Chapter 2

Preliminaries

The behavior of deforming objects is the topic of continuum mechanics, a branch of mathematics that tries to capture physical phenomena of continuous media in precise mathematical formulations. One branch of continuum mechanics, nonlinear elasticity, provides the mathematical description of how objects deform. Since deformation is fundamental to the topic of the thesis, we will review elasticity in Section 2.1, drawing on the book by Antman [4]. An introduction to the tensor notation used is given in Appendix A.

Continuum mechanics describes materials in terms of partial differential equations. Solving such equations will be done by the Finite Element Method, so this broad category of solution strategies is discussed in Section 2.2. We will work towards the specific technique that we shall use: the Rayleigh-Ritz method for variational problems, using linear interpolation on tetrahedra. This section is loosely based on the introductory chapters of the books by Zienkiewicz and Taylor [103] and Braess [14] on the subject.

The Finite Element Method (FEM) is a discretization method. It transforms a continuous, infinite-dimensional problem into systems of equations with a finite number of variables. For mechanical problems, the FEM discretizes the equations of motion, hence it delivers a system of ordinary differential equations, i.e., equations where time still has a role. There are two ways to deal with these systems: compute the evolution of the system, or try to find the final equilibrium solution directly.

If the final state of the system is all that matters, a *static* method can be used. By assuming that velocity and acceleration are null, the system of differential equations is changed into a normal system of equations. For many mechanical problems, these equations can be stated in terms of finding minimum energy solutions. Hence, we will discuss a number of minimization algorithms in Section 2.4. A more extensive treatment of unconstrained optimization algorithms can be found in the work of Nocedal [75] and Fletcher [41].

If transient effects do matter, then the evolution of the differential equations must be calculated using a *time-integration* method. Section 2.5 discusses a popular time-integration method for mechanical problems, based on the book by Zienkiewicz and Taylor [103].

Basically, our problems come from the simulation of soft tissue. Although simulating the full mechanical characteristics of soft tissue is not possible in an interactive setting, it is instructive to study exactly what kind of characteristics are ignored in our simulations. Section 2.6 briefly discusses a few mechanical properties of living tissue, drawing on the book by Fung [44].

All sections in this chapter tend towards simplification. This is not surprising: the constraints of an interactive simulation do not allow for much sophistication. This chapter also states many results without proving their correctness, and it is in some parts deliberately vague. Since the subjects are too extensive to fit a full discussion in this thesis, the reader is referred to the books mentioned above for more detailed information.

2.1 Continuum mechanics

Continuum mechanics describes the behavior of material objects when they are subjected to loading. The basic assumption is that the objects and their behavior may be described using continous quantities: bodies occupy a continuous region in 3D space, and have continuous motions.

We describe a mechanical or material body as being a subset $\mathcal{B}$ of $\mathbb{R}^3$. We call this set $\mathcal{B}$ the reference configuration. As time progresses, the body may occupy different positions in space: every point of $\mathcal{B}$ moves to some location, depending on the time t. Hence we may describe a deformation by a function $\mathbf{p} : \mathcal{B} \times \mathbb{R} \to \mathbb{R}^3$.

Such a function $\mathbf{p}(\mathbf{z}, t)$ must satisfy at least two requirements to represent a valid motion. First, a body may not penetrate itself: no two points of the body may occupy the same position when deformed. Second, the body can not be folded inside out. In other words, the function $\mathbf{p}(\cdot, t)$ must preserve orientation for all t. Since the volume change of the deformed object is measured by $\det(\mathbf{p}_{\mathbf{z}}(\mathbf{z}, t))$, we require that this quantity be positive for all t. We shall encounter the derivative $\mathbf{p}_{\mathbf{z}}$ much more often, and therefore we give it a name and a notation: the 2-tensor field $\mathbf{p}_{\mathbf{z}}(\mathbf{z}, t)$ is called the deformation gradient, and is denoted by $\mathbf{F}$. It is illustrated for a 2D example in Figure 2.1.

When we talk of deformation, usually we refer to motions $\mathbf{p}$ which locally change the shape of the body. Isometric transformations (rigid movements) do not interest us, therefore we shall use a measure of shape change which filters out these rigid motions. Let

$$\mathbf{C} = \mathbf{F}^* \cdot \mathbf{F}. \tag{2.1}$$

This tensor is called the *(right Cauchy-)Green deformation tensor*, and it is invariant under rotations. This deformation tensor measures the elongation of a fiber running in a particular direction: if we have an infinitesimally long fiber with direction running from $\mathbf{z}$ to $\mathbf{z} + \mathbf{h}$, then its length after deformation is measured by $\sqrt{\mathbf{h} \cdot \mathbf{C} \cdot \mathbf{h}}$. The length change of such a fiber attains extremes when $\mathbf{h}$ is an eigenvector of $\mathbf{C}$. Since $\mathbf{C}$ is symmetric, it has three orthogonal eigenvectors, called the *principal axes of strain*. The tensor $\mathbf{C}$ is also visualized in Figure 2.1.

Solid bodies resist movement. This property is called inertia, and inertia is measured by the *mass* of a body. We assume that a body $\mathcal{B}$ has a density $\rho(\mathbf{z})$ in each point $\mathbf{z}$, and

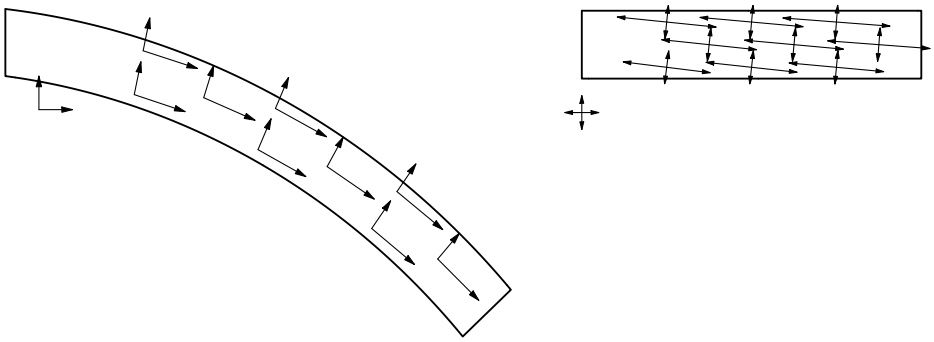

Figure 2.1: This figure shows a deformation of a beam in 2D (reference configuration on the right), with vectors of the deformation gradient (left) and principal axes of strain (right). The identity mapping is represented in the lower left of each picture. The motion used is $\mathbf{p}((x,y)) = -ce_x + \mathbf{R}(-\alpha x)(y+c)e_y - \beta x e_y$, where $\mathbf{R}(\phi)$ is the mapping for rotation over angle ϕ, and e_1, e_2 is the unit basis in $\mathbb{R}^2$. This corresponds to a combination of bending, dilation and shearing. The deformation gradient $\mathbf{F}$ is really defined in points of the reference configuration, but we show $\mathbf{F}$ in the corresponding deformed points $\mathbf{p}(z)$.

the mass of a some part $\mathcal{A}$ of a body is given by the volume integral

$$\int_{\mathcal{A}} \rho(z)\, dv(z). \tag{2.2}$$

A part $\mathcal{A}$ of a body $\mathcal{B}$ may be subject to force. Forces take the form of either traction on the surfaces or body forces, which are applied to the volume of the body. Surface tractions are denoted by $\mathbf{t}$. Body forces (which can be either electromagnetic or gravity forces) are denoted by $\mathbf{f}$,

The balance of linear momentum, which is also known as Newton's law, postulates that resultant forces equal acceleration. It may be formulated as follows.

$$\begin{aligned} &\int_{\mathcal{A}} \mathbf{f}(z,t) dv(z) + \int_{\partial\mathcal{A}} \mathbf{t}(z,t;\partial\mathcal{A})\, da(z) \\ &\quad = \frac{d}{dt}\int_{\mathcal{A}} \mathbf{p}_t(z,t)\rho(z)\, dv(z), \qquad \mathcal{A} \subset \mathcal{B}. \end{aligned} \tag{2.3}$$

The left side of the equation represents body force plus boundary tractions. The right side of the equation is the derivative of linear momentum. This equation holds for every subset $\mathcal{A}$ of the body $\mathcal{B}$. The traction t on a surface $\partial\mathcal{A}$ depends only on z through the surface normal n on $\partial\mathcal{A}$. This relation is linear, so there is a 2-tensor $\mathbf{T}$ defined on the body $\mathcal{B}$, such that

$$\mathbf{t}(z,t;\partial\mathcal{A}) = \mathbf{T}(z,t)\cdot n, \qquad z \in \partial\mathcal{A}, n \perp \partial\mathcal{A}. \tag{2.4}$$

This theorem is called Cauchy's stress theorem, and the mapping $\mathbf{T}$ is called the *first Piola-Kirchoff stress tensor*. The vector $\mathbf{n}$ is the outward pointing surface normal of $\partial\mathcal{A}$ at point $\mathbf{z}$ in the reference configuration.

Equation (2.3) integrates variables over parts of $\mathcal{B}$, hence it is called *weak*. If $\mathbf{p}$ is continuously differentiable, then the balance of momentum can be rewritten in a *strong*, localized version

$$\operatorname{div}\mathbf{T}(\mathbf{z},t) + \mathbf{f}(\mathbf{z},t) = \rho(\mathbf{z})\mathbf{p}_{tt}(\mathbf{z},t), \qquad \mathbf{z} \in \mathcal{B}. \tag{2.5}$$

The operator div is the divergence of a 2-tensor field. It is defined by $\operatorname{div}\mathbf{T} = \partial\mathbf{T}/\partial\mathbf{z} : \mathbf{I}$, where $\cdot : \cdot$ is the inner product for 2-tensors.

The balance of angular momentum asserts that resultant couples equal angular acceleration. If we assume that no electromagnetic effects take place, then resultant couples are always null, and the tensor $\mathbf{T} \cdot \mathbf{F}^*$ is symmetric. This leads us to introduce the tensor $\mathbf{S}$, called the second Piola-Kirchoff stress tensor, by defining

$$\mathbf{S} = \mathbf{F}^{-1} \cdot \mathbf{T}. \tag{2.6}$$

This tensor is symmetric. If $\boldsymbol{e}_1$ is the normal of a material plane in $\mathcal{B}$, and $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$ span the plane in $\mathcal{B}$, then the matrix entries of $\mathbf{S}$ contain the traction on that plane, $\mathbf{T}\boldsymbol{e}_1$, but expressed in deformed basis vectors $\{\mathbf{F}\boldsymbol{e}_1, \mathbf{F}\boldsymbol{e}_2, \mathbf{F}\boldsymbol{e}_3\}$.

Constitutive equations define the relations between stress and deformation. A constitutive equation is given by a function $\hat{\mathbf{T}}$, such that

$$\mathbf{T} = \hat{\mathbf{T}}(\mathbf{p}(\cdot,\cdot), \mathbf{z}, t), \qquad \mathbf{z} \in \mathcal{B}. \tag{2.7}$$

The dependency on $\mathbf{p}$ is understood to be causal: the value of $\hat{\mathbf{T}}$ on some time t_0 only depends on values of $\mathbf{p}$ before t_0. The hat on $\hat{\mathbf{T}}$ stresses the difference between the tensor field $\mathbf{T}$, a physical quantity that could be measured in physical realizations of the situations described, and $\hat{\mathbf{T}}$, a function that expresses the mathematical relation between two tensor fields (in this case, the field $\mathbf{T}$ and $\mathbf{p}(\cdot,\cdot)$). For the stress tensor $\mathbf{S}$ we use the same convention.

The description in Equation (2.7) is too general to be of practical use, and therefore we directly restrict ourselves to *simple materials*, where the stress at a point $\mathbf{z}$ only depends on the values of $\mathbf{p}$ in a neighborhood of $\mathbf{z}$, as measured by the first derivatives $\mathbf{p}_z$. We get

$$\mathbf{T} = \hat{\mathbf{T}}(\mathbf{p}(\mathbf{z},\cdot), \mathbf{p}_z(\mathbf{z},\cdot), \mathbf{z}, t), \tag{2.8}$$

for some function $\hat{\mathbf{T}}$. Again the dependency on $\mathbf{p}$ and its derivatives is understood to be causal.

Constitutive equations should be independent of the choice of a time and coordinate system. This property is called *frame-indifference*. For a simple material, this implies that

$$\hat{\mathbf{T}}(\mathbf{F}(\mathbf{z},\cdot), \mathbf{z}) = \mathbf{R}\hat{\mathbf{T}}(\mathbf{U}(\mathbf{z},\cdot), \mathbf{z}), \tag{2.9}$$

$$\mathbf{U}(\mathbf{z},\cdot) = \sqrt{\mathbf{F}(\mathbf{z},\cdot)^* \cdot \mathbf{F}(\mathbf{z},\cdot)}, \tag{2.10}$$

$$\mathbf{R} = \mathbf{F}(\mathbf{z},t) \cdot \mathbf{U}^{-1}(\mathbf{z},t). \tag{2.11}$$

The tensors $\mathbf{U}$ and $\mathbf{R}$ form the polar decomposition of $\mathbf{F}$: $\mathbf{F} = \mathbf{R} \cdot \mathbf{U}$, where $\mathbf{R}$ is a rotation, and $\mathbf{U}$ is a positive definite symmetric matrix. The dependency on $\mathbf{U}(z, \cdot)$ is meant to be causal.

By substituting Equation (2.6) into (2.9), we can derive the following description for a frame indifferent constitutive equation which uses symmetric tensors only:

$$\mathbf{S}(z, t) = \hat{\mathbf{S}}(\mathbf{C}(z, \cdot), z).$$

A further simplification can be done if we assume that the material responds to deformations equally in all directions. Then for all rotation mappings $\mathbf{Q}$ we have

$$\mathbf{Q} \cdot \hat{\mathbf{S}}(\mathbf{C}, z) \cdot \mathbf{Q}^* = \hat{\mathbf{S}}(\mathbf{Q} \cdot \mathbf{C} \cdot \mathbf{Q}^*, z).$$

Such material is called *isotropic* or *hemitropic*. If not, then the material is called *aelotropic* or *anisotropic*. Examples of materials that are aelotropic are those that contain networks of regularly arranged fibers.

The tensor $\mathbf{C}$ is symmetric, and can therefore be characterized up to rotations by its three eigenvalues, or equivalently, by its three invariants. If λ_1, λ_2 and λ_3 are the eigenvalues of $\mathbf{A}$, then the invariants ι_1, ι_2 and ι_3 are defined as follows:

$$\begin{aligned} \iota_1(\mathbf{A}) &= \lambda_1 + \lambda_2 + \lambda_3, \\ \iota_2(\mathbf{A}) &= \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1, \\ \iota_3(\mathbf{A}) &= \lambda_1\lambda_2\lambda_3. \end{aligned}$$

The invariants can be determined without computing the eigenvalues of $\mathbf{A}$. We have

$$\begin{aligned} \iota_1(\mathbf{A}) &= \operatorname{trace}(\mathbf{A}), \\ \iota_2(\mathbf{A}) &= \frac{1}{2}((\operatorname{trace}\mathbf{A})^2 - \operatorname{trace}(\mathbf{A}^* \cdot \mathbf{A})), \\ \iota_3(\mathbf{A}) &= \det\mathbf{A}. \end{aligned}$$

The trace of a matrix $\mathbf{H}$ can be defined in terms of the inner product for 2-tensors: $\operatorname{trace}(\mathbf{H}) = \mathbf{H} : \mathbf{I}$. In a matrix representation, the trace of a matrix is the sum of the diagonal elements.

The following representation theory should come as no surpise: a hemitropic stress function $\hat{\mathbf{S}}$ depends on $\mathbf{C}$ only through the invariants ι_1, ι_2 and ι_3 of $\mathbf{C}$. For hyperelastic media, this means that the energy density W must also be a function of the invariants:

$$W = \hat{W}(\iota_1, \iota_2, \iota_3, z). \tag{2.12}$$

Some materials have restriction on the ways in which they can deform. The most important example of this is *incompressibility*. A material is incompressible if deformations conserve volume locally. This constraint may be expressed as $\det\mathbf{C} = 1$. Constitutive equations describing incompressible material can be derived by a limit process. This limit process introduces a new variable, the pressure p, that serves to balance any elastic force that tries to violate the constraint. Examples of incompressible material are those containing lots of fluid.

We can distinguish many classes of material by restricting the dependency of $\mathbf{T}$ on $\mathbf{p}$ even further. We will restrict ourselves to *elastic* materials. These are materials where $\mathbf{T}$ depends only on $\mathbf{p}_z$, and not on t or $\mathbf{p}$ itself. The dependency on $\mathbf{p}_z$ does not take the past history of $\mathbf{p}_z$ into account, but only its value at time t. In other words, this is material that satisfies

$$\mathbf{T} = \hat{\mathbf{T}}(\mathbf{p}_z(z, t), z), \tag{2.13}$$

for some function $\hat{\mathbf{T}}$. Such a material is also called *Cauchy-elastic*. An elastic material is hyperelastic, or *Green-elastic*, if there is a scalar function W called stored energy function, such that

$$\hat{\mathbf{T}}(\mathbf{F}, z, t) = \frac{\partial W(\mathbf{F}, z)}{\partial \mathbf{F}}, \tag{2.14}$$

or equivalently

$$\hat{\mathbf{S}}(\mathbf{C}, z, t) = 2\frac{\partial W(\mathbf{C}, z)}{\partial \mathbf{C}}. \tag{2.15}$$

If the stress is independent of temperature, or temperature is held constant, then the equilibrium response of any elastic material is hyperelastic.

These considerations give us some constraints on functions that can be used as $\hat{\mathbf{T}}$ and W, but not all such functions lead to descriptions that realistically describe existing materials. Mathematical conditions exist that express basic properties of materials, e.g., elastic forces should oppose deformations, and extreme deformations should lead to extreme tensions within the material. This still leaves a lot of freedom in specifying constitutive equations, i.e. selecting W or $\hat{\mathbf{T}}$ functions. Finding a constitutive equation for a given material is far from trivial. A variety of theoretical models exist for different classes of material (e.g. metals and rubbers), but they are usually only validated by fragmentary experimental results [5].

The equations presented so far are inherently nonlinear: the strain tensor $\mathbf{C}$ is nonlinear in $\mathbf{p}$. This ensures that solutions to the equations for $\mathbf{p}$ will not be linear in the given conditions, i.e. the boundary conditions and the body forces. However, a linear approximation to the elasticity does exist. It has three virtues: it is simpler from a conceptual point of view, it is computationally simpler, and relatively uncomplicated proofs of existence and unicity of the solution can be given.

If we assume that deformations are small, then we can assume that the stress strain relation is linear, or equivalently, W is quadratic in $\mathbf{C}$. This is called the *St. Venant-Kirchoff* material model, and it is the simplest material model available. It will be used in Chapter 4. Since W is quadratic in this model, must have the following form.

$$W(\iota_1, \iota_2) = c_0 + c_1\iota_1 + c_2\iota_1^2 + c_3\iota_2. \tag{2.16}$$

The choice of c_0 is irrelevant, so we set $c_0 = 0$. The constants $c_1, \ldots, c_3$ are dependent: the reference configuration should be stress-free, so we can eliminate one constant by imposing $\mathbf{S} = \mathbf{0}$ when $\mathbf{C} = \mathbf{I}$. Two constants remain. We can express the energy density thus obtained as follows:

$$W(\iota_1, \iota_2) = \frac{1}{2}\left(\left(-\mu - \frac{3\lambda}{2}\right)\iota_1 + \left(\frac{\lambda}{4} + \frac{\mu}{2}\right)\iota_1^2 - \mu\iota_2\right). \tag{2.17}$$

The stress tensor is linear in $\mathbf{C}$,

$$\mathbf{S} = \mu(\mathbf{C} - \mathbf{I}) + \frac{1}{2}\lambda(\iota_1 - 3)\mathbf{I}.$$

Hence the tension in the material is proportional to the deformation, as measured by $\mathbf{C}-\mathbf{I}$, with additional tension caused by volume changes, which are measured by (ι_1-3) for small deformations.

The parameters μ and λ are called Lamé parameters. Material properties are usually expressed in parameters that have a macroscopically more intuitive definition: the Young modulus measures the resistance to stretching. Its notation is E, and we have

$$E = \mu\frac{(3\lambda + 2\mu)}{\lambda + \mu}. \tag{2.18}$$

The unit of E is Pascal (N/m^2). The Poisson ratio is the ratio between the transverse contraction and longitudinal stretching. We have

$$\nu = \frac{\lambda}{2(\lambda + \mu)}. \tag{2.19}$$

Since ν is a ratio, it is dimensionless. For physical reasons, we have $0 \leq \nu < 1/2$. If ν tends to $1/2$, then $\lambda \to \infty$, and the material becomes incompressible. The meaning of E and ν is illustrated in Figure 2.2.

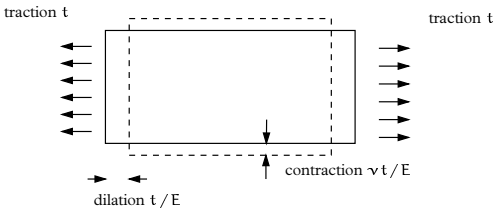

Figure 2.2: A uniformly distributed traction leads to uniform stresses and strains. Material in the direction of the load expands inversely proportional to the Young modulus E. The contraction in the transversal direction is ν times the dilation, where ν is the Poisson ratio.

We introduce the displacement $\mathbf{u}$ of a point. The displacement $\mathbf{u}$ of $\boldsymbol{z}$ at time t is defined by

$$\mathbf{u}(\boldsymbol{z}, t) = \mathbf{p}(\boldsymbol{z}, t) - \boldsymbol{z}. \tag{2.20}$$

This expression vanishes in the reference configuration. For rotations, $\mathbf{u}$ generally is nonzero. We write $\mathbf{G}$ for $\frac{\partial \mathbf{u}}{\partial \boldsymbol{z}}$. We can linearize both the definition of strain in Equation (2.1), and the constitutive relations in (2.7). Linearizing the strain tensor is also called the linear geometry assumption. We have

$$\mathbf{C} = \mathbf{F}^* \cdot \mathbf{F} = (\mathbf{I} + \mathbf{G})^* \cdot (\mathbf{I} + \mathbf{G}) = \mathbf{I} + \mathbf{G}^* + \mathbf{G} + \mathbf{G}^* \cdot \mathbf{G}.$$

When we assume that $\mathbf{G}$ is small, then we can neglect the quadratic term $\mathbf{G}^* \cdot \mathbf{G}$, obtaining

$$\mathbf{C} \approx \mathbf{I} + \mathbf{G}^* + \mathbf{G}.$$

Another measure for deformation is the material strain tensor $\mathbf{E}$, which is defined by

$$\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{I}). \tag{2.21}$$

The tensor $\mathbf{E}$ disappears in the reference configuration. Let $\mathcal{E}$ be the linear geometry approximation of $\mathbf{E}$, then we have

$$\begin{aligned} \mathcal{E} &= \frac{1}{2}(\mathbf{G} + \mathbf{G}^*), \\ \tilde{W}(\mathcal{E}) &= \mu \operatorname{trace}(\mathcal{E}^* \mathcal{E}) + \frac{1}{2}\lambda(\operatorname{trace} \mathcal{E})^2, \\ \mathbf{S} &= 2\mu\mathcal{E} + \lambda \mathbf{I}(\operatorname{trace} \mathcal{E}). \end{aligned} \tag{2.22}$$

When both strain and material linearizations are combined, stresses are linear in the displacement. If deformations are small, then $\partial \mathbf{u}/\partial \mathbf{z} = \mathcal{O}(\varepsilon)$ for some small $\varepsilon > 0$. In this case $\mathcal{O}(\varepsilon^2)$ terms are neglible and both $\mathcal{E}$ and $\mathbf{S}$ are also $\mathcal{O}(\varepsilon)$. It follows that we can equate $\mathbf{S}$ and $\mathbf{T}$ since

$$\mathbf{T} = \mathbf{F} \cdot \mathbf{S} = (\mathbf{I} + \mathcal{O}(\varepsilon)) \cdot \mathbf{S} = \mathbf{S} + \mathcal{O}(\varepsilon^2). \tag{2.23}$$

This combination is called linear elasticity. Suppose that the displacements are fixed on a subset of the boundary with non-zero area, say $\partial\mathcal{B}_0$, and tractions are specified on the rest of the boundary, say $\partial\mathcal{B}_3$ and $V(\mathcal{B})$ is the set of functions with square-integrable derivatives that satisfy the boundary condition on $\mathcal{B}_0$, then then the following problem has a unique solution.

$$\min_{u \in V(\mathcal{B})} \int_{\mathcal{B}} \left(\tilde{W}(\mathcal{E}(\mathbf{u})) - \mathbf{f} \cdot \mathbf{u} \right) \, dv(\mathbf{z}) + \int_{\partial\mathcal{B}_3} \mathbf{t} \cdot \mathbf{u} \, da(\mathbf{z}), \tag{2.24}$$

$$\mathbf{u} = \mathbf{0}, \qquad \text{on } \partial\mathcal{B}_0. \tag{2.25}$$

The solution is weak: the space $V(\mathcal{B})$ consists of functions from $\mathcal{B}$ to $\mathbb{R}^3$ with square-integrable derivatives that satisfy the boundary condition on $\mathcal{B}_0$.

2.2 Finite Element Method

Mathematical modeling processes such as the one in the previous section, yield a collection of partial differential equations. The unknown variable in such an equation is a sufficiently smooth function defined on the domain $\mathcal{B}$ of the equation. In the general case it is not possible to compute a solution in closed form to these problems. Therefore, approximative schemes are necessary. The Finite Element Method (FEM) is an approximative scheme, and it is very popular for mechanical analysis. In this section, we shall briefly explain the essentials of this method, and work towards the simple FEM scheme that we will use (linear tetrahedral elements).

Let us suppose, for the moment, that the partial differential equation is mechanical, set in $\mathbb{R}^3$, and that the equation we wish to solve states that the resultant force $\mathbf{r}$ for the displacement $\hat{\mathbf{u}}$ is zero on a domain $\mathcal{B} \subset \mathbb{R}^3$. In other words,

$$\text{find } \hat{\mathbf{u}} \in V(\mathcal{B}), \text{ such that } \mathbf{r}(\hat{\mathbf{u}}) = \mathbf{0}. \tag{2.26}$$

The set $V(\mathcal{B})$ is a linear space of candidate solution functions with suitable differentiability and boundary conditions. The functions in $V(\mathcal{B})$ represent displacements, so they take on values from $\mathbb{R}^3$. Equilibrium solutions of the equations of motion in (2.5) have such a form.

The space $V(\mathcal{B})$ is infinite-dimensional, and to make finding a solution tractable we search the solution in a smaller, finite-dimensional subspace of $V(\mathcal{B})$. Such subspaces are typically created by interpolating functions $V(\mathcal{B})$ using points from $\mathcal{B}$ spaced at distance h. Since h is a parameter, the notation for the subspace is $V_h(\mathcal{B})$. We obtain the following problem.

$$\text{Find } \tilde{\mathbf{u}} \in V_h(\mathcal{B}), \text{ such that } \mathbf{r}(\tilde{\mathbf{u}}) = \mathbf{0}. \tag{2.27}$$

The space V_h is much smaller than V, so it is unlikely that it contains a solution $\tilde{\mathbf{u}}$ for which (2.27) holds exactly. All we can really hope for is that $\mathbf{r}(\tilde{\mathbf{u}})$ is as small as possible. To measure this, we weigh $\mathbf{r}$ with functions $\boldsymbol{w}$ from a finite-dimensional space W_h, and state the problem as follows.

$$\begin{aligned} &\text{find } \tilde{\mathbf{u}} \in V_h(\mathcal{B}), \text{ such that for all } \tilde{\boldsymbol{w}} \in W_h \\ &\qquad \int_{\mathcal{B}} \tilde{\boldsymbol{w}}(z) \cdot (\mathbf{r}(\tilde{\mathbf{u}}))(z)\, dv(z) = 0. \end{aligned} \tag{2.28}$$

Formulations where the equations are integrated over sets in $\mathbb{R}^3$ are also called *weak* formulations. The weak equations of motion only require functions whose derivative is square integrable on $\mathcal{B}$ in the Lesbesgue sense. This explains why a V_h consisting of piecewise linear functions is sufficient for computing a solution: the derivatives of such functions are not defined everywhere, but they are square integrable.

Formulations such as Equation (2.26) set a condition that should hold in every point of the domain $\mathcal{B}$. Equation (2.5) is a second order partial differential equation, which suggests that candidate solutions should be twice differentiable, i.e. $V(\mathcal{B}) = C_2(\mathcal{B})$. Since this is more restrictive than the weak form of the equations, we call such a form *strong*.

The space W_h is finite-dimensional, and hence has a basis $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$ for some $k \in \mathbb{N}$. We may therefore rephrase this equation in terms of $\boldsymbol{w}_j$ as

$$\int_{\mathcal{B}} \boldsymbol{w}_j(z) \cdot (\mathbf{r}(\tilde{\mathbf{u}}))(z)\, dv(z) = 0, \quad j = 1, \ldots, k.$$

In some types of electromechanical calculations, and in the linear simplifications of the continuum mechanics of Section 2.1, it may happen that $\mathbf{r}$ is affine linear, say

$$\mathbf{r}(\mathbf{u}) = \mathbf{L}(\mathbf{u}) - \mathbf{f}.$$

Here, $\mathbf{L}$ is a differentiation operator. The space $V_h(\mathcal{B})$ is finite-dimensional, and hence has a basis $\mathbf{u}_1, \ldots, \mathbf{u}_l$ for some $l \in \mathbb{N}$. If we expand $\tilde{\mathbf{u}}$ in this, then the linearity of $\mathbf{r}$ yields the following set of equations:

$$\tilde{\mathbf{u}} = \sum_i \alpha_i \mathbf{u}_i$$
$$\sum_i \alpha_i \int_{\mathcal{B}} \boldsymbol{w}_j(z) \cdot (\mathbf{L}\tilde{\mathbf{u}}_i)(z)\, dv(z) = \int_{\mathcal{B}} \boldsymbol{w}_j(z) \cdot \mathbf{f}(z)\, dv(z), \qquad 1 \le j \le k$$

If we look closely, we see that the above equation simply is a linear system of size $l \times k$. Let $K \in \mathbb{R}^{l\times k}$, and $f \in \mathbb{R}^k$ be defined by

$$\begin{aligned} K_{ij} &= \int_{\mathcal{B}} \boldsymbol{w}_j(z) \cdot (\mathbf{L}\mathbf{u}_i)(z) dv(z), \quad i = 1, \ldots, l, j = 1, \ldots, k \\ f_j &= \int_{\mathcal{B}} \boldsymbol{w}_j \cdot \mathbf{f}\, dv(z), \quad j = 1, \ldots, k, \end{aligned} \tag{2.29}$$

then the solution with zero weighted residual is given by

$$\tilde{\mathbf{u}} = \sum_{i=1}^{l} \alpha_i \mathbf{u}_i,$$
$$K\alpha = f, \qquad \alpha = (\alpha_1, \ldots, a_l).$$

If W_h is large enough then we could find $\tilde{\mathbf{u}}$, for example by solving the least-squares problem

$$K^T K\alpha = K^T f.$$

The matrix K is often called *stiffness matrix*, especially if the original problem is mechanical. This process of finding the solution by weighting the error ("residual") is called the *weighted residual* method, or *Petrov-Galerkin* process. The functions $\mathbf{u}_j$ are called *shape functions*, and $\boldsymbol{w}_j$ are *trial functions*.

In some applications, the differential operator is self-adjoint: the integral over $\mathcal{B}$ defines an inner product,

$$(\mathbf{x}, \mathbf{y}) = \int_{\mathcal{B}} \mathbf{x}(z) \cdot \mathbf{y}(z)\, dv(z), \qquad \mathbf{x}, \mathbf{y} \in V(\mathcal{B}).$$

If the operator L is self-adjoint, i.e., if $(\boldsymbol{w}, \mathbf{L}\mathbf{u}) = (\mathbf{L}^*\boldsymbol{w}, \mathbf{u}) = (\mathbf{L}\boldsymbol{w}, \mathbf{u})$, then we can take $W_h = V_h$. The weak problem can be translated into a linear system similar to (2.29). In this case, the stiffness matrix will be symmetric. This solution process is called the *Bubnov-Galerkin* method. Additionally, if $\mathbf{L}$ is positive definite, i.e. $(\mathbf{L}\mathbf{u}, \mathbf{u}) \ge 0$ for all $\mathbf{u} \in V(\mathcal{B})$, then so will be the matrix K. Both symmetry and positive-definiteness help in solving the numerical system. For the remainder of this thesis, we assume that $W_h = V_h$.

In many formulations, the residual $\mathbf{r}(\mathbf{u})$ is related to some virtual work functional $\Pi(\cdot)$. For example, the internal energy may equal the work done by the residual force

plus work done by the boundary tractions:

$$\Pi(\mathbf{u}) = \int_{\mathcal{B}} \mathbf{u} \cdot \mathbf{r}(\mathbf{u}) \, dv(z) + \int_{\partial \mathcal{B}} \mathbf{u} \cdot \mathbf{t} \, da(z).$$

In this case, the solution $\mathbf{u}$ that we seek minimizes $\Pi(\mathbf{u})$ over $V(\mathcal{B})$. An approximation of this minimization problem is found again by seeking it in a smaller, finite-dimensional space of V_h. If we are given a basis $\mathbf{u}_1, \ldots, \mathbf{u}_l$ of this space, then the coefficients $\alpha_1, \ldots, \alpha_l$ can be found by solving

$$\min_{\alpha_1, \ldots, \alpha_l} \Pi(\sum_{i=1}^{l} \alpha_i \mathbf{u}_i).$$

This approach is called the *Rayleigh-Ritz* method. We can express the problem as a system of equations by setting the derivative of Π to 0,

$$\frac{\partial \Pi(\sum_i \alpha_i \mathbf{u}_i)}{\partial \alpha_i} = 0, \quad i = 1, \ldots, l. \tag{2.30}$$

If Π is a quadratic form, then its derivative is linear, and (2.30) is exactly the linear system produced by the Bubnov-Galerkin method. Algorithms to solve these systems are discussed in Section 2.4.

We have discussed a general scheme for going from a problem in a space $V(\mathcal{B})$ with infinite dimension to an approximation in a subspace V_h. If we select a V_h and a basis $\mathbf{u}_1, \ldots, \mathbf{u}_l$ to span V_h, we can assemble a system of equations like (2.29) or (2.30).

The support of a function $\boldsymbol{\nu} \in V(\mathcal{B})$, denoted by $\operatorname{supp} \boldsymbol{\nu}$, is the subset of $\mathcal{B}$ where $\boldsymbol{\nu}$ takes on non-zero values. It is advantageous to select shape functions whose supports are as much as possible disjoint. If $\mathbf{u} = \mathbf{0}$ on a region, then we have $\mathbf{L}\mathbf{u} = \mathbf{0}$ on that region, so if $\operatorname{supp} \mathbf{u}_i$ and $\operatorname{supp} \mathbf{u}_j$ are disjoint, then $(\mathbf{u}_i, \mathbf{L}\mathbf{u}_j) = 0$. In other words, having disjoint supports promotes sparsity of the stiffness matrix K.

The most popular approach to obtain V_h—the method that is generally called the 'Finite Element Method'—is to subdivide $\mathcal{B}$ into a collection of geometric primitives such as triangles, quadrilaterals (in 2D) or tetrahedra, hexahedrals or other solids (in 3D). Let the partition be denoted by $\bigcup_j \Omega_j$, and select an integer $p \geq 1$. Every function from V_h is then taken to be a polynomial of degree at most p on every Ω_k. Conditions are set such that all functions have the continuity over the boundaries of each element appropriate for the problem being solved.

There is still ample choice of a basis. Recall that we want to make $\operatorname{supp} \mathbf{u}_j$ as small as possible. One way to enforce this, is to select a set points in the elements, called nodes, that uniquely determine the value of the polynomial on that element. For example, if we take the piecewise linear functions defined on mesh of triangles, then these functions form a finite-dimensional space V_h. The vertices of the triangles form nodes for this finite element basis. Elements that include higher order polynomials, might also have nodes located on the midpoints of edges, or in the interior of the elements.

If we are given function values in each node, the piecewise polynomial is uniquely determined on each element. Since nodes on the boundaries of edges are shared by adjoining elements, continuity of the functions is ensured by construction. Continuity

in the derivatives is not required for the weak problem formulation of the elasticity equations.

By pinpointing nodes in a mesh, we can form a natural basis of V_h, which is called the *nodal* basis. It is the collection of functions that take the value 1 on a single node, and 0 on all other nodes. Such a function is identically zero on all elements that do not contain the non-zero node. The hat-functions from Figure 2.3 form the nodal basis for triangles in 2D.

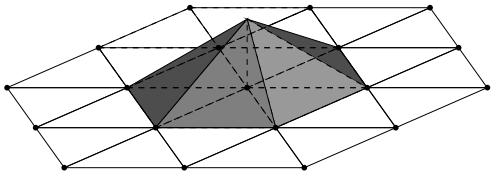

Figure 2.3: Hat functions are a nodal basis for linear triangles

If the derivatives in the differential equation have constant coefficients, i.e. if they are independent of the spatial coordinate $\mathbf{z}$, then the matrices in Equation (2.29) and the derivatives in (2.30) be calculated analytically, so numerical integration methods are not needed.

The relative simplicity of finite elements comes at a price: the difficulty in obtaining a solution is transferred to the problem of subdividing $\mathcal{B}$. The quality of the FEM approximation, as well as the speed and quality of a numerical method depends on the quality of the tessellation of $\mathcal{B}$: subdivision should contain nicely shaped elements. For example, in a triangle subdivision, the triangles should have neither very small nor very large angles [87]. In general, FEM approximation can only be successful if we can generate good meshes for the domain of the problem, and for complex 3D shapes, the task of generating good meshes is a rich source of interesting problems.

We have been sloppy in discussing the traction boundary conditions on purpose; in a displacement-based FE formulation, boundary conditions are added as body forces for boundary nodes, and both can be treated integrally.

2.3 Linear tetrahedra for hyperelastic materials

In the remain of this thesis, we will use linear elements. The displacement function $\mathbf{u}$ is interpolated with a piecewise linear approximation. The interpolation conditions are put at the vertices of the tetrahedral mesh. In other words, we want to have a linear function, and it should map $\mathbf{z}_j \in \mathbb{R}^3$ to $\mathbf{q}_j \in \mathbb{R}^3$ for $j = 1, \ldots, 4$. These conditions uniquely determine a linear affine function. It takes the form

$$\mathbf{z} \mapsto \mathbf{A} \cdot \mathbf{z} + \mathbf{b}.$$

The following function satisfies our interpolation requirements

$$z \mapsto \mathbf{Q} \cdot \mathbf{Z}^{-1} \cdot (z - z_4) + \mathbf{q}_4, \tag{2.31}$$

$$\mathbf{Z} = (z_1 - z_4) \otimes \boldsymbol{e}_1 + (z_2 - z_4) \otimes \boldsymbol{e}_2 + (z_3 - z_4) \otimes \boldsymbol{e}_3, \tag{2.32}$$

$$\mathbf{Q} = (\mathbf{q}_1 - \mathbf{q}_4) \otimes \boldsymbol{e}_1 + (\mathbf{q}_2 - \mathbf{q}_4) \otimes \boldsymbol{e}_2 + (\mathbf{q}_3 - \mathbf{q}_4) \otimes \boldsymbol{e}_3. \tag{2.33}$$

Here, $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ is taken to be an orthonormal base of $\mathbb{R}^n$.

Where the tensor product $\mathbf{a} \otimes \mathbf{b}$ is the 2-tensor defined by $(\mathbf{a} \otimes \mathbf{b}) \cdot \mathbf{x} = \mathbf{a}(\mathbf{b} \cdot \mathbf{x})$. The derivative of this function with respect to z is constant across an element, and is given by

$$\mathbf{Q} \cdot \mathbf{Z}^{-1}. \tag{2.34}$$

The deformations are given in terms of displacements at every node of the mesh, so in effect, we use a linear interpolation for the displacements. For such an interpolation, we can interpret the derivatives of the elastic energy as elastic forces concentrated at nodes of the mesh. We can view the elastic force in a single node as the compound result of the elastic forces of individual tetrahedra incident with that node.

We take a hyperelastic model as a starting point for deriving stresses. This means that we should choose a function $W(\mathbf{C})$, and find its derivative to obtain the stress:

$$\mathbf{T} = \mathbf{F} \cdot \mathbf{S} = \mathbf{F} \cdot \frac{\partial W}{\partial \mathbf{C}}. \tag{2.35}$$

We can formulate weak equations of motion from the strong version presented in Equation (2.5). Let $\boldsymbol{w}$ be a test function, then in every point of $\mathcal{B}$ we have

$$(\operatorname{div} \mathbf{T}) \cdot \boldsymbol{w} + \mathbf{f} \cdot \boldsymbol{w} = \rho \mathbf{p}_{tt} \cdot \boldsymbol{w}. \tag{2.36}$$

Since the divergence satisfies

$$\operatorname{div}(\boldsymbol{A} \cdot \mathbf{b}) = \boldsymbol{A} : \operatorname{grad} \mathbf{b} + \operatorname{div}(\boldsymbol{A}) \cdot \mathbf{b}, \qquad \boldsymbol{A} \in \operatorname{Lin}, \mathbf{b} \in \mathbb{R}^3,$$

we can integrate (2.36) over $\mathcal{B}$ and obtain

$$\int_{\mathcal{B}} \operatorname{div}(\mathbf{T} \cdot \boldsymbol{w}) \, \mathrm{d}\nu(z) - \int_{\mathcal{B}} \mathbf{T} : \operatorname{grad} \boldsymbol{w} \, \mathrm{d}\nu(z) + \int_{\mathcal{B}} \mathbf{f} \cdot \boldsymbol{w} \, \mathrm{d}\nu(z) = \int_{\mathcal{B}} \rho \mathbf{p}_{tt} \cdot \boldsymbol{w} \, \mathrm{d}\nu(z).$$

The first term may be rewritten using Green's theorem from Equation (A.7), yielding

$$\int_{\partial \mathcal{B}} \mathbf{t} \cdot \boldsymbol{w} \, \mathrm{d}\nu(z) - \int_{\mathcal{B}} \mathbf{T} : \operatorname{grad} \boldsymbol{w} \, \mathrm{d}\nu(z) + \int_{\mathcal{B}} \mathbf{f} \cdot \boldsymbol{w} \, \mathrm{d}\nu(z) = \int_{\mathcal{B}} \rho \mathbf{p}_{tt} \cdot \boldsymbol{w} \, \mathrm{d}\nu(z). \tag{2.37}$$

If we discretize the shape functions and test functions with the same finite element space, then the second term may be computed element by element. For a linearly interpolated tetrahedron $\mathbf{F}$ (and therefore $\mathbf{T}$) and $\operatorname{grad} \boldsymbol{w}$ are constant across a tetrahedron, so we have

$$\int_{\tau} \mathbf{T} : \operatorname{grad} \boldsymbol{w} \, \mathrm{d}\nu(z) = \nu(\tau)(\mathbf{T} : \operatorname{grad} \boldsymbol{w}),$$

where $\nu(\tau)$ is the volume of τ in the reference configuration.

Let $\mathbf{q}$ be an element of a nodal basis: $\mathbf{q} = \boldsymbol{e}_k$ in node j of the tetrahedron and 0 in others. According to (2.34), we have $\operatorname{grad} \mathbf{q} = \mathbf{Q} \cdot \mathbf{Z}^{-1}$. We get

$$\begin{aligned} \mathbf{T} : \operatorname{grad} \mathbf{q} &= \mathbf{T} : \mathbf{Q}\mathbf{Z}^{-1} \\ &= \mathbf{T} \cdot \mathbf{Z}^{-*} : \mathbf{Q} \\ &= \begin{cases} \mathbf{T} \cdot \mathbf{Z}^{-*} : \boldsymbol{e}_k \otimes \boldsymbol{e}_j & j \leq 3 \\ -\mathbf{T} \cdot \mathbf{Z}^{-*} : \boldsymbol{e}_k \otimes (\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3) & j = 4. \end{cases} \end{aligned}$$

Here $(\mathbf{Z}^{-1})^*$ is denoted with $\mathbf{Z}^{-*}$. We can interpret this as follows: the elastic forces $\mathbf{f}_1, \ldots, \mathbf{f}_3$ on nodes $1, 2$ and 3 are given by the columns of the matrix representation of

$$-\nu(\tau)(\mathbf{T} \cdot \mathbf{Z}^{-*}) \tag{2.38}$$

with respect to the unit basis. The elastic force on node 4 is $-\mathbf{f}_1 - \mathbf{f}_2 - \mathbf{f}_3$, which implies that the tetrahedron is in equilibrium.

When we assume a body made of hyperelastic material in equilibrium, then we can put $\mathbf{p}_{tt} = 0$. The left hand side of Equation (2.37) is the directional derivative of the potential energy of the system. If we set $\hat{W}(\boldsymbol{z}) = W(\mathbf{C}(\boldsymbol{z}), \boldsymbol{z})$, then that equation is equivalent to

$$\frac{\partial}{\partial \varepsilon} \left(\int_{\mathcal{B}} \hat{W}(\mathbf{p} + \varepsilon \boldsymbol{w}) \, dv - \int_{\mathcal{B}} (\mathbf{p} + \varepsilon \boldsymbol{w}) \cdot \mathbf{f} \, dv - \int_{\partial \mathcal{B}} (\mathbf{p} + \varepsilon \boldsymbol{w}) \cdot \mathbf{t} \, dv \right) = 0. \tag{2.39}$$

In other words, the potential energy ("virtual work") is stationary in $\mathbf{p}$, since moving by an infinitesimal motion $\varepsilon \boldsymbol{w}$ does not change the potential energy of the system.

2.4 Unconstrained optimization

Equation (2.39) describes the solution of a static problem as the stationary point of a virtual work function. Assuming that these stationary points are stable, it follows that we can find equilibrium solutions to mechanical problems by finding the minimum of a virtual work function Π. Inspired by Equation (2.39), we define the potential energy of the system discretized by a FEM basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ as follows.

$$\begin{gathered} \Pi(\alpha_1, \ldots, \alpha_n) = \int_{\mathcal{B}} \hat{W}(\tilde{\mathbf{u}}) \, dv - \int_{\mathcal{B}} (\tilde{\mathbf{u}}) \cdot \mathbf{f} \, dv - \int_{\partial \mathcal{B}} (\tilde{\mathbf{u}}) \cdot \mathbf{t} \, dv, \\ \text{where } \tilde{\mathbf{u}} = \sum_{j=1}^{n} \alpha_j \mathbf{u}_j. \end{gathered} \tag{2.40}$$

The static deformation problem may simply by formulated as finding the solution to

$$\min_{x \in \mathbb{R}^n} \Pi(x).$$

The field of unconstrained optimization studies the algorithms that are used to solve such systems. In this section we discuss three algorithms. The most basic problem is a quadratic Π, and the standard algorithm is the Conjugate Gradient algorithm. For more complex functions, the nonlinear Conjugate Gradient algorithm and Truncated Newton methods may be used, which are discussed in the following sections.

2.4.1 Linear Conjugate Gradient method

For linear elasticity problems, the stiffness matrix K is symmetric positive definite, and for such problems, the most popular optimization method to use is the conjugate gradient method (dubbed CG throughout this thesis). Suppose that we have the following functional $\Pi : \mathbb{R}^n \to \mathbb{R}$,

$$\Pi(x) = \frac{1}{2} x^T K x - b^T x, \tag{2.41}$$

$$\min_{x \in \mathbb{R}^n} \Pi(x), \tag{2.42}$$

where $K \in \mathbb{R}^{n \times n}$ is a given symmetric positive definite matrix, and $b \in \mathbb{R}^n$ a given vector. This corresponds with the potential energy in a linear elastic FEM problem. The minimum is given by an x for which the gradient $\partial \Pi / \partial x = Kx - b$ vanishes. The direction of steepest descent is the negative gradient $b - Kx$, which we call *residual*. It is denoted by r.

We define the following inner product and associated norm on $\mathbb{R}^n$. Let $H \in \mathbb{R}^{n \times n}$ be a positive definite symmetric matrix, then let

$$(x, y) = x^T y, \qquad (x, y)_H = (x, Hy), \tag{2.43}$$

$$\|x\| = \sqrt{(x, y)}, \qquad \|x\|_H = \sqrt{(x, y)_H}. \tag{2.44}$$

The norm $\| \cdot \|_K$ is also called the energy norm.

One basis for an iterative algorithm is the line-search model: starting from some approximation $x_k \in \mathbb{R}^n$, we obtain a new solution x_{k+1} by selecting a search direction $d_k \in \mathbb{R}^n$, and searching in that direction, i.e.

$$x_{k+1} = x_k + \alpha_k d_k,$$

The choice for α_k more or less follows from the search direction d_k chosen, so an iterative optimization algorithm is characterized by its choice for d_k.

The conjugate gradient algorithm is the most popular algorithm for the case that K is positive definite. It computes the search direction d_k as a combination of the last search direction d_{k-1} and the current residual. The CG algorithm is given by the following pseudo code, which assumes that a starting solution x_0 is known.

$k \leftarrow 0$
$r_0 \leftarrow b - Kx_0$
while $\|r_k\|$ too large:
  if $k = 0$:
    $\beta_k \leftarrow 0$
  else:
    $\beta_k \leftarrow \|r_k\|^2 / \|r_{k-1}\|^2$
  $d_k \leftarrow r_k + \beta_k d_{k-1}$
  $\alpha_k \leftarrow \|r_k\|^2 / (d_k, Kd_k)$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$

$r_{k+1} \leftarrow r_k - \alpha_k K d_k$
$k \leftarrow k + 1$

Notice that we can find α_k and update the residual using only one matrix-vector product. The vectors are indexed for clarity, but it is not necessary to store the vectors for each step separately. The CG algorithm can be implemented with four vectors of storage.

The convergence characteristics of CG is a well-researched [97]. Here, we sketch the convergence analysis of CG following the exposition by Axelsson and Barker [6]. The sequence of r_k, d_k and x_k generated by this algorithm satisfies the following properties:

- The residuals are orthogonal: if $i \neq j$, then $(r_i, r_j) = 0$.
- The search directions are K-orthogonal, or K-*conjugate*: if $i \neq j$ then $(d_i, d_j)_K = 0$.
- The sequence of r_k is optimal in the following sense: let

  $$W_k := \operatorname{span}\left\{K^1 r_0, \ldots, K^k r_0\right\},$$

  then $\|r_k\|_{K^{-1}} = \min_{r \in r_0 + W_k} \|r\|_{K^{-1}}$, or equivalently, x_k minimizes $\|Kx - b\|_{K^{-1}}$ over $x \in x_0 + K^{-1} W_k$.

  Let $\hat{x}$ be the exact solution to the problem, then $\|x_k - \hat{x}\|_K = \|r_k\|_{K^{-1}}$, so CG minimizes the solution error in the energy norm in each step.
- In exact arithmetic, the algorithm converges in at most n steps, otherwise the set $\{r_0, \ldots, r_n\}$ would be a set of $n + 1$ orthogonal non-zero vectors.

Since $K^j r_0$ is a polynomial of K applied to r_0, the space W_k can be written in terms of polynomials. Let $\mathbb{P}_l$ be the space of polynomials of degree at most l, then we have

$$W_k = \{\, q(K) K r_0 : q \in \mathbb{P}_{k-1} \,\}.$$

Similarly, we can rephrase the fact that r_k is chosen optimally from $r_0 + W_k$ as follows:

$$\|r_k\|_{K^{-1}} = \min_{r \in r_0 + W_k} \|r\|_{K^{-1}} \tag{2.45}$$

$$= \min_{p \in \mathbb{P}_k, P(0)=1} \|P(K) r_0\|_{K^{-1}}. \tag{2.46}$$

The matrix K is symmetric and positive definite, so it has orthonormal eigenvectors v_i for $i = 1, \ldots, n$ with eigenvalues $0 < \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$. We can expand r_0 in eigenvectors,

$$r_0 = \sum_i (v_i, r_0) v_i,$$

and rewrite (2.46) as

$$\min_{p \in \mathbb{P}_k, p(0)=1} \sqrt{\sum_i (v_i, r_0)^2 p(\lambda_i)^2 / \lambda_i}.$$

If we can bound $|p(\lambda_j)|$ for $j = 1, \ldots, n$, then this minimum is also bounded: suppose that $|p(\lambda_j)| \leq M$ for $j = 1, \ldots, n$, then

$$\sqrt{\sum_i (v_i, r_0)^2 p(\lambda_i)^2 / \lambda_i} \leq M \sqrt{\sum_i (v_i, r_0)^2 / \lambda_i} = M \|r_0\|_{K^{-1}}.$$

In the general case, the eigenvalues of K are distributed in the interval $[\lambda_1, \lambda_n]$, and Chebychev polynomials give a bound on $|p(\lambda)|$ for $\lambda \in [\lambda_1, \lambda_n]$. This leads to the following estimate of $\|r_k\|_{K^{-1}}$:

$$\|r_k\|_{K^{-1}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|r_0\|_{K^{-1}}, \qquad \kappa = \lambda_n / \lambda_1.$$

If we want to improve the starting solution x_0 by a factor ε, then the number of steps necessary is less than

$$\frac{1}{2} \sqrt{\kappa} \ln \frac{2}{\varepsilon} + 1, \qquad \kappa = \lambda_n / \lambda_1. \tag{2.47}$$

The number κ is the condition number of the matrix K, which also bounds the accuracy of a solution computed with finite precision arithmetic. A method to speed up the CG iteration is *preconditioning*. This method transforms the iteration so it runs on the related problem

$$\begin{aligned} E^{-1} K E^{-T} y &= c, \\ c &= E^{-1} b, \\ y &= E^T x. \end{aligned}$$

The condition number for this problem is $\mathrm{cond}(E^{-1} K E^{-T}) = \mathrm{cond}((EE^T)^{-1} K)$. If this condition number is lower than $\mathrm{cond}(K)$ then solving this problem requires less iterations. During every step, a system of the form $(EE^T)x = y$ must be solved, so if (EE^T) has a sufficiently simple form, then this will improve the performance of the CG algorithm.

If K has less than n eigenvalues, say $k < n$, then a kth degree polynomial q exists such that $q(\lambda_j) = 0$. The optimal r will be null, and $\|r_k\|_{K^{-1}} = 0$; in other words: the algorithm will converge within k steps. This observation illustrates that the exact distribution of the eigenvalues of K plays a large role in the convergence behavior. We give a final example which will become relevant in Chapter 4. Suppose that all eigenvalues but the m largest of K are bounded by some constant γ, then we have

$$\|r_k\|_{K^{-1}} \leq 2 \left(\frac{\sqrt{\kappa'} - 1}{\sqrt{\kappa'} + 1} \right)^{k-m} \|r_0\|_{K^{-1}},$$

and where $\kappa' = \gamma / \lambda_1$. The number of steps necessary for a factor ε reduction is less than

$$\frac{1}{2} \sqrt{\kappa'} \ln \left(\frac{2}{\varepsilon} \right) + m + 1.$$

In other words, the magnitude of a few extremely large but isolated eigenvalues does not affect the performance of CG.

2.4.2 Non-linear conjugate gradients

The basic iteration of CG consists of three steps: determining a new search direction, determining the optimal step, and finding the residual in the new point. There is nothing inherently quadratic about the last two steps, so if some search direction is given, then the CG algorithm can also be performed for non-quadratic Π. We refer to this algorithm as the nonlinear CG algorithm. It can be expressed as follows.

$k \leftarrow 0$
$r_0 \leftarrow -\operatorname{grad}\Pi(x_0)$
while $\|r_k\|$ is too large:
  select d_k
  find α_k such that $\Pi(x_k + \alpha_k d_k)$ minimal
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $r_{k+1} \leftarrow -\operatorname{grad}\Pi(x_{k+1})$

For selecting d_k, the conjugate gradient algorithm uses

$$d_k = \begin{cases} r_k, & k = 0, \\ r_k + \beta_k d_{k-1}, & k > 0. \end{cases}$$

For selecting β_k in the nonlinear case, a number of different recipes have been proposed: Fletcher-Reeves,

$$\beta_{FR} = \frac{\|r_k\|^2}{\|r_{k-1}\|}, \tag{2.48}$$

and Polak-Ribière

$$\beta_{PR} = \max\{0, \frac{r_k^T(r_k - r_{k-1})}{\|r_{k-1}\|}\}. \tag{2.49}$$

The convergence theory of linear CG algorithms is fairly complete. In contrast, little is known on the convergence for non-quadratic Π. Both β-selection strategies are equivalent with linear CG when applied to a quadratic Π, and they are known to converge to a stationary point if the line search for α_k is sufficiently precise [75]. In practice, Polak-Ribière is known to perform better than Fletcher-Reeves. This has been attributed to the fact that an unsuccessful Polak-Ribière step yields $\beta_{PR} \approx 0$, which in effect resets the search direction to the direction of steepest descent. However, a convincing explanation of the success of Polak-Ribière is still lacking [42].

2.4.3 Truncated Newton

A function attains an unconstrained optimum at a stationary point, i.e. when $\partial\Pi/\partial x = 0$. This is a nonlinear system of equations, and root-finding techniques can be used to solve it. The Newton-Raphson iteration is a classic method for finding roots of equations [41]. Let $K(x)$ denote the Hessian of the energy function Π in the point $x \in \mathbb{R}^n$, i.e.

$$K(x) = \frac{\partial^2\Pi}{\partial u^2}(x). \tag{2.50}$$

The algorithm can then be expressed as follows.

$k \leftarrow 0$
$r_k \leftarrow (\partial\Pi/\partial x)(x_0)$
while r_k is too large:
  solve $K(x_0)d_k = r_k \qquad (*)$
  $x_{k+1} \leftarrow x_k + d_k.$
  $r_{k+1} \leftarrow (\partial\Pi/\partial x)(x_{k+1})$

The solution step in $(*)$ can be done by a linear CG algorithm if $K(x_0)$ is a positive definite matrix. For elasticity problems, this is normally the case around stable equilibrium solutions. Newton-Raphson with CG as an inner loop is also called the *Truncated Newton* or *Truncated Conjugate Gradient* method. The precision of the solution for the inner loop is of minor importance, so low tolerances can be used [42].

The process has a quadratic convergence, i.e. if $\hat{x}$ is the exact solution to a problem, then

$$\|\hat{x} - x_{k+1}\| \leq c\|\hat{x} - x_k\|^2,$$

for some positive constant c: the process converges superlinearly. This is attractive when high precision of the solution is required: close to a solution, the number of correct digits doubles at each step. The Newton-Raphson iteration is characteristic of superlinear convergence: any algorithm that converges superlinearly must have search directions that approximate Newton search directions [33]. In Chapter 4, the Truncated Newton method will be used to compute solutions with high precision.

2.5 Time integration

Equation (2.37) involves $\mathbf{p}_{tt}$, so it is time-dependent, and $\mathbf{u}$ also depends on time. When an approximation $\tilde{\mathbf{u}}$ of $\mathbf{u}$ is expanded in shape functions $\boldsymbol{w}_i$, for $i = 1, \ldots, n$, then the coefficients are also time dependent:

$$\tilde{\mathbf{u}}(z, t) = \sum_i u_i(t)\boldsymbol{w}_i(z).$$

The Finite Element method transforms the PDE into a system of ordinary differential equations (ODEs), and the solution process requires us to find the evolution of u over time:

$$s(u(\cdot), t) + f^{ex}(t) = M\ddot{u}. \tag{2.51}$$

The function $f^{ex}(t)$ represents body forces and tractions combined, and the $\mathbb{R}^{n\times n}$ matrix M is called *mass matrix*. For shape functions $(\boldsymbol{w}_j)_j$, it follows from

$$M_{ij} = \int_{\mathcal{B}} \boldsymbol{w}_i \cdot \boldsymbol{w}_j \rho \, dv. \tag{2.52}$$

The function s in (2.51) represents internal forces within the material. In a general mechanical formulation these forces may depend on the history of u. This leads to a *viscoelastic* problem. In hyperelastic materials internal stresses do not depend on history of u. Internal forces are elastic forces f^{el}, that depend on u at time t. Energy is generally dissipated by adding a friction term f^{fr} that depends on $\dot{u}$, yielding

$$f^{el}(u(t)) + f^{fr}(\dot{u}(t)) + f^{ex}(t) = M\ddot{u} \tag{2.53}$$

For the linear case (where both elastic and frictional forces are linear in u), a closed form analytic solution to (2.53) exists. For large problems or nonlinear problems, computing that solution is not possible or practical. In these cases, numerical methods must be used. A *numerical integration scheme* or *time integration scheme* computes the evolution of an approximate solution numerically. The process proceeds by advancing the time variable by an increment Δt, called the *time step*. By using Equation (2.51), the configuration at time $t + \Delta t$ is estimated given the situation at time t. If the recipe specifies $u(t+\Delta t)$ and $\dot{u}(t+\Delta t)$ as an unknown in a system of equations involving $f^{el}(u(t+\Delta t)$ then we call the method *implicit*. An *explicit* method uses the forces at time t to predict $u(t + \Delta t)$.

Computing the next result in an implicit method involves solving a large system, which is costly. Advancing a time step in an explicit methods requires much less calculations. The price paid for this simplicity is *conditional stability*. If the dynamic system includes phenomena which evolve quicker than the approximate solution itself, these will be mistaken for exponentially increasing components of the solution. This is called *instability*, and typically results in blow-up of the solution. Conditional stability for mechanical problems is expressed through the Courant-Friedrichs-Lewy criterion: the time step must be smaller than the critical time step Δt_{crit}:

$$\Delta t \leq \Delta t_{crit} \sim h/c \tag{2.54}$$

Here c is the wave speed in the medium, and h the element size. The quantity h/c is the time that a wave needs to propagate across an element of size h. The proportionality constant depends on the problem and the integration scheme used.

During the rest of the discussion, we will consider the linearized FEM equations. These can be specified as

$$M\ddot{u} + C\dot{u} + Ku - f^{ex} = 0, \tag{2.55}$$

The $\mathbb{R}^{n\times n}$ matrix C is the damping matrix. Measuring physically realistic values for C is hard, therefore C is set often set to a linear combination of M and K. This is called *Rayleigh-damping*.

We follow Chapter 17 from Zienkiewicz and Taylor [103] for discussing popular second order schemes for integrating FEM systems. These are the so-called Generalized Newmark (GN) methods and the related *weighted residual* (denoted by SS). Both approaches expand u in a truncated Taylor series in t, and estimate the highest order term using the differential equation. Both methods have similar precision and stability properties, and both can be formulated for j-th order problems, estimating the Taylor series up to the p-th derivative; the methods are denoted as SSpj (for the weighted residual form) and GNpj (for the Generalized Newmark).

The SS22 method is applicable to second order ODEs resulting from a FEM discretization. It estimates the coefficients of a Taylor series expansion of u up to order 2 around the chosen time t, and the accumulated error in u is $\mathcal{O}(\Delta t)$. It starts with estimating weighted averages of u and its derivative at time t

$$\bar{u} = u(t) + \theta_1 \Delta t \dot{u}(t), \tag{2.56}$$

$$\bar{\dot{u}} = \dot{u}. \tag{2.57}$$

The number θ_1 is a weighting parameter. The differential equation produces an equation that estimates the second derivative, weighed by θ_2,

$$(M + \theta_1 \Delta t C + K\frac{1}{2}\theta_2 \Delta t^2)\bar{\ddot{u}} + (C\bar{\dot{u}}) + K\bar{u} + \bar{f} = 0. \tag{2.58}$$

The estimated derivatives can then be used to find $u(t + \Delta t)$

$$u(t + \Delta t) = u(t) + \Delta t \dot{u}(t) + \frac{1}{2}\Delta t^2 \bar{\ddot{u}}, \tag{2.59}$$

$$\dot{u}(t + \Delta t) = \dot{u}(t) + \Delta t \bar{\ddot{u}}. \tag{2.60}$$

The properties of this method are controlled by the values of θ_1 and θ_2. If $\theta_2 = 0$, then we call the method explicit. This explicit method is only stable on the condition that $\theta_1 \geq 1/2$. When $\theta_1 > \frac{1}{2}$, then the $\Delta t_{crit} = \mathcal{O}(h^2)$, where h is the shortest edge length. This is undesirable, since small h values are necessary to obtain an accurate discretization. If $\theta_1 = \frac{1}{2}$, then

$$\Delta t_{crit} = \frac{2}{\sqrt{3}}\frac{h}{c}, \tag{2.61}$$

where $c = \sqrt{\frac{\lambda+2\mu}{\rho}}$ is the wave speed in the medium.

The central difference method of integration has exactly the same stability and error properties as the explicit SS22 method with $\theta_1 = 1/2$. Instead of adding the velocity as an extra variable, it adds the position at time $t - \Delta t$ as a variable: it is a *multi-step* method. Derivatives in Equation (2.55) are replaced by explicit differences, yielding the equation:

$$M\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta t^2} + C\frac{u_{i+1} - u_{i-1}}{\Delta t} + Ku_i + f_i = 0, \tag{2.62}$$

where $u_k \in \mathbb{R}^n$ is approximation for the value of u at time $t + k\Delta t$. It is less trivial to change the value of the time-step during a simulation.

Both explicit SS22 and central differences are particularly efficient if M and C are diagonal. In this case, the mass and damping are redistributed to be concentrated in the nodes: node j has one quarter of the mass of all the tetrahedrons incident with j. This process is called *mass lumping*. If it is applied to the damping matrix, this is called *lumped damping*. Mass lumping is not strictly realistic, but it has been shown to yield precise results, and enlarges the critical time step. The SS22 scheme with lumped masses and lumped damping for nonlinear elasticity problems is examined in Chapter 4.

2.6 Mechanics of soft tissue

In the rest of the thesis we wil concentrate on compressible hyperelastic aelotropic material models. For living tissue this is a simplification of reality. In this section, we briefly explain what effects are neglected by this simplification. It is partially based on the book by Fung [44].

The term soft tissue includes a variety of tissue types in the body. The mechanical characteristics of this tissue are determined by connective tissue. The materials that contribute to the mechanics of the tissue include the following: *Elastine* is a rubbery biological material. Its loading and unloading cycles are almost equal, meaning that it is almost perfectly elastic. This material is found in elastic tissues, such as skin, artery walls, lung tissue. It also helps keep the skin smooth; elastin production stops after puberty. *Collagen* is a biological construction material. It forms the load bearing material in soft and hard tissue. It is a major component of tendons, bone, skin, and blood vessels. Collagen and elastine can form fibers. If these fibers have some dominant orientation, the tissue will behave differently in different directions. The material then is *aelotropic* or *anisotropic*. The fibers are suspended together with cells in a watery gel called *ground substance*. Since water is incompressible, many tissue types are also (nearly) incompressible.

The relation between load and deformation (stress and strain) follows the path shown in Figure 2.4. The stress response of biological tissue can be divided into three trajectories: at small loads (OA), the stress is exponential in the strain. For larger loads, the stress is linear in the strain (AB). Finally, in the third trajectory (BC), the tissue is almost stressed to failure, and reacts nonlinearly. Normal tissue loads fall into the first region.

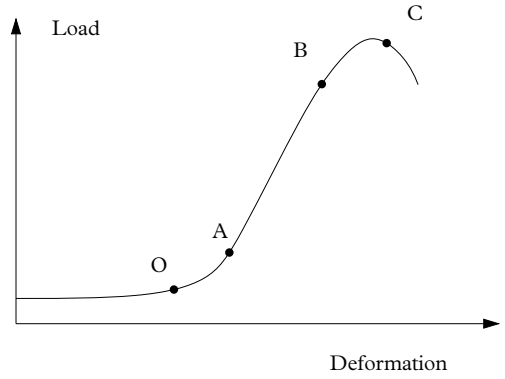

Figure 2.4: Stress response of a rabbit limb tendon, after Fung [44].

When tissue is stretched it offers more resistance than during a following unload. This phenomenon is called *hysteresis*, and it is an example of a viscoelastic effect: stresses in the material depend on the history of the deformation. When tissue is stressed with a constant load, then after the initial elastic response, the tissue will slowly distend further. This process is known as *creep*. A related phenomenon is *stress relax-*

ation: when a tissue specimen is loaded and then held at a constant elongation, stresses within the tissue decrease. This process is rather slow, taking minutes to many hours before a steady state is reached.

When tissue is loaded and then unloaded, its elastic properties change: the tissue becomes softer. When this cyclical loading is repeated often enough, the difference between the cycles disappears, and the deformation converges. The tissue is now said to be *preconditioned*. This processes is illustrated in Figure 2.5.

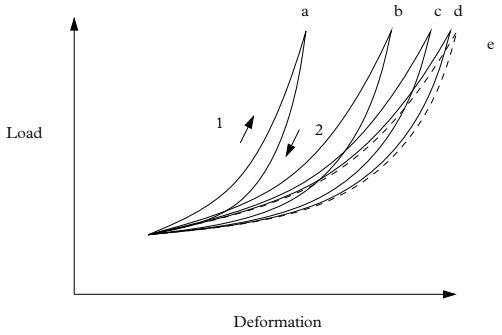

Figure 2.5: During loading (1), tissue offers more resistance than during unloading (2). If such a cycle is repeated, tissue will become softer (a, b, c, d) until the response converges (e).

The first completely 3D description of soft tissue elasticity was given by Veronda and Westman [98]. They used some simplifications to overcome the difficulties posed by the complex mechanical behavior of tissue. They have proposed constitutive equations of soft-tissue on the basis of measurements on cat skin. Anisotropy and viscoelastic effects were handled by measuring only the loading step during a uniaxial stretching test. This gave sufficient experimental data to derive an isotropic hyperelastic constitutive equation. Material parameters were extracted from the same data set by fitting the derived stress/strain curves to the data. The Veronda-Westmann material model will be used in Chapter 4.

A more accurate description is quasi-linear visco-elasticity [44], which accounts for visco-elastic effects by modeling tissue as a superposition of materials with different relaxation times. Kauer [57] used this model to measure tissue elasticity of ex-vivo pig kidney and in- and ex-vivo human uteri. This was done using aspiration experiments: during surgery, a tube was placed over the tissue to be measured. A partial vacuum was created which caused a bulge to form inside the tube. The evolution of this bulge was recorded with a camera, and stored. The conditions of the experiment were repeated in FEM simulation. Material parameters were determined by searching for those settings that recreated the experiment accurately.