GEO 441 Notes

Max Chien

Fall 2025

Contents

Introduction

This document contains notes taken for the class GEO 441: Computational Geophysics at Princeton University, taken in the Spring 2025 semester. These notes are primarily based on lectures by Professor Jeroen Tromp. This class covers finite-difference, finite-element, and spectral methods for numerical solutions to the wave and heat equations. Since these notes were primarily taken live, they may contains typos or errors.

Chapter 1

Continuum Mechanics and the Equations of Motion

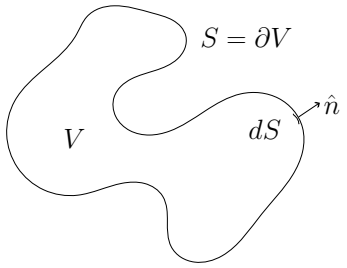
In this class, we will primarily focus on the wave and heat equations, which are important in the study of geophysics, and more broadly, continuum mechanics. As such, we will begin with an introduction to basic continuum mechanics to better understand the role of the differential equations we study.

Continuum mechanics are primarily governed by four conservation laws:

1. Conservation of mass,
2. Conservation of linear momentum,
3. Conservation of angular momentum,
4. Conservation of energy.

The wave and heat equations arise as a result of (2) and (4), respectively, but in actual applications it is often the case that coupled systems of conservation laws must be solved.

1.1 Conservation of Mass



We consider a "comoving volume" V. By "comoving volume", one can imagine a bag of some fluid mass deposited in a river, which can be deformed as it moves, but nevertheless

maintains a constant mass throughout. We also denote the surface of V by $S = \partial V$, and for small surface elements $\mathrm{d}S$ we denote the unit outward normal vector by $\hat{n}$.[1]

We also adopt the Einstein summation convention, in which repeated indices that are not otherwise used are implied to be summed over:

$$\vec{u} = u^i e_i$$

If we consider a change of basis to some new basis $\{e'_1, e'_2\}$, this can then be written as

$$\vec{u} = u^{i'} e'_i$$

where $u^{i'}$ denotes the ith component of $\vec{u}$ in the new basis.

While $\vec{u}$ is invariant under change of basis, the components are of course not. The way that they transform under change of basis is given by the change of basis matrix $\boldsymbol{\lambda}$, and this relationship is expressed under Einstein summation notation by

$$u^i = \lambda^i_{i'} u^{i'}$$
$$e_i = \lambda^i_{i'} e'_i$$

The reverse transformation may be denoted by $\boldsymbol{\Lambda}$. The fact that they are inverses may be expressed by the equation

$$\lambda^i_{i'} \Lambda^{i'}_j = \delta^i_j$$

where δ^i_j is the Kronecker delta (in coordinates, the RHS is the identity matrix). This then allows us to express the reverse relationships for change of basis:

$$u^{i'} = \Lambda^{i'}_i u^i$$
$$e'_i = \Lambda^{i'}_i e_i$$

Now, to formalize the notion of the mass of V, we first consider the mass density, considered as a function $\rho(\vec{x}, t)$ of both space and time (with respect to some coordinate system). For an infinitesimal volume element $\mathrm{d}V$, the mass of the volume is given by $\rho\,\mathrm{d}V$. Notice that the dimensions of mass density is

$$[\rho] = \frac{\mathrm{kg}}{\mathrm{m}^3}$$

so that the dimensions of mass are

$$[\rho]\,[\mathrm{d}V] = \mathrm{kg}$$

More generally, the mass of V is given by integrating against mass density,

$$M = \int_V \rho\,\mathrm{d}V$$

[1] In this course we adopt the convention that a vector is denoted by $\vec{v}$, a unit vector by $\hat{v}$, and the ith component of a vector by v_i or v^i. (The distinction is the distinction between covariant and contravariant indices, but is not necessary for this course). Moreover, we denote the standard basis vectors in the x and y directions by $e_x = \hat{x}$ and $e_y = \hat{y}$, respectively.

In Cartesian coordinates this is

$$M = \int_V \rho(x, y, z, t)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

Notice that the integrand is time dependent. Moreover, we allow V to deform over time as well, so that this equation might be more appropriately written as

$$M(t) = \int_{V(t)} \rho(x, y, z, t)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

Then the conservation of mass law is expressed as the ODE

$$0 = \frac{\mathrm{d}M}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)} \rho\,\mathrm{d}V$$

If V is constant (that is, if we allow for no deformation), then Feynman's trick give us

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \int_V \frac{\partial \rho}{\partial t}\,\mathrm{d}V$$

However, because V is time-dependent, this fails to hold. Instead, we first appeal to the single-dimensional case by considering Leibniz's rule, which handles integration with time-dependent limits and integrand of the form

$$I(t) = \int_{a(t)}^{b(t)} f(x, t)\,\mathrm{d}x$$

In this case, by considering I as the area under the curve, it is clear that (at least for continuous a, b) the value $\frac{\mathrm{d}I}{\mathrm{d}t}$ must take into account both the values $\frac{\partial f}{\partial t}|_{[a,b]}$, but also the area which is added or removed by the change in a, b.

Theorem 1.1: Leibniz's Rule

Let $f(x, t)$ be jointly continuous with $\frac{\partial}{\partial t}f(x, t)$ also jointly continuous in some region given by $a(t) \le x \le b(t)$, $t_0 \le t \le t_1$. If a, b are both continuously differentiable, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{a(t)}^{b(t)} f(x, t)\,\mathrm{d}x\right) = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(x, t)\,\mathrm{d}x + f(b(t), t)\frac{\mathrm{d}b}{\mathrm{d}t}(t) - f(a(t), t)\frac{\mathrm{d}a}{\mathrm{d}t}(t)$$

This can be derived using the limit formulation of the derivative by writing

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\left[\int_{a(t+\Delta t)}^{b(t+\Delta t)} f(x, t + \Delta t)\,\mathrm{d}x - \int_{a(t)}^{b(t)} f(x, t)\,\mathrm{d}x\right]$$

As a first order approximation for the change in area if the integration limits are constant, Feynman's rule holds and we have

$$\int_{a(t)}^{b(t)} \frac{1}{\Delta t}\lim_{\Delta t \to 0}\left[f(x, t + \Delta t) - f(x)\right]\mathrm{d}x + O((\Delta t)^2) = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(x, t)\,\mathrm{d}x$$

At the upper limit, f is also near constant, so the change in area is approximated to first order by

$$f(b(t),t)\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[b(t+\Delta t)-b(t)\right]=f(b(t),t)\frac{\mathrm{d}b}{\mathrm{d}t}(t)$$

The lower limit is similar with a negative sign. Combining the three approximations, we get

$$\frac{\mathrm{d}I}{\mathrm{d}t}=\int_{a(t)}^{b(t)}\frac{\partial f}{\partial t}(x,t)\,\mathrm{d}x+f(b(t),t)\frac{\mathrm{d}b}{\mathrm{d}t}(t)-f(a(t),t)\frac{\mathrm{d}a}{\mathrm{d}t}(t)$$

Now, we return to the case of our comoving volume. Taking inspiration from Leibniz's rule, the main term that we have to adjust in the 2-dimensional case is the change in boundary area. This is approximated by considering the volume over which a surface element moves within an infinitesimal time interval.

For a given surface element $\mathrm{d}S(t)$, we consider both the associated normal $\hat{n}(t)$ and the velocity vector $\vec{v}$. Then the component of the velocity of $\mathrm{d}S(t)$ in the normal direction is given by

$$\vec{v}\cdot\hat{n}(t)=v^i(t)n^i(t)$$

Note that, as usual we also define the length of u by

$$\|\vec{u}\|^2=(u^i)^2$$

Now, the flux of mass through $\mathrm{d}S(t)$ in the period $[t,t+\Delta t]$ is then

$$\rho|_{\mathrm{d}S(t)}\vec{v}\cdot\hat{n}$$

Then we can now include the correct error term to calculate $\frac{\mathrm{d}M}{\mathrm{d}t}$:

$$\frac{\mathrm{d}M}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)}\rho\,\mathrm{d}V=\int_{V(t)}\frac{\partial\rho}{\partial t}\,\mathrm{d}V+\int_{S(t)}\rho\vec{v}\cdot\hat{n}\,\mathrm{d}S$$

(where S is equipped with the outward-facing orientation). Lastly, we can replace the second term with an integral over $V(t)$ using the divergence theorem:

$$\int_S\vec{u}\cdot\hat{n}\,\mathrm{d}S=\int_V\nabla\cdot\vec{u}\,\mathrm{d}V$$

We combine the integrals:

$$\frac{\mathrm{d}M}{\mathrm{d}t}=\int_{V(t)}\left[\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\vec{v})\right]\mathrm{d}V$$

Note that the divergence is taken against $\rho\vec{v}$, since this is the quantity which is dotted against $\hat{n}$.

Because the integral must be zero for all possible V, the integrand is identically zero. Thus we express the conservation of mass law for a comoving volume (also known as the **continuity equation**) by

$$\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\vec{v})=0$$

We can expand this using summation notation as

$$\partial_t \rho + v^i \nabla_i \rho + \rho \nabla_i v^i = 0$$

The first two terms $\partial_t \rho + v^i \nabla_i \rho$ is known as the **material derivative**

$$D_t \rho = \partial_t \rho + \vec{v} \cdot \nabla \rho$$

where the first term is the local change in density, and the second is the advection term (which is the directional derivative of the density in the direction of velocity). In other words, the rate of change of local mass along a path is given by the pointwise rate of change together with the change given by the motion of the path against the gradient. We then rephrase the continuity equation as

$$D_t \rho + \rho \nabla \cdot \vec{v} = 0$$

or equivalently

$$\frac{1}{\rho} D_t \rho = -\nabla \cdot \vec{v}$$

This essentially says that the relative change in density along a path is the negative of the velocity divergence. This makes sense because when divergence is positive, mass is moving away and density decreases, while density increases with velocity divergence is negative. In particular, if the mass is incompressible, $\nabla \cdot \vec{v} = 0$, so that density is constant along any path. In this case, we don't need to worry about conservation of mass.

1.2 Conservation of Linear Momentum

Linear momentum is given by the product of mass with velocity. In continuum mechanics this is given by $\rho \vec{v} \, \mathrm{d}V$. Thus the total momentum of a volume is simply

$$p = \int_{V(t)} \rho \vec{v} \, \mathrm{d}V$$

The statement of conservation of linear momentum is essentially that the only way to change linear momentum is to apply (external) forces to our volume. This is basically Newton's second law, written as $\vec{F} = \dot{p}$. One can consider a body force $\vec{f}$ which pulls on small volume elements $\mathrm{d}V$. We can also consider forces $\vec{t}$ which act only on the surface of the volume. Thus we write

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{V(t)} \rho \vec{v} \, \mathrm{d}V = \int_{V(t)} f \, \mathrm{d}V + \int_{S(t)} \vec{t} \, \mathrm{d}S$$

We can differentiate the left hand side the same way as we did in the conservation mass equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{V(t)} \rho \vec{v} \, \mathrm{d}V = \int_{V(t)} \partial_t (\rho \vec{v}) \, \mathrm{d}V + \int_{S(t)} (\rho \vec{v}) \hat{n} \cdot \vec{v} \, \mathrm{d}S$$

To conceptualize the surface-acting forces, we consider the **stress tensor**, which is a rank 2 tensor (or matrix) $\mathbf{T}$ such that $\mathbf{T} \cdot \vec{n}$ gives the traction force on $\mathrm{d}S$, if the unit outward normal of $\mathrm{d}S$ is $\hat{n}$. In indices, this is

$$t_i = T_{ij} \hat{n}_j \, \mathrm{d}S$$

(Note that in general $T_{ij}\hat{n}_j \neq \hat{n}_j T_{ji}$, but this is true if $\mathbf{T}$ is a symmetric tensor). Then the right hand side of our equation is writen as

$$\int_{V(t)} \vec{f}\,\mathrm{d}V + \int_{S(t)} \mathbf{T}\cdot\hat{n}\,\mathrm{d}S$$

Once again we use the divergence theorem to convert these to volume integrals, so that our equation is given by

$$\int_{V(t)} \left[\partial_t(\rho\vec{v}) + \nabla\cdot(\rho\vec{v}\otimes\vec{v})\right]\mathrm{d}V = \int_{V(t)} \left[\vec{f} + \nabla\cdot\mathbf{T}\right]\mathrm{d}V$$

Note that both integrals are vector quantities. In components, the integrand on the left can be given by

$$\partial_t \rho v^i + \nabla_j \cdot \left(\rho v^i v^j\right)$$

(By convention, the divergence theorem is written in indices as $\int_S u^i \hat{n}_i\,\mathrm{d}S = \int_V \nabla_i u^i\,\mathrm{d}V$).

Similarly, the divergence of $\mathbf{T}$ is given by contracting the gradient against the last index of $\mathbf{T}$, so that the integrand on the right is given in indices by

$$f^i + \nabla_j \cdot T^{ij}$$

Equating the integrands again, the conservation of linear momentum law is thus given by

$$\partial_t(\rho\vec{v}) + \nabla\cdot(\rho\vec{v}\otimes\vec{v}) = \nabla\cdot\mathbf{T} + \vec{f}$$

Some equivalent formulations are

$$\partial_t(\rho\vec{v}) = \nabla\cdot(\mathbf{T} - \rho\vec{v}\otimes\vec{v}) + \vec{f}$$
$$\partial_t(\rho\vec{v}) + \nabla\cdot(\rho\vec{v}\otimes\vec{v} - \mathbf{T}) = \vec{f}$$

The last formulation is the Eulerian form, which expresses the conservation law as the pointwise time derivative of a quantity plus its flux being equated to the source term.

Expressing this with the chain rule gives

$$(\partial_t\rho)\vec{v} + \rho\partial_t\vec{v} + \nabla\cdot(\rho\vec{v})\vec{v} + \rho\vec{v}\nabla\cdot\vec{v} = \nabla\cdot\mathbf{T} + \vec{f}$$

The first and third term are zero by conservation of mass. Thus this is equivalent to

$$\rho\left(\partial_t\vec{v} + \vec{v}\nabla\cdot\vec{v}\right) = \nabla\cdot\mathbf{T} + \vec{f}$$

The parenthetical term is again the material derivative, this time of velocity, so this is

$$\rho D_t\vec{v} = \nabla\cdot\mathbf{T} + \vec{f}$$

As formulated, the coupling of the conservation of mass and momentum laws gives four scalar equations. Even if body forces are given, this leaves as unknowns the mass density, velocity, and stress tensor. Thus we need constitutive relationships, which express some of these (particularly the stress tensor) in terms of the others in order to solve these. This

makes sense given that the actual results will depend on material properties, which are specified in the stress tensor but nowhere else.

To do this, we consider stress and strain. Fix some origin point and let $\vec{x}$ denote the starting point of some particle. Let $\vec{r}(\vec{x},t)$ denote the position of particle $\vec{x}$ at time t. By definition $\vec{r}(\vec{x},0)=\vec{x}$. Define $\vec{s}(\vec{x},t)=\vec{r}(\vec{x},t)-\vec{x}$ to be the displacement vector. Suppose we consider two initially neighboring particles $\vec{x},\vec{x}+\mathrm{d}\vec{x}$. As time progresses, their displacement becomes $\mathrm{d}\vec{r}=\vec{r}(\vec{x}+\mathrm{d}\vec{x},t)-\vec{r}(\vec{x},t)$. We take the first order Taylor expansion:

$$\mathrm{d}\vec{r}\approx\vec{r}(\vec{x},t)+\mathrm{d}\vec{x}\cdot\nabla\vec{r}(\vec{x},t)-\vec{r}(\vec{x},t)=\mathrm{d}x^i\nabla_i\vec{r}=\mathrm{d}\vec{x}\cdot\nabla\vec{r}$$

We can express this as a tensor by

$$\nabla_j r^i\,\mathrm{d}x^j=F^i_j\,\mathrm{d}x^j$$

where $F^i_j=\nabla_j r^i$, or equivalently $\mathbf{F}=(\nabla\vec{r})^T$. The tensor $\mathbf{F}$ is known as the **deformation gradient**. Recalling that $\vec{r}=\vec{x}+\vec{s}$, we have

$$F=[\nabla(\vec{x}+\vec{s})]^T=[\nabla\vec{x}+\nabla s]^T$$

Since $\nabla\vec{x}$ is taken against $\vec{x}$ itself, its matrix formulation is just the identity:

$$\mathbf{I}=\nabla\vec{x}=\hat{x}\otimes\hat{x}+\hat{y}\otimes\hat{y}+\hat{z}\otimes\hat{z}=(\delta_{ij})$$

In summary, we can write

$$\mathbf{F}=\mathbf{I}+(\nabla\vec{s})^T$$

which is the identity plus the transpose of the displacement gradient. Physically, the displacement gradient represents the separation or convergence of material, or equivalently the deviation from uniform motion. Noting that

$$\mathrm{d}\vec{r}=\mathbf{F}\cdot\mathrm{d}\vec{x}$$

we have

$$\mathrm{d}\vec{r}=\left[\mathbf{I}+(\nabla\vec{s})^T\right]\mathrm{d}\vec{x}=\mathrm{d}\vec{x}+(\nabla\vec{s})^T\cdot\mathrm{d}\vec{x}$$

In general the tensor may not be symmetric; however we can always decompose a matrix into its symmetric and antisymmetric parts as

$$\mathbf{A}=\frac{1}{2}\left(\mathbf{A}+\mathbf{A}^T\right)+\frac{1}{2}\left(\mathbf{A}-\mathbf{A}^T\right)$$

In particular, we can write F as

$$\mathbf{F}=\mathbf{I}+\boldsymbol{\varepsilon}+\boldsymbol{\omega}$$

where $\boldsymbol{\varepsilon},\boldsymbol{\omega}$ are the symmetric and antisymmetric parts of $(\nabla\vec{s})^T$, respectively. $\boldsymbol{\varepsilon}$ is called the **strain** and $\boldsymbol{\omega}$ the **vorticity**. In other words, $\boldsymbol{\varepsilon}$ denotes the linear deviation from uniform displacement, or the linear deformation, and $\boldsymbol{\omega}$ denotes the twisting component.

Note that this implies the following:

$$\mathrm{tr}(\boldsymbol{\omega})=0$$
$$\mathrm{tr}(\boldsymbol{\varepsilon})=\mathrm{tr}(\nabla\vec{s})=\nabla\cdot\vec{s}$$

So $\operatorname{tr}(\boldsymbol{\varepsilon})$ can be seen to measure the local density or volume change.

It is shown in homework that we can calculate

$$\boldsymbol{\omega} \cdot \mathrm{d}\vec{x} = \frac{1}{2} (\nabla \times \vec{s}) \times \mathrm{d}\vec{x}$$

Thus we have

$$\mathrm{d}\vec{r} = \mathbf{F} \cdot \mathrm{d}\vec{x} = (\mathbf{I} + \boldsymbol{\varepsilon} + \boldsymbol{\omega}) = \mathrm{d}\vec{x} + \boldsymbol{\varepsilon} \cdot \mathrm{d}\vec{x} + \boldsymbol{\omega} \cdot \mathrm{d}\vec{x}$$

Which essentially says that final change in position differs by original change position by some linear strain component $\boldsymbol{\varepsilon} \cdot \mathrm{d}\vec{x}$, and rotationally by $\boldsymbol{\omega} \cdot \mathrm{d}\vec{x}$.

Based on this physical interpretation, it is clear that applying a stress force on the exterior of a body should impart a strain on the interior. If we relate the two, this can help us reduce the dimensionality of our PDE. One possible assumption is **Hooke's law**, which postulates that this is a linear relationship.

In other words, for each component T_{ij} of the stress tensor, there should be coefficients $a_{ij}, \ldots, f_{ij}$ such that

$$T_{ij} = a_{ij}\varepsilon_{11} + b_{ij}\varepsilon_{12} + \ldots + f_{ij}\varepsilon_{33}$$

(Note that there are only six degrees of freedom since $\boldsymbol{\varepsilon}$ is symmetric). These coefficients can be collected in a fourth-order tensor called the **elastic tensor**. This is summarized as

$$T_{ij} = c_{ijkl}\varepsilon_{kl}$$

Returning to the conservation of momentum law and substitute this relationship, we have

$$\rho \partial_t^2 \vec{s} = \nabla \cdot (\mathbf{c} : \boldsymbol{\varepsilon}) + \vec{f}$$

or in components,

$$\rho \partial_t^2 s^i = \nabla_j \cdot \left(c^{ijkl} \varepsilon_{kl} \right) + f^i$$

A priori, we have not really reduced the dimensionality, since $\mathbf{c}$ has 81 components. However, conservation of angular momentum forces the stress tensor to be symmetric, and conservation of energy gives symmetricity across the first two and last two indices. This reduces the number of independent components to 21.

We now investigate further the importance of Hooke's law in developing the wave equation. Consider the one-dimensional case of Hooke's law, which can be imagined by a spring of length L and spring constant k. If it is given an initial displacement $\Delta\ell$, then Hooke's law says that the spring force is given by $F = k\Delta L$.

By graphing the displacement at $t = 0$ against the position along the spring x, the displacement linearly increases from 0 to $\Delta\ell$. In other words,

$$s(x) = \frac{\Delta\ell}{L} x$$

so the strain is given by

$$\varepsilon = \frac{\mathrm{d}}{\mathrm{d}x} s = \frac{\Delta\ell}{L}$$

which is therefore constant along the spring. Then the stress is linear in ε, so that

$$\sigma = \mu\varepsilon = \mu\frac{\Delta\ell}{L} = \frac{\mu}{kL}F$$

Relating this back to stress forces, the force against a unit area is the stress force:

$$\sigma = \frac{F}{A}$$

so that

$$\sigma = \frac{\mu A}{kL}\sigma$$

or

$$\mu = \frac{kL}{A}$$

1.3 The 1D Wave Equation

Here we develop the one-dimensional wave equation PDE as a consequence of conservation of linear momentum. Imagine a horizontal string of length L, and suppose that the string experiences perpendicular displacement given by $\vec{s}(x,t) = s(x,t)\hat{y}$ in the vertical direction. The gradient of $\vec{s}$ is given by

$$\nabla\vec{s} = \hat{x}\otimes\hat{y}\partial_x s$$

The strain is the symmetric part, which is therefore given by

$$\boldsymbol{\varepsilon} = \frac{1}{2}\partial_x s\left(\hat{x}\otimes\hat{y} + \hat{y}\otimes\hat{x}\right)$$

We apply Hooke's law to linearly relate stress and strain:

$$\mathbf{T} = 2\mu\boldsymbol{\varepsilon} = T_{xy}\left(\hat{x}\otimes\hat{y} + \hat{y}\otimes\hat{x}\right)$$

(The factor of 2 is conventional). It is thus clear that $\mathbf{T}$ has to be symmetric in the 1D case under Hooke's law, so that

$$T_{xy} = \mu\partial_x s$$

To calculate the divergence of $\mathbf{T}$, we have

$$\nabla\cdot\mathbf{T} = \partial_x\left(\mu\partial_x s\right)\hat{y}$$

Since the acceleration is also vertical, it is given by

$$\rho\partial_t^2\vec{s} = \rho\partial_t^2 s\hat{y}$$

Plugging this into the conservation of momentum equation, we get

$$\rho\partial_t^2 s = \partial_x\left(\mu\partial_x s\right)$$

Note that a priori we allow the **shear modulus** μ to vary over the string. However, if it is constant, then we can conclude

$$\partial_t^2 s = \beta^2 \partial_x^2 s$$
$$\beta = \sqrt{\frac{\mu}{\rho}}$$

where β is the shear wave speed.

Let us now consider the propagation of sound waves through fluids. In a fluid, the traction must be perpendicular to the surface, so that

$$\vec{t} \sim -p\hat{n}\,\mathrm{d}S$$

where p is the pressure. For an isotropic fluid, the forces are the same in all directions and only governed by pressure, so that the stress tensor can be written as

$$\mathbf{T} = -p\mathbf{I}$$

In a fluid, the pressure fluctuations are thus governed by the strain

$$\operatorname{tr}\boldsymbol{\varepsilon} = \nabla \cdot \vec{s}$$

This is expressed as

$$p = -\kappa \nabla \cdot \vec{s}$$

where κ is the **bulk modulus** or incompressibility of the fluid.

Under isotropy, the stress is completely governed by the shear modulus and bulk modulus, which reduces from 81 to 2 parameters. Hooke's law can be written as

$$\mathbf{T} = \kappa \operatorname{tr}(\boldsymbol{\varepsilon})\mathbf{I} + 2\mu\mathbf{d}$$

where $\mathbf{d}$ is the deviatoric strain tensor, which is essentially the traceless part of the strain: $\mathbf{d} = \boldsymbol{\varepsilon} - \frac{1}{3}\operatorname{tr}(\boldsymbol{\varepsilon})\mathbf{I}$.

Using these relations, we have

$$\rho\partial_t^2 \vec{s} = -\nabla p$$

Or alternately,

$$\partial_t^2 \vec{s} = -\frac{1}{\rho}$$
$$\partial_t^2 \nabla \cdot \vec{s} = -\nabla \cdot \left(\frac{1}{\rho \nabla p}\right)$$
$$\frac{1}{\kappa}\partial_t^2 p = \nabla \cdot \left(\frac{1}{\rho}\nabla p\right)$$

Under constant density assupmtions, we have the **acoustic wave equation**

$$\partial_t^2 p = c^2 \partial_x^2 p$$

where

$$c = \sqrt{\frac{\kappa}{\rho}}$$

is the sound wave speed.

1.4 Conservation of Angular Momentum

Since a small quantity of linear momentum can be calculated as $\rho\vec{v}\,\mathrm{d}V$, a small quantity of angular momentum is given by $\vec{r}\times\rho\vec{v}\,\mathrm{d}V$. The total angular momentum of a comoving volume is therefore

$$\int_{V(t)}\vec{r}\times\rho\vec{v}\,\mathrm{d}V$$

As with linear momentum, we can express the changes in angular momentum as a sum of body torques and surface torques:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)}\vec{r}\times\rho\vec{v}\,\mathrm{d}V=\int_{V(t)}\vec{r}\times\vec{f}\,\mathrm{d}V+\int_{S(t)}\vec{r}\times\vec{t}\,\mathrm{d}S$$

We proceed on the left side as before, differentiating and applying the divergence theorem:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)}\vec{r}\times\rho\vec{v}\,\mathrm{d}V=\int_{V(t)}\left[\partial_t\left(\vec{r}\times\rho\vec{v}\right)+\nabla\cdot\left(\rho\vec{r}\times\vec{v}\otimes\vec{v}\right)\right]\mathrm{d}V$$

On the right, we have

$$\int_{V(t)}\vec{r}\times\vec{f}\,\mathrm{d}V+\int_{S(t)}\vec{r}\times\vec{t}\,\mathrm{d}S=\int_{V(t)}\left[\vec{r}\times\vec{f}+\nabla\cdot\left(\vec{r}\times\mathbf{T}\right)\right]\mathrm{d}V$$

(Note that $\vec{r}\times\mathbf{T}$ should be interpreted as taking the cross product against the first index of $\mathbf{T}$.) On the left hand side, there is a $\partial_t\rho+\nabla\cdot(\rho\vec{v})$ term, which is zero by conservation of lienar momentum. Thus the final expression of conservation of angular momentum is

$$\rho D_t(\vec{r}\times\vec{v})-\nabla\cdot(\vec{r}\times\mathbf{T})=\vec{r}\times\vec{f}$$

For the material derivative of angular velocity, we note that $D_t\vec{r}=\vec{v}$, so that we can pull the $\vec{r}\times$ outside:

$$\vec{r}\times\left(\rho D_t\vec{v}-\nabla\cdot\mathbf{T}-\vec{f}\right)-\boldsymbol{\epsilon}:\mathbf{T}=\vec{0}$$

where $\boldsymbol{\epsilon}$ is the rank three alternating tensor. By conservation of linear momentum,

$$\rho D_t\vec{v}=\nabla\cdot\mathbf{T}+\vec{f}$$

so that the entire first term vanishes. Thus we conclude that

$$\boldsymbol{\epsilon}:\mathbf{T}=\vec{0}$$

In other words, this tells us that

$$\epsilon_{ijk}T_{kj}=0$$

for all i, which implies that $T_{23}=T_{32}$ and similarly $T_{13}=T_{31}, T_{12}=T_{21}$. Therefore under conservation of angular momentum and linear momentum, the stress tensor has to be symmetric. Thus the elastic tensor c_{ijkl} is symmetric in i,j, as well as k,l since ε is also symmetric. Thus we have 6 independent components in i,j and 6 in k,ℓ, so there are 36 independent components.

1.5 Conservation of Energy

Our final conservation law is conservation of energy, which leads to the heat equation.

The principle of conservation of energy essentially says that energy content is changed by work done. Considering again our comoving volume V. For any point particle with mass $\rho\,\mathrm{d}V$, the kinetic energy is $\frac{1}{2}\rho\,\mathrm{d}V\|\vec{v}\|^2$, so that the total kinetic energy is given by

$$KE = \frac{1}{2}\int_{V(t)} \rho\|\vec{v}\|^2\,\mathrm{d}V$$

To calculate the internal or potential energy term, simply consolidate all the internal energies into a term $\rho U\,\mathrm{d}V$, where U is the potential energy per unit mass. This gives

$$PE = \int_{V(t)} \rho U\,\mathrm{d}V$$

Therefore the total energy is expressed as

$$E = \int_{V(t)} \rho\left(\frac{1}{2}\|\vec{v}\|^2 + U\right)\mathrm{d}V$$

As with the previous conservation laws, we can relate the relate of change of the energy to the forces applied to our volume. In this case, work is calculated by considering the forces as

$$\int_{V(t)} \vec{v}\cdot\vec{f}\,\mathrm{d}V + \int_{S(t)} \vec{v}\cdot\vec{t}\,\mathrm{d}S$$

However, we also need to consider internal production of heat within the volume, for instance due to radioactivity. Similarly we need to consider heat fluxes out of the volume. These terms are given by

$$\int_{V(t)} h\,\mathrm{d}V - \int_{S(t)} \vec{H}\cdot\hat{n}\,\mathrm{d}S$$

where $\vec{H}$ is the heat flux out of V. Thus we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)} \rho\left(\frac{1}{2}\|\vec{v}\|^2 + U\right)\mathrm{d}V = \int_{V(t)} \vec{v}\cdot\vec{f}\,\mathrm{d}V + \int_{S(t)} \vec{v}\cdot(\mathbf{T}\cdot\hat{n})\,\mathrm{d}S + \int_{V(t)} h\,\mathrm{d}V - \int_{S(t)} \vec{H}\cdot\hat{n}\,\mathrm{d}S$$

Applying the same strategy as before, we have

$$\begin{aligned}\int_{V(t)} \rho D_t\left(\frac{1}{2}\|\vec{v}\|^2 + U\right)\mathrm{d}V &= \int_{V(t)} \left[\vec{v}\cdot\vec{f} + h + \nabla\cdot\left(\vec{v}\cdot\mathbf{T} - \vec{H}\right)\right]\mathrm{d}V \\ \implies \rho D_t\left(\frac{1}{2}\|\vec{v}\|^2 + U\right) + \nabla\cdot\left(\vec{H} - \vec{v}\cdot\mathbf{T}\right) &= h + \vec{v}\cdot\vec{f}\end{aligned}$$

Recall that one formulation of conservation of linear momentum was $\rho D_t\vec{v} - \nabla\cdot\mathbf{T} = \vec{f}$. Thus

$$\vec{v}\cdot\left(\rho D_t\vec{v} - \nabla\cdot\mathbf{T} - \vec{f}\right) = 0$$

Removing these terms from the equation, we get

$$\rho D_t U + \nabla \cdot \vec{H} = \mathbf{T} : \nabla \vec{v} + h$$

When $\mathbf{T}$ is symmetric, the contraction $\mathbf{T} : \nabla \vec{v}$ leaves only the symmetric part of $\nabla \vec{v}$ (the general principle is that a symmetric tensor contracted with an antisymmetric tensor gives zero). Thus we could replace $\nabla \vec{v}$ in the above with $\mathbf{D}$, where $\mathbf{D} = \frac{1}{2}\left[(\nabla \vec{v})^T + \nabla \vec{v}\right]$:

$$\rho D_t U + \nabla \cdot \vec{H} = \mathbf{T} : \mathbf{D} + h$$

In the case of waves propagating through elastic materials, particularly seismic waves, the rate of heat flux and heat production are negligible. Linearizing the wave equation, we have

$$\rho \partial_t U = \mathbf{T} : \nabla \partial_t \vec{s}$$

Applying Hooke's law, we write

$$\rho \partial_t U = (\mathbf{c} : \nabla \vec{s}) : \nabla \partial_t \vec{s} = c_{ijkl} \partial_k s_l \partial_i \partial_t s_j = \frac{1}{2} \partial_t \left(\partial_i s_j c_{ijkl} \partial_k s_l \right)$$
$$= \frac{1}{2} \partial_t \left(\varepsilon_{ij} c_{ijkl} \varepsilon_{kl} \right)$$

Integrating against time, we have

$$\rho U = \rho_0 U_0 + \frac{1}{2} \varepsilon_{ij} c_{ijkl} \varepsilon_{kl} = \rho_0 U_0 + \frac{1}{2} \boldsymbol{\varepsilon} : \mathbf{c} : \boldsymbol{\varepsilon}$$

1.6 The 1D Heat Equation

To develop the heat equation, consider a particle with zero initial velocity. Then conservation of energy gives

$$\rho \partial_t U + \nabla \cdot \vec{H} = h$$

In order to continue deriving this equation, we need assumptions on U. One possible assumption is **caloric equation of state**, which says that U may be expressed as a function $U(\theta)$ solely of temperature. We define the specific heat capacity at constant volume V to be

$$c_V = \frac{\mathrm{d}U}{\mathrm{d}\theta}$$

Fourier's law says that heat fluxes against the temperature gradient:

$$\vec{H} = -K \nabla \theta$$

(Under anisotropic conditions we may assume that there is cross-gradient heat flux; in this case we replace K with a tensor and contraction $\mathbf{K}\cdot$). Plugging this in, we arrive at the **heat equation** or diffusion equation

$$\rho c_V \partial_t \theta = \nabla \cdot (K \nabla \theta) + h$$

In one dimension this is

$$\rho c_v \partial_t \theta = \partial_x (K \partial_x \theta) + h$$

Chapter 2

Strong Methods

Having introduced our model equations, we now turn to methods for numerically solving differential equations. We first begin with strong methods, which solve for solutions to the non-integrated form of the desired differential equation. Both of our model equations are linaer second order PDEs:

$$\begin{cases} \rho\partial_t^2 s = \partial_x\left(\mu\partial_x s\right) \\ \partial_t\theta = \partial_x\left(\alpha\partial_x\theta\right) + h \end{cases}$$

The most general form of a linear second order PDE is given by

$$A\partial_t^2 s + 2B\partial_t\partial_x s + c\partial_x^2 s + D\partial_t s + E\partial_x s + Fs + G = 0$$

Taking inspiration from conic sections we consider the discriminant of the second derivatives, given by $B^2 - AC$. When this quantity is positive the PDE is called **hyperbolic**; when it is negative the PDE is **elliptic**; and when it is zero it is **parabolic**.

2.1 The Finite Difference Method

The finite difference method is essentially rooted in the Taylor series. Essentially we can simulate the evolution of our system forward in time by simply linearizing and taking small steps forward in time:

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{f(x+\Delta x) - f(x)}{\Delta x} + O(\Delta x)$$

This allows us to estimate the derivative of f at a point, so long as we know the values of f at points close to x. This is called the forward difference approximation. Of course we may approximate from below as well (backward difference approximation):

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{f(x) - f(x-\Delta x)}{\Delta x} + O(\Delta x)$$

A third estimate combines the above approximations:

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x} + O((\Delta x)^2)$$

This is called the centered-difference scheme, and it exhibits quadratic error, since approximating from both sides allows the linear term to cancel.

To calculate second derivatives, we can use the following first order approximation:

$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = \frac{f(x+\Delta x) - 2f(x) + f(x-\Delta x)}{\Delta x^2} + O(\Delta x)$$

The finite difference method then uses a discretization of the relevant sample space. Essentially we define a finite grid of points that we will compute. Afterward, we discretize the time steps as well and progress it forward.

Consider the 1D wave equation. Discretizing the sample space, we approximate our finite differences as

$$\partial_t^2 u(x_i, t_n) \approx \frac{u(x_i, t_n + \Delta t) - 2u(x_i, t_n) + u(x_i, t_n - \Delta t)}{\Delta t^2} = \frac{1}{\Delta t^2}\left[u_i^{n+1} - 2u_i^n + u_i^{n-1}\right]$$
$$\partial_x^2 u(x_i, t_n) \approx \frac{1}{\Delta x^2}\left[u_{i+1}^n - 2u_i^n + u_{i-1}^n\right]$$

In the homogeneous case, our PDE is

$$\partial_t^2 u(x,t) = c^2 \partial_x^2 u(x,t)$$

Substituting in, we have

$$u_i^{n+1} = \frac{c^2 \Delta t^2}{\Delta x^2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) + 2u_i^n - u_i^{n-1}$$

Now, it is important that we specify the behavior of the boundary conditions. Some options include the Dirichlet boundary conditions, which corresponds to a fixed boundary that satisfies $u(0,t) = 0$ for all t. On the other hand, we can pick the Neumann boundary conditions, which allow the boundary free movement. In other words, it expreiences no stress, so that $T(0,t) = 0$.

We also need to set initial conditions. It suffices to define $u(x,0)$ and $\partial_t u(x,0)$, which is just initial position and velocity.

2.2 Stability Analysis

Let us analyze the stability of our approximation in the second order homogeneous case. Since we are considering waves, we can suppose our solution is a plane wave on the space-time grid. In other words, we assume it is of the form

$$u_j^n = A^n e^{ikj\Delta x}$$

where k is the wave number $k = 2\pi/\lambda$ for λ the wavelength. Then substituting this into the discretization, we have

$$A^2 - 2A + 1 = A\left(e^{ik\Delta x} - 2 + e^{-k\Delta x}\right)C^2 = 2A(\cos k\Delta x - 1)C^2 = -4AC^2 \sin^2\left(\frac{1}{2}k\Delta x\right)$$

where

$$C = \frac{c\Delta t}{\Delta x}$$

The average value of $\sin^2$ is $\frac{1}{2}$. If we substitute this in, then the whole expression becomes

$$A^2 - 2A(1 - C^2) + 1 = 0$$

This is a constraint that must be satisfied for the plane wave to be simulated. Solving, we have

$$A = 1 - C^2 \pm \sqrt{(1 - C^2)^2 - 1}$$

We need $A \leq 1$, otherwise our expression for s_j^n scales with $A^n \to \infty$. This occurs precisely when $0 < c^2 \leq 1$, so our condition (known as the **Courant condition**) is

$$C \leq 1 \implies \Delta t \leq \frac{\Delta x}{c}$$

We can simplify the second order wave equation by depressing it to a first order system:

$$\rho(x)\partial_t^2 u(x,t) = \partial_x[\kappa(x)\partial_x u(x,t)]$$

becomes

$$\begin{cases} \rho(x)\partial_t v(x,t) = \partial_x T(x,t) \\ \partial_t T(x,t) = \kappa(x)\partial_x v(x,t) \end{cases}$$

where

$$\begin{cases} T(x,t) \coloneqq \kappa(x)\partial_x u(x,t) \\ v(x,t) = \partial_t u(x,t) \end{cases}$$

In this case we only have first order derivatives, so we can use the forward/backward difference or centered difference approximations. For the sake of demonstration, suppose first that we choose to use the centered difference method in space, and the forward difference method in time. In this case we have

$$v_j^{n+1} = v_j^n + \frac{\Delta t}{\rho_j}\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x}$$
$$T_j^{n+1} = T_j^n + \Delta t \cdot \kappa_j \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x}$$
$$u_j^{n+1} = u_j^n + \Delta t \cdot v_j^n$$

If we again take our plane wave solution, we can write

$$v_j^n = v_0 A^n e^{ikj\Delta x}$$
$$T_j^n = T_0 A^n e^{ikj\Delta x}$$

Plugging this into our relation (and assuming the homogeneous case for simplicity) , we can express this in matrix form as

$$\begin{bmatrix} 1 - A & \frac{i\Delta t}{\rho\Delta x}\sin k\Delta x \\ \frac{i\Delta t \cdot \kappa}{\Delta x}\sin k\Delta x & 1 - A \end{bmatrix} \begin{bmatrix} v_0 \\ T_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In order to have solutions for nontrivial v_0, T_0, our matrix needs to have zero determinant. The determinant is given by

$$\begin{vmatrix} 1-A & \frac{i\Delta t}{\rho \Delta x}\sin k\Delta x \\ \frac{i\Delta t \cdot \kappa}{\Delta x}\sin k\Delta x & 1-A \end{vmatrix} = (1-A)^2 + C^2 \sin^2 k\Delta x$$

with

$$C = c\frac{\Delta t}{\Delta x}$$

The solutions are

$$A = 1 \pm iC \sin k\Delta x$$

In order for us to avoid blowup, we need to have $|A| \leq 1$. However, in this case we have

$$|A| = \sqrt{1 + C^2 \sin k\Delta x} > 1$$

Thus we have illustrated that a scheme with a centered difference approximation in space and forward difference in time is always unstable, reagrdless of how small a timestep we choose.

To avoid this, we use a centered difference method in time instead of forward difference. This is given by

$$\begin{aligned} v_j^{n+1} &= v_j^{n-1} + \frac{2\Delta t}{\rho_j}\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} \\ T_j^{n+1} &= T_j^{n-1} + 2\Delta t \cdot \kappa_j \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} \\ u_j^{n+1} &= u_j^{n-1} + 2\Delta t \cdot v_j^n \end{aligned}$$

Under this scheme, we have the equation

$$\begin{vmatrix} 1-A^2 & \frac{2i\Delta t}{\rho \Delta x}A\sin k\Delta x \\ \frac{2i\Delta t \cdot \kappa}{\Delta x}A\sin k\Delta x & 1-A^2 \end{vmatrix} = (1-A)^2 + 4C^2A^2 \sin^2 k\Delta x$$

Applying the same determinant logic, we have

$$(1-A^2)^2 + 4C^2A^2 \sin^2 k\Delta x = 0$$

Making the substitution $\sin^2 \approx \frac{1}{2}$ as in the second order case, we have solutions with $A \leq 1$ so long as $C \leq 1$, which is precisely the same Courant condition.

One way to remedy the unstable scheme (with forward difference in time and centered difference in space) is to approximate v_j^n and T_j^n using their spatial averages:

$$\begin{aligned} v_j^{n+1} &= \frac{1}{2}\left(v_{j+1}^n + v_{j-1}^n\right) + \frac{\Delta t}{\rho_j}\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} \\ T_j^{n+1} &= \frac{1}{2}\left(T_{j+1}^n + T_{j-1}^n\right) + \Delta t \cdot \mu_j \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} \end{aligned}$$

This is known as the **Lax-Friedrich method**, which is also stable if $C \leq 1$.

2.3 Grid Dispersion

In a homogeneous solutions, velocity and stress satisfy the plane wave relations

$$v = v_0 \exp\left[i(kx - \omega t)\right]$$
$$T = T_0 \exp\left[i(kx - \omega t)\right]$$

The plane waves satisfy these equations when

$$\omega = kc = k\sqrt{\frac{\kappa}{\rho}}$$

In discretized form, we have

$$v_j^n = v_0 \exp\left[i(kj\Delta x - \omega n\Delta t)\right]$$
$$T_j^n = v_0 \exp\left[i(kj\Delta x - \omega n\Delta t)\right]$$

Substituting into the discretized wave equation, we have

$$\begin{bmatrix} -\sin\omega\Delta t & \frac{\Delta t}{\rho\Delta x}\sin k\Delta x \\ \frac{\Delta t\cdot\kappa}{\Delta x}\sin k\Delta x & -\sin\omega\Delta t \end{bmatrix} \begin{bmatrix} v_0 \\ T_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

With the vanishing determinant condition, we have the condition

$$\sin^2(\omega\Delta t) = C^2 \sin^2(k\Delta x)$$

This relates ω to k, which we can then plug into the formula for the phase speed on the grid:

$$c^{\text{grid}} = \frac{\omega}{k} = \frac{1}{k\Delta t}\arcsin\left[C\sin k\Delta x\right]$$

In particular, this is not in general equal to $c = \sqrt{\frac{\kappa}{\rho}}$, though as $\Delta x \to 0, \Delta t \to 0$, $c^{\text{grid}} \to c$. In other words, the propagation of the waves is dependent on the choice of $\Delta x, \Delta t$. This phenomenon is known as **grid dispersion**.

2.4 Staggered Grids

In the staggered grid method, we double the resolution of the grid to consider half-length time and space steps, so that our indices range over $\ldots, j-1, j-1/2, j, j+1/2, j+1, \ldots$. We then evaluate velocity on the grid with time values $n-1/2, n+1/2, \ldots$, and evaluate the stress on a grid staggered in space, with indices $j-1/2, j+1/2, \ldots$. Thus our new relation becomes

$$v_j^{n+1/2} = v_j^{n-1/2} + \frac{\Delta t}{\rho_j \Delta x}\left[\sigma_{j+1/2}^n - j_{j-1/2}^n\right]$$
$$\sigma_{j+1/2}^{n+1} = \sigma_{j+1/2}^n + \frac{\mu_{j+1/2}\Delta t}{\Delta x}\left[v_{j+1}^{n+1/2} - v_j^{n+1/2}\right]$$

Notice that the density, which is evaluated with velocity, is defined on the normal grid, while the shear modulus is evluated with stress on the staggered grid.

Now performing the same stability analysis as for the grid dispersion problem, we obtain the equation

$$\sin^2(\omega\Delta t/2) = C^2 \sin(k\Delta x/2)$$

resulting in a grid wave speed of

$$c_{\text{grid}} = \frac{1}{k\Delta t/2} \arcsin\left(C \sin(k\Delta x/2)\right)$$

In other words, we make the same number of calculations but have the benefit of only experiencing grid dispersion for half the step size.

2.5 Shallow Water Waves and Grid Anisotropy

We will next investigate another numerical artifact known as grid anisotropy, which occurs when considering plane waves in two dimensions. As an example, we can consider shallow-water waves, which are waves where the wavespeed is governed by the depth of a shallow ocean basin:

$$c = \sqrt{gh}$$

(Note that this implies that as a wave reaches the shore, it slows down, thus increasing its amplitude). From here the wave equation is given by the reasonable 2D-analogue of the 1D wave equation:

$$\partial_t^2 s = c^2\left(\partial_x^2 s + \partial_y^2 s\right)$$

We use a first-order approximation for second derivatives:

$$\partial_t^2 s = \frac{s_{j,k}^{n+1} - 2s_{j,k}^n + s_{j,k}^{n-1}}{(\Delta t)^2} + O(\Delta t)$$

with similar expressions for $\partial_x^2 s, \partial_y^2 s$. If we substitute this into the wave equation, we have

$$s_{j,k}^{n+1} - 2s_{j,k}^n + s_{j,k}^{n-1} = \left(\frac{\Delta t c_{j,k}}{\Delta x}\right)^2 \left(s_{j+1,k}^n - 2s_{j,k}^n + s_{j-1,k}^n\right) + \left(\frac{\Delta t c_{j,k}}{\Delta y}\right)^2 \left(s_{j,k+1}^n - 2s_{j,k}^n + s_{j,k-1}^n\right)$$

To analyze this system, we can consider a plane wave given by

$$s_{j,k}^n = \exp\left[i(k_x j\Delta x + k_Y k\Delta y - \omega n\Delta t)\right]$$

This satisfies the wave equation so long as

$$c = \frac{\omega}{k} = \frac{\omega}{\sqrt{k_x^2 + k_y^2}}$$

Our discretized plane wave results in the equation

$$\sin\left(\frac{1}{2}\omega\Delta t\right) = c\Delta t\left[\frac{1}{(\Delta x)^2}\sin^2\left(\frac{1}{2}k_x\Delta x\right) + \frac{1}{(\Delta y)^2}\sin^2\left(\frac{1}{2}k_y\Delta y\right)\right]^{1/2}$$

Solving for ω, we have

$$c_{\text{grid}} = \frac{\omega}{k} = \frac{2}{k\Delta t}\arcsin\left(c\Delta t\left[\frac{1}{(\Delta x)^2}\sin^2\left(\frac{1}{2}k_x\Delta x\right) + \frac{1}{(\Delta y)^2}\sin^2\left(\frac{1}{2}k_y\Delta y\right)\right]^{1/2}\right)$$

Again this is in general different from c and we only have equality in the limit $\Delta t, \Delta x, \Delta y \to 0$, so that this scheme again suffers from grid dispersion.

Differentiating ω with respect to $k_x k_y$, we find that the grid group speed has components given by

$$U_x^{\text{grid}} = \frac{\partial\omega}{\partial k_x} = \frac{c^2\Delta t}{\Delta x}\frac{\sin(k_x\Delta x)}{\sin(\omega\Delta t)}$$
$$U_y^{\text{grid}} = \frac{\partial\omega}{\partial k_y} = \frac{c^2\Delta t}{\Delta y}\frac{\sin(k_y\Delta y)}{\sin(\omega\Delta t)}$$

This set of equations implies that the group speed is dependent on wave direction, which is an undesirable phenomenon known as **grid anisotropy**.

2.6 The Heat Equation

Recall that in a one-dimensional, homogeneous medium with no heating, the heat equation is given by

$$\partial_t\theta = \alpha\partial_x^2\theta$$

Applying forward difference approximations in time and centered difference in space, we approximate this as

$$\partial_t\theta(x,t) = \frac{\theta(x,t+\Delta t) - \theta(x,t)}{\Delta t} + O(\Delta t)$$
$$\partial_x^2\theta(x,t) = \frac{\theta(x+\Delta x,t) - 2\theta(x,t) + \theta(x-\Delta x,t)}{(\Delta x)^2} + O(\Delta x)$$

Discretizing this on a grid, we obtain the scheme

$$\theta_j^{n+1} = \theta_j^n + \frac{\alpha\Delta t}{(\Delta x)^2}\left(\theta_{j+1}^n - 2\theta_j^n + \theta_{j-1}^n\right)$$

If we substitute the solution

$$\theta_j^n = A^n\exp(ikj\Delta x)$$

we then obtain the solution

$$A = 1 - 4\frac{\alpha\Delta t}{(\Delta x)^2}\sin^2\left(\frac{1}{2}k\Delta x\right)$$

In order to ensure $|A| < 1$, and again using the average bound $\sin^2 \equiv 1/2$, we need

$$\Delta t \leq \frac{(\Delta x)^2}{2\alpha}$$

While this does suggest that stable simulations exist, the timestep is quadratic in space steps. Moreover, dimesnional analysis of the heat equation suggests that the characteristic diffusion time is given by

$$\tau \sim \frac{L^2}{\alpha}$$

In order to resolve spatial accuracy, we will need $L \gg \Delta x$, so that the number of timesteps needed to simulate a characteristic diffusion time will very large:

$$N_t = \frac{\tau}{\Delta t} \sim \frac{L^2}{\alpha \Delta t} \geq \frac{N_x^2 (\Delta x)^2 2\alpha}{\alpha (\Delta x)^2} = 2N_x^2$$

In other words, the time increments needed scale quadratically with the number of spatial increments.

One way to solve this is to use a backward difference in time instead of a forward difference:

$$\partial_t \theta(x,t) = \frac{\theta(x,t) - \theta(x,t-\Delta t)}{\Delta t} + O(\Delta t)$$
$$\partial_x^2 \theta(x,t) = \frac{\theta(x+\Delta x,t) - 2\theta(x,t) + \theta(x-\Delta x,t)}{(\Delta x)^2} + O(\Delta x)$$

This results in the equation

$$-\frac{\alpha \Delta t}{(\Delta x)^2}\theta_{j-1}^n + \left[1 + 2\frac{\alpha \Delta t}{(\Delta x)^2}\right]\theta_j^n - \frac{\alpha \Delta t}{(\Delta x)^2}\theta_{j+1}^n = \theta_j^{n-1}$$

By considering the set of such equations over $j \in [2, N_x - 1]$ (and using appropriate replacement approximations at the boundaries $j = 1, N_t$), we obtain a linear system for $\{\theta_j^n\}_j$ in terms of $\{\theta_j^{n-1}\}_j$. In particular, the matrix for this system is *tridiagonal*, meaning the nonzero terms are on the main diagonal, the diagonal directly above it, and directly below it. This can be solved using various matrix solving libraries.

Again substituting our sample solution, we this time have the equation

$$A = \left[1 + 4\frac{\alpha \Delta t}{(\Delta x)^2}\sin^2\left(\frac{1}{2}k\Delta x\right)\right]^{-1}$$

In this case the power of -1 ensures that we always have $|A| < 1$, so that this scheme is unconditionally stable. This allows us to avoid using excessively many timesteps; however it is computationally expensive since it requires the solution to a linear system at every timestep. This scheme is an example of an **implicit scheme**.

2.7 The Crank-Nicolsen Scheme

The Crank-Nicolsen scheme is a scheme for solving the heat equation that uses a second order accurate centered difference scheme, while averages both the current and previous time with a first-order centered difference to calculate the spatial derivative:

$$\partial_x^2\theta \approx \left(\theta_{j+1}^n - 2\theta_j^n + \theta_{j-1}^n + \theta_{j+1}^{n-1} - 2\theta_j^{n-1} + \theta_{j-1}^{n-1}\right)$$

This provides the equation

$$\begin{aligned} -\frac{\alpha\Delta t}{2(\Delta x)^2}\theta_{j-1}^n + \left[1 + \frac{\alpha\Delta t}{(\Delta x)^2}\right]\theta_j^n - \frac{\alpha\Delta t}{2(\Delta x)^2}\theta_{j-1}^n \\ = \frac{\alpha\Delta t}{2(\Delta x)^2}\theta_{j-1}^{n-1} + \left[1 - \frac{\alpha\Delta t}{(\Delta x)^2}\right]\theta_j^{n-1} + \frac{\alpha\Delta t}{2(\Delta x)^2}\theta_{j+1}^{n-1} \end{aligned}$$

Once again we have a linear system represented by a tridiagonal matrix. Substituting in our reference solution, we then obtain

$$A = \frac{1 - 2\frac{\alpha\Delta t}{(\Delta x)^2}\sin^2\left(\frac{1}{2}k\Delta x\right)}{1 + 2\frac{\alpha\Delta t}{(\Delta x)^2}\sin^2\left(\frac{1}{2}k\Delta x\right)}$$

Here we find that the Crank-Nicolsen scheme is unconditionally stable, just like our previous impilcit scheme.

2.8 The Psuedospectral Method

Having covered the finite difference method, we now cover the psuedospectral method. While the finite difference method uses local information with high-gridpoint resolution in order to extract information about the Taylor series, the psuedospectral method leverages the Fourier transform to incorporate information from the entire space.

Recall that the Fourier transform of a function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$\tilde{f}(k) = \int_{-\infty}^{\infty} f(x)\exp(-ikx)\,\mathrm{d}x$$

and moreover that $f, \tilde{f}$ are dual to one another in the sense that

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}(k)\exp(ikx)\,\mathrm{d}k$$

Moreover if we differentiate f with respect to x, we note that in Fourier space this simply amounts to multiplication:

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}(k)ik\exp(ikx)\,\mathrm{d}k$$

In order to discretize this, we assume that our domain is some line segment with length L, and we divide it into intervals of Δx. We then divide the wave numbers into the same number of intervals:

$$x_j = j\Delta x, j = 0, \dots, N-1$$
$$k_\ell = \ell \Delta k, \ell = 0, \dots, N-1$$

Now recall that for a wave with wavelength λ

$$k = \frac{2\pi}{\lambda}$$

The highest resolution wave we can resolve on this domain has wavelength $N\Delta x$, and this corresponds to the smallest wavenumber, so we have a wavenumber spacing of

$$\Delta k = \frac{2\pi}{N\Delta x}$$

Note that this gives the identity

$$\Delta k \Delta x = \frac{2\pi}{N}$$

Thus our discretized Fourier transform gives

$$\tilde{f}(\ell \Delta k) = \Delta x \sum_{j=0}^{N-1} f(j\Delta x) \exp(-2\pi i \ell j / N)$$

and the inverse is given by

$$f(j\Delta x) = \frac{1}{N\Delta x} \sum_{\ell=0}^{N-1} \tilde{f}(\ell \Delta k) \exp(2\pi i \ell j / N)$$

(where we have used the fact that $\frac{\Delta k}{2\pi} = \frac{1}{N\Delta x}$). Taking the derivative in this discretized sense, we have

$$\frac{\mathrm{d}f}{\mathrm{d}x}(j\Delta x) = \frac{1}{N\Delta x} \sum_{\ell=0}^{N-1} i\ell \Delta k \tilde{f}(\ell \Delta k) \exp(2\pi i \ell j / N)$$

The fact that this calculation sums over all of the gridpoints implies that this approximation is also the highest order spatial derivative that is possible to resolve on the grid. Also, the existence of the **Fast Fourier Transform** (FFT) algorithm means that this can run in $N \log N$ time.

2.9 Psuedospectral Grid Dispersion

We now revisit the wave equation

$$\begin{cases} \partial_t v = \frac{1}{\rho} \partial_x \sigma \\ \partial_t \sigma = \mu \partial_x v \end{cases}$$

If we use a second order finite difference in time and a pseudospectral method in space, we get the relations

$$\frac{v_j^{n+1} - v_j^{n-1}}{2\Delta t} = \frac{1}{\rho_j} \frac{1}{N\Delta x} \sum_{\ell=0}^{N-1} i\ell\Delta k \tilde{\sigma}_\ell^n \exp(2\pi i \ell j/N)$$
$$\frac{\sigma_j^{n+1} - \sigma_j^{n-1}}{2\Delta t} = \mu_j \frac{1}{N\Delta x} \sum_{\ell=0}^{N-1} i\ell\Delta k \tilde{v}_\ell^n \exp(2\pi i \ell j/N)$$

where $\tilde{\sigma}, \tilde{v}$ are calculated at each timestep using the FFT algorithm.

It is worth considering what advantages are brought by this method over the finite difference method. In particular we investigate the issues of grid dispersion and anisotropy we observed before. Taking the Fourier transform on both sides, we have

$$\begin{cases} \dfrac{\tilde{v}_\ell^{n+1} - \tilde{v}_\ell^{n-1}}{2\Delta t} = \dfrac{1}{\rho} i\ell\Delta k \tilde{\sigma}_\ell^n \\ \dfrac{\tilde{\sigma}_\ell^{n+1} - \tilde{\sigma}_\ell^{n-1}}{2\Delta t} = \mu i\ell\Delta k \tilde{v}_\ell^n \end{cases}$$

If we introduce a standing waveg

$$\tilde{v}_\ell^n = \tilde{v} \exp(i\omega n\Delta t)$$
$$\tilde{\sigma}_j^n = \tilde{\sigma}_j \exp(i\omega n\Delta t)$$

Substituting this standing wave into our system, we obtain the matrix relation

$$\begin{bmatrix} \sin(\omega\Delta t) & -\frac{\ell\Delta k\Delta t}{\rho} \\ -\mu\ell\Delta k\Delta t & \sin(\omega\Delta t) \end{bmatrix} \begin{bmatrix} \tilde{v}_\ell \\ \tilde{\sigma}_\ell \end{bmatrix} = \vec{0}$$

The determinant condition gives us

$$\sin(\omega\Delta t) = \beta\ell\Delta k\Delta t = \frac{2\pi\beta\ell\Delta t}{N\Delta x}$$

where

$$\beta = \sqrt{\frac{\mu}{\ell}}$$

The highest value of ℓ we can observe is N, so the CFL condition on the timesteps is

$$C \leq 2\pi$$

where

$$C = \beta\frac{\Delta x}{\Delta t}$$

Immediately we observe that this condition is less strict than the $C \leq 1$ condition we obtained from the finite difference method.

Also, we note that the wavespeed is given by

$$c^{\text{grid}} = \frac{\omega}{k} = \frac{1}{k\Delta t}\arcsin(\beta k \Delta t)$$

Importantly, the grid wavespeed is not dependent on Δx as a result of this psuedospectral method, eliminating this component of grid dispersion. Because the time scheme is still finite difference, there is still grid dispersion from the time domain.

The group speed is given by

$$U = \frac{\partial \omega}{\partial k} = \frac{\beta}{\cos(\omega \Delta t)}$$

which of course is also not dependent on Δx. The phase speed should be β, but we observe an error factor of $\cos(\omega \Delta t)$. Thus we again see that the group speed differs slightly as a result of the time scheme.

In theory the psuedospectral method is extremely powerful in terms of information extracted per point. However, it runs into issues with integrating complicated boundary conditions, or with gridding irregular domains (such as surfaces with nontrivial topography). Lastly, the psuedospectral method fails to parallelize well on modern GPUs in higher dimensional cases.

Chapter 3

Weak Methods

So far we have examined strong methods, which search for strong solutions to the relevant PDEs. In PDE theory it is also of interest to search for weak solutions. A weak solution involves integrating both sides of a PDE against an arbitrarily chosen test function in one of the dimensions.

3.1 Weak Solutions

For instance, consider the wave equation

$$\rho\partial_t^2 s = \partial_x(\mu\partial_x s)$$

Suppose briefly that the spatial domain is $x \in [0,1]$. Let $\tilde{s}$ be an arbitrarily chosen **test function**. Then a strong solution to the PDE certainly satisfies

$$\begin{aligned}\int_0^1 \tilde{s}\rho\partial_t^2 s &= \int_0^1 \tilde{s}\partial_x(\mu\partial_x s)\,\mathrm{d}x \\ &= \int_0^1 \partial_x(\mu\tilde{s}s)\,\mathrm{d}x - \int_0^1 \mu\partial_x\tilde{s}\partial_x s\,\mathrm{d}x \\ &= -\int_0^1 \mu\partial_x\tilde{s}\partial_x s\,\mathrm{d}x + \mu\tilde{s}s\Big|_{x=0}^{x=1}\end{aligned}$$

Similarly for the heat equation

$$\partial_t\theta = \partial_x(\alpha\partial_x\theta) + h$$

we have

$$\int_0^1 \tilde{\theta}\partial_t\theta\,\mathrm{d}x = -\int_0^1 \alpha\partial_x\tilde{\theta}\partial_x\theta\,\mathrm{d}x + \int_0^1 \tilde{\theta}h\,\mathrm{d}x + \alpha\tilde{\theta}\partial_x\theta\Big|_{x=0}^{x=1}$$

Thus we see that a strong method satisfies an integrated form of the equation, for any test function. A weak solution is said to be one which satisfies the integrated form for any test function. In particular a strong solution is a weak solution but the converse is not generally true.

Consider the static heat equation

$$\partial_x^2 \theta + h = 0$$

We also impose the boundary conditions $\theta(1) = \theta_1$ and $\theta_x \theta(0) = -H_0$.

A strong solution is given by

$$\theta(x) = \theta_1 + (1-x)H_0 + \int_x^1 \int_0^y h(z)\,\mathrm{d}z\,\mathrm{d}y$$

This can be verified by simply observing that the derivatives are given by

$$\partial_x \theta = -H_0 - \int_0^x h(z)\,\mathrm{d}z$$

so that

$$\partial_x^2 \theta = -h(x)$$

It is also clear that the boundary conditions are satisfied.

If we instead look at the weak formulation of this problem (referring to our previous derivation of the weak heat equation and setting $\partial_t \theta = 0$), we need

$$-\int_0^1 (\partial_x \tilde{\theta})(\partial_x \theta)\,\mathrm{d}x + \int_0^1 \tilde{\theta} h\,\mathrm{d}x + \tilde{\theta}(1)\partial_x\theta(1) - \tilde{\theta}(0)\partial_x(0) = 0$$

If we choose test functions $\tilde{\theta}$ such that $\tilde{\theta}(1) = 0$ then we are given

$$-\int_0^1 (\partial_x \tilde{\theta})(\partial_x \theta)\,\mathrm{d}x + \int_0^1 \tilde{\theta} h\,\mathrm{d}x + H_0\tilde{\theta}(0) = 0$$

The finite element method is a way of discretizing weak formulations of PDEs, in the sense that the solution space is approximated with a finite dimensional subspace given some basis vectors. Specifically, we will assume that both our solution θ and the test functions $\tilde{\theta}$ are linear combinations of some finite number of basis functions with small support (called **shape functions**) $N_1, \ldots, N_n$, each of which satisfies $N_i(1) = 0$. We will also define a final shape funtion N_{n+1} such that $N_{n+1}(1) = 1$, so that we can represent solutions that satisfy the boundary condition on the right.

We will also discretize the spatial domain into some finite number of elements, and on each of these elements we will locally solve the PDE with the basis functions. Then expanding θ in terms of the basis functions, we have

$$\theta = \sum_{i=1}^{n} d_i N_i + \theta_1 N_{n+1}$$

Now, for any test function which is also a linear combination

$$\tilde{\theta} = \sum_{i=1}^{n} c_i N_i$$

the weak form is given by

$$-\sum_{i,j}^{n} c_i d_j \int_0^1 (\partial_x N_i)(\partial_x N_j)\,\mathrm{d}x - \sum_{i}^{n} c_i \theta_1 \int_0^1 (\partial_x N_i)(\partial_x N_{n+1})\,\mathrm{d}x \\ + \sum_{i}^{n} c_i \int_0^1 N_i h\,\mathrm{d}x + \sum_{i}^{n} c_i H_0 N_i(0) = 0$$

The goal is to solve for the coefficients d_i. Since the integrals are independent of the coefficients d_i, we can compute them ahead of time and represent this whole equation as a matrix expression

$$Kd = F$$

where K is an $n \times n$ matrix (called the **stiffness matrix** or **diffusivity matrix**) given by

$$K_{ij} = \int_0^1 (\partial_x N_i)(\partial_x N_j)\,\mathrm{d}x$$

and F is a $n \times 1$ vector given by

$$F_i = \int_0^1 N_i h\,\mathrm{d}x + H_0 N_i(0) - \theta_1 \int_0^1 (\partial_x N_i)(\partial_x N_{n+1})\,\mathrm{d}x$$

Inverting K then allows us to compute the coefficients d_i, and hence our approximated solution θ.

3.2 The Finite Element Method

A convenient choice of shape functions are piecewise linear functions which take the value 1 on a specific node and are zero at each other node. For instance, if we have just a single element, with the two nodes $x = 0, 1$, then we can pick the shape functions

$$N_1 = 1 - x$$
$$N_2 = x$$

With more nodes, the shape functions will in general be triangular. It is also helpful to note that even if the spacing of the nodes is not constant, each individual element is similar, up to some constant scaling value. This motivates us to view the computations as more of a local computation, by mapping each element back to a **reference element** which has boundaries $[-1, 1]$. For a specific element $[x_A, x_{A+1}]$, we can map it back to this reference element via

$$x(\xi) = x_A M_1(\xi) + x_{A+1} M_2(\xi)$$
$$M_i(\xi) = \frac{1}{2}\left[1 + (-1)^i \xi\right]$$

By choosing triangular shape functions, we ensure that for any given element, only two shape functions are supported on the element. Thus the global stiffness matrix reduces to a 2×2 local matrix k^i, which is related to the global matrix by

$$\begin{aligned} K_{11} &= k_{11}^1 \\ K_{ii} &= k_{22}^{i-1} + k_{11}^i \\ K_{i-1,i} &= k_{12}^{i-1} \end{aligned}$$

Then to calculate the coefficients of k^i, we just need to change variables from x into ξ ($a, b \in \{1, 2\}$)

$$k_{ab}^i = \int_{x_i}^{x_{i+1}} (\partial_x N_a)(\partial_x N_b)\,\mathrm{d}x = \int_{-1}^{1} (\partial_\xi N_a)(\partial_\xi N_b)\frac{\mathrm{d}\xi}{\mathrm{d}x}\,\mathrm{d}\xi = \frac{(-1)^{a+b}}{\Delta x_i}$$

where $\Delta x_i = x_{x+1} - x_i$, so that

$$k^i = \frac{1}{\Delta x_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

We can follow a similar remapping process for the heat source term, creating local 2×1 vectors f^i such that

$$\begin{aligned} F_1 &= f_1^1 + H_0 \\ F_i &= f_2^{i-1} + f_1^i \\ F_n &= f_2^{n-1} + f_1^n + \frac{\theta_1}{\Delta x_n} \end{aligned}$$

where the entries are given by

$$f_a^i = \int_{x_i}^{x_{i+1}} N_a h\,\mathrm{d}x = \frac{1}{2}\Delta x_i \int_{-1}^{1} N_a h\,\mathrm{d}\xi$$

Returning to the weak form of the heat equation, we split the integral into the components over each finite element:

$$\int_0^1 (\partial_x \tilde{\theta})(\partial_x \theta)\,\mathrm{d}x = \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} (\partial_x \tilde{\theta})(\partial_x \theta)\,\mathrm{d}x = \sum_{i=1}^{n} \int_{-1}^{1} (\partial_x \tilde{\theta})(\partial_x \theta)\frac{\mathrm{d}\xi}{\mathrm{d}x}\,\mathrm{d}\xi$$

and

$$\int_0^1 \tilde{\theta} h\,\mathrm{d}x = \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} \tilde{\theta} h\,\mathrm{d}x = \sum_{i=1}^{n} \int_{-1}^{1} \tilde{\theta} h \frac{\mathrm{d}\xi}{\mathrm{d}x}\,\mathrm{d}\xi$$

When integrating over ξ, we have performed a pullback to the reference element by

$$\begin{cases} \tilde{\theta}(x(\xi)) = c_1 M_1(\xi) + c_2 M_2(\xi) \\ \theta(x(\xi)) = d_1 M_1(\xi) + d_2 M_2(\xi) \end{cases}$$

We can also pull back the derivatives:

$$\partial_x \tilde{\theta} = c_1 \partial_\xi N_1(\xi) \frac{\mathrm{d}\xi}{\mathrm{d}x} + c_2 \partial_\xi N_2(\xi) \frac{\mathrm{d}\xi}{\mathrm{d}x}$$
$$\partial_x \theta = d_1 \partial_\xi N_1(\xi) \frac{\mathrm{d}\xi}{\mathrm{d}x} + d_2 \partial_\xi N_2(\xi) \frac{\mathrm{d}\xi}{\mathrm{d}x}$$

The explicit expressions for the derivatives can be substituted back into the integral form to give

$$\int_0^1 (\partial_x \tilde{\theta})(\partial_x \theta)\,\mathrm{d}x = \sum_{i=1}^{n} \sum_{j,k\in\{1,2\}} c_j d_k \int_{-1}^{1} \partial_\xi M_j M_k \frac{\mathrm{d}\xi}{\mathrm{d}x}\,\mathrm{d}\xi = \sum_{i=1}^{n} \sum_{j,k\in\{1,2\}} c_j d_k k^i_{jk}$$

where k^i is once again the local stiffness matrix. So we have recovered the idea that the global stiffness matrix may be obtained by piecing together the local stiffness matrices. Of course, this process will in practice require a kind of atlas which uniquely identifies local nodes, based on their relative coordinates for a given element, with a global identifier. For instance in the case of linearly spaced nodes, the ath node for element i (where $a = 1, 2$) can be mapped to the global position by

$$\psi(a, i) = i + a - 1$$

3.3 Dynamic Heat Equation

In the previous section we were only concerned with solving a static equation, where the entire space could be discretized once, without worrying about a time derivative. In particular this allowed us to ignore the issue of "stepping forward" in time with the finite element method. To study this we will consider the dynamic heat equation, where the weak form is given by

$$\int_0^1 \tilde{\theta}\partial_t \theta\,\mathrm{d}x = -\int_0^1 \alpha(\partial_x \tilde{\theta})(\partial - x\theta)\,\mathrm{d}x + \int_0^1 \tilde{\theta} h\,\mathrm{d}x + \alpha\tilde{\theta}\partial_x\theta\Big|_0^1$$

with the boundary conditions given as

$$\begin{cases} \theta(1,t) = \theta_1 \\ \partial_x \theta(0,t) = -H_0 \end{cases}$$

and the initial condition

$$\theta(x, 0) = \theta_0(x)$$

Once again we can assume that the test function and solution both lie in the span of some basis functions:

$$\theta(x,t) = \sum_{i=1}^{n} d_i(t) N_i(x) + \theta_1 N_{n+1}(x)$$
$$\tilde{\theta}(x,t) = \sum_{i=1}^{n} c_i(t) N_i(x)$$

where this time the coefficients vary in time, to represent the solution's evolution through the function space over time. Collecting terms, we now have an extra set of time-dependent terms, expressed as

$$\sum_{j=1}^{n} M_{ij}\dot{d}_j + \sum_{j=1}^{n} K_{ij}d_j = F_i$$

or in matrix notation as

$$M\dot{d} + Kd = F$$

(Here $\dot{d}$ denotes the vector $(\partial_t d_j)_j$). The coefficients of our new matrix M, called the **capacity matrix**, are given by

$$M_{ij} = \int_0^1 N_i N_j \,\mathrm{d}x$$

and the diffusivity matrix is only slightly changed as

$$K_{ij} = \int_0^1 \alpha(\partial_x N_i)(\partial_x N_j)\,\mathrm{d}x$$

The goal now is to understand the right way to implement the time dependence of this system. We can consider a **generalized trapezoidal** time scheme, which parameterizes a trapezoidal rule for interpolation:

$$\begin{aligned} M\dot{d}_{n+1} + Kd_{n+1} &= F_{n+1} \\ d_{n+1} &= d_n + \Delta t \dot{d}_{n+\eta} \\ \dot{d}_{n+\eta} &= (1-\eta)\dot{d}_n + \eta\dot{d}_{n+1} \end{aligned}$$

We can implement this using a **predictor-corrector scheme**. This strategy defines a predictor $\tilde{d}$ by

$$\tilde{d}_{n+1} = d_n + (1-\eta)\Delta t \dot{d}_n$$

Using the predictor, we can then solve for $\dot{d}_{n+1}$ in

$$(M + \eta\Delta t K)\dot{d}_{n+1} = F_{n+1} - K\tilde{d}_{n+1}$$

and then correct to calculate d_{n+1} as

$$d_{n+1} = \tilde{d}_{n+1} + \eta\Delta t\dot{d}_{n+1}$$

The choice of η recovers many of our previously used time schemes. For instance, $\eta = 0$ is the forward difference method, $\eta = 1/2$ is the central difference, and for $\eta = 1$ we have the backward difference. For $\eta \geq 1/2$ this method is unconditionally stable.

3.4 Meshes

In the 1D case, the only choice we have in designing our node points is the spacing between them. However, in the 2D and 3D cases there are many more choices we can make. In particular, elements need not be intervals; they can be quadrilaterals, tetrahedra, hexahedra, and so on. Computational tools exist for taking in models and designing a mesh based on the model. In any case, the important part about implementing FEM in these cases is that each mesh element needs a map back to the typical reference element. This can be accomplished by interpolating relative to given anchor points, where the map between the element and reference element is known for the anchor points. For instance, a quadrilateral needs at least four anchors (the vertices) to specify a map, but more anchors may be used in order to better capture meshes that are not as flat locally.

This can be adapted to account for time dependence. In the classical finite element method, the same shape functions are used to specify the local geometry and to approximate the field. When we discuss the spectral element we will see how this can be avoided.

One choice of mesh is to pick hexahedral elements. These are specified using the **Lagrange polynomials** as shape functions. These are given, after choosing $n+1$ points $\xi_0, \xi_1, \ldots, \xi_n$, by

$$\ell_\alpha^n(\xi) = \prod_{\beta \neq \alpha} \frac{\xi - \xi_\beta}{\xi_\alpha - \xi_\beta}$$

Inspection shows that $\ell_\alpha^n(\xi_\beta) = \delta_{\alpha\beta}$. For a degree 1 approximation with two anchor points $\xi_0 = -1$, $\xi_1 = +1$, the polynomials are precisely the shape functions we had before:

$$\ell_0^1(\xi) = \frac{1}{2}(1 - \xi)$$
$$\ell_1^1(\xi) = \frac{1}{2}(1 + \xi)$$

For the degree 2 polynomials with anchors $\xi_0 = -1$, $\xi_1 = 0$, $\xi_2 = +1$, the polynomials are given by

$$\ell_0^2(\xi) = \frac{1}{2}\xi(\xi - 1)$$
$$\ell_1^2(\xi) = 1 - \xi^2$$
$$\ell_2^2(\xi) = \frac{1}{2}\xi(\xi + 1)$$

These are the unique quadratics which satisfy the property that they are zero at each anchor point except a single point, where they are 1. By degree arguments it is clear that this process should yield uniquely defined polynomials for each order.

Now, given that we have chosen some set of shape functions as a basis, we will interpolate arbitrary functions on an element by pulling back from the reference element:

$$f(\vec{x}(\vec{\xi})) = \sum_{a=1}^{N} f_a N_a(\vec{\xi})$$

This in turn allows us to determine what the gradient operator looks like in our choice of function space. Specifically, we have (using index notation)

$$\nabla f = \hat{x}_i \partial_{x_i} f = \hat{x}_i f_a (\partial_{\xi_j} N_a)(\partial_{x_i} \xi_j)$$

The derivatives $\partial_{x_i} \xi_j$ may be calculated from the knowledge that for a given element e, the map from the reference element is

$$\vec{x_e}(\vec{\xi}) = \sum_{a=1}^{N} \vec{x_a} N_a(\xi) \implies \partial_{\xi_j} (x_e)_i = \sum_{a=1}^{N} \vec{x_a} \partial_{\xi_j} N_a(\vec{\xi})$$

As a result, each element is associated with a Jacobian

$$J_e = \left(\frac{\partial (x_e)_i}{\partial \xi_j} \right)_{ij}$$

which can be inverted to yield the required values of $\partial_{x_i} \xi_j$. This calculation only needs to be computed once and may be stored for each element. In the case of the Lagrange polynomials, there is a tradeoff between the number of elements and the degree of the polynomials, both of which can increase the strength of the approximation. Some influences of this choice include GPU rasterization, which motivates the chocie of polynomials with degrees that maximize the efficiency of GPU performance.

As in our study of the piecewise linear shape functions, we then need to assemble the local calculatinos into a global calculation. For instance, on meshes which have multiple anchor points, some anchor points may be shared between elements, and each element's contribution must be added to compute the global field value. This requires a map that uniquely identifies anchors that are shared between elements, so that the appropriate valeus may be added.

Another important aspect of meshing for large simulations (such as global-scale simulations) is the partitioning of meshes into submeshes for the purpose of parallel processing. In this case contiguous sets of nodes are passed to separate compute clusters with separate memory. In order to integrate each of these submeshes, boundary information needs to be passed between the clusters in order to progress in time. This is typically accomplished with a **message passing interface** (MPI), which is a protocol for passing the boundary data between nodes.

Because MPI needs to be applied at the end of each timestep, it can significnatly slow down computation if the MPI time is significant relative to the compute time for one timestep. This motivates the use of simulations where the time it takes to compute a timestep is much longer than the MPI time.

3.5 The Spectral Element Method

The spectral element method is a variant on the finite element method, which is particularly valuable in geophysics because it captures wave behavior especially well, for instance being less susceptible to grid dispersion and anisotropy.

In the spectral element method, field function spaces are approximated with the basis elements given by the Lagrange polynomials of degree n, ℓ_α^n. The degree is typically around $n \in [4, 10]$. On the other hand, the geometry is relatively simple so low order shape functions suffice. The interpolation points for the Lagrange polynomials are chosen as **Gauss-Lobatto-Legendre points** (GLL). In one dimension, these points are defined to be the roots of the function

$$(1 - \xi^2)P_n'(\xi) = 0$$

where P_n is the Legendre polynomial of degree n. This ensures that there are precisely $n+1$ interpolation points, all lying in $[-1, 1]$, and including ± 1.

For an example of why this choice is important, we consider again the weak dynamic heat equation:

$$\int_0^1 \tilde{\theta}\partial_t\theta \,\mathrm{d}x = -\int_0^1 \alpha(\partial_x\tilde{\theta})(\partial_x\theta)\,\mathrm{d}x + \int_0^1 \tilde{\theta}h\,\mathrm{d}x + \alpha\tilde{\theta}\partial_x\theta\Big|_0^1$$

Since we are working with the spectral element method, we fix some degree n, let $\xi_0, \ldots, \xi_n$ be the GLL points of degree n, and expand the functions $\tilde{\theta}, \theta$ (suppose for now we just have one element) in terms of our basis:

$$\tilde{\theta}(x(\xi)) = \sum_{\alpha=0}^{n} \tilde{\theta}_\alpha \ell_\alpha^n(\xi)$$
$$\theta(x(\xi), t) = \sum_{\beta=0}^{n} \theta_\beta(t)\ell_\beta^n(\xi)$$

Now, we can pull integrals back to the reference element:

$$\int_0^1 f(x)\,\mathrm{d}x = \int_{-1}^{+1} f(x(\xi))|J(\xi)|\,\mathrm{d}\xi$$

To approximate this integral, we use **GLL quadrature**, which says that this integral may be calculated by evaluating $|J(\xi_\alpha)|$ at the GLL points and taking a weighted sum:

$$\int_{-1}^{+1} f(x(\xi))|J(\xi)|\,\mathrm{d}\xi \approx \sum_{\gamma=0}^{n} w_\gamma |J(\xi_\gamma)| f_\gamma$$

The weights are given by

$$w_\alpha = \int_{-1}^{1} \ell_\alpha^n(\xi)\,\mathrm{d}\xi$$

The ability to approximate integrals with sums is an important motivation for the use of the spectral element method. In particular, because our functions are in terms of Lagrange polynomials, the f_γ coefficients will just be the Kronecker delta, vastly simplifying the integrals. While the advantage of the spectral element method is that GLL quadrature makes integral computations very easy, the choice of quadrature does result in some loss of numerical integration accuracy relative to other methods like Gaussian quadrature.

To see this, consider the capacitance term. Using the pullback and GLL quadrature, we approximate this as

$$\begin{aligned}\int_0^1 \tilde{\theta}\partial_t\theta \,\mathrm{d}x &= \int_{-1}^1 \tilde{\theta}(x(\xi))\partial_t\theta(x(\xi),t)|J(\xi)|\,\mathrm{d}\xi \\ &= \sum_{\gamma=0}^n w_\gamma |J(\xi_\gamma)| \left(\sum_{\alpha=0}^n \tilde{\theta}_\alpha \underbrace{\ell_\alpha^n(\xi_\gamma)}_{\delta_{\alpha\gamma}}\right)\left(\sum_{\beta=0}^n \partial_t\theta_\beta(t)\underbrace{\ell_\beta^n(\xi_\gamma)}_{\delta_{\beta\gamma}}\right) \\ &= \sum_{\gamma=0}^n w_\gamma |J(\xi_\gamma)|\tilde{\theta}_\gamma\partial_t\theta_\gamma(t)\end{aligned}$$

Because this represents mutiplication of the $\theta_\gamma(t)$ vector with a diagonal matrix, we can maintain an explicit timescheme in time rather than needing to expensively invert a matrix at each time point.

3.6 Spectral Element Method for the Wave Equation

We now consider the application of the spectral element method to the wave equation. The weak form of the wave equation is given by

$$\int_0^1 \rho\tilde{s}\partial_t^2 s \,\mathrm{d}t = -\int_0^1 \mu(\partial_x\tilde{s})(\partial_x s)\,\mathrm{d}x + \mu\tilde{s}\partial_x s\Big|_0^1$$

Since we are working in the spectral element method, we write our time-constant test function and the time-dependent solution in terms of the Lagrange polynomials of degree N passing through the points $\xi_0,\ldots,\xi_N$ in the reference element:

$$\begin{aligned}\tilde{s}(x(\xi)) &= \sum_{\alpha=0}^N \tilde{s}_\alpha \ell_\alpha^N(\xi) \\ s(x(\xi),t) &= \sum_{\beta=0}^N s_\beta(t)\ell_\beta^N(\xi)\end{aligned}$$

Using this, we write the left hand side of the wave equation on the reference element as

$$\begin{aligned}\int_0^1 \rho(x)\tilde{s}(x)\partial_t^2 s(x,t)\,\mathrm{d}x &= \int_{-1}^1 \rho(x(\xi))\tilde{s}(x(\xi))\partial_t^2 s(x(\xi),t)J(\xi)\,\mathrm{d}\xi \\ &= \sum_{\gamma=0}^N w_\gamma\rho(x_\gamma)J(x_\gamma)\left(\sum_{\alpha=0}^N \tilde{s}_\alpha\ell_\alpha^N(\xi_\gamma)\right)\left(\sum_{\beta=0}^N \ddot{s}_\beta(t)\ell_\beta^N(\xi_\gamma)\right) \\ &= \sum_{\gamma=0}^N w_\gamma\rho(x_\gamma)J(x_\gamma)\tilde{s}_\gamma\ddot{s}_\gamma(t) \\ &= \tilde{s}^T M^e(t)\ddot{s}\end{aligned}$$

where the (local) mass matrix $M^e(t)$ is diagonal and satisfies

$$M^e_{\gamma\gamma}(t) = w_\gamma \rho(x_\gamma) J(x_\gamma) \ddot{s}_\gamma(t)$$

The first term on the right hand isde is calculated by

$$\int_0^1 \mu(x)\partial_x \tilde{s}(x)\partial_x s(x,t)\,\mathrm{d}x = \int_0^1 \mu(x(\xi))\partial_x \tilde{s}(x(\xi))\partial_x s(x(\xi),t) J\,\mathrm{d}\xi$$
$$= \sum_{\gamma=0}^N w_\gamma \mu(x_\gamma) J_\gamma \left(\sum_{\alpha=0}^N \tilde{s}_\alpha \left(\ell_\alpha^N\right)'(\xi_\gamma) \frac{\mathrm{d}\xi}{\mathrm{d}x}(\xi_\gamma) \right) \left(\sum_{\beta=0}^N s_\beta(t) \left(\ell_\beta^N\right)'(\xi_\gamma) \frac{\mathrm{d}\xi}{\mathrm{d}x}(\xi_\gamma) \right)$$
$$= \tilde{s}^T K^e s$$

where

$$K^e_{\alpha\beta} = \sum_{\gamma=0}^N \frac{w_\gamma \mu(x_\gamma)}{J_\gamma} \left(\ell_\alpha^N\right)'(\xi_\gamma) \left(\ell_\beta^N\right)'(\xi_\gamma)$$

Putting this together, this is true for all $\tilde{s}$ so we are given a system of equations

$$\sum_{\beta=0}^N m_{\alpha\beta}\ddot{s}_\beta + k_{\alpha\beta} s_\beta = f_\alpha \qquad \alpha = 0, \ldots, N$$

Now, given the relation between $\ddot{s}$ and s, we need to figure out how to progress the system in time. For the heat equation we had a relation between $\dot{s}$ and s and used the predictor-corrector scheme to advance in time; in this case we will require a different scheme.

One such method is the **Newmark time scheme**. This provides solutions for equations of the form

$$M\ddot{s} + Ks = F$$

where M, K are not time-dependent. To do so, interpolates using both $\dot{s}$ and $\ddot{s}$ concurrently. With γ, β as paramters, we define

$$s_{n+1} = s_n + \Delta t \dot{s}_n + \frac{1}{2}(\Delta t)^2 \left[(1-2\beta)\ddot{s}_n + 2\beta \ddot{s}_{n+1}\right]$$
$$\dot{s}_{n+1} = \dot{s}_n + (1-\gamma)\Delta t \ddot{s}_n + \gamma \Delta t \ddot{s}_{n+1}$$
$$\ddot{s}_{n+1} = M^{-1}(F_{n+1} - K s_{n+1})$$

Choosing $\gamma = \frac{1}{2}, \beta = 0$ results in an explicit scheme. To implement this scheme, we progress in steps. First, we calculate a predictor for each value:

$$s_{n+1} = s_n + \Delta t \dot{s}_n + \frac{1}{2}(\Delta t)^2 \ddot{s}_n$$
$$\dot{s}^*_{n+1} = \dot{s}_n + \frac{1}{2}\Delta t \ddot{s}_n$$
$$\ddot{s}^*_{n+1} = 0$$

Now, we calculate

$$\Delta\alpha = M^{-1} K s_{n+1}$$

which we use to correct the predictors:

$$\ddot{s}_{n+1} = \ddot{s}^*_{n+1} + \Delta\alpha = \Delta\alpha$$
$$\dot{s}_{n+1} = \dot{s}^*_{n+1} + \frac{1}{2}\Delta t \ddot{s}_{n+1}$$

3.7 Spectral Element Method in 3D

Now suppose we wish to implement the spectral element method in 3 dimensions. We will need to do a bit of backgroudn so that we can adapt this method.

Rather than letting the reference element be $[-1, 1]$, we will now let the reference element be the cube $[-1, 1]^3$, and allow our actual elements to be arbitrary hexahedral elements. If we write $\vec{\xi} = (\mu, \eta, \zeta)$, then can set our basis elements to be the product of Lagrange polynomials in the orthogonal directions, so that any function is written as

$$f(\vec{x}(\vec{\xi})) = \sum_{\alpha=0}^{N_\alpha} \sum_{\beta=0}^{N_\beta} \sum_{\gamma=0}^{N_\gamma} f_{\alpha\beta\gamma} \ell_\alpha^{N_\alpha}(\mu) \ell_\beta^{N_\beta}(\eta) \ell_\gamma^{N_\gamma}(\zeta)$$

where

$$f_{\alpha\beta\gamma} = f(\vec{x}(\mu_\alpha, \eta_\beta, \zeta_\gamma))$$

In order to take the gradient of a function on the reference element, we write

$$\nabla f(\mu, \eta, \zeta) = \sum_{i=1}^{3} \hat{x}_i \sum_{\alpha,\beta,\gamma} f_{\alpha\beta\gamma} \left[\left(\ell_\alpha^{N_\alpha}\right)'(\mu) \frac{\mathrm{d}\mu}{\mathrm{d}x_i} \ell_\beta^{N_\beta}(\eta) \ell_\gamma^{N_\gamma}(\zeta) \right.$$
$$\left. + \left(\ell_\beta^{N_\beta}\right)'(\eta) \frac{\mathrm{d}\eta}{\mathrm{d}x_i} \ell_\alpha^{N_\alpha}(\mu) \ell_\gamma^{N_\gamma}(\zeta) + \left(\ell_\gamma^{N_\gamma}\right)'(\zeta) \frac{\mathrm{d}\zeta}{\mathrm{d}x_i} \ell_\alpha^{N_\alpha}(\mu) \ell_\beta^{N_\beta}(\eta) \right]$$

As in the 1D case, this simplifies immensely when we evaluate at the GLL points.

Also as before, we express the geometry in terms of a separate set of shape functions:

$$\vec{x}(\vec{\xi}) = \sum_{\alpha,\beta,\gamma} \vec{x}^a N_a(\vec{\xi})$$

To perform quadrature in three dimensions, we calculate iterated integrals:

$$\int_V f(\vec{x})\, \mathrm{d}^3\vec{x} = \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} f(\vec{x}(\vec{\xi})) \left|\frac{\partial \vec{x}}{\partial \vec{\xi}}\right| \mathrm{d}^3\vec{\xi}$$
$$= \sum_{\alpha,\beta,\gamma} w_\alpha w_\beta w_\gamma f_{\alpha\beta\gamma} J_{\alpha\beta\gamma}$$

We now demonstrate the application of these tools to the 3D wave equation. First, recall that the strong form of the wave equation is

$$\rho \partial_t^2 \vec{s} = \nabla \cdot \boldsymbol{\sigma} + \vec{f}$$

where Hooke's law gives the relation

$$\boldsymbol{\sigma} = \mathbf{c} : \boldsymbol{\epsilon}$$

where

$$\boldsymbol{\epsilon} = \frac{1}{2}\left[\nabla \vec{s} + (\nabla \vec{s})^T\right]$$

To derive the weak form, we integrate by parts after multiplying by a test function $\tilde{\vec{s}}$:

$$\int_V \rho \tilde{\vec{s}} \cdot \partial_t^2 \vec{s} \, \mathrm{d}V = \int_V \boldsymbol{\sigma} : \nabla \tilde{\vec{s}} \, \mathrm{d}V - \int_S \tilde{\vec{s}} \boldsymbol{\sigma} \cdot \hat{n} \, \mathrm{d}S + \int_V \tilde{\vec{s}} \cdot \vec{f} \, \mathrm{d}V$$

This is an extremely convenient form for seismology, because the typical boundary condition for seismology is the stress-free condition $\boldsymbol{\sigma} : \hat{n} = 0$. This is easily implemented here simply by dropping the second term on the right. Isolating the boundary conditions like this also allows phase transitions over property discontinuities to be addressed by meshing elements to connect on the boundary. At the connection points, different properties can be assigned to the same point based on which element is being integrated over, thus implementing the discontinuity.

Now, we use the expansions of $\vec{s}$ and $\tilde{\vec{s}}$ in terms of the Lagrange polynomials:

$$\tilde{\vec{s}}(\vec{x}(\vec{\xi})) = \sum_{i=1}^{3} \hat{x}_i \sum_{\alpha=0}^{N_\alpha} \sum_{\beta=0}^{N_\beta} \sum_{\gamma=0}^{N_\gamma} \tilde{s}_i^{\alpha\beta\gamma} \ell_\alpha^{N_\alpha}(\mu) \ell_\beta^{N_\beta}(\eta) \ell_\gamma^{N_\gamma}(\zeta)$$

$$\vec{s}(\vec{x}(\vec{\xi}), t) = \sum_{i=1}^{3} \hat{x}_i \sum_{\alpha,\beta,\gamma} s_i^{\alpha\beta\gamma}(t) \ell_\alpha^{N_\alpha}(\mu) \ell_\beta^{N_\beta}(\eta) \ell_\gamma^{N_\gamma}(\zeta)$$

Now using GLL quadrature, the left hand side of this equality becomes

$$\int_V \rho \tilde{\vec{s}} \cdot \partial_t^2 \vec{s} \, \mathrm{d}V = \sum_{\alpha,\beta,\gamma} w_\alpha w_\beta w_\gamma J_{\alpha\beta\gamma} \rho_{\alpha\beta\gamma} \sum_{i=1}^{3} \tilde{s}_i^{\alpha\beta\gamma} \ddot{s}_i^{\alpha\beta\gamma}(t)$$

If we select $\tilde{\vec{s}}$ such that exactly one of $\tilde{s}_i^{\alpha\beta\gamma}$ is nonzero, then will end up with a system of equations so that this expression looks like the matrix multiplication

$$M \ddot{\vec{s}}(t)$$
$$M_{\alpha\beta\gamma,\alpha\beta\gamma} = w_\alpha w_\beta w_\gamma J_{\alpha\beta\gamma} \rho_{\alpha\beta\gamma}$$

where M is a diagonal matrix.

Now, for the first term on the right hand side, we calculate

$$\begin{aligned}\int_V \boldsymbol{\sigma} : \nabla \tilde{\vec{s}} \, \mathrm{d}V &= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \boldsymbol{\sigma} : \nabla_{\vec{x}} \tilde{\vec{s}} J \, \mathrm{d}^3\xi \\ &= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \mathbf{P} : \left(\nabla_{\vec{\xi}} \tilde{\vec{s}}\right)^T \mathrm{d}^3\xi\end{aligned}$$

where **P** is the first **Piola-Kirchhoff stress** given by

$$\mathbf{P}(\vec{\xi}, t) = J(\vec{\xi}, t)\boldsymbol{\sigma}(\vec{x}(\vec{\xi}), t) \cdot \mathbf{F}^{-1}(\vec{\xi}, t)$$

where

$$\mathbf{F}^{-1} = \left(\frac{\partial \vec{\xi}}{\partial \vec{x}}\right)^T$$

Definitions