Appendix A
Elementary Mathematical Concepts

Throughout this document, boldface symbols imply vectors or tensors (matrices in our analyses).

### A.1 Vector Products

For the inner product of two vectors (first-order tensors) $\mathbf{u}$ and $\mathbf{v}$ we have in three dimensions

$$\mathbf{u} \cdot \mathbf{v} = u_i v_j = u_1 v_1 + u_2 v_2 + u_3 v_3 = |\mathbf{u}| |\mathbf{v}| \cos \theta,$$

(A.1)

in Cartesian bases

where $|\mathbf{u}| = \sqrt{u_1^2 + u_2^2 + u_3^2}$ and where Einstein index summation notation is used.

Two vectors are said to be orthogonal if $\mathbf{u} \cdot \mathbf{v} = 0$. The cross (vector) product of two vectors is

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} e_1 & e_2 & e_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = |\mathbf{u}| |\mathbf{v}| \sin \theta \mathbf{n},$$

(A.2)

where $\mathbf{n}$ is the unit normal to the plane formed by the vectors $\mathbf{u}$ and $\mathbf{v}$. The triple product of three vectors is

$$\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = \begin{vmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = (\mathbf{w} \times \mathbf{u}) \cdot \mathbf{v}$$

(A.3)

This represents the volume of a parallelepiped formed by the three vectors.
A.2 Vector Calculus

We have the following elementary operations:

- The divergence of a vector (a contraction to a scalar) is defined by
  \[ \nabla \cdot \mathbf{u} = u_{i,i} \]  
  (A.4)
  whereas for a second-order tensor (a contraction to a vector):
  \[ \nabla \cdot \mathbf{A} \text{ has components of } A_{ij,j}. \]  
  (A.5)

- The gradient of a vector (a dilation to a second-order tensor) is:
  \[ \nabla \mathbf{u} \text{ has components of } u_{i,j}, \]  
  (A.6)
  whereas for a second-order tensor (a dilation to a third-order tensor):
  \[ \nabla \mathbf{A} \text{ has components of } A_{ij,k}. \]  
  (A.7)

- The gradient of a scalar (a dilation to a vector) is:
  \[ \nabla \phi \text{ has components of } \phi_{i}. \]  
  (A.8)

The scalar product of two second-order tensors, for example, the gradients of first-order vectors, is defined as

\[ \nabla \mathbf{v} : \nabla \mathbf{u} = \frac{\partial v_i}{\partial x_j} \frac{\partial u_i}{\partial x_j} \overset{\text{def}}{=} v_{i,j} u_{i,j} \quad i, j = 1, 2, 3, \]  
(A.9)

in Cartesian bases

where \( \partial u_i / \partial x_j, \partial v_i / \partial x_j \) are partial derivatives of \( u_i \) and \( v_i \), and where \( u_i, v_i \) are the Cartesian components of \( \mathbf{u} \) and \( \mathbf{v} \) and

\[ \nabla \mathbf{u} \cdot \mathbf{n} \text{ has components of } u_{i,j} n_j, \quad i, j = 1, 2, 3. \]  
(A.10)

in Cartesian bases

- The divergence theorem for vectors is

\[ \int_{\Omega} \nabla \cdot \mathbf{u} \, d\Omega = \int_{\partial \Omega} \mathbf{u} \cdot \mathbf{n} \, dA = \int_{\Omega} u_{i,j} \, d\Omega = \int_{\partial \Omega} u_{i} n_{i} \, dA \]  
(A.11)

and analogously for tensors

\[ \int_{\Omega} \nabla \cdot \mathbf{B} \, d\Omega = \int_{\partial \Omega} \mathbf{B} \cdot \mathbf{n} \, dA = \int_{\Omega} B_{ij,j} \, d\Omega = \int_{\partial \Omega} B_{ij} n_{j} \, dA, \]  
(A.12)

where \( \mathbf{n} \) is the outward normal to the bounding surface.

These standard operations arise throughout the analysis.
A.3 Interpretation of the Gradient of Functionals

The elementary concepts to follow are important for understanding iterative solvers. Consider a surface in space defined by

\[ \Pi(x_1, x_2, \ldots x_N) = C. \]  \hspace{1cm} (A.13)

Consider a unit vector \( b \), and the inner product, forming the directional derivative (the rate of change of \( \Pi \) in the direction of \( b \)):

\[ \nabla \Pi \cdot b = ||b|| ||\nabla \Pi|| \cos \gamma. \]  \hspace{1cm} (A.14)

When \( \gamma = 0 \), the directional derivative is maximized, in other words when \( b \) and \( \nabla \Pi \) are colinear. Since we can represent curves on the surface defined by \( \Pi = C \) by a position vector (\( t \) is a parameter)

\[ r = x_1(t)e_1 + x_2(t)e_2 + \ldots + x_N(t)e_N, \]  \hspace{1cm} (A.15)

the tangent is

\[ \frac{dr}{dt} = \frac{dx_1}{dt} e_1 + \frac{dx_2}{dt} e_2 + \ldots + \frac{dx_N}{dt} e_N. \]  \hspace{1cm} (A.16)

If we take

\[ \frac{d\Pi}{dt} = 0 = \nabla \Pi \cdot \frac{dr}{dt} = \frac{\partial \Pi}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial \Pi}{\partial x_2} \frac{dx_2}{dt} + \ldots + \frac{\partial \Pi}{\partial x_N} \frac{dx_N}{dt}, \]  \hspace{1cm} (A.17)

we immediately see that \( \nabla \Pi \) is normal to the surface and represents the direction of maximum change in the normal direction.

A.4 Matrix Manipulations

Throughout the next few definitions, we consider the matrix \([A]\). The matrix \([A]\) is said to be symmetric if \([A] = [A]^T\) and skew-symmetric if \([A] = -[A]^T\). A first-order contraction (inner product) of two matrices is defined by

\[ A \cdot B = [A][B] \] has components of \( A_{ij}B_{jk} = C_{ik} \) \hspace{1cm} (A.18)

where it is clear that the range of the inner index \( j \) must be the same for \([A]\) and \([B]\). The second-order inner product of two matrices is

\[ A : B = A_{ij}B_{ij} = tr([A]^T[B]) \] \hspace{1cm} (A.19)
The rule of transposes for the product of two matrices is
\[(A\,|\,B)^T = B^T A^T.\] (A.20)

The rule of inverses for two invertible \(n \times n\) matrices is
\[(A\,|\,B)^{-1} = B^{-1} A^{-1} \quad [A]^{-1}[A] = [A][A]^{-1} = [I] \] (A.21)
where \([I]\) is the identity matrix. Clearly, \([A]^{-1}\) exists only when \(\text{det}[A] \neq 0\).

**A.4.1 Determinant**

Some properties of the determinant (where \([A]\) is a \(3 \times 3\) matrix):

\[
[A] \overset{\text{def}}{=} \begin{bmatrix}
  A_{11} & A_{12} & A_{13} \\
  A_{21} & A_{22} & A_{23} \\
  A_{31} & A_{32} & A_{33}
\end{bmatrix}
\] (A.22)

are

\[
\text{det}[A] = A_{11}(A_{22}A_{33} - A_{32}A_{23}) - A_{12}(A_{21}A_{33} - A_{31}A_{23}) + A_{13}(A_{21}A_{32} - A_{31}A_{22}), \\
\text{det}[I] = 1, \quad \text{det} \alpha [A] = \alpha^3 \text{det}[A], \quad \alpha = \text{scalar}, \\
\text{det}[A] [B] = \text{det}[A] \text{det}[B], \quad \text{det}[A]^T = \text{det}[A], \quad \text{det}[A]^{-1} = \frac{1}{\text{det}[A]}.
\]

An important use of the determinant is in forming the inverse by

\[
[A]^{-1} = \frac{\text{adj}[A]}{\text{det}[A]}, \quad \text{adj}[A] \overset{\text{def}}{=} \begin{bmatrix}
  C_{11} & C_{12} & C_{13} \\
  C_{21} & C_{22} & C_{23} \\
  C_{31} & C_{32} & C_{33}
\end{bmatrix}^T, \quad (A.23)
\]

where the so-called cofactors are

\[
C_{11} = A_{22}A_{33} - A_{32}A_{23} \quad C_{12} = -(A_{21}A_{33} - A_{31}A_{23}) \\
C_{13} = A_{21}A_{32} - A_{31}A_{22} \quad C_{21} = -(A_{12}A_{33} - A_{32}A_{13}) \\
C_{22} = A_{11}A_{33} - A_{31}A_{13} \quad C_{23} = -(A_{11}A_{32} - A_{31}A_{12}) \\
C_{31} = A_{12}A_{23} - A_{22}A_{13} \quad C_{32} = -(A_{11}A_{23} - A_{21}A_{13}) \\
C_{33} = A_{11}A_{22} - A_{21}A_{12}
\] (A.24)
A.4.2 Eigenvalues

The mathematical definition of an eigenvalue, a scalar denoted $\Lambda$, and eigenvector, a vector denoted $\mathcal{E}$, of a matrix $[A]$ is

$$[A][\mathcal{E}] = \Lambda[\mathcal{E}] \quad (A.25)$$

Some main properties to remember about eigenvalues and eigenvectors are:

1. If $[A]$ $(n \times n)$ has $n$ linearly independent eigenvectors then it is diagonalizable by a matrix formed by columns of the eigenvectors. In the case of a $3 \times 3$ matrix,

$$
\begin{bmatrix}
A_1 & 0 & 0 \\
0 & A_2 & 0 \\
0 & 0 & A_3 \\
\end{bmatrix}
= 
\begin{bmatrix}
\mathcal{E}_1^{(1)} & \mathcal{E}_1^{(2)} & \mathcal{E}_1^{(3)} \\
\mathcal{E}_2^{(1)} & \mathcal{E}_2^{(2)} & \mathcal{E}_2^{(3)} \\
\mathcal{E}_3^{(1)} & \mathcal{E}_3^{(2)} & \mathcal{E}_3^{(3)} \\
\end{bmatrix}^{-1}
\begin{bmatrix}
A_{11} & A_{12} & A_{13} \\
A_{21} & A_{22} & A_{23} \\
A_{31} & A_{32} & A_{33} \\
\end{bmatrix}
\begin{bmatrix}
\mathcal{E}_1^{(1)} & \mathcal{E}_1^{(2)} & \mathcal{E}_1^{(3)} \\
\mathcal{E}_2^{(1)} & \mathcal{E}_2^{(2)} & \mathcal{E}_2^{(3)} \\
\mathcal{E}_3^{(1)} & \mathcal{E}_3^{(2)} & \mathcal{E}_3^{(3)} \\
\end{bmatrix}
\quad (A.26)
$$

2. If $[A]$ $(n \times n)$ has $n$ distinct eigenvalues then the eigenvectors are linearly independent.

3. If $[A]$ $(n \times n)$ is symmetric then its eigenvalues are real. If the eigenvalues are distinct, then the eigenvectors are orthogonal.

A quadratic form is defined as $\{x\}^T [A]\{x\}$, and is positive when $[A]$ has positive eigenvalues. Explicitly, for a $3 \times 3$ matrix, we have

$$\{x\}^T [A]\{x\} \text{ def } = 
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
\end{bmatrix}
\begin{bmatrix}
A_{11} & A_{12} & A_{13} \\
A_{21} & A_{22} & A_{23} \\
A_{31} & A_{32} & A_{33} \\
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
\end{bmatrix}. \quad (A.27)$$

A matrix $[A]$ is said to be positive definite if the quadratic form is positive for all nonzero vectors $x$. Clearly, from Eq. A.26 a positive definite matrix must have positive eigenvalues.

Remark: If we set the determinant $det[A - \Lambda \mathbf{1}] = 0$, it can be shown that the so-called characteristic polynomial is, for example for a $3 \times 3$ matrix:

$$
det (A - \Lambda \mathbf{1}) = -\Lambda^3 + I_A \Lambda^2 - II_A \Lambda + III_A = 0, \quad (A.28)
$$

where

$$
I_A = tr(A) = A_1 + A_2 + A_3 \\
II_A = \frac{1}{2} ((tr(A))^2 - tr(A^2)) = A_1 A_2 + A_2 A_3 + A_1 A_3 \\
III_A = det(A) = \frac{1}{6} ((tr(A))^3 + 2tr A^3 - 3(tr A^2) (tr A)) = A_1 A_2 A_3. \quad (A.29)
$$

Since $I_A$, $II_A$, and $III_A$ can be written in terms of $tr A$, which is invariant under frame rotation, they too are invariant under frame rotation.
A.4.3 Coordinate Transformations

To perform a coordinate transform for a $3 \times 3$ matrix $[A]$ from one Cartesian coordinate system to another (denoted with a $(\hat{\cdot})$), we apply a transformation matrix $[Q]$ (Fig. A.1):

$$[\hat{A}] = [Q][A][Q]^{-1} \quad (A.30)$$

In three dimensions, the standard axes rotators are, about the $x_1$ axis

$$Rot(x_1) \overset{\text{def}}{=} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta_1 & \sin \theta_1 \\ 0 & -\sin \theta_1 & \cos \theta_1 \end{bmatrix}, \quad (A.31)$$

about the $x_2$ axis

$$Rot(x_2) \overset{\text{def}}{=} \begin{bmatrix} \cos \theta_2 & 0 & -\sin \theta_2 \\ 0 & 1 & 0 \\ \sin \theta_2 & 0 & \cos \theta_2 \end{bmatrix}, \quad (A.32)$$

and about the $x_3$ axis

$$Rot(x_3) \overset{\text{def}}{=} \begin{bmatrix} \cos \theta_3 & \sin \theta_3 & 0 \\ -\sin \theta_3 & \cos \theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (A.33)$$

The standard axes reflectors are, with respect to the $x_2 - x_3$ plane

$$Ref(x_1) \overset{\text{def}}{=} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (A.34)$$

**Fig. A.1** Top: reflection with respect to the $x_2 - x_3$ plane. Bottom: rotation with respect to the $x_3$ axis.
with respect to the $x_1 - x_3$ plane

$$\text{Ref}(x_2) \overset{\text{def}}{=} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$  \hspace{1cm} (A.35)

with respect to the $x_1 - x_2$ plane

$$\text{Ref}(x_3) \overset{\text{def}}{=} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$  \hspace{1cm} (A.36)
In this chapter, we provide the reader with basic background information for field equations of interest.

### B.1 Deformations

The term deformation refers to a change in the shape of a continuum between a reference configuration and a current configuration. In the reference configuration, a representative particle of a continuum occupies a point \( P \) in space and has the position vector (Fig. B.1)

\[
X = X_1 e_1 + X_2 e_2 + X_3 e_3,
\]

(B.1)

where \( e_1, e_2, e_3 \) is a Cartesian reference triad, and \( X_1, X_2, X_3 \) (with center \( O \)) can be thought of as labels for a material point. Sometimes the coordinates or labels \( (X_1, X_2, X_3) \) are called the referential or material coordinates. In the current configuration, the particle originally located at point \( P \) (at time \( t = 0 \)) is located at point \( P' \) and can be also expressed in terms of another position vector \( x \), with coordinates \( (x_1, x_2, x_3) \). These are called the current coordinates. In this framework, the displacement is \( u = x - X \) for a point originally at \( X \) and with final coordinates \( x \).

When a continuum undergoes deformation (or flow), its points move along various paths in space. This motion may be expressed as a function of \( X \) and \( t \) as

\[
x(X, t) = u(X, t) + X(t),
\]

(B.2)

\footnote{Frequently, analysts consider the referential configuration to be fixed in time thus, \( X \neq X(t) \). We shall adopt this in the present work.}
which gives the present location of a point at time $t$, written in terms of the reference coordinates $X_1, X_2, X_3$. The previous position vector may be interpreted as a mapping of the initial configuration onto the current configuration. In classical approaches, it is assumed that such a mapping is one-to-one and continuous, with continuous partial derivatives to whatever order required. The description of motion or deformation expressed previously is known as the Lagrangian formulation. Alternatively, if the independent variables are the coordinates $x$ and time $t$, then $x(x_1, x_2, x_3, t) = u(x_1, x_2, x_3, t) + X(x_1, x_2, x_3, t)$, and the formulation is denoted as Eulerian (Fig. B.1).

Partial differentiation of the displacement vector $u = x - X$, with respect to $X$, produces the following displacement gradient:

$$\nabla_X u = F - 1,$$

where

$$F \stackrel{\text{def}}{=} \nabla_X x \overset{\text{def}}{=} \frac{\partial x}{\partial X} = \begin{bmatrix} \frac{\partial x_1}{\partial X_1} & \frac{\partial x_1}{\partial X_2} & \frac{\partial x_1}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{bmatrix}. \quad \text{(B.4)}$$

$F$ is known as the material deformation gradient.

Now, consider the length of a differential element in the reference configuration $dX$ and $dx$ in the current configuration, $dx = \nabla_X x \cdot dX = F \cdot dX$. Taking the difference in the squared magnitudes of these elements yields

$$dx \cdot dx - dX \cdot dX = (\nabla_X x \cdot dX) \cdot (\nabla_X x \cdot dX) - dX \cdot dX = dX \cdot (F^T \cdot F - 1) \cdot dX \overset{\text{def}}{=} 2 dX \cdot E \cdot dX. \quad \text{(B.5)}$$

Equation (B.5) defines the so-called Lagrangian strain tensor

$$E \overset{\text{def}}{=} \frac{1}{2} (F^T \cdot F - 1) = \frac{1}{2} [\nabla_X u + (\nabla_X u)^T + (\nabla_X u)^T \cdot \nabla_X u]. \quad \text{(B.6)}$$

**Fig. B.1** Different descriptions of a deforming body.
Remark: It should be clear that $dx$ can be reinterpreted as the result of a mapping $F \cdot dX \rightarrow dx$ or a change in configuration (reference to current). One may develop the so-called Eulerian formulations, employing the current configuration coordinates to generate Eulerian strain tensor measures. An important quantity is the Jacobian of the deformation gradient, $J \overset{\text{def}}{=} \det F$, which relates differential volumes in the reference configuration ($d\omega_0$) to differential volumes in the current configuration ($d\omega$) via $d\omega = J d\omega_0$. The Jacobian of the deformation gradient must remain positive, otherwise we obtain physically impossible “negative” volumes. For more details, we refer the reader to the texts of Malvern [1], Gurtin [2], Chandrasekharahia and Debnath [3].

### B.2 Equilibrium/Kinetics of Solid Continua

The balance of linear momentum in the deformed (current) configuration is

$$
\int_{\partial\omega} \boldsymbol{t} \, da + \int_{\omega} \rho \boldsymbol{b} \, d\omega = \frac{d}{dt} \int_{\omega} \rho \dot{\boldsymbol{u}} \, d\omega, \quad (B.7)
$$

where $\omega \subset \Omega$ is an arbitrary portion of the continuum, with boundary $\partial\omega$, $\rho$ is the material density, $\boldsymbol{b}$ is the body force per unit mass, and $\dot{\boldsymbol{u}}$ is the time derivative of the displacement. The force densities, $\boldsymbol{t}$, are commonly referred to as “surface forces” or tractions.

### B.2.1 Postulates on Volume and Surface Quantities

Now, consider a tetrahedron in equilibrium, as shown in Fig. B.2, where a balance of forces yields

$$
\begin{align*}
t^{(n)} A^{(n)} + t^{(-1)} A^{(1)} + t^{(-2)} A^{(2)} + t^{(-3)} A^{(3)} + \rho \boldsymbol{b} \Delta V &= \rho \Delta V \ddot{\boldsymbol{u}}, \quad (B.8)
\end{align*}
$$

where $\Delta A^{(n)}$ is the surface area of the face of the tetrahedron with normal $\boldsymbol{n}$, and $\Delta V$ is the tetrahedron volume. As the distance ($h$) between the tetrahedron base (located at (0,0,0)) and the surface center goes to zero ($h \rightarrow 0$), we have $\Delta A^{(n)} \rightarrow 0 \Rightarrow \frac{\Delta V}{\Delta A^{(n)}} \rightarrow 0$. Geometrically, we have $\frac{\Delta A^{(i)}}{\Delta A^{(n)}} = \cos(x_i, x_n) \overset{\text{def}}{=} n_i$, and therefore $t^{(n)} + t^{(-1)} \cos(x_1, x_n) + t^{(-2)} \cos(x_2, x_n) + t^{(-3)} \cos(x_3, x_n) = 0$. It is clear that forces on the surface areas could be decomposed into three linearly independent components. It is convenient to introduce the concept of stress at a point, representing the surface forces there, pictorially represented by a cube surrounding a point. The fundamental issue that must be resolved is the characterization of these surface forces.
We can represent the surface force density vector, the so-called traction, on a surface by the component representation:

\[ t^{(i)} \overset{\text{def}}{=} \begin{bmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \sigma_{i3} \end{bmatrix}, \quad \text{(B.9)} \]

where the second index represents the direction of the component and the first index represents components of the normal to corresponding coordinate plane. Henceforth, we will drop the superscript notation of \( t^{(n)} \), where it is implicit that \( t \overset{\text{def}}{=} t^{(n)} = \sigma^T \cdot n \), where

\[ \sigma \overset{\text{def}}{=} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}, \quad \text{(B.10)} \]

or explicitly \( (t^{(1)} = -t^{(-1)}, t^{(2)} = -t^{(-2)}, t^{(3)} = -t^{(-3)}) \)

\[ t = t^{(1)} n_1 + t^{(2)} n_2 + t^{(3)} n_3 = \sigma^T \cdot n = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}^T \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}, \quad \text{(B.11)} \]

where \( \sigma \) is the so-called Cauchy stress tensor.

Remark: In the absence of couple stresses, a balance of angular momentum implies a symmetry of stress, \( \sigma = \sigma^T \), and thus the difference in notations becomes immaterial. Explicitly, starting with an angular momentum balance, under the assumptions that no infinitesimal “micro-moments” or so-called couple-stresses exist, then it can be shown that the stress tensor must be symmetric, i.e.,

\[ \int_{\partial \omega} x \times t \, da + \int_{\omega} x \times \rho b \, d\omega = \frac{d}{dt} \int_{\omega} x \times \rho \dot{u} \, d\omega; \] that is, \( \sigma^T = \sigma \). It is somewhat easier to consider a differential element, such as in Fig. B.2, and to simply sum moments about the center. Doing this, one immediately obtains \( \sigma_{12} = \sigma_{21}, \sigma_{23} = \sigma_{32} \) and \( \sigma_{13} = \sigma_{31} \). Consequently, \( t = \sigma \cdot n = \sigma^T \cdot n \).
B.2.2 Balance Law Formulations

Substitution of Eq. B.11 into Eq. B.7 yields (\( \omega \subset \Omega \))

\[
\int_{\partial \omega} \sigma \cdot n \, da + \int_{\omega} \rho b \, d\omega = \frac{d}{dt} \int_{\omega} \rho \dot{u} \, d\omega. \tag{B.12}
\]

A relationship can be determined between the densities in the current and reference configurations,

\[
\int_{\omega} \rho d\omega = \int_{\omega_0} \rho J d\omega_0 = \int_{\omega_0} \rho_0 d\omega_0. \nonumber
\]

Therefore, the Jacobian can also be interpreted as the ratio of material densities at a point. Since the volume is arbitrary, we can assume that \( \rho J = \rho_0 \) holds at every point in the body. Therefore, we may write

\[
\frac{d}{dt} (\rho_0) = \frac{d}{dt} (\rho J) = 0, \quad \text{when the system is mass conservative over time.}
\]

This leads to writing the last term in Eq. B.12 as

\[
\frac{d}{dt} \int_{\omega} \rho \dot{u} \, d\omega = \int_{\omega_0} \frac{d}{dt} \rho J \dot{u} \, d\omega_0 + \int_{\omega_0} \rho \ddot{u} J \, d\omega_0 = \int_{\omega} \rho \ddot{u} \, d\omega.
\]

From Gauss’s divergence theorem, and an implicit assumption that \( \sigma \) is differentiable, we have

\[
\int_{\omega} (\nabla x \cdot \sigma + \rho b - \rho \ddot{u}) \, d\omega = 0. \tag{B.13}
\]

B.3 Referential Descriptions of Balance Laws and Nanson’s Formula

Although we will not consider finite deformation problems in this monograph, some important concepts will be useful later in the context of mapping from one configuration to the next. In many cases it is quite difficult to perform a stress analysis, for finite deformation solid mechanics problems, in the current configuration, primarily because it is unknown a priori. Therefore all quantities are usually transformed (“pulled”) back to the original coordinates, the referential frame. Therefore, it is preferable to think of a formulation in terms of the referential fixed coordinated \( X \), a so-called Lagrangian formulation. With this in mind there are two commonly used referential measures of stresses. We start by a purely mathematical result, leading to the so-called Nanson formula for transformation of surface elements. Consider the cross product of two differential line elements in a current configuration, \( dx^{(1)} \times dx^{(2)} = (F \cdot dX^{(1)}) \times (F \cdot dX^{(2)}) \). An important vector identity (see Chandriashakiah and Debnath [3]) for a tensor \( T \) and two first-order vectors \( a \) and \( b \) is

\[
(T \cdot a) \times (T \cdot b) = T^* \cdot (a \times b),
\]

where the \( T^* \) is the transpose of the adjoint defined by

\[
T^* \stackrel{\text{def}}{=} (det T)T^{-T}. \nonumber
\]

This leads to

\[
(det T)1 = T^T \cdot T^*. \nonumber
\]

Applying the result we have

\[
dx^{(1)} \times dx^{(2)} = F^* \cdot (dX^{(1)} \times dX^{(2)}) \quad \text{and} \quad F^T \cdot (dx^{(1)} \times dx^{(2)}) = (det F)1 \cdot (dX^{(1)} \times dX^{(2)}). \nonumber
\]

This leads to

\[
F^T \cdot nda = (det F)n_0 da_0. \nonumber
\]

This is the so-called Nanson formula. Knowing this, we now formulate the equations of equilibrium in the current or reference configuration (Fig. B.3).
Consider two surface elements: one on the current configuration and one on a reference configuration. If we form a new kind of stress tensor, call it $P$, such that the amount of force is the same we have $P \cdot n_0 da_0 = \sigma \cdot nda = \sigma \cdot F^{-T} (det F) \cdot n_0 da_0$ which implies $P = \sigma \cdot F^{-T} (det F)$. The tensor $P$ is called the first Piola–Kirchhoff stress and gives the actual force on the current area, but calculated per unit area of reference area. However, it is not symmetric, and this sometimes causes difficulties in an analysis. Therefore, we symmetrize it by $F^{-1} \cdot P = S = S^T = F^{-1} \cdot \sigma \cdot F^{-T} (det F)$. The tensor $S$ is called the second Piola–Kirchhoff stress. By definition we have $\int_{\partial \omega_0} n_0 \cdot P da_0 = \int_{\partial \omega} n \cdot \sigma da$, and thus

$$\int_{\partial \omega_0} n_0 \cdot P da_0 + \int_{\omega_0} f J d\omega_0 = \int_{\omega_0} \rho_0 \frac{d\dot{u}}{dt} d\omega_0, \quad (B.14)$$

and therefore

$$\int_{\omega_0} \nabla_X \cdot P d\omega_0 + \int_{\omega_0} f J d\omega_0 = \int_{\omega_0} \rho_0 \frac{d\dot{u}}{dt} d\omega_0. \quad (B.15)$$

Since $P = F \cdot S$, $\int_{\omega_0} \nabla_X \cdot (F \cdot S) d\omega_0 + \int_{\omega_0} f J d\omega_0 = \int_{\omega_0} \rho_0 \frac{d\dot{u}}{dt} d\omega_0$. Since the control volume is arbitrary, we have

$$\nabla_X \cdot P + f J = \rho_0 \frac{d\dot{u}}{dt} \quad \text{or} \quad \nabla_X \cdot (F \cdot S) + f J = \rho_0 \frac{d\dot{u}}{dt}. \quad (B.16)$$
B.4 The First Law of Thermodynamics/An Energy Balance

The interconversions of mechanical, thermal, and chemical energy in a system are governed by the first law of thermodynamics, which states that the time rate of change of the total energy, \( K + I \), is equal to the mechanical power, \( P \), and the net heat supplied, \( H + Q \), i.e.,

\[
\frac{d}{dt}(K + I) = P + H + Q.
\]

Here the kinetic energy of a subvolume of material contained in \( \Omega \), denoted \( \omega \), is

\[
K \overset{\text{def}}{=} \int_{\omega} \frac{1}{2} \rho \dot{u} \cdot \dot{u} \, d\omega;
\]

the power (rate of work) of the external forces acting on \( \omega \) is given by

\[
P \overset{\text{def}}{=} \int_{\omega} \rho \dot{b} \cdot \dot{u} \, d\omega + \int_{\partial \omega} \sigma \cdot n \cdot \dot{u} \, da;
\]

the heat flow into the volume by conduction is

\[
Q \overset{\text{def}}{=} -\int_{\partial \omega} q \cdot n \, da = -\int_{\omega} \nabla_x \cdot q \, d\omega,
\]

\( q \) being the heat flux; the heat generated due to sources, such as chemical reactions, is

\[
H \overset{\text{def}}{=} \int_{\omega} \rho z \, d\omega,
\]

\( z \) being the reaction source rate per unit mass; and the internal energy is

\[
I \overset{\text{def}}{=} \int_{\omega} \rho w \, d\omega,
\]

\( w \) being the internal energy per unit mass. Differentiating the kinetic energy yields

\[
\frac{dK}{dt} = \int_{\omega_0} \frac{1}{2} \rho \dot{u} \cdot \dot{u} \, d\omega_0 = \int_{\omega_0} \frac{1}{2} \left( \frac{d}{dt} \rho \right) \dot{u} \cdot \dot{u} \, d\omega_0 + \int_{\omega} \rho \frac{d}{dt} \frac{1}{2} \dot{u} \cdot \dot{u} \, d\omega = \int_{\omega} \rho \dot{u} \cdot \ddot{u} \, d\omega,
\]

(B.17)

where we have assumed that the mass in the system is constant. We also have

\[
\frac{dI}{dt} = \int_{\omega} \rho \dot{w} \, d\omega = \int_{\omega_0} \frac{1}{2} \rho J \dot{u} \cdot \dot{u} \, d\omega_0 = \int_{\omega_0} \frac{1}{2} \left( \frac{d}{dt} \rho J \right) \dot{u} \cdot \dot{u} \, d\omega_0 + \int_{\omega} \rho \frac{d}{dt} \frac{1}{2} \dot{u} \cdot \dot{u} \, d\omega = \int_{\omega} \rho \dot{w} \, d\omega = \int_{\omega} \rho \ddot{w} \, d\omega.
\]

(B.18)

By using the divergence theorem, we obtain

\[
\int_{\partial \omega} \sigma \cdot n \cdot \dot{u} \, da = \int_{\omega} \nabla_x \cdot (\sigma \cdot \dot{u}) \, d\omega = \int_{\omega} (\nabla_x \cdot \sigma) \cdot \dot{u} \, d\omega + \int_{\omega} \sigma : \nabla_x \dot{u} \, d\omega.
\]

(B.19)

Combining the results, and enforcing a balance of linear momentum, leads to

\[
\int_{\omega} \left( \rho \ddot{w} + \dot{u} \cdot (\rho \ddot{u} - \nabla_x \cdot \sigma - \rho \dot{b}) - \sigma : \nabla_x \dot{u} + \nabla_x \cdot q - \rho z \right) \, d\omega = 0.
\]

(B.20)

Since the volume \( \omega \) is arbitrary, the integrand must hold locally and we have

\[
\rho \dddot{w} - \sigma : \nabla_x \dot{u} + \nabla_x \cdot q - \rho z = 0.
\]

(B.21)

When dealing with multifield problems, this equation is used extensively.
B.5  Linearly Elastic Constitutive Equations

We now discuss relationships between the stress and strain, so-called material laws or constitutive relations for linearly elastic cases (infinitesimal deformations).

B.5.1 The Infinitesimal Strain Case

In infinitesimal deformation theory, the displacement gradient components are considered small enough that higher-order terms such as $(\nabla X u)^T \cdot \nabla X u$ and $(\nabla_x u)^T \cdot \nabla_x u$ can be neglected in the strain measure $E = \frac{1}{2} (\nabla X u + (\nabla X u)^T + (\nabla_x u)^T \cdot \nabla_x u)$, leading to $E \approx \epsilon = \frac{1}{2} [\nabla X u + (\nabla X u)^T]$. If the displacement gradients are small compared with unity, $\epsilon$ coincides closely with $E$. If we assume that $\frac{\partial}{\partial X} \approx \frac{\partial}{\partial x}$, we may use $E$ or $\epsilon$ interchangeably. Usually $\epsilon$ is the symbol used for infinitesimal strains. Furthermore, to avoid confusion, when using models employing the geometrically linear infinitesimal strain assumption we use the symbol of $\nabla$ with no $X$ or $x$ subscript. Hence, the infinitesimal strains are defined by

$$\epsilon = \frac{1}{2} (\nabla u + (\nabla u)^T). \quad (B.22)$$

B.5.2 Linear Elastic Constitutive Laws

If we neglect thermal effects, Eq. B.21 implies $\rho \dot{w} = \sigma : \nabla \dot{u}$ which, in the infinitesimal strain linearly elastic case, is $\rho \dot{w} = \sigma : \dot{\epsilon}$. From the chain rule of differentiation we have

$$\rho \dot{w} = \rho \frac{\partial w}{\partial \epsilon} \frac{d\epsilon}{dt} = \sigma : \dot{\epsilon} \Rightarrow \sigma = \rho \frac{\partial w}{\partial \epsilon}. \quad (B.23)$$

The starting point to develop a constitutive theory is to assume a stored elastic energy function exists, a function denoted $W \equiv \rho w$, which depends only on the mechanical deformation. The simplest function that fulfills $\sigma = \rho \frac{\partial w}{\partial \epsilon}$ is $W = \frac{1}{2} \epsilon : IE : \epsilon$, where $IE$ is the fourth rank elasticity tensor. Such a function satisfies the intuitive physical requirement that, for any small strain from an undeformed state, energy must be stored in the material. Alternatively, a small strain material law can be derived from $\sigma = \frac{\partial W}{\partial \epsilon}$ and $W \approx c_0 + c_1 : \epsilon + \frac{1}{2} \epsilon : IE : \epsilon + ...$ which implies $\sigma \approx c_1 + IE : \epsilon + ...$. We are free to set $c_0 = 0$ (it is arbitrary) in order to have zero strain energy at zero strain, and furthermore, we assume that no stresses exist in the reference state ($c_1 = 0$). With these assumptions, we obtain the familiar relation

$$\sigma = IE : \epsilon. \quad (B.24)$$

This is a linear relation between stresses and strains. The existence of a strictly positive stored energy function in the reference configuration implies that the linear
Appendix B: Basic Continuum Mechanics

elasticity tensor must have positive eigenvalues at every point in the body. Typically, different materials are classified according to the number of independent components in $\mathbf{IE}$. In theory, $\mathbf{IE}$ has 81 components, since it is a fourth-order tensor relating 9 components of stress to strain. However, the number of components can be reduced to 36 since the stress and strain tensors are symmetric. This is observed from the matrix representation of $\mathbf{IE}$:

$$
\begin{array}{cccccccc}
\epsilon_{11} & E_{1111} & E_{1122} & E_{1133} & E_{1112} & E_{1123} & E_{1113} \\
\epsilon_{22} & E_{2211} & E_{2222} & E_{2233} & E_{2212} & E_{2223} & E_{2213} \\
\epsilon_{33} & E_{3311} & E_{3322} & E_{3333} & E_{3312} & E_{3323} & E_{3313} \\
\epsilon_{12} & E_{1211} & E_{1222} & E_{1233} & E_{1212} & E_{1223} & E_{1213} \\
\epsilon_{23} & E_{2311} & E_{2322} & E_{2333} & E_{2312} & E_{2323} & E_{2313} \\
\epsilon_{31} & E_{3111} & E_{3122} & E_{3133} & E_{3112} & E_{3123} & E_{3113} \\
\end{array}
$$

The existence of a scalar energy function forces $\mathbf{IE}$ to be symmetric since the strains are symmetric, in other words $W = \frac{1}{2} \varepsilon : \mathbf{IE} : \varepsilon = \frac{1}{2} (\varepsilon : \mathbf{IE} : \varepsilon) = \frac{1}{2} \varepsilon : \mathbf{IE}^T : \varepsilon$, which implies $\mathbf{IE}^T = \mathbf{IE}$. Consequently, $\mathbf{IE}$ has only 21 independent components. The nonnegativity of $W$ imposes the restriction that $\mathbf{IE}$ remains positive definite. At this point, based on many factors that depend on the material microstructure, it can be shown that the components of $\mathbf{IE}$ may be written in terms of anywhere between 21 and 2 independent parameters. Accordingly, for isotropic materials, we have two planes of symmetry and an infinite number of planes of directional independence (two free components), yielding

$$
\mathbf{IE} \overset{\text{def}}{=} \begin{bmatrix}
\kappa + \frac{4}{3} \mu & \kappa - \frac{2}{3} \mu & -\frac{2}{3} \mu & 0 & 0 & 0 \\
\kappa - \frac{2}{3} \mu & \kappa + \frac{4}{3} \mu & -\frac{2}{3} \mu & 0 & 0 & 0 \\
\kappa - \frac{2}{3} \mu & -\frac{2}{3} \mu & \kappa + \frac{4}{3} \mu & 0 & 0 & 0 \\
0 & 0 & 0 & \mu & 0 & 0 \\
0 & 0 & 0 & 0 & \mu & 0 \\
0 & 0 & 0 & 0 & 0 & \mu \\
\end{bmatrix}
$$

In this case, we have

$$
\mathbf{IE} : \varepsilon = 3\kappa \frac{tr \varepsilon}{3} 1 + 2\mu \varepsilon' \Rightarrow \varepsilon : \mathbf{IE} : \varepsilon = 9\kappa \left( \frac{tr \varepsilon}{3} \right)^2 + 2\mu \varepsilon' : \varepsilon',
$$

where $tr \varepsilon = \varepsilon_{ij} i_j$ and $\varepsilon' = \varepsilon - \frac{1}{3} (tr \varepsilon) 1$ is the deviatoric strain. The eigenvalues of an isotropic elasticity tensor are $(3\kappa, 2\mu, 2\mu, \mu, \mu, \mu)$. Therefore, we must have $\kappa > 0$ and $\mu > 0$ to retain positive definiteness of $\mathbf{IE}$. All of the material components of $\mathbf{IE}$ may be spatially variable, as in the case of composite media.

---

2The symbol $\{\cdot\}$ is used to indicate the matrix notation equivalent to a tensor form, while $\{\cdot\}$ is used to indicate the vector representation.
B.5.3 Material Component Interpretation

There are a variety of ways to write isotropic constitutive laws, each time with a physically meaningful pair of material values.

Splitting the strain

It is sometimes important to split infinitesimal strains into two physically meaningful parts

$$\epsilon = \frac{tr \epsilon}{3} + (\epsilon - \frac{tr \epsilon}{3})_1. \quad (B.28)$$

An expansion of the Jacobian of the deformation gradient yields

$$J = \det (1 + \nabla X u) \approx 1 + tr \nabla X u + O(\nabla X u) = 1 + tr \epsilon + \ldots.$$ Therefore, with infinitesimal strains, \((1 + tr \epsilon)d \omega_0 = d \omega\) and we can write

$$tr \epsilon = \frac{d \omega - d \omega_0}{d \omega_0}.$$ Hence, \(tr \epsilon\) is associated with the volumetric part of the deformation. Furthermore, since \(\epsilon' \equiv \epsilon - \frac{tr \epsilon}{3} \mathbf{1}\), the so-called strain deviator describes distortion in the material.

Infinitesimal strain material laws

The stress \(\sigma\) can be split into two parts (dilatational and a deviatoric):

$$\sigma = \frac{tr \sigma}{3} \mathbf{1} + (\sigma - \frac{tr \sigma}{3} \mathbf{1}) \equiv -p \mathbf{1} + \sigma', \quad (B.29)$$

where we call the symbol \(p\) the hydrostatic pressure and \(\sigma'\) the stress deviator. With \((B.27)\) we write

$$p = -3 \kappa \left(\frac{tr \epsilon}{3}\right) \quad \text{and} \quad \sigma' = 2 \mu \epsilon'.$$ 

This is one form of Hooke’s Law. The resistance to change in the volume is measured by \(\kappa\). We note that \(\left(\frac{tr \sigma}{3} \mathbf{1}\right)' = 0\), which indicates that this part of the stress produces no distortion.

Another fundamental form of Hooke’s law is

$$\sigma = \frac{E}{1 + \nu} \left(\epsilon + \frac{\nu}{1 - 2\nu} (tr \epsilon) \mathbf{1}\right). \quad (B.31)$$

and the inverse form

$$\epsilon = \frac{1 + \nu}{E} \sigma - \frac{\nu}{E} (tr \sigma) \mathbf{1}. \quad (B.32)$$

To interpret the material values, consider an idealized uniaxial tension test (pulled in the \(x_1\) direction inducing a uniform stress state) where \(\sigma_{12} = \sigma_{13} = \sigma_{23} = 0\), which implies \(\epsilon_{12} = \epsilon_{13} = \epsilon_{23} = 0\). Also, we have \(\sigma_{22} = \sigma_{33} = 0\). Under these conditions we have \(\sigma_{11} = E \epsilon_{11}\) and \(\epsilon_{22} = \epsilon_{33} = -\nu \epsilon_{11}\). Therefore, \(E\), Young’s modulus, is the ratio of the uniaxial stress to the corresponding strain component. The Poisson ratio, \(\nu\), is the ratio of the transverse strains to the uniaxial strain.
Another commonly used set of stress–strain forms are the Lamé relations:

\[ \sigma = \lambda (tr \epsilon) \mathbf{1} + 2\mu \epsilon \quad \text{or} \quad \epsilon = -\frac{\lambda}{2\mu(3\lambda + 2\mu)} (tr \sigma) \mathbf{1} + \frac{\sigma}{2\mu}. \]  

(B.33)

To interpret the material values, consider a homogeneous pressure test (uniform stress), where \( \sigma_{12} = \sigma_{13} = \sigma_{23} = 0 \), and where \( \sigma_{11} = \sigma_{22} = \sigma_{33} \). Under these conditions, we have

\[
\kappa = \lambda + \frac{2}{3} \mu = \frac{E}{3(1-2\nu)} \quad \text{and} \quad \mu = \frac{E}{2(1+\nu)},
\]

(B.34)

and consequently

\[
\frac{\kappa}{\mu} = \frac{2(1+\nu)}{3(1-2\nu)}.
\]

(B.35)

We observe that \( \frac{\kappa}{\mu} \to \infty \) implies \( \nu \to \frac{1}{2} \), and \( \frac{\kappa}{\mu} \to 0 \) implies \( \Rightarrow \nu \to -1 \). Therefore, since both \( \kappa \) and \( \mu \) must be positive and finite, this implies \(-1 < \nu < 1/2\) and \(0 < E < \infty\). For example, some polymeric foams exhibit \( \nu < 0 \), steels \( \nu \approx 0.3 \), and some forms of rubber have \( \nu \to 1/2 \). We note that \( \lambda \) can be positive or negative. For more details, see Malvern [1], Gurtin [2], Chandrasekharaiah and Debnath [3].

### B.6 Related Physical Concepts

In closing, we briefly consider two other commonly encountered physical scenarios which are formally related to mechanical equilibrium.

#### B.6.1 Heat Conduction

We recall from our thermodynamic analysis the first law in the current configuration

\[ \rho \dot{w} - \sigma : \nabla \dot{u} + \nabla \cdot q - \rho z = 0, \]  

(B.36)

or in the reference configuration as

\[ \rho_0 \dot{w} - S : \dot{E} + \nabla \cdot q_0 - \rho_0 z = 0, \]

(B.37)

where \( q_0 = q J \cdot F^{-T} \). When (1) the deformations are ignored, \( u = 0 \), thus \( S : \dot{E} = 0 \), (2) the stored energy is purely thermal, described by \( \rho_0 \dot{w} = \rho_0 C \dot{\theta} \), where \( C \) is the heat capacity, (3) the reactions are zero, \( \rho_0 z = 0 \), (4) the variation in time is ignored, i.e., steady-state, and (5) \( q_0 = -k \nabla \theta \), where the conductivity tensor...
\( \mathbf{IK} \in \mathbb{IR}^{3 \times 3} \) is a spatially varying symmetric bounded positive definite tensor-valued function, then we arrive at the familiar equation of linear heat conduction:

\[
\nabla_X \cdot (\mathbf{IK} \cdot \nabla_X \theta) = \rho_0 C \dot{\theta}.
\]

(B.38)

If the variation in time is ignored, i.e., steady-state conditions are enforced:

\[
\nabla_X \cdot (\mathbf{IK} \cdot \nabla_X \theta) = 0.
\]

(B.39)

**B.6.2 Solid-State Diffusion-Reaction**

Consider a structure which occupies an open bounded domain in \( \Omega \in \mathbb{IR}^3 \), with boundary \( \partial \Omega \). The boundary consists of \( \Gamma_c \) and \( \Gamma_g \), where the solute concentrations \( (c) \) and solute fluxes are, respectively, specified. The diffusive properties of the heterogeneous material are characterized by a spatially varying diffusivity \( \mathbf{ID}_0 \in \mathbb{IR}^{3 \times 3} \), which is assumed to be a symmetric bounded positive definite tensor-valued function. The mass balance for a small diffusing species, denoted by the normalized concentration of the solute \( \bar{c} \) (molecules per unit volume), in an arbitrary subvolume of material contained within \( \Omega \), denoted \( \omega \), consists of a storage term \( (\dot{\bar{c}}) \), a reaction term \( \dot{s} \), and an inward normal flux term \( (-G \cdot n) \), leading to \( \int_\omega (\dot{\bar{c}} + \dot{s}) d\omega = -\int_{\partial\omega} G \cdot n \, da \). It is a classical *stoichiometrically inexact* approximation to assume that the diffusing species reacts (is created or destroyed) in a manner such that the rate of production of the reactant \( (s) \) is directly proportional to the concentration of the diffusing species itself and the rate of change of the diffusing species,

\[
\dot{\bar{c}} = \tau \bar{c} + \varpi \dot{\bar{c}}.
\]

Here, \( \tau = \tau_0 e^{-\frac{Q}{R \theta}} \) and \( \varpi = \varpi_0 e^{-\frac{Q}{R \theta}} \), where \( \tau_0 \) and \( \varpi_0 \) are rate constants, \( Q \) and \( Q \) \( (Q \neq \bar{Q}) \) are activation energies per mole of diffusive species, \( R \) is the universal gas constant, and \( \theta \) is the temperature. Upon substitution of these relations into the conservation law for the diffusing species, and after using the divergence theorem, since the volume \( \omega \) is arbitrary, one has a Fickian diffusion-reaction model in strong form, assuming \( G = -\mathbf{ID} \cdot \nabla_X c \)

\[
\dot{c} = \nabla_X \cdot (\mathbf{ID} \cdot \nabla_X c) - \tau c - \varpi \dot{c} \Rightarrow \dot{c}(1 + \varpi) = \nabla_X \cdot (\mathbf{ID} \cdot \nabla_X c) - \tau c. \quad (B.40)
\]

When \( \tau_0 > 0 \), the diffusing species is destroyed as it reacts, while \( \tau_0 < 0 \) means that the diffusing species is created as it reacts, i.e., an autocatalytic or “chain” reaction occurs. We will only consider the nonautocatalytic case in this work. Also, depending on the sign of \( \varpi_0 \), effectively the process will have an accelerated or decelerated diffusivity as well as accelerated or decelerated reactivity. In Eq. E.14, \( \mathbf{ID} \) is the diffusivity tensor (area per unit time). If we ignore reactions and time dependency, and assume that the domain is not deforming, we then arrive at the familiar

\[
\nabla_X \cdot (\mathbf{ID} \cdot \nabla_X c) = 0. \quad (B.41)
\]
B.6.3 Conservation Law Families

In summary we have the following related linearized steady-state forms (with no body forces in mechanical equilibrium)

\[
\begin{align*}
\nabla_X \cdot (IE : \nabla_X u) &= 0, \\
\nabla_X \cdot (IK \cdot \nabla_X \theta) &= 0, \\
\nabla_X \cdot (ID \cdot \nabla_X c) &= 0,
\end{align*}
\]

which stem from the following coupled, time-transient, nonlinear equations:

\[
\begin{align*}
\nabla_x \cdot \sigma + \rho b &= \rho \ddot{u} \\
\nabla_x \cdot q - \sigma : \nabla_x \dot{u} - \rho z &= -\rho \dot{w}, \\
\nabla_x \cdot G + \tau c + \omega \dot{c} &= -\dot{c},
\end{align*}
\]

From this point forth, we consider infinitesimal deformations, and we drop the explicit reference to differentiation with respect to $X$ or $x$, under the assumption that they are one and the same at infinitesimal deformations

\[
\begin{align*}
\nabla \cdot (IE : \nabla u) &= 0, \\
\nabla \cdot (IK \cdot \nabla \theta) &= 0, \\
\nabla \cdot (ID \cdot \nabla c) &= 0.
\end{align*}
\]

Furthermore, we shall use the notation $x$ to indicate the location of a point in space, under the assumption of infinitesimal deformations, where the difference between $X$ and $x$ is considered insignificant.
Appendix C
Convergence of Recursive Iterative Schemes

Recursive iterative schemes arise frequently in computational mechanics, for example, in implicit time-stepping, domain decomposition, etc. To understand the convergence of such iterative schemes, consider a general system of coupled equations given by

\[ \mathcal{A}(s) = \mathcal{F}, \quad (C.1) \]

where \( s \) is a solution, and where it is assumed that the operator, \( \mathcal{A} \), is invertible. One desires that the sequence of iterated solutions, \( s^I, I = 1, 2, \ldots \), converge to \( \mathcal{A}^{-1}(\mathcal{F}) \) as \( I \to \infty \). It is assumed that the \( I \)th iterate can be represented by some arbitrary function \( s^I = T^I(\mathcal{A}, \mathcal{F}, s^{I-1}) \). One makes the following split

\[ s^I = G^I(s^{I-1}) + r^I. \quad (C.2) \]

For this method to be useful the exact solution should be reproduced. In other words, when \( s = \mathcal{A}^{-1}(\mathcal{F}) \), then

\[ s = \mathcal{A}^{-1}(\mathcal{F}) = G^I(\mathcal{A}^{-1}(\mathcal{F})) + r^I. \quad (C.3) \]

Therefore, one has the following consistency condition

\[ r^I = \mathcal{A}^{-1}(\mathcal{F}) - G^I(\mathcal{A}^{-1}(\mathcal{F})), \quad (C.4) \]

and as a consequence,

\[ s^I = G^I(s^{I-1}) + \mathcal{A}^{-1}(\mathcal{F}) - G^I(\mathcal{A}^{-1}(\mathcal{F})). \quad (C.5) \]

Convergence of the iteration can be studied by defining the error vector:

\[ e^I = s^I - s = s^I - \mathcal{A}^{-1}(\mathcal{F}) = G^I(s^{I-1}) + \mathcal{A}^{-1}(\mathcal{F}) - G^I(\mathcal{A}^{-1}(\mathcal{F})) - \mathcal{A}^{-1}(\mathcal{F}) = G^I(s^{I-1}) - G^I(\mathcal{A}^{-1}(\mathcal{F})). \quad (C.6) \]
One sees that, if $G^I$ is linear and invertible, the above reduces to

$$e^I = G^I (s^{I-1} - A^{-1}(F)) = G^I (e^{I-1}). \quad (C.7)$$

Therefore, if the spectral radius of $G^I$, i.e., the magnitude of its largest eigenvalue, is less than unity for each iteration $I$, then $e^I \to 0$ for any arbitrary starting solution $s^{I=0}$ as $I \to \infty$. 
D.1 Sample Problem 1

• (a) Concisely explain the classical Galerkin method (without weak form) by considering a simple one-dimensional differential equation, written in the following abstract form

\[ A(u) = f, \]  

(D.1)

where \( u(0) = c_1 \) and \( u(L) = c_2 \).

• (b) Analytically solve the following two-point boundary value problem:

\[ \frac{d^2 u}{dx^2} + bu + \sin(ax) = 0 \quad (BC's: u(0) = 0, u(L) = 0), \]  

(D.2)

with domain size \((0, L)\) and where \(a > 0\) and \(b > 0\) are constants.

• (c) Using Galerkin’s method (no weak form), with the approximation

\[ u(x) \approx u^{app}(x) = a_1(x - xL) + a_2(x - xL)^2 + a_3\sin(6\pi x/L) + a_4\sin(8\pi x/L) \]  

(D.3)

generate the system of equations needed to determine \(a_1\) through \(a_5\) to approximately solve Eq. D.2. Set up the integrals. **Put in matrix form, but do not solve.**

• (d) Give expressions for the errors in the \(H_1\), \(H_2\), and \(H_3\) norms, respectively? Which error measure is higher and why?

• (e) List two major difficulties encountered when using the classical Galerkin method for more complex problems, which led us eventually to the finite element method.
(f) Repeat for
\[
d^2u \over dx^2 + au = e^{bx}
\]  \quad (BC's : u(0) = 0, u(L) = 0), \quad (D.4)

and
\[
u(x) \approx u^{app}(x) = a_1(x^4 - xL^3) + a_2(x^3 - xL^2) + a_3(x^2 - xL) + a_4(x - L),
\]  \quad (D.5)

\section*{D.2 Sample Problem 2}

If you were given the following
\[
E {d^2u \over dx^2} + ku = 0,
\]  \quad (D.6)
\[u(0) = c_1, \quad u(L) = c_2,
\]  
what is a quick way to determine (approximately) the minimum elements would one need to capture the basic physics if \(E\) and \(k\) where positive constants.

\section*{D.3 Sample Problem 3}

Consider the boundary value problem (Fig. D.1), with domain \(\Omega = (0, L)\):

\[
\frac{d}{dx} \left( E \frac{du}{dx} \right) + k_1 \frac{du}{dx} + k_2 u = \cos(14\pi x / L)
\]
\[
E \frac{du}{dx} (x = 0) = c_1,
\]
\[
u(x = L) = c_2,
\]  \quad (D.7)

where \(E > 0, k_1, \) and \(k_2\) are constants.

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{Fig.D.1.png}
\caption{A 1D structure}
\end{figure}
• (a) Derive the weak form (step-by-step, without the penalty approach).
• (b) If \( N \) equally sized linear elements were used, set up the matrix system of equations that one would need to solve. Draw the shape functions on the master element. Explicitly determine the element by element contributions (leave in integral form). In other words, derive the linear algebraic system
\[
[K](a) = \{R\}. \tag{D.8}
\]
Explicitly write \([K]^e\), \(\{a\}\) and \(\{R\}^e\). Use isoparametric mappings and make all calculations over the master element. Do not evaluate the integrals.
• (c) If \( N \) equally sized quadratic elements were used, set up the matrix system of equations that one would need to solve. Draw the shape functions on the master element. Explicitly determine the element-by-element contributions (leave in integral form). In other words, derive the linear algebraic system
\[
[K](a) = \{R\}. \tag{D.9}
\]
Explicitly write \([K]\), \(\{a\}\), and \(\{R\}\). Use isoparametric mappings and make all calculations over the master element. Do not evaluate the integrals.
• (d) If \( N \) equally sized cubic elements were used, set up the matrix system of equations that one would need to solve. Draw the shape functions on the master element. Explicitly determine the element by element contributions (leave in integral form). In other words, derive the linear algebraic system
\[
[K](a) = \{R\}. \tag{D.10}
\]
Explicitly write \([K]\), \(\{a\}\), and \(\{R\}\). Use isoparametric mappings and make all calculations over the master element. Do not evaluate the integrals.
• (e) How many Gauss points would be needed to evaluate each of the needed integrals using linear element shape functions? How many Gauss points would be needed to evaluate each of the needed integrals using quadratic element shape functions? How many Gauss points would be needed to evaluate each of the needed integrals, using cubic element shape functions?
• (f) Using \( N \) linear elements, if one used a direct (Gaussian solver) how much would it “cost” to solve? Under what conditions could you use an element-by-element Conjugate Gradient solver? How much would it cost? Explain any expressions that you write down.
• (g) Using \( N \) quadratic elements, if one used a direct (Gaussian solver) how much would it “cost” to solve? Under what conditions could you use an element-by-element Conjugate Gradient solver? How much would it cost? Explain any expressions that you write down.
• (h) Using \( N \) cubic elements, if one used a direct (Gaussian solver) how much would it “cost” to solve? Under what conditions could you use an element-by-element Conjugate Gradient solver? How much would it cost? Explain any expressions that you write down.
• (i) Assuming that you know the true solution, $u$, give an expression for the error in the $H^1(0, L)$ norm.
• (j) For the previous part of this problem, repeat parts (a, b, c, d) with the penalty method to apply the specified (primal/Dirichlet) boundary conditions. It is adequate to simply show the modifications.

### D.4 Sample Problem 4

Given the following integral

$$\int_{4}^{14} \left( (x - 11)^5 + (x - 1)^2 + 12 \right) dx, \quad (D.11)$$

using Gaussian quadrature, and assuming that the standard Gauss point weights and locations are given to you, show how you would evaluate it. Indicate how many quadrature points you will need to integrate it exactly.

### D.5 Sample Problem 5

Given the following integral

$$I = \int_{\Omega} (ax^3 + by^5) d\Omega, \quad (D.12)$$

where $\Omega$ is shown above and using a isoparametric mapping for a quadratic element (Fig. D.2).

![Fig. D.2 A quadratic element](image-url)
• (a) Derive the quadratic shape functions for the standard 2D (9-node) square master element.
• (b) For the following 2D element, derive the isoparametric mapping for this element shown in the figure.
• (c) Calculate the deformation gradient matrix $\mathbf{F}$ for the given mapping and the determinant (Jacobian $J = \det \mathbf{F}$).
• (d) Set up the integral in the master domain.
• (e) Set up the approximation of this integral with Gaussian quadrature.
• (f) What is the minimum number of Gauss points that are needed in each direction?

**D.6 Sample Problem 6**

• (a) Carefully, derive the best approximation theorem for $B(u, v) = L(v)$ and support the claim that the FEM solution is “the best possible” (specify in which norm). Include a detailed diagram of the spaces of approximations for $u$, $v$, $uh$, and $vh$.
• (b) What is the (true) potential $\mathcal{J}(u)$ for that problem?
• (c) Explain why the potential is important to monitor in the finite element method for this class of problems.
• (d) How could one could use the potential to determine the constant $C$ in the error estimation expression:

$$||u - uh||_{E(Ω)} \leq Ch \quad \text{(D.13)}$$

by using two successively finer meshes. Be very explicit.
• (e) Given

$$\frac{d}{dx} \left( E \left( \frac{du}{dx} \right) \right) + Ku = f, \quad \text{(D.14)}$$

and boundary conditions $u(0) = c_1$, $\frac{du}{dx}|_{x=L} = c_2$. Derive the weak form:

$$B(u, v) = L(v), \quad \text{(D.15)}$$

and the potential $\mathcal{J}(u)$.
• (f) For this system, under what conditions will the operator $\sqrt{B(u, u)}$ violate being a norm?

**D.7 Sample Problem 7**

Consider the following $(N \times N)$ matrix equation:

$$[K][a] = [R]. \quad \text{(D.16)}$$
• (a) When can one apply the Conjugate Gradient Method to obtain a solution?
• (b) Explicitly, write down what we are trying to minimize to obtain a solution.
• (c) Derive the ingredients needed for the Conjugate Gradient Method.
• (d) If $[K]$ happened to come from a one-dimensional finite element discretization using linear elements, how much cheaper would it be to solve using a CG element by element solver than a direct Gaussian approach that does not exploit the element-by-element structure.
• (e) Define the condition number of $[K]$. How does your answer in (d) depend on the condition number?
• (f) Given the following matrix:

$$ [K] = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 5 & 0 \\ 1 & 0 & 6 \end{bmatrix}, \quad (D.17) $$

what is the condition number?
• (g) Given the previous matrix and $R = (1, 1, 1)^T$ and $a^i=0 = (2, 2, 2)$, perform two complete CG iterations.
• (h) Now use a diagonal preconditioner (like in class), and repeat part (g).
• (i) What is the condition number of the new system?
• Repeat for:

$$ [K] = \begin{bmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad (D.18) $$

• Repeat for:

$$ [K] = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & 0 \\ 0 & 0 & 8 \end{bmatrix}, \quad (D.19) $$

• Repeat for:

$$ [K] = \begin{bmatrix} 5 & 0 & 3 \\ 0 & 4 & 0 \\ 3 & 0 & 2 \end{bmatrix}, \quad (D.20) $$

• Repeat for:

$$ [K] = \begin{bmatrix} 3 & 0 & 5 \\ 0 & 4 & 0 \\ 5 & 0 & 10 \end{bmatrix}, \quad (D.21) $$

D.8 Sample Problem 8

- (a) Explain the concept of an isoparametric map in 1D, 2D, and 3D and indicate the mathematical condition one must avoid to have “good” elements.
- (b) Using an isoparametric map, construct the mapping function for each of the 2D elements on in Fig. D.3.
- (c) What, if anything, is wrong with the following elements? Explicitly show what you mean mathematically.
- (d) Determine the Jacobian of the mapping for each element.
- (e) Repeat for the 2D elements in Fig. D.4.
D.9 Sample Problem 9

- (a) Give the proper connectivity functions for the mesh in Fig. D.5. Use the standard counterclockwise for the local numbering on the master element (Tables D.1, D.2, and D.3).
- (b) Using the standard counterclockwise local numbering for the master element, is there anything wrong with the following connectivity? If so, what problem will occur?

Fig. D.5 Hypothetical mapping

Table D.1 Above: local/global numbers for elements (YOU HAVE 4 ELEMENTS)

<table>
<thead>
<tr>
<th>Local number</th>
<th>EL # 1-GLOB</th>
<th>EL # 2-GLOB</th>
<th>EL # 3-GLOB</th>
<th>EL # 4-GLOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

Table D.2 Above: local/global numbers for elements for part (a)

<table>
<thead>
<tr>
<th>Local</th>
<th>EL # 1-GLOB</th>
<th>EL # 2-GLOB</th>
<th>EL # 3-GLOB</th>
<th>EL # 4-GLOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table D.3  Above: local/global numbers for elements for part (b)

<table>
<thead>
<tr>
<th>Local</th>
<th>EL # 1-GLOB</th>
<th>EL # 2-GLOB</th>
<th>EL # 3-GLOB</th>
<th>EL # 4-GLOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>

Fig. D.6  First hypothetical mapping

D.10  Sample Problem 10

- (a) (5 points) Derive the linear shape (basis) functions for the standard 2D square master element (Fig. D.6).
- (b) (5 points) For the following 2D element (Fig. D.7), derive the isoparametric map.
- (c) (5 points) What, if anything, is wrong with the following elements? Explicitly show what you mean mathematically. In the second element nodes 1 and 4 coincide.

D.11  Sample Problem 11

- (a) Derive the quadratic shape functions for the standard 2D (9-node) square master element in Fig. D.8.
- (b) For the following 2D element, derive the isoparametric mapping for this element shown in the figure.
- (c) Using the isoparametric mapping, what is the Jacobian of the mapping for this element?
Fig. D.7 Second hypothetical mapping

Fig. D.8 Quadratic 2D element

D.12 Sample Problem 12

With the standard isoparametric mapping for a trilinear cube element:

\[ x(\zeta_1, \zeta_2, \zeta_3) = \sum_{i=1}^{8} X_i \phi_i(\zeta_1, \zeta_2, \zeta_3) \]  \hspace{1cm} (D.22)
and
\[ y(\zeta_1, \zeta_2, \zeta_3) = \sum_{i=1}^{8} Y_i \phi_i(\zeta_1, \zeta_2, \zeta_3) \]  
(D.23)

and
\[ z(\zeta_1, \zeta_2, \zeta_3) = \sum_{i=1}^{8} Z_i \phi_i(\zeta_1, \zeta_2, \zeta_3) \]  
(D.24)

where \( \phi_i \) are the standard shape functions, e.g., \( \phi_i = \frac{1}{8} (1 \pm \zeta_1)(1 \pm \zeta_2)(1 \pm \zeta_3) \) and \( X_i, Y_i, Z_i \) are the nodal positions. Describe in detail how to solve for \( \zeta_1, \zeta_2, \zeta_3 \) with Newton’s method.

---

**D.13 Sample Problem 13**

Consider an ODE given by
\[ \dot{u} + 4e^{u+1} = \sin(\omega t) \]  
(D.25)

where \( C > 0 \) and \( \omega \) are a constant and \( u(0) = k \).

- (a) Set up the solution process using an explicit time-stepping scheme.
- (b) Set up the solution process using an implicit time-stepping scheme and Newton’s method.
- (c) What are the general differences (pros/cons) between explicit and implicit methods. Support claims with examples.
- (d) Repeat for:
\[ \dot{u} + 5(u + 1)^{11} = \cos(\omega t) \]  
(D.26)

where \( C > 0 \) and \( \omega \) are a constant and \( u(0) = k \).
- (e) Repeat for:
\[ \dot{u} - C(u + 5)^5 = 0 \]  
(D.27)

where \( C > 0 \) is a constant and \( u(0) = k \).
- (f) Repeat for:
\[ \dot{u} - C(cosu + 1)^4 = 0 \]  
(D.28)

where \( C > 0 \) is a constant and \( u(0) = k \).
D.14 Sample Problem 14

You are given an \((L \times L \times L)\) block (Fig. D.9) with scalar diffusivity \(D(x, y, z) > 0\), where the concentration is governed by

\[
\dot{c} = \nabla \cdot (D \nabla c) + f - \tau c,  \tag{D.29}
\]

where \(f = f(x, y, z)\) is given data (sources), reaction coefficient \(\tau = \tau(x, y, z)\) and with the initial condition \(c(t = 0, x, y, z) = c_0(x, y, z)\). It is externally flux loaded on a portion of its surface \(\Gamma_g\)

\[
(D \nabla c) \cdot n = g, \tag{D.30}
\]

where \(g\) is given and \(n\) is the outward surface normal, while it has as specified concentration on a portion of its surface \(\Gamma_c\)

\[
c = c_o. \tag{D.31}
\]

Note: The union of \(\Gamma_c\) and \(\Gamma_g\) comprises the entire boundary of the body.

- (a) Now develop a weak form for the problem, providing all the steps and assumptions necessary. Derive it with and without the penalty method. Explain the differences.
- (b) Use a finite difference approximation in time, and give the modified weak form for the implicit method with and without the penalty method. Explain the differences.
- (c) Substitute the basis functions into the weak form, and use the penalty method to apply concentration boundary conditions. Be very explicit in indicating what integral terms contribute to which “matrix” and “vector” terms in the system that you would eventually have to solve (Just set it up) with and without the penalty method.

Fig. D.9 A block experiencing diffusion
• (d) Now use a finite difference approximation in time, and give the modified weak form for the explicit method.
• (e) Substitute the basis functions into the weak form, and use the penalty method to apply concentration boundary conditions. Be very explicit in indicating what integral terms contribute to which “matrix” and “vector” terms in the system that you would eventually have to solve (Just set it up) with and without the penalty method.

D.15 Sample Problem 15

Consider the standard heat conduction problem in two dimensions (Fig. D.10):

\[ \nabla \cdot (K \cdot \nabla T) + f = 0 \quad (D.32) \]

that you are to solve on the following two-dimensional mesh:

The Dirichlet boundary condition has a temperature \( T = T^* \), and the Neumann (flux) conditions are applied to every other surface \( g = (K \cdot \nabla T) \cdot n \).

• (a) Derive the weak form of the problem, imposing the Dirichlet boundary conditions directly. Be sure to provide details on the spaces that the functions live in.
• (b) Sketch the matrix system to be solved, showing how the boundary conditions are applied. Just put X’s to denote nonzero entries.
• (c) Derive the weak form of the problem, imposing the Dirichlet boundary conditions using the penalty method. Be sure to detail from the spaces that the functions live in. You only need to state the additional terms added to the form from part (a).
• (d) Sketch the matrix system to be solved, showing how the boundary conditions are applied. State and illustrate the changes caused by the penalty terms.

Fig. D.10 An arch experiencing heat transfer
(e) Suppose you are using the Conjugate Gradient Method in both cases. How does solving the matrix system change when using the penalty method? State two differences.

D.16 Sample Problem 16

Consider the following linear elasticity problem, called “problem 1”:

\[
\text{problem #1} : \nabla \cdot (IE : \nabla u) + f = 0, \quad (D.33)
\]

where \( f \) is a given source term and the following heat conduction problem, called “problem 2”:

\[
\text{problem #2} : \nabla \cdot (IK \cdot \nabla T) + w = 0 \quad (D.34)
\]

where \( w \) is a given source term.

For each problem, consider the three-dimensional block where we use \textbf{linear} brick elements resulting in \( 3N-1 \) elements in the \( x \)-direction, \( 3N-1 \) elements in the \( y \)-direction, and \( 3N-1 \) elements in the \( z \)-direction, where \( N \) is given.

(a) How much does it cost if we use a regular Gaussian solver for problem 1 and problem 2?
(b) How much does it cost if we use a regular CG solver for problem 1 and problem 2?
(c) How much does it cost if we use a element-by-element CG solver for problem 1 and problem 2?

Now consider the same problem with the same finite element mesh, but now the mesh is broken up using a \( 5 \times 5 \times 5 \) subdomain decompositions and distributed across 125 processors:

(d) Draw a picture of the system and the decomposition.
(e) How much does it cost if we use the domain decomposition process and a regular Gaussian solver in each subdomain for problem 1 and problem 2?

D.17 Sample Problem 17

Given:

\[
\frac{d}{dx} \left( A \left( \frac{du}{dx} \right)^n \right) + u^q + f = 0, \quad (D.35)
\]

\[ u(0) = c_1, \quad \frac{du}{dx} \big|_{x=L} = c_2 \]
• (a) Derive the weak form.
• (b) Derive linearized weak form for the system in part (a) around $u = u_0$.
• (c) If one were to attempt to apply the Conjugate Gradient Method to solve the linearized system, for what combination of $A, u_0, p, and q$ will the CG method fail?
Appendix E
Selected Computer Projects

The projects in this chapter are selected from exams given over the last 15 years UC Berkeley.

E.1 Assignment Format

- All assignments must be typed—nothing handwritten.
- Be concise—shorter is better—provided you do not delete essential information.
- You are encouraged to talk and work with one another. Please see me for any problems, theoretical, coding etc.
- Introduction to the problem: explain it to a layman.
- Objectives: what are the goals?
- Your procedure: Brief explanation, flow-charts, difficulties, assumptions, etc.
- Findings: figures, plots and tables. Make sure they are readable.
- Observations and discussion: some interpretation and insight into the results.
- Appendix: the messy stuff like your code or raw data.

E.2 Sample Project 1: The Basics of FEM

- Solve the following boundary value problem, with domain \( \Omega = (0, L) \), analytically:
\[
\frac{d}{ds} \left( E \frac{du}{ds} \right) = k^2 \sin(\frac{2\pi k s}{L})
\]

\( E = \text{given constant} = 0.1 \)

\( k = \text{given constant} \)

\( L = 1 \)

\( u(0) = \Delta_1 = \text{given constant} = 0 \)

\( u(L) = \Delta_2 = \text{given constant} = 1 \)

(E.1)

- Now solve this with the finite element method using linear equal-sized elements. In order to achieve,

\[
e^N \overset{\text{def}}{=} \frac{||u-u^N||_{E(\Omega)}}{||u||_{E(\Omega)}} \leq TOL = 0.05,
\]

(E.2)

\[
||u||_{E(\Omega)} \overset{\text{def}}{=} \sqrt{\int_{\Omega} \frac{du}{dx} E \frac{du}{dx} dx}
\]

How many finite elements \((N)\) are needed for

\[
\begin{array}{c}
k = 1 \Rightarrow N = ? \\
k = 2 \Rightarrow N = ? \\
k = 4 \Rightarrow N = ? \\
k = 8 \Rightarrow N = ? \\
k = 16 \Rightarrow N = ? \\
k = 32 \Rightarrow N = ?
\end{array}
\]

(E.3)

You should set up a general matrix equation and solve it using Gaussian elimination. Later we will use other types of more efficient solvers. Plot the numerical solutions for \( N = 2, 4, 8, 16, \ldots \), for each \( k \), along with the exact solution. Also make a plot of the \( e^N \) for each \( k \).

**Remarks:** You should write a general one-dimensional code where you specify the number of elements. Your code should partition the domain automatically. However, if you want to make the code more general (for future assignments), you should put in the following features:

- element endpoint locations (different sized elements),
- the possibility for different material values for each element \((E(x))\).
E.3 Sample Project 2: Higher-Order Elements

- Consider the following boundary value problem, with domain $\Omega = (0, L)$:

$$\frac{d}{dx} \left( E \frac{du}{dx} \right) = x k^3 \cos \left( \frac{2\pi k x}{L} \right)$$

$E = 0.2$

$k = 12$

$L = 1$

$u(0) = \Delta_1 = \text{given constant} = 3$

$u(L) = \Delta_2 = \text{given constant} = -1$

(E.4)

- Solve this with the finite element method using order $p$ equal-sized elements. In order to achieve

$$e^N \overset{\text{def}}{=} \frac{||u - u^N||_{E(\Omega)}}{||u||_{E(\Omega)}} \leq TOL = 0.04,$$

$$||u||_{E(\Omega)} \overset{\text{def}}{=} \sqrt{\int_{\Omega} \frac{du}{dx} E \frac{du}{dx} dx}$$

(E.5)

how many finite elements ($N$) are needed for

$$p = 1 \Rightarrow N = ?$$

$$p = 2 \Rightarrow N = ?$$

$$p = 3 \Rightarrow N = ?$$

(E.6)

- Plot the numerical solutions for several values of $N$, for each $p$, along with the exact solution.
- Plot $e^N$ as a function of the element size $h$ for each $p$.
- Plot $e^N$ as a function of the number of degrees of freedom for each $p$.
- Determine the relationship between the error and the element size for each $p$.
- Note: Please be careful with the quadrature order...you will need higher order Gauss rules for quadratic and cubic elements.
E.4 Sample Project 3: Potential and Efficient Solution Techniques

- Solve the following boundary value problem, with domain $\Omega = (0, L)$, analytically:

$$
\frac{d}{dx} \left( E(x) \frac{du}{dx} \right) = x k^3 \cos \left( \frac{2\pi k x}{L} \right)
$$

$E(x) = 10$ different segments (see below)

$k = 12, \ L = 1, \ u(0) = -0.3, \ u(L) = 0.7$

For $E$

| $0.0 < x < 0.1$ | $E = 2.5$ |
| $0.1 < x < 0.2$ | $E = 1.0$ |
| $0.2 < x < 0.3$ | $E = 1.75$ |
| $0.3 < x < 0.4$ | $E = 1.25$ |
| $0.4 < x < 0.5$ | $E = 2.75$ |
| $0.5 < x < 0.6$ | $E = 3.75$ |
| $0.6 < x < 0.7$ | $E = 2.25$ |
| $0.7 < x < 0.8$ | $E = 0.75$ |
| $0.8 < x < 0.9$ | $E = 2.0$ |
| $0.9 < x < 1.0$ | $E = 1.0$ |

- Solve this with the finite element method using linear equal-sized elements. Use 100, 1000, and 10000 elements. You are to write a preconditioned Conjugate Gradient solver. Use the diagonal preconditioning given in the notes. The data storage is to be done element by element (symmetric), and the matrix vector multiplication is to be done element by element.
- You are to plot the solution (nodal values) for each $N$.
- You are to plot

$$
E_N \stackrel{\text{def}}{=} \frac{||u - u^N||_{E(\Omega)}}{||u||_{E(\Omega)}},
$$

$$
||u||_{E(\Omega)} \stackrel{\text{def}}{=} \sqrt{\int_{\Omega} \frac{du}{dx} E \frac{du}{dx} dx},
$$

for each $N$. 

(E.7)
• You are to plot

\[ \text{Potential energy} = J(u^N) \]  

(E.10)

for each \( N \).

• You are to plot the number of PCG solver iterations for each \( N \) for a stopping tolerance of 0.000001.

• Use a Gauss integration rule of level 5.

• Check your Conjugate Gradient generated results against a regular Gaussian solver, for example the one available in MATLAB.

---

**E.5 Sample Project 4: Error Estimation and Adaptive Meshing**

**Using the Exact Solution as a Test**

• Consider the boundary value problem \( \frac{d}{dx} \left( E \frac{du}{dx} \right) = f(x) \), \( E = 1 \), with domain \( \Omega = (0, L) \), \( L = 1 \), and solution \( u(x) = \cos(10\pi x^5) \).

• Compute the finite element solution \( u^N \) to this problem using linear equal-sized elements. Determine how many elements are needed in order to achieve

\[ e^N \defeq \frac{||u - u^N||_{E(\Omega)}}{||u||_{E(\Omega)}} \leq TOL = 0.05, \]

\[ ||u||_{E(\Omega)} \defeq \sqrt{\int_\Omega \left( \frac{du}{dx} E \frac{du}{dx} \right) dx} \]

• Plot \( I \) versus \( A_I \), where

\[ A_I^2 \defeq \frac{1}{\Omega_I} ||u - u^N||^2_{E(\Omega_I)} \]

Here \( I \) is the element index, \( h_I \) is the length of element \( I \), and

\[ ||u||^2_{E(\Omega_I)} \defeq \int_{\Omega_I} \frac{du}{dx} E \frac{du}{dx} dx. \]

• Modify your code from HW 1 so that it can automatically refine the mesh the following criterion:

• Refine the mesh (by dividing elements into two) until \( A_I < TOL_E \) for all \( I \). Use this criterion to refine your mesh, starting with \( N = 20 \) equal-sized elements:

  – Determine how many elements are needed to achieve \( A_I < TOL_E = 0.05 \) for all \( I \).
  – Plot the final solution, together with the exact solution.
– Tabulate the final number of elements that fall into each of the initial 20 elements.
– Plot $X_I$ versus $A_I$ for the final solution ($X_I =$ position of node $I$).

### E.6 Sample Project 5: 3D Formulations for Elasticity

You are given a tubular multiphase structure (Fig. E.1) with an elasticity of $\mathbf{E}(x, y, z)$, and with dimensions shown in the figure. It is clamped on one end and externally traction loaded everywhere else, including on the interior surface. The small deformation of the body is governed by (strong form):

$$
\nabla \cdot (\mathbf{E} : \nabla \mathbf{u}) + \rho \mathbf{b} = 0
$$  \hspace{1cm} (E.11)

where $\mathbf{E}$ and $\rho$ are spatially variable and where $\mathbf{b} = \mathbf{b}(x, y, z)$ is given data.

- Develop a weak form, providing all the steps and assumptions necessary. Carefully define the spaces of approximation.
- Develop a finite element weak form. Carefully define the spaces of approximation.
- Develop a finite element weak statement using the penalty method. Carefully define the spaces of approximation.
- Derive the equations for element stiffness matrices (be explicit) and load vectors. Thereafter, describe how the global stiffness matrix and load vector are generated, using the penalty method. Use trilinear subspatial approximations. There are different kinds of loading on the surfaces, so be very explicit as to what each of the individual stiffness matrices and right-hand-side vectors look like, as well as a generic element that is not on the surface.
- Write a mesh generator. Explicitly explain how it works and, in particular, the connectivity function. Use $N_t$ elements in the thickness direction, $N_e$ elements in

![Fig. E.1 3D structure](image)
the circumferential direction, and $N_{\theta}$ elements in the $\theta$ direction for each semi-circular portion. For the given figure $N_t = 3$, $N_c = 4$, and $N_{\theta} = 8$ generate the mesh - show it.

- If one were to use a Conjugate Gradient solver, theoretically how many operation counts would be needed to solve this problem for a mesh of $N_t$ elements in the thickness direction, $N_c$ elements in the circumferential direction, and $N_{\theta}$ elements in the $\theta$ direction

### E.7 Sample Project 6: Implementation of the Finite Element Method in 2D

- Solve the following boundary value problem, on an arch-shaped domain, using the finite element method:

  \[
  \nabla \cdot (K \nabla T) + f = 0, \\
  T = T_0 \text{ along } \theta = \pi \\
  -K\nabla T \cdot n = q_0(r) \text{ along } \theta = 0 \\
  -K\nabla T \cdot n = 0 \text{ along } r = r_i, r_o \\
  K = K_1 \text{ for } ||x - x_c|| \leq r_c \\
  K = K_2 \text{ for } ||x - x_c|| > r_c 
  \]

These equations describe a thermal physics problem of the two-phase structure that is shown in Fig. E.2.

- You are to generate a mesh of the domain for $r_i = 3$ and $r_o = 4$. Use $N_r \times N_{\theta}$ quadrilateral elements. For example, in Fig. E.3, $N_r = 3$ and $N_{\theta} = 12$. Write a finite element code to solve this problem with bilinear shape functions. For elements that have material discontinuities use a $5 \times 5$ Gaussian integration rule, otherwise, use a $2 \times 2$ rule.

**Fig. E.2** An arch
Fig. E.3  The proposed mesh for the arch

- Solve the problem both by strictly enforcing the boundary conditions and by using the penalty method on $\Gamma_T$ (the part of the boundary where the temperature is prescribed). Explain how your choice of penalty parameter affects the results.
- To verify that your code works properly, solve this with $k_1 = k_2 = 1$, $T_0 = 100$, $q_0(r) = \frac{40}{r}$, and $f(r, \theta) = \frac{80}{r^2} \sin(2\theta)$. Determine the exact solution for this problem (Hint: the solution is independent of $r$). Include a plot of your solution for $N_r = 10$ and $N_\theta = 80$.
- Solve the problem with $k_1 = 10^{-3}$, $k_2 = 1$, $T_0 = 110$, $q_0(r) = \frac{20}{r}$, $f(r, \theta) = \frac{40}{r^2} \sin(2\theta)$, $r_c = 0.40$, and $x_c$ is given by $(r = 3.5, \theta = \pi/2)$. Include a plot of your solution for $N_r = 50$ and $N_\theta = 400$.

E.8  Sample Project 7: Time-Dependent Problems

Part 1: formulation

You are given a two-phase (two material) structure (Fig. E.4), comprised of a two semicircular rings, with scalar diffusivity $D(x, y, z)$ and with dimensions shown in the figure. It is externally flux loaded on a portion of the top surface and one end surface, while the other end surface has as specified concentration. All other
The physics of the body is described by a simplified version of the diffusion-reaction equation, which in strong form is:

$$\nabla \cdot (D \nabla c) - \tau c + f = 0,$$  \hspace{1cm} (E.12)

where $D$ is a nonconstant, positive scalar function and where $f = f(x, y, z)$ is given data (sources).

- Now develop a weak form for the statement, providing all the steps and assumptions necessary.
- Develop a finite element weak statement. Carefully define the spaces of approximation.
- Develop a finite element weak statement using the penalty method. Carefully define the spaces of approximation.
- Derive explicit equations for element stiffness matrices and load vectors. Thereafter, describe how the global stiffness matrix and load vector are generated, using the penalty method. Use trilinear subspatial approximations. There are different kinds of loading on the surfaces, so be very explicit as to what each of the individual stiffness matrices and right-hand-side vectors look like, as well as a generic...
element that is not on the surface. Also, pay attention to the fact that elements may or may not have discontinuities when using Gaussian integration (consider both cases).

- Using your mesh generator, modify it to handle this new problem. Explicitly explain how it works and, in particular, the connectivity function. Use $N_t$ elements in the thickness direction, $N_c$ elements in the circumferential direction, and $N_{\theta}$ elements in the $\theta$ direction for each semicircular portion.

- If one were to use a Conjugate Gradient solver, theoretically how many operation counts would be needed to solve this problem for a mesh of $N_t$ elements in the thickness direction, $N_c$ elements in the circumferential direction, and $N_{\theta}$ elements in the $\theta$ direction.

- Now consider the time-transient case. The body has the same boundary conditions as before, with the initial condition that $c(t = 0, x, y, z) = c_0(x, y, z)$. The governing equation is

$$\nabla \cdot (D \nabla c) - \tau c + f = \dot{c}, \quad (E.13)$$

Develop a finite element weak statement. Carefully define the spaces of approximation. Use the finite difference approximation that we have used this semester for the time-dependent term.

- Finally, given that this is a three-dimensional problem, with heterogeneous coefficients, it is most likely you will need a large number of elements to solve it. Suppose that your machine has only enough memory to allow you to solve a wedge (sector) of $0 \leq \theta \leq \theta^* \ll \pi$ degrees, but that you need to solve the entire $0 \leq \theta \leq \pi$ problem. How would you break the problem and solve it? Give an overall algorithm.

**Part 2: Implementation in 1D**

- Solve the following boundary value problem ($L = 1$)

\[
\frac{\partial c}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial c}{\partial x} \right) - \tau c \\
D(x) = 10 \text{ different segments (see below)} \\
\tau(x) = 10 \text{ different segments (see below)} \\
c(x = 0, t) = 0.5 \\
D \frac{\partial c}{\partial x}(x = L, t) = 5 \times 10^{-6} \\
c(x, t = 0) = 0.5 \quad 0 < x < L
\]
with 100 elements ($\delta x = 0.01$) and set the total amount of time to be $T = 6500$ s. Use an implicit (Backward Euler) time-stepping scheme. Solve with the following time-step sizes: (I) $\delta t = \frac{T}{100}$, (II) $\delta t = \frac{T}{1000}$, and (III) $\delta t = \frac{T}{10000}$ with

\[
\begin{align*}
\text{For } 0.0 < x < 0.1 & \quad D = 2.4 \times 10^{-6} \\
\text{For } 0.1 < x < 0.2 & \quad D = 2.0 \times 10^{-6} \\
\text{For } 0.2 < x < 0.3 & \quad D = 1.5 \times 10^{-6} \\
\text{For } 0.3 < x < 0.4 & \quad D = 0.6 \times 10^{-6} \\
\text{For } 0.4 < x < 0.5 & \quad D = 1.3 \times 10^{-6} \\
\text{For } 0.5 < x < 0.6 & \quad D = 0.14 \times 10^{-6} \\
\text{For } 0.6 < x < 0.7 & \quad D = 1.1 \times 10^{-6} \\
\text{For } 0.7 < x < 0.8 & \quad D = 2.2 \times 10^{-6} \\
\text{For } 0.8 < x < 0.9 & \quad D = 2.0 \times 10^{-6} \\
\text{For } 0.9 < x < 1.0 & \quad D = 1.5 \times 10^{-6}
\end{align*}
\]  
(E.15)

\[
\begin{align*}
\text{For } 0.0 < x < 0.1 & \quad \tau = 1.2 \times 10^{-3} \\
\text{For } 0.1 < x < 0.2 & \quad \tau = 0.8 \times 10^{-3} \\
\text{For } 0.2 < x < 0.3 & \quad \tau = 0.3 \times 10^{-3} \\
\text{For } 0.3 < x < 0.4 & \quad \tau = 1.4 \times 10^{-3} \\
\text{For } 0.4 < x < 0.5 & \quad \tau = 1.15 \times 10^{-3} \\
\text{For } 0.5 < x < 0.6 & \quad \tau = 0.75 \times 10^{-3} \\
\text{For } 0.6 < x < 0.7 & \quad \tau = 0.35 \times 10^{-3} \\
\text{For } 0.7 < x < 0.8 & \quad \tau = 0.85 \times 10^{-3} \\
\text{For } 0.8 < x < 0.9 & \quad \tau = 1.25 \times 10^{-3} \\
\text{For } 0.9 < x < 1.0 & \quad \tau = 2.0 \times 10^{-3}
\end{align*}
\]  
(E.16)

- Clearly show the evolution of solution with time, together with the final time solution. Also, comment on the effect of time-step size on the solution.

**References**