Tensors, continuum mechanics, and the finite element method

Alexander Cumberworth

April 14, 2022

Tensors

A tensor can be defined as a multilinear map that takes vectors and covectors and outputs real numbers,

$$\vec{T} : V^* \times \ldots \times V^* \times V \times \ldots \times V \to \mathbb{R}, \tag{1}$$

where V is a vector space, V^* is the dual vector space of V, $V \times V$ is the Cartesian product of two vector spaces, and $\mathbb{R}$ are the real numbers. The order, degree, or rank of a tensor refers to the number of vectors and covectors that a tensor maps from. A vector that is invariant to a change in coordinates is an order-1 tensor, while a similarly invariant scalar is an order-0 tensor. Because the term rank has another meaning in the context of matrices, we will only use order or degree to refer to this property henceforth. Tensors allow physical laws to be expressed in a coordinate independent manner.

There exists a number of different notations for tensors; here we will primarily use Einstein notation, a type of index notation. The components of the tensors in some basis set are indexed by superscripts and subscripts. Superscripts imply that the quantity follows a covariant coordinate transformation law, while subscripts imply the quantity follows a contravariant transformation law. However, in the context of continuum mechanics, the vector spaces are typically Euclidean and expressed with orthonormal basis sets, which allows us to ignore the subtleties of dual vector spaces, and treat everything as a vector. We will do so for the remainder of this document, and follow the convention of writing all indices as subscripts. We can then more narrowly define a tensor of order n as

$$\vec{T} : \underbrace{\mathbb{R}^3 \times \ldots \times \mathbb{R}^3}_{n \text{ times}} \to \mathbb{R}. \tag{2}$$

An order-2 tensor, for example, may be written in terms of its components in some basis $\vec{\hat{e}}$ as $T(\vec{\hat{e}}_i, \vec{\hat{e}}_j) = T_{ij}$. For vectors $\vec{v}$ and $\vec{w}$ as arguments, we can write $T(\vec{v}, \vec{w}) = T(v_i\vec{\hat{e}}_i, w_j\vec{\hat{e}}_j) = v_i w_j T_{ij}$.

Tensor addition extends the addition operation of vectors, such that, without loss of generality, for some order-2 tensors $\vec{A}$, $\vec{B}$, and $\vec{C}$ in some basis,

$$C_{ij} = A_{ij} + B_{ij}. \tag{3}$$

Multiplication by a scalar is also trivially extended to tensors,

$$c\vec{T}(\vec{\hat{e}}_i, \vec{\hat{e}}_j) = cT_{ij}. \tag{4}$$

Transposition of an order-2 tensor can be defined here as switching the order of the arguments to the mapping,

$$T^{\intercal}_{ij} = \vec{T}(\vec{\hat{e}}_i, \vec{\hat{e}}_j)^{\intercal} = \vec{T}(\vec{\hat{e}}_j, \vec{\hat{e}}_i) = T_{ji} \tag{5}$$

Three additional operators can be defined that do not have a clear analogy with vectors. The tensor product combines two tensors $\vec{A}$ and $\vec{B}$ of order n and m, respectively, to produce a tensor $\vec{C}$ of order $n + m$,

$$\vec{C} = \vec{A} \otimes \vec{B}. \tag{6}$$

This allows us to write tensors in terms of tensor products of basis vectors. If $m = n = 2$, then,

$$\begin{aligned}\vec{C} &= (A_{ij}\vec{\hat{e}}_i \otimes \vec{\hat{e}}_j) \otimes (B_{kl}\vec{\hat{e}}_k \otimes \vec{\hat{e}}_l) \\ &= A_{ij}B_{kl}\vec{\hat{e}}_i \otimes \vec{\hat{e}}_j \otimes \vec{\hat{e}}_k \otimes \vec{\hat{e}}_l \\ &= C_{ijkl}\vec{\hat{e}}_i \otimes \vec{\hat{e}}_j \otimes \vec{\hat{e}}_k \otimes \vec{\hat{e}}_l. \end{aligned} \tag{7}$$

Contraction is an operation that allows the order of a tensor to be lowered. Contraction involves pairing two of the input vectors and summing over them. When written in a component-free form, there is no standard notation for the contraction operation; here a contraction between indices i and k for an order-4 tensor $\vec{C}$, without loss of generality, will be denoted as

$$\mathrm{Cont}_{ik}\vec{C} = C_{ijil}\vec{\hat{e}}_j \otimes \vec{\hat{e}}_l. \tag{8}$$

With our assumption of Euclidean space, this pairing can be made more concrete, and becomes the standard dot product between the basis vectors. Partial application of the tensor with respect to one or more of its vector arguments can be done with combinations of the tensor product and contraction operations. For example, if $\vec{T}$ is an order-2 tensor and $\vec{w}$ a vector, $\vec{T}$ can be applied to $\vec{w}$ to produce a vector $\vec{v}$,

$$\begin{aligned}\vec{T}\vec{w} &= \mathrm{Cont}_{jk}[(T_{ij}\vec{\hat{e}}_i \otimes \vec{\hat{e}}_j) \otimes (w_k\vec{\hat{e}}_k)] \\ &= T_{ij}w_k\vec{\hat{e}}_i(\vec{\hat{e}}_j \cdot \vec{\hat{e}}_k) \\ &= T_{ij}w_k\vec{\hat{e}}_i\delta_{jk} \\ &= T_{ij}w_j\vec{\hat{e}}_i \\ &= v_i\vec{\hat{e}}_i \\ &= \vec{v}, \end{aligned} \tag{9}$$

where δ_{jk} is the Kroneker delta placing the tensor. We have explicitly shown contraction as a dot product, and by placing the tensor and vector (and more generally two tensors) directly adjacent we imply a tensor product followed by the contraction of the adjacent pair of indices.

Tensors themselves are elements of a vector space. As we are assuming Euclidean space for the vector spaces that comprise the vector space the tensors are a part of, we can define inner products for these spaces. For order-2 tensors, we can define the inner product the double dot product,

$$\begin{aligned}\langle\vec{A},\vec{B}\rangle &= \vec{A} : \vec{B}\\ &= A_{ij}B_{ij}.\end{aligned} \tag{10}$$

The norm is then

$$\begin{aligned}|\vec{T}| &= \sqrt{\vec{T} : \vec{T}}\\ &= \sqrt{T_{ij}T_{ij}}.\end{aligned} \tag{11}$$

It is also straightforward to define the determinant of order-2 tensors, with

$$\det(\vec{T}) = \epsilon_{ijk}T_{1i}T_{2k}T_{3j}, \tag{12}$$

where ϵ_{ijk} is the Levi-Civita, or permutation, symbol. The trace of an order-2 tensor is simply the contraction of the two indices,

$$\mathrm{tr}(\vec{T}) = T_{ii}. \tag{13}$$

The above operators also extend to the concept of tensor fields. With tensor fields we can also take derivatives with respect to the variables that the tensor depends on. The gradient of a tensor $\vec{T}(\vec{x})$, where $\vec{x}$ is some vector and we continue to use the assumption of Euclidean space and an orthonormal basis, is defined as

$$\nabla\vec{T}(\vec{x}) = \frac{\partial\vec{T}(\vec{x})}{\partial x_i} \otimes \vec{e}_i. \tag{14}$$

For an order-2 tensor,

$$\nabla\vec{T}(\vec{x}) = \frac{\partial T_{ij}}{\partial x_k}\vec{e}_i \otimes \vec{e}_j \otimes \vec{e}_k. \tag{15}$$

The divergence of a tensor, again assuming Euclidean space and an orthonormal basis, is defined as

$$\nabla \cdot \vec{T}(\vec{x}) = \frac{\partial \vec{T}(\vec{x})}{\partial x_i} \cdot \vec{\hat{e}}_i. \tag{16}$$

For an order-2 tensor,

$$\nabla \cdot \vec{T}(\vec{x}) = \frac{\partial T_{ij}}{\partial x_j} \vec{\hat{e}}_i. \tag{17}$$

Note that the gradient operator acts to increase the order of the tensor by one, while the divergence operator acts to decrease the order of the tensor by one. It is not uncommon for the definitions of the gradient and the tensor to be defined such that the tensor product is applied from the left rather than from the right (we have written the divergence as a dot product, but as discussed above this can be expanded as a tensor product followed by a contraction).

A change of coordinates of tensor fields requires some additional considerations. To derive the general formula would require a discussion of differential geometry, so we will simply give the results here. For a vector field $\vec{v}(\vec{x})$, to express the same vector on the field variable $\vec{y} = \vec{y}(\vec{x})$,

$$\begin{aligned}
\vec{v}(\vec{y}) &= \frac{\partial y_i}{\partial x_j} v_j(\vec{x}) \vec{\hat{e}}_i^{\,\prime} \\
&= \left[\frac{\partial y_i}{\partial x_j} \vec{\hat{e}}_i^{\,\prime} \otimes \vec{\hat{e}}_j \right] \left[v_k(\vec{x}) \vec{\hat{e}}_k \right] \\
&= \vec{F} \vec{v}(\vec{x}),
\end{aligned} \tag{18}$$

where $\vec{\hat{e}}_i^{\,\prime}$ are the basis vectors chosen for expressing $\vec{y}$, and $\vec{\hat{e}}_j$ are the basis vectors chosen to express $\vec{x}$. $\vec{F}$ is not necessarily a tensor, but as we restrict ourselves to Cartesian coordinate systems, it does behave as one. For a tensor field, we must apply $\vec{F}$ a number of times equal to the order of the tensor. For an order-2 tensor,

$$\vec{T}(\vec{x}) = \vec{F} \vec{T}(\vec{y}) \vec{F}^{\intercal}. \tag{19}$$

Solid mechanics and elasticity

In continuum mechanics, it is common to refer to two different frames of reference. The first is known as the spatial, or Eulerian description. In the spatial description, quantities are written as functions of the position in the physical space the system exists in. The second is known as the material, or Lagrangian, description. In material description, quantities are instead written as functions of the position in the material being described. To do so, a reference configuration is used, typically the undeformed configuration before any forces have been applied to the system. The spatial description is commonly used in fluid mechanics, as one is typically interested in properties of the fluid at a particular point in space. In contrast, the material description is commonly used in solid mechanics, as one is typically interested in properties of the solid at a particular point of the system, regardless of where it has moved to after deformation in space.

Kinematics is the framework of describing how the system deforms. It is independent from the forces that actually lead to the deformation, in that it is simply a mathematical description of the state of the system. The spatial position vectors, their bases, and their indices will be written in lower case,

$$\vec{x} = x_i \vec{\hat{e}}_i, \tag{20}$$

while the material position vectors, their basis vectors, and their indices will be written in upper case,

$$\vec{X} = X_I \vec{\hat{E}}_I. \tag{21}$$

Here, we will only consider Cartesian coordinates, and so the basis vectors of the material and reference frames will be the same, i.e. $\vec{\hat{e}}_i = \vec{\hat{E}}_i$. The spatial vectors are defined on a domain B, the set of points that the deformed system exists in, while the material vectors are defined on a domain B_0, the set of points that the reference configuration exists in. We can write the spatial vectors as a function of the material vectors, and vice versa with the deformation mapping, where by deformation mapping we mean $\phi(\vec{X}) = \vec{x}(\vec{X})$ or its inverse $\phi^{-1}(\vec{x}) = \vec{X}(\vec{x})$. The deformation of the system in the material reference can be described by a deformation vector field,

$$\begin{aligned} \vec{u}(\vec{X}) &= \vec{x}(\vec{X}) - \vec{X} \\ \vec{u}(\vec{X}) &= [x_I(\vec{X}) - X_I]\vec{\hat{E}}_I, \end{aligned} \tag{22}$$

while in the spatial reference it can be described as

$$\begin{aligned}\vec{u}(\vec{x}) &= \vec{x} - \vec{X}(\vec{x}) \\ \vec{u}(\vec{x}) &= [x_i - X_i(\vec{x})]\vec{\hat{e}}_i,\end{aligned} \tag{23}$$

although we will not use this form.

While here we do use the same basis set in both the material and spatial reference, we will continue to differentiate between the two. This is because we will generally assume the field quality of many variables, writing them without function notation, and use the convention of writing the basis vectors to reflect whether quality is in the material or spatial reference frame. To be clear, it is important to keep track of whether a field is a directly a function of $\vec{x}$ or $\vec{X}$, rather than indirectly through a deformation mapping. For example, there is a difference between $\vec{u}(\vec{X} = \vec{x}) = \vec{x}(\vec{x}) - \vec{x}$, which is not correct, and $\vec{u}(\vec{X}[\vec{x}]) = \vec{x}(\vec{X}[\vec{x}]) - \vec{X}(\vec{x})$, in which we write the deformation as a direct function $\vec{X}$ but use $\vec{x}$ as the independent variable.

The deformation gradient is a tensor field of both $\vec{X}$ and $\vec{x}$, and is defined as

$$\begin{aligned}\vec{F} &= \nabla_0 \vec{x} \\ &= \frac{\partial x_i}{\partial X_J} \vec{\hat{e}}_i \otimes \vec{\hat{E}}_J \\ &= F_{iJ} \vec{\hat{e}}_i \otimes \vec{\hat{E}}_J,\end{aligned} \tag{24}$$

where ∇_0 is the gradient with respect to the material coordinates. Note that the deformation tensor is also the tensor defined in eq. (18) for the general changing of coordinates of the underlying field. This means that it can be used to convert a field from a material reference frame to spatial reference frame. The inverse deformation gradient,

$$\vec{F}^{-1} = \frac{\partial X_I}{\partial x_j} \vec{\hat{E}}_I \otimes \vec{\hat{e}}_j, \tag{25}$$

can be used to convert from a spatial reference frame to a material reference frame. The deformation gradient can written in terms of the deformation with simple substitution,

$$
\begin{aligned}
\vec{F} &= \nabla_0(\vec{u} + \vec{X}) \\
&= \nabla_0\vec{u} + \nabla_0\vec{X} \\
&= \frac{\partial u_i}{\partial X_J}\vec{\hat{e}}_i \otimes \vec{\hat{E}}_J + \frac{\partial X_I}{\partial X_J}\vec{\hat{E}}_I \otimes \vec{\hat{E}}_J \\
&= \frac{\partial u_i}{\partial X_J}\vec{\hat{e}}_i \otimes \vec{\hat{E}}_J + \delta_{IJ}\vec{\hat{E}}_I \otimes \vec{\hat{E}}_J \\
&= \nabla_0\vec{u} + \vec{I},
\end{aligned}
\tag{26}
$$

where $\vec{I}$ is the order-2 identity tensor. The deformation tensor can also be considered as a map between infinitesimal position vectors in the spatial and material description,

$$
\mathrm{d}\vec{x} = \vec{F}\mathrm{d}\vec{X}. \tag{27}
$$

It is also important to be able to describe how distances between points in the material and spatial descriptions are related. If ds and dS are infinitesimal distances in the spatial and material description, respectively, then

$$
\begin{aligned}
(\mathrm{d}s)^2 &= \mathrm{d}\vec{x}\mathrm{d}\vec{x} \\
&= (\vec{F}\mathrm{d}\vec{X})(\vec{F}\mathrm{d}\vec{X}) \\
&= (\mathrm{d}\vec{X}\vec{F}^\intercal)(\vec{F}\mathrm{d}\vec{X}) \\
&= \mathrm{d}\vec{X}(\vec{F}^\intercal\vec{F})\mathrm{d}\vec{X} \\
&= \mathrm{d}\vec{X}([F_{iJ}\vec{\hat{E}}_J \otimes \vec{\hat{e}}_i][F_{kL}\vec{\hat{e}}_k \otimes \vec{\hat{E}}_L])\mathrm{d}\vec{X} \\
&= \mathrm{d}\vec{X}(F_{iJ}F_{kL}[\vec{\hat{e}}_i \cdot \vec{\hat{e}}_k]\vec{\hat{E}}_J \otimes \vec{\hat{E}}_L)\mathrm{d}\vec{X} \\
&= \mathrm{d}\vec{X}(F_{iJ}F_{kL}\delta_{ik}\vec{\hat{E}}_J \otimes \vec{\hat{E}}_L)\mathrm{d}\vec{X} \\
&= \mathrm{d}\vec{X}(F_{iJ}F_{iL}\vec{\hat{E}}_J \otimes \vec{\hat{E}}_L)\mathrm{d}\vec{X} \\
&= \mathrm{d}\vec{X}\vec{C}\mathrm{d}\vec{X},
\end{aligned}
\tag{28}
$$

where $\vec{C}$ is known as the right Cauchy-Green tensor, and

$$\begin{aligned}
(\mathrm{d}s)^2 - (\mathrm{d}S)^2 &= \mathrm{d}\vec{X}\vec{C}\mathrm{d}\vec{X} - \mathrm{d}\vec{X}\mathrm{d}\vec{X} \\
&= \mathrm{d}\vec{X}(\vec{C}\mathrm{d}\vec{X} - \mathrm{d}\vec{X}) \\
&= 2\mathrm{d}\vec{X}\left[\frac{1}{2}(\vec{C} - \vec{I})\right]\mathrm{d}\vec{X} \\
&= 2\mathrm{d}\vec{X}\vec{E}\mathrm{d}\vec{X}, \qquad (29)
\end{aligned}$$

where $\vec{E}$ is the Green-Lagrange strain tensor. The Green-Lagrange strain tensor can be expressed directly as a function of the deformation field,

$$\begin{aligned}
\vec{E} &= \frac{1}{2}(\vec{C} - \vec{I}) \\
&= \frac{1}{2}(\vec{F}^{\intercal}\vec{F} - \vec{I}) \\
&= \frac{1}{2}\left[(\nabla_0\vec{u} + \vec{I})^{\intercal}(\nabla_0\vec{u} + \vec{I}) - \vec{I}\right] \\
&= \frac{1}{2}\left[\nabla_0\vec{u} + (\nabla_0\vec{u})^{\intercal} + (\nabla_0\vec{u})^{\intercal}(\nabla_0\vec{u})\right] \\
&= \frac{1}{2}\left(\frac{\partial u_I}{\partial X_J} + \frac{\partial u_J}{\partial X_I} + \frac{\partial u_K}{\partial X_I}\frac{\partial u_K}{\partial X_J}\right)\hat{\vec{E}}_I \otimes \hat{\vec{E}}_J. \qquad (30)
\end{aligned}$$

The Green-Lagrange strain tensor contains a term that is nonlinear in $\vec{u}$. If the strain is small enough, or $|\nabla_0\vec{u}| << 1$, the infinitesimal strain assumption may be used to linearize the tensor. Under this assumption, the Green-Lagrange strain tensor is approximately equal to the infinitesimal strain tensor,

$$\begin{aligned}
\epsilon &= \frac{1}{2}\left[\nabla_0\vec{u} + (\nabla_0\vec{u})^{\intercal}\right] \\
&= \frac{1}{2}\left(\frac{\partial u_I}{\partial X_J} + \frac{\partial u_J}{\partial X_I}\right)\hat{\vec{E}}_I \otimes \hat{\vec{E}}_J. \qquad (31)
\end{aligned}$$

The difference between taking the gradient with respect to the material or the spatial derivatives also becomes insignificant, so $\nabla_0\vec{u} \approx \nabla\vec{u}$.

A description of the forces on the system begins with a definition of the stress vector field $\vec{t}^{\hat{n}}(\vec{x})$ in the spatial reference frame, which is the force $\vec{f}^{\hat{n}}(\vec{x})$ per area a for a particular surface with normal vector $\vec{n}$,

$$\vec{t}^{\hat{n}}(\vec{x}) = \lim_{\Delta a \to 0} \frac{\Delta \vec{f}^{\hat{n}}(\vec{x})}{\Delta a}. \tag{32}$$

Stress vectors defined on boundary surfaces of the system are known as traction vectors. The Cauchy stress tensor $\vec{\sigma}$ is an order-2 tensor defined such that

$$\vec{t}^{\hat{n}} = \vec{\sigma}\vec{n}. \tag{33}$$

In other words, partial application of the tensor to the surface normal maps to the stress vector on that surface. The stress tensor field contains all the stress information for the system. The Cauchy stress tensor can also be written as the sum of the tensor product of the stress vectors on the surfaces normal to the basis vectors,

$$\vec{\sigma} = \vec{t}^{\hat{e}_i} \otimes \vec{e}_i. \tag{34}$$

The stress vector can can be decomposed into the sum of the normal stress $\vec{t}^{\hat{n}}_{\text{nn}}$ and the shear stress $\vec{t}^{\hat{n}}_{\text{ns}}$,

$$\begin{aligned} \vec{t}^{\hat{n}} &= \vec{t}^{\hat{n}}_{\text{nn}} + \vec{t}^{\hat{n}}_{\text{ns}} \\ &= (\vec{t}^{\hat{n}} \cdot \vec{n})\vec{n} + \vec{n} \times (\vec{t}^{\hat{n}} \times \vec{n}). \end{aligned} \tag{35}$$

The Cauchy stress tensor is in the spatial reference frame, and gives the force per unit deformed area. In other words, it gives the stress vectors for the deformed configuration. To work in the material description, it is necessary to have stress vectors that give a transformed force per unit undeformed area. These stress vectors are not the true physical stresses, but rather are convenient mathematical constructions, pseudo stresses, which allow us to discuss forces relative to the reference configuration. The first Piola-Kirchhoff stress tensor gives the force per unit undeformed area, but the force vectors are otherwise untransformed. The resulting pseudo stress vectors $\vec{T}^{\hat{N}}$ on undeformed surface $\vec{N}$ of area

A are related to the stress vectors as

$$\begin{aligned} \mathrm{d}\vec{f}^{\hat{n}} &= \vec{t}^{\hat{n}}\mathrm{d}a \\ &= \vec{T}^{\vec{\hat{N}}}\mathrm{d}A, \end{aligned} \tag{36}$$

while the pseudo-stress vector can be produced from the first Piola-Kirchhoff stress tensor $\vec{P}$ with

$$\vec{T}^{\vec{\hat{N}}} = \vec{P}\vec{\hat{N}}. \tag{37}$$

This stress tensor is a mixed tensor; it applies to one vector in the material reference, the surface normal vector, and another in the spatial reference, the direction in the spatial reference that the stress is desired in. Because we are changing the coordinates of part of the tensor, the first Piola-Kirchhoff stress tensor can be related to the Cauchy stress tensor with the deformation gradient, creating a stress tensor that can be applied to surface normal vectors in the material reference. We additionally need to scale the force as we want the force per undeformed unit area, and it can be be shown that this involves multiplying by the determinant of the deformation vector,

$$\vec{P} = \det(\vec{F})\vec{\sigma}\vec{F}^{-\intercal}. \tag{38}$$

However, we want a stress tensor that is fully in the material reference frame. In general, a transformed force can then be related to the untransformed force as

$$\mathrm{d}\mathfrak{F} = \vec{F}^{-1}\mathrm{d}\vec{f} \tag{39}$$

and an associated pseudo stress tensor defined as

$$\mathrm{d}\mathfrak{F}^{\vec{\hat{N}}} = \vec{\widetilde{T}}^{\vec{\hat{N}}}\mathrm{d}A. \tag{40}$$

The second Piola-Kirchhoff stress tensor $\vec{S}$ produces the desired pseudo-stress vectors that give the transformed force per undeformed area,

$$\vec{\tilde{T}}^{\vec{\hat{N}}} = \vec{S}\vec{\hat{N}}. \tag{41}$$

The second Piola-Kirchoff tress tensors is related to the first Piola-Kirchoff stress and tensor and the Cauchy stress tensor as

$$\begin{aligned}\vec{S} &= \vec{F}^{-1}\vec{P} \\ &= \det(\vec{F})\vec{F}^{-1}\vec{\sigma}\vec{F}^{-\intercal}.\end{aligned} \tag{42}$$

Under the infinitesimal strain approximation, because $\nabla_0\vec{u} \approx \nabla\vec{u}$, the difference between the different stress tensors becomes insignificant, so $\vec{S} \approx \vec{\sigma}$.

The stress-strain relationships can only be derived empirically. These are also known as constitutive relationships, and are material specific. Here, we are focused on elastic materials, and particularly hyperelastic materials, which are a subset of Cauchy elastic materials. In Cauchy elastic materials, the stress field only depends on the current state. For hyperelastic materials, it is possible to define a Helmholtz free-energy potential. If this Helmholtz free-energy potential is only a function of a deformation or strain tensor, then it is referred to as the strain energy density function, W, which has units of energy per unit mass.

To derive a linear relationship between the stress and strain, we can do a Taylor expansion of the strain energy density function. If we also assume infinitesimal strain, we can write the strain energy density as a function of the infinitesimal strain tensor. Then, if we retain only the first three terms while expanding around an unstrained configuration, $\vec{\epsilon} = 0$,

$$W(\vec{\epsilon}) = C_0 + C_{ij}\epsilon_{ij} + \frac{1}{2}C_{ijkl}\epsilon_{ij}\epsilon_{kl}, \tag{43}$$

where C_0 is a constant scalar, C_{ij} are the components of a constant vector, and C_{ijkl} are the components of a constant order-4 tensor. We can set $C_0 = 0$, as this is just a reference energy. It is possible to show with thermodynamic arguments and a number of axioms and constraints that

$$\vec{\sigma} = \frac{\partial W(\vec{\epsilon})}{\partial\vec{\epsilon}}. \tag{44}$$

If we take the derivative, then

$$\begin{aligned}\vec{\sigma} &= \left[C_{ij}\frac{\partial \epsilon_{ij}}{\partial \epsilon_{mn}} + \frac{1}{2}C_{ijkl}\left(\frac{\partial \epsilon_{ij}}{\partial \epsilon_{mn}}\epsilon_{kl} + \epsilon_{ij}\frac{\partial \epsilon_{kl}}{\partial \epsilon_{mn}}\right)\right]\vec{e}_m \otimes \vec{e}_n \\ &= \left[C_{ij}\delta_{im}\delta_{jn} + \frac{1}{2}C_{ijkl}(\delta_{im}\delta_{jn}\epsilon_{kl} + \epsilon_{ij}\delta_{km}\delta_{ln})\right]\vec{e}_m \otimes \vec{e}_n \\ &= \left[C_{mn} + \frac{1}{2}(C_{mnkl}\epsilon_{kl} + C_{ijmn}\epsilon_{ij})\right]\vec{e}_m \otimes \vec{e}_n \\ &= (C_{mn} + C_{mnkl}\epsilon_{kl})\vec{e}_m \otimes \vec{e}_n, \end{aligned} \tag{45}$$

where in the final step we have defined the constant order-4 tensor to have major symmetry, i.e. $C_{mnkl} = C_{klmn}$, for convenience. Because ϵ is symmetric, the constant order-4 tensor will also have minor symmetry, i.e. $C_{mnkl} = C_{nmkl}$. We can see that $C_{mn}\vec{e}_m \otimes \vec{e}_n$ is the residual stress, and so if our reference configuration is unstressed, it is 0. Finally, the order-4 tensor $C_{mnkl}\vec{e}_m \otimes \vec{e}_n$ is referred to as the stiffness tensor.

We can narrow our focus further, as we only consider isotropic materials, to simplify the constitutive relations. An isotropic tensor is one in which the components are invariant under transformations, and can be expressed as

$$C_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + \kappa(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}), \tag{46}$$

where λ, μ, and κ are known as the Lamé parameters. Because the stiffness tensor has minor symmetry, the third term will cancel, leading to

$$\begin{aligned} W(\vec{\epsilon}) &= \frac{1}{2}(\lambda\delta_{ij}\delta_{kl} + 2\mu\delta_{ik}\delta_{jl})\epsilon_{ij}\epsilon_{kl} \\ &= \frac{1}{2}\lambda\epsilon_{ii}\epsilon_{kk} + \mu\epsilon_{il}\epsilon_{il} \\ &= \frac{1}{2}\lambda \mathrm{tr}(\vec{\epsilon})^2 + \mu\vec{\epsilon} : \vec{\epsilon} \end{aligned} \tag{47}$$

and

$$\begin{aligned} \vec{\sigma} &= [(\lambda\delta_{ij}\delta_{kl} + 2\mu\delta_{ik}\delta_{jl})\epsilon_{kl}]\vec{e}_i \otimes \vec{e}_j \\ &= (\lambda\delta_{ij}\epsilon_{kk} + 2\mu\epsilon_{ij})\vec{e}_i \otimes \vec{e}_j \\ &= \lambda \mathrm{tr}(\vec{\epsilon})\vec{I} + 2\mu\vec{\epsilon}. \end{aligned} \tag{48}$$

Typically the equations are written in terms of the Young's modulus E and the Poisson's ratio ν, which have a more direct link to experimentally measurable quantities, and are related to the Lamé parameters with

$$\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)} \tag{49}$$

and

$$\mu = \frac{E}{2(1+\nu)}. \tag{50}$$

To begin to include nonlinearity, we can remove the assumption of infinitesimal strain and use the Green-Lagrange strain tensor in the above expressions. This is known as geometrically nonlinear elasticity, or the Saint Venant-Kirchhoff model. The constitutive relationship itself is still linear, although if large strains are being used, the approximation of using only the three terms of the expansion of the energy is likely to become poor. The strain energy density becomes

$$W(\vec{E}) = \frac{1}{2}\lambda \mathrm{tr}(\vec{E})^2 + \mu \vec{E} : \vec{E}, \tag{51}$$

and the stress tensor, now the second Piola-Kirchhoff stess tensor, is

$$\vec{S} = \lambda \mathrm{tr}(\vec{E})\vec{I} + 2\mu\vec{E}. \tag{52}$$

From the balance of linear momentum, it can be shown that in a material description

$$\nabla_0 \cdot (\vec{F}\vec{S}) + \rho_0 \vec{f} = \rho_0 \frac{\partial^2 \vec{u}}{\partial \vec{t}^2}, \tag{53}$$

where $\vec{f}$ is the body force, and all tensors (including vectors) are fields on $\vec{X}$. Here we are only interested in static problems, so the right-hand-side is zero. We will also not be applying any body forces, so we will also set that to zero. From the balance of angular momentum it can be shown that the stress tensor is symmetric, or

$$\vec{S} = \vec{S}^{\intercal}. \tag{54}$$

For a linear elastic model, we can substitute in for the stress tensor, and then for the deformation and strain tensors to produce a second order PDE in terms of the deformation field $\vec{u}$,

$$\begin{aligned} 0 &= \nabla \cdot [\vec{F}(\lambda \mathrm{tr}(\vec{\epsilon})\vec{I} + 2\mu\vec{\epsilon})] \\ &= \nabla \cdot [\lambda \mathrm{tr}(\vec{\epsilon})\vec{F} + 2\mu\vec{F}\vec{\epsilon}] \\ &= \nabla \cdot \left[\frac{\lambda}{2}\mathrm{tr}(\nabla\vec{u} + [\nabla\vec{u}]^\intercal)(\nabla\vec{u} + \vec{I}) + \mu(\nabla\vec{u} + \vec{I})(\nabla\vec{u} + [\nabla\vec{u}]^\intercal)\right]. \end{aligned} \tag{55}$$

Because we are making the infinitesimal strain assumption, we ignore the difference between derivatives with respect to the material coordinates and those with respect to the spatial coordinates, and have written everything in lower case for simplicity. As the expression becomes quite long, it is best to consider it in pieces. For the trace,

$$\begin{aligned} \mathrm{tr}(\nabla\vec{u} + [\nabla\vec{u}]^\intercal) &= \frac{\partial u_i}{\partial x_i} + \frac{\partial u_i}{\partial x_i} \\ &= 2\frac{\partial u_i}{\partial x_i}. \end{aligned} \tag{56}$$

For the whole first term T_1,

$$\begin{aligned} T_1 &= \nabla \cdot \left[\frac{\lambda}{2}\mathrm{tr}(\nabla\vec{u} + [\nabla\vec{u}]^\intercal)(\nabla\vec{u} + \vec{I})\right] \\ &= \left[\frac{\partial}{\partial x_i}\vec{\hat{e}}_i\right]\left[\lambda\frac{\partial u_j}{\partial x_j}(\frac{\partial u_l}{\partial x_k} + \delta_{kl})\vec{\hat{e}}_l \otimes \vec{\hat{e}}_k\right] \\ &= \lambda\left[\frac{\partial}{\partial x_k}\left(\frac{\partial u_j}{\partial x_j}\frac{\partial u_l}{\partial x_k}\right) + \frac{\partial^2 u_j}{\partial x_l \partial x_j}\right]\vec{\hat{e}}_l \\ &\approx \lambda\left[\frac{\partial^2 u_j}{\partial x_l \partial x_j}\right]\vec{\hat{e}}_l \\ &= \lambda\nabla(\nabla \cdot \vec{u}), \end{aligned} \tag{57}$$

where we have discarded terms involving multiples of derivatives of $\vec{u}$ as they become insignificant under the infinitesimal strain assumption. For the second term T_2,

$$\begin{aligned}
T_2 &= \nabla \cdot [\mu(\nabla\vec{u} + \vec{I})(\nabla\vec{u} + [\nabla\vec{u}]^\intercal)] \\
&= \nabla \cdot \mu[\nabla\vec{u}\nabla\vec{u} + \nabla\vec{u}(\nabla\vec{u})^\intercal + \nabla\vec{u} + (\nabla\vec{u})^\intercal] \\
&\approx \nabla \cdot \mu[\nabla\vec{u} + (\nabla\vec{u})^\intercal] \\
&= \mu\left[\frac{\partial}{\partial x_i}\vec{e}_i\right]\left[\frac{\partial u_j}{\partial x_k} + \frac{\partial u_k}{\partial x_j}\right]\vec{e}_j \otimes \vec{e}_k \\
&= \mu\left(\frac{\partial^2 u_j}{\partial x_k^2} + \frac{\partial u_k}{\partial x_k \partial x_j}\right)\vec{e}_j \\
&= \mu[\nabla \cdot \nabla\vec{u} + \nabla(\nabla \cdot \vec{u})], \qquad (58)
\end{aligned}$$

where we have again dropped terms with multiples of derivatives of $\vec{u}$. Together this gives

$$\mu\nabla \cdot \nabla\vec{u} + (\lambda + \mu)\nabla(\nabla \cdot \vec{u}) = 0. \qquad (59)$$

If instead of linear elasticity, we wish to derive an equivalent form for a geometrically nonlinear elastic model, the expression become much more cumbersome. It can be convenient to work instead with an expression for the strain energy density function. With this, the total energy of the system can be calculated by integrating over the total system volume. The energy will also have contributions from work done by external body and traction forces applied to the system, so two additional integration terms are needed,

$$\Pi(\vec{u}) = \int_\Omega W(\vec{u})\mathrm{d}\vec{x} - \int_\Omega \vec{f} \cdot \vec{u}\mathrm{d}\vec{x} - \int_{\partial\Omega} \vec{t} \cdot \vec{u}\mathrm{d}s, \qquad (60)$$

where Ω is the system volume, $\partial\Omega$ is the boundary, and $\vec{t}$ is the traction force. From this expression, using Hamilton's principle and variational calculus, it is possible to derive the second order partial differential equation (PDE) that results from directly using the balance of momentum. However, as we will see later, when estimating a solution with the finite element method (FEM), it is actually a variational form that is required.

Euler-Bernoulli beam theory

It is generally not possible to find a closed form solution for the deformation field without making some assumptions about the kinematics. Euler-Bernoulli beam theory is one such set of assumptions that can be applied when the bending and stretching of a linearly elastic, isotropic beam is being modeled. The theory is not internally consistent, and yet it has found extensive use in modeling such systems. Essentially, it amounts to assuming that the beam can be considered as a series of rigid plates along the beam axis which rotate about another axis perpendicular to the beam axis such that on one side of the beam there is compression and on the other side extension. The plates do not slide relative to each other, so lines that are perpendicular to the plates stay parallel with each other upon bending. It also assumes that this bending happens within a plane, which then allows the problem to be reduced to a 1D one. If the beam axis is along the x axis, and the bending occurs in the xy plane, then the deformation field corresponding to these assumptions can be written as

$$\vec{u}(X) = \left[a(X) - Y\frac{\mathrm{d}w(X)}{\mathrm{d}X}\right]\vec{\hat{e}}_x + w(X)\vec{\hat{e}}_y, \tag{61}$$

where $a(X)$ and $w(X)$ are unknown functions. We have written the derivative with respect to X here as it is more convenient.

If we plug these back into the expression for the infinitesimal strain tensor, we will find that the ϵ_{yx} component is 0, consistent with the assumptions made of the kinematics. Then,

$$\vec{\epsilon} = \left[\frac{\mathrm{d}a(X)}{\mathrm{d}X} - Y\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\right]\vec{\hat{e}}_x \otimes \vec{\hat{e}}_x. \tag{62}$$

The stress tensor is then

$$\begin{aligned}\vec{\sigma} &= (\lambda + 2\mu)\left[\frac{\mathrm{d}a(X)}{\mathrm{d}X} - Y\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\right]\vec{\hat{e}}_x \otimes \vec{\hat{e}}_x \\ &= E\left[\frac{\mathrm{d}a(X)}{\mathrm{d}X} - Y\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\right]\vec{\hat{e}}_x \otimes \vec{\hat{e}}_x,\end{aligned} \tag{63}$$

where the second line follows because Poisson's ratio $\nu = 0$ under the assumptions of beam theory.

If we tried to plug this into the expression for the balance of momentum, we would find the equations could not be satisfied if there are any forces applied to the system in the y direction to bend the beam, as all other terms related to the force are 0. Euler-Bernoulli beam theory proceeds by carrying out a force and moment balance at each point along the beam axis. Assume the beam has a force applied axially at each point by $f(X)$ with units of force per unit length, and a force applied transversely at each point by $q(X)$, also with units of force per unit length. If $N(X)$ is the normal reactive force, $V(X)$ is the transverse, or shear, reactive force, and $M(X)$ is the reactive moment around the z axis, then for each element of the beam along the beam axis

$$\frac{\mathrm{d}N(X)}{\mathrm{d}X} + f(X) = 0 \tag{64}$$

$$\frac{\mathrm{d}V(X)}{\mathrm{d}X} + q(X) = 0 \tag{65}$$

$$V(x) - \frac{\mathrm{d}M(X)}{\mathrm{d}X} = 0. \tag{66}$$

These reactive forces and moment can be written in terms of the stress tensor,

$$N(X) = \int_A \sigma_{xx}\mathrm{d}A, \tag{67}$$

$$V(X) = \int_A \sigma_{xy}\mathrm{d}A, \tag{68}$$

$$M(X) = \int_A Y\sigma_{xx}\mathrm{d}A, \tag{69}$$

Under the assumptions made, eq. (68) implies that $V(X)$ must be 0, and yet that clearly cannot be as it would again lead to an equilibrium condition that cannot be met in eq. (65).

The key trick in Euler-Bernoulli beam theory is to avoid this by defining the shear force in terms of the moment with eq. (66). If we substitute that into eq. (65), then

$$\frac{\mathrm{d}^2M(X)}{\mathrm{d}X^2} - q(X) = 0. \tag{70}$$

If we first substitute for σ_{xx} in the expressions in eq. (67),

$$\begin{aligned}
N(X) &= \int_A E\left[\frac{\mathrm{d}a(X)}{\mathrm{d}X} + Y\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\right]\mathrm{d}A \\
&= E\frac{\mathrm{d}a(X)}{\mathrm{d}X}\int_A \mathrm{d}A + E\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\int_A Y\mathrm{d}A \\
&= EA\frac{\mathrm{d}a(X)}{\mathrm{d}X}, \qquad (71)
\end{aligned}$$

and eq. (69),

$$\begin{aligned}
M(X) &= \int_A YE\left[\frac{\mathrm{d}a(X)}{\mathrm{d}X} + Y\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\right]\mathrm{d}A \\
&= E\frac{\mathrm{d}a(X)}{\mathrm{d}X}\int_A Y\mathrm{d}A + E\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\int_A Y^2\mathrm{d}A \\
&= EI\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}, \qquad (72)
\end{aligned}$$

where I is known as the second moment of the area, or the moment of inertia, then we can further substitute these expressions into eq. (64),

$$\frac{\mathrm{d}}{\mathrm{d}X}\left[EA\frac{\mathrm{d}a(X)}{\mathrm{d}X}\right] = -f(X), \qquad (73)$$

and eq. (70),

$$\frac{\mathrm{d}^2}{\mathrm{d}X^2}\left[EI\frac{\mathrm{d}^2w(X)}{\mathrm{d}X^2}\right] = q(X), \qquad (74)$$

to derive differential equations that can be independently solved for the unknown functions $a(X)$ and $w(X)$, and thus the deformation vector.

It is easy to show that if no axial forces are applied, then $a(X) = 0$. This also implies that the neutral line of the beam, that part which undergoes neither compression nor extension upon bending, will be in the same X position upon bending, which further implies that the contour length will increase. With small strains, this difference remains insignificant.

In solving for $w(X)$, which is often referred to as the deflection, if EI, which is also referred to as the flexular rigidity, is assumed to be constant along the beam, then we can factor these constants out of the derivatives. If the only forces applied to the system are at the boundary, then we can further simplify the expression involving the deflection to

$$EI\frac{\mathrm{d}^4 w(X)}{\mathrm{d}X^4} = 0, \tag{75}$$

and integrate to arrive at the following four equations with four unknowns,

$$V = EIC_1, \tag{76}$$

$$M(X) = -EI(C_1X + C_2), \tag{77}$$

$$\frac{\mathrm{d}w(X)}{\mathrm{d}X} = \frac{C_1X^2}{2} + C_2X + C_3, \tag{78}$$

$$w(X) = \frac{C_1X^3}{6} + \frac{C_2X^2}{2} + C_3X + C_4, \tag{79}$$

where C_1, C_2, C_3, and C_4 are integration constants. With four boundary conditions, we can solve for the integration constants.

Finite element method

The FEM allows approximate solutions to differential equations to be found. More precisely, the FEM provides a method to approximate weak forms of differential equations. While this method can be applied to a broad range of problems, here we will restrict our focus to 2nd order elliptic PDEs with vector valued solutions. We will also restrict further to problems involving Dirichlet boundary conditions that are essential, i.e., setting the solution values at the boundary. Let $\mathcal{L}(\vec{u})$ be some differential equation for which we want a solution for $\vec{u}$, where $\vec{u} = \vec{u}(\vec{x})$ is an element of some function space. Then, the strong formulation is simply

$$\begin{aligned} \mathcal{L}(\vec{u}) &= 0 & \vec{x} \in \Omega, \\ \vec{u} &= g(\vec{x}) & \vec{x} \in \partial\Omega, \end{aligned} \tag{80}$$

where Ω is the domain, $\partial\Omega$ is the boundary of the domain, and $g(\vec{x})$ is a function that gives the value of the solution at the boundary. Using the Poisson equation as a simple example, we have

$$\begin{aligned} \nabla \cdot \nabla \vec{u} + \vec{f} = 0 &\quad \vec{x} \in \Omega, \\ \vec{u} = 0 &\quad \vec{x} \in \partial\Omega, \end{aligned} \tag{81}$$

where $\vec{f} = \vec{f}(\vec{x})$ is some vector valued function.

Instead of having a function across its entire domain being equated to 0, it can be more convenient to set a functional of the differential equation to 0, where this is done for each function in some infinite set. In the weak formulation, the functional is the inner product of the differential equation with some test function,

$$\langle \vec{v}, \mathcal{L}(\vec{u}) \rangle = 0 \quad \forall \vec{v} \in V, \tag{82}$$

where V is the space that the test functions exist in. The solution space and the test function space are constructed such that the boundary conditions are enforced. As the spaces involved are function spaces, the inner product will involve some integral over the domain. For the Poisson equation, this becomes

$$\begin{aligned} \langle \vec{v}, \nabla \cdot \nabla \vec{u} + \vec{f} \rangle = 0 &\quad \forall \vec{v} \in V \\ \int_\Omega \vec{v} \cdot (\nabla \cdot \nabla \vec{u}) \mathrm{d}\vec{x} + \int_\Omega \vec{v} \cdot \vec{f} \mathrm{d}\vec{x} = 0 &\quad \forall \vec{v} \in V. \end{aligned} \tag{83}$$

The advantage of the weak formulation here has to do with the fact that integration by parts can be applied, which essentially shifts one of the derivatives off of the solution and to the test function. By doing this, it is possible to derive solutions that are not second-order differentiable, or in more general terms, are not solutions to the original strong form of the problem. To illustrate this, for the first term of the Poisson equation in eq. (83) we have

$$\begin{aligned}\int_\Omega \vec{v}\cdot(\nabla\cdot\nabla\vec{u})\mathrm{d}\vec{x} &= \int_\Omega v_i\frac{\partial^2 u_i}{\partial x_j^2}\mathrm{d}\vec{x}\\ &= \int_\Omega \frac{\partial}{\partial x_i}\left(v_j\frac{\partial u_j}{\partial x_i}\right)\mathrm{d}\vec{x} - \int_\Omega \frac{\partial v_i}{\partial x_j}\frac{\partial u_i}{\partial x_j}\mathrm{d}\vec{x}\\ &= \int_{\partial\Omega} v_j\frac{\partial u_j}{\partial x_i}n_i\mathrm{d}s - \int_\Omega \frac{\partial v_i}{\partial x_j}\frac{\partial u_i}{\partial x_j}\mathrm{d}\vec{x}\\ &= \int_{\partial\Omega}(\vec{v}\cdot\nabla\vec{u})\cdot\vec{n}\mathrm{d}s - \int_\Omega \nabla\vec{v} : \nabla\vec{u}\mathrm{d}\vec{x},\end{aligned} \tag{84}$$

where we have integrated by parts and then applied the divergence theorem. Because of the boundary conditions here, $\nabla\vec{u} = 0$ everywhere on the boundary, so the first term disappears. Overall this gives

$$\int_\Omega \nabla\vec{v} : \nabla\vec{u}\mathrm{d}\vec{x} = \int_\Omega \vec{v}\cdot\vec{f}\mathrm{d}\vec{x} \quad \forall\vec{v}\in V. \tag{85}$$

This is also commonly written as a bilinear term A and a linear term L,

$$A(\nabla\vec{v},\nabla\vec{u}) = L(\vec{v}) \quad \forall\vec{v}\in V. \tag{86}$$

This form is in fact the general form that integration by parts of the weak form is intended to give, with a bilinear term in $\nabla\vec{v}$ and $\nabla\vec{u}$, and a linear term in $\vec{v}$. It will be referred to here as the abstract weak form.

In the FEM, $\vec{u}(\vec{x})$ is approximated with a finite sum of local basis functions,

$$\vec{u}(\vec{x}) = \sum_i^N c_i\vec{\phi}_i(\vec{x}), \tag{87}$$

where $\vec{\phi}_i$ are the basis functions, c_i are their coefficients, and N is the number of basis functions. In the Galerkin version of FEM, the test functions are selected from the same space as the approximate solution is taken from.[1] More specifically, they are restricted to the finite set of the N basis functions used to express $\vec{u}$. The abstract weak form can

[1] Actually, the test functions are from the same space that the solution is from with homogeneous boundary conditions, if the solution is decomposed into one for the homogeneous part, and a part that adds the difference between that and the solution with the nonhomogeneous boundary conditions.

then be written as

$$\begin{aligned}A(\nabla[d_i\vec{\phi}_i], \nabla[c_j\vec{\phi}_j]) &= L(d_i\vec{\phi}_i)\\A(\nabla\vec{\phi}_i, \nabla\vec{\phi}_j)c_j &= L(\vec{\phi}_i)\\\vec{A}\vec{c} &= \vec{L},\end{aligned}\tag{88}$$

where now the problem of estimating a solution becomes a linear one in which we must solve for the vector of coefficients $\vec{c}$, where the elements of matrix $\vec{A}$ and vector $\vec{L}$ can be calculated once a basis set has been selected.

By local basis functions, we mean that the problem domain is partitioned into cells, and the basis functions are defined on a cell level. The cells are defined by a list of vertices associated with them that demarcate their boundaries. In addition to these vertices, there may be additional nodes associated with a cell for use in defining the basis functions. These nodes can be shared between more than one cell, or internal to a particular cell. Together, the set of cells is referred to as the mesh.

Because we are using the weak formulation, we can select a function space that is only first order differentiable. Here and typically, piecewise polynomials are used that are based on Lagrange polynomials. Lagrange polynomials are defined such that each basis function is 1 at only a single node, and 0 at all other nodes. A Lagrange polynomial is defined as

$$\psi_i(x) = \prod_{j\neq i}^{n} \frac{x - x_j}{x_i - x_j},\tag{89}$$

where x_i are the coordinates of the nodes and n is the number nodes being used, with the resulting polynomial being of degree $n - 1$. The Lagrange polynomials are defined on the level of individual cells, and the nodes are indexed locally to the cells, as well as having a global index. For 1D problem and an internal node in cell Ω_a,

$$\phi_i(x) = \begin{cases}\psi_{i,a}(x) & \text{if } x \in \Omega_a\\ 0 & \text{if } x \notin \Omega_a,\end{cases}\tag{90}$$

where $\psi_{i,a}$ is the Lagrange polynomial defined on cell a with global index i. For a node shared between two cells Ω_a and Ω_b,

$$\phi_i(x) = \begin{cases} \psi_{i,a} & \text{if } x \in \Omega_a, x \neq x_i \\ \psi_{i,b} & \text{if } x \in \Omega_b, x \neq x_i \\ 1 & \text{if } x = x_i \\ 0 & \text{if } x \notin \Omega_a \cup \Omega_b, \end{cases} \tag{91}$$

where the check for $x = 1$ is simply to avoid ambiguity in the expression above, as both $\psi_{i,a}$ and $\psi_{i,b}$ are equal to one at this point. For a 3D problem with a scalar solution, the basis functions are the tensor products of the 1D versions,

$$\phi_{ijk}(\vec{x}) = \phi_i(x)\phi_j(y)\phi_k(z). \tag{92}$$

For a vector valued solution, the basis functions for the scalar case are copied for each component,

$$\vec{\phi}_{ijkl}(\vec{x}) = \phi_i(x)\phi_j(y)\phi_k(z)\vec{\hat{e}}_l. \tag{93}$$

Typically, the nodes are indexed with a single number, and the basis functions then written with a single index, as in eq. (87).

We are also interested in solving problems that begin with a variational formulation, rather than a strong form PDE. This is the case when starting with a potential energy function and applying a variational approach, rather than beginning with a balance of momentum. According to Hamilton's principle for conservative forces,

$$\delta I = \delta \int_{t_1}^{t_2} L dt = 0, \tag{94}$$

where $L = T - \Pi$ is the Lagrangian, T is the total kinetic energy, and Π is the total potential energy. As we are dealing with static problems, $T = 0$, and the potential does not depend on time, so

$$\delta\Pi = 0. \tag{95}$$

For hyperelastic materials,

$$\begin{aligned}\delta\Pi &= \delta\int_{\Omega} W \mathrm{d}\vec{x} - \delta\int_{\Omega} \vec{f}\cdot\vec{u}\mathrm{d}\vec{x} - \int_{\partial\Omega}\vec{t}\cdot\vec{u}\mathrm{d}s \\ &= \int_{\Omega}\delta W\mathrm{d}\vec{x} - \int_{\Omega}(\delta\vec{f}\cdot\vec{u} + \vec{f}\cdot\delta\vec{u})\mathrm{d}\vec{x} - \\ &\qquad \int_{\partial\Omega}(\delta\vec{t}\cdot\vec{u} + \vec{t}\cdot\delta\vec{u})\mathrm{d}s \\ &= \int_{\Omega}\delta W\mathrm{d}\vec{x} - \int_{\Omega}\vec{f}\cdot\delta\vec{u}\mathrm{d}\vec{x} - \int_{\partial\Omega}\vec{t}\cdot\delta\vec{u}\mathrm{d}s, \end{aligned} \tag{96}$$

where the $\delta\vec{f}$ and $\delta\vec{t}$ terms are eliminated as they do not depend on $\vec{u}$. As we do not consider body forces or traction force boundary conditions, we can further simplify to

$$\delta\Pi = \int_{\Omega}\delta W\mathrm{d}\vec{x}. \tag{97}$$

To derive the weak formulation needed for the FEM, we will need to replace the $\delta\vec{u}$ terms with the test function. As an example, we can derive the weak form for linear elasticity from its energy density function,

$$\begin{aligned}\delta W(\vec{\epsilon}) &= \frac{1}{2}\lambda\delta(\epsilon_{ii}\epsilon_{jj}) + \mu\delta(\epsilon_{ij}\epsilon_{ij}) \\ &= \lambda\epsilon_{ii}\delta\epsilon_{jj} + 2\mu\epsilon_{ij}\delta\epsilon_{ij} \\ &= \lambda\epsilon_{ii}\delta_{jk}\delta\epsilon_{jk} + 2\mu\epsilon_{ij}\delta\epsilon_{ij}. \end{aligned} \tag{98}$$

We then must find the variation of the infinitesimal strain tensor,

$$\begin{aligned}\delta\vec{\epsilon} &= \frac{1}{2}\delta\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{x_i}\right)\vec{e}_i\otimes\vec{e}_j \\ \delta\vec{\epsilon} &= \frac{1}{2}\left(\frac{\partial\delta u_i}{\partial x_j} + \frac{\partial\delta u_j}{x_i}\right)\vec{e}_i\otimes\vec{e}_j. \end{aligned} \tag{99}$$

If we plug this back into the expression for the strain energy density function,

$$\delta W(\vec{\epsilon}) = \lambda\left(\frac{\partial u_i}{\partial x_i}\frac{\partial\delta u_j}{\partial x_j}\right) + \mu\left(\left[\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right]\left[\frac{\partial\delta u_i}{\partial x_j} + \frac{\partial\delta u_j}{\partial x_i}\right]\right). \tag{100}$$

Plugging this back into the total energy integral, we now have an expression that is bilinear in $\vec{u}$ and $\delta\vec{u}$, so if we then replace $\delta\vec{u}$ with a test function, we have the desired weak formulation.

To use the FEM for nonlinear problems, a method of converting the problem into a series of converging linear problems is needed. Fixed point iteration methods are typically used. The general idea of a fixed point iteration is to first choose a function $g(\vec{u})$ such that $\vec{u} = g(\vec{u})$ is also a solution to the nonlinear differential equation $\mathcal{L}(\vec{u}) = 0$. The function chosen must have the property that it converges to the solution with iterative updates to $\vec{u}$. Then, the general iteration is $\vec{u}_{k+1} = g(\vec{u}_k)$. Newton's method is one particular selection for $g(\vec{u})$, where

$$\begin{aligned} g(\vec{u}) &= \vec{u} - [\mathcal{L}'(\vec{u})]^{-1}\mathcal{L}(\vec{u}) \\ &\longrightarrow \vec{u}_{k+1} = \vec{u}_k - [\mathcal{L}'(\vec{u})]^{-1}\mathcal{L}(\vec{u}) \\ &\longrightarrow \mathcal{L}'(\vec{u}_k)\delta\vec{u}_k = -\mathcal{L}(\vec{u}_k), \end{aligned} \tag{101}$$

where $\delta\vec{u} = \vec{u}_{k+1} - \vec{u}_k$, and $\mathcal{L}'(\vec{u})$ is the directional derivative of the differential equation, which is itself a linear operator that takes the direction of the derivative as an argument. We have used $\delta\vec{u}$ as the notation for the update to the solution as it will take the place of the variation of $\vec{u}$, which comes about from taking the directional derivative, once $\mathcal{L}'(\vec{u})$ is applied to it. We can then construct the weak formulation of this expression to use the FEM to solve for $\delta\vec{u}$,

$$\langle \vec{v}, \mathcal{L}'(\vec{u}_k)\delta\vec{u}_k \rangle = -\langle \vec{v}, \mathcal{L}(\vec{u}_k) \rangle, \tag{102}$$

and integrate by parts to arrive at the abstract weak form,

$$A(\nabla\vec{v}, \nabla\delta\vec{u}) = L(\vec{v}). \tag{103}$$

Actually carrying out calculations with the FEM involves many further details at a level below the problem formulation. Creating a mesh for simple geometries can be done by hand, but more complex geometries may require algorithms to decide where to place nodes and connect them into cells. In calculating the elements of the bilinear and linear terms of the weak formulation, much time can be saved by performing the integrals once on a reference cell and mapping to the particular geometry of each cell in the system. The integrals themselves are typically not able to be solved analytically, and so some numerical integration algorithm is usually employed. Once the linear system has been assembled, some solver must be employed to calculate the solution coefficients. While a direct solver might be used, if the system is large, it may be necessary to apply an iterative solver like the conjugate gradient method instead.

Boundary conditions are implemented as constraints on the solution. It is common to refer to the basis functions and associated coefficients as the degrees of freedom. In general, if a constraint on the solution can be expressed as a linear combination of the degrees of freedom in the system, then it is possible to incorporate the constraints into the FEM estimate. The boundary conditions and any other conditions can be combined into a matrix of weights that is applied to the solution coefficient vector and set equal to the inhomogeneities, if inhomogeneous constraints are included. It is then possible to modify the linear equation for the unconstrained system to incorporate the constraints and solve for the desired constrained solution.