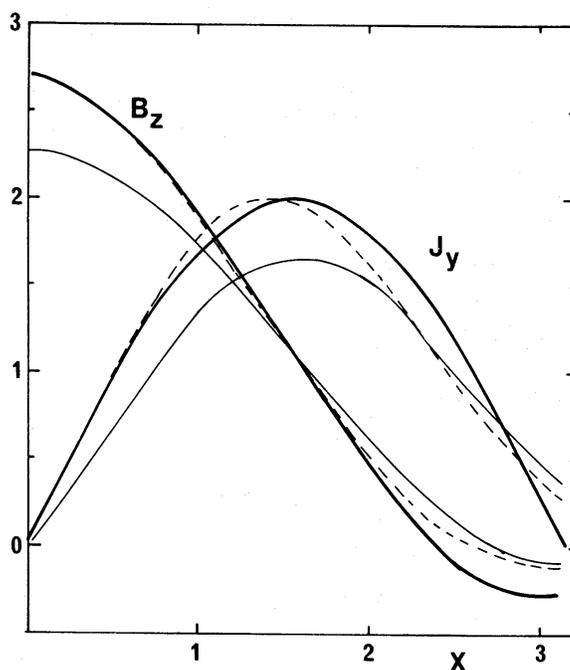
Figure 5. Poloidal magnetic field and current in a polytrope of index 1. Perturbation analysis (Monaghan 1965) shown thus: ———. The smoothed initial field and current shown thus: ------. The final static field and current shown thus: ———.
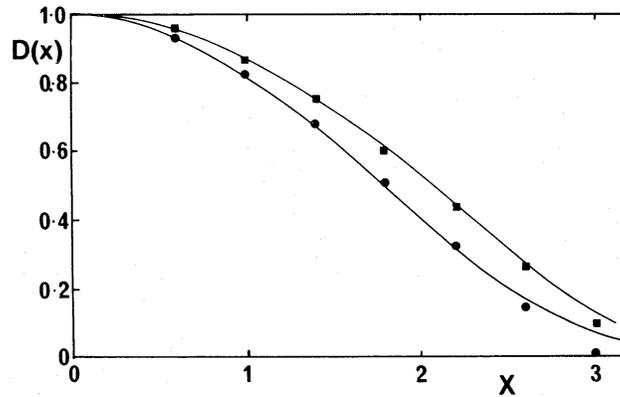
Figure 6. Density profiles for a polytrope of index 1 with a dipole poloidal field. Perturbation analysis (Monaghan 1965) shown thus: polar density: ● ● ● ●; equatorial density: ■ ■ ■. The SPH density shown thus: ———. The full field smoothing method has been used.

the first-order perturbation equations. The poloidal field was taken from Monaghan (1965) and the toroidal field from Roxburgh (1966).

In Fig. 5 we show the initial field and current on the x axis calculated from the analytical expression for a dipole field in a polytrope of index 1. Also shown is the initial and final smoothed field and current calculated according to the procedure of Section 2.5. The agreement between the initial field and its smoothed equivalent is very good. We believe it could be further improved by adjusting the smoothing parameter b, or by adopting a different value of this parameter for each component of the magnetic field.

The analytical equilibrium field is based on a first-order perturbation analysis which assumes the field can be constructed from a non-perturbed density. Since we find density perturbations of ~ 10 per cent, we expect the final field to differ from the first-order perturbation results by quantities of this order. The difference between the initial and final field and current shown in Fig. 5 is therefore not unexpected.

In Fig. 6 the equatorial and polar density profiles are shown for both the present numerical calculations, and for the first-order perturbation results. Since our models have $D(0) \neq 1$, and a field and current which differ from the analytical one by approximately a scale factor, the relation between η and the factors ϖ and k of Monaghan (1965) is approximately

$$\varpi = \frac{\eta}{n k^2 D(0)^{1+1/n}} \left(\frac{\text{computed } B_z(0)}{\text{analytical } B_z(0)} \right)^2 .$$

The agreement between the first-order perturbation results and our numerical results is very good. Small changes, of the order of 10 per cent of the deviations from the unperturbed density, are to be expected because of errors in the perturbation method, but this has a negligible effect on the density profiles.

The toroidal field investigated is zero outside the polytrope and the smoothed field can be obtained satisfactorily without using the importance sampling device of Section 2.5. With 40 particles in a polytrope of index 1 the initial smoothed field reproduced the analytical field to within $\lesssim 5$ per cent.

During the calculations the constancy of the magnetic flux was monitored and found to remain constant to within 2 per cent.

6 Computational requirements

All of the sequences described in this paper were stored in 60—90K bytes of core storage in an IBM 360/165. A typical $N = 40$ sequence with no magnetic field requires about 0.25 s per

time step and about 200 time steps (50 s) per model. The time step is fixed by calculating the minimum of h/v_{max}, $(h/F_{max})^{1/2}$ and $h/D^{1/2n}$ where v_{max} is the maximum particle velocity, F_{max} the maximum force and the last criterion is based on the speed of sound. The minimum is then multiplied by a constant. If this constant is in the range 0.1−0.5 the integration is stable without requiring excessive time. The time can be halved by storing the current forces and densities at the particle positions. We are currently performing calculations for dynamical sequences leading to fission. These require approximately four minutes of computing time for an 80-particle configuration.

Conclusions

The results of this study show that the smoothed particle method is a simple technique which gives satisfactory results for oscillating polytropes, and for polytropes which relax from a non-spherical initial state to a spherical final state. Rotation and magnetic fields may be included without difficulty, and the comparison with the perturbation results shows that moderate distortion can be reproduced accurately. Structure on a finer scale or greater accuracy can always be obtained by increasing the number of particles and by using the devices known to improve Monte Carlo integration methods.

Acknowledgment

References

Bartlett, M. S., 1963. *Sankhya (A)*, **25**, 245.
Boneva, L. I., Kendall, D. & Stepanov, I., 1971. *J. R. stat. Soc.*, **33**, 1.
Chandrasekhar, S., 1939. *An introduction to the study of stellar structure*, p. 87. University of Chicago Press.
Hammersley, J. M. & Handscomb, D. C., 1964. *Monte Carlo methods*, Methuen, London.
Kopal, Z., 1938. *Mon. Not. R. astr. Soc.*, **99**, 33.
Monaghan, J. J., 1965. *Mon. Not. R. astr. Soc.*, **131**, 105.
Monaghan, J. J. & Roxburgh, I. W., 1965. *Mon. Not. R. astr. Soc.*, **131**, 13.
Parzen, E., 1962. *Ann. Math. Stat.*, **33**, 1065.
Pasta, J. R. & Ulam, S., 1959. *Mathematical tables and other aids to computation*, **13**, 1.
Roxburgh, I. W., 1966. *Mon. Not. R. astr. Soc.*, **132**, 347.

Appendix

In one dimension the simplest representation of a sample is by a histogram. To smooth the histogram the constraints of minimizing the slope, while retaining reproducibility of the data, can be used. The resulting smoothing function is the delta spline of Boneva, Kendall & Stepanov (1971).

In three dimensions there are various possible generalizations. Our experiments have been based on the following.

Around a sample point construct the unit ball, i.e. the sphere of unit radius. This is one generalization of the unit histogram. Surround the ball by concentric shells of radius $r_i = i$. Now construct the spherical delta spline $S(r)$ by the rules

$$\text{Min } 4\pi \int_{r_i}^{r_{i+1}} \left(\frac{\partial S}{\partial r}\right)^2 r^2 dr \text{ with } 4\pi \int_{r_i}^{r_{i+1}} S r^2 dr = \delta_{i0} \qquad i = 1, 2, 3.$$

These rules ensure minimization of the slope with the constraint that the integral over all space is equal to the contribution from within the unit ball. The resulting set of equations is easily solved and the spherical delta spline is found to oscillate with an exponentially decreasing amplitude.

An alternative we haven't experimented with is based on the subdivision of the space into cubes. Then the function to be minimized is $(\partial^3 S/\partial x \partial y \partial z)^2$ and the delta spline becomes just a product of one-dimensional delta splines.

Smoothed particle hydrodynamics

J J Monaghan

School of Mathematical Sciences, Monash University, Vic 3800, Australia

E-mail: joe.monaghan@sci.monash.edu.au

Abstract

In this review the theory and application of Smoothed particle hydrodynamics (SPH) since its inception in 1977 are discussed. Emphasis is placed on the strengths and weaknesses, the analogy with particle dynamics and the numerous areas where SPH has been successfully applied.

Contents

1. Introduction

Smoothed particle hydrodynamics (SPH) is a method for obtaining approximate numerical solutions of the equations of fluid dynamics by replacing the fluid with a set of particles. For the mathematician, the particles are just interpolation points from which properties of the fluid can be calculated. For the physicist, the SPH particles are material particles which can be treated like any other particle system. Either way, the method has a number of attractive features. The first of these is that pure advection is treated exactly. For example, if the particles are given a colour, and the velocity is specified, the transport of colour by the particle system is exact. Modern finite difference methods give reasonable results for advection but the algorithms are not Galilean invariant so that, when a large constant velocity is superposed, the results can be badly corrupted. The second advantage is that with more than one material, each described by its own set of particles, interface problems are often trivial for SPH but difficult for finite difference schemes. The third advantage is that particle methods bridge the gap between the continuum and fragmentation in a natural way. As a consequence, the best current method for the study of brittle fracture and subsequent fragmentation in damaged solids is SPH (see, e.g. Benz and Asphaug (1994, 1995)). A fourth advantage is that the resolution can be made to depend on position and time, which makes the method very attractive for most astrophysical and many geophysical problems. Fifth, SPH has the computational advantage, particularly in problems involving fragments, drops or stars, that the computation is only where the matter is, with a consequent reduction in storage and calculation. Finally, because of the close similarity between SPH and molecular dynamics, it is often possible to include complex physics easily.

Although the idea of using particles is natural, it is not obvious which interactions between the particles will faithfully reproduce the equations of fluid dynamics or continuum mechanics. One way of doing this was proposed by Bob Gingold and myself (Gingold and Monaghan (1977) where the term SPH was coined) and independently by Lucy (1977). Gingold and Monaghan derived the equations of motion using a kernel estimation technique, pioneered by statisticians, to estimate probability densities from sample values (Rosenblatt (1956), Parzen (1962) and, for a general discussion, see Boneva *et al* (1971)). When applied to interpolation, this yielded an estimate of a function at any point using the values of the function at the particles. This estimate of the function could be differentiated exactly provided the kernel was differentiable. In this way, the gradient terms required for the equations of fluid dynamics could be written in terms of the properties of the particles. Because of its close relation to the statistical ideas, Gingold and Monaghan (1977) described the method as a Monte Carlo method, as did Lucy (1977) who had, in effect, re-discovered the statistical technique. However, in subsequent papers (e.g. Gingold and Monaghan (1978)), it was discovered that the errors were much smaller than the predicted probability estimates. Gingold and Monaghan realized that the particle number density was not equivalent to a probability density because the fluctuations predicted by probability theory require energy, which is not available from the equations of motion. This is particularly easy to see in the case of static equilibrium as the system moves to a minimum energy state in which large voids do not occur, since they require higher energy. In a dynamical problem more disorder can occur but only to the extent allowed by the dynamical equations.

The original papers (Gingold and Monaghan 1977, Lucy 1977) proposed numerical schemes which did not conserve linear and angular momentum exactly, but gave good results for a class of astrophysical problems that were considered too difficult for the techniques available at the time. The basic SPH algorithm was improved to conserve linear and angular momentum exactly using the particle equivalent of the Lagrangian for a compressible non-dissipative fluid (Gingold and Monaghan 1982). In this way, the similarities between SPH and

molecular dynamics were made clearer. Recent studies by Hoover (1998) and Hoover *et al* (2004) explore the correspondence between SPH and molecular dynamics.

Since SPH models a fluid as a mechanical and thermodynamical particle system, it is natural to derive the SPH equations for non-dissipative flow from a Lagrangian. The equations for the early SPH simulations of binary fission and instabilities were derived from Lagrangians (Gingold and Monaghan 1978, 1979, 1980). These Lagrangians took into account the smoothing length (the same for each particle) which was a function of the coordinates. More recent examples include Lagrangians which incorporate a resolution length for each particle (Springel and Hernquist 2002, Monaghan 2002), a relativistic Lagrangian (Monaghan and Price 2001), a Lagrangian for MHD problems (Price and Monaghan 2004a, 2004b) and a Lagrangian for SPH compressible turbulence (Monaghan 2002). In addition, Bonet and his colleagues (Bonet and Lok 1999, Bonet and Kulasegaram 2000, 2001) have used Lagrangians for the SPH simulation of elastic materials. The advantage of a Lagrangian is that it not only guarantees conservation of momentum and energy, but also ensures that the particle system retains much of the geometric structure of the continuum system in the phase space of the particles. This includes Liouville's theorem and the Poincare invariants. In addition, as noted by Dirac, basing the equations of motion on a Lagrangian allows new physical interactions to be included consistently.

The comments made by Von Neumann in 1944 (see Von Neumann (1944)), in connection with the use of the particle methods to model shocks, are relevant to SPH. To paraphrase his remarks:

> *The particle method is not only an approximation of the continuum fluid equations, but also gives the rigorous equations for a particle system which approximates the molecular system underlying, and more fundamental than the continuum equations.*

When combined with a simple but effective viscosity, and a form of the thermal energy equation that guarantees that the viscous dissipation increases both the thermal energy and the entropy, a variety of shock problems have been studied (Monaghan and Gingold 1983, Monaghan 1997, Price and Monaghan 2004a). The SPH algorithm gives very satisfactory results for shocks though they are not as accurate as those obtained from well-designed Riemann solvers and other modern techniques—although these have their own set of problems, especially when approximate Riemann solvers are used (Quirk 1994). Sharpness is often overated as a measure of the fidelity of simulations. Real shocks are only a few mean free paths thick so that, in a typical shock tube of 2 m length, $\sim 10^7$ finite difference cells, in each direction, would be required in a finite difference code to resolve the shock. However, most codes can afford only 10^3 cells along each coordinate so that their numerical shock widths are 10^4 times greater than the actual shock width. Therefore, the discussion about which code gives the sharpest shocks is irrelevant; they are all outstandingly bad. What are relevant are the pre- and post-shock values of the physical variables. SPH is able to obtain these as accurately as desired in one dimension, but in two and three dimensions SPH shocks, using current algorithms, can be noisy. In astrophysical problems, this should not be a cause for concern because the flows are invariably turbulent and the noise created in an SPH shock is small relative to that owing to turbulence.

In problems involving very small perturbations, the lower accuracy of SPH makes finite difference methods preferable. However, it has advantages which show up in those fluid problems where the perturbations are large. The first of these is that complex physics can often be included with little effort and effective codes produced in days, whereas finite difference codes would take many months or years to write. The second is that the SPH method can be easily extended to include a resolution which varies in space and time. That is, each particle has

its own resolution length (Gingold and Monaghan (1982), see their section 4). It is, therefore, ideal for astrophysical problems where enormous variations in the relevant length scales are common (see, e.g. the simulation of the formation of the Moon (Benz *et al* (1986), or the star formation studies of Bate *et al* (1995) and Bate *et al* (2003) or the binary neutron star collisions of Rosswog and Davies (2002)). Furthermore, the SPH method combines easily with the particle methods used for star systems and is a natural tool for cosmological simulations, in particular (see, e.g. Hernquist and Katz (1989), Couchman *et al* (1995), Springel and Hernquist (2002) and Marri and White (2003)).

Because SPH is essentially a technique for approximating the continuum equations, it can be used for a wide range of fluid dynamical problems. Although the initial applications were to gas dynamic problems, it has also been applied to problems in incompressible flow by treating that flow as slightly compressible with an appropriate equation of state (Monaghan 1994). Using the same idea waves, breaking on arbitrary structures (Monaghan *et al* 2004, Colagrossi and Landrini 2003) as well as the more classical problems of waves on beaches (Monaghan and Kos 1999) could be simulated. Colagrossi (2004) has made a detailed study of the application of SPH to breaking waves, where an accurate boundary element method could be used up to the point where the wave curls over to touch the water surface in the front. The SPH calculations agree with the boundary element method up to the point that it can be used, and thereafter the SPH method gives good agreement with the experiment. Colagrossi (2004) also shows that the SPH simulation of sloshing tanks and the bow waves produced by certain ship hulls are in good agreement with the experiment. Simulations of liquid metal moulding (Cleary and Ha 2002) also show good agreement with the experiment.

Another class of problems suitable for the SPH algorithm arise in elasticity and fracture. Libersky and Petschek (1991) derived and applied the SPH equations for elasticity. Benz and Asphaug (1994, 1995) showed how SPH could be applied to the fracture of brittle solids, where it gives much better results than the finite element or the finite difference methods. These methods have been applied to the breakup of planetesimals and the formation of asteroid families (Michel *et al* 2004). In these simulations, the ease with which the SPH particles can describe the transition from a continuum to a set of fragments gives it a computational edge over other numerical methods. Commerical software packages (e.g. Dyna3D and Autodyn) for simulating impact now incorporate SPH. Elastic SPH also provides a simple and robust technique for simulating complex fracture in geological rock formations and in brittle materials (Gray *et al* 2001, Gray and Monaghan 2004). SPH is also being used in virtual reality surgery (see, e.g. the work of M Mueller, S Schirm and M Teschner at the Computer Graphics Laboratory ETH, Zurich).

In many of these problems *a priori* estimates of the accuracy of SPH interpolation suggest that the simulations would give results which would be too inaccurate for most problems. As a consequence, a technique called Moving Least Squares (Dilts 1999) was developed to produce a particle code with perfect linear interpolation. However, the disadvantages are that conservation is lost and the method is considerably slower than the standard SPH. Furthermore, in practice, as noted earlier, the low accuracy predicted from interpolation errors usually does not occur. For example, Colagrossi (2004) shows that, for the complex evolution of a patch of fluid, the SPH results are as good as those from the level set method, and often surpass them. Part of the reason may be that, for non-dissipative problems, the equations follow directly from a Lagrangian, which retains many of the properties of the original continuum Lagrangian.

In problems involving heat conduction, Cleary and Monaghan (1999) showed that the SPH simulations, which conserve thermal energy and guarantee that the entropy increased, were very accurate even though the particles' positions were disordered and the thermal conductivity discontinuous. These results together with those mentioned earlier show that if SPH equations

are set up such that they satisfy the fundamental conservation laws, the results are much better than would be deduced from consideration of the interpolation alone.

The reader may find the early reviews of SPH (Monaghan 1992, Benz 1990) useful. A different aspect of SPH is detailed in the website www.nextlimit.com, which shows a wide variety of SPH simulations of fluids for both scientific problems and for video and film special effects (In the third film of the trilogy 'Lord of the Rings', Nextlimit used SPH to simulate Gollum falling into the lava.)

2. Interpolation

The equations of fluid dynamics have the form

$$\frac{\mathrm{d}A}{\mathrm{d}t} = f(A, \nabla A, \boldsymbol{r}), \tag{2.1}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla \tag{2.2}$$

is the Lagrangian derivative, or the derivative following the motion. It is worth noting that the characteristics of this differential operator are the particle trajectories.

In the equations of fluid dynamics, the rates of change of physical quantities require spatial derivatives of physical quantities. The key step in any computational fluid dynamics algorithm is to approximate these derivatives using information from a finite number of points. In finite difference methods, the points are the vertices of a mesh. In the SPH method, the interpolating points are particles which move with the flow, and the interpolation of any quantity, at any point in space, is based on kernel estimation.

2.1. Integral and summation interpolants and their kernels

SPH interpolation of a quantity A, which is a function of the spatial coordinates, is based on the integral interpolant

$$A_{\mathrm{I}}(\boldsymbol{r}) = \int A(\boldsymbol{r}') W(\boldsymbol{r} - \boldsymbol{r}', h) \, \mathrm{d}\boldsymbol{r}', \tag{2.3}$$

where the function W is the kernel and $\mathrm{d}\boldsymbol{r}'$ is a differential volume element . The interpolant reproduces A exactly if the kernel is a delta function. In practice, the kernels are functions which tend to the delta function as the length scale h tends to zero. They are normalized to 1 so that the constants are interpolated exactly. An example in one dimension x is the Gaussian kernel $W(x, h) = \exp(-x^2/h^2)/(h\sqrt{\pi})$. The Gaussian kernel was used by Gingold and Monaghan (1977) and a kernel with continuous second derivatives of the form $W(r, h) = (105/(16\pi h^3)(1 - r/h)^3(1 + 3r/h)$ in $0 \leqslant r \leqslant h$, and zero otherwise, was used by Lucy (1977) for his three-dimensional calculations. The most commonly used kernels are based on Schoenberg (1946) M_n splines, which are piece-wise continuous functions with compact support having the derivatives up to $(n - 2)$ continuous. They can be defined by the Fourier transform

$$M_n(x, h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\frac{\sin kh/2}{kh/2}\right)^n \cos(kx) \, \mathrm{d}k \tag{2.4}$$

and algebraic forms are given by Schoenberg (1946) and Monaghan (1985b). The M_2 spline with $q = |x|/h$, is

$$M_2(x) = \begin{cases} 1 - q, & \text{for } 0 \leqslant q \leqslant 1, \\ 0, & \text{for } q \geqslant 1. \end{cases} \tag{2.5}$$

M_2 gives linear interpolation but its first derivative is discontinuous. In its product form it gives what is called equal area interpolation (Hockney and Eastwood 1988). A commonly used kernel is the M_4 kernel (called the cubic spline because it is a piecewise cubic polynomial). It has the form,

$$M_4(x) = \begin{cases} \frac{1}{6}[(2-q)^3 - 4(1-q)^3], & \text{for } 0 \leqslant q \leqslant 1, \\ \frac{1}{6}(2-q)^3, & \text{for } 1 \leqslant q \leqslant 2, \\ 0, & \text{for } q > 2. \end{cases} \tag{2.6}$$

The SPH kernel associated with $M_n(x)$ in one dimension is $W(x, h) = M_n(x)/h$. In $\hat{d}$ dimensions the same functional forms are used but they are multiplied by $1/h^{\hat{d}}$ and by a constant to ensure they are normalized in the new space. For example, the factor $1/6$ in the cubic spline (2.6) is replaced by $15/(14\pi)$ in two dimensions and by $1/(4\pi)$, in three dimensions. Higher order interpolation using splines was studied by Monaghan (1985a). The higher order kernels perform very well for equi-spaced particles, but they require a cancellation of positive and negative contributions which is less likely when the particles are disordered. Furthermore, many of the desirable features of SPH involving positive definite dissipation terms are lost when higher order kernels are used because the gradient of the kernel changes sign. Schoenberg (1946) also discusses a class of smoothing kernels with Fourier transforms which have a Gaussian decay with increasing k. These have not been used in simulations.

Alternative kernels have been studied by Fulk and Quinn (1996) in one dimension. According to their measures, no kernel is significantly better than the cubic spline. Price (2004a) has studied the effect of changing the joining points in one dimension without finding a kernel significantly better than the cubic spline. In higher dimensions it is not clear whether optimum interpolation is obtained with equi-spaced joining points for the piece-wise polynomials of the M_n functions. It would be interesting to study either the functions, or their Fourier transforms, when the joining points are allowed to be arbitrary. It may be that equal volumes should be cut by the slices between the joining points.

To apply this interpolation to a fluid, we divide it into a set of small mass elements. The element a will have a mass m_a, density ρ_a and position r_a. The value of A at particle a is denoted by A_a. The interpolation integral can be written as

$$\int \frac{A(r')}{\rho(r')} \rho(r') \, dr', \tag{2.7}$$

where an element of mass is $\rho \, dr'$. The integral can then be approximated by a summation over the mass elements. This gives the summation interpolant

$$A_s(r) = \sum_b m_b \frac{A_b}{\rho_b} W(r - r_b, h), \tag{2.8}$$

where the summation is over all the particles but, in practice, it is only over near neighbours because W falls off rapidly with distance. Typically, h is close to the particle spacing and the kernel W is effectively zero beyond a distance $2h$ (as in the case of the kernel based on the cubic spline M_4). In practice, we choose kernels which have compact support, i.e. they vanish at a finite distance.

As an example of the use of kernel estimation, suppose A is the density ρ. The interpolation formula then gives the following estimate for the density at a point r

$$\rho(r) = \sum_b m_b W(r - r_b, h), \tag{2.9}$$

which shows how the mass of a set of particles is smoothed to produce the estimated density. The reader who is familiar with the technique of estimating probability densities from sample

points (Rosenblatt 1956, Parzen 1962) will see that our formula for the density is the same as their formulae for the probability density but with m_b replaced by $1/N$, where N is the number of sample points.

If h is constant, we can integrate the density estimate to give

$$\int \rho(\boldsymbol{r})\,\mathrm{d}\tau = \sum_b m_b = M, \tag{2.10}$$

which shows that mass is conserved exactly. If we allow h to vary, the integral is no longer exactly M, but the total mass is conserved because it is carried by the particles.

2.2. First derivatives

The SPH formulation allows derivatives to be estimated easily. If W is a differentiable function then (2.8) can be differentiated exactly to give

$$\frac{\partial A_{\mathrm{s}}}{\partial x} = \sum_b m_b \frac{A_b}{\rho_b} \frac{\partial W}{\partial x}. \tag{2.11}$$

In SPH the derivative is, therefore, found by an *exact derivative* of an approximate function. However, this form of the derivative does not vanish if A is constant. A simple way to ensure that it does vanish if A is constant is to write

$$\frac{\partial A}{\partial x} = \frac{1}{\Phi}\left(\frac{\partial(\Phi A)}{\partial x} - A\frac{\partial \Phi}{\partial x}\right), \tag{2.12}$$

where Φ is any differentiable function. The SPH form of (2.12) is

$$\left(\frac{\partial A}{\partial x}\right)_a = \frac{1}{\Phi_a}\sum_b m_b \frac{\Phi_b}{\rho_b}(A_b - A_a)\frac{\partial W_{ab}}{\partial x_a}, \tag{2.13}$$

which vanishes if A is constant. In this expression, and elsewhere, W_{ab} denotes $W(\boldsymbol{r}_a - \boldsymbol{r}_b, h)$. Different choices of Φ give all the versions of derivatives in the literature. For example, choosing $\Phi = 1$ gives

$$\frac{\partial A_a}{\partial x_a} = \sum_b \frac{m_b}{\rho_b}(A_b - A_a)\frac{\partial W_{ab}}{\partial x_a} \tag{2.14}$$

and choosing $\Phi = \rho$,

$$\frac{\partial A_a}{\partial x_a} = \frac{1}{\rho_a}\sum_b m_b(A_b - A_a)\frac{\partial W_{ab}}{\partial x_a}. \tag{2.15}$$

These results have immediate application to the convergence equation (often called the continuity equation, but in this review it will be called the *convergence* equation since $-\nabla \cdot \boldsymbol{v}$ is the opposite of *divergence*)

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho\nabla \cdot \boldsymbol{v}. \tag{2.16}$$

Generalizing the previous expressions for derivatives to approximate $\nabla \cdot \boldsymbol{v}$ we find

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \rho_a \sum_b \frac{m_b}{\rho_b}\boldsymbol{v}_{ab} \cdot \nabla_a W_{ab} \tag{2.17}$$

and

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \sum_b m_b \boldsymbol{v}_{ab} \cdot \nabla_a W_{ab}, \tag{2.18}$$

where v_{ab} denotes $v_a - v_b$ and ∇_a denotes the gradient taken with respect to the coordinates of particle a. This equation is the time derivative of the summation form of the density (2.9).

If (2.17) is compared with (2.18) it will be seen that the former involves ρ explicitly in the summation, whereas the latter does not. Both expressions vanish, as they should, when the velocity is constant. However, when the system involves two or more fluids with large density ratios in contact, the expression (2.17) with ρ in the summation is more accurate (Colagrossi 2004). The reason is that near an interface the summation for $\nabla \cdot v$ for one type of fluid SPH particle involves contributions from the other fluid. If we imagine the other fluid being changed for a fluid with exactly the same velocity field, and exactly the same particle positions but different density, we would still want the same estimate of $\nabla \cdot v$. However, with (2.18) the mass elements will be changed and the estimate will be different, but if (2.17) is used the ratio of mass to density will be constant and $\nabla \cdot v$ will not change. In practice, it turns out that either (2.17) or (2.18) can be used for density ratios $\leqslant 2$, but for larger density ratios it is better to use (2.17). The Lagrangian approach, which we consider later, requires that these equations for the rate of change of density with time be included as constraints. As a result, the form of the pressure forces changes with the form chosen for the density convergence equation.

Although the focus in the previous analysis has been on designing interpolation formula to achieve satisfactory accuracy it is natural with particle methods to interpret the formula in terms of interactions between SPH particles. In the present case we expect that as particles get closer their density will increase. In particular, any two particles moving closer together should give a positive contribution to their density. In either form of the convergence equation, we can write

$$\nabla_a W_{ab} = \boldsymbol{r}_{ab} F_{ab}, \tag{2.19}$$

where $F_{ab} \leqslant 0$ is a function of $|\boldsymbol{r}_{ab}|$. The contribution of particle b to the density of particle a in (2.18) is then

$$\rho_a \frac{m_b}{\rho_b} \boldsymbol{v}_{ab} \cdot \boldsymbol{r}_{ab} F_{ab} \tag{2.20}$$

and if the particles a and b are approaching each other (so that $\boldsymbol{v}_{ab} \cdot \boldsymbol{r}_{ab} \leqslant 0$) the contribution to the density change is positive definite as expected. The same is true for (2.17).

2.3. Second derivatives

As in the case of first derivatives, second derivatives can be estimated by differentiating an SPH interpolant twice. For example, in a heat conduction problem in one dimension, the second derivative of the temperature T at the position of particle a can be estimated by

$$\left(\frac{\mathrm{d}^2 T}{\mathrm{d}x^2}\right)_a = \sum_b m_b T_b \frac{\mathrm{d}^2 W(x_a - x_b, h)}{\mathrm{d}x_a^2}. \tag{2.21}$$

However, this expression has a number of disadvantages. First, it is very sensitive to particle disorder. Second, the transfer of heat to particle a from particle b may be positive or negative depending on their separation because the second derivative of the kernel can change sign. Physics tells us that a hot particle should transfer heat to a cold particle no matter what the separation. Another disadvantage is that this expression will not result in conservation of thermal energy in an adiabatic enclosure.

A much better approach (Brookshaw 1985, Cleary and Monaghan 1999) is to begin with an integral approximation to the second derivative. For example, starting with

$$I = \int (\kappa(\boldsymbol{r}) + \kappa(\boldsymbol{r}'))(T(\boldsymbol{r}) - T(\boldsymbol{r}')) F(|\boldsymbol{r} - \boldsymbol{r}'|) \,\mathrm{d}\boldsymbol{r}', \tag{2.22}$$

where $\boldsymbol{q}F(|\boldsymbol{q}|) = \nabla W(\boldsymbol{q}, h)$, expanding $\kappa(\boldsymbol{r}')$ and $T(\boldsymbol{r}')$ in a Taylor series about $\boldsymbol{r}$, and keeping up to second order terms, we find

$$I = \nabla \cdot (\kappa \nabla T) + O(h^2). \tag{2.23}$$

The SPH form of I for particle a is

$$I = \sum_b \frac{m_b}{\rho_b}(\kappa_a + \kappa_b)(T_a - T_b)F_{ab}. \tag{2.24}$$

Because $F \leqslant 0$ this expression has the property that if $T_a > T_b$, then heat will flow from particle a to b (that is the contribution to $\mathrm{d}T_a/\mathrm{d}t < 0$) and vice versa. Other second derivatives can be calculated using similar integral expressions.

2.3.1. Second derivatives in two dimensions.

To obtain second derivatives integrals of the form

$$J_{xx} = \int \frac{\Delta x \Delta x}{\Delta r^2}(\kappa(\boldsymbol{r}) + \kappa(\boldsymbol{r}'))(T(\boldsymbol{r}) - T(\boldsymbol{r}'))F \, \mathrm{d}\boldsymbol{r}' \tag{2.25}$$

are used (Español and Revenga 2003). Here $\Delta x = x - x'$ and $\Delta r = |\boldsymbol{r} - \boldsymbol{r}'|$. Expanding the κ and T terms in a Taylor series gives, to $O(h^2)$,

$$J_{xx} = \kappa \left(\tfrac{3}{4}T_{xx} + \tfrac{1}{4}T_{yy}\right) + \tfrac{3}{4}\kappa_x T_x + \tfrac{1}{4}\kappa_y T_y \tag{2.26}$$

and

$$J_{yy} = \kappa \left(\tfrac{3}{4}T_{yy} + \tfrac{1}{4}T_{xx}\right) + \tfrac{3}{4}\kappa_y T_y + \tfrac{1}{4}\kappa_x T_x, \tag{2.27}$$

such that

$$J_{xx} + J_{yy} = \nabla \cdot (\kappa \nabla T). \tag{2.28}$$

Furthermore,

$$J_{xy} = \tfrac{1}{4}(2\kappa T_{xy} + T_x \kappa_y + T_y \kappa_x). \tag{2.29}$$

If we construct the J integrals taking $\kappa = 1$, we get estimates for the second derivatives of T in the form (now using tensor notation for the coordinates denoted by x^i and $\Delta x^i = (x^i - x'^i)$)

$$\frac{\partial^2 T}{\partial x^i \partial x^j} = \int \left[4\frac{\Delta x^i \Delta x^j}{\Delta r^2} - \delta^{ij}\right](T(\boldsymbol{r}) - T(\boldsymbol{r}'))F \, \mathrm{d}\boldsymbol{r}' \tag{2.30}$$

or, in SPH form

$$\left(\frac{\partial^2 T}{\partial x^i \partial x^j}\right)_a = \sum_b \frac{m_b}{\rho_b}\left(4\frac{\Delta x^i \Delta x^j}{\Delta r^2} - \delta^{ij}\right)(T_a - T_b)F_{ab}. \tag{2.31}$$

2.3.2. Second derivatives in three dimensions.

With the same definition of J_{xx} as before, but now integrating over three dimensions we find

$$J_{xx} = \tfrac{1}{5}\kappa(3T_{xx} + T_{yy} + T_{zz}) + \tfrac{1}{5}(3\kappa_x T_x + \kappa_y T_y + \kappa_z T_z), \tag{2.32}$$

with similar expressions for J_{yy} and J_{zz}. The integral J_{xy} becomes

$$J_{xy} = \tfrac{1}{5}(2\kappa T_{xy} + T_x \kappa_y + T_y \kappa_x). \tag{2.33}$$

These results show that

$$J_{xx} + J_{yy} + J_{zz} = \nabla \cdot (\kappa \nabla T) + O(h^2). \tag{2.34}$$
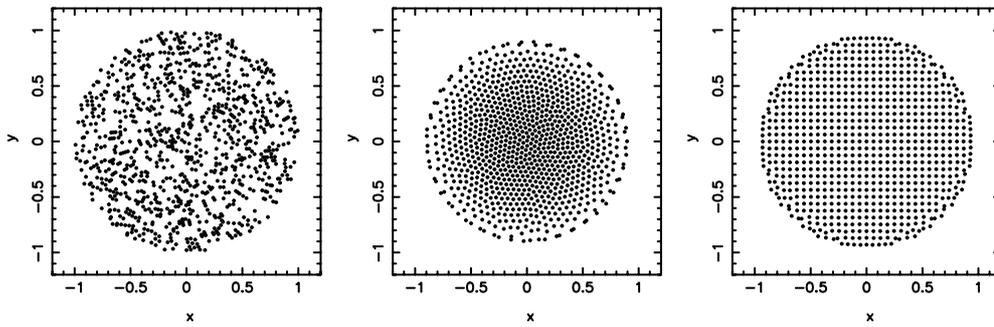
Figure 1. The frame on the left shows SPH particles placed at random according to a constant probability density within a unit circle. Note the large voids. The middle frame shows the positions of equal mass particles settled down in a Toy star potential starting with the particle positions in the left frame. The right-hand frame shows variable mass SPH particles also settled down in a Toy star potential. These latter particles were initially placed on a cubic lattice with cell length Δp and given a mass $\rho(\Delta p)^2$ using the theoretical static density. The cubic spline kernel was used and h was calculated selfconsistently (see section 4).

A similar expression to (2.30) with the factor 4 replaced with 5 in the integral, gives the second derivatives of T. These results generalize those of Español and Revenga (2003) who work out the case where κ is a constant. The application of these derivatives to problems involving viscous effects and thermal conduction will be considered later.

2.4. Errors in the integral interpolant

It is not easy to estimate the errors in the SPH equations from first principles because the particles get disordered during motion. The errors depend on the type of disorder which, in turn, depends on the dynamics. One approach to estimating the errors is to begin with particles on a regular lattice, then give each particle a random shift in position (Colagrossi 2004). However, this kind of short wavelength disorder does not usually occur if the particle spacing is much smaller than the dominant length scales of the motion. For example, if the particles are damped to an equilibrium, they fall into a nearly regular cell structure which depends, in general, on the kernel being used and the masses of the particles (see, e.g. figure 1). If the particles are in motion, for example, in a breaking wave, most of the particles are in a type of nearly ordered array associated with shearing a regular array of particles. The end result is that SPH simulations are much more accurate than the interpolation of quantities from randomly disordered particle arrays would suggest. For that reason, it is better to run carefully designed test cases to assess the accuracy of an SPH simulation. However, it is still interesting to study the kernel interpolation on a regular array of points.

Starting with the integral interpolant in one dimension

$$A_{\mathrm{I}}(x) = \int A(x')W(x-x',h)\,\mathrm{d}x' = A(x) + \int (A(x')-A(x))W(x-x',h)\,\mathrm{d}x'. \qquad (2.35)$$

The error can be estimated by a Taylor series expansion of $A(x')$. Assuming $W(q,h)$ is an even function of q, the interpolant gives

$$A_{\mathrm{I}}(x) = A(x) + \frac{\sigma h^2}{2}\frac{\mathrm{d}^2 A(x)}{\mathrm{d}x^2}\cdots, \qquad (2.36)$$

where σ is a constant dependent on the kernel. The integral interpolant, therefore, gives at least a second order interpolation. The interpolation is better if σ is zero, in which case higher

order terms in the Taylor series expansion must be included. The third order term vanishes because of symmetry leaving a possible fourth order term. All these results assume that the integrals can be extended to the entire volume within the support of the kernel. If this is not possible, for example, near a boundary, the error is larger.

Monaghan (1985a) gave a technique for constructing higher order kernels from lower order kernels using a variant of Richardson extrapolation. An example is the kernel

$$W(x, h) = \frac{1}{h\sqrt{\pi}} \left(\frac{3}{2} - \frac{x^2}{h^2} \right) e^{-x^2/h^2}, \tag{2.37}$$

which is based on the Gaussian. For this kernel, the integral interpolant is accurate to $O(h^4)$. This kernel changes sign; a necessary feature of higher order interpolation. Unfortunately this may have unwanted side effects, including the possibility that the density might become negative near a strong shock. It would, however, be possible to use a high order kernel using a switch from high to low order kernels near shocks. Such a technique has been used but not fully explored.

2.5. Errors in the summation interpolant

If the particles are equi-spaced in one dimension, we can easily estimate the errors in the summation interpolant using the Poisson summation formula

$$\sum_{j=-\infty}^{\infty} f(j) = \int_{-\infty}^{\infty} f(j)\, \mathrm{d}j + 2 \sum_{r=1}^{\infty} \int_{-\infty}^{\infty} \cos(2\pi r j) f(j)\, \mathrm{d}j, \tag{2.38}$$

where, on the right-hand side, j is treated as a continuous quantity.

Consider the interpolation of the function $g(x) = \alpha + \beta x$ with the particles equi-spaced with spacing Δ along an infinite line so that $\rho = 1$ and $m = \Delta$. The SPH interpolation formula gives, at $x_a = a\Delta$, the following expression for g at the point $x = a\Delta$.

$$\Delta \sum_{j=-\infty}^{\infty} (\alpha + \beta j \Delta) W(a\Delta - j\Delta, h). \tag{2.39}$$

If the origin is shifted to the point $a\Delta$ and the Poisson summation formula is used together with the assumption that the kernel is an even function, (2.39) becomes

$$(\alpha + \beta a \Delta) \left(\int_{-\infty}^{\infty} W(q, h)\, \mathrm{d}q + 2 \int_{-\infty}^{\infty} \cos\left(\frac{2\pi q}{\Delta} \right) W(q, h)\, \mathrm{d}q + \cdots \right). \tag{2.40}$$

This formula shows how the error depends on the Fourier transform of the kernel (Schoenberg 1946). If the kernel is a Gaussian, the previous expression becomes

$$(\alpha + \beta a \Delta)(1 + 2 e^{-\pi^2 h^2/\Delta^2} + \cdots) \tag{2.41}$$

In this simple case, we conclude that the SPH summation interpolant does not even interpolate a constant exactly, but the error is exponentially small and is negligible if $h > \Delta$. If we have any sufficiently smooth kernel the Fourier transform decreases rapidly and the error can be made negligible. The frequently used cubic spline kernel gives the following expression for the previous interpolation:

$$(\alpha + \beta a \Delta) \left(1 + 2 \left(\frac{\sin \pi h/\Delta}{\pi h/\Delta} \right)^4 + \cdots \right). \tag{2.42}$$

In this case the dominant error terms vanish if $h = \Delta$ and are small if $h > \Delta$.

Of greater interest are the errors in the derivative. For the previous case, and using (2.13) we find $\mathrm{d}g/\mathrm{d}x$ is estimated by

$$\beta\Delta\sum_{j=-\Delta\infty}^{\infty}j\Delta\frac{\partial W(a\Delta-j\Delta,h)}{\partial x_a}. \tag{2.43}$$

Using the Poisson summation formula again and shifting the origin of the summation to a, we find the derivative is given by

$$\beta\left(1-\int_{\infty}^{\infty}q\frac{\partial W}{\partial q}\cos\left(\frac{2\pi q}{\Delta}\right)\mathrm{d}q+\cdots\right). \tag{2.44}$$

The error now involves the Fourier transform of the gradient of the kernel and is larger than in the case of the function interpolation. In the case of the Gaussian, the errors are again exponentially small and are negligible if $h > \Delta$. We conclude from these results (which may be easily extended to 2 or more dimensions) that the SPH interpolation is as accurate as desired provided the particles are equi-spaced in an infinite space. This has led some researchers (Chaniotis *et al* 2002) to use re-meshing strategies for SPH, and their simulations of homogeneous fluids give very good results. At fixed boundaries they use one-sided interpolation which works well. However, boundaries, such as free surface liquid problems, then require special care as do multi-phase and multi-material problems.

2.6. Errors when the particles are disordered

During the course of an SPH calculation the particles become disordered. The exact form of this disorder depends on the dynamics. When Bob Gingold and I first ran the SPH calculations, we thought that the disorder could be described by a probability distribution proportional to the mass density and that the errors could be estimated in the same way as a Monte Carlo estimate. In particular, we expected that the errors arising from fluctuations would be $\sim 1/\sqrt{N}$, where N is the number of particles. However, the errors were much smaller than this estimate would suggest. The reason for the smaller errors, as mentioned earlier, is that the probability estimates allow fluctuations which are inconsistent with the dynamics. The result is that the SPH particles are disordered, but in an *orderly* way.

For example, the left frame of figure 1 shows the positions of 971 particles with equal mass placed at random within a unit circle. The middle frame shows the same particles after they have been allowed to evolve in a simple linear force field, where the equation of motion is

$$\frac{\mathrm{d}\boldsymbol{v}_a}{\mathrm{d}t}=-\nu\boldsymbol{v}_a-\sum_b m_b\left(\frac{P_a}{\rho_a^2}+\frac{P_b}{\rho_b^2}\right)\nabla_a W_{ab}-\boldsymbol{r}_a, \tag{2.45}$$

where the term $-\nu\boldsymbol{v}$ damps the motion, the terms involving the pressure P approximate the pressure gradient (discussed in detail in section 3) and the last term is the body force. In this force field the exact density varies with radius r according to $(1-r^2)$. The particles are still disordered but the large voids and concentrations appearing in the left frame of figure 1 have disappeared, and the disorder is far from random. The set of particles gives a density field shown in figure 2. Because the density is accurate, we can deduce that the gradients of the pressure field are accurately computed in spite of the disorder in the particles.

The simulation can be set up differently by choosing particle positions for particles which begin on a lattice of square cells with sides of length Δp. The particles have mass $\rho\Delta p^2$, which varies over the domain. Only the particles that have a radius $r < (1-0.5\Delta p)$ are kept such that no particle has zero mass. In the previous case the particles had equal mass and their spacing varied. Now the particles have different masses but equal initial spacing. After evolving the particles with the same damping as before, the particle positions settle into the state shown in
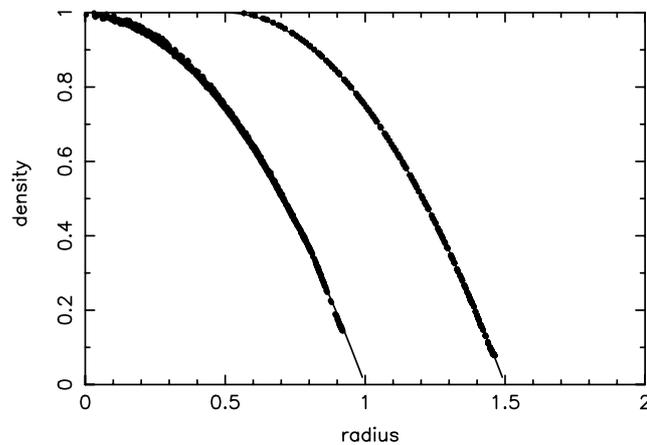
Figure 2. The density radius relation for an SPH simulation of a gas in a linear force field. The exact results are shown by the curved lines and the SPH results by the filled circles. The left-hand curve is for the equal mass particle case and the right-hand curve (shifted for clarity) is for the variable mass particles with nearly equal separation.

the right-hand frame of figure 1. As expected, by analogy with atomic systems, the difference in the force between pairs produces a different particle equilibrium configuration.

Since the disorder depends on the dynamics, it is not possible to make traditional error estimates like those used for finite differences or finite elements. In fact, the previous examples show that, at least for equilibrium configurations, the SPH particles seek the best positions for a given interpolation formula. This is a profoundly different picture from that for finite differences, where the best interpolation formula is sought for a given grid. For this reason, estimates of SPH calculations have had to depend on comparisons with known solutions, comparisons with experiments or by studying how the error varies with particle number (see, e.g. Cleary and Monaghan (1999)). These comparisons show that it is possible to achieve very accurate results with SPH. An example is given in figure 3, where the function $r^2 \exp(-6r^2)$ is interpolated using the cubic spline and the interpolation formula (2.8) with the particle positions shown in the central frame of figure 1.

The calculation of derivatives is less accurate except for the calculation of the density derivatives from the pressure force term in the equation of motion. That derivative is accurate because the particles are forced to move to an equilibrium position where the density gradient is determined accurately to balance the applied force which is $\propto \boldsymbol{r}$. If the derivative with respect to x of $(r^2 - 1)$ is calculated using (2.14), the results are shown in figure 4. The lower curve is for the case of equal mass particles and the upper curve for variable mass (the graphs are shifted by one for clarity). Only the particles within 0.9 of the outer radius were used for these plots. These particles comprise 96% of the mass. The mean square error in the gradient is 0.02.

One reason for the accuracy of SPH despite the disorder in dynamical problems is that it is possible to devise SPH algorithms so that they conserve important quantities like momentum and energy. The importance of this conservation shows up in simple problems involving the integration of ordinary differential equations. Suppose, for example, that we wish to integrate the equations for a binary star system with the stars treated as points and we are offered either a Verlet symplectic integrator (since the system is Hamiltonian) or a standard fourth order Runge–Kutta integrator. The Runge–Kutta scheme is of higher order so that, if we use the same time step in each case, a numerical analyst might argue that the Runge–Kutta will give more *accurate* results than the Verlet integrator. However, the Runge–Kutta scheme
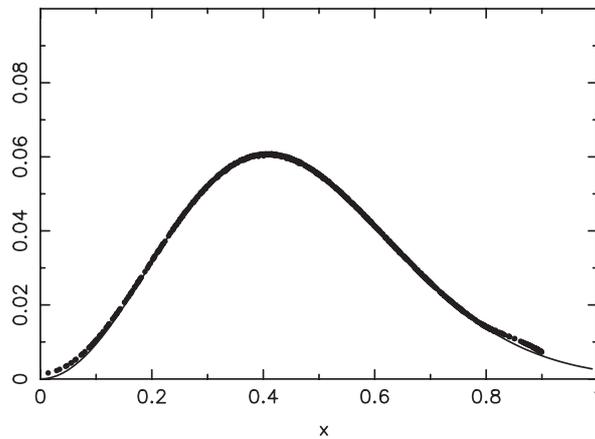
Figure 3. The function $r^2 \exp(-6r^2)$ interpolated using a cubic spline and the particle distribution shown in figure 2. The continuous line is the exact result. The dots show the SPH results.
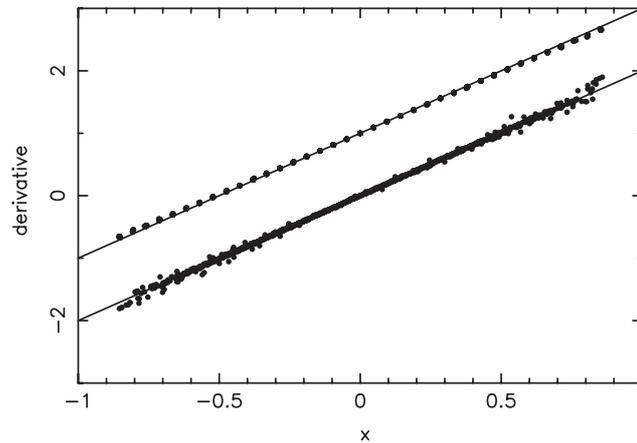
Figure 4. The SPH derivative with respect to x of $(r^2 - 1)$ using (2.14). The upper points are for the particles with initially equal spacing. They have been shifted by 1.0 for clarity. The lower points are for the case of equal mass particles. The points included have $r < 0.9$ and contain 96% of the mass.

produces a less accurate orbit. The effect is more extreme as the eccentricity gets closer to 1. The problem arises because the standard fourth order Runge–Kutta does not conserve angular momentum (it is also not reversible, whereas the system is). On the other hand, the symplectic integrator, which is a lower order integrator, gives much better results because it conserves angular momentum and is reversible. In this example the order of the integrator is less important than the conservation. It turns out that in SPH simulations, and in molecular dynamics, integrators which give very good conservation are to be preferred over higher order integrators which do not have good conservation properties. For these reasons it is preferable to write the gradient terms of SPH algorithms so that conservation is very accurate.

An example of the accuracy of SPH in a complex evolution of a liquid is due to Colagrossi (2004) and shown in figure 5. The liquid is initially in the shape of a square. The initial velocity field for a square with initial side length L is

$$(v_x, v_y) = (V(e^{-(4y/L)^2} - e^{-4}), -V(e^{-(4x/L)^2} - e^{-4})), \tag{2.46}$$
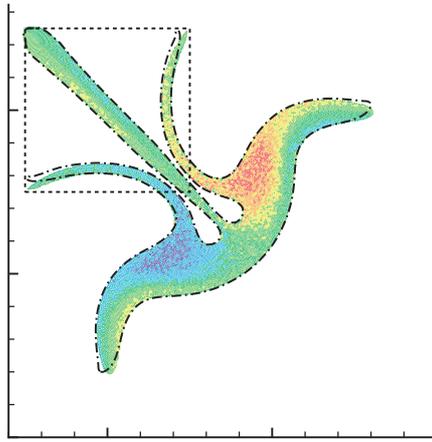
Figure 5. The evolution of an initial square of liquid (- - - -) computed using SPH and a combination
of level set and finite difference methods (— · —, Colagrossi (2004)).

(This figure is in colour only in the electronic version)

has spatially varying vorticity and produces severe distortion of the original square. The
vertices initially have zero velocity and three of the vortices remain at rest. The dash–dot lines
show the position of the outer boundary calculated using a combination of level set and finite
difference techniques. The agreement between the two methods is remarkably good and shows
that SPH is capable of simulating complex flows very satisfactorily.

An alternative approach to accuracy is taken by Vila and Lanson and their colleagues
(Ben Moussa *et al* 1999) who extend the idea of Johnson and Beissel (1996) to use normalized
kernels, but do so within the framework of an estimate of error bounds. This approach is
more mathematical than the applied mathematical approach described in this review. As a
result, these authors can get rigorous bounds on errors but with coefficients which cannot
be determined accurately. Kahan (1980), in a witty discussion of the problems of estimating
errors, comments on the pessimistic nature of error bounds and the options for estimating them
accurately, in the following terms:

> 'Both options are often so pessimistic and so costly that most people prefer to take their
> chances with computations carried out with precisions believed, rightly or wrongly,
> to exceed by far what is necessary. Their attitude makes sense; they would rather
> believe the error to be negligible than know how big it isn't'.

However, the analysis of Villa and Lanson has led them to a re-appraisal of the SPH
normalization of Johnson and Beissel with promising results.

3. SPH Euler equations

The Euler equations are the equations for the rates of change of velocity, density and position,
namely,

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla P + \boldsymbol{g}, \tag{3.1}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho\nabla\cdot\boldsymbol{v}, \tag{3.2}$$

$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \boldsymbol{v}, \tag{3.3}$$

where v is the velocity, ρ the density, P the pressure and g is the body force per unit mass. In this equation the time derivative is the derivative following the motion. In general, P is a function of ρ and the thermal energy but in the present case where there is no dissipation, the pressure can be taken as a function of ρ and the entropy per unit mass s, which remains unchanged during the motion. In some cases we will assume the entropy is the same for all particles, but, in general, each particle could have a different entropy.

The equation for the rate of change of density and its SPH equivalent have been discussed earlier. The SPH acceleration equation is discussed in the following sections.

3.1. The SPH acceleration equation

The original forms of SPH (Gingold and Monaghan 1977, Lucy 1977) converted the acceleration equation into SPH by writing

$$(\nabla P)_a = \sum_b m_b \frac{P_b}{\rho_b} \nabla_a W_{ab}, \tag{3.4}$$

such that

$$\frac{\mathrm{d} v_a}{\mathrm{d} t} = -\frac{1}{\rho_a} \sum_b m_b \frac{P_b}{\rho_b} \nabla_a W_{ab}. \tag{3.5}$$

However, (3.5) does not conserve linear or angular momentum exactly, since the force on particle a owing to b is not equal and opposite to the force on b owing to a or

$$\frac{m_a m_b P_b}{\rho_a \rho_b} \nabla_a W_{ab} \neq -\frac{m_a m_b P_a}{\rho_a \rho_b} \nabla_b W_{ab}, \tag{3.6}$$

because $P_a \neq P_b$. Note that $\nabla_a W_{ab} = -\nabla_b W_{ab}$.

To write the acceleration equation in a form which conserves linear and angular momentum the original approach was to make use of a Lagrangian (Gingold and Monaghan (1978, 1979) and in more detail Gingold and Monaghan (1982)). However, the same result is obtained by noting that

$$\frac{\nabla P}{\rho} = \nabla \left(\frac{P}{\rho} \right) + \frac{P}{\rho^2} \nabla \rho. \tag{3.7}$$

Using the SPH interpolation rules, (3.7) becomes

$$\frac{\mathrm{d} v_a}{\mathrm{d} t} = -\sum_b m_b \left(\frac{P_b}{\rho_b^2} + \frac{P_a}{\rho_a^2} \right) \nabla_a W_{ab}. \tag{3.8}$$

Writing

$$\nabla_a W_{ab} = r_{ab} F_{ab}, \tag{3.9}$$

where F_{ab} is a scalar function of $|r_a - r_b|$, the force on a owing to b is then

$$m_a m_b \left(\frac{P_b}{\rho_b^2} + \frac{P_a}{\rho_a^2} \right) r_{ab} F_{ab}, \tag{3.10}$$

which is equal and opposite to the force on b owing to a. As a consequence, linear and angular momentum are conserved exactly if h is constant or a symmetric function of a and b. It is possible to maintain this conservation even when h is allowed to vary (see later).

This pair force is actually a disguised many-body force because the pressure and density depend on the distribution of the particles and, in general, the resolution length also depends on the particle number density. The result is that, in general, the dynamics of an SPH system differs from an atomic or molecular system which can be approximated by pure pair forces.

3.2. The energy equations

The assumptions of the Euler equation do not require the time rate of change of thermal energy to be calculated. However, it is convenient to convert the non-dissipative rate of change of thermal energy to its SPH form. From the first law of thermodynamics

$$T \, \mathrm{d}s = \mathrm{d}u + P \, \mathrm{d}V, \tag{3.11}$$

$$= \mathrm{d}u - \frac{P}{\rho^2} \, \mathrm{d}\rho, \tag{3.12}$$

where s is the entropy and all quantities are per unit mass the time rate of change of thermal energy is

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{P}{\rho^2} \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\frac{P}{\rho^2} \nabla \cdot \boldsymbol{v}. \tag{3.13}$$

Using the SPH form for $\nabla \cdot \boldsymbol{v}$ given earlier, the previous equation can be written either as

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\rho_a^2} \sum_b m_b \boldsymbol{v}_{ab} \cdot \nabla_a W_{ab} \tag{3.14}$$

or

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\rho_a} \sum_b \frac{m_b}{\rho_b} \boldsymbol{v}_{ab} \cdot \nabla_a W_{ab}. \tag{3.15}$$

A good general principle when writing SPH equations is to approximate the same quantity is the same way in all the equations. For example, in the equation for the rate of change of thermal energy, the particular expression for $\nabla \cdot \boldsymbol{v}$ should be the same as that used in the time rate of change of the density.

In addition to an equation for the thermal energy, it is useful to consider the equation for the thermokinetic energy per unit mass defined by

$$\hat{e} = \tfrac{1}{2} v^2 + u. \tag{3.16}$$

The rate of change of $\hat{e}$ with time can be deduced from equations for the acceleration and the rate of change of u. The continuum equation derived in this way is

$$\frac{\mathrm{d}\hat{e}}{\mathrm{d}t} = -\frac{1}{\rho} \nabla \cdot (P\boldsymbol{v}). \tag{3.17}$$

Following the same procedure, but now using the SPH equations we find

$$\frac{\mathrm{d}\hat{e}_a}{\mathrm{d}t} = -\sum_b m_b \left(\frac{P_a \boldsymbol{v}_b}{\rho_a^2} + \frac{P_b \boldsymbol{v}_a}{\rho_b^2} \right) \cdot \nabla_a W_{ab}. \tag{3.18}$$

The continuum limit of this SPH equation is

$$\frac{\mathrm{d}\hat{e}}{\mathrm{d}t} = -\frac{P}{\rho^2} \nabla \cdot (\rho \boldsymbol{v}) - \boldsymbol{v} \cdot \nabla \left(\frac{P}{\rho} \right) = -\frac{1}{\rho} \nabla \cdot (P\boldsymbol{v}). \tag{3.19}$$

Calculations of shock phenomena with finite difference methods often use the thermokinetic energy equation rather than the thermal energy equation because it ensures conservation of the energy. Furthermore, in relativistic problems, it is natural to work with momentum and energy equations which guarantee conservation of momentum and thermokinetic energy. Because of the symmetry of the SPH equation, the rate of change with time of the total thermokinetic energy $\sum_a m_a \hat{e}_a$ is zero.

4. Resolution varying in space and time

In the original calculations of Gingold and Monaghan (1977), each particle had the same h proportional to $(\langle r^2 \rangle - \langle r \rangle^2)^{1/2}$ where, for example, $\langle r^2 \rangle$ denotes the mass average

$$\langle r^2 \rangle = \frac{\sum_b m_b r_b^2}{\sum_b m_b}. \tag{4.1}$$

During a simulation, h is then automatically increased as the particle system expands and decreased as it contracts. In their binary fission calculations, Gingold and Monaghan (1978) used an h proportional to the inverse of the gravitational energy of the system. These two choices were crude attempts to automatically match the resolution length h to the scale of the system. Gingold and Monaghan (1982) suggested that it would be preferable to allow h_a for any particle a to be related to the density according to

$$h_a = \sigma \left(\frac{m_a}{\rho_a} \right)^{1/d}, \tag{4.2}$$

where d is the number of dimensions and σ is a constant ~ 1.3. This has proved to be a powerful and robust way of specifying the resolution length h. It automatically gives SPH a resolution which varies in time and space and, if used consistently, leads to SPH equations which can be derived from a Lagrangian.

If the density is determined by summation, the density for particle a can be written as

$$\rho_a = \sum_b m_b W_{ab}(h_a). \tag{4.3}$$

The usual approach in the literature is either to calculate h_a at any time using the current value of ρ_a (estimated from the SPH summation), or to calculate h_a from the rate of change of density according to

$$\frac{\mathrm{d} \ln h}{\mathrm{d}t} = -\frac{1}{d} \frac{\mathrm{d} \ln \rho}{\mathrm{d}t}. \tag{4.4}$$

Various techniques may then be used to adjust the h_a. For example, Steinmetz and Mueller (1993) average the local density and use this to change h. Another often used alternative is to adjust h so that each particle has a constant number of neighbours (Hernquist and Katz 1989).

Ideally, h should be determined from the summation equations so that it is consistent with the density obtained from the summation (Monaghan 2002). Equation (4.3) is a non-linear equation for the single variable ρ_a, which can be solved rapidly by point iteration possibly combined with a Newton–Raphson scheme. For example, in the case of a Toy star potential, starting with random positions in the left frame of figure 1, the mean square error in solving (4.3) is reduced by a factor 10 each point iteration, and one iteration is often sufficient. Further iterations are only required for a sub-set of the particles and the time required for extra iterations is not much (Price 2004b).

In some problems it might be necessary to replace (4.2) by a formula that limits how large or small h can become. For example, an upper bound on h_a when ρ_a becomes very small is desirable to prevent strong interactions between a very low and a very high density region. This can be achieved if (4.2) is replaced by

$$h_a = \sigma \left(\frac{m_a}{A + \rho_a} \right)^{1/d}, \tag{4.5}$$

where A is a suitable constant. A lower bound can be, similarly, included. In all cases (4.3) can be solved consistently.

5. Lagrangian equations

The Lagrangian L for the non-dissipative motion of a fluid in a potential $\Phi(r)$ per unit mass is (Eckart 1960)

$$L = \int \rho \left(\frac{1}{2} v^2 - u(\rho, s) - \Phi \right) dr, \tag{5.1}$$

where v is the velocity, u the thermal energy per unit mass, ρ the density and s is the entropy. We assume that the entropy of each element of fluid remains constant, though each particle can have a different entropy. SPH Lagrangian equations of motion have been obtained by Springel and Hernquist (2002) using a constraint on the mass within a sphere of radius h_a about particle a and by Monaghan (2002) assuming a functional relation between h and ρ. In this review we use the latter approach.

The SPH form of Eckart's Lagrangian is

$$L = \sum_b m_b \left(\frac{1}{2} v_b^2 - u(\rho_b, s_b) - \Phi_b \right). \tag{5.2}$$

From Lagrange's equations for particle a

$$\frac{d}{dt} \left(\frac{\partial L}{\partial v_a} \right) - \frac{\partial L}{\partial r_a} = 0, \tag{5.3}$$

we find

$$\frac{dv_a}{dt} = -\sum_b m_b \left(\frac{\partial u}{\partial \rho} \right)_s \frac{\partial \rho_b}{\partial r_a} - \frac{\partial \Phi_a}{\partial r_a}. \tag{5.4}$$

From the first law of thermodynamics

$$\left(\frac{\partial u}{\partial \rho} \right) = \frac{P}{\rho^2}. \tag{5.5}$$

From the SPH summation for the density (2.9) (assuming h is a function of ρ as in (4.2)),

$$\Omega_b \frac{\partial \rho_b}{\partial r_a} = \sum_c m_c \nabla_a W_{ac}(h_a) \delta_{ab} - m_a \nabla_b W_{ab}(h_b), \tag{5.6}$$

where the gradient of W_{ab} is taken keeping h constant, δ_{ab} is the Kronecker delta, and

$$\Omega_b = 1 - H_b \sum_c m_c \frac{\partial W_{bc}(h_b)}{\partial h_b}. \tag{5.7}$$

Here H_b denotes $\partial h_b / \partial \rho_b$.

Using these results (5.4) becomes

$$\frac{dv_a}{dt} = -\sum_b m_b \left(\frac{P_a}{\Omega_a \rho_a^2} \nabla_a W_{ab}(h_a) + \frac{P_b}{\Omega_b \rho_b^2} \nabla_a W_{ab}(h_b) \right) + g_a, \tag{5.8}$$

where δ_{ab} is a Kronecker delta, g_a is the force/mass owing to the potential Φ and ∇_a denotes the gradient taken with respect to the coordinates of particle a.

In the case of equi-spaced particles in one dimension, Ω can be estimated using the Poisson summation formula. We find

$$\Omega = 1 + 2h \frac{\partial \tilde{W}}{\partial h}, \tag{5.9}$$

where $\tilde{W}$ is the Fourier transform of W. For the case of a Gaussian kernel

$$\Omega = 1 - \left(\frac{2\pi h}{\Delta} \right)^2 e^{-(\pi h/\Delta)^2}. \tag{5.10}$$
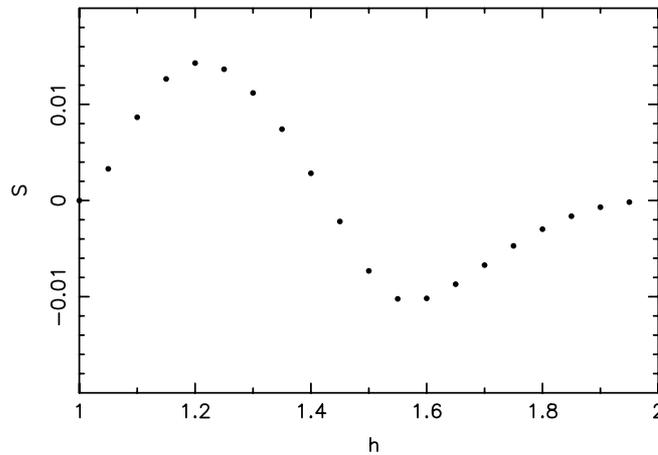
Figure 6. The function S for the cubic spline with equi-spaced particles. The values of h are scaled to the particle spacing.

Since $\pi h/\Delta \sim 4$ this result shows that for the Gaussian kernel and equi-spaced particles Ω is very close to 1. The cubic spline estimate of

$$S = \frac{h}{\rho} \sum_c m_c \frac{\partial W_{bc}(h_b)}{\partial h_b} \tag{5.11}$$

in one dimension is shown in figure 6. The value of S for the cubic spline is larger than for the Gaussian.

However, when ρ varies significantly, Ω can vary significantly and it must be included to give accurate wave propagation. Finally we note (see Monaghan (2002), Price and Monaghan (2004a)) that the rate of change of density with time (2.18), when h is a function of ρ, becomes

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \frac{1}{\Omega_a} \sum_b m_b v_{ab} \cdot \nabla_a W_{ab}(h_a), \tag{5.12}$$

where the gradient is taken with h_a constant. Similarly, the rate of change of thermal energy per unit mass is

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\Omega_a \rho_a^2} \sum_b m_b v_{ab} \cdot \nabla_a W_{ab}(h_a). \tag{5.13}$$

5.1. Conservation laws

The conservation laws can be deduced either from the equations of motion or from the invariance of the Lagrangian to infinitesimal transformations.

5.1.1. Momentum conservation. Linear and angular momentum will be conserved, provided the Lagrangian (5.2) is invariant to translations and rotations. Because the SPH density and therefore the thermal energy term (with constant entropy) is invariant to these transformations, so the Lagrangian, the fluid dynamical terms, therefore, conserve linear and angular momentum. If the force terms owing to the potential are also invariant to the transformations (this is true of self-gravity), the entire system will conserve momentum. The same result follows from (5.8), using

$$\nabla_a W_{ab}(h_a) = r_{ab} F_{ab}(h_a), \tag{5.14}$$

where $F_{ab}(h_a) = F(|r_{ab}|, h_a)$. From the symmetry of the interaction terms the linear and angular momentum are exactly conserved. In addition, because there is no explicit time dependence in L, the energy is conserved. The SPH system, therefore, mimics a system of molecules with forces between their line of centres, but with the difference that the strength of the interaction, through P and ρ, and its geometric dependence through the kernel, depends on the positions of other particles.

5.1.2. Circulation. Kelvin (see, e.g. Lamb (1932)) showed that for an inviscid fluid with $P = P(\rho)$, and conservative body forces, the integral of the velocity around any closed path

$$C_K = \oint v \cdot dr \tag{5.15}$$

is constant. This conservation law is really infinitely many conservation laws since there are infinitely many closed curves. The constancy of circulation has been found useful in many hydrodynamic and atmospheric problems, and it is also applicable in astrophysical problems involving the dynamics of isothermal or adiabatic gas.

We can recover the circulation conservation directly from our SPH system. Consider a fluid where all the particles have the same mass and imagine a necklace of particles. If the particles have the same entropy (so that the necklace lies in a constant entropy surface) then nothing will change if each particle is shifted to its neighbour's positions always moving in the same sense around the necklace. With a proviso to be considered below, the dynamics should be unchanged. We can interpret this as requiring the change in the Lagrangian to be zero.

In this case, if a particle label on the necklace is ℓ, the change in position and velocity of the ℓth particle will be $\delta r_\ell = (r_{\ell+1} - r_\ell)$ and $\delta v_\ell = (v_{\ell+1} - v_\ell)$, respectively. The change in the Lagrangian to first order is then

$$\delta L = \sum_\ell \left(\frac{\partial L}{\partial r_\ell} \cdot \delta r_\ell + \frac{\partial L}{\partial v_\ell} \cdot \delta v_\ell \right), \tag{5.16}$$

where now the summation only applies to the particles around the necklace. Using the previous expressions for δr_ℓ and δv_ℓ together with Lagrange's equations to replace $\partial L/\partial r_\ell$ by $d(\partial L/\partial v_\ell)/dt$, and assuming the particle masses are equal, results in

$$\delta L = m \frac{d}{dt} \sum_\ell v_\ell \cdot (r_{\ell+1} - r_\ell) = 0, \tag{5.17}$$

which must be zero if there is no change in the dynamics. We conclude that

$$C = \sum_\ell v_\ell \cdot (r_{\ell+1} - r_\ell) \tag{5.18}$$

is constant and this is true regardless of the necklace in the constant entropy surface. This result is a discrete version of Kelvin's theorem. We can get the same result, but with opposite sign, by going around the necklace in the opposite sense. If we combine the two (changing the sign of the second because the integral is in the reverse sense) we get

$$C = \frac{1}{2} \sum_\ell v_\ell \cdot (r_{\ell+1} - r_{\ell-1}), \tag{5.19}$$

which is a more accurate estimate of the circulation.

This result is, in general, only approximate because the changes in position and velocity to get from one place in the necklace to its neighbour are discrete, whereas exact conservation is only true when the transformations are infinitesimal. However, as Frank and Reich (2003) show for the case $P = K\rho^2$, where K is a constant, the SPH equations lead to very accurate

conservation of circulation, provided the necklace is defined by a large set of *tracer particles*. These tracer particles have negligible mass and interact only with the real SPH particles. The pressure force can, therefore, be written as the derivative of a potential and it follows (with ℓ denoting a tracer label and noting that the sum over $\boldsymbol{v}_\ell \cdot (\boldsymbol{v}_{\ell+1} - \boldsymbol{v}_{\ell-1})$ vanishes) that

$$\frac{\mathrm{d}C}{\mathrm{d}t} = -\sum_\ell (\boldsymbol{r}_{\ell+1} - \boldsymbol{r}_{\ell-1}) \cdot \nabla_\ell \Psi_\ell, \tag{5.20}$$

where

$$\Psi_\ell = K \sum_b m_b W(\boldsymbol{r}_\ell - \boldsymbol{r}_b, h). \tag{5.21}$$

If the number of tracer particles is made sufficiently large, the summation over ℓ becomes arbitrarily close to a line integral of a potential function around a closed loop and this vanishes. An interesting conclusion from this result is that the tracer particles have enough information from the real SPH particles to define their velocity and position so that the circulation is constant to high accuracy. The same argument can be extended to more complicated barotropic equations of state and applied to molecules or to clusters of stars.

The circulation of a fluid also appears in the work of Feynman on vortices in liquid helium and the necklace transformation was used by him to determine the quantization of circulation. For our present purposes we follow Feynman's review article (Feynman 1957). In that review he suggests a simple form of the wave function for a set of N identical helium atoms. If the entire system moves as a rigid body then the wave function Ψ is given by

$$\Psi = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\sum_j \boldsymbol{r}_j}\Phi, \tag{5.22}$$

where $\boldsymbol{r}_j$ is the position vector of particle j and $N\hbar\boldsymbol{k}$ is the momentum of the system. The function Φ is the ground state wave function. Feynman then argues that if the velocity is varying slowly then the wave function in a region must be close to the wave function of the atoms moving at a uniform velocity. As a result, the wave function for the entire fluid is expected to be similar to

$$\Psi = \mathrm{e}^{\mathrm{i}\sum_j m\boldsymbol{v}_j\cdot\boldsymbol{r}_j}\Phi, \tag{5.23}$$

where m is the mass of a helium atom. Feynman argues that the wave function must be invariant to the necklace transformation. When the particles are shifted around the necklace he finds that the change in the phase is given by

$$\frac{1}{\hbar}\sum_j m\boldsymbol{v} \cdot \Delta\boldsymbol{r}_j. \tag{5.24}$$

The wave function will be invariant if this phase is a multiple of 2π. Accordingly, we can write

$$\sum_j \boldsymbol{v}_j \cdot \Delta\boldsymbol{r}_j = \frac{2\pi\hbar n}{m}, \tag{5.25}$$

where n is an integer. Thus, circulation is quantized.

Finally we note that the circulation invariant contains a topological quantity, the loop around which the circulation is calculated and a dynamical quantity, the velocity. Many numerical codes in astrophysics can guarantee satisfactory accuracy for the velocity, but few can guarantee the same accuracy for the circulation because the numerical codes cannot follow the tangling of the loop.

5.2. The Lagrangian with constraints

In the simplest form of the SPH equations, ρ is defined by a summation over kernels. However, as suggested in section 2, there may be advantages in working with other forms of the density convergence equation; for example,

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \rho_a \sum_b \frac{m_b}{\rho_b} \boldsymbol{v}_{ab} \cdot \nabla_a W_{ab}. \tag{5.26}$$

The action principle requires that

$$S = \int L \, \mathrm{d}t \tag{5.27}$$

is stationary for arbitrary and infinitesimal variations $\delta\boldsymbol{r}$ in the coordinates and corresponding variations $\delta\boldsymbol{v}$ in the velocities. These variations are related by

$$\frac{\mathrm{d}\delta\boldsymbol{r}}{\mathrm{d}t} = \delta\boldsymbol{v}. \tag{5.28}$$

Suppose that the only non-zero variation is $\delta\boldsymbol{r}_a$. The first order change in S is

$$\delta S = \int \left(m_a \boldsymbol{v}_a \cdot \delta\boldsymbol{v}_a - \sum_b m_b \frac{\partial u(\rho_b, s)}{\partial \rho_b} \frac{\delta\rho_b}{\delta\boldsymbol{r}_a} \cdot \delta\boldsymbol{r}_a \right) \mathrm{d}t, \tag{5.29}$$

where $\delta\rho_b/\delta\boldsymbol{r}_a$ denotes the Lagrangian change in ρ_b when the position of particle a changes by $\delta\boldsymbol{r}_a$ at time t. From (5.12) the change in ρ_b (assuming the variation in h can be neglected) is

$$\delta\rho_b = \rho_b \sum_c \frac{m_c}{\rho_c} (\delta\boldsymbol{r}_b - \delta\boldsymbol{r}_c) \cdot \nabla_b W_{bc}(h_b) \tag{5.30}$$

and, therefore,

$$\frac{\delta\rho_b}{\delta\boldsymbol{r}_a} = \rho_b \sum_c \frac{m_c}{\rho_c} (\delta_{ab} - \delta_{ac}) \nabla_b W_{bc}(h_b). \tag{5.31}$$

If this expression is substituted into the integral for δS, and the velocity term is integrated by parts (recalling that $\mathrm{d}\delta\boldsymbol{r}/\mathrm{d}t = \delta\boldsymbol{v}$), the variational principle gives

$$\frac{\mathrm{d}\boldsymbol{v}_a}{\mathrm{d}t} = -\sum_b \frac{m_b}{\rho_a \rho_b} \left(P_a \nabla_a W_{ab}(h_a) + P_b \nabla_a W_{ab}(h_b) \right). \tag{5.32}$$

This is the acceleration equation that is consistent with the convergence equation (5.21). This procedure can be generalized (for details see Price (2004a)) by writing the convergence equation as

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \left(\frac{\rho}{\Phi} \right) \Phi \nabla \cdot \boldsymbol{v}, \tag{5.33}$$

where Φ is an arbitrary function. We can write (5.28) as

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \frac{\rho}{\Phi} (\nabla \cdot (\Phi \boldsymbol{v}) - \boldsymbol{v} \cdot \nabla \Phi). \tag{5.34}$$

If the SPH form of (5.29) is used as a constraint, the action principle gives

$$\frac{\mathrm{d}\boldsymbol{v}_a}{\mathrm{d}t} = -\sum_b \frac{m_b}{\rho_a \rho_b} \left(\frac{P_a \Phi_b}{\Phi_a} \nabla_a W_{ab}(h_a) + \frac{P_b \Phi_a}{\Phi_b} \nabla_a W_{ab}(h_b) \right). \tag{5.35}$$

If $\Phi = \rho$, then the first form of the acceleration equation is recovered. If $\Phi = 1$, the second form is recovered. If we choose $\Phi = \sqrt{P}$, then the acceleration equation becomes

$$\frac{\mathrm{d}\boldsymbol{v}_a}{\mathrm{d}t} = -\sum_b m_b \frac{\sqrt{P_a P_b}}{\rho_a \rho_b} \left(\nabla_a W_{ab}(h_a) + \nabla_a W_{ab}(h_b) \right). \tag{5.36}$$

The advantages of choosing this last form consistently with $\Phi = \sqrt{P}$ in the convergence equation have not been analysed. These various forms of the acceleration equation have the same conservation properties.

5.3. Time integration in the absence of dissipation

Because the SPH algorithm reduces the original continuum partial differential equations to sets of ordinary differential equations, any stable time stepping algorithm for ordinary differential equations can be used. However, when there is no dissipation, the properties of the Lagrangian description can be preserved using a symplectic integrator (see, e.g. Leimkuhler *et al* (1997)). A simple example is the Verlet second order integrator which, for the one-dimensional system

$$\frac{dq}{dt} = v, \tag{5.37}$$

$$\frac{dv}{dt} = f(q), \tag{5.38}$$

takes the form (for constant time step δt)

$$q^{1/2} = q^0 + \tfrac{1}{2}\delta t v^0, \tag{5.39}$$

$$v^1 = v^0 + \delta t f(q^{1/2}), \tag{5.40}$$

$$q^1 = q^{1/2} + \tfrac{1}{2}\delta t v^1, \tag{5.41}$$

where a^0, $a^{1/2}$ and a^1 denote the values of a at the start of a step, halfway and at the end of the step, respectively.

In the case where there are n coordinates $q_1, q_2, q_3, \ldots, q_n$ with velocities $v_1, v_2, v_3, \ldots, v_n$ we get

$$q_i^{1/2} = q_i^0 + \tfrac{1}{2}\delta t v_i^0, \tag{5.42}$$

$$v_i^1 = v^0 + \delta t f(q_1^{1/2}, q_1^{1/2}, q_1^{1/2}, \ldots), \tag{5.43}$$

$$q_i^1 = q_i^{1/2} + \tfrac{1}{2}\delta t v_i^1. \tag{5.44}$$

In an SPH calculation, δt will depend on the speed of sound which, for a non-dissipative system depends on the density and, therefore, on the coordinates. In this case (5.39)–(5.41) are replaced with the following steps where the first half step has a different time step from the second half step.

$$q_i^{1/2} = q_i^0 + \tfrac{1}{2}\delta t^0 v_i^0, \tag{5.45}$$

$$v_i^{1/2} = v_i^0 + \tfrac{1}{2}\delta t^0 f(q_1^{1/2}, q_2^{1/2}, q_3^{1/2}, \ldots), \tag{5.46}$$

$$v_i^1 = v_i^{1/2} + \delta t^1 f(q^{1/2}), \tag{5.47}$$

$$q_i^1 = q_i^{1/2} + \tfrac{1}{2}\delta t^1 v_i^1, \tag{5.48}$$

where, for example,

$$\delta t^{1/2} = \tfrac{1}{2}(\delta t^0 + \delta t^1), \tag{5.49}$$

or, the frequently used

$$\frac{2}{\delta t^{1/2}} = \frac{1}{\delta t^0} + \frac{1}{\delta t^1}, \tag{5.50}$$

so that, with $\delta t^{1/2}$ calculated from the mid-point coordinate values, δt^1 can be calculated for the second half of the time stepping. This algorithm is reversible in time. The time stepping

for the n coordinate system (5.42)–(5.44) can be replaced in the same way. An alternative form with the same accuracy is

$$v^{1/2} = v^0 + \tfrac{1}{2}\delta t f^0, \tag{5.51}$$

$$q^1 = q^0 + \delta t v^{1/2}, \tag{5.52}$$

$$v^1 = v^{1/2} + \tfrac{1}{2}\delta t f^1, \tag{5.53}$$

which can be compared with (5.39)–(5.41). The latter is often referred to as the drift–kick–drift form, whereas the steps (5.51)–(5.53) are referred to as the kick–drift–kick form. The kick is the change in the velocity by the force. The drift is the change in the coordinate moving with the initial velocity. In some cases it may be useful to have the forces evaluated at the end of the step as in the kick–drift–kick form.

It is possible to show that the symplectic integrator equations (5.42)–(5.44) are equivalent to using the Lagrangian

$$L = \sum_i \frac{1}{2} m_i v_i^2 - \Phi - \frac{\delta t^2}{12} \left(\sum_j m_j f_j^2 + \frac{1}{2} \sum_j \sum_k m_j \dot{q}_j \dot{q}_k \frac{\partial f_j}{\partial q_k} \right) + O(\delta t^4), \tag{5.54}$$

or, equivalently, the Hamiltonian

$$H = \sum_i \frac{1}{2} m_i v_i^2 + \Phi + \frac{\delta t^2}{12} \left(\sum_j m_j f_j^2 + \frac{1}{2} \sum_j \sum_k m_j \dot{q}_j \dot{q}_k \frac{\partial f_j}{\partial q_k} \right) + O(\delta t^4), \tag{5.55}$$

where Φ is the potential energy such that $f_i = \partial \Phi / \partial q_i$. In an SPH calculation, Φ is given by

$$\Phi = \sum_j m_j u_j + \Psi, \tag{5.56}$$

where Ψ is the potential of any body force. As a consequence, the Hamiltonian, and therefore the energy, will not show a secular increase or decrease with time. Note that the double summation term can be written

$$\sum_j \sum_k \dot{q}_j \dot{q}_k \frac{\partial f_j}{\partial q_k} = \sum_j v_j \frac{\mathrm{d} f_j}{\mathrm{d} t}. \tag{5.57}$$

Since $\mathrm{d} f_j / \mathrm{d} t$ can be estimated from f_j at two time steps, the contribution of this double summation can be computed at little cost.

The advantages of using symplectic integrators for molecular dynamics has been discussed by many authors (see, e.g. Leimkuhler *et al* (1997)).

6. Applications of the Euler equations

The most common application of the SPH equations without dissipation is to small oscillations. The simplest of these is the oscillation of an infinite, one-dimensional gas with constant initial density. The analysis in the case of constant h has been given by Monaghan (1989) and Morris (1996). However, we will give the dispersion relation appropriate for h a function of ρ. A more complicated example is the oscillation of a Toy star in one dimension. This case is important because it mimics the oscillations of a star and is more difficult because the eigen functions vary sharply near the surface where the density goes to zero.

6.1. Dispersion relation for an infinite one-dimensional gas

Consider an SPH system that consists of an infinite set of particles in one dimension with initial spacing Δ. They are perturbed by a velocity much less than the speed of sound $c_{\rm s}$. Let the unperturbed quantities (shown by an over bar) and the space and time variation of all perturbed quantities be proportional to

$$\exp {\rm i}(k\bar{x}_a - \omega t), \tag{6.1}$$

where $\bar{x}_a = a\Delta$ is the unperturbed position of particle a. In an unpublished work, I have shown that the linearized one-dimensional SPH equations of motion, with $h \propto 1/\rho$ and $P = K\rho^\gamma$, give the dispersion relation

$$\omega^2 = \frac{c_{\rm s}^2}{\gamma\bar{\Omega}} \left[\frac{(\gamma - 2)\Phi^2}{\omega} + 2\Psi - \frac{h}{\bar{\Omega}} \frac{\partial \Phi^2}{\partial h} - \frac{\ell\Phi^2}{\bar{\Omega}^2} \right]. \tag{6.2}$$

The functions Φ, Ψ and ℓ are defined by

$$\Phi = \Delta \sum_c \sin (k\bar{x}_c) \frac{\partial W(\bar{x}_c, h)}{\partial \bar{x}_c}, \tag{6.3}$$

$$\Psi = \Delta \sum_c [1 - \cos (k\bar{x}_c)] \frac{\partial^2 W(\bar{x}_c, h)}{\partial h^2} \tag{6.4}$$

and

$$\ell = -2h\Delta \sum_c \frac{\partial W(\bar{x}_c, h)}{\partial h} - h^2\Delta \sum_c \frac{\partial^2 W(\bar{x}_c, h)}{\partial h^2} \tag{6.5}$$

and $c_{\rm s}$ is the adiabatic sound speed.

If the wavelength is much larger than Δ (as in many simulations, where the wave length is typically 100Δ), the summation can be replaced by an integration. We find

$$\Phi = \int_\infty^\infty \sin (kx) \frac{{\rm d}W}{{\rm d}x} {\rm d}x = -k\tilde{W}(k, h), \tag{6.6}$$

where $\tilde{W}(k, h)$ is the Fourier Transform of the kernel, and

$$\Psi = \int_\infty^\infty (1 - \cos (kx)) \frac{{\rm d}^2 W}{{\rm d}x^2} {\rm d}x = k^2\tilde{W}(k, h). \tag{6.7}$$

In this limit the dispersion relation becomes

$$\omega^2 = \frac{c_{\rm s}^2 k^2}{\gamma\Omega} \left[\frac{(\gamma - 2)\tilde{W}^2(k, h)}{\Omega} + 2\tilde{W}(k, h) - \frac{h}{\Omega} \frac{\partial \tilde{W}^2}{\partial h} - \frac{\ell\tilde{W}^2}{\Omega^2} \right]. \tag{6.8}$$

The Fourier transform of the one-dimensional Gaussian kernel and the spline kernels are

$${\rm e}^{-(hk/2)^2} \tag{6.9}$$

and

$$\left(\frac{\sin (hk/2)}{hk/2} \right)^n, \tag{6.10}$$

respectively, where the latter is obtained from (2.4) (note that the cubic spline has $n = 4$). If $kh < 1$, we can approximate each of these by
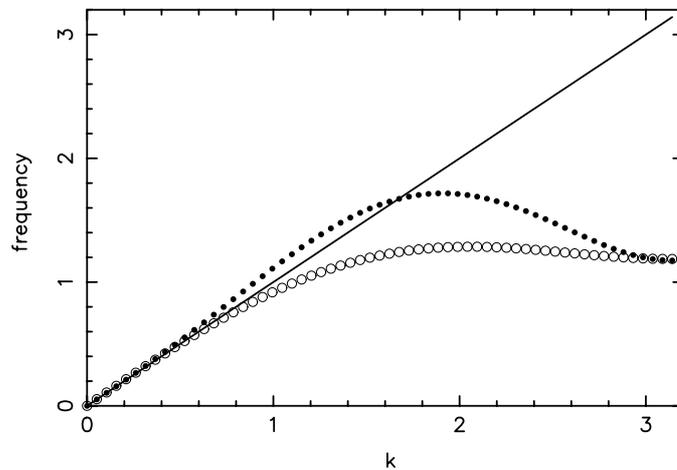
$$1 - \beta h^2 k^2, \tag{6.11}$$

Figure 7. The SPH dispersion relation for sound waves in one dimension using the cubic spline and taking $\gamma = 5/3$ with full account of the variation of h with ρ. The black dots show the results when the variation of h with density is included. The open circles show the results when h is fixed.

with $\beta = 1/4$ for the Gaussian and $\beta = 1/6$ for the cubic spline. Using this approximation, and the further approximation $\Omega = 1$ and $\ell \sim 0$, the dispersion relation when h varies with ρ becomes

$$\omega^2 = k^2 c_s^2 \left[1 + \frac{2\beta h^2 k^2 (3 - \gamma)}{\gamma} \right]. \tag{6.12}$$

If the contributions from the variation of h with ρ are neglected, the dispersion relation becomes

$$\omega^2 = k^2 c_s^2 \left[1 + \frac{2\beta h^2 k^2 (1 - \gamma)}{\gamma} \right]. \tag{6.13}$$

This shows that, if h is constant, and γ lies in the normal range $1 \leqslant \gamma \leqslant 3$, the dispersion curve lies below the exact line $\omega = c_s k$, whereas if the variation of h is included the dispersion curve is above the exact line. The dispersion relation for the cubic spline in the case where, initially, $h = 1.2$ and $\gamma = 5/3$ is shown in figure 7. The variation of the dispersion relation is in agreement with the previous results when k is sufficiently small. In addition, we note that the dispersion relation for the case of varying h always lies above that for the case where h is constant. Both dispersion curves have the same limit when $k = \pi/\Delta$ because, for this k, $\Phi = 0$ and $\Psi = 8\Delta(\partial^2 W/\partial x^2)$ evaluated at $x = \Delta$. More accurate dispersion relations can be obtained if kernels which interpolate at a higher order are used. However, as mentioned earlier, these kernels are not satisfactory for shocks unless, in a dynamical calculation, there is a switch from lower order interpolation near shocks (e.g. where the cubic spline could be used) to a higher order interpolating kernel, elsewhere. An alternative is to use velocity smoothing (see later) with a suitable coefficient to cancel the error terms.

Figure 8 shows a velocity field after 4000 steps, for a one-dimensional gas with $\gamma = 1.4$, which was begun with density constant and velocity $0.05c_s \sin(2\pi x)$. The velocity was reversed after 2000 steps. The total time of the simulation is equivalent to 13.7 periods. The SPH results were calculated using 100 particles with constant $h = 0.013$. The exact result (the reversed initial velocity) is shown by the continuous line, which is difficult to see because it passes through the points from the SPH simulation shown by filled symbols. Figure 9 shows the results for an initial velocity $0.05c_s \sin(10\pi x)$. In this case the integration time is equivalent to 68.5 periods. The agreement between the SPH results and the exact values is excellent.
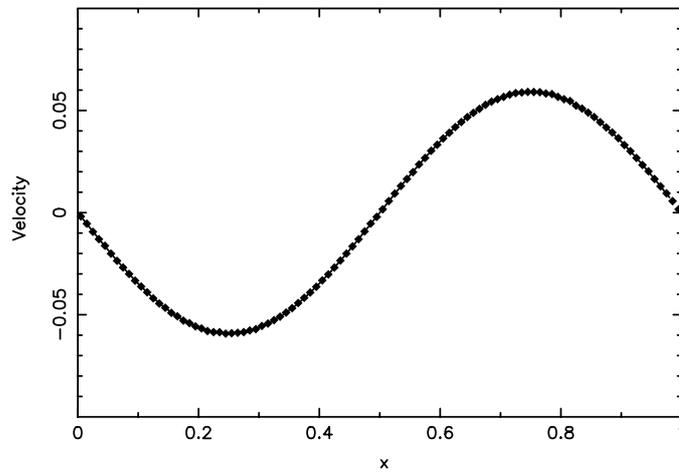
Figure 8. The velocity field for an oscillation in a one-dimensional gas at step 4000. The integration was performed using a Verlet symplectic integrator with 100 SPH points and the motion was reversed at step 2000. The SPH results are shown by filled symbols and the exact results by a continuous line (which is difficult to see because it passes through the SPH points).
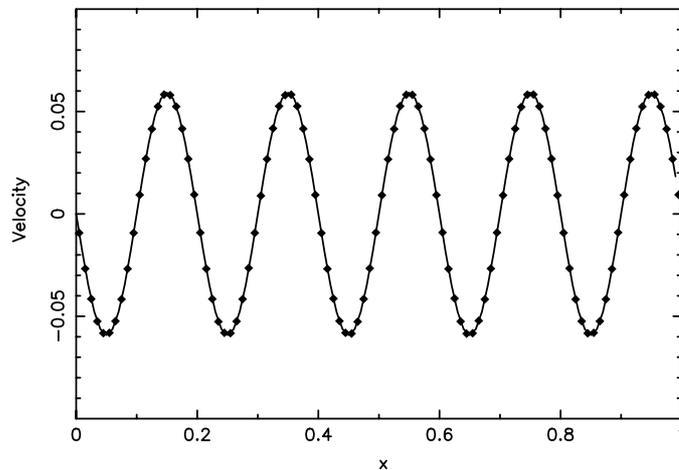
Figure 9. The velocity field for the conditions of the previous figure except that the initial velocity is $0.05c_s \sin(10\pi x)$.

6.2. Toy star oscillations

The usual tests in computational gas dynamics involve systems with rigid or periodic boundaries as in the previous test. These boundaries are quite useful for testing algorithms for industrial fluid dynamics. However, in astrophysics a more realistic test case is a finite mass of gas pulled together by gravity or a force which mimics gravity. The region outside the gas then has zero density. When finite difference methods are used for the dynamics of such a system they often give poor results because they do not handle the outer region of the gas moving into a vacuum. However, they do not present difficulties for particle methods such as SPH.

A useful class of such test problems are the Toy stars considered by Monaghan and Price (2004). The self-gravity is replaced by an attractive force proportional to the distance

and along the line of centres of any two particles. This force is the simplest many-body force. It was discovered by Newton, who pointed out that if two particles attract each other with a linear force then they move as if attracted to the centre of mass of the pair (see Chandrasekhar (1995) for a modern interpretation of Newton's Principia and, in particular, Newton's proposition LXIV, which discusses this force).

If there are N particles attracting each other with a force proportional to the separation, and directed along the line joining pairs of particles, then each particle moves as if it is independent of the others. The force appears as a linear force towards the centre of mass of the N particles (the particles, therefore, move in a common oscillator potential). In the case of two particles in three dimensions, the trajectories are closed Liassajous figures. A gaseous system with this force has a number of attractive features for testing algorithms for fluid dynamics. The linear modes of oscillation can be calculated easily and there is an exact non-linear solution where the velocity is a linear function of the coordinates but a non-linear function of time. This solution can be calculated very accurately by integrating a small number of ordinary differential equations and the results provide an excellent test of any computational fluid dynamics algorithm.

The simplest version of the Toy star assumes that pressure P is given in terms of the density ρ by $P = K\rho^2$, where K is a constant. This makes the problem analogous to the problem of shallow water motion in paraboloidal basins. There is extensive literature on this problem including the seminal papers of Ball (1963) and the general analysis by Holm (1991).

6.3. Toy stars in one dimension

Suppose that we have an isolated group of N particles in one dimension interacting with linear forces such that the force on particle j owing to particle k is $\nu m_j m_k (x_k - x_j)$. The potential energy is

$$\Phi = \frac{1}{4} \nu \sum_j \sum_k m_j m_k (x_j - x_k)^2, \tag{6.14}$$

The equation of motion of the jth particle is then

$$m_j \frac{\mathrm{d}^2 x_j}{\mathrm{d}t^2} = -\nu m_j \sum_k m_k (x_j - x_k). \tag{6.15}$$

However, the centre of mass can be chosen as the origin, so the equation of motion becomes

$$\frac{\mathrm{d}^2 x_j}{\mathrm{d}t^2} = -\nu M x_j, \tag{6.16}$$

where M is the total mass. The motion of the N-body system is therefore identical to the independent motion of each particle in a harmonic potential. In the following, we replace $M\nu$ by Ω^2. The acceleration equation for a one-dimensional gaseous toy star with velocity v, density ρ and pressure P is

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{1}{\rho} \frac{\partial P}{\partial x} - \Omega^2 x. \tag{6.17}$$

Solutions can be found for $P = K\rho^\gamma$ with K constant and any $\gamma \geqslant 1$. If $\gamma = 2$, the equations are identical in form to those for the shallow water equations with density replacing the water depth. The equilibrium quantities can be easily calculated and, when the equilibrium is disturbed by velocities that are small compared with the speed of sound, the equations can be linearized. The errors in the linearization may, however, be large near the surface where the speed of sound and the pressure fall to zero. The velocity eigenfunctions are Gegenbauer
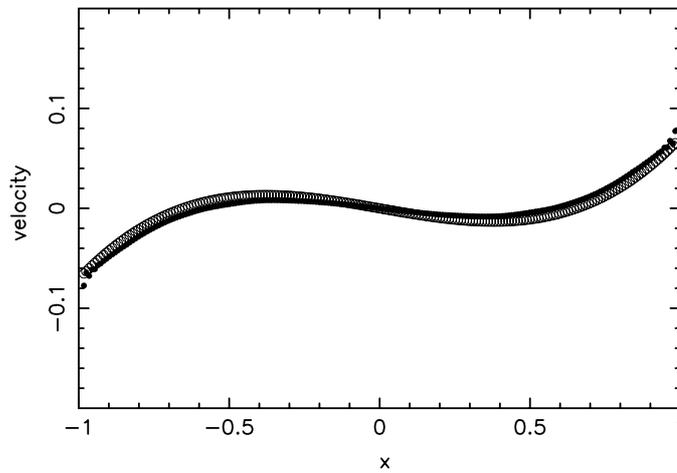
Figure 10. The velocity field for the Toy star oscillating with the velocity field in mode 3. The SPH results are shown by the filled symbols and the exact result by the circles.
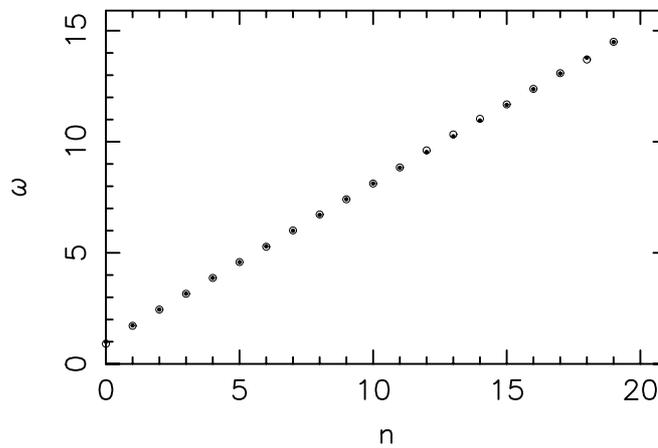
Figure 11. The SPH frequencies for Toy star eigenfunctions shown by filled circles compared with the exact results shown by open circles.

polynomials and the density eigenfunctions are Legendre polynomials. To simulate high order oscillations a large number of particles must be used to ensure that the resolution length is much smaller than the separation of the modes. If 400 particles are used then modes up to the twentieth can be simulated with high accuracy. The frequencies are always very accurate but the errors in the eigenfunctions are less accurate especially near the boundary. The velocity field for mode 3 is shown in figure 10 after 4 oscillation periods. The agreement with the perturbation solution is very good. A comparison between the SPH and the exact frequencies are shown in figure 11 for the first 20 modes. The agreement between theory and computation is excellent. These results show that the SPH method is able to accurately reproduce rather delicate and small oscillations in one dimension.

An attractive feature of the Toy stars is that exact non-linear solutions can be found. For the one-dimensional case with $P = K\rho^\gamma$ the solution has the form

$$v = A(t)x, \tag{6.18}$$

with

$$\rho^{(\gamma-1)} = H(t) - C(t)x^2, \tag{6.19}$$

so that the time-dependent radius of the toy star is $\sqrt{H/C}$. Substitution into the equations of motion and equating powers of x gives a set of ordinary differential equations for A, H and C. These can be integrated with high accuracy and compared with direct simulations by SPH. The agreement between the SPH results and the exact solution is again excellent.

The generalization of Toy stars to 2 and 3 dimensions is straightforward although the details become more complicated. Solutions can also be found with magnetic fields (see Monaghan and Price (2004)) who solve the one-dimensional MHD case.

7. Heat conduction and matter diffusion

The efficient solution of the heat conduction equation is fundamental for dissipative processes since similar techniques can be used for viscous dissipation or matter diffusion. An advantage of the SPH equations for these dissipative problems, as in the purely mechanical case, is that they can be written in such a way that they mimic fundamental properties of the system and allow complicated physics to be handled in a straightforward way. Appropriate forms of these equations have been derived (Brookshaw 1985, Cleary and Monaghan 1999, Monaghan *et al* 2005) and applied to a wide variety of heat conduction problems, including the Stefan problem and the freezing of alloy solutions (Monaghan *et al* 2005) and problems involving radiative transfer in the diffusion approximation (Whitehouse and Bate 2004).

7.1. The SPH heat conduction equation

A convenient form of the heat conduction equation without heat sources or sinks is

$$c_p \frac{\mathrm{d}T}{\mathrm{d}t} = \frac{1}{\rho}\nabla(\kappa\nabla T), \tag{7.1}$$

where T is the absolute temperature, c_p the heat capacity per unit mass at constant pressure, ρ the density, κ the coefficient of thermal conductivity and $\mathrm{d}/\mathrm{d}t$ the derivative following the motion. The spatial derivatives can be determined using the results of section 2.3, and the SPH form of (7.1) is

$$c_{p,a}\frac{\mathrm{d}T_a}{\mathrm{d}t} = \sum_b \frac{m_b}{\rho_a\rho_b}(\kappa_a + \kappa_b)(T_a - T_b)F_{ab}. \tag{7.2}$$

This equation shows that the contribution of particle b to the rate of change of T_a is positive if $T_b > T_a$ because $F_{ab} \leqslant 0$, i.e. the heat flows from the hotter element of the fluid to the cooler element as expected. As mentioned in section 2.3, this fundamental requirement could not be guaranteed if the second derivatives of the interpolation formula for T were calculated directly.

Equation (7.2) does not guarantee that the heat flux will be continuous when κ is discontinuous. Cleary and Monaghan (1999) show from an analysis of the finite difference case that this problem can be solved by replacing $(\kappa_a + \kappa_b)$ in (7.2) with

$$\frac{4\kappa_a\kappa_b}{(\kappa_a + \kappa_b)}. \tag{7.3}$$

The heat flux is then continuous even with jumps by a factor 10^3 in κ across 3 particle spacings. A slightly different κ term, based on similar ideas, gives satisfactory results for jumps in κ by a factor 10^9 (Parshikov and Medin 2002). However, because the very simple form (7.3) gives

excellent results for the normal range of material properties, the final SPH heat conduction equation is, therefore,

$$c_{p,a} \frac{\mathrm{d}T_a}{\mathrm{d}t} = \sum_b \frac{m_b}{\rho_a \rho_b} \frac{4\kappa_a \kappa_b}{(\kappa_a + \kappa_b)} (T_a - T_b) F_{ab}. \tag{7.4}$$

Cleary and Monaghan (1999) showed that this SPH form of the heat conduction equation had similar accuracy to finite difference methods and was not sensitive to the particle disorder that occurs in some SPH calculations. In addition, heat conduction problems with discontinuous κ, and with κ varying with T, were accurately integrated. Whitehouse and Bate (2004) studied heat conduction by radiation in the diffusion approximation obtaining accurate results for test problems.

If the particles are thermally isolated (so they can only exchange heat amongst themselves) then (7.2) shows (noting $F_{ab} = F_{ba}$), that the total heat content

$$\sum_a m_a c_{p,a} T_a \tag{7.5}$$

is constant.

7.2. Heat conduction with sources or sinks

When the system contains point sources or sinks, (7.1) becomes

$$\rho c_p \frac{\mathrm{d}T}{\mathrm{d}t} = \nabla(\kappa \nabla T) + \sum_k Q_k \delta(\boldsymbol{r} - \boldsymbol{R}_k), \tag{7.6}$$

where Q_k denotes the strength of the source or sink and is negative for a sink. $\boldsymbol{R}_k$ denotes the position of source/sink k and δ denotes a Dirac delta function. The SPH equation corresponding to (7.6) becomes

$$c_{p,a} \frac{\mathrm{d}T_a}{\mathrm{d}t} = \sum_b \frac{m_b}{\rho_a \rho_b} \frac{4\kappa_a \kappa_b}{(\kappa_a + \kappa_b)} (T_a - T_b) F_{ab} + \frac{1}{\rho_a} \sum_k Q_k \zeta_k W(\boldsymbol{r}_a - \boldsymbol{R}_k), \tag{7.7}$$

where the delta function has been replaced by a smoothing kernel, which is consistent with the smoothing of the original continuum equation and, to ensure that the rate of change of thermal energy owing to the source is correct, a normalizing factor ζ_k for source k defined by

$$\frac{1}{\zeta_k} = \sum_b \frac{m_b}{\rho_b} W(\boldsymbol{r}_b - \boldsymbol{R}_k, h), \tag{7.8}$$

has been introduced. The right-hand side is an SPH estimate of the constant 1 at the position of the source. From (7.7) the rate of change of thermal energy is

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\sum_a m_a c_{p,a} T_a \right) = \sum_k Q_k, \tag{7.9}$$

as expected.

7.3. Salt diffusion

Denoting the mass fraction of salt by C so that the mass of salt in a mass M of liquid is CM, the diffusion of the salt is given by an equation similar in form to the heat conduction equation, namely,

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{1}{\rho} \nabla(D\nabla C), \tag{7.10}$$

where D is the coefficient of diffusion with dimensions of $ML^{-1}T^{-1}$. The SPH form of this equation follows in the same way as for the heat conduction equation. The SPH equation for the rate of change of the concentration C_a of particle a is given by

$$\frac{dC_a}{dt} = \sum_b \frac{m_b}{\rho_a \rho_b} \frac{4 D_a D_b}{(D_a + D_b)} (C_a - C_b) F_{ab}. \tag{7.11}$$

The combination of D in the SPH equation ensures that the flux of material across an interface between two materials with different diffusion coefficients is constant. The total mass of salt is conserved by the SPH equation.

7.4. The increase of entropy

The SPH conduction equation results in entropy increasing in the absence of heat sinks. If S is the total entropy of the system then

$$\frac{dS}{dt} = \sum_a m_a \frac{ds_a}{dt} = \sum_a \frac{m_a}{T_a} \frac{dq_a}{dt}, \tag{7.12}$$

where s_a is the entropy/mass of particle a, q_a is the heat content/mass of particle a and T is the absolute temperature. From equations (7.4) and (7.12), with an interchange of labels the change of entropy with time can be written as

$$\frac{dS}{dt} = \frac{1}{2} \sum_a \sum_b \frac{m_a m_b}{\rho_a \rho_b} \frac{4 \kappa_a \kappa_b}{(\kappa_a + \kappa_b)} \left(\frac{1}{T_a} - \frac{1}{T_b} \right) (T_a - T_b) F_{ab}. \tag{7.13}$$

Since $F_{ab} \leqslant 0$ we deduce that $dS/dt \geqslant 0$.

When the composition changes there is a further contribution to the entropy. To deduce this we first divide (7.11) by C_a. If the resulting equation is summed over a, and added to the same expression with the labels interchanged, the following positive definite quantity is obtained.

$$\frac{d}{dt} \sum_a m_a \ln C_a = \sum_a \sum_b m_a m_b \frac{4 D_a D_b}{(D_a + D_b)} \left(\frac{1}{C_a} - \frac{1}{C_b} \right) \frac{(C_a - C_b)}{\rho_a \rho_b} F_{ab} \geqslant 0. \tag{7.14}$$

This quantity is the increase in entropy resulting from composition changes.

7.5. Boundary and interface conditions

There is no need to place a special condition on the gradient of the temperature at the boundary to satisfy these conditions if SPH is used. If all the boundaries are adiabatic, then the particles interact amongst themselves and the symmetry of the SPH conduction equation ensures that the system conserves its thermal energy as shown earlier. If one or more boundary curves have fixed temperatures, the SPH particles on the boundaries are included in the heat conduction equation so that the heat transferred to the boundary during a time step can be calculated. After this is done the temperatures of the boundary particles are set back to the specified boundary temperatures for the next time step. The heat transferred to the boundary particle can be calculated from the temperature change. Cleary and Monaghan (1999) noted that near the boundaries, the SPH interpolation can give errors of a few per cent, and they made corrections to the density near the boundary to compensate for this. As noted earlier, SPH calculations do not need special interface conditions. The SPH particles exchange heat and material with neighbouring particles whether they are of the same or different phases.
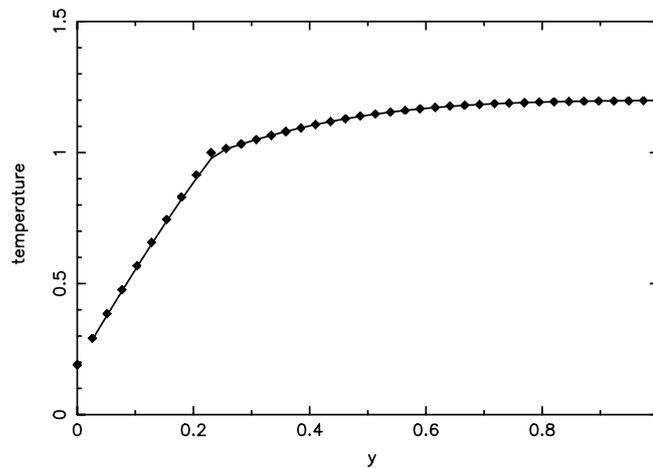
Figure 12. The temperature against the distance from the cooling boundary for a two-dimensional Stefan problem. The system is periodic in the x direction. The exact results are shown by the solid line and the SPH results by the solid diamonds. The change of slope shows the interface between solid and liquid.

7.6. The Stefan problem

An interesting application of SPH is to the Stefan problem where a pure substance is cooled sufficiently for it to freeze. In the standard treatment of this problem the following condition is required at the interface:

$$\kappa_1 \left(\frac{\mathrm{d}T}{\mathrm{d}y} \right)_1 - \kappa_2 \left(\frac{\mathrm{d}T}{\mathrm{d}y} \right)_2 = \rho L \frac{\mathrm{d}Y}{\mathrm{d}t}, \tag{7.15}$$

where L is the latent heat/mass and $\mathrm{d}Y/\mathrm{d}t$ is the rate of change of the position of the interface (Carslaw and Jaeger 1990). The condition expresses the fact that the difference in the heat flux on each side of the interface supplies the heat to change the phase. In this formulation the position of the interface is one of the unknowns.

The SPH treatment of the freezing is very simple. Initially, the SPH particles are assumed to be liquid and tagged with an integer to denote liquid particles and given the material properties of the liquid. As heat is conducted from the liquid some liquid SPH particles reach the solidification temperature T_m. The heat per unit mass q, lost by these particles after this time, is then stored and their temperatures are kept at T_m. If particle a is in this condition then, when q_a reaches L, the integer tag is changed to that for the solid phase, and the properties of this phase (thermal conductivity and heat capacity) are assigned to this particle. Between the solid particles and the liquid particles there is a region where the particles have reached the solidification temperature but have, not yet had their latent heat fully extracted. An example of the SPH solution of a one-dimensional Stefan problem is shown in figure 12, and for an axisymmetric Stefan problem with a heat sink in figure 13 (both taken from Monaghan *et al* (2005)). The agreement between the SPH results and the exact result (Carslaw and Jaeger 1990) is excellent.

8. Viscosity

The first use of viscosity in SPH equations was by Lucy (1977) who introduced an artificial bulk viscosity to prevent a slow build-up of acoustic energy from integration errors in an SPH simulation. A different, and more effective viscosity, which conserves linear and angular
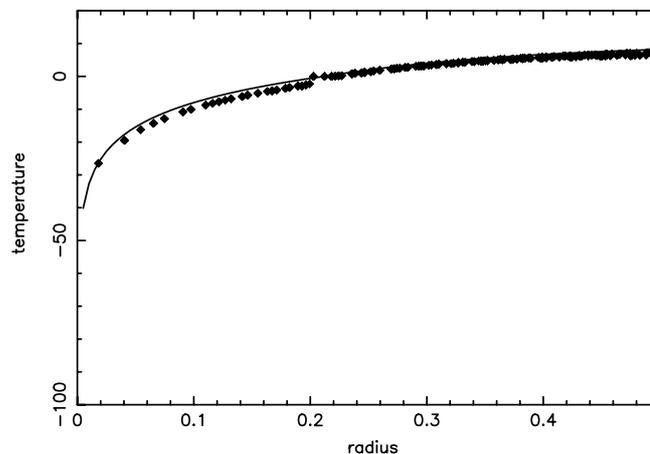
Figure 13. The temperature against the radius for the case of freezing is induced by a line sink in an axisymmetric system. The exact results are shown by the continuous line and the SPH results by symbols. Despite the particles being on a rectangular grid, the variation of the temperature is close to radial. Note that the small group of particles which have reached the freezing temperature but have not yet become ice particles, form a small horizontal line at a radius of approximately 0.22.

momentum was suggested and tested by Monaghan and Gingold (1983). The results obtained using this viscosity in a wide variety of shock problems involving gases, liquids and solids (Libersky and Petschek 1991) and, with a different version for a relativistic gas (Chow and Monaghan 1997) in one dimension, is in good agreement with the theory. With reference to shock problems, SPH does not give the widths of shock fronts as accurately as the methods based on Riemann solvers with similar resolution; however, no current method gives the width of a shock front accurately, since the width of real shock fronts is only a few molecular mean-free paths. Typical resolutions in numerical simulations are a factor 10^4 greater. The key is to get the pre- and post-shock values correct and SPH is capable of producing these to any degree of desired accuracy.

In two and more dimensions it is more difficult for SPH to match the accuracy of modern finite difference codes, but its advantage is that it is independant of the special properties of the ideal gas equation, which are built into the finite difference codes. Consequently, SPH can be used when the equation of state is complicated and Riemann solutions are unavailable (approximate linear solutions could be used but they are unreliable (Quirk (1994))).

The viscosity of real fluids can be implemented using ideas similar to those used for the artificial viscosity and for heat conduction (Cleary 1998, Cleary and Ha 2002). Applications have been made to a low Reynolds number flow (Morris *et al* 1997) and to systems involving more than one fluid in contact. An alternative approach is to calculate the velocity derivatives in the viscous term using SPH methods (Takeda *et al* 1994, Watkins *et al* 1996, Chaniotis *et al* 2002). These forms of the viscous stress tensor conserve linear momentum but not angular momentum. In many industrial fluid dynamics problems, the exact conservation of angular momentum is not an issue and the work of Chaniotis *et al* (2002) shows that SPH (together with a re-meshing strategy) gives excellent results.

8.1. Artificial viscosity

As its name suggests, artificial viscosity bears no relation to real viscosities, but is designed to allow shock phenomena to be simulated, or simply to stabilize a numerical algorithm. Artificial

viscosities are often constructed analogously to real gas viscosities, replacing the mean free path with the resolution length. The Navier–Stokes acceleration equation for viscous flow has the form

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + \frac{1}{\rho}\left[\frac{\partial}{\partial x_k}\left(\eta\left(\frac{\partial v_i}{\partial x_k} + \frac{\partial v_k}{\partial x_i} - \frac{2}{3}\delta_{ik}\nabla\cdot\boldsymbol{v}\right)\right) + \frac{\partial}{\partial x_i}(\zeta\nabla\cdot\boldsymbol{v})\right],\tag{8.1}$$

where η is the shear viscosity coefficient and ζ is the bulk viscosity, which is required when the internal degrees of freedom of the molecules of the fluid are activated during the flow. These viscosity coefficients are, in general, functions of temperature and density. For a monatomic gas $\eta \sim \frac{1}{3}\rho\lambda c_{\mathrm{s}}$, where λ is the mean free path and c_{s} is the speed of sound.

The viscous terms could be estimated directly using the SPH interpolation formula but, as in the case of heat conduction, this leads to equations, which do not conserve linear and angular momentum, and do not guarantee that the viscous dissipation will increase the entropy. Monaghan and Gingold (1983) devised a viscosity by simple arguments about its form and its relation to gas viscosity. The viscous term, denoted by Π_{ab} is added to the pressure terms in SPH equations to give

$$\frac{\mathrm{d}v_a}{\mathrm{d}t} = -\sum_b m_b\left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2} + \Pi_{ab}\right)\nabla_a W_{ab},\tag{8.2}$$

where

$$\Pi_{ab} = -\nu\left(\frac{\boldsymbol{v}_{ab}\cdot\boldsymbol{r}_{ab}}{r_{ab}^2 + \epsilon\bar{h}_{ab}^2}\right)\tag{8.3}$$

and $\epsilon \sim 0.01$ is introduced to prevent a singularity when $r_{ab} = 0$ and ν is defined by

$$\nu = \frac{\alpha\bar{h}_{ab}\bar{c}_{ab}}{\bar{\rho}_{ab}},\tag{8.4}$$

where, for example, $\bar{h}_{ab} = (h_a + h_b)/2$. A further generalization, which has not been explored, but which gives a higher order viscosity, is to multiply the previous viscosity by any power of

$$\left|\frac{\boldsymbol{v}_{ab}\cdot\boldsymbol{r}_{ab}}{\bar{c}_{ab}}\right|.\tag{8.5}$$

The artificial viscosity term Π_{ab} is a Galilean invariant and vanishes for rigid rotation. When two particles approach each other, the artificial viscosity produces a repulsive force between the particles. When they recede from each other the force is attractive.

The SPH viscosity can be related to a continuum viscosity by converting the summation to integrals. The x component of the acceleration equation has the viscous contribution

$$f_x = \sum_b m_b\frac{\alpha\bar{c}_{ab}\bar{h}_{ab}}{\bar{\rho}_{ab}}\frac{\boldsymbol{v}_{ab}\cdot\boldsymbol{r}_{ab}}{r_{ab}^2 + \epsilon\bar{h}_{ab}^2}(x_a - x_b)F_{ab}.\tag{8.6}$$

If the $\epsilon\bar{h}_{ab}^2$ in the denominator is dropped this integral can be written as a sum of terms similar in form to those considered in section 2.3. If α, c, h and ρ are constant the continuum equivalent of f_x in two dimensions is

$$f_x = \alpha h c\left(\frac{3}{8}v_{xx}^x + \frac{1}{8}v_{yy}^x + \frac{1}{4}v_{xy}^y\right),\tag{8.7}$$

where v_{xy}^y denotes $\partial^2 v^x/\partial x\,\partial y$ and v^x denotes the x component of the velocity with a similar notation for the other terms. This shows that the shear viscosity coefficient $\eta = \rho\alpha h c/8$ and the bulk viscosity coefficient $\zeta = 5\eta/3$. Similar analysis in three dimensions shows that $\eta = \rho\alpha h c/10$ and $\zeta = 5\eta/3$.

If there are rapid changes in the parameters then the same argument used for the case of heat conduction with discontinuous thermal conductivity can be used. If we define (for two dimensions)

$$\mu_a = \tfrac{1}{8}\alpha_a h_a c_a \rho_a \tag{8.8}$$

and define a new Π_{ab} according to

$$\Pi_{ab} = -\frac{16\mu_a\mu_b}{\rho_a\rho_b(\mu_a + \mu_b)}\left(\frac{\boldsymbol{v}_{ab}\cdot\boldsymbol{r}_{ab}}{r_{ab}^2 + \epsilon\bar{h}_{ab}^2}\right), \tag{8.9}$$

we produce a viscosity term which can be used for real viscosities and maintains the continuity of viscous stress accurately. This SPH viscosity was proposed by Cleary (1998) and applied to the simulation of viscous liquids in flow modelling for casting processes (see, e.g. Cleary and Ha (2002) and the references therein). Cleary determined the coefficient by numerical experiment and found that coefficient 16 should be replaced by 19.8. In three dimensions (where the factor 1/8 in (8.8) is replaced by 1/10) the analysis suggests a coefficient of 20.

In the case of shock tube problems, it is usual to turn the viscosity on for approaching particles and turn it off for receding particles. In this way, the viscosity is used for shocks and not rarefactions. Unfortunately, in astrophysical calculations, this rule means that the viscosity is turned on when the density increases in the shock-free regions, for example, when gravity pulls gas together.

When the viscosity term Π_{ab} was first used (Monaghan and Gingold 1983) it was found to work well for shocks of moderate strength. However, in astrophysical calculations involving colliding gas clouds, where the Mach number can be very high, it was found that particles from one cloud could stream between the particles of the other cloud. Generally, this streaming is limited to a few particle spacings, and is, therefore, not a severe problem; however, it should not occur at all. To prevent it, an extra term was added to ν which then took the form (Monaghan 1992)

$$\nu = \frac{\bar{h}_{ab}}{\bar{\rho}_{ab}}\left(\alpha\bar{c}_{ab} - \beta\frac{\bar{h}_{ab}\boldsymbol{v}_{ab}\cdot\boldsymbol{r}_{ab}}{r_{ab}^2 + \epsilon\bar{h}_{ab}^2}\right). \tag{8.10}$$

This form of ν, and hence Π_{ab}, evolved through various forms starting with the work of Lattanzio *et al* (1985) on interstellar cloud collisions. Good results have been obtained with the choice $\alpha = 1$ and $\beta = 2$. This form of the viscosity, though changed in details, is found naturally by considering aspects of the dissipative term in shock solutions based on Riemann solvers (Monaghan 1997). In this case

$$\Pi_{ab} = -\frac{Kv_{\text{sig}}(\boldsymbol{v}_{ab}\cdot\boldsymbol{r}_{ab})}{\bar{\rho}_{ab}|r_{ab}|}, \tag{8.11}$$

where $K \sim 0.5$. The signal velocity v_{sig} is defined by

$$v_{\text{sig}} = c_a + c_b - \beta\boldsymbol{v}_{ab}\cdot\hat{\boldsymbol{r}} \tag{8.12}$$

where $\hat{\boldsymbol{r}} = \boldsymbol{r}_{ab}/|\boldsymbol{r}_{ab}|$ and $\beta \sim 4$. The signal velocity can be interpreted as follows. If the fluid is at rest we estimate the speed at which a sound wave from a approaches a sound wave from b as $(c_a + c_b)$. The extra term represents the change in speed if the fluids at a and b are moving relative to each other. The fact that this must be a Galilean invariant, and would vanish if they have the same velocity or rotate rigidly, leads directly to the form shown. This is discussed further by Monaghan (1997).

8.2. Viscous heating and the energy equations

Viscosity dissipates the flow and transfers energy from kinetic to thermal. The contribution to the thermal energy is always positive. Owing to the way in which the SPH viscosity was derived, viscous dissipation is best obtained directly from the SPH equations. By taking the scalar product of v_a and the acceleration equation, multiplying by m_a and summing over a, the viscous contribution to the rate of change of thermal energy can be identified (Monaghan and Gingold 1983, Monaghan 1997). The final result is the thermal energy equation

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\Omega_a \rho_a^2} \sum_b m_b v_{ab} \cdot \nabla_a W_{ab} + \frac{1}{2} \sum_a m_a \sum_b m_b \Pi_{ab} v_{ab} \cdot \nabla_a W_{ab}. \tag{8.13}$$

Referring now to the definition of Π_{ab}, for example to (8.3), the contribution to the viscous dissipation of particle a from b can be written as (using the same definition of $F_{ab} \leqslant 0$ as before)

$$-\left(\frac{\alpha \bar{h}_{ab} \bar{c}_{ab}}{\bar{\rho}_{ab}}\right) \frac{F_{ab}(v_{ab} \cdot r_{ab})^2}{r_{ab}^2 + \eta^2}, \tag{8.14}$$

which is $\geqslant 0$. This confirms that the SPH dissipation increases the thermal energy as it should. In addition to increasing the thermal energy, the viscous dissipation should increase the total entropy of the system. From the first law of thermodynamics

$$T\frac{\mathrm{d}s}{\mathrm{d}t} = \mathrm{d}u - \frac{P}{\rho^2}\,\mathrm{d}\rho. \tag{8.15}$$

In SPH form this becomes

$$T_a\frac{\mathrm{d}s_a}{\mathrm{d}t} = \frac{1}{2} \sum_b m_b \Pi_{ab} v_{ab} \cdot \nabla_a W_{ab} \tag{8.16}$$

and from the previous results the change in the entropy of any particle owing to viscous dissipation is positive.

8.3. Dissipation and the thermokinetic energy equation

It was shown earlier (section 3) that the thermokinetic equation takes the form

$$\frac{\mathrm{d}\hat{e}_a}{\mathrm{d}t} = -\sum_b m_b \left(\frac{P_a v_b}{\rho_a^2} + \frac{P_b v_a}{\rho_b^2}\right) \cdot \nabla_a W_{ab}, \tag{8.17}$$

where $\hat{e}_a = \frac{1}{2}v_a^2 + u_a$. In order to use this equation for shock phenomena it is necessary to add dissipative terms. Although these could be deduced by beginning with the definition of $\hat{e}$ and SPH equations for the derivatives, it is more convenient to be guided by ideas from Riemann solvers (Monaghan 1997). Hence, we need to add a dissipative term Υ_{ab} to the pressure–velocity terms in (8.17) where

$$\Upsilon_{ab} = -\frac{K v_{\mathrm{sig}}(a,b)(e_a^* - e_b^*)\hat{r}}{\bar{\rho}_{ab}} \tag{8.18}$$

and

$$e_a^* = \frac{1}{2}(v_a \cdot \hat{r})^2 + u_a, \tag{8.19}$$

where $\hat{r} = r_{ab}/|r_{ab}|$. Replacing the actual kinetic energy with the kinetic term using the velocity along the line joining the particles a and b guarantees that the contribution to the thermal energy from viscous dissipation will be positive, and that the entropy will increase with time (Monaghan 1997). It is often assumed that the constant K and the signal velocity

v_{sig} are the same as in the dissipative term (8.11) in the acceleration equation but that is not necessary. Starting with the equation for the rate of change of $\hat{e}$

$$\frac{\mathrm{d}\hat{e}_a}{\mathrm{d}t} = -\sum_b m_b \left(\frac{P_a \boldsymbol{v}_b}{\rho_a^2} + \frac{P_b \boldsymbol{v}_a}{\rho_b^2} + \Upsilon_{ab} \right) \cdot \nabla_a W_{ab}, \tag{8.20}$$

it is possible to deduce the rate of change of thermal energy (Monaghan 1997). This takes the following form

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\rho_a^2} \sum_b m_b \boldsymbol{v}_{ab} \cdot \nabla_a W_{ab} + \text{dissipative term}, \tag{8.21}$$

where the dissipative term is

$$\sum_b m_b \frac{K v_{\text{sig}}(a,b)}{\bar{\rho}_{ab}} \left(u_a - u_b - \frac{1}{2}(\boldsymbol{v} \cdot \hat{\boldsymbol{r}})^2 \right) |\boldsymbol{r}_{ab}| F_{ab}. \tag{8.22}$$

The terms involving u, namely,

$$\sum_b m_b \frac{K v_{\text{sig}}(a,b)}{\bar{\rho}_{ab}} (u_a - u_b) |\boldsymbol{r}_{ab}| F_{ab} \tag{8.23}$$

give heat diffusion. This expression, is a variant on the heat diffusion term described in section 7.1 and has similar properties. The diffusion coefficient is proportional to $v_{\text{sig}}(a,b)|\boldsymbol{r}_{ab}|$. A heat diffusion conduction term was used in the thermal energy by Lattanzio and Monaghan (1991) in their discussion of fragmenting molecular gas clouds.

8.4. Reducing artificial dissipation

Artificial dissipation is very successful for handling shocks but it can be too large in other parts of the flow. For example, artificial viscous dissipation increases the Reynolds number of a flow, artificially, with the result that, for example, the Kelvin–Helmoltz shear instabilities are heavily diffused. Balsara (1995) suggested reducing viscous dissipation by multiplying Π_{ab} by the factor

$$\frac{|\nabla \cdot \boldsymbol{v}|}{|\nabla \cdot \boldsymbol{v}| + |\nabla \times \boldsymbol{v}|}, \tag{8.24}$$

made symmetric, for example, by replacing $\nabla \cdot \boldsymbol{v}$ by the average for the interacting pair of particles. Colagrossi (2004) found that it is preferable to replace the previous factor by

$$\frac{|\nabla \cdot \boldsymbol{v}|_{ab}}{|\nabla \cdot \boldsymbol{v}|_{ab} + \sqrt{E^{ij} E^{ij}} + 10^{-4} \bar{c}_{ab}/h}, \tag{8.25}$$

where c is the speed of sound and the rate of strain tensor E^{ij} is defined by

$$E^{ij} = \frac{1}{2} \left(\frac{\partial v^i}{\partial x^j} + \frac{\partial v^j}{\partial x^i} \right). \tag{8.26}$$

The indices denote cartesian tensors and the summation convention is used in evaluating $E^{ij} E^{ij}$. Colagrossi (2004) found that the replacement of $\nabla \times \boldsymbol{v}$ with the term involving E^{ij} gave improved results for problems involving slightly compressible fluids. A particularly impressive example being the rotation in two dimensions of a rotating square of water in an otherwise empty space.

Another very useful approach is to note that the dissipation terms have the same coefficients K and v_{sig} in both Π_{ab} and Υ_{ab}. In general, different coefficients, or signal speeds could be used for the viscous and the thermal energy terms. Furthermore, each particle can have its own

coefficient determined by the conditions it encounters. Morris and Monaghan (1997) explored this idea for the artificial viscous terms in gas dynamics where it is desirable to reduce the viscosity away from shocks. In finite difference calculations this is achieved by switches based on the first and second spatial derivatives of physical variables such as the momentum flux. However, spatial derivatives sit uncomfortably with particle methods for which time derivatives are more natural. The basic problem is like trying to predict the onset of a stock market crash from the time variation of the market. For shock simulation the coefficient α in (8.4) or equivalently K in (8.11) should be different for each particle and should change with time according to the conditions the particle is in, becoming large at shocks, but relaxing back to a small value when the flow is calmer. A simple way to do this for a typical particle a is to determine its α_a from the equation

$$\frac{\mathrm{d}\alpha_a}{\mathrm{d}t} = -\frac{(\alpha_a - \alpha_0)}{\tau} + S_a, \tag{8.27}$$

where τ is a suitable time scale $\propto h/c_{\mathrm{s}}$, $\alpha_0 \sim 0.1$ is the ambient value of α, and S_a is a source term. S should increase as the particle approaches a shock in such a way that α increases to approximately 1. Morris and Monaghan (1997) discuss a choice of $S \propto \nabla \cdot \boldsymbol{v}$ and show that it gives good results. Rosswog *et al* (2000) take

$$S = \mathrm{Max}(-\nabla \cdot \boldsymbol{v}, 0)(2 - \alpha). \tag{8.28}$$

Not only are the good results for the shocks retained, but elsewhere in the flow the viscosity is also reduced by approximately a factor of 10. However, in many astrophysical problems, where a collapse of gas clouds occurs, $-\nabla \cdot \boldsymbol{v}$ can increase without the occurrence of shocks. It would, therefore, be desirable to relate S to some other quantity related to the change of entropy.

Price (2004a) suggested that the artificial thermal conductivity should also vary with each particle and proposed a similar equation to that for α but using a source term proportional to $|\nabla \sqrt{u}|$. In terms of (8.18) it means splitting the dissipative term into a viscous part and a heat conduction part and using a different K for each. Price obtains improved results for shock tube phenomena especially near contact discontinuities. In the same way he tested a dissipation term for the magnetic fields in MHD simulations and found it improved his SPH simulations. Whether or not the onset of a shock could be predicted more satisfactorily with higher derivatives is an open question.

9. Applications to shock and rarefaction problems

There have been widespread applications of SPH to shocks in gases, liquids and metals (see, e.g. Libersky and Petschek (1991), Johnson *et al* (1996), Monaghan (1997)). There is only space here to describe some elementary examples from gas dynamics.

The first case we consider is the rarefaction wave. This can be set up by placing SPH particles in the region $-0.5 \leqslant x \leqslant 0.5$ with uniform separation Δx and the density $\rho = 1$. For this example, we use 200 particles and set $\gamma = 1.4$, the initial $h = 1.5\Delta x$, and the thermal energy/mass is 2. The SPH acceleration, continuity and thermal energy equation were integrated. In figure 14 the velocity field for $x \geqslant 0$ is shown. The exact velocity field is shown by the solid line and the SPH results are shown by solid diamonds. The agreement between the two is excellent.

We now consider the shock tube used by Sod (1978) as a test for numerical techniques. The system is one-dimensional with uniform conditions on each side of a diaphragm which breaks at $t = 0$. To the left of the diaphragm ($x < 0$) the conditions are $\rho, P, v, \gamma = 1.0$, 1.0, 0.0, 1.4 and to the right (0.125, 0.1, 0.0, 1.4). The evolved system consists of (from the left), the undisturbed original conditions, a rarefaction, a contact discontinuity and a shock.
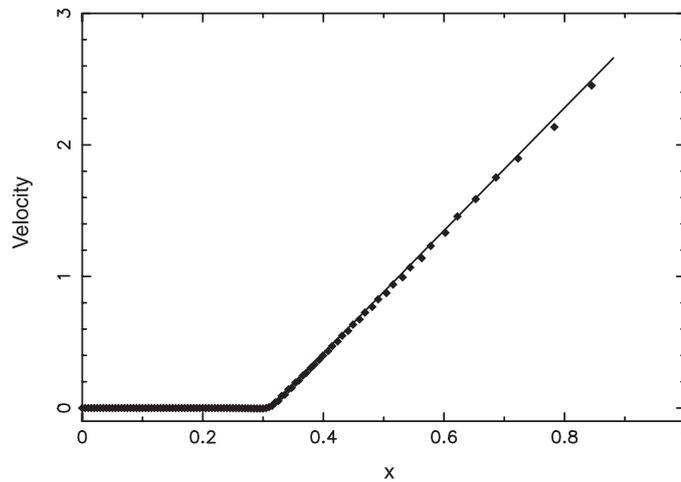
Figure 14. The velocity field for the one-dimensional rarefaction waves from the expansion of uniform gas initially in the region $-0.5 \leqslant x \leqslant 0.5$ is shown. The results for the right half $x \geqslant 0$ of the domain are also shown. The exact velocity field is shown by the solid line and the SPH results by the solid diamonds.
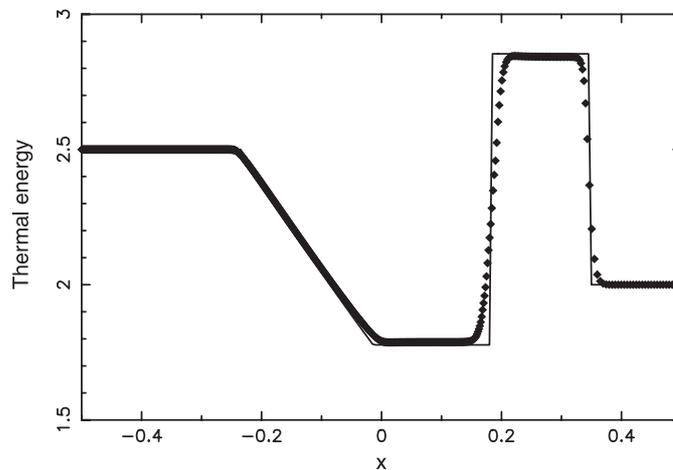

Figure 15. The thermal energy for a shock tube where the initial density ratio is $8:1$. The particles have equal mass but the spacing is a factor of 8 smaller to the left of the initial diaphragm. The simulation uses 100 particles to the right of the initial diaphragm and 800 particles to the left. The exact results are shown by the continuous line and the SPH results by solid diamonds.

Between the shock and the rarefaction the pressure and velocity are constant. The density and thermal energy change discontinuously at the contact discontinuity. In this simulation the viscosity (8.3) with (8.10) was used with the coefficients: $\alpha = 1$ and $\beta = 2$. The particles have equal mass. Since there is an initial discontinuity in all the properties other than the initial velocity, the density and thermal energy are smoothed at the interface (as in Monaghan (1997)). Hence to be consistent with the particles having constant mass and the density being smooth, we must smooth the spacing. Price (2004a) finds that using h, calculated consistently with the density, according to (4.2) and (4.3) gives better results. The thermal energy is shown in figure 15 and the velocity field in figure 16. The solid lines show the exact results. The
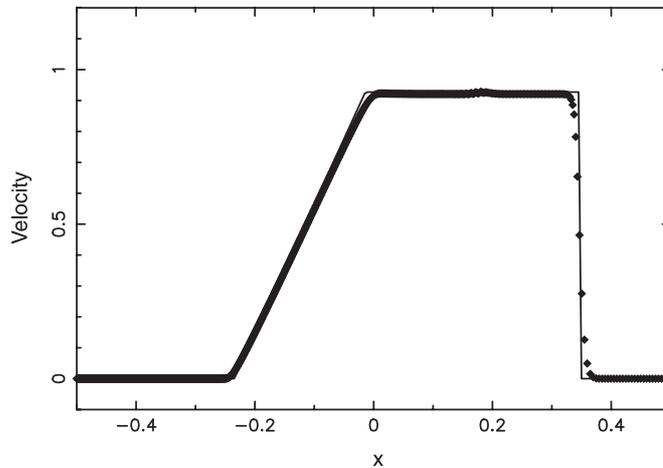
Figure 16. The velocity field for a shock tube. The exact results are shown by the continuous line and the SPH results by solid diamonds. The value of h and the particle spacing becomes smaller as the gas passes through the shock. The width is, therefore, ~ 3 initial particle spacings. Note the variation in the velocity owing to a slight jump in the pressure at the contact discontinuity.

post-shock values are accurate to within 1%, though the shock fronts are broader than the comparable Riemann solver shocks. The actual broadening is smaller than the number of particles across the shock would suggest because, on entering the shock, the particles move closer together, and h becomes smaller.

10. Applications of SPH to liquids

A liquid such as water is slightly compressible but, for many fluid dynamical problems, it can be approximated by an artificial incompressible fluid, and this is the basis of most of the finite difference numerical algorithms for liquids. An alternative approach, better suited to SPH, is to approximate the liquid by an artificial fluid which is slightly compressible. All that is required is that the speed of sound be large enough for the density fluctuations to be negligible (Monaghan 1994). The equation of state most frequently used is due to Cole (1948), (see also Batchelor (1974)) which, when atmospheric pressure is negligible, has the form

$$P = B \left(\left(\frac{\rho}{\rho_0} \right)^{\gamma} - 1 \right), \tag{10.1}$$

where ρ_0 is a reference density, $\gamma \sim 7$ and B is chosen so that the speed of sound is large enough to keep the relative density fluctuation $|\delta\rho|/\rho$ small. Since

$$\frac{|\delta\rho|}{\rho} \sim \frac{v^2}{c_s^2}, \tag{10.2}$$

where v is the maximum speed of the fluid we can ensure $|\delta\rho|/\rho \sim 0.01$ if $v/c_s < 0.1$. The speed of sound at the reference density is

$$c_s^2 = \frac{\gamma B}{\rho_0}. \tag{10.3}$$

Therefore, if $B = 100\rho_0 v^2/\gamma$, the relative density fluctuations should be ~ 0.01. This requires an estimate of the maximum speed to be found which is often very easy to do. An example

of the results that can be achieved with SPH is shown in figure 5, in section 2, where the SPH calculations of Colagrossi (2004) are compared with those from a combination of level set and finite difference methods. This figure shows the evolution of liquid which is initially in the shape of a square and is distorted by a velocity field with spatially varying viscosity.

10.1. Boundaries

Most problems involving liquids also involve boundaries which may be fixed or moving or they might represent the surfaces of rigid bodies, wholly or partially, within the fluid. These boundaries may be handled easily by replacing the boundary with particles which interact with the fluid with prescribed forces. In this way, complicated problems involving fluids interacting with rigid bodies (which may float) and contained within an arbitrarily moving rigid body can be treated easily. An example would be the dynamics of a damaged car ferry with water pouring into the decks containing the cars.

Let f_{ka} be the force per unit mass on boundary particle k due to fluid particle a. To ensure that linear and angular momentum of the entire system is conserved in the absence of external forces, the force on a due to k must be equal and opposite to the force on k due to a. The most obvious way to specify the forces would be to use a Lennard–Jones force acting between the centres of the particles (Monaghan 1994). However, the large variation in the force on a particle moving parallel to the boundary causes large disturbance to flow near a boundary. A better procedure is the following (Monaghan *et al* 2004).

Consider the interaction between a fluid particle a and a boundary particle k where the local unit normal to the boundary is $\boldsymbol{n}_k$. If the distance measured normal to the boundary, from the boundary particle to the fluid particle, is denoted by y and the tangential distance by x, then a suitable form for the force per unit mass on boundary particle k due to fluid particle a is

$$\boldsymbol{f}_{ka} = -\frac{m_a}{m_a + m_k} B(x, y)\boldsymbol{n}_k, \tag{10.4}$$

where $B(x, y)$ is chosen to ensure that B rapidly increases as y decreases towards zero (to prevent penetration of the walls) and the variation with x ensures that the force on a particle moving parallel to the wall is constant. The total force/unit mass on boundary particle k due to all fluid particles is then $\boldsymbol{f}_k = \sum_a \boldsymbol{f}_{ka}$. The force per unit mass on fluid particle a due to boundary particle k is

$$\boldsymbol{f}_{ak} = \frac{m_k}{m_a + m_k} B(x, y)\boldsymbol{n}_k, \tag{10.5}$$

so that the forces $m_k \boldsymbol{f}_{ka} = -m_a \boldsymbol{f}_{ak}$ are equal and opposite. The total force per unit mass on fluid particle a from all boundary particles is $\boldsymbol{f}_a = \sum_k \boldsymbol{f}_{ak}$.

The equation of motion of a fluid particle a is then

$$\frac{\mathrm{d}v_a}{\mathrm{d}t} = -\sum_b m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2} + \Pi_{ab} \right) \nabla_a W_{ab} + \boldsymbol{f}_a. \tag{10.6}$$

10.2. Motion of a rigid body interacting with a liquid

If the system consists of a liquid containing a rigid body with centre of mass $\boldsymbol{R}$ and centre of mass velocity $\boldsymbol{V}$, the equations of motion of this body are first, the equation for the motion of the centre of mass

$$M\frac{\mathrm{d}\boldsymbol{V}}{\mathrm{d}t} = \sum_k m_k \boldsymbol{f}_k, \tag{10.7}$$

where the summation over k only refers to the boundary particles on the surface of the rigid body. Second, the equation for the angular velocity Ω about the centre of mass which, in the case of 2D motion, is

$$I\frac{d\Omega}{dt} = \tau,\tag{10.8}$$

where I is the moment of inertia (a scalar for the present case) and τ is the total torque about the centre of mass. The torque can be calculated from the forces on the boundary particles of the rigid body. We then get

$$I\frac{d\Omega}{dt} = \sum_k m_k(\boldsymbol{r}_k - \boldsymbol{R}) \times \boldsymbol{f}_k,\tag{10.9}$$

where the direction of Ω is perpendicular to the plane of the motion and $\boldsymbol{R}$ is the position of the centre of mass. The rigid body boundary particles move as part of the rigid body so that the change in position of boundary particle k is given by

$$\frac{d\boldsymbol{r}_k}{dt} = \boldsymbol{V} + \Omega \times (\boldsymbol{r}_k - \boldsymbol{R}).\tag{10.10}$$

From (10.6) and (10.7) we get

$$\frac{d}{dt}\left(\sum_a m_a \boldsymbol{v}_a + M\boldsymbol{V}\right) = 0,\tag{10.11}$$

because the pair forces in each term cancel. Linear momentum is, therefore, conserved. To prove the conservation of angular momentum is a little more complicated. Since it has not been given in the literature, I now give it here.

The rate of change of the angular momentum of the rigid body about a fixed origin, when the motion takes place in a plane, is

$$\frac{d\boldsymbol{J}}{dt} = M\boldsymbol{R} \times \frac{d\boldsymbol{V}}{dt} + I\frac{d\Omega}{dt}.\tag{10.12}$$

Using the previous equations the right-hand side becomes

$$\boldsymbol{R} \times \sum_k m_k \boldsymbol{f}_k + \sum_k m_k(\boldsymbol{r}_k - \boldsymbol{R}) \times \boldsymbol{f}_k = \sum_k m_k \boldsymbol{r}_k \times \boldsymbol{f}_k\tag{10.13}$$

$$= \sum_k \sum_a m_k \boldsymbol{r}_k \times \boldsymbol{f}_{ka}.\tag{10.14}$$

The rate of change of the angular momentum of the liquid is

$$\sum_a m_a \boldsymbol{r}_a \times \frac{d\boldsymbol{v}_a}{dt} = \sum_a \sum_k m_a \boldsymbol{r}_a \times \boldsymbol{f}_{ak},\tag{10.15}$$

because the pressure forces give zero net contribution to the total angular momentum of the fluid as we showed earlier. Adding the rate of change of angular momentum of the rigid body and the liquid and recalling that $m_k \boldsymbol{f}_{ka} = -m_a \boldsymbol{f}_{ak}$ gives

$$\sum_k \sum_a m_a(\boldsymbol{r}_a - \boldsymbol{r}_k) \times \boldsymbol{f}_{ka} = \sum_k \sum_a \frac{m_a m_k}{m_a + m_k}(\boldsymbol{r}_a - \boldsymbol{r}_k) \times \hat{\boldsymbol{n}}_k B(x, y).\tag{10.16}$$

Now consider the contribution to the rate of change of angular momentum from a liquid particle a. Suppose this particle lies between two boundary particles k and $(k+1)$ and suppose the tangential distance to k is x and to $(k+1)$ is $(1-x)$ assuming the unit of length is the separation of the boundary particles. The contribution from the previous summation is then

$$xB(x, y) - (1-x)B(1-x, y) = 0\tag{10.17}$$

provided $B(x, y) = (1 - x)\Gamma(y)$ which is the choice made in the next section. The rate of change of total angular momentum is, therefore, zero. This proof requires that the masses of the boundary particles are equal. If they are not equal then the forces must be scaled so that the torque from neighbouring boundary particles vanishes.

10.3. The boundary force

Monaghan *et al* (2004) write $B(x, y)$ as a product $\Gamma(y)\chi(x)$ where the function $\chi(x)$ is defined by

$$\chi(x) = \begin{cases} \left(1 - \dfrac{x}{\Delta p}\right), & \text{if } 0 < x < \Delta p, \\ 0, & \text{otherwise,} \end{cases}$$

where Δp is the boundary particle spacing. This factor ensures that a fluid particle moving parallel to the wall will always feel the same force because, when it is between any two boundary particles, the total force from them is constant, regardless of where it lies between them.

The essential condition on the function $\Gamma(y)$ is that it increases as y decreases to prevent penetration of the wall. Monaghan *et al* (2004) choose a form related to the gradient of the cubic spline with the argument $q = y/h$. The gradient of the cubic spline has a maximum at $q = 2/3$. For $0 < q < 2/3$ they replace the value of the gradient by its maximum. Thus,

$$\Gamma(y) = \beta \begin{cases} \frac{2}{3}, & \text{if } 0 < q < \frac{2}{3}, \\ (2q - \frac{3}{2}q^2), & \text{if } \frac{2}{3} < q < 1, \\ \frac{1}{2}(2 - q)^2, & \text{if } 1 < q < 2, \\ 0, & \text{otherwise,} \end{cases}$$

where β is $0.02\, c_s^2/y$. This term is an estimate of the maximum force/mass necessary to stop a particle moving at the estimated maximum speed. The factor $1/y$ ensures that a faster moving particle can be stopped. Other details concerning boundaries, including the treatment of corners, are given by Monaghan *et al* (2004). If there is more than one rigid body interacting with the fluid, then the same methods can be used but now there may be an interaction between the bodies. There are, therefore, two types of boundary forces, namely, boundary–fluid and boundary–boundary. The best choice of boundary force is not known.

Other authors (e.g. Colagrossi and Landrini 2003) prefer to replace the rigid boundaries by ghost particles. This has advantages when the geometry is simple because the use of ghost particles gives less disturbance to the fluid. However, for complicated geometries, for example, those describing engine body parts in liquid metal moulding (Cleary and Ha 2002), or in geophysical flows, boundary particles are easier to use and can be more accurate. A simple generalization is to allow the boundary particles to have a different interaction with different fluids. For example, in the case of a dusty gas, the dust and gas SPH particles could interact with the boundary particles with different forces.

10.4. Applications to rigid bodies in water

The early applications were to bores, dam collapse, wave makers and breaking waves, though not to a high accuracy because only a small number of particles were used (Monaghan 1994). Further applications, with comparisons between SPH and experimental results for waves on beaches were made by Monaghan and Kos (1999), who also studied the generation of solitary waves by dropping boxes (Monaghan and Kos 2000) or by sliding boxes down ramps

Monaghan *et al* (2004). The mechanism by which a rising bubble could sink a ship was studied using experiments and SPH simulations by May and Monaghan (2003). Colagrossi and Landrini (2003) describe applications to more than one fluid in dam break (see also Colicchio *et al* (2002)). They also considered the rise of bubbles in water and the effect of air on wave breaking. Their work incorporates a number of improvements for these problems including a periodic re-initialization of the density field based on moving least squares interpolation (Belytschko *et al* 1998), and a generalized Balsara correction (discussed in section 9). Because of the contrast in density they use the acceleration equation (5.35) with $\Phi = 1$, and the convergence equation (5.26).

Sloshing tanks have been studied by Colagrossi (2004) and Colagrossi *et al* (2003) who found that the SPH simulations revealed an aspect of the sloshing not noticed previously. Subsequent specially designed experiments confirmed this prediction.

10.5. Turbulence

There have been limited studies of turbulence using SPH. Studies of wave breaking by Colagrossi (2004), and Landrini *et al* (2003) show that detailed properties of the complex vortices resulting from wave breaking can be recovered using SPH. A fully Lagrangian turbulence model based on the Lagrangian averaged alpha model (Holm 1999), Mohseni *et al* 2003) has been worked out (Monaghan 2002, 2004), but no comparisons have been made with other, more traditional, methods. In the SPH Lagrangian averaged model a typical particle a is moved with the XSPH smoothed velocity $\hat{v}_a$ (Monaghan 1989). This was originally defined by

$$\hat{v}_a = v_a + \epsilon \sum_b m_b \frac{(v_b - v_a)}{\bar{\rho}_{ab}} W_{ab}, \qquad (10.18)$$

where $\epsilon \sim 0.5$ is constant and the kernel need not be the same as the kernel used in calculating the density. This smoothed velocity brings the particle velocity closer to the average velocity in its neighbourhood and reduces the particle disorder. Moving the particles with the new velocity does not change the linear or angular momentum. However, if the particles are moving with the smoothed velocity, energy is not conserved. To conserve it, and bring the algorithm into agreement with the alpha model of turbulence, we replace (10.18) with

$$\hat{v}_a = v_a + \epsilon \sum_b m_b \frac{(\hat{v}_b - \hat{v}_a)}{\bar{\rho}_{ab}} W_{ab}. \qquad (10.19)$$

Moving particles with this velocity still conserves linear and angular momentum. In the continuum limit the previous equation becomes

$$\hat{v}_a = v_a + \frac{1}{2} \frac{\epsilon}{\rho} \nabla^j \left(\rho \nabla^j \hat{v}_a \right) \left(\int q^2 W(q, h) \, dq \right), \qquad (10.20)$$

where $\nabla^j = \partial/\partial x^j$. To compare with the continuum α model define

$$\alpha_{\text{turb}} = \frac{1}{2} \epsilon \int q^2 W(q, h) \, dq \sim \epsilon h^2, \qquad (10.21)$$

so that (10.20) agrees, when ρ is constant, with Holm (1999).

The smoothing algorithm is similar to a discrete time-stepping of a diffusion equation. For example, the diffusion equation

$$\frac{dv}{d\tau} = \frac{\kappa}{\rho} \nabla^j (\rho \nabla^j v) \qquad (10.22)$$

can be approximated by the implicit, discrete time stepping

$$v^{n+1} = v^n + \delta\tau \frac{\kappa}{\rho} \nabla^j (\rho \nabla^j v^{n+1}). \tag{10.23}$$

If $\hat{v}$ is identified with v^{n+1}, v with v^n, the two equations (10.20) and (10.23) become equivalent. Because the implicit smoothing is stable it can be used for any value of $\epsilon > 0$. However, in practice, this implicit equation must be solved by iteration and several iterations may be needed if $\epsilon > 1$.

To complete the dynamics the most convenient approach is to use a Lagrangian (Monaghan 2002).

$$L = \sum_b m_b \left(\frac{1}{2} \hat{v}_b \cdot v_b - u_b - \Phi_b \right), \tag{10.24}$$

where Φ is a potential energy. The kinetic energy term can be written as

$$\frac{1}{2} \sum_b m_b \hat{v}_b \cdot \hat{v}_b + \frac{\epsilon}{4} \sum_a \sum_b m_a m_b \frac{\hat{v}_{ab}^2}{\bar{\rho}_{ab}} W_{ab}, \tag{10.25}$$

from which the canonical momentum of particle a can be calculated. Remarkably, it is just $m_a v_a$.

The smoothing of the velocity makes the Lagrangian averaged model similar to the large Eddy simulation method. However, the Lagrangian leads to a different set of equations from those used in LES simulations and variable resolution length is built into the equations. An interesting feature of these equations is that, in the absence of any dissipation, they result in the energy being redistributed so that the energy transfer to short length scales is impeded (Mohseni *et al* 2003, Monaghan 2004). Various aspects of these equations are discussed by Monaghan (2002, 2004). There is a need to apply the SPH turbulence model to standard problems such as turbulence decay in two- and three-dimensional boxes. An interesting astrophysical example to study would be the turbulence in toy stars.

Particle methods lead naturally to the idea of studying turbulence along the lines of statistical mechanics, that is in terms of the velocity and spatial distributions of the particles. No detailed work on this has appeared in the literature though there has been some analysis of probability distributions using SPH (Welton 1998, Welton and Pope 1997).

10.6. Multiphase flow

It is straightforward to include more than one fluid in SPH simulations. Each fluid has its own set of SPH particles with an appropriate equation of state. All the SPH particles are used in the summations. If the fluids are incompressible, the technique described earlier, where the speed of sound is artificial, and sufficiently large to make density fluctuations negligible, can be used. Gravity currents flowing into a stratified fluid have been studied using both experiment and simulation (Monaghan *et al* 1999) and air–water interactions have been simulated successfully by Colagrossi and Landrini (2003). Dusty gas occurs in both astrophysics and in volcanic outbursts. A formulation of SPH suitable for dusty gas (Monaghan and Kocharyan 1995) is available but no applications have appeared in the literature, yet.

11. Elasticity and fracture

The equations of elastic dynamics are the acceleration equation

$$\frac{\mathrm{d}v^i}{\mathrm{d}t} = \frac{1}{\rho} \frac{\partial \sigma^{ij}}{\partial x^j} + g^i, \tag{11.1}$$

where the stress tensor is given by

$$\sigma^{ij} = -P\delta^{ij} + S^{ij}, \tag{11.2}$$

S^{ij} is the deviatoric stress and g^i denotes the ith component of a body force per unit mass. In linear elastic theory, the deviatoric stress can be obtained from Hooke's law with shear modulus μ

$$\frac{\mathrm{d}S^{ij}}{\mathrm{d}t} = 2\mu\left(\dot{\epsilon}^{ij} - \frac{1}{3}\delta^{ij}\dot{\epsilon}^{kk}\right) + S^{ik}R^{jk} + R^{ik}S^{kj}, \tag{11.3}$$

where

$$\dot{\epsilon}^{ij} = \frac{1}{2}\left(\frac{\partial v^i}{\partial x^j} + \frac{\partial v^j}{\partial x^i}\right) \tag{11.4}$$

and

$$R^{ij} = \frac{1}{2}\left(\frac{\partial v^i}{\partial x^j} - \frac{\partial v^j}{\partial x^i}\right). \tag{11.5}$$

Alternative laws for the time change of the deviatoric stress (Ellero *et al* 2002) can be used without any change in the formalism. The pressure P is normally obtained from the Tillotson or Mie Gruniessen equation of state. The elastic equations can be converted into SPH form following the principles already established. In particular, the acceleration equation becomes

$$\frac{\mathrm{d}v_a^i}{\mathrm{d}t} = \sum_b m_b\left(\frac{\sigma_a^{ij}}{\rho_a^2} + \frac{\sigma_b^{ij}}{\rho_b^2} + \Pi_{ab}\right)\frac{\partial W_{ab}}{\partial x_a^j} + g^i \tag{11.6}$$

and the velocity derivatives in the equation for the deviatoric stress and the tensor R^{ij} can be estimated using the methods in section 2.2.

The elastic dynamical equations were first studied by Libersky and Petschek (1991). A comprehensive discussion by Randles and Libersky (1996) covers many aspects of elastic SPH. The elastic equations were combined with an elastic fracture model (Grady and Kipp 1987, Benz and Asphaug 1994, 1995) in order to study asteroid/comet and planetesimal collisions (Michel *et al* (2004) give a comprehensive review of this work). The brittle fracture model of Grady and Kipp (1987) is based on Griffiths theory of fracture. A set of flaws (cracks) is assigned to the SPH particles at random, according to the Weibull distribution. Depending on the flaw, tension may or may not cause it to grow. The growth is associated with the local damage quantified by a damage parameter D. When D is zero it means that the material is perfectly elastic and when D increases to 1 the material is completely damaged and the contribution of the deviatoric stress is zero. The precise way in which the flaws are assigned, and the equation for D, are discussed in detail by Benz and Aspaugh.

Since material carries its damage with it, Lagrangian models like SPH are uniquely designed to model fracture. SPH, in particular, gives a good description of the fragments and provides a natural transition from the continuum to the fragmented state. This method has also been used to study fracture in and around volcanoes (Gray and Monaghan 2004).

In the initial application of SPH to elastic problems it was noticed that, under tension, particles tended to clump in pairs. This instability was first analysed by Phillips and Monaghan (1985) in the context of magneto-hydrodynamics. They showed that the tension which always exists in magnetic fields can cause an instability. The instability was re-discovered by Swegle *et al* (1995) in the context of elastic simulations and called the tensile instability. Various methods have been proposed to eliminate it from SPH simulations. The most successful has been the artificial stress method (Monaghan 2000) and Gray *et al* (2001) which also includes using the XSPH (Monaghan 1989) smoothed velocity. Others include

additional stress points (Dyka *et al* 1997), and the correction of the kernels to give exact linear interpolation (Dilts 1999, Bonet and Lok 1999, Bonet and Kulasegaram 2000, 2001). One skeleton in the SPH closet is that the normal SPH elastic equations do not conserve angular momentum. A spectacular example is a rotating elastic wheel which loses its rotation after one rotation or less. Hoover *et al* (2004) show that by using strong XSPH smoothing the loss of angular momentum could be reduced. Atomic models of elasticity conserve angular momentum exactly and it would be worth investigating whether an SPH elastic model can be based on the atomic models with the SPH particles mimicking atoms.

12. Special and general relativistic SPH

The continuum equations for special relativity can be derived from the derivative of the energy momentum tensor

$$\frac{\partial T^{\mu\nu}}{\partial x^\nu} = 0. \tag{12.1}$$

For a non-dissipative gas of baryons each with rest mass m_0, $T^{\mu\nu}$

$$T^{\mu\nu} = (nm_0c^2 + nu + P)U^\mu U^\nu + P\eta^{\mu\nu}, \tag{12.2}$$

where U^μ is a 4-velocity and c is the speed of light. In the following, the velocity unit is c and the energy unit is m_0c^2. In (12.1) n and u are the baryon number density and the energy/baryon in the rest frame of the element of fluid they refer to. The metric tensor $\eta^{\mu\nu}$ has the signature $(-1, 1, 1, 1)$. The resulting equations can be solved using a variety of computational algorithms. When the gas is ideal (i.e. $P = (\Gamma - 1)nu$) excellent results have been obtained using Riemann methods (Marti and Mueller 2003). For more complicated equations of state, for example, those that are used to mimic heavy ion collisions (Amsden *et al* 1978), particle in cell (PIC) methods have been used (Amsden *et al* 1978). In this latter case, after the collision, the rapidly expanding pion gas ceases to behave like a continuum fluid and behaves more like a set of particles in a process called 'freeze-out'. This situation would be very easy to simulate using SPH because it handles the transition from continuum fluid to particles seamlessly.

SPH equations for special relativity can be derived either from the continuum equations (Mann 1991, Laguna *et al* 1993, Chow and Monaghan 1997) or from a Lagrangian (Monaghan and Price 2001). The Lagrangian is

$$L = -\int T^{\mu\nu}U_\mu U_\nu \, d\boldsymbol{r}, \tag{12.3}$$

or

$$L = -\int n(1+u) \, d\boldsymbol{r}. \tag{12.4}$$

The SPH formalism can be set up in a selected frame, conveniently called the computing frame. In this frame the baryon number density is

$$N = nU^0 = n\gamma = n/\sqrt{(1-v^2)}. \tag{12.5}$$

Using standard SPH interpolation but replacing the mass m_b for SPH particle b by the number of baryons ν_b, Monaghan and Price (2001) show that the Lagrangian becomes

$$L = -\sum_b \nu_b \sqrt{(1-v_b^2)}(1+u_b), \tag{12.6}$$

with

$$N(\boldsymbol{r}) = \sum_b \nu_b W(|\boldsymbol{r} - \boldsymbol{r}_b|). \tag{12.7}$$

The Lagrangian equations then give the acceleration equation

$$\frac{\mathrm{d}\boldsymbol{p}_a}{\mathrm{d}t} = -\nu_a \sum_b \nu_b \left(\frac{P_a}{N_a^2} + \frac{P_b}{N_b^2} \right) \nabla_a W_{ab}, \tag{12.8}$$

where the canonical momentum $\boldsymbol{p}_a$ is given by

$$\boldsymbol{p}_a = \nu_a \left(1 + u_a + \frac{P_a}{n_a} \right). \tag{12.9}$$

This equation is identical to that obtained from the continuum equations by Chow and Monaghan (1997). To apply the SPH equations to strong shocks it is necessary to add dissipative terms. The dissipative terms for the non-relativistic case can be chosen by analogy with the actual viscosity and heat conduction of a gas. However, in the relativistic case there are no accurate relativistic dissipative terms (if they existed they would involve relativistic fields) to act as a guide The approach of Eckart (1940), and Landau and Lifshitz (1993) gives dissipative terms which are unstable (Hiscock and Lindblom 1985), while Carter's approach fails to correctly describe the non-relativistic gases (Olson and Hiscock 1990). Chow and Monaghan (1997), therefore, based their dissipation terms on those chosen for Riemann problems which, in the SPH form, are similar to those worked out by Amsden *et al* (1978) by considering baryon scattering. These dissipation terms are very effective and give a degree of accuracy comparable to methods based on Riemann solvers. The disappointing early SPH calculations of Mann (1991) and Laguna *et al* (1993) can be attributed to their poor choice of artificial viscosity. No attempt has yet been made to solve these problems with h and N, calculated consistently as described earlier for h and ρ.

The general relativistic equations for fluid dynamics in a specified metric can also be obtained from a Lagrangian. The resulting SPH equations (Monaghan and Price 2001) differ from those of Siegler and Riffert (2000) which do not conserve momentum. The shock calculations of Siegler and Riffert (2000) show unphysical jumps at the contact discontinuity. These are due to the lack of heat conduction in the dissipative terms. At present no satisfactory dissipative terms have appeared in the literature. One obvious approach would be to use the signal velocities found for Riemann solvers, then construct dissipative terms along the lines of those used by Chow and Monaghan (1997).

13. Prospects for the future

The features of SPH which make it an effective computational algorithm are ultimately due to the fact that it can be derived from a Lagrangian and has the conservation properties of a Lagrangian system. As a result, the conservation of momentum and energy together with the approximate invariant of the circulation follow naturally. However, SPH also conserves composition, that is, each particle carries its composition unchanged unless the material the particle represents undergoes chemical transformations. This property of carrying composition unchanged has not been fully exploited despite its importance in both industry and astronomy. In the latter case the extent to which elements are fully mixed in clusters of stars is known from observation, but has not been studied with simulations.

Another attractive feature of SPH is that the resolution adjusts smoothly to changes in the density, but there is no reason, other than computational efficiency, why the resolution could

not be changed in response to steep local gradients in other quantities, for example, temperature gradients. Preliminary steps have already been taken. These are:

- Direct splitting Kitsionas and Whitworth (2002).
- Adjustment of *h* and particle position Børve *et al* (2001).
- Regridding of the particles Chaniotis *et al* (2002).

All of these have defects which may be overcome. The first has not been tested for splitting and merging in problems where the split–merge rule depends, for example, on the temperature gradient, nor has it been tested for liquids with a stiff equation of state where density perturbations may lead to large pressure changes. The second is computationally intensive, however, that may be the price one has to pay for a better algorithm. The third has not been extensively tested with the split–merge rule based on a variety of gradients and, in current formulations, leads to excess diffusion, but this can be expected to be greatly reduced in the future.

To achieve a robust SPH algorithm for splitting and merging, it might be useful to reflect on what happens in nature. For example, when a gas moves into a region of high temperature the atoms smoothly ionize, producing more particles, then smoothly recombine if they enter a cooler state. This is exactly the process that would be natural for SPH and it could be implemented by allowing a particle to split, but placing the new particles close together so the effect on the flow will initially be negligible. The original particle, now less massive, could be tagged to provide a nucleus for merging. It mimics the role of the ion in the ionized gas example. The flow would gradually spread the new particles because of their slightly different velocities. The merging could occur by allowing the split particles to be attracted to the tagged particles. This would be a continuous process similar to the way ions and electrons in an ionized gas combine when cooled.

The next class of software advances would be the development of more efficient strategies for handling very low Mach number flows. These are required for industrial, geological and oceanographic hydrodynamic problems, and for simulating the dynamics of elastic materials. Recent work (Hu and Adams, preprint (2005)) has produced accurate and robust SPH algorithms for multi-phase flow including surface tension effects involving three fluids, but the maximum density ratio of the fluids is 100, which is an order of magnitude less than the air–water density ratio. This work improves on that of Colagrossi and Landrini described in this review. If the efficient strategies can be found then the low Mach number flows in geology could be handled efficiently. At present, the most straightforward application of SPH is in the volcanic outbursts as these are often close to the speed of sound. An implicit code would enable SPH codes to be devised for plate tectonics. In marine engineering we could look forward to simulations of water–metal impact leading to breakage and providing information about the stability of ships, especially those containing dynamically significant moving parts, as in a car ferry.

Within the category of software development we mention two, the first being concerned with MHD problems. We can expect significant advances in the next few years as the technique of Børve *et al* (2001) is made more efficient, and that of Price and Monaghan (2004a) is improved. The second is concerned with the multi-scaling problems, where calculations at the atomic level are linked to macroscopic dynamics. Many researchers have noted that SPH allows a seamless transition from the continuum to the fragments in problems involving fracture and splashing fluids. It is natural, therefore, to predict that SPH will provide an effective approach to multi-scaling simulations.

Finally, it is worth noting that apart from the advances in software there have been significant advances in hardware as well. New chips (FPGA), in particular, can be programmed

to implement the SPH summations in hardware. This will lead to SPH simulations which are extremely fast and will make previously difficult problems trivial.

References

Amsden A A, Goldhaber A S, Harlow F H and Nix J R 1978 Relativistic two-fluid model of nucleus–nucleus collsions *Phys. Rev.* C **17** 2080–96

Ball F K 1963 Some general theorems concerning the finite motion of a shallow liquid lying on a paraboloid *J. Fluid Mech.* **17** 240–56

Balsara D S 1995 von Neumann stability analysis of smooth particle hydrodynamics—suggestions for optimal algorithms *J. Comput. Phys.* **121** 357–72

Batchelor G K 1974 *An Introduction to Fluid Mechanics* (Cambridge: Cambridge University Press)

Bate M R, Bonnell I A and Bromm V 2003 The formation of a star cluster predicting the properties of stars and Brown Dwarfs *Mon. Not. R. Astron. Soc.* **399** 577–99

Bate M R, Bonnell I A and Price N M 1995 Modelling accretion in protobinary systems *Mon. Not. R. Astron. Soc.* **277** 362–76

Belytschko T, Krongauz Y, Organ D and Gerlack G 1998 On the completeness of meshfree particle methods *Int. J. Numer. Methods Eng.* **43** 785–819

Ben Moussa B, Lanson N and Vila J P 1999 Convergence of meshless methods for conservations laws: applications to Euler equations *Int. Ser. Numer. Math.* **129** 31–40

Benz W 1990 Smoothed particle hydrodynamics—a review *The Numerical Modelling of Nonlinear Stellar Pulsations* ed J R Buchler (Dordrecht: Kluwer) pp 269–88

Benz W and Asphaug E 1994 Impact simulations with fracture: I. Method and tests *Icarus* **1233** 98–116

Benz W and Asphaug E 1995 Simulations of brittle solids using smoothed particle hydrodynamics *Comput. Phys. Commun.* **87** 253–65

Benz W, Slattery W L and Cameron A G W 1986 The origin of the moon and the single impact hypothesis *Icarus* **66** 515–35

Bonet J and Kulasegaram S 2000 Correction and stabilization of smooth particle hydrodynamics methods with applications in metal forming simulations *Int. J. Numer. Methods Eng.* **47** 1189–214

Bonet J and Kulasegaram S 2001 Remarks on tension instability of Eulerian and Lagrangian corrected smooth particle hydrodynamics (CSPH) methods *Int. J. Numer. Methods Eng.* **52** 1203–20

Bonet J and Lok T-S L 1999 Variational and momentum preservation aspects of smooth particle hydrodynamic formulations *Comput. Methods Appl. Mech. Eng.* **180** 97–115

Boneva L I, Kendall D and Stepanov I 1971 Spline transformations: three new diagnostic aids for statistical data analysis *J. R. Stat. Soc.* B **33** 1–37

Børve S, Omang M and Trulsen J 2001 Regularized smoothed particle hydrodynamics: a new approach to simulating magnetohydrodynamic shocks *Astrophys. J* **561** 82–93

Brookshaw L 1985 A method of calculating radiative heat diffusion in particle simulations *Proc. Astron. Soc. Aust.* **6** 207–10

Carslaw H S and Jaeger J C 1990 *Conduction of Heat in Solids* (Oxford: Oxford University Press)

Chaniotis A K, Poulikakos D and Kououtsakos P 2002 Remeshed smoothed particle hydrodynamics for the simulations of viscous and heat conducting flows *J. Comput. Phys.* **182** 67–90

Chandrasekhar S 1995 *Newton's Principia for the Common Reader* (Oxford: Clarendon)

Chow E and Monaghan J J 1997 Ultrarelativistic SPH *J. Comput. Phys.* **134** 296–305

Cleary P W 1998 Modelling confined multi-material heat and mass flows using SPH *Appl. Math. Modelling* **22** 981–93

Cleary P W and Ha J 2002 Flow modelling in casting processes *Appl. Math. Modelling* **26** 171–90

Cleary P W and Monaghan J J 1999 Conduction modelling using smoothed particle hydrodynamics *J. Comput. Phys.* **148** 227–64

Colagrossi A 2004 Dottorato di Ricerca in Meccanica Teorica ed Applicata XVI CICLO A meshless Lagrangian method for free-surface and interface flows with fragmentation *PhD Thesis* Universita di Roma, La Sapienza

Colagrossi A and Landrini M 2003 Numerical simulation of interfacial flows by smoothed particle hydrodynamics *J. Comput. Phys.* **191** 448–75

Colagrossi A, Lugni C, Douset V, Bertram V and Faltinsen O 2003 Numerical and experimental study of sloshing in partially filled rectangular tanks *6th Numerical Towing Tank Symp. (Rome, Italy)*

Cole R H 1948 *Underwater Explosions* (Princeton, NJ: Princeton University Press)

Colicchio G, Colagrossi A, Greco M and Landrini M 2002 Free surface flow after a dam break *Ship Technol. Res.* **49** 95–104

Couchman H M P, Thomas P A and Pearce F R 1995 HYDRA: an adaptive mesh implementation of P^3M—SPH *Astrophys. J.* **452** 797–813

Dilts G A 1999 Moving least squares hydrodynamics: consistency and stability *Int. J. Numer. Methods* **44** 1115–55

Dyka C T, Randles P W and Ingel R P 1997 Stress points for tension instability in SPH *Int. J. Numer. Methods Eng.* **40** 2325–41

Eckart C 1940 The thermodynamics of irreversible processes: III. Relativistic theory of the simple fluid *Phys. Rev.* **58** 919–24

Eckart C 1960 Variation principles of hydrodynamics *Phys. Fluids* **3** 421–7

Ellero M, Kroger M and Hess S 2002 Viscoelastic flows studied by smoothed particle dynamics *J. Non-Newtonian Fluid Mech.* **105** 35–51

Español P and Revenga M 2003 Smoothed dissipative particle dynamics *Phys. Rev.* E **67** 026705

Feynman R P 1957 Applications of quantum mechanics to liquid helium *Low Temp. Phys.* **1** 17–53

Frank J and Reich S 2003 Conservation properties of smoothed particle hydrodynamics applied to the shallow-water equations *BIT* **43** 40–54

Fulk D A and Quinn D W 1996 An analysis of 1-D smoothed particle hydrodynamics kernels *J. Comput. Phys.* **126** 165–80

Gingold R A and Monaghan J J 1977 Smoothed particle hydrodynamics: theory and application to non-spherical stars *Mon. Not. R. Astron. Soc.* **181** 375–89

Gingold R A and Monaghan J J 1978 Binary fission in damped rotating polytropes *Mon. Not. R. Astron. Soc.* **184** 481–99

Gingold R A and Monaghan J J 1979 A numerical study of the Roche and Darwin problems for polytropic stars *Mon. Not. R. Astron. Soc.* **188** 45–58

Gingold R A and Monaghan J J 1980 The Roche problem for polytropes in central orbits *Mon. Not. R. Astron. Soc.* **191** 897–924

Gingold R A and Monaghan J J 1982 Kernel estimates as a basis for general particle methods in hydrodynamics *J. Comput. Phys.* **46** 429–53

Grady D E and Kipp M E 1987 Dynamic rock fragmentation *Fracture Mechanics of Rock* (New York: Academic) chapter 10, pp 429–73

Gray J, Monaghan J J and Swift R P 2001 SPH elastic dynamics *Comput. Methods Appl. Mech. Eng.* **190** 6641–62

Gray J A and Monaghan J J 2004 Numerical modelling of stress fields and fracture around magma chambers *Volcanology Geothermal Res.* **135** 259–83

Hernquist L and Katz N 1989 TREESPH—A unification of SPH with the hierarchical tree method *Astrophys. J.* (Suppl.) **70** 419–46

Hiscock W A and Lindblom L 1985 Generic instabilities in first order dissipative relativistic fluid theories *Phys. Rev.* **31** 725–33

Hockney R W and Eastwood J W 1988 *Computer Simulation Using Particles* (Bristol: Hilger)

Holm D 1991 Elliptical vortices and integrable Hamiltonian dynamics of the rotating shallow-water equations *J. Fluid Mech.* **227** 393–406

Holm D 1999 Fluctuation effects on 3D Lagrangian mean and Eulerian mean notion *Physica* D **133** 215–69

Hoover W G 1998 Isomorphism linking smooth particles and embedded atoms *Physica* A **260** 244–54

Hoover W G, Hoover C C and Merritt E C 2004 Smooth particle applied mechanics: conservation of angular momentum with tensile stability and velocity averaging *Phys. Rev.* E **69** 016702-1-10

Hu X Y and Adams N A 2005 A multi-phase SPH method for macroscopic and mesoscopic flows *J. Comput. Phys.* at press

Johnson G R and Beissel S R 1996 Normalized smoothing functions for impact computations *Int. J. Numer. Methods Eng.* **569** 501–18

Johnson G R, Stryk R A and Beissel S R 1996 SPH for high velocity impact computations *Comput. Methods Appl. Mech. Eng.* **139** 347–73

Kahan W H 1980 IEEE floating point standard ed K L E Nickel *Interval Mathematics* (New York: Academic)

Kitsionas S and Whitworth A P 2002 Smoothed particle hydrodynamics with particle splitting applied to self-gravitating flows *Mon. Not. R. Astron. Soc.* **330** 129–36

Laguna L D, Miller W A and Zureck W H 1993 Smooth particle hydrodynamics near a black hole *Astrophys. J* **404** 678–85

Lamb H 1932 *Hydrodynamics* (Cambridge: Cambridge University Press)

Landau L D and Lifshitz E M 1993 *Fluid Mechanics* 2nd edn (Oxford: Pergamon) p 512

Landrini M, Colagrossi L and Faltinsen O 2003 Sloshing in 2D flows by the SPH method *Proc. 8th Int. Conf. on Numerical Ship Hydrodynamics (Pusan, Korea)*

Lattanzio J C and Monaghan J J 1991 A simulation of the collapse and fragmentation of cooling molecular clouds *Mon. Not. R. Astron. Soc.* **375** 177–89

Lattanzio J C, Monaghan J J, Pongracic H and Schwarz M P 1985 Interstellar cloud collisions *Mon. Not. R. Astron. Soc.* **215** 125–47

Leimkuhler B J, Reich S and Skeel R D 1997 Integration methods for molecular dynamics *IMA Volume in Mathematics and its Applications* Vol 82 ed K Schulten and J Mesirov (Springer) pp 161–86

Libersky L and Petschek A G 1991 Smooth particle hydrodynamics with strength of materials *Advances in Free Lagrange Methods* ed H E Trease and M J Fritts (Springer)

Lucy L B 1977 A numerical approach to the testing of the fission hypothesis *Astron. J.* **82** 1013–24

Mann P J 1991 A relativistic smoothed particle hydrodynamics code tested with the shock tube *Comput. Phys. Commun.* **67** 245–60

Marri S and White S D M 2003 Smoothed particle hydrodynamics for galaxy-formation simulations: improved treatments of multiphase gas, of star formation and of supernovae feedback *Mon. Not. R. Astron. Soc.* **345** 561–74

Marti J M and Mueller E 2003 Numerical hydrodynamics in special relativity *Living Rev. (Electronic)* **7**

May D and Monaghan J J 2003 Can a single bubble sink a ship *Am. J. Phys.* **71** 842–9

Michel P, Benz W and Richardson D 2004 Catastrophic disruption of asteroids and family formation: a review of numerical simulations, including both fragmentation and gravitational reaccumulation *Planet. Space Sci.* **52** 1109–17

Mohseni K, Kosovic B, Shkoller S and Marsden J E 2003 Numerical simulations of the Lagrangian averaged Navier–Stokes equations for homogeneous isotropic turbulence *Phys. Fluids* A **15** 524–44

Monaghan J J 1985a Extrapolating B splines for interpolation *J. Comput. Phys.* **60** 253–62

Monaghan J J 1985b Particle methods for hydrodynamics *Comput. Phys. Rep.* **3** 71–124

Monaghan J J 1989 On the problem of penetration in particle methods *J. Comput. Phys.* **82** 1–15

Monaghan J J 1992 Smoothed particle hydrodynamics *Ann. Rev. Astron. Astrophys.* **30** 543–74

Monaghan J J 1994 Simulating free surface flows with SPH *J. Comput. Phys.* **110** 399–406

Monaghan J J 1997 SPH and Riemann solvers *J. Comput. Phys.* **136** 298–307

Monaghan J J 2000 SPH without a tensile instability *J. Comput. Phys.* **159** 290–311

Monaghan J J 2002 SPH compressible turbulence *Mon. Not. R. Astron. Soc.* **335** 843–52

Monaghan J J 2004 Energy distribution in a particle alpha model *J. Turbul.* **12** 1

Monaghan J J, Cas R F, Kos A and Hallworth M 1999 Gravity currents descending a ramp in a stratified tank *J. Fluid Mech.* **379** 36–9

Monaghan J J and Gingold R A 1983 Shock simulation by the particle method SPH *J. Comput. Phys.* **52** 374–89

Monaghan J J, Huppert H E and Worster M G 2005 Solidification using smoothed particle hydrodynamics *J. Comput. Phys.* **206** 684–705

Monaghan J J and Kocharyan A 1995 SPH simulation of multi phase flow *Comput. Phys. Commun.* **87** 225–35

Monaghan J J and Kos A 1999 Solitary waves on a Cretan beach *J. Waterways Port Coastal Ocean Eng.* **1111** 145–54

Monaghan J J and Kos A 2000 Scott Russell's wave generator *Phys. Fluids* A **12** 622–30

Monaghan J J, Kos A and Issa N 2004 Fluid motion generated by impact *J. Waterway Port Coastal Ocean Eng.* **129** 250–9

Monaghan J J and Price D J 2001 Variational principles for relativistic smoothed particle hydrodynamics *Mon. Not. R. Astron. Soc.* **328** 381–92

Monaghan J J and Price D J 2004 Toy stars in one dimension *Mon. Not. R. Astron. Soc.* **350** 1449–56

Morris J P 1996 Analysis of smoothed particle hydrodynamics with applications *PhD Thesis* Monash University, Melbourne, Australia

Morris J P, Fox P J and Zhu Y J 1997 Modelling low Reynolds number incompressible flows using SPH *J. Comput. Phys.* **136** 214–26

Morris J P and Monaghan J J 1997 A switch to reduce SPH viscosity *J. Comput. Phys.* **136** 41–50

Olson T S and Hiscock W A 1990 Stability, causality and hyperbolicity in Carter's 'regular' theory of relativistic heat conducting fluids *Phys. Rev.* D **41** 3687–95

Parshikov A N and Medin S A 2002 Smoothed particle hydrodynamics using interparticle contact algorithms *J. Comput. Phys.* **180** 358–82

Parzen E 1962 On estimations of a probability density and mode *Ann. Math. Stat.* **33** 1065–76

Phillips G J and Monaghan J J 1985 A numerical method for three-dimensional simulations of collapsing, isothermal, magnetic gas clouds *Mon. Not. R. Astron. Soc.* **216** 883–95

Price D J and Monaghan J J 2004a Smoothed particle magnetohydrodynamics: I. Algorithms and tests in one dimension *Mon. Not. R. Astron. Soc.* **348** 123–38

Price D J and Monaghan J J 2004b Smoothed particle magnetohydrodynamics: II. Variational principles and variable smoothing length terms *Mon. Not. R. Astron. Soc.* **348** 139–52

Price D J 2004a Magnetic fields in astrophysics *PhD Thesis* University of Cambridge, Cambridge, UK www.astro.ex.ac.uk/people/dprice

Price D J 2004b Private communication

Quirk J J 1994 A contribution to the great Riemann solver debate *Int. J. Numer. Methods Fluids* **18** 555–74

Randles P W and Libersky L 1996 Smoothed particle hydrodynamics some recent improvements and applications *Comput. Methods Appl. Mech. Eng.* **139** 375–408

Rosenblatt M 1956 Remarks on some nonparametric estimates of a density function *Ann. Math. Stat.* **27** 832–7

Rosswog S and Davies M B 2002 High resolution calculations of merging neutron stars-I. Model description and hydrodynamic evolution *Mon. Not. R. Astron. Soc.* **334** 481–97

Rosswog S, Davies M B, Thielemann F-K and Piran T 2000 Merging neutron stars: asymmetric systems *Astron. Astrophys.* **360** 171–84

Schoenberg I J 1946 Contributions to the problem of approximation of equidistant data by analytic functions: part A *Q. Appl. Math.* **IV** 45–99

Siegler S and Riffert H 2000 Smoothed particle hydrodynamics simulations of ultra-relativistic shocks with artificial viscosity *Astrophys. J.* **531** 1053–66

Sod G A 1978 A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws *J. Comput. Phys.* **27** 1–31

Springel V and Hernquist L 2002 Cosmological smoothed particle hydrodynamics simulations: the entropy equation *Mon. Not. R. Astron. Soc.* **333** 649–64

Steinmetz M and Mueller E 1993 On the capabilities and limits of smoothed particle hydrodynamics *Astron. Astrophys.* **268** 391–410

Swegle J, Hicks J and Attaway S 1995 Smoothed particle hydrodynamics stability analysis *J. Comput. Phys.* **116** 123–34

Takeda H, Miyama S and Sekiya M 1994 Numerical simulation of viscosity using SPH *Prog. Theor. Phys.* **92** 939–60

Von Neumann J 1944 Proposal and analysis of a new numerical method for the treatment of hydrodynamical shock problems *Von Neumann Collected Works* ed A Taub (Oxford: Pergamon)

Watkins S J, Bhattal A S, Francis N, Turner J A and Whitworth A P 1996 A new prescription for viscosity in smoothed particle hydrodynamics *Astron. Astrophys.* (Suppl.) **119** 177–87

Welton W 1998 Two-dimensional PDF/SPH simulations of compressible turbulent flows *J. Comput. Phys.* **139** 410–43

Welton W and Pope S 1997 PDF models of compressible turbulence using SPH *J. Comput. Phys.* **134** 150–68

Whitehouse S C and Bate M R 2004 Smoothed particle hydrodynamics with radiative transfer in the flux limited approximation *Mon. Not. R. Astron. Soc.* **353** 1078–94