

# The Finite Element Method in Geodynamics

C. Thieulot

January 15, 2019

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Acknowledgments . . . . .	4
1.2	Essential literature . . . . .	4
1.3	Installation . . . . .	4
<b>2</b>	<b>The physical equations of Fluid Dynamics</b>	<b>5</b>
2.1	The heat transport equation - energy conservation equation . . . . .	5
2.2	The momentum conservation equations . . . . .	6
2.3	The mass conservation equations . . . . .	6
2.4	The equations in ASPECT manual . . . . .	6
2.5	the Boussinesq approximation: an Incompressible flow . . . . .	8
<b>3</b>	<b>The building blocks of the Finite Element Method</b>	<b>9</b>
3.1	Numerical integration . . . . .	9
3.1.1	in 1D - theory . . . . .	9
3.1.2	in 1D - examples . . . . .	11
3.1.3	in 2D/3D - theory . . . . .	12
3.2	The mesh . . . . .	12
3.3	A bit of FE terminology . . . . .	12
3.4	Elements and basis functions in 1D . . . . .	13
3.4.1	Linear basis functions ( $Q_1$ ) . . . . .	13
3.4.2	Quadratic basis functions ( $Q_2$ ) . . . . .	13
3.4.3	Cubic basis functions ( $Q_3$ ) . . . . .	14
3.5	Elements and basis functions in 2D . . . . .	16
3.5.1	Bilinear basis functions in 2D ( $Q_1$ ) . . . . .	17
3.5.2	Biquadratic basis functions in 2D ( $Q_2$ ) . . . . .	19
3.5.3	Bicubic basis functions in 2D ( $Q_3$ ) . . . . .	20
<b>4</b>	<b>Solving the Stokes equations with the FEM</b>	<b>22</b>
4.1	strong and weak forms . . . . .	22
4.2	The penalty approach . . . . .	22
4.3	The mixed FEM . . . . .	25
<b>5</b>	<b>Solving the elastic equations with the FEM</b>	<b>26</b>
<b>6</b>	<b>Additional techniques</b>	<b>27</b>
6.1	The method of manufactured solutions . . . . .	28
6.2	Assigning values to quadrature points . . . . .	28
6.3	The value of the timestep . . . . .	31
6.4	Picard and Newton . . . . .	31
6.5	The SUPG formulation for the energy equation . . . . .	31
6.6	Tracking materials and/or interfaces . . . . .	31
6.7	Dealing with a free surface . . . . .	31

6.8	Matrix (Sparse) storage . . . . .	32
6.8.1	2D domain - One degree of freedom per node . . . . .	32
6.8.2	2D domain - Two degrees of freedom per node . . . . .	33
6.8.3	in fieldstone . . . . .	34
6.9	Mesh generation . . . . .	35
6.10	Visco-Plasticity . . . . .	38
6.10.1	Tensor invariants . . . . .	38
6.10.2	Scalar viscoplasticity . . . . .	39
6.10.3	about the yield stress value $Y$ . . . . .	39
6.11	Pressure smoothing . . . . .	40
6.12	Pressure scaling . . . . .	42
6.13	Pressure normalisation . . . . .	43
6.14	Convergence criterion for nonlinear iterations . . . . .	44
6.15	The choice of solvers . . . . .	45
6.15.1	The Schur complement approach . . . . .	45
6.16	The GMRES approach . . . . .	47
6.17	The consistent boundary flux (CBF) . . . . .	48
6.18	Static condensation . . . . .	49
6.19	Exporting data to vtk format . . . . .	50
7	<b>fieldstone: simple analytical solution</b>	53
8	<b>fieldstone: Stokes sphere</b>	56
9	<b>fieldstone: Convection in a 2D box</b>	57
10	<b>fieldstone: The lid driven cavity</b>	60
10.1	the lid driven cavity problem ( $ldc=0$ ) . . . . .	60
10.2	the lid driven cavity problem - regularisation I ( $ldc=1$ ) . . . . .	60
10.3	the lid driven cavity problem - regularisation II ( $ldc=2$ ) . . . . .	60
11	<b>fieldstone: solcx benchmark</b>	62
12	<b>fieldstone: solkz benchmark</b>	64
13	<b>fieldstone: solvi benchmark</b>	65
14	<b>fieldstone: the indentor benchmark</b>	67
15	<b>fieldstone: the annulus benchmark</b>	69
16	<b>fieldstone: stokes sphere (3D) - penalty</b>	71
17	<b>fieldstone: stokes sphere (3D) - mixed formulation</b>	72
18	<b>fieldstone: consistent pressure recovery</b>	73
19	<b>fieldstone: the Particle in Cell technique (1) - the effect of averaging</b>	75
20	<b>fieldstone: solving the full saddle point problem</b>	79
21	<b>fieldstone: solving the full saddle point problem in 3D</b>	81
21.0.1	Constant viscosity . . . . .	83
21.0.2	Variable viscosity . . . . .	84
22	<b>fieldstone: solving the full saddle point problem with <math>Q_2 \times Q_1</math> elements</b>	86
23	<b>fieldstone: solving the full saddle point problem with <math>Q_3 \times Q_2</math> elements</b>	88

<b>24 fieldstone: the Busse benchmark</b>	<b>90</b>
<b>25 fieldstone: The non-conforming <math>Q_1 \times P_0</math> element</b>	<b>92</b>
<b>26 fieldstone: The stabilised <math>Q_1 \times Q_1</math> element</b>	<b>93</b>
<b>27 fieldstone: compressible flow (1)</b>	<b>94</b>
<b>28 fieldstone: compressible flow (2)</b>	<b>97</b>
28.1 The physics . . . . .	97
28.2 The numerics . . . . .	97
28.3 The experimental setup . . . . .	99
28.4 Scaling . . . . .	99
28.5 Conservation of energy 1 . . . . .	100
28.5.1 under BA and EBA approximations . . . . .	100
28.5.2 under no approximation at all . . . . .	101
28.6 Conservation of energy 2 . . . . .	101
28.7 The problem of the onset of convection . . . . .	102
28.8 results - BA - $Ra = 10^4$ . . . . .	104
28.9 results - BA - $Ra = 10^5$ . . . . .	108
28.10 results - BA - $Ra = 10^6$ . . . . .	110
28.11 results - EBA - $Ra = 10^4$ . . . . .	111
28.12 results - EBA - $Ra = 10^5$ . . . . .	115
28.13 Onset of convection . . . . .	117
<b>29 fieldstone: Rayleigh-Taylor instability (1)</b>	<b>118</b>
<b>30 fieldstone: Slab detachment benchmark (1)</b>	<b>120</b>

# 1 Introduction

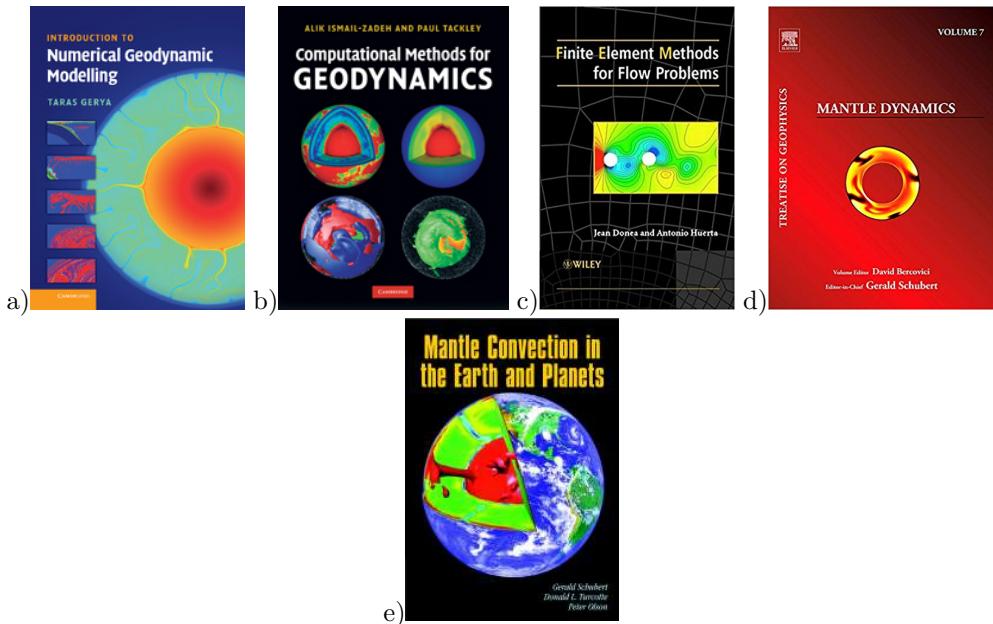
## WARNING: this is work in progress

practical hands-on approach  
as little as possible jargon  
no mathematical proof  
no optimised codes (readability over efficiency). avoiding as much as possible to have to look elsewhere.  
very sequential, so unavoidable repetitions (jacobian, shape functions)  
FE is one of several methods.  
All the python scripts and this document are freely available at <https://github.com/cedrikt/fieldstone>.

### 1.1 Acknowledgments

Jean Braun, Philippe Fullsack, Arie van den Berg. Lukas van de Wiel. Robert Myhill. Menno, Anne Too many BSc and MSc students to name individually, although Job Mos did produce the very first version of fieldstone as part of his MSc thesis. The ASPECT team in general and Wolfgang Bangerth and Timo Heister in particular.

### 1.2 Essential literature



### 1.3 Installation

```
python3.6 -m pip install --user numpy scipy matplotlib
```

## 2 The physical equations of Fluid Dynamics

Symbol	meaning	unit
$t$	Time	s
$x, y, z$	Cartesian coordinates	m
$\mathbf{v}$	velocity vector	$\text{m} \cdot \text{s}^{-1}$
$\rho$	mass density	$\text{kg/m}^3$
$\eta$	dynamic viscosity	$\text{Pa} \cdot \text{s}$
$\lambda$	penalty parameter	$\text{Pa} \cdot \text{s}$
$T$	temperature	K
$\nabla$	gradient operator	$\text{m}^{-1}$
$\nabla \cdot$	divergence operator	$\text{m}^{-1}$
$p$	pressure	Pa
$\dot{\epsilon}(\mathbf{v})$	strain rate tensor	$\text{s}^{-1}$
$\alpha$	thermal expansion coefficient	$\text{K}^{-1}$
$k$	thermal conductivity	$\text{W}/(\text{m} \cdot \text{K})$
$C_p$	Heat capacity	J/K
$H$	intrinsic specific heat production	W/kg
$\beta_T$	isothermal compressibility	$\text{Pa}^{-1}$
$\boldsymbol{\tau}$	deviatoric stress tensor	Pa
$\boldsymbol{\sigma}$	full stress tensor	Pa

### 2.1 The heat transport equation - energy conservation equation

Let us start from the heat transport equation as shown in Schubert, Turcotte and Olson [74]:

$$\rho C_p \frac{DT}{Dt} - \alpha T \frac{Dp}{Dt} = \nabla \cdot k \nabla T + \Phi + \rho H$$

with  $D/Dt$  being the total derivatives so that

$$\frac{DT}{Dt} = \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \quad \frac{Dp}{Dt} = \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p$$

Solving for temperature, this equation is often rewritten as follows:

$$\rho C_p \frac{DT}{Dt} - \nabla \cdot k \nabla T = \alpha T \frac{Dp}{Dt} + \Phi + \rho H$$

A note on the shear heating term  $\Phi$ : In many publications,  $\Phi$  is given by  $\Phi = \tau_{ij} \partial_j u_i = \boldsymbol{\tau} : \nabla \mathbf{v}$ .

$$\begin{aligned}
\Phi &= \tau_{ij} \partial_j u_i \\
&= 2\eta \dot{\epsilon}_{ij}^d \partial_j u_i \\
&= 2\eta \frac{1}{2} (\dot{\epsilon}_{ij}^d \partial_j u_i + \dot{\epsilon}_{ji}^d \partial_i u_j) \\
&= 2\eta \frac{1}{2} (\dot{\epsilon}_{ij}^d \partial_j u_i + \dot{\epsilon}_{ij}^d \partial_i u_j) \\
&= 2\eta \dot{\epsilon}_{ij}^d \frac{1}{2} (\partial_j u_i + \partial_i u_j) \\
&= 2\eta \dot{\epsilon}_{ij}^d \dot{\epsilon}_{ij} \\
&= 2\eta \dot{\epsilon}^d : \dot{\epsilon} \\
&= 2\eta \dot{\epsilon}^d : \left( \dot{\epsilon}^d + \frac{1}{3} (\nabla \cdot \mathbf{v}) \mathbf{1} \right) \\
&= 2\eta \dot{\epsilon}^d : \dot{\epsilon}^d + 2\eta \dot{\epsilon}^d : \mathbf{1} (\nabla \cdot \mathbf{v}) \\
&= 2\eta \dot{\epsilon}^d : \dot{\epsilon}^d
\end{aligned} \tag{1}$$

Finally

$$\Phi = \boldsymbol{\tau} : \nabla \mathbf{v} = 2\eta \dot{\epsilon}^d : \dot{\epsilon}^d = 2\eta ((\dot{\epsilon}_{xx}^d)^2 + (\dot{\epsilon}_{yy}^d)^2 + 2(\dot{\epsilon}_{xy}^d)^2)$$

## 2.2 The momentum conservation equations

Because the Prandlt number is virtually zero in Earth science applications the Navier Stokes equations reduce to the Stokes equation:

$$\nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{g} = 0$$

Since

$$\boldsymbol{\sigma} = -p\mathbf{1} + \boldsymbol{\tau}$$

it also writes

$$-\nabla p + \nabla \cdot \boldsymbol{\tau} + \rho \mathbf{g} = 0$$

Using the relationship  $\boldsymbol{\tau} = 2\eta\dot{\boldsymbol{\varepsilon}}^d$  we arrive at

$$-\nabla p + \nabla \cdot (2\eta\dot{\boldsymbol{\varepsilon}}^d) + \rho \mathbf{g} = 0$$

## 2.3 The mass conservation equations

The mass conservation equation is given by

$$\frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{v} = 0$$

or,

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

In the case of an incompressible flow, then  $\partial \rho / \partial t = 0$  and  $\nabla \rho = 0$ , i.e.  $D\rho/Dt = 0$  and the remaining equation is simply:

$$\nabla \cdot \mathbf{v} = 0$$

## 2.4 The equations in ASPECT manual

The following is lifted off the ASPECT manual. We focus on the system of equations in a  $d = 2$ - or  $d = 3$ -dimensional domain  $\Omega$  that describes the motion of a highly viscous fluid driven by differences in the gravitational force due to a density that depends on the temperature. In the following, we largely follow the exposition of this material in Schubert, Turcotte and Olson [74].

Specifically, we consider the following set of equations for velocity  $\mathbf{u}$ , pressure  $p$  and temperature  $T$ :

$$-\nabla \cdot \left[ 2\eta \left( \dot{\boldsymbol{\varepsilon}}(\mathbf{v}) - \frac{1}{3}(\nabla \cdot \mathbf{v})\mathbf{1} \right) \right] + \nabla p = \rho \mathbf{g} \quad \text{in } \Omega, \quad (2)$$

$$\nabla \cdot (\rho \mathbf{v}) = 0 \quad \text{in } \Omega, \quad (3)$$

$$\begin{aligned} \rho C_p \left( \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right) - \nabla \cdot k \nabla T &= \rho H \\ &+ 2\eta \left( \dot{\boldsymbol{\varepsilon}}(\mathbf{v}) - \frac{1}{3}(\nabla \cdot \mathbf{v})\mathbf{1} \right) : \left( \dot{\boldsymbol{\varepsilon}}(\mathbf{v}) - \frac{1}{3}(\nabla \cdot \mathbf{v})\mathbf{1} \right) \\ &+ \alpha T (\mathbf{v} \cdot \nabla p) \end{aligned} \quad (4)$$

in  $\Omega$ ,

where  $\dot{\boldsymbol{\varepsilon}}(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T)$  is the symmetric gradient of the velocity (often called the *strain rate*).

In this set of equations, (104) and (105) represent the compressible Stokes equations in which  $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$  is the velocity field and  $p = p(\mathbf{x}, t)$  the pressure field. Both fields depend on space  $\mathbf{x}$  and time  $t$ . Fluid flow is driven by the gravity force that acts on the fluid and that is proportional to both the density of the fluid and the strength of the gravitational pull.

Coupled to this Stokes system is equation (106) for the temperature field  $T = T(\mathbf{x}, t)$  that contains heat conduction terms as well as advection with the flow velocity  $\mathbf{v}$ . The right hand side terms of this equation correspond to

- internal heat production for example due to radioactive decay;
- friction (shear) heating;
- adiabatic compression of material;

In order to arrive at the set of equations that ASPECT solves, we need to

- neglect the  $\partial p / \partial t$ . **WHY?**
- neglect the  $\partial \rho / \partial t$ . **WHY?**

from equations above.

---

Also, their definition of the shear heating term  $\Phi$  is:

$$\Phi = k_B(\nabla \cdot \mathbf{v})^2 + 2\eta\dot{\epsilon}^d : \dot{\epsilon}^d$$

For many fluids the bulk viscosity  $k_B$  is very small and is often taken to be zero, an assumption known as the Stokes assumption:  $k_B = \lambda + 2\eta/3 = 0$ . Note that  $\eta$  is the dynamic viscosity and  $\lambda$  the second viscosity. Also,

$$\boldsymbol{\tau} = 2\eta\dot{\epsilon} + \lambda(\nabla \cdot \mathbf{v})\mathbf{1}$$

but since  $k_B = \lambda + 2\eta/3 = 0$ , then  $\lambda = -2\eta/3$  so

$$\boldsymbol{\tau} = 2\eta\dot{\epsilon} - \frac{2}{3}\eta(\nabla \cdot \mathbf{v})\mathbf{1} = 2\eta\dot{\epsilon}^d$$

## 2.5 the Boussinesq approximation: an Incompressible flow

[from aspect manual] The Boussinesq approximation assumes that the density can be considered constant in all occurrences in the equations with the exception of the buoyancy term on the right hand side of (104). The primary result of this assumption is that the continuity equation (105) will now read

$$\nabla \cdot \mathbf{v} = 0$$

This implies that the strain rate tensor is deviatoric. Under the Boussinesq approximation, the equations are much simplified:

$$-\nabla \cdot [2\eta\dot{\epsilon}(\mathbf{v})] + \nabla p = \rho\mathbf{g} \quad \text{in } \Omega, \quad (5)$$

$$\nabla \cdot (\rho\mathbf{v}) = 0 \quad \text{in } \Omega, \quad (6)$$

$$\rho_0 C_p \left( \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right) - \nabla \cdot k \nabla T = \rho H \quad \text{in } \Omega \quad (7)$$

Note that all terms on the rhs of the temperature equations have disappeared, with the exception of the source term.

### 3 The building blocks of the Finite Element Method

#### 3.1 Numerical integration

As we will see later, using the Finite Element method to solve problems involves computing integrals which are more often than not too complex to be computed analytically/exactly. We will then need to compute them numerically.

[wiki] In essence, the basic problem in numerical integration is to compute an approximate solution to a definite integral

$$\int_a^b f(x)dx$$

to a given degree of accuracy. This problem has been widely studied [?] and we know that if  $f(x)$  is a smooth function, and the domain of integration is bounded, there are many methods for approximating the integral to the desired precision.

There are several reasons for carrying out numerical integration.

- The integrand  $f(x)$  may be known only at certain points, such as obtained by sampling. Some embedded systems and other computer applications may need numerical integration for this reason.
- A formula for the integrand may be known, but it may be difficult or impossible to find an antiderivative that is an elementary function. An example of such an integrand is  $f(x) = \exp(-x^2)$ , the antiderivative of which (the error function, times a constant) cannot be written in elementary form.
- It may be possible to find an antiderivative symbolically, but it may be easier to compute a numerical approximation than to compute the antiderivative. That may be the case if the antiderivative is given as an infinite series or product, or if its evaluation requires a special function that is not available.

##### 3.1.1 in 1D - theory

The simplest method of this type is to let the interpolating function be a constant function (a polynomial of degree zero) that passes through the point  $((a+b)/2, f((a+b)/2))$ .

This is called the midpoint rule or rectangle rule.

$$\int_a^b f(x)dx \simeq (b-a)f\left(\frac{a+b}{2}\right)$$

insert here figure

The interpolating function may be a straight line (an affine function, i.e. a polynomial of degree 1) passing through the points  $(a, f(a))$  and  $(b, f(b))$ .

This is called the trapezoidal rule.

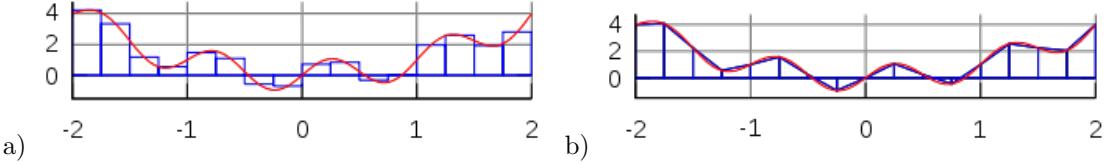
$$\int_a^b f(x)dx \simeq (b-a)\frac{f(a) + f(b)}{2}$$

insert here figure

For either one of these rules, we can make a more accurate approximation by breaking up the interval  $[a, b]$  into some number  $n$  of subintervals, computing an approximation for each subinterval, then adding up all the results. This is called a composite rule, extended rule, or iterated rule. For example, the composite trapezoidal rule can be stated as

$$\int_a^b f(x)dx \simeq \frac{b-a}{n} \left( \frac{f(a)}{2} + \sum_{k=1}^{n-1} f(a + k \frac{b-a}{n}) + \frac{f(b)}{2} \right)$$

where the subintervals have the form  $[kh, (k+1)h]$ , with  $h = (b-a)/n$  and  $k = 0, 1, 2, \dots, n-1$ .



The interval  $[-2, 2]$  is broken into 16 sub-intervals. The blue lines correspond to the approximation of the red curve by means of a) the midpoint rule, b) the trapezoidal rule.

There are several algorithms for numerical integration (also commonly called 'numerical quadrature', or simply 'quadrature'). Interpolation with polynomials evaluated at equally spaced points in  $[a, b]$  yields the NewtonCotes formulas, of which the rectangle rule and the trapezoidal rule are examples. If we allow the intervals between interpolation points to vary, we find another group of quadrature formulas, such as the Gauss(ian) quadrature formulas. A Gaussian quadrature rule is typically more accurate than a NewtonCotes rule, which requires the same number of function evaluations, if the integrand is smooth (i.e., if it is sufficiently differentiable).

An  $n$ -point Gaussian quadrature rule, named after Carl Friedrich Gauss, is a quadrature rule constructed to yield an exact result for polynomials of degree  $2n - 1$  or less by a suitable choice of the points  $x_i$  and weights  $w_i$  for  $i = 1, \dots, n$ .

The domain of integration for such a rule is conventionally taken as  $[-1, 1]$ , so the rule is stated as

$$\int_{-1}^{+1} f(x) dx = \sum_{i_q=1}^n w_{i_q} f(x_{i_q})$$

In this formula the  $x_{i_q}$  coordinate is the  $i$ -th root of the Legendre polynomial  $P_n(x)$ .

It is important to note that a Gaussian quadrature will only produce good results if the function  $f(x)$  is well approximated by a polynomial function within the range  $[-1, 1]$ . As a consequence, the method is not, for example, suitable for functions with singularities.

Number of points, $n$	Points, $x_i$	Weights, $w_i$
1	0	2
2	$\pm\sqrt{\frac{1}{3}}$	1
3	0	$\frac{8}{9}$
	$\pm\sqrt{\frac{3}{5}}$	$\frac{5}{9}$
4	$\pm\sqrt{\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18+\sqrt{30}}{36}$
	$\pm\sqrt{\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18-\sqrt{30}}{36}$
5	0	$\frac{128}{225}$
	$\pm\frac{1}{3}\sqrt{5 - 2\sqrt{\frac{10}{7}}}$	$\frac{322+13\sqrt{70}}{900}$
	$\pm\frac{1}{3}\sqrt{5 + 2\sqrt{\frac{10}{7}}}$	$\frac{322-13\sqrt{70}}{900}$

Gauss-Legendre points and their weights.

As shown in the above table, it can be shown that the weight values must fulfill the following condition:

$$\sum_{i_q} w_{i_q} = 2 \tag{8}$$

and it is worth noting that all quadrature point coordinates are symmetrical around the origin.

Since most quadrature formula are only valid on a specific interval, we now must address the problem of their use outside of such intervals. The solution turns out to be quite simple: one must carry out a change of variables from the interval  $[a, b]$  to  $[-1, 1]$ .

We then consider the reduced coordinate  $r \in [-1, 1]$  such that

$$r = \frac{2}{b-a}(x-a) - 1$$

This relationship can be reversed such that when  $r$  is known, its equivalent coordinate  $x \in [a, b]$  can be computed:

$$x = \frac{b-a}{2}(1+r) + a$$

From this it follows that

$$dx = \frac{b-a}{2}dr$$

and then

$$\int_a^b f(x)dx = \frac{b-a}{2} \int_{-1}^{+1} f(r)dr \simeq \frac{b-a}{2} \sum_{i_q=1}^n w_{i_q} f(r_{i_q})$$

### 3.1.2 in 1D - examples

**example 1** Since we know how to carry out any required change of variables, we choose for simplicity  $a = -1$ ,  $b = +1$ . Let us take for example  $f(x) = \pi$ . Then we can compute the integral of this function over the interval  $[a, b]$  exactly:

$$I = \int_{-1}^{+1} f(x)dx = \pi \int_{-1}^{+1} dx = 2\pi$$

We can now use a Gauss-Legendre formula to compute this same integral:

$$I_{gq} = \int_{-1}^{+1} f(x)dx = \sum_{i_q=1}^{n_q} w_{i_q} f(x_{i_q}) = \sum_{i_q=1}^{n_q} w_{i_q} \pi = \pi \underbrace{\sum_{i_q=1}^{n_q} w_{i_q}}_{=2} = 2\pi$$

where we have used the property of the weight values of Eq.(8). Since the actual number of points was never specified, this result is valid for all quadrature rules.

**example 2** Let us now take  $f(x) = mx + p$  and repeat the same exercise:

$$I = \int_{-1}^{+1} f(x)dx = \int_{-1}^{+1} (mx + p)dx = [\frac{1}{2}mx^2 + px]_{-1}^{+1} = 2p$$

$$I_{gq} = \int_{-1}^{+1} f(x)dx = \sum_{i_q=1}^{n_q} w_{i_q} f(x_{i_q}) = \sum_{i_q=1}^{n_q} w_{i_q} (mx_{i_q} + p) = m \underbrace{\sum_{i_q=1}^{n_q} w_{i_q} x_{i_q}}_{=0} + p \underbrace{\sum_{i_q=1}^{n_q} w_{i_q}}_{=2} = 2p$$

since the quadrature points are symmetric w.r.t. to zero on the x-axis. Once again the quadrature is able to compute the exact value of this integral: this makes sense since an  $n$ -point rule exactly integrates a  $2n - 1$  order polynomial such that a 1 point quadrature exactly integrates a first order polynomial like the one above.

**example 3** Let us now take  $f(x) = x^2$ . We have

$$I = \int_{-1}^{+1} f(x)dx = \int_{-1}^{+1} x^2 dx = [\frac{1}{3}x^3]_{-1}^{+1} = \frac{2}{3}$$

and

$$I_{gq} = \int_{-1}^{+1} f(x)dx = \sum_{i_q=1}^{n_q} w_{i_q} f(x_{i_q}) = \sum_{i_q=1}^{n_q} w_{i_q} x_{i_q}^2$$

- $n_q = 1$ :  $x_{i_q}^{(1)} = 0$ ,  $w_{i_q} = 2$ .  $I_{gq} = 0$
- $n_q = 2$ :  $x_q^{(1)} = -1/\sqrt{3}$ ,  $x_q^{(2)} = 1/\sqrt{3}$ ,  $w_q^{(1)} = w_q^{(2)} = 1$ .  $I_{gq} = \frac{2}{3}$
- It also works  $\forall n_q > 2$  !

### 3.1.3 in 2D/3D - theory

Let us now turn to a two-dimensional integral of the form

$$I = \int_{-1}^{+1} \int_{-1}^{+1} f(x, y) dx dy$$

The equivalent Gaussian quadrature writes:

$$I_{gq} \simeq \sum_{i_q=1}^{n_q} \sum_{j_q=1}^{n_q} f(x_{i_q}, y_{j_q}) w_{i_q} w_{j_q}$$

## 3.2 The mesh

### 3.3 A bit of FE terminology

We introduce here some terminology for efficient element descriptions [37]:

- For triangles/tetrahedra, the designation  $P_m \times P_n$  means that each component of the velocity is approximated by continuous piecewise complete Polynomials of degree  $m$  and pressure by continuous piecewise complete Polynomials of degree  $n$ . For example  $P_2 \times P_1$  means

$$u \sim a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2$$

with similar approximations for  $v$ , and

$$p \sim b_1 + b_2x + b_3y$$

Both velocity and pressure are continuous across element boundaries, and each triangular element contains 6 velocity nodes and three pressure nodes.

- For the same families,  $P_m \times P_{-n}$  is as above, except that pressure is approximated via piecewise *discontinuous* polynomials of degree  $n$ . For instance,  $P_2 \times P_{-1}$  is the same as  $P_2P_1$  except that pressure is now an independent linear function in each element and therefore discontinuous at element boundaries.
- For quadrilaterals/hexahedra, the designation  $Q_m \times Q_n$  means that each component of the velocity is approximated by a continuous piecewise polynomial of degree  $m$  in each direction on the quadrilateral and likewise for pressure, except that the polynomial is of degree  $n$ . For instance,  $Q_2 \times Q_1$  means

$$u \sim a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2 + a_7x^2y + a_8xy^2 + a_9x^2y^2$$

and

$$p \sim b_1 + b_2x + b_3y + b_4xy$$

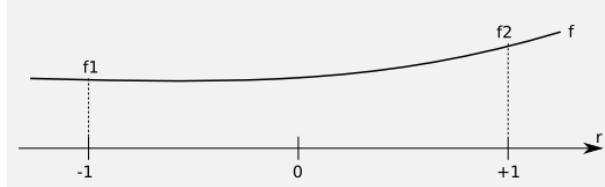
- For these same families,  $Q_m \times Q_{-n}$  is as above, except that the pressure approximation is not continuous at element boundaries.
- Again for the same families,  $Q_m \times P_{-n}$  indicates the same velocity approximation with a pressure approximation that is a discontinuous complete piecewise polynomial of degree  $n$  (not of degree  $n$  in each direction !)
- The designation  $P_m^+$  or  $Q_m^+$  means that some sort of bubble function was added to the polynomial approximation for the velocity. You may also find the term 'enriched element' in the literature.
- Finally, for  $n = 0$ , we have piecewise-constant pressure, and we omit the minus sign for simplicity.

Another point which needs to be clarified is the use of so-called 'conforming elements' (or 'non-conforming elements'). Following again [37], conforming velocity elements are those for which the basis functions for a subset of  $H^1$  for the continuous problem (the first derivatives and their squares are integrable in  $\Omega$ ). For instance, the rotated  $Q_1 \times P_0$  element of Rannacher and Turek (see section 25) is such that the velocity is discontinuous across element edges, so that the derivative does not exist there. Another typical example of non-conforming element is the Crouzeix-Raviart element [18].

## 3.4 Elements and basis functions in 1D

### 3.4.1 Linear basis functions ( $Q_1$ )

Let  $f(r)$  be a  $C^1$  function on the interval  $[-1 : 1]$  with  $f(-1) = f_1$  and  $f(1) = f_2$ .



Let us assume that the function  $f(r)$  is to be approximated on  $[-1, 1]$  by the first order polynomial

$$f(r) = a + br \quad (9)$$

Then it must fulfill

$$\begin{aligned} f(r = -1) &= a - b = f_1 \\ f(r = +1) &= a + b = f_2 \end{aligned}$$

This leads to

$$a = \frac{1}{2}(f_1 + f_2) \quad b = \frac{1}{2}(-f_1 + f_2)$$

and then replacing  $a, b$  in Eq. (9) by the above values gets

$$f(r) = \left[ \frac{1}{2}(1 - r) \right] f_1 + \left[ \frac{1}{2}(1 + r) \right] f_2$$

or

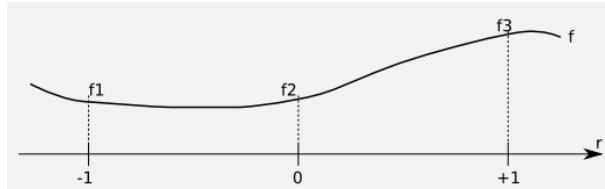
$$f(r) = \sum_{i=1}^2 N_i(r) f_i$$

with

$$\begin{aligned} N_1(r) &= \frac{1}{2}(1 - r) \\ N_2(r) &= \frac{1}{2}(1 + r) \end{aligned} \quad (10)$$

### 3.4.2 Quadratic basis functions ( $Q_2$ )

Let  $f(r)$  be a  $C^1$  function on the interval  $[-1 : 1]$  with  $f(-1) = f_1$ ,  $f(0) = f_2$  and  $f(1) = f_3$ .



Let us assume that the function  $f(r)$  is to be approximated on  $[-1, 1]$  by the second order polynomial

$$f(r) = a + br + cr^2 \quad (11)$$

Then it must fulfill

$$\begin{aligned} f(r = -1) &= a - b + c = f_1 \\ f(r = 0) &= a = f_2 \\ f(r = +1) &= a + b + c = f_3 \end{aligned}$$

This leads to

$$a = f_2 \quad b = \frac{1}{2}(-f_1 + f_3) \quad c = \frac{1}{2}(f_1 + f_3 - 2f_2)$$

and then replacing  $a, b, c$  in Eq. (11) by the above values on gets

$$f(r) = \left[ \frac{1}{2}r(r-1) \right] f_1 + (1-r^2)f_2 + \left[ \frac{1}{2}r(r+1) \right] f_3$$

or,

$$f(r) = \sum_{i=1}^3 N_i(r) f_i$$

with

$\begin{aligned} N_1(r) &= \frac{1}{2}r(r-1) \\ N_2(r) &= (1-r^2) \\ N_3(r) &= \frac{1}{2}r(r+1) \end{aligned}$	(12)
---	------

### 3.4.3 Cubic basis functions ( $Q_3$ )

The 1D basis polynomial is given by

$$f(r) = a + br + cr^2 + dr^3$$

with the nodes at position -1,-1/3, +1/3 and +1.

$$\begin{aligned} f(-1) &= a - b + c - d = f_1 \\ f(-1/3) &= a - \frac{b}{3} + \frac{c}{9} - \frac{d}{27} = f_2 \\ f(+1/3) &= a - \frac{b}{3} + \frac{c}{9} - \frac{d}{27} = f_3 \\ f(+1) &= a + b + c + d = f_4 \end{aligned}$$

Adding the first and fourth equation and the second and third, one arrives at

$$f_1 + f_4 = 2a + 2c \quad f_2 + f_3 = 2a + \frac{2c}{9}$$

and finally:

$$\begin{aligned} a &= \frac{1}{16}(-f_1 + 9f_2 + 9f_3 - f_4) \\ c &= \frac{9}{16}(f_1 - f_2 - f_3 + f_4) \end{aligned}$$

Combining the original 4 equations in a different way yields

$$2b + 2d = f_4 - f_1 \quad \frac{2b}{3} + \frac{2d}{27} = f_3 - f_2$$

so that

$$\begin{aligned} b &= \frac{1}{16}(f_1 - 27f_2 + 27f_3 - f_4) \\ d &= \frac{9}{16}(-f_1 + 3f_2 - 3f_3 + f_4) \end{aligned}$$

Finally,

$$\begin{aligned}
f(r) &= a + b + cr^2 + dr^3 \\
&= \frac{1}{16}(-1 + r + 9r^2 - 9r^3)f_1 \\
&\quad + \frac{1}{16}(9 - 27r - 9r^2 + 27r^3)f_2 \\
&\quad + \frac{1}{16}(9 + 27r - 9r^2 - 27r^3)f_3 \\
&\quad + \frac{1}{16}(-1 - r + 9r^2 + 9r^3)f_4 \\
&= \sum_{i=1}^4 N_i(r)f_i
\end{aligned}$$

where

	$N_1 = \frac{1}{16}(-1 + r + 9r^2 - 9r^3)$
	$N_2 = \frac{1}{16}(9 - 27r - 9r^2 + 27r^3)$
	$N_3 = \frac{1}{16}(9 + 27r - 9r^2 - 27r^3)$
	$N_4 = \frac{1}{16}(-1 - r + 9r^2 + 9r^3)$

Verification:

- Let us assume  $f(r) = C$ , then

$$\hat{f}(r) = \sum N_i(r)f_i = \sum_i N_i C = C \sum_i N_i = C$$

so that a constant function is exactly reproduced, as expected.

- Let us assume  $f(r) = r$ , then  $f_1 = -1$ ,  $f_2 = -1/3$ ,  $f_3 = 1/3$  and  $f_4 = +1$ . We then have

$$\begin{aligned}
\hat{f}(r) &= \sum N_i(r)f_i \\
&= -N_1(r) - \frac{1}{3}N_2(r) + \frac{1}{3}N_3(r) + N_4(r) \\
&= [-( -1 + r + 9r^2 - 9r^3) \\
&\quad - \frac{1}{3}(9 - 27r - 9r^2 + 27r^3) \\
&\quad + \frac{1}{3}(9 + 27r - 9r^2 - 27r^3) \\
&\quad + (-1 - r + 9r^2 + 9r^3)]/16 \\
&= [-r + 9r + 9r - r]/16 + \dots 0... \\
&= r
\end{aligned} \tag{13}$$

The basis functions derivative are given by

$$\begin{aligned}
\frac{\partial N_1}{\partial r} &= \frac{1}{16}(1 + 18r - 27r^2) \\
\frac{\partial N_2}{\partial r} &= \frac{1}{16}(-27 - 18r + 81r^2) \\
\frac{\partial N_3}{\partial r} &= \frac{1}{16}(+27 - 18r - 81r^2) \\
\frac{\partial N_4}{\partial r} &= \frac{1}{16}(-1 + 18r + 27r^2)
\end{aligned}$$

Verification:

- Let us assume  $f(r) = C$ , then

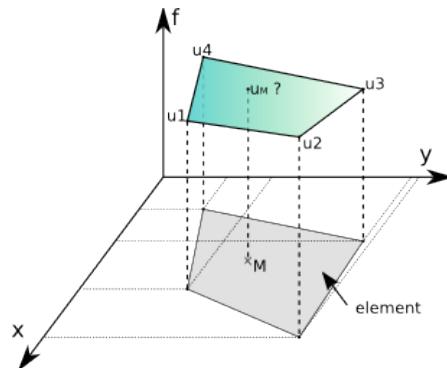
$$\begin{aligned}
\frac{\partial \hat{f}}{\partial r} &= \sum_i \frac{\partial N_i}{\partial r} f_i \\
&= C \sum_i \frac{\partial N_i}{\partial r} \\
&= \frac{C}{16} [(1 + 18r - 27r^2) \\
&\quad + (-27 - 18r + 81r^2) \\
&\quad + (+27 - 18r - 81r^2) \\
&\quad + (-1 + 18r + 27r^2)] \\
&= 0
\end{aligned}$$

- Let us assume  $f(r) = r$ , then  $f_1 = -1$ ,  $f_2 = -1/3$ ,  $f_3 = 1/3$  and  $f_4 = +1$ . We then have

$$\begin{aligned}
\frac{\partial \hat{f}}{\partial r} &= \sum_i \frac{\partial N_i}{\partial r} f_i \\
&= \frac{1}{16} [-(1 + 18r - 27r^2) \\
&\quad - \frac{1}{3}(-27 - 18r + 81r^2) \\
&\quad + \frac{1}{3}(+27 - 18r - 81r^2) \\
&\quad + (-1 + 18r + 27r^2)] \\
&= \frac{1}{16} [-2 + 18 + 54r^2 - 54r^2] \\
&= 1
\end{aligned}$$

### 3.5 Elements and basis functions in 2D

Let us for a moment consider a single quadrilateral element in the  $xy$ -plane, as shown on the following figure:



Let us assume that we know the values of a given field  $u$  at the vertices. For a given point  $M$  inside the element in the plane, what is the value of the field  $u$  at this point? It makes sense to postulate that  $u_M = u(x_M, y_M)$  will be given by

$$u_M = \phi(u_1, u_2, u_3, u_4, x_M, y_M)$$

where  $\phi$  is a function to be determined. Although  $\phi$  is not unique, we can decide to express the value  $u_M$  as a weighed sum of the values at the vertices  $u_i$ . One option could be to assign all four vertices the same weight, say  $1/4$  so that  $u_M = (u_1 + u_2 + u_3 + u_4)/4$ , i.e.  $u_M$  is simply given by the arithmetic mean of the vertices values. This approach suffers from a major drawback as it does not use the location of point  $M$  inside the element. For instance, when  $(x_M, y_M) \rightarrow (x_2, y_2)$  we expect  $u_M \rightarrow u_2$ .

In light of this, we could now assume that the weights would depend on the position of  $M$  in a continuous fashion:

$$u(x_M, y_M) = \sum_{i=1}^4 N_i(x_M, y_M) u_i$$

where the  $N_i$  are continuous ("well behaved") functions which have the property:

$$N_i(x_j, y_j) = \delta_{ij}$$

or, in other words:

$$N_3(x_1, y_1) = 0 \quad (14)$$

$$N_3(x_2, y_2) = 0 \quad (15)$$

$$N_3(x_3, y_3) = 1 \quad (16)$$

$$N_3(x_4, y_4) = 0 \quad (17)$$

The functions  $N_i$  are commonly called basis functions.

Omitting the  $M$  subscripts for any point inside the element, the velocity components  $u$  and  $v$  are given by:

$$\hat{u}(x, y) = \sum_{i=1}^4 N_i(x, y) u_i \quad (18)$$

$$\hat{v}(x, y) = \sum_{i=1}^4 N_i(x, y) v_i \quad (19)$$

Rather interestingly, one can now easily compute velocity gradients (and therefore the strain rate tensor) since we have assumed the basis functions to be "well behaved" (in this case differentiable):

$$\dot{\epsilon}_{xx}(x, y) = \frac{\partial u}{\partial x} = \sum_{i=1}^4 \frac{\partial N_i}{\partial x} u_i \quad (20)$$

$$\dot{\epsilon}_{yy}(x, y) = \frac{\partial v}{\partial y} = \sum_{i=1}^4 \frac{\partial N_i}{\partial y} v_i \quad (21)$$

$$\dot{\epsilon}_{xy}(x, y) = \frac{1}{2} \frac{\partial u}{\partial y} + \frac{1}{2} \frac{\partial v}{\partial x} = \frac{1}{2} \sum_{i=1}^4 \frac{\partial N_i}{\partial y} u_i + \frac{1}{2} \sum_{i=1}^4 \frac{\partial N_i}{\partial x} v_i \quad (22)$$

How we actually obtain the exact form of the basis functions is explained in the coming section.

### 3.5.1 Bilinear basis functions in 2D ( $Q_1$ )

In this section, we place ourselves in the most favorable case, i.e. the element is a square defined by  $-1 < r < 1$ ,  $-1 < s < 1$  in the Cartesian coordinates system  $(r, s)$ :

```

3=====2
|       | (r_0,s_0)=(-1,-1)
|       | (r_1,s_1)=(+1,-1)
|       | (r_2,s_2)=(+1,+1)
|       | (r_3,s_3)=(-1,+1)
|
0=====1

```

This element is commonly called the reference element. How we go from the  $(x, y)$  coordinate system to the  $(r, s)$  once and vice versa will be dealt later on. For now, the basis functions in the above reference element and in the reduced coordinates system  $(r, s)$  are given by:

$$\begin{aligned}
N_1(r, s) &= 0.25(1 - r)(1 - s) \\
N_2(r, s) &= 0.25(1 + r)(1 - s) \\
N_3(r, s) &= 0.25(1 + r)(1 + s) \\
N_4(r, s) &= 0.25(1 - r)(1 + s)
\end{aligned}$$

The partial derivatives of these functions with respect to  $r$  and  $s$  automatically follow:

$$\begin{array}{ll}
\frac{\partial N_1}{\partial r}(r, s) = -0.25(1 - s) & \frac{\partial N_1}{\partial s}(r, s) = -0.25(1 - r) \\
\frac{\partial N_2}{\partial r}(r, s) = +0.25(1 - s) & \frac{\partial N_2}{\partial s}(r, s) = -0.25(1 + r) \\
\frac{\partial N_3}{\partial r}(r, s) = +0.25(1 + s) & \frac{\partial N_3}{\partial s}(r, s) = +0.25(1 + r) \\
\frac{\partial N_4}{\partial r}(r, s) = -0.25(1 + s) & \frac{\partial N_4}{\partial s}(r, s) = +0.25(1 - r)
\end{array}$$

Let us go back to Eq.(19). And let us assume that the function  $v(r, s) = C$  so that  $v_i = C$  for  $i = 1, 2, 3, 4$ . It then follows that

$$\hat{v}(r, s) = \sum_{i=1}^4 N_i(r, s) v_i = C \sum_{i=1}^4 N_i(r, s) = C[N_1(r, s) + N_2(r, s) + N_3(r, s) + N_4(r, s)] = C$$

This is a very important property: if the  $v$  function used to assign values at the vertices is constant, then the value of  $\hat{v}$  *anywhere* in the element is exactly  $C$ . If we now turn to the derivatives of  $v$  with respect to  $r$  and  $s$ :

$$\frac{\partial \hat{v}}{\partial r}(r, s) = \sum_{i=1}^4 \frac{\partial N_i}{\partial r}(r, s) v_i = C \sum_{i=1}^4 \frac{\partial N_i}{\partial r}(r, s) = C[-0.25(1 - s) + 0.25(1 - s) + 0.25(1 + s) - 0.25(1 + s)] = 0$$

$$\frac{\partial \hat{v}}{\partial s}(r, s) = \sum_{i=1}^4 \frac{\partial N_i}{\partial s}(r, s) v_i = C \sum_{i=1}^4 \frac{\partial N_i}{\partial s}(r, s) = C[-0.25(1 - r) - 0.25(1 + r) + 0.25(1 + r) + 0.25(1 - r)] = 0$$

We reassuringly find that the derivative of a constant field anywhere in the element is exactly zero.

If we now choose  $v(r, s) = ar + bs$  with  $a$  and  $b$  two constant scalars, we find:

$$\hat{v}(r, s) = \sum_{i=1}^4 N_i(r, s) v_i \quad (23)$$

$$= \sum_{i=1}^4 N_i(r, s)(ar_i + bs_i) \quad (24)$$

$$= \underbrace{a \sum_{i=1}^4 N_i(r, s)r_i}_{r} + \underbrace{b \sum_{i=1}^4 N_i(r, s)s_i}_{s} \quad (25)$$

$$\begin{aligned} &= a [0.25(1-r)(1-s)(-1) + 0.25(1+r)(1-s)(+1) + 0.25(1+r)(1+s)(+1) + 0.25(1-r)(1+s)(-1)] \\ &+ b [0.25(1-r)(1-s)(-1) + 0.25(1+r)(1-s)(-1) + 0.25(1+r)(1+s)(+1) + 0.25(1-r)(1+s)(+1)] \\ &= a [-0.25(1-r)(1-s) + 0.25(1+r)(1-s) + 0.25(1+r)(1+s) - 0.25(1-r)(1+s)] \\ &+ b [-0.25(1-r)(1-s) - 0.25(1+r)(1-s) + 0.25(1+r)(1+s) + 0.25(1-r)(1+s)] \\ &= ar + bs \end{aligned} \quad (26)$$

**verify above eq.** This set of bilinear shape functions is therefore capable of exactly representing a bilinear field. The derivatives are:

$$\frac{\partial \hat{v}}{\partial r}(r, s) = \sum_{i=1}^4 \frac{\partial N_i}{\partial r}(r, s) v_i \quad (27)$$

$$= a \sum_{i=1}^4 \frac{\partial N_i}{\partial r}(r, s)r_i + b \sum_{i=1}^4 \frac{\partial N_i}{\partial r}(r, s)s_i \quad (28)$$

$$\begin{aligned} &= a [-0.25(1-s)(-1) + 0.25(1-s)(+1) + 0.25(1+s)(+1) - 0.25(1+s)(-1)] \\ &+ b [-0.25(1-s)(-1) + 0.25(1-s)(-1) + 0.25(1+s)(+1) - 0.25(1+s)(+1)] \\ &= \frac{a}{4} [(1-s) + (1-s) + (1+s) + (1+s)] \\ &+ \frac{b}{4} [(1-s) - (1-s) + (1+s) - (1+s)] \\ &= a \end{aligned} \quad (29)$$

Here again, we find that the derivative of the bilinear field inside the element is exact:  $\frac{\partial \hat{v}}{\partial r} = \frac{\partial v}{\partial r}$ .

However, following the same methodology as above, one can easily prove that this is no more true for polynomials of degree strivtly higher than 1. This fact has serious consequences: if the solution to the problem at hand is for instance a parabola, the  $Q_1$  shape functions cannot represent the solution properly, but only by approximating the parabola in each element by a line. As we will see later,  $Q_2$  basis functions can remedy this problem by containing themselves quadratic terms.

### 3.5.2 Biquadratic basis functions in 2D ( $Q_2$ )

Inside an element the local numbering of the nodes is as follows:

```
3=====6=====2
|       |       |   (r_0,s_0)=(-1,-1)   (r_4,s_4)=( 0,-1)
|       |       |   (r_1,s_1)=(+1,-1)   (r_5,s_5)=(+1, 0)
7=====8=====5   (r_2,s_2)=(+1,+1)   (r_6,s_6)=( 0,+1)
|       |       |   (r_3,s_3)=(-1,+1)   (r_7,s_7)=(-1, 0)
|       |       |                           (r_8,s_8)=( 0, 0)
0=====4=====1
```

The velocity shape functions are then given by:

$$\begin{aligned}
N_0(r, s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s-1) \\
N_1(r, s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s-1) \\
N_2(r, s) &= \frac{1}{2}r(r+1)\frac{1}{2}s(s+1) \\
N_3(r, s) &= \frac{1}{2}r(r-1)\frac{1}{2}s(s+1) \\
N_4(r, s) &= (1-r^2)\frac{1}{2}s(s-1) \\
N_5(r, s) &= \frac{1}{2}r(r+1)(1-s^2) \\
N_6(r, s) &= (1-r^2)\frac{1}{2}s(s+1) \\
N_7(r, s) &= \frac{1}{2}r(r-1)(1-s^2) \\
N_8(r, s) &= (1-r^2)(1-s^2)
\end{aligned}$$

and their derivatives by:

$$\begin{aligned}
\frac{\partial N_0}{\partial r} &= \frac{1}{2}(2r-1)\frac{1}{2}s(s-1) & \frac{\partial N_0}{\partial s} &= \frac{1}{2}r(r-1)\frac{1}{2}(2s-1) \\
\frac{\partial N_1}{\partial r} &= \frac{1}{2}(2r+1)\frac{1}{2}s(s-1) & \frac{\partial N_1}{\partial s} &= \frac{1}{2}r(r+1)\frac{1}{2}(2s-1) \\
\frac{\partial N_2}{\partial r} &= \frac{1}{2}(2r+1)\frac{1}{2}s(s+1) & \frac{\partial N_2}{\partial s} &= \frac{1}{2}r(r+1)\frac{1}{2}(2s+1) \\
\frac{\partial N_3}{\partial r} &= \frac{1}{2}(2r-1)\frac{1}{2}s(s+1) & \frac{\partial N_3}{\partial s} &= \frac{1}{2}r(r-1)\frac{1}{2}(2s+1) \\
\frac{\partial N_4}{\partial r} &= (-2r)\frac{1}{2}s(s-1) & \frac{\partial N_4}{\partial s} &= (1-r^2)\frac{1}{2}(2s-1) \\
\frac{\partial N_5}{\partial r} &= \frac{1}{2}(2r+1)(1-s^2) & \frac{\partial N_5}{\partial s} &= \frac{1}{2}r(r+1)(-2s) \\
\frac{\partial N_6}{\partial r} &= (-2r)\frac{1}{2}s(s+1) & \frac{\partial N_6}{\partial s} &= (1-r^2)\frac{1}{2}(2s+1) \\
\frac{\partial N_7}{\partial r} &= \frac{1}{2}(2r-1)(1-s^2) & \frac{\partial N_7}{\partial s} &= \frac{1}{2}r(r-1)(-2s) \\
\frac{\partial N_8}{\partial r} &= (-2r)(1-s^2) & \frac{\partial N_8}{\partial s} &= (1-r^2)(-2s)
\end{aligned}$$

### 3.5.3 Bicubic basis functions in 2D ( $Q_3$ )

Inside an element the local numbering of the nodes is as follows:

```

12====13====14====15      (r,s)_{{00}}=(-1,-1)      (r,s)_{{08}}=(-1,+1/3)
||    ||    ||    ||      (r,s)_{{01}}=(-1/3,-1)    (r,s)_{{09}}=(-1/3,+1/3)
08====09====10====11      (r,s)_{{02}}=(+1/3,-1)    (r,s)_{{10}}=(+1/3,+1/3)
||    ||    ||    ||      (r,s)_{{03}}=(+1,-1)      (r,s)_{{11}}=(+1,+1/3)
04====05====06====07      (r,s)_{{04}}=(-1,-1/3)    (r,s)_{{12}}=(-1,+1)
||    ||    ||    ||      (r,s)_{{05}}=(-1/3,-1/3)  (r,s)_{{13}}=(-1/3,+1)
00====01====02====03      (r,s)_{{06}}=(+1/3,-1/3)  (r,s)_{{14}}=(+1/3,+1)
                           (r,s)_{{07}}=(+1,-1/3)    (r,s)_{{15}}=(+1,+1)

```

The velocity shape functions are given by:

$$\begin{aligned}
 N_1(r) &= (-1 + r + 9r^2 - 9r^3)/16 & N_1(t) &= (-1 + t + 9t^2 - 9t^3)/16 \\
 N_2(r) &= (+9 - 27r - 9r^2 + 27r^3)/16 & N_2(t) &= (+9 - 27t - 9t^2 + 27t^3)/16 \\
 N_3(r) &= (+9 + 27r - 9r^2 - 27r^3)/16 & N_3(t) &= (+9 + 27t - 9t^2 - 27t^3)/16 \\
 N_4(r) &= (-1 - r + 9r^2 + 9r^3)/16 & N_4(t) &= (-1 - t + 9t^2 + 9t^3)/16
 \end{aligned}$$

$N_{01}(r, s) = N_1(r)N_1(s) = (-1 + r + 9r^2 - 9r^3)/16 * (-1 + t + 9s^2 - 9s^3)/16$	
$N_{02}(r, s) = N_2(r)N_1(s) = (+9 - 27r - 9r^2 + 27r^3)/16 * (-1 + t + 9s^2 - 9s^3)/16$	
$N_{03}(r, s) = N_3(r)N_1(s) = (+9 + 27r - 9r^2 - 27r^3)/16 * (-1 + t + 9s^2 - 9s^3)/16$	
$N_{04}(r, s) = N_4(r)N_1(s) = (-1 - r + 9r^2 + 9r^3)/16 * (-1 + t + 9s^2 - 9s^3)/16$	
$N_{05}(r, s) = N_1(r)N_2(s) = (-1 + r + 9r^2 - 9r^3)/16 * (9 - 27s - 9s^2 + 27s^3)/16$	
$N_{06}(r, s) = N_2(r)N_2(s) = (+9 - 27r - 9r^2 + 27r^3)/16 * (9 - 27s - 9s^2 + 27s^3)/16$	
$N_{07}(r, s) = N_3(r)N_2(s) = (+9 + 27r - 9r^2 - 27r^3)/16 * (9 - 27s - 9s^2 + 27s^3)/16$	
$N_{08}(r, s) = N_4(r)N_2(s) = (-1 - r + 9r^2 + 9r^3)/16 * (9 - 27s - 9s^2 + 27s^3)/16$	
$N_{09}(r, s) = N_1(r)N_3(s) =$	(30)
$N_{10}(r, s) = N_2(r)N_3(s) =$	(31)
$N_{11}(r, s) = N_3(r)N_3(s) =$	(32)
$N_{12}(r, s) = N_4(r)N_3(s) =$	(33)
$N_{13}(r, s) = N_1(r)N_4(s) =$	(34)
$N_{14}(r, s) = N_2(r)N_4(s) =$	(35)
$N_{15}(r, s) = N_3(r)N_4(s) =$	(36)
$N_{16}(r, s) = N_4(r)N_4(s) =$	(37)

## 4 Solving the Stokes equations with the FEM

In the case of an incompressible flow, we have seen that the continuity (mass conservation) equation takes the simple form  $\nabla \cdot \mathbf{v} = 0$ . In other word flow takes place under the constraint that the divergence of its velocity field is exactly zero everywhere (solenoidal constraint), i.e. it is divergence free.

We see that the pressure in the momentum equation is then a degree of freedom which is needed to satisfy the incompressibility constraint (and it is not related to any constitutive equation) [25]. In other words the pressure is acting as a Lagrange multiplier of the incompressibility constraint.

Various approaches have been proposed in the literature to deal with the incompressibility constraint but we will only focus on the penalty method (section 4.2) and the so-called mixed finite element method 4.3.

### 4.1 strong and weak forms

The strong form consists of the governing equation and the boundary conditions, i.e. the mass, momentum and energy conservation equations supplemented with Dirichlet and/or Neumann boundary conditions on (parts of) the boundary.

To develop the finite element formulation, the partial differential equations must be restated in an integral form called the weak form. In essence the PDEs are first multiplied by an arbitrary function and integrated over the domain.

### 4.2 The penalty approach

In order to impose the incompressibility constraint, two widely used procedures are available, namely the Lagrange multiplier method and the penalty method [5, 41]. The latter is implemented in ELEFANT, which allows for the elimination of the pressure variable from the momentum equation (resulting in a reduction of the matrix size).

Mathematical details on the origin and validity of the penalty approach applied to the Stokes problem can for instance be found in [19], [66] or [39].

The penalty formulation of the mass conservation equation is based on a relaxation of the incompressibility constraint and writes

$$\nabla \cdot \mathbf{v} + \frac{p}{\lambda} = 0 \quad (38)$$

where  $\lambda$  is the penalty parameter, that can be interpreted (and has the same dimension) as a bulk viscosity. It is equivalent to say that the material is weakly compressible. It can be shown that if one chooses  $\lambda$  to be a sufficiently large number, the continuity equation  $\nabla \cdot \mathbf{v} = 0$  will be approximately satisfied in the finite element solution. The value of  $\lambda$  is often recommended to be 6 to 7 orders of magnitude larger than the shear viscosity [27, 42].

Equation (38) can be used to eliminate the pressure in Eq. (??) so that the mass and momentum conservation equations fuse to become :

$$\nabla \cdot (2\eta \dot{\epsilon}(\mathbf{v})) + \lambda \nabla (\nabla \cdot \mathbf{v}) = \rho \mathbf{g} = 0 \quad (39)$$

[55] have established the equivalence for incompressible problems between the reduced integration of the penalty term and a mixed Finite Element approach if the pressure nodes coincide with the integration points of the reduced rule.

In the end, the elimination of the pressure unknown in the Stokes equations replaces the original saddle-point Stokes problem [6] by an elliptical problem, which leads to a symmetric positive definite (SPD) FEM matrix. This is the major benefit of the penalized approach over the full indefinite solver with the velocity-pressure variables. Indeed, the SPD character of the matrix lends itself to efficient solving strategies and is less memory-demanding since it is sufficient to store only the upper half of the matrix including the diagonal [36].

The stress tensor  $\boldsymbol{\sigma}$  is symmetric (*i.e.*  $\sigma_{ij} = \sigma_{ji}$ ). For simplicity I will now focus on a Stokes flow in two dimensions.

list codes which  
use this ap-  
proach

Since the penalty formulation is only valid for incompressible flows, then  $\dot{\epsilon} = \dot{\epsilon}^d$  so that the  $d$  superscript is omitted in what follows. The stress tensor can also be cast in vector format:

$$\begin{aligned}\begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} &= \begin{pmatrix} -p \\ -p \\ 0 \end{pmatrix} + 2\eta \begin{pmatrix} \dot{\epsilon}_{xx} \\ \dot{\epsilon}_{yy} \\ \dot{\epsilon}_{xy} \end{pmatrix} \\ &= \lambda \begin{pmatrix} \dot{\epsilon}_{xx} + \dot{\epsilon}_{yy} \\ \dot{\epsilon}_{xx} + \dot{\epsilon}_{yy} \\ 0 \end{pmatrix} + 2\eta \begin{pmatrix} \dot{\epsilon}_{xx} \\ \dot{\epsilon}_{yy} \\ \dot{\epsilon}_{xy} \end{pmatrix} \\ &= \left[ \underbrace{\lambda \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{\mathbf{K}} + \underbrace{\eta \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_C \right] \cdot \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{pmatrix}\end{aligned}$$

Remember that

$$\frac{\partial u}{\partial x} = \sum_{i=1}^4 \frac{\partial N_i}{\partial x} u_i \quad \frac{\partial v}{\partial y} = \sum_{i=1}^4 \frac{\partial N_i}{\partial y} v_i$$

$$\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = \sum_{i=1}^4 \frac{\partial N_i}{\partial y} u_i + \sum_{i=1}^4 \frac{\partial N_i}{\partial x} v_i$$

so that

$$\begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial N_1}{\partial x} & 0 & \frac{\partial N_2}{\partial x} & 0 & \frac{\partial N_3}{\partial x} & 0 & \frac{\partial N_4}{\partial x} & 0 \\ 0 & \frac{\partial N_1}{\partial y} & 0 & \frac{\partial N_2}{\partial y} & 0 & \frac{\partial N_3}{\partial y} & 0 & \frac{\partial N_4}{\partial y} \\ \frac{\partial N_1}{\partial y} & \frac{\partial N_1}{\partial x} & \frac{\partial N_2}{\partial y} & \frac{\partial N_2}{\partial x} & \frac{\partial N_3}{\partial y} & \frac{\partial N_3}{\partial x} & \frac{\partial N_4}{\partial y} & \frac{\partial N_4}{\partial x} \end{pmatrix}}_B \cdot \underbrace{\begin{pmatrix} u1 \\ v1 \\ u2 \\ v2 \\ u3 \\ v3 \\ u4 \\ v4 \end{pmatrix}}_V$$

Finally,

$$\vec{\sigma} = \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} = (\lambda \mathbf{K} + \eta \mathbf{C}) \cdot \mathbf{B} \cdot \mathbf{V}$$

We will now establish the weak form of the momentum conservation equation. We start again from

$$\nabla \cdot \boldsymbol{\sigma} + \mathbf{b} = \mathbf{0}$$

For the  $N_i$ 's 'regular enough', we can write:

$$\int_{\Omega_e} N_i \nabla \cdot \boldsymbol{\sigma} d\Omega + \int_{\Omega_e} N_i \mathbf{b} d\Omega = 0$$

We can integrate by parts and drop the surface term<sup>1</sup>:

$$\int_{\Omega_e} \nabla N_i \cdot \boldsymbol{\sigma} d\Omega = \int_{\Omega_e} N_i \mathbf{b} d\Omega$$

or,

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial N_i}{\partial x} & 0 & \frac{\partial N_i}{\partial y} \\ 0 & \frac{\partial N_i}{\partial y} & \frac{\partial N_i}{\partial x} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} d\Omega = \int_{\Omega_e} N_i \mathbf{b} d\Omega$$

---

<sup>1</sup>We will come back to this at a later stage

Let  $i = 1, 2, 3, 4$  and stack the resulting four equations on top of one another.

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial N_1}{\partial x} & 0 & \frac{\partial N_1}{\partial y} \\ 0 & \frac{\partial N_1}{\partial y} & \frac{\partial N_1}{\partial x} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} d\Omega = \int_{\Omega_e} N_1 \begin{pmatrix} b_x \\ b_y \end{pmatrix} d\Omega \quad (40)$$

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial N_2}{\partial x} & 0 & \frac{\partial N_2}{\partial y} \\ 0 & \frac{\partial N_2}{\partial y} & \frac{\partial N_2}{\partial x} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} d\Omega = \int_{\Omega_e} N_2 \begin{pmatrix} b_x \\ b_y \end{pmatrix} d\Omega \quad (41)$$

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial N_3}{\partial x} & 0 & \frac{\partial N_3}{\partial y} \\ 0 & \frac{\partial N_3}{\partial y} & \frac{\partial N_3}{\partial x} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} d\Omega = \int_{\Omega_e} N_3 \begin{pmatrix} b_x \\ b_y \end{pmatrix} d\Omega \quad (42)$$

$$\int_{\Omega_e} \begin{pmatrix} \frac{\partial N_4}{\partial x} & 0 & \frac{\partial N_4}{\partial y} \\ 0 & \frac{\partial N_4}{\partial y} & \frac{\partial N_4}{\partial x} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} d\Omega = \int_{\Omega_e} N_4 \begin{pmatrix} b_x \\ b_y \end{pmatrix} d\Omega \quad (43)$$

We easily recognize  $\mathbf{B}^T$  inside the integrals! Let us define

$$\mathbf{N}_b^T = (N_1 b_x, N_1 b_y, \dots, N_4 b_x, N_4 b_y)$$

then we can write

$$\int_{\Omega_e} \mathbf{B}^T \cdot \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix} d\Omega = \int_{\Omega_e} \mathbf{N}_b d\Omega$$

and finally:

$$\int_{\Omega_e} \mathbf{B}^T \cdot [\lambda \mathbf{K} + \eta \mathbf{C}] \cdot \mathbf{B} \cdot \mathbf{V} d\Omega = \int_{\Omega_e} \mathbf{N}_b d\Omega$$

Since  $\mathbf{V}$  contains the velocities at the corners, it does not depend on the  $x$  or  $y$  coordinates so it can be taking outside of the integral:

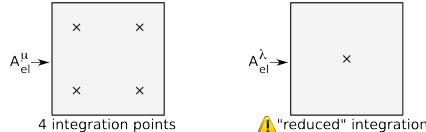
$$\underbrace{\left( \int_{\Omega_e} \mathbf{B}^T \cdot [\lambda \mathbf{K} + \eta \mathbf{C}] \cdot \mathbf{B} d\Omega \right)}_{A_{el}(8 \times 8)} \cdot \underbrace{\mathbf{V}}_{(8 \times 1)} = \underbrace{\int_{\Omega_e} \mathbf{N}_b d\Omega}_{B_{el}(8 \times 1)}$$

or,

$$\left[ \underbrace{\left( \int_{\Omega_e} \lambda \mathbf{B}^T \cdot \mathbf{K} \cdot \mathbf{B} d\Omega \right)}_{A_{el}^\lambda(8 \times 8)} + \underbrace{\left( \int_{\Omega_e} \eta \mathbf{B}^T \cdot \mathbf{C} \cdot \mathbf{B} d\Omega \right)}_{A_{el}^\eta(8 \times 8)} \right] \cdot \underbrace{\mathbf{V}}_{(8 \times 1)} = \underbrace{\int_{\Omega_e} \mathbf{N}_b d\Omega}_{B_{el}(8 \times 1)}$$

### INTEGRATION - MAPPING

1. partition domain  $\Omega$  into elements  $\Omega_e$ ,  $e = 1, \dots, n_{el}$ .
2. loop over elements and for each element compute  $\mathbf{A}_{el}$ ,  $\mathbf{B}_{el}$



3. a node belongs to several elements  
→ need to assemble  $\mathbf{A}_{el}$  and  $\mathbf{B}_{el}$  in  $\mathbf{A}$ ,  $\mathbf{B}$
4. apply boundary conditions
5. solve system:  $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{B}$
6. visualise/analyse  $\mathbf{x}$

### 4.3 The mixed FEM

## 5 Solving the elastic equations with the FEM

## 6 Additional techniques

## 6.1 The method of manufactured solutions

The method of manufactured solutions is a relatively simple way of carrying out code verification. In essence, one postulates a solution for the PDE at hand (as well as the proper boundary conditions), inserts it in the PDE and computes the corresponding source term. The same source term and boundary conditions will then be used in a numerical simulation so that the computed solution can be compared with the (postulated) true analytical solution.

Examples of this approach are to be found in [25, 13, ?].

```
> python_codes/fieldstone
> python_codes/fieldstone_saddlepoint
> python_codes/fieldstone_saddlepoint_q2q1
> python_codes/fieldstone_saddlepoint_q3q2
> python_codes/fieldstone_burstedde
```

## 6.2 Assigning values to quadrature points

As we have seen in Section ??, the building of the elemental matrix and rhs requires (at least) to assign a density and viscosity value to each quadrature point inside the element. Depending on the type of modelling, this task can prove more complex than one might expect and have large consequences on the solution accuracy.

Here are several options:

- The simplest way (which is often used for benchmarks) consists in computing the 'real' coordinates  $(x_q, y_q, z_q)$  of a given quadrature point based on its reduced coordinates  $(r_q, s_q, t_q)$ , and passing these coordinates to a function which returns density and/or viscosity at this location. For instance, for the Stokes sphere:

```
def rho(x,y):
    if (x-.5)**2+(y-0.5)**2<0.123**2:
        val=2.
    else:
        val=1.
    return val

def mu(x,y):
    if (x-.5)**2+(y-0.5)**2<0.123**2:
        val=1.e2
    else:
        val=1.
    return val
```

This is very simple, but it has been shown to potentially be problematic. In essence, it can introduce very large contrasts inside a single element and perturb the quadrature. Please read section 3.3 of [40] and/or have a look at the section titled "Averaging material properties" in the ASPECT manual.

- another similar approach consists in assigning a density and viscosity value to the nodes of the FE mesh first, and then using these nodal values to assign values to the quadrature points. Very often ,and quite logically, the shape functions are used to this effect. Indeed we have seen before that for any point  $(r, s, t)$  inside an element we have

$$f_h(r, s, t) = \sum_i^m f_i N_i(r, s, t)$$

where the  $f_i$  are the nodal values and the  $N_i$  the corresponding basis functions.

In the case of linear elements ( $Q_1$  basis functions), this is straightforward. In fact, the basis functions  $N_i$  can be seen as moving weights: the closer the point is to a node, the higher the weight (basis function value).

However, this is quite another story for quadratic elements ( $Q_2$  basis functions). In order to illustrate the problem, let us consider a 1D problem. The basis functions are

$$N_1(r) = \frac{1}{2}r(r-1) \quad N_2(r) = 1 - r^2 \quad N_3(r) = \frac{1}{2}r(r+1)$$

Let us further assign:  $\rho_1 = \rho_2 = 0$  and  $\rho_3 = 1$ . Then

$$\rho_h(r) = \sum_i^m \rho_i N_i(r) = N_3(r)$$

There lies the core of the problem: the  $N_3(r)$  basis function is negative for  $r \in [-1, 0]$ . This means that the quadrature point in this interval will be assigned a negative density, which is nonsensical and numerically problematic!

**use 2X Q1. write about it !**

The above methods work fine as long as the domain contains a single material. As soon as there are multiple fluids in the domain a special technique is needed to track either the fluids themselves or their interfaces. Let us start with markers. We are then confronted to the infernal trio (a *menage à trois?*) which is present for each element, composed of its nodes, its markers and its quadrature points.

Each marker carries the material information (density and viscosity). This information must ultimately be projected onto the quadrature points. Two main options are possible: an algorithm is designed and projects the marker-based fields onto the quadrature points directly or the marker fields are first projected onto the FE nodes and then onto the quadrature points using the techniques above.

---

At a given time, every element  $e$  contains  $n^e$  markers. During the FE matrix building process, viscosity and density values are needed at the quadrature points. One therefore needs to project the values carried by the markers at these locations. Several approaches are currently in use in the community and the topic has been investigated by [23] and [28] for instance.

ELEFANT adopts a simple approach: viscosity and density are considered to be elemental values, i.e. all the markers within a given element contribute to assign a unique constant density and viscosity value to the element by means of an averaging scheme.

While it is common in the literature to treat the so-called arithmetic, geometric and harmonic means as separate averagings, I hereby wish to introduce the notion of generalised mean, which is a family of functions for aggregating sets of numbers that include as special cases the arithmetic, geometric and harmonic means.

If  $p$  is a non-zero real number, we can define the generalised mean (or power mean) with exponent  $p$  of the positive real numbers  $a_1, \dots, a_n$  as:

$$M_p(a_1, \dots, a_n) = \left( \frac{1}{n} \sum_{i=1}^n a_i^p \right)^{1/p} \quad (44)$$

and it is trivial to verify that we then have the special cases:

$$M_{-\infty} = \lim_{p \rightarrow -\infty} M_p = \min(a_1, \dots, a_n) \quad (\text{minimum}) \quad (45)$$

$$M_{-1} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n}} \quad (\text{harm. avrg.}) \quad (46)$$

$$M_0 = \lim_{p \rightarrow 0} M_p = \left( \prod_{i=1}^n a_i \right)^{1/n} \quad (\text{geom. avrg.}) \quad (47)$$

$$M_{+1} = \frac{1}{n} \sum_{i=1}^n a_i \quad (\text{arithm. avrg.}) \quad (48)$$

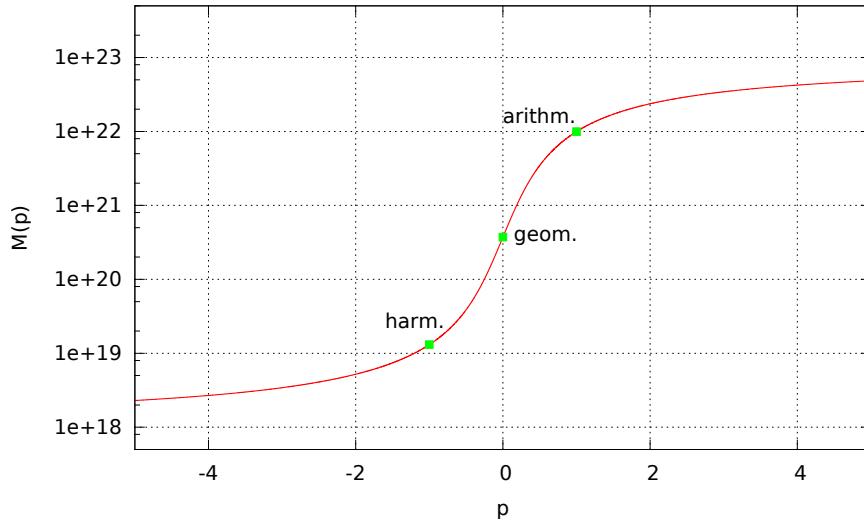
$$M_{+2} = \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2} \quad (\text{root mean square}) \quad (49)$$

$$M_{+\infty} = \lim_{p \rightarrow +\infty} M_p = \max(a_1, \dots, a_n) \quad (\text{maximum}) \quad (50)$$

Note that the proofs of the limit convergence are given in [12].

An interesting property of the generalised mean is as follows: for two real values  $p$  and  $q$ , if  $p < q$  then  $M_p \leq M_q$ . This property has for instance been illustrated in Fig. 20 of [72].

One can then for instance look at the generalised mean of a randomly generated set of 1000 viscosity values within  $10^{18} Pa.s$  and  $10^{23} Pa.s$  for  $-5 \leq p \leq 5$ . Results are shown in the figure hereunder and the arithmetic, geometric and harmonic values are indicated too. The function  $M_p$  assumes an arctangent-like shape: very low values of  $p$  will ultimately yield the minimum viscosity in the array while very high values will yield its maximum. In between, the transition is smooth and occurs essentially for  $|p| \leq 5$ .



```
> python_codes/fieldstone_markers_avrg
```

- 6.3 The value of the timestep**
- 6.4 Picard and Newton**
- 6.5 The SUPG formulation for the energy equation**
- 6.6 Tracking materials and/or interfaces**
- 6.7 Dealing with a free surface**

## 6.8 Matrix (Sparse) storage

The FE matrix is the result of the assembly process of all elemental matrices. Its size can become quite large when the resolution is being increased (from thousands of lines/columns to tens of millions).

One important property of the matrix is its sparsity. Typically less than 1% of the matrix terms is not zero and this means that the matrix storage can and should be optimised. Clever storage formats were designed early on since the amount of RAM memory in computers was the limiting factor 3 or 4 decades ago. [68]

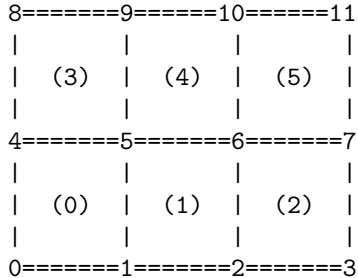
There are several standard formats:

- compressed sparse row (CSR) format
- compressed sparse column format (CSC)
- the Coordinate Format (COO)
- Skyline Storage Format
- ...

I focus on the CSR format in what follows.

### 6.8.1 2D domain - One degree of freedom per node

Let us consider again the  $3 \times 2$  element grid which counts 12 nodes.



In the case there is only a single degree of freedom per node, the assembled FEM matrix will look like this:

$$\left( \begin{array}{cccccc|cc} X & X & & X & X & & & \\ X & X & X & X & X & X & & \\ & X & X & X & X & X & X & \\ & & X & X & & X & X & \\ X & X & & X & X & & X & X \\ X & X & X & X & X & X & X & X \\ & X & X & X & X & X & X & X \\ & & X & X & & X & X & \\ & & & X & X & X & X & X \\ & & & & X & X & X & X \\ & & & & & X & X & X \end{array} \right)$$

where the  $X$  stand for non-zero terms. This matrix structure stems from the fact that

- node 0 sees nodes 0,1,4,5
- node 1 sees nodes 0,1,2,4,5,6
- node 2 sees nodes 1,2,3,5,6,7
- ...

- node 5 sees nodes 0,1,2,4,5,6,8,9,10
- ...
- node 10 sees nodes 5,6,7,9,10,11
- node 11 sees nodes 6,7,10,11

In light thereof, we have

- 4 corner nodes which have 4 neighbours (counting themselves)
- $2(nnx-2)$  nodes which have 6 neighbours
- $2(nny-2)$  nodes which have 6 neighbours
- $(nnx-2) \times (nny-2)$  nodes which have 9 neighbours

In total, the number of non-zero terms in the matrix is then:

$$NZ = 4 \times 4 + 4 \times 6 + 2 \times 6 + 2 \times 9 = 70$$

In general, we would then have:

$$NZ = 4 \times 4 + [2(nnx - 2) + 2(nny - 2)] \times 6 + (nnx - 2)(nny - 2) \times 9$$

Let us temporarily assume  $nnx = nny = n$ . Then the matrix size (total number of unknowns) is  $N = n^2$  and

$$NZ = 16 + 24(n - 2) + 9(n - 2)^2$$

A full matrix array would contain  $N^2 = n^4$  terms. The ratio of  $NZ$  (the actual number of reals to store) to the full matrix size (the number of reals a full matrix contains) is then

$$R = \frac{16 + 24(n - 2) + 9(n - 2)^2}{n^4}$$

It is then obvious that when  $n$  is large enough  $R \sim 1/n^2$ .

CSR stores the nonzeros of the matrix row by row, in a single indexed array A of double precision numbers. Another array COLIND contains the column index of each corresponding entry in the A array. A third integer array RWPTR contains pointers to the beginning of each row, which an additional pointer to the first index following the nonzeros of the matrix A. A and COLIND have length NZ and RWPTR has length N+1.

In the case of the here-above matrix, the arrays COLIND and RWPTR will look like:

$$COLIND = (0, 1, 4, 5, 0, 1, 2, 4, 5, 6, 1, 2, 3, 5, 6, 7, \dots, 6, 7, 10, 11)$$

$$RWPTR = (0, 4, 10, 16, \dots)$$

### 6.8.2 2D domain - Two degrees of freedom per node

When there are now two degrees of freedom per node, such as in the case of the Stokes equation in two-dimensions, the size of the  $\mathbb{K}$  matrix is given by

NfemV=nnp*ndofV
-----------------

In the case of the small grid above, we have NfemV=24. Elemental matrices are now  $8 \times 8$  in size.

We still have

- 4 corner nodes which have 4 ,neighbours,
- $2(nnx-2)$  nodes which have 6 neighbours
- $2(nny-2)$  nodes which have 6 neighbours

- $(nnx-2) \times (nny-2)$  nodes which have 9 neighbours,

but now each degree of freedom from a node sees the other two degrees of freedom of another node too. In that case, the number of nonzeros has been multiplied by four and the assembled FEM matrix looks like:

Note that the degrees of freedom are organised as follows:

$$(u_0, v_0, u_1, v_1, u_2, v_2, \dots, u_{11}, v_{11})$$

In general, we would then have:

$$NZ = 4[4 \times 4 + [2(nn_x - 2) + 2(nn_y - 2)] \times 6 + (nn_x - 2)(nn_y - 2) \times 9]$$

and in the case of the small grid, the number of non-zero terms in the matrix is then:

$$NZ = 4[4 \times 4 + 4 \times 6 + 2 \times 6 + 2 \times 9] = 280$$

In the case of the here-above matrix, the arrays COLIND and RWPTR will look like:

$$COLIND = (0, 1, 2, 3, 8, 9, 10, 11, 0, 1, 2, 3, 8, 9, 10, 11, \dots)$$

$$RW PTR = (0, 8, 16, 28, \dots)$$

### 6.8.3 in fieldstone

The majority of the codes have the FE matrix being a full array

```
a_mat = np.zeros((Nfem,Nfem), dtype=np.float64)
```

and it is converted to CSR format on the fly in the solve phase:

```
sol = sps.linalg.spsolve(sps.csr_matrix(a_mat),rhs)
```

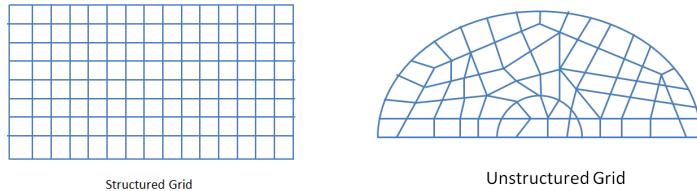
Note that linked list storages can be used (`lil_matrix`). Substantial memory savings but much longer compute times.

## 6.9 Mesh generation

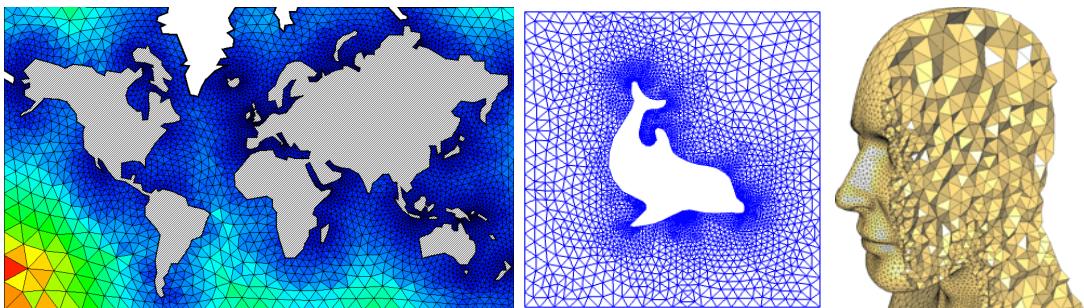
Before basis functions can be defined and PDEs can be discretised and solved we must first tessellate the domain with polygons, e.g. triangles and quadrilaterals in 2D, tetrahedra, prisms and hexahedra in 3D.

When the domain is itself simple (e.g. a rectangle, a sphere, ...) the mesh (or grid) can be (more or less) easily produced and the connectivity array filled with straightforward algorithms [82]. However, real life applications can involve extremely complex geometries (e.g. a bridge, a human spine, a car chassis and body, etc ...) and dedicated algorithms/softwares must be used (see [84, 31, 90]).

We usually distinguish between two broad classes of grids: structured grids (with a regular connectivity) and unstructured grids (with an irregular connectivity).



On the following figure are shown various triangle- tetrahedron-based meshes. These are obviously better suited for simulations of complex geometries:



add more examples coming from geo

Let us now focus on the case of a rectangular computational domain of size  $L_x \times L_y$  with a regular mesh composed of  $\text{nelx} \times \text{nely} = \text{nel}$  quadrilaterals. There are then  $\text{nnx} \times \text{nny} = \text{nnp}$  grid points. The elements are of size  $hx \times hy$  with  $hx = L_x / \text{nelx}$ .

We have no reason to come up with an irregular/illogical node numbering so we can number nodes row by row or column by column as shown on the example hereunder of a  $3 \times 2$  grid:

8=====9=====10=====11	2=====5=====8=====11
(3)   (4)   (5)	(1)   (3)   (5)
4=====5=====6=====7	1=====4=====7=====10
(0)   (1)   (2)	(0)   (2)   (4)
0=====1=====2=====3	0=====3=====6=====9

"row by row"

"column by column"

The numbering of the elements themselves could be done in a somewhat chaotic way but we follow the numbering of the nodes for simplicity. The row by row option is the adopted one in **fieldstone** and the coordinates of the points are computed as follows:

```
x = np.empty(nnp, dtype=np.float64)
y = np.empty(nnp, dtype=np.float64)
```

```

counter = 0
for j in range(0,nnx):
    for i in range(0,nnx):
        x[counter]=i*hx
        y[counter]=j*hy
        counter += 1

```

The inner loop has *i* ranging from 0 to *nnx*-1 first for *j*=0, 1, ... up to *nnx*-1 which indeed corresponds to the row by row numbering.

We now turn to the connectivity. As mentioned before, this is a structured mesh so that the so-called connectivity array, named `icon` in our case, can be filled easily. For each element we need to store the node identities of its vertices. Since there are *nel* elements and *m*=4 corners, this is a *m*×*nel* array. The algorithm goes as follows:

```

icon =np.zeros((m, nel), dtype=np.int16)
counter = 0
for j in range(0,nely):
    for i in range(0,nelx):
        icon[0,counter] = i + j * nnx
        icon[1,counter] = i + 1 + j * nnx
        icon[2,counter] = i + 1 + (j + 1) * nnx
        icon[3,counter] = i + (j + 1) * nnx
        counter += 1

```

In the case of the  $3 \times 2$  mesh, the `icon` is filled as follows:

element id→	0	1	2	3	4	5
node id↓						
0	0	1	2	4	5	6
1	1	2	3	5	6	7
2	5	6	7	9	10	11
3	4	5	6	8	9	10

It is to be understood as follows: element #4 is composed of nodes 5, 6, 10 and 9. Note that nodes are always stored in a counter clockwise manner, starting at the bottom left. This is very important since the corresponding basis functions and their derivatives will be labelled accordingly.

In three dimensions things are very similar. The mesh now counts *nelx*×*nely*×*nelz*=*nel* elements which represent a cuboid of size *Lx*×*Ly*×*Lz*. The position of the nodes is obtained as follows:

```

x = np.empty(nnp, dtype=np.float64)
y = np.empty(nnp, dtype=np.float64)
z = np.empty(nnp, dtype=np.float64)
counter=0
for i in range(0,nnx):
    for j in range(0,nny):
        for k in range(0,nnz):
            x[counter]=i*hx
            y[counter]=j*hy
            z[counter]=k*hz
            counter += 1

```

The connectivity array is now of size *m*×*nel* with *m*=8:

```

icon =np.zeros((m, nel), dtype=np.int16)
counter = 0
for i in range(0,nelx):
    for j in range(0,nely):
        for k in range(0,nelz):
            icon[0,counter]=nny*nnz*(i )+nnz*(j )+k
            icon[1,counter]=nny*nnz*(i+1)+nnz*(j )+k
            icon[2,counter]=nny*nnz*(i+1)+nnz*(j+1)+k
            icon[3,counter]=nny*nnz*(i )+nnz*(j+1)+k
            icon[4,counter]=nny*nnz*(i+1)+nnz*(j )+k+1
            icon[5,counter]=nny*nnz*(i+1)+nnz*(j )+k+1
            icon[6,counter]=nny*nnz*(i+1)+nnz*(j+1)+k+1
            icon[7,counter]=nny*nnz*(i )+nnz*(j+1)+k+1
            counter += 1

```

produce drawing of node numbering

## 6.10 Visco-Plasticity

### 6.10.1 Tensor invariants

Before we dive into the world of nonlinear rheologies it is necessary to introduce the concept of tensor invariants since they are needed further on. Unfortunately there are many different notations used in the literature and these can prove to be confusing.

Given a tensor  $\mathbf{T}$ , one can compute its (moment) invariants as follows:

- first invariant :

$$\begin{aligned} T_I|^{2D} &= \text{Tr}[\mathbf{T}] = T_{xx} + T_{yy} \\ T_I|^{3D} &= \text{Tr}[\mathbf{T}] = T_{xx} + T_{yy} + T_{zz} \end{aligned}$$

- second invariant :

$$\begin{aligned} T_{II}|^{2D} &= \frac{1}{2} \text{Tr}[\mathbf{T}^2] = \frac{1}{2} \sum_{ij} T_{ij} T_{ji} = \frac{1}{2} (T_{xx}^2 + T_{yy}^2) + T_{xy}^2 \\ T_{II}|^{3D} &= \frac{1}{2} \text{Tr}[\mathbf{T}^2] = \frac{1}{2} \sum_{ij} T_{ij} T_{ji} = \frac{1}{2} (T_{xx}^2 + T_{yy}^2 + T_{zz}^2) + T_{xy}^2 + T_{xz}^2 + T_{yz}^2 \end{aligned}$$

- third invariant :

$$T_{III} = \frac{1}{3} \text{Tr}[\mathbf{T}^3] = \frac{1}{3} \sum_{ijk} T_{ij} T_{jk} T_{ki}$$

The implementation of the plasticity criterions relies essentially on the second invariants of the (deviatoric) stress  $\boldsymbol{\tau}$  and the (deviatoric) strainrate tensors  $\dot{\boldsymbol{\varepsilon}}$ :

$$\begin{aligned} \tau_{II}|^{2D} &= \frac{1}{2} (\tau_{xx}^2 + \tau_{yy}^2) + \tau_{xy}^2 \\ &= \frac{1}{4} (\sigma_{xx} - \sigma_{yy})^2 + \sigma_{xy}^2 \\ &= \frac{1}{4} (\sigma_1 - \sigma_2)^2 \\ \tau_{II}|^{3D} &= \frac{1}{2} (\tau_{xx}^2 + \tau_{yy}^2 + \tau_{zz}^2) + \tau_{xy}^2 + \tau_{xz}^2 + \tau_{yz}^2 \\ &= \frac{1}{6} [(\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{yy} - \sigma_{zz})^2 + (\sigma_{xx} - \sigma_{zz})^2] + \sigma_{xy}^2 + \sigma_{xz}^2 + \sigma_{yz}^2 \\ &= \frac{1}{6} [(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2] \\ \varepsilon_{II}|^{2D} &= \frac{1}{2} [(\dot{\varepsilon}_{xx}^d)^2 + (\dot{\varepsilon}_{yy}^d)^2] + (\dot{\varepsilon}_{xy}^d)^2 \\ &= \frac{1}{2} \left[ \frac{1}{4} (\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + \frac{1}{4} (\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{xx})^2 \right] + \dot{\varepsilon}_{xy}^2 \\ &= \frac{1}{4} (\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + \dot{\varepsilon}_{xy}^2 \\ \varepsilon_{II}|^{3D} &= \frac{1}{2} [(\dot{\varepsilon}_{xx}^d)^2 + (\dot{\varepsilon}_{yy}^d)^2 + (\dot{\varepsilon}_{zz}^d)^2] + (\dot{\varepsilon}_{xy}^d)^2 + (\dot{\varepsilon}_{xz}^d)^2 + (\dot{\varepsilon}_{yz}^d)^2 \\ &= \frac{1}{6} [(\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{yy})^2 + (\dot{\varepsilon}_{yy} - \dot{\varepsilon}_{zz})^2 + (\dot{\varepsilon}_{xx} - \dot{\varepsilon}_{zz})^2] + \dot{\varepsilon}_{xy}^2 + \dot{\varepsilon}_{xz}^2 + \dot{\varepsilon}_{yz}^2 \end{aligned}$$

Note that these (second) invariants are almost always used under a square root so we define:

$$\underline{\tau}_{II} = \sqrt{\tau_{II}} \quad \dot{\underline{\varepsilon}}_{II} = \sqrt{\dot{\varepsilon}_{II}}$$

Note that these quantities have the same dimensions as their tensor counterparts, i.e. Pa for stresses and  $\text{s}^{-1}$  for strain rates.

### 6.10.2 Scalar viscoplasticity

This formulation is quite easy to implement. It is widely used, e.g. [89, 83, 75], and relies on the assumption that a scalar quantity  $\eta_p$  (the 'effective plastic viscosity') exists such that the deviatoric stress tensor

$$\boldsymbol{\tau} = 2\eta_p \dot{\boldsymbol{\varepsilon}} \quad (51)$$

is bounded by some yield stress value  $Y$ . From Eq. (51) it follows that  $\underline{\tau}_{II} = 2\eta_p \dot{\underline{\varepsilon}}_{II} = Y$  which yields

$$\eta_p = \frac{Y}{2\dot{\underline{\varepsilon}}_{II}}$$

This approach has also been coined the Viscosity Rescaling Method (VRM) [46].

insert here the rederivation 2.1.1 of spmw16

It is at this stage important to realise that (i) in areas where the strainrate is low, the resulting effective viscosity will be large, and (ii) in areas where the strainrate is high, the resulting effective viscosity will be low. This is not without consequences since (effective) viscosity contrasts up to 8-10 orders of magnitude have been observed/obtained with this formulation and it makes the FE matrix very stiff, leading to (iterative) solver convergence issues. In order to contain these viscosity contrasts one usually resorts to viscosity limiters  $\eta_{min}$  and  $\eta_{max}$  such that

$$\eta_{min} \leq \eta_p \leq \eta_{max}$$

Caution must be taken when choosing both values as they may influence the final results.

▷ `python_codes/fieldstone_indentor`

### 6.10.3 about the yield stress value $Y$

In geodynamics the yield stress value is often given as a simple function. It can be constant (in space and time) and in this case we are dealing with a von Mises plasticity yield criterion. . We simply assume  $Y_{vM} = C$  where  $C$  is a constant cohesion independent of pressure, strainrate, deformation history, etc ...

Another model is often used: the Drucker-Prager plasticity model. A friction angle  $\phi$  is then introduced and the yield value  $Y$  takes the form

$$Y_{DP} = p \sin \phi + C \cos \phi$$

and therefore depends on the pressure  $p$ . Because  $\phi$  is with the range  $[0^\circ, 45^\circ]$ ,  $Y$  is found to increase with depth (since the lithostatic pressure often dominates the overpressure).

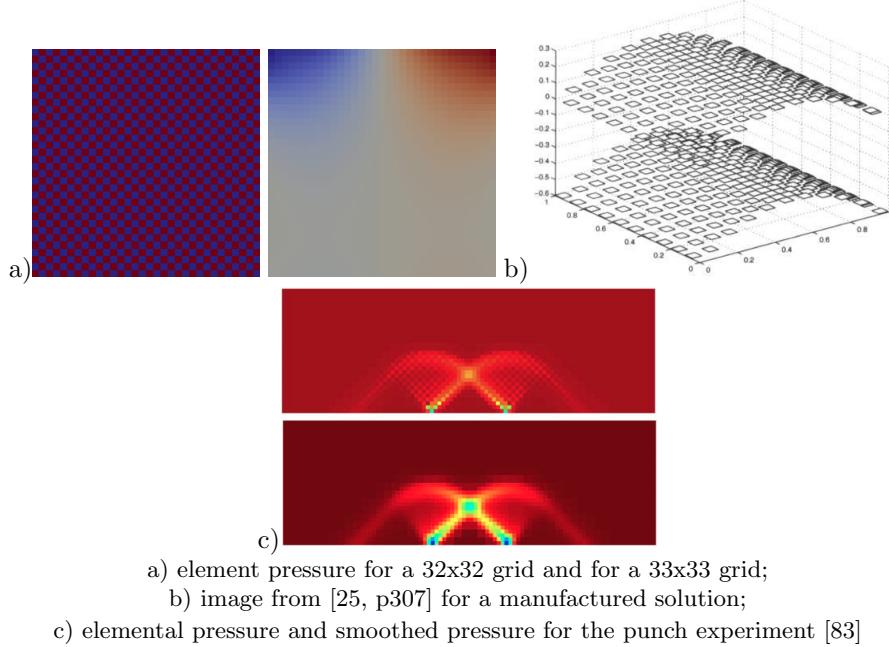
Note that a slightly modified version of this plasticity model has been used: the total pressure  $p$  is then replaced by the lithostatic pressure  $p_{lith}$ .

## 6.11 Pressure smoothing

It has been widely documented that the use of the  $Q_1 \times P_0$  element is not without problems. Aside from the consequences it has on the FE matrix properties, we will here focus on another unavoidable side effect: the spurious pressure checkerboard modes.

These modes have been thoroughly analysed [38, 15, 69, 70]. They can be filtered out [15] or simply smoothed [50].

On the following figure (a,b), pressure fields for the lid driven cavity experiment are presented for both an even and un-even number of elements. We see that the amplitude of the modes can sometimes be so large that the 'real' pressure is not visible and that something as simple as the number of elements in the domain can trigger those or not at all.



The easiest post-processing step that can be used (especially when a regular grid is used) is explained in [83]: "The element-to-node interpolation is performed by averaging the elemental values from elements common to each node; the node-to-element interpolation is performed by averaging the nodal values element-by-element. This method is not only very efficient but produces a smoothing of the pressure that is adapted to the local density of the octree. Note that these two steps can be repeated until a satisfying level of smoothness (and diffusion) of the pressure field is attained."

In the codes which rely on the  $Q_1 \times P_0$  element, the (elemental) pressure is simply defined as

```
p=np.zeros(nel, dtype=np.float64)
```

while the nodal pressure is then defined as

```
q=np.zeros(nnp, dtype=np.float64)
```

The element-to-node algorithm is then simply (in 2D):

```
count=np.zeros(nnp, dtype=np.int16)
for iel in range(0,nel):
    q[icon[0,iel]]+=p[iel]
    q[icon[1,iel]]+=p[iel]
    q[icon[2,iel]]+=p[iel]
    q[icon[3,iel]]+=p[iel]
    count[icon[0,iel]]+=1
    count[icon[1,iel]]+=1
    count[icon[2,iel]]+=1
    count[icon[3,iel]]+=1
q=q/count
```

[produce figure to explain this](#)

[link to proto paper](#)

[link to least square and nodal derivatives](#)

## 6.12 Pressure scaling

As perfectly explained in the step 32 of deal.ii<sup>2</sup>, we often need to scale the  $\mathbb{G}$  term since it is many orders of magnitude smaller than  $\mathbb{K}$ , which introduces large inaccuracies in the solving process to the point that the solution is nonsensical. This scaling coefficient is  $\eta/L$ . We start from

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} f \\ h \end{pmatrix}$$

and introduce the scaling coefficient as follows (which in fact does not alter the solution at all):

$$\begin{pmatrix} \mathbb{K} & \frac{\eta}{L}\mathbb{G} \\ \frac{\eta}{L}\mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \underline{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} f \\ \frac{\eta}{L}h \end{pmatrix}$$

We then end up with the modified Stokes system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \underline{\mathbb{G}}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \underline{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} f \\ \underline{h} \end{pmatrix}$$

where

$$\underline{\mathbb{G}} = \frac{\eta}{L}\mathbb{G} \quad \underline{\mathcal{P}} = \frac{L}{\eta}\mathcal{P} \quad \underline{h} = \frac{\eta}{L}h$$

After the solve phase, we recover the real pressure with  $\mathcal{P} = \frac{\eta}{L}\underline{\mathcal{P}}$ .

---

<sup>2</sup>[https://www.dealii.org/9.0.0/doxygen/deal.II/step\\_32.html](https://www.dealii.org/9.0.0/doxygen/deal.II/step_32.html)

### 6.13 Pressure normalisation

When Dirichlet boundary conditions are imposed everywhere on the boundary, pressure is only present by its gradient in the equations. It is thus determined up to an arbitrary constant. In such a case, one commonly impose the average of the pressure over the whole domain or on a subset of the boundary to be have a zero average, i.e.

$$\int_{\Omega} pdV = 0$$

Another possibility is to impose the pressure value at a single node.

Let us assume that we are using  $Q_1 \times P_0$  elements. Then the pressure is constant inside each element. The integral above becomes:

$$\int_{\Omega} pdV = \sum_e \int_{\Omega_e} pdV = \sum_e p_e \int_{\Omega_e} dV = \sum_e p_e A_e =$$

where the sum runs over all elements  $e$  of area  $A_e$ . This can be rewritten

$$\mathbb{L}^T \mathcal{P} = 0$$

and it is a constraint on the pressure. As we have seen before ??, we can associate to it a Lagrange multiplier  $\lambda$  so that we must solve the modified Stokes system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} & 0 \\ \mathbb{G}^T & 0 & \mathbb{L} \\ 0 & \mathbb{L}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \mathcal{P} \\ \lambda \end{pmatrix} = \begin{pmatrix} f \\ h \\ 0 \end{pmatrix}$$

When higher order spaces are used for pressure (continuous or discontinuous) one must then carry out the above integration numerically by means of (usually) a Gauss-Legendre quadrature.

```

▷ python_codes/fieldstone_stokes_sphere_3D_saddlepoint/
▷ python_codes/fieldstone_burstedde/
▷ python_codes/fieldstone_busse/
▷ python_codes/fieldstone_RTinstability1/
▷ python_codes/fieldstone_saddlepoint_q2q1/
▷ python_codes/fieldstone_compressible2/
▷ python_codes/fieldstone_compressible1/
▷ python_codes/fieldstone_saddlepoint_q3q2/

```

## 6.14 Convergence criterion for nonlinear iterations

## 6.15 The choice of solvers

Let us first look at the penalty formulation. In this case we are only solving for velocity since pressure is recovered in a post-processing step. We also know that the penalty factor is many orders of magnitude higher than the viscosity and in combination with the use of the  $Q_1 \times P_0$  element the resulting matrix condition number is very high so that the use of iterative solvers is precluded. Indeed codes such as SOPALE [32], DOUAR [9], or FANTOM [81] relying on the penalty formulation all use direct solvers (BLKFCT, MUMPS, WSMP).

The main advantage of direct solvers is used in this case: They can solve ill-conditioned matrices. However memory requirements for the storage of number of nonzeros in the Cholesky matrix grow very fast as the number of equations/grid size increases, especially in 3D, to the point that even modern computers with tens of Gb of RAM cannot deal with a  $100^3$  element mesh. This explains why direct solvers are often used for 2D problems and rarely in 3D with noticeable exceptions [83, 91, 10, 54, 2, 3, 4, 88, 60].

In light of all this, let us start again from the (full) Stokes system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} f \\ h \end{pmatrix}$$

We need to solve this system in order to obtain the solution, i.e. the  $\mathcal{V}$  and  $\mathcal{P}$  vectors. But how?

Unfortunately, this question is not simple to answer and the 'right' depends on many parameters. How big is the matrix ? what is the condition number of the matrix  $\mathbb{K}$  ? Is it symmetric ? (i.e. how are compressible terms handled?).

### 6.15.1 The Schur complement approach

Let us write the above system as two equations:

$$\mathbb{K}\mathcal{V} + \mathbb{G}\mathcal{P} = f \quad (52)$$

$$\mathbb{G}^T\mathcal{V} = h \quad (53)$$

The first line can be re-written  $\mathcal{V} = \mathbb{K}^{-1}(f - \mathbb{G}\mathcal{P})$  and can be inserted in the second:

$$\mathbb{G}^T\mathcal{V} = \mathbb{G}^T[\mathbb{K}^{-1}(f - \mathbb{G}\mathcal{P})] = h$$

or,

$$\mathbb{G}^T\mathbb{K}^{-1}\mathbb{G}\mathcal{P} = \mathbb{G}^T\mathbb{K}^{-1}f - h$$

The matrix  $\mathbb{S} = \mathbb{G}^T\mathbb{K}^{-1}\mathbb{G}$  is called the Schur complement. It is Symmetric (since  $\mathbb{K}$  is symmetric) and Positive-Definite<sup>3</sup> (SPD) if  $Ker(\mathbb{G}) = 0$ . [look in donea-huerta book for details](#) Having solved this equation (we have obtained  $\mathcal{P}$ ), the velocity can be recovered by solving  $\mathbb{K}\mathcal{V} = f - \mathbb{G}\mathcal{P}$ . For now, let us assume that we have built the  $\mathbb{S}$  matrix and the right hand side  $\tilde{f} = \mathbb{G}^T\mathbb{K}^{-1}f - h$ . We must solve  $\mathbb{S}\mathcal{P} = \tilde{f}$ .

Since  $\mathbb{S}$  is SPD, the Conjugate Gradient (CG) method is very appropriate to solve this system. Indeed, looking at the definition of Wikipedia: *In mathematics, the conjugate gradient method is an algorithm for the numerical solution of particular systems of linear equations, namely those whose matrix is symmetric and positive-definite. The conjugate gradient method is often implemented as an iterative algorithm, applicable to sparse systems that are too large to be handled by a direct implementation or other direct methods such as the Cholesky decomposition. Large sparse systems often arise when numerically solving partial differential equations or optimization problems.*

The Conjugate Gradient algorithm is as follows:

<sup>3</sup> $M$  positive definite  $\iff x^T M x > 0 \forall x \in \mathbb{R}^n \setminus \mathbf{0}$

```

r0 := b - Ax0
if r0 is sufficiently small, then return x0 as the result
p0 := r0
k := 0
repeat
     $\alpha_k := \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k}$ 
     $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k$ 
     $\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A} \mathbf{p}_k$ 
    if  $\mathbf{r}_{k+1}$  is sufficiently small, then exit loop
     $\beta_k := \frac{\mathbf{r}_{k+1}^\top \mathbf{r}_{k+1}}{\mathbf{r}_k^\top \mathbf{r}_k}$ 
     $\mathbf{p}_{k+1} := \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k$ 
    k := k + 1
end repeat
return  $\mathbf{x}_{k+1}$  as the result

```

Algorithm obtained from [https://en.wikipedia.org/wiki/Conjugate\\_gradient\\_method](https://en.wikipedia.org/wiki/Conjugate_gradient_method)

Let us look at this algorithm up close. The parts which may prove to be somewhat tricky are those involving the matrix (in our case the Schur complement). We start the iterations with a guess pressure  $\mathcal{P}_0$  ( and an initial guess velocity which could be obtained by solving  $\mathbb{K}\mathcal{V}_0 = f - \mathbb{G}\mathcal{P}_0$ ).

$$r_0 = \tilde{f} - \mathbb{S}\mathcal{P}_0 \quad (54)$$

$$= \mathbb{G}^T \mathbb{K}^{-1} f - h - (\mathbb{G}^T \mathbb{K}^{-1} \mathbb{G}) \mathcal{P}_0 \quad (55)$$

$$= \mathbb{G}^T \mathbb{K}^{-1} (f - \mathbb{G}\mathcal{P}_0) - h \quad (56)$$

$$= \mathbb{G}^T \mathbb{K}^{-1} \mathbb{K}\mathcal{V}_0 - h \quad (57)$$

$$= \mathbb{G}^T \mathcal{V}_0 - h \quad (58)$$

$$= \mathbb{G}^T \mathcal{V}_0 - h \quad (59)$$

We now turn to the  $\alpha_k$  coefficient:

$$\alpha_k = \frac{r_k^\top r_k}{p_k^\top \mathbb{S} p_k} = \frac{r_k^\top r_k}{p_k \mathbb{G}^T \mathbb{K}^{-1} \mathbb{G} p_k} = \frac{r_k^\top r_k}{(\mathbb{G} p_k)^\top \mathbb{K}^{-1} (\mathbb{G} p_k)}$$

We then define  $\tilde{p}_k = \mathbb{G} p_k$ , so that  $\alpha_k$  can be computed as follows:

1. compute  $\tilde{p}_k = \mathbb{G} p_k$
2. solve  $\mathbb{K}d_k = \tilde{p}_k$
3. compute  $\alpha_k = (r_k^\top r_k) / (\tilde{p}_k^\top d_k)$

Then we need to look at the term  $\mathbb{S}p_k$ :

$$\mathbb{S}p_k = \mathbb{G}^T \mathbb{K}^{-1} \mathbb{G} p_k = \mathbb{G}^T \mathbb{K}^{-1} \tilde{p}_k = \mathbb{G}^T d_k$$

We can then rewrite the CG algorithm as follows:

- $r_0 = \mathbb{G}^T \mathcal{V}_0 - h$
- if  $r_0$  is sufficiently small, then return  $(\mathcal{V}_0, \mathcal{P}_0)$  as the result
- $p_0 = r_0$
- $k = 0$
- repeat
  - compute  $\tilde{p}_k = \mathbb{G} p_k$
  - solve  $\mathbb{K}d_k = \tilde{p}_k$
  - compute  $\alpha_k = (r_k^\top r_k) / (\tilde{p}_k^\top d_k)$

- $\mathcal{P}_{k+1} = \mathcal{P}_k + \alpha_k p_k$
- $r_{k+1} = r_k - \alpha_k \mathbb{G}^T d_k$
- if  $r_{k+1}$  is sufficiently small, then exit loop
- $\beta_k = (r_{k+1}^T r_{k+1}) / (r_k^T r_k)$
- $p_{k+1} = r_{k+1} + \beta_k p_k$
- $k = k + 1$

- return  $\mathcal{P}_{k+1}$  as result

We see that we have managed to solve the Schur complement equation with the Conjugate Gradient method without ever building the matrix  $\mathbb{S}$ . Having obtained the pressure solution, we can easily recover the corresponding velocity with  $\mathbb{K}\mathcal{V}_{k+1} = f - \mathbb{G}\mathcal{P}_{k+1}$ . However, this is rather unfortunate because it requires yet another solve with the  $\mathbb{K}$  matrix. As it turns out, we can slightly alter the above algorithm to have it update the velocity as well so that this last solve is unnecessary.

We have

$$\mathcal{V}_{k+1} = \mathbb{K}^{-1}(f - \mathbb{G}\mathcal{P}_{k+1}) = \mathbb{K}^{-1}(f - \mathbb{G}(\mathcal{P}_k + \alpha_k p_k)) = \mathbb{K}^{-1}(f - \mathbb{G}\mathcal{P}_k) - \alpha_k \mathbb{K}^{-1} \mathbb{G} p_k = \mathcal{V}_k - \alpha_k \mathbb{K}^{-1} \tilde{p}_k = \mathcal{V}_k - \alpha_k d_k$$

and we can insert this minor extra calculation inside the algorithm and get the velocity solution nearly for free. The final CG algorithm is then

- $r_0 = \mathbb{G}^T \mathcal{V}_0 - h$
  - if  $r_0$  is sufficiently small, then return  $(\mathcal{V}_0, \mathcal{P}_0)$  as the result
  - $p_0 = r_0$
  - $k = 0$
  - repeat
    - compute  $\tilde{p}_k = \mathbb{G} p_k$
    - solve  $\mathbb{K} d_k = \tilde{p}_k$
    - compute  $\alpha_k = (r_k^T r_k) / (\tilde{p}_k^T d_k)$
    - $\mathcal{P}_{k+1} = \mathcal{P}_k + \alpha_k p_k$
    - $\mathcal{V}_{k+1} = \mathcal{V}_k - \alpha_k d_k$
    - $r_{k+1} = r_k - \alpha_k \mathbb{G}^T d_k$
    - if  $r_{k+1}$  is sufficiently small, then exit loop
    - $\beta_k = (r_{k+1}^T r_{k+1}) / (r_k^T r_k)$
    - $p_{k+1} = r_{k+1} + \beta_k p_k$
    - $k = k + 1$
  - return  $\mathcal{P}_{k+1}$  as result

This iterative algorithm will converge to the solution with a rate which depends on the condition number of the  $\mathbb{S}$  matrix, which is not easily obtainable since  $\mathbb{S}$  is never built. Also, we know that large viscosity contrasts in the domain will influence this too. Finally, we notice that this algorithm requires one solve with matrix  $\mathbb{K}$  per iteration but says nothing about the method employed to do so.

One thing we know improves the convergence of any iterative solver is the use of a preconditioner matrix.

**PCG**

## 6.16 The GMRES approach

## 6.17 The consistent boundary flux (CBF)

Original paper: [57, 94]

Find all papers citing it.

[95]

in ASPECT: aspect/source/postprocess//dynamic\_topography.cc

## 6.18 Static condensation

## 6.19 Exporting data to vtk format

This format seems to be the universally accepted format for 2D and 3D visualisation in Computational Geodynamics. Such files can be opened with free softwares such as Paraview<sup>4</sup>, MayaVi<sup>5</sup> or Visit<sup>6</sup>.

Unfortunately it is my experience that no simple tutorial exists about how to build such files. There is an official document which describes the vtk format<sup>7</sup> but it delivers the information in a convoluted way. I therefore describe hereafter how **fieldstone** builds the vtk files.

I hereunder show vtk file corresponding to the 3x2 grid presented earlier 6.9. In this particular example there are:

- 12 nodes and 6 elements
- 1 elemental field: the pressure  $p$ )
- 2 nodal fields: 1 scalar (the smoothed pressure  $q$ ), 1 vector (the velocity field  $u, v, 0$ )

Note that vtk files are inherently 3D so that even in the case of a 2D simulation the  $z$ -coordinate of the points and for instance their  $z$ -velocity component must be provided. The file, usually called *solution.vtu* starts with a header:

```
<VTKFile type='UnstructuredGrid' version='0.1' byte_order='BigEndian'>
<UnstructuredGrid>
<Piece NumberOfPoints='12' NumberOfCells='6'>
```

We then proceed to write the node coordinates as follows:

```
<Points>
<DataArray type='Float32' NumberOfComponents='3' Format='ascii'>
0.000000e+00 0.000000e+00 0.000000e+00
3.333333e-01 0.000000e+00 0.000000e+00
6.666667e-01 0.000000e+00 0.000000e+00
1.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 5.000000e-01 0.000000e+00
3.333333e-01 5.000000e-01 0.000000e+00
6.666667e-01 5.000000e-01 0.000000e+00
1.000000e+00 5.000000e-01 0.000000e+00
0.000000e+00 1.000000e+00 0.000000e+00
3.333333e-01 1.000000e+00 0.000000e+00
6.666667e-01 1.000000e+00 0.000000e+00
1.000000e+00 1.000000e+00 0.000000e+00
</DataArray>
</Points>
```

These are followed by the elemental field(s):

```
<CellData Scalars='scalars'>
<DataArray type='Float32' Name='p' Format='ascii'>
-1.333333e+00
-3.104414e-10
1.333333e+00
-1.333333e+00
8.278417e-17
1.333333e+00
</DataArray>
</CellData>
```

Nodal quantities are written next:

```
<PointData Scalars='scalars'>
<DataArray type='Float32' NumberOfComponents='3' Name='velocity' Format='ascii'>
0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00
```

<sup>4</sup><https://www.paraview.org/>

<sup>5</sup><https://docs.enthought.com/mayavi/mayavi/>

<sup>6</sup><https://wci.llnl.gov/simulation/computer-codes/visit/>

<sup>7</sup><https://www.vtk.org/wp-content/uploads/2015/04/file-formats.pdf>

```

0.000000e+00 0.000000e+00 0.000000e+00
8.888885e-08 -8.278405e-24 0.000000e+00
8.888885e-08 1.655682e-23 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00
1.000000e+00 0.000000e+00 0.000000e+00
</DataArray>
<DataArray type='Float32' NumberOfComponents='1' Name='q' Format='ascii'>
-1.333333e+00
-6.666664e-01
6.666664e-01
1.333333e+00
-1.333333e+00
-6.666664e-01
6.666664e-01
1.333333e+00
-1.333333e+00
-6.666664e-01
6.666664e-01
1.333333e+00
</DataArray>
</PointData>

```

To these informations we must append 3 more datasets. The first one is the connectivity, the second one is the offsets and the third one is the type. The first one is trivial since said connectivity is needed for the Finite Elements. The second must be understood as follows: when reading the connectivity information in a linear manner the offset values indicate the beginning of each element (omitting the zero value). The third simply is the type of element as given in the vtk format document (9 corresponds to a generic quadrilateral with an internal numbering consistent with ours).

```

<Cells>
<DataArray type='Int32' Name='connectivity' Format='ascii'>
0 1 5 4
1 2 6 5
2 3 7 6
4 5 9 8
5 6 10 9
6 7 11 10
</DataArray>
<DataArray type='Int32' Name='offsets' Format='ascii'>
4
8
12
16
20
24
</DataArray>
<DataArray type='Int32' Name='types' Format='ascii'>
9
9
9
9
9
9
</DataArray>
</Cells>

```

The file is then closed with

```

</Piece>
</UnstructuredGrid>
</VTKFile>

```

The *solution.vtu* file can then be opened with ParaView, MayaVi or Visit and the reader is advised to find tutorials online on how to install and use these softwares.

**To Do:**

write about impose bc on el matrix  
full compressible  
total energy calculations  
constraints  
compositions, marker chain  
van keken initial value with deformed mesh  
free-slip bc on annulus and sphere . See for example p540 Gresho and Sani book.  
non-linear rheologies (two layer brick spmw16, tosn15)  
Picard vs Newton  
markers  
Schur complement approach  
periodic boundary conditions  
open boundary conditions  
free surface  
SUPG  
zaleski disk advection  
Busse convection pb, compare with aspect  
cvi !!!  
pure elastic  
including phase changes (w. R. Myhill)  
discontinuous galerkin  
nonlinear poiseuille  
formatting of code style  
navier-stokes ? (LUKAS)  
compute strainrate in middle of element or at quad point for punch?  
GEO1442 code  
GEO1442 indenter setup in plane ?  
in/out flow on sides for lith modelling  
Fehlberg RK advection  
Problems to solve:  
colorscale  
better yet simple matrix storage ?  
write Scott about matching compressible2 setup with his paper  
deal with large matrices.

## 7 fieldstone: simple analytical solution

From [25]. In order to illustrate the behavior of selected mixed finite elements in the solution of stationary Stokes flow, we consider a two-dimensional problem in the square domain  $\Omega = [0, 1] \times [0, 1]$ , which possesses a closed-form analytical solution. The problem consists of determining the velocity field  $\mathbf{v} = (u, v)$  and the pressure  $p$  such that

$$-\nu \Delta \mathbf{v} + \nabla p = \mathbf{b} \quad \text{in } \Omega$$

$$\nabla \cdot \mathbf{v} = 0 \quad \text{in } \Omega$$

$$\mathbf{v} = \mathbf{0} \quad \text{on } \Gamma$$

where the fluid viscosity is taken as  $\nu = 1$ . The components of the body force  $\mathbf{b}$  are prescribed as

$$\begin{aligned} b_x &= (12 - 24y)x^4 + (-24 + 48y)x^3 + (-48y + 72y^2 - 48y^3 + 12)x^2 \\ &\quad + (-2 + 24y - 72y^2 + 48y^3)x + 1 - 4y + 12y^2 - 8y^3 \\ b_y &= (8 - 48y + 48y^2)x^3 + (-12 + 72y - 72y^2)x^2 \\ &\quad + (4 - 24y + 48y^2 - 48y^3 + 24y^4)x - 12y^2 + 24y^3 - 12y^4 \end{aligned}$$

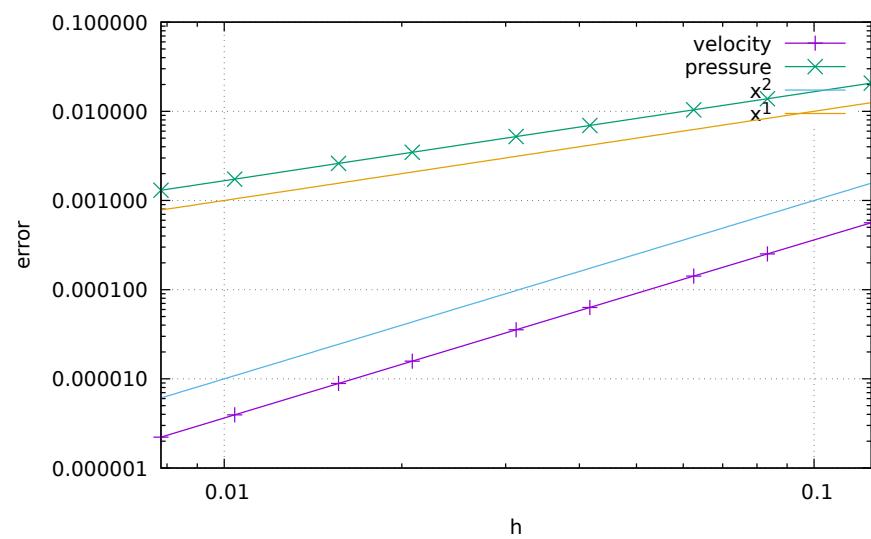
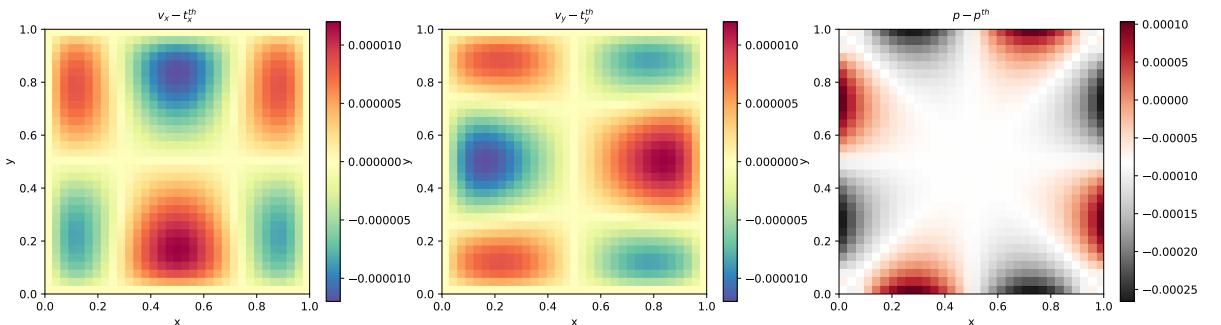
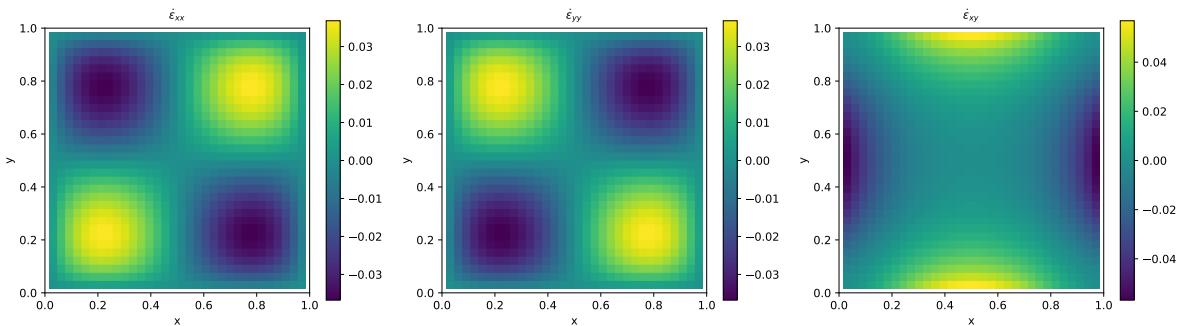
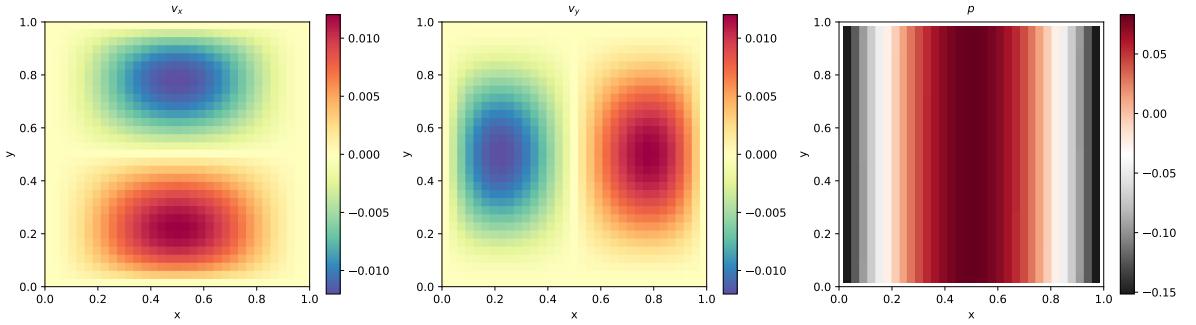
With this prescribed body force, the exact solution is

$$\begin{aligned} u(x, y) &= x^2(1-x)^2(2y-6y^2+4y^3) \\ v(x, y) &= -y^2(1-y)^2(2x-6x^2+4x^3) \\ p(x, y) &= x(1-x) - 1/6 \end{aligned}$$

Note that the pressure obeys  $\int_{\Omega} p \, d\Omega = 0$

### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (no-slip)
- direct solver
- isothermal
- isoviscous
- analytical solution



Quadratic convergence for velocity error, linear convergence for pressure error, as expected.

ToDo:

pressure normalisation?

different cmat, a la schmalholz

To go further:

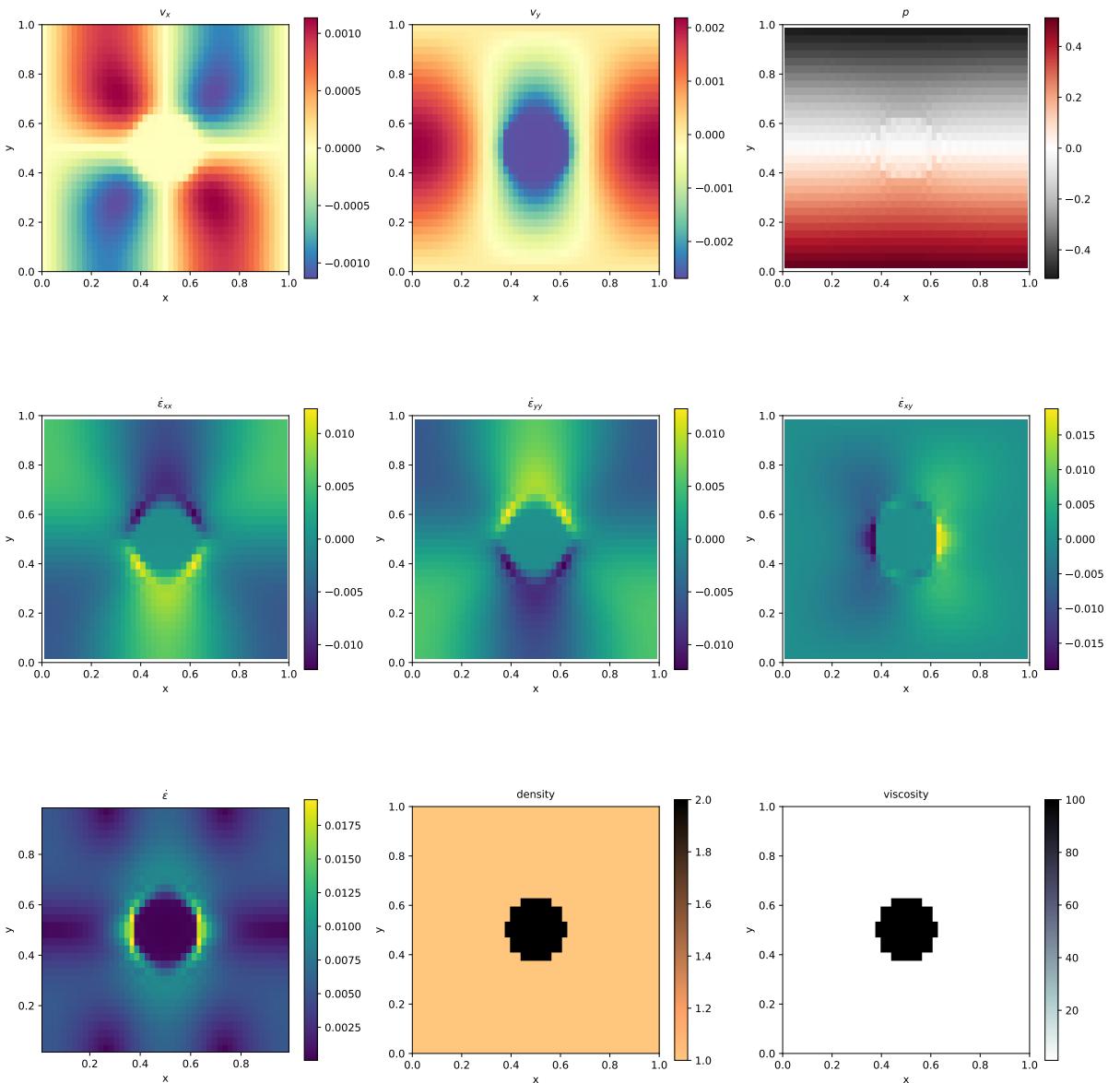
1. make your own analytical solution

## 8 fieldstone: Stokes sphere

Viscosity and density directly computed at the quadrature points.

### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (free-slip)
- direct solver
- isothermal
- non-isoviscous
- buoyancy-driven flow



## 9 fieldstone: Convection in a 2D box

This benchmark deals with the 2-D thermal convection of a fluid of infinite Prandtl number in a rectangular closed cell. In what follows, I carry out the case 1a, 1b, and 1c experiments as shown in [8]: steady convection with constant viscosity in a square box.

The temperature is fixed to zero on top and to  $\Delta T$  at the bottom, with reflecting symmetry at the sidewalls (i.e.  $\partial_x T = 0$ ) and there are no internal heat sources. Free-slip conditions are implemented on all boundaries.

The Rayleigh number is given by

$$Ra = \frac{\alpha g_y \Delta T h^3}{\kappa \nu} = \frac{\alpha g_y \Delta T h^3 \rho^2 c_p}{k \mu} \quad (60)$$

In what follows, I use the following parameter values:  $L_x = L_y = 1, \rho_0 = c_P = k = \mu = 1, T_0 = 0, \alpha = 10^{-2}, g = 10^2 Ra$  and I run the model with  $Ra = 10^4, 10^5$  and  $10^6$ .

The initial temperature field is given by

$$T(x, y) = (1 - y) - 0.01 \cos(\pi x) \sin(\pi y) \quad (61)$$

The perturbation in the initial temperature fields leads to a perturbation of the density field and sets the fluid in motion.

Depending on the initial Rayleigh number, the system ultimately reaches a steady state after some time.

The Nusselt number (i.e. the mean surface temperature gradient over mean bottom temperature) is computed as follows [8]:

$$Nu = L_y \frac{\int \frac{\partial T}{\partial y} (y = L_y) dx}{\int T(y = 0) dx} \quad (62)$$

Note that in our case the denominator is equal to 1 since  $L_x = 1$  and the temperature at the bottom is prescribed to be 1.

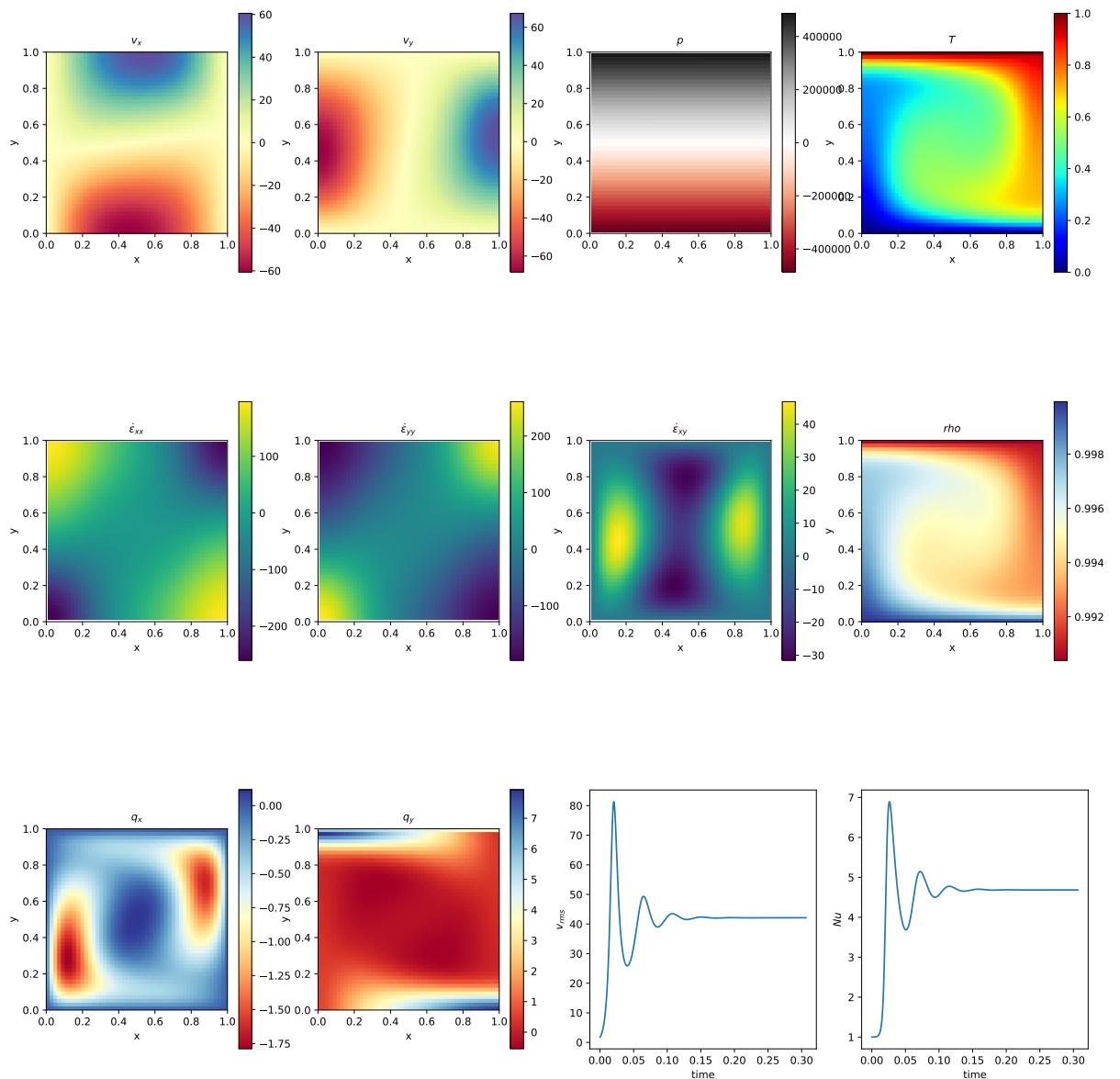
Finally, the steady state root mean square velocity and Nusselt number measurements are indicated in Table ?? alongside those of [8] and [77]. (Note that this benchmark was also carried out and published in other publications [85, 1, 34, 20, 52] but since they did not provide a complete set of measurement values, they are not included in the table.)

		Blankenbach et al	Tackley [77]
$Ra = 10^4$	$V_{rms}$	$42.864947 \pm 0.000020$	42.775
	$Nu$	$4.884409 \pm 0.000010$	4.878
$Ra = 10^5$	$V_{rms}$	$193.21454 \pm 0.00010$	193.11
	$Nu$	$10.534095 \pm 0.000010$	10.531
$Ra = 10^6$	$V_{rms}$	$833.98977 \pm 0.00020$	833.55
	$Nu$	$21.972465 \pm 0.000020$	21.998

Steady state Nusselt number  $Nu$  and  $V_{rms}$  measurements as reported in the literature.

## features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (free-slip)
- Boussinesq approximation
- direct solver
- non-isothermal
- buoyancy-driven flow
- isoviscous
- CFL-condition



ToDo:

- implement steady state criterion
- reach steady state
- do  $\text{Ra}=1\text{e}4, 1\text{e}5, 1\text{e}6$
- plot against blankenbach paper and aspect
- look at critical Ra number

## 10 fieldstone: The lid driven cavity

The lid driven cavity is a famous Computational Fluid Dynamics test case and has been studied in countless publications with a wealth of numerical techniques (see [30] for a succinct review) and also in the laboratory [48].

It models a plane flow of an isothermal isoviscous fluid in a rectangular (usually square) lid-driven cavity. The boundary conditions are indicated in the Fig. ??a. The gravity is set to zero.

### 10.1 the lid driven cavity problem (ldc=0)

In the standard case, the upper side of the cavity moves in its own plane at unit speed, while the other sides are fixed. This thereby introduces a discontinuity in the boundary conditions at the two upper corners of the cavity and yields an uncertainty as to which boundary (side or top) the corner points belong to. In this version of the code the top corner nodes are considered to be part of the lid. If these are excluded the recovered pressure showcases an extremely large checkboard pattern.

This benchmark is usually discussed in the context of low to very high Reynolds number with the full Navier-Stokes equations being solved (with the noticeable exception of [69, 70, 15, 29] which focus on the Stokes equation). In the case of the incompressible Stokes flow, the absence of inertia renders this problem instantaneous so that only one time step is needed.

### 10.2 the lid driven cavity problem - regularisation I (ldc=1)

We avoid the top corner nodes issue altogether by prescribing the horizontal velocity of the lid as follows:

$$u(x) = x^2(1-x)^2. \quad (63)$$

In this case the velocity and its first derivative is continuous at the corners. This is the so-called regularised lid-driven cavity problem [63].

### 10.3 the lid driven cavity problem - regularisation II (ldc=2)

Another regularisation was presented in [22]. Here, a regularized lid driven cavity is studied which is consistent in the sense that  $\nabla \cdot \mathbf{v} = 0$  holds also at the corners of the domain. There are no-slip conditions at the boundaries  $x = 0$ ,  $x = 1$ , and  $y = 0$ .

The velocity at  $y = 1$  is given by

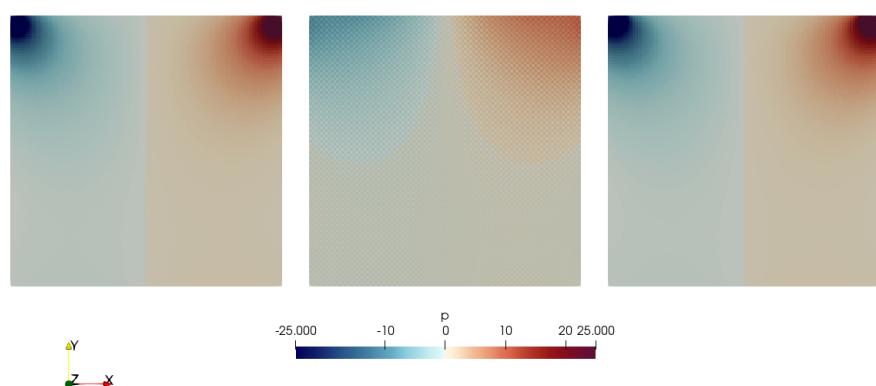
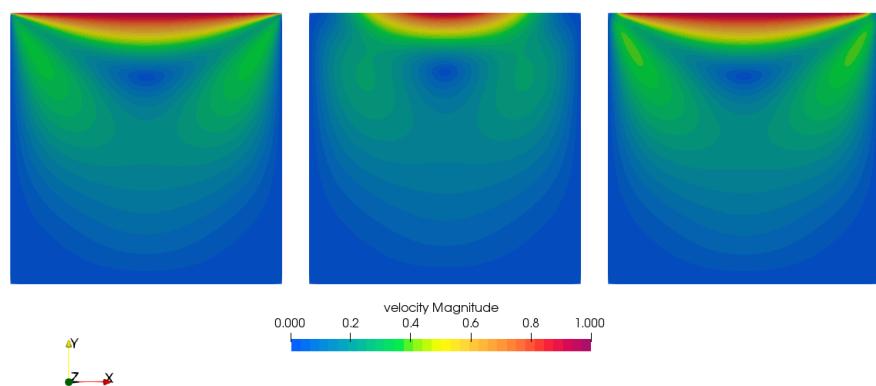
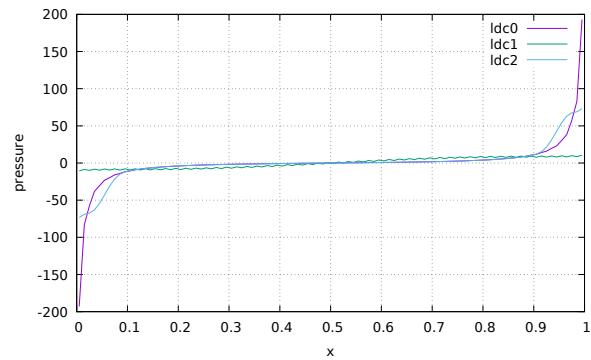
$$\begin{aligned} u(x) &= 1 - \frac{1}{4} \left( 1 - \cos\left(\frac{x_1 - x}{x_1}\pi\right) \right)^2 & x \in [0, x_1] \\ u(x) &= 1 & x \in [x_1, 1 - x_1] \\ u(x) &= 1 - \frac{1}{4} \left( 1 - \cos\left(\frac{x - (1 - x_1)}{x_1}\pi\right) \right)^2 & x \in [1 - x_1, 1] \end{aligned} \quad (64)$$

Results are obtained with  $x_1 = 0.1$ .

#### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- isothermal
- isoviscous

A 100x100 element grid is used. No-slip boundary conditions are prescribed on sides and bottom. A zero vertical velocity is prescribed at the top and the exact form of the prescribed horizontal velocity is controlled by the ldc parameter.



## 11 fieldstone: solcx benchmark

The SolCx benchmark is intended to test the accuracy of the solution to a problem that has a large jump in the viscosity along a line through the domain. Such situations are common in geophysics: for example, the viscosity in a cold, subducting slab is much larger than in the surrounding, relatively hot mantle material.

The SolCx benchmark computes the Stokes flow field of a fluid driven by spatial density variations, subject to a spatially variable viscosity. Specifically, the domain is  $\Omega = [0, 1]^2$ , gravity is  $\mathbf{g} = (0, -1)^T$  and the density is given by

$$\rho(x, y) = \sin(\pi y) \cos(\pi x) \quad (65)$$

Boundary conditions are free slip on all of the sides of the domain and the temperature plays no role in this benchmark. The viscosity is prescribed as follows:

$$\mu(x, y) = \begin{cases} 1 & \text{for } x < 0.5 \\ 10^6 & \text{for } x > 0.5 \end{cases} \quad (66)$$

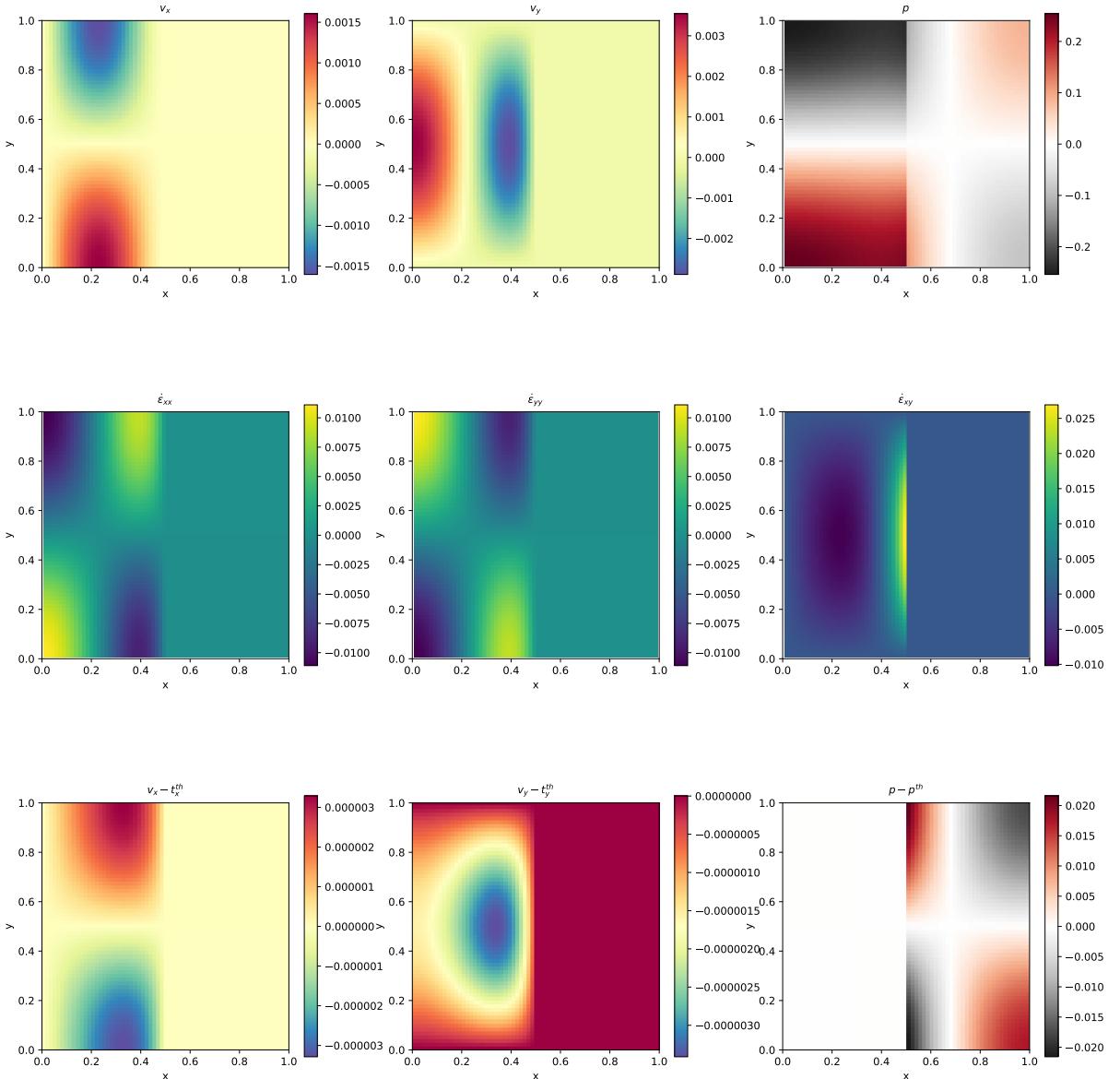
Note the strongly discontinuous viscosity field yields a stagnant flow in the right half of the domain and thereby yields a pressure discontinuity along the interface.

The SolCx benchmark was previously used in [28] (references to earlier uses of the benchmark are available there) and its analytic solution is given in [93]. It has been carried out in [49] and [35]. Note that the source code which evaluates the velocity and pressure fields for both SolCx and SolKz is distributed as part of the open source package Underworld ([59], <http://underworldproject.org>).

In this particular example, the viscosity is computed analytically at the quadrature points (i.e. tracers are not used to attribute a viscosity to the element). If the number of elements is even in any direction, all elements (and their associated quadrature points) have a constant viscosity(1 or  $10^6$ ). If it is odd, then the elements situated at the viscosity jump have half their integration points with  $\mu = 1$  and half with  $\mu = 10^6$  (which is a pathological case since the used quadrature rule inside elements cannot represent accurately such a jump).

### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (free-slip)
- direct solver
- isothermal
- non-isoviscous
- analytical solution



What we learn from this

## 12 fieldstone: solkz benchmark

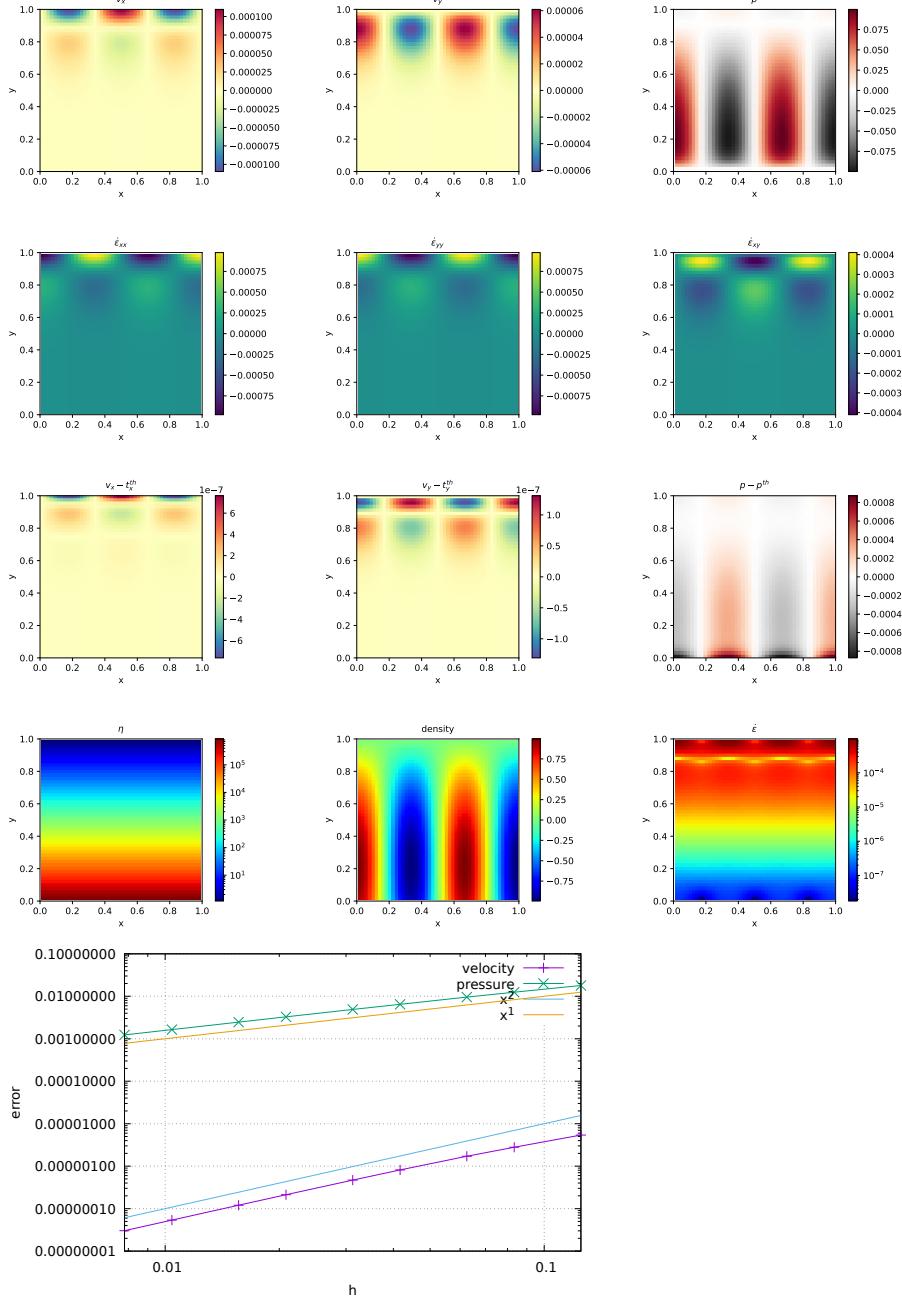
The SolKz benchmark [67] is similar to the SolCx benchmark, but the viscosity is now a function of the space coordinates:

$$\mu(y) = \exp(By) \quad \text{with} \quad B = 13.8155 \quad (67)$$

It is however not a discontinuous function but grows exponentially with the vertical coordinate so that its overall variation is again  $10^6$ . The forcing is again chosen by imposing a spatially variable density variation as follows:

$$\rho(x, y) = \sin(2y) \cos(3\pi x) \quad (68)$$

Free slip boundary conditions are imposed on all sides of the domain. This benchmark is presented in [93] as well and is studied in [28] and [35].



## 13 fieldstone: solvi benchmark

Following SolCx and SolKz, the SolVi inclusion benchmark solves a problem with a discontinuous viscosity field, but in this case the viscosity field is chosen in such a way that the discontinuity is along a circle. Given the regular nature of the grid used by a majority of codes and the present one, this ensures that the discontinuity in the viscosity never aligns to cell boundaries. This in turns leads to almost discontinuous pressures along the interface which are difficult to represent accurately. [73] derived a simple analytic solution for the pressure and velocity fields for a circular inclusion under simple shear and it was used in [23], [76], [28], [49] and [35].

Because of the symmetry of the problem, we only have to solve over the top right quarter of the domain (see Fig. ??a).

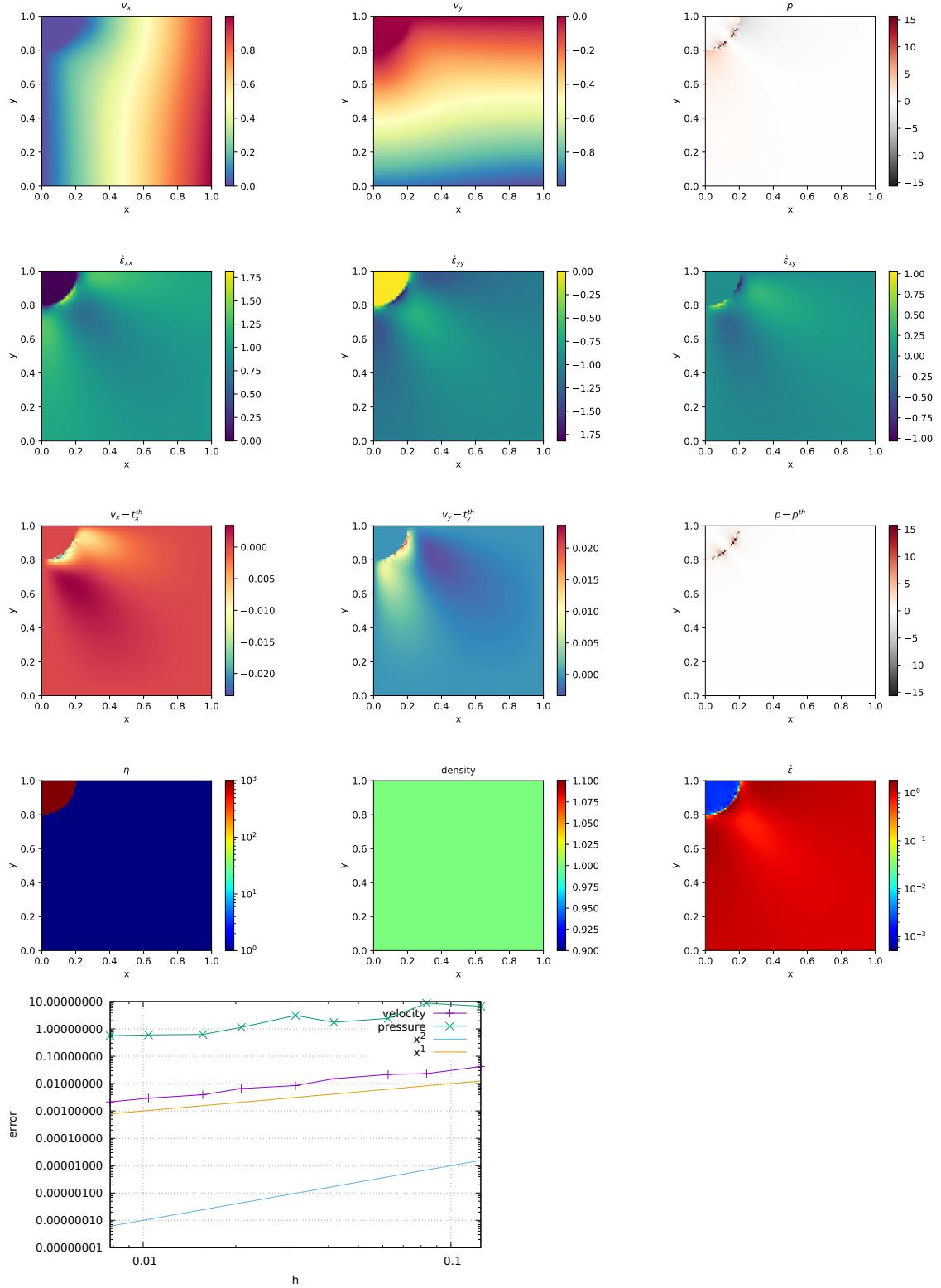
The analytical solution requires a strain rate boundary condition (e.g., pure shear) to be applied far away from the inclusion. In order to avoid using very large domains and/or dealing with this type of boundary condition altogether, the analytical solution is evaluated and imposed on the boundaries of the domain. By doing so, the truncation error introduced while discretizing the strain rate boundary condition is removed.

A characteristic of the analytic solution is that the pressure is zero inside the inclusion, while outside it follows the relation

$$p_m = 4\dot{\epsilon} \frac{\mu_m(\mu_i - \mu_m)}{\mu_i + \mu_m} \frac{r_i^2}{r^2} \cos(2\theta) \quad (69)$$

where  $\mu_i = 10^3$  is the viscosity of the inclusion and  $\mu_m = 1$  is the viscosity of the background media,  $\theta = \tan^{-1}(y/x)$ , and  $\dot{\epsilon} = 1$  is the applied strain rate.

[23] thoroughly investigated this problem with various numerical methods (FEM, FDM), with and without tracers, and conclusively showed how various averagings lead to different results. [28] obtained a first order convergence for both pressure and velocity, while [49] and [35] showed that the use of adaptive mesh refinement in respectively the FEM and FDM yields convergence rates which depend on refinement strategies.



## 14 fieldstone: the indentor benchmark

The punch benchmark is one of the few boundary value problems involving plastic solids for which there exists an exact solution. Such solutions are usually either for highly simplified geometries (spherical or axial symmetry, for instance) or simplified material models (such as rigid plastic solids) [46].

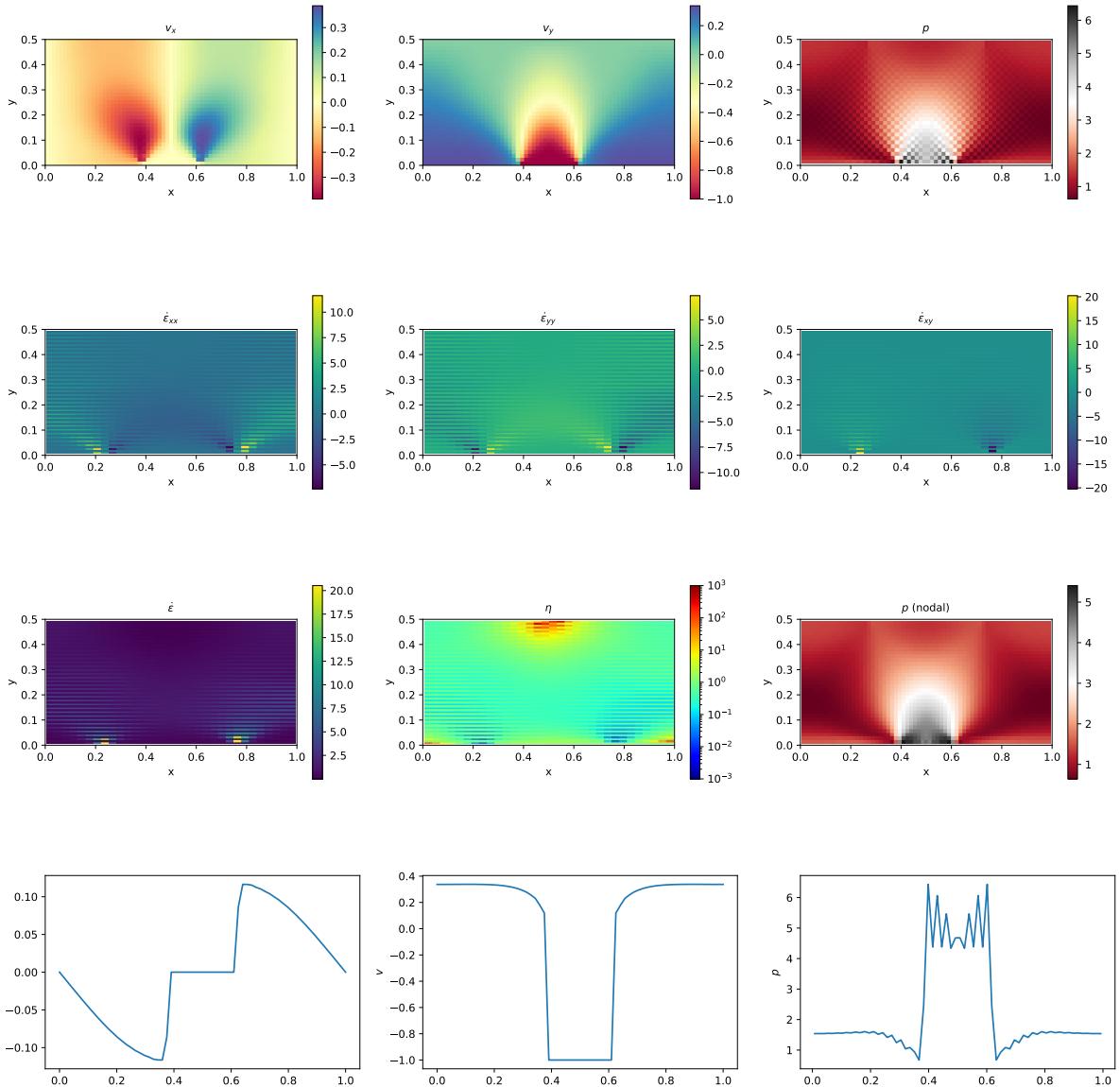
In this experiment, a rigid punch indents a rigid plastic half space; the slip line field theory gives exact solutions as shown in Fig. ??a. The plane strain formulation of the equations and the detailed solution to the problem were derived in the Appendix of [83] and are also presented in [33].

The two dimensional punch problem has been extensively studied numerically for the past 40 years [98, 97, 17, 16, 43, 92, 11, 65] and has been used to draw a parallel with the tectonics of eastern China in the context of the India-Eurasia collision [79, 58]. It is also worth noting that it has been carried out in one form or another in series of analogue modelling articles concerning the same region, with a rigid indenter colliding with a rheologically stratified lithosphere [62, 21, 45].

Numerically, the one-time step punch experiment is performed on a two-dimensional domain of purely plastic von Mises material. Given that the von Mises rheology yield criterion does not depend on pressure, the density of the material and/or the gravity vector is set to zero. Sides are set to free slip boundary conditions, the bottom to no slip, while a vertical velocity  $(0, -v_p)$  is prescribed at the top boundary for nodes whose  $x$  coordinate is within  $[L_x/2 - \delta/2, L_x/2 + \delta/2]$ .

The following parameters are used:  $L_x = 1$ ,  $L_y = 0.5$ ,  $\mu_{min} = 10^{-3}$ ,  $\mu_{max} = 10^3$ ,  $v_p = 1$ ,  $\delta = 0.123456789$  and the yield value of the material is set to  $k = 1$ .

The analytical solution predicts that the angle of the shear bands stemming from the sides of the punch is  $\pi/4$ , that the pressure right under the punch is  $1 + \pi$ , and that the velocity of the rigid blocks on each side of the punch is  $v_p/\sqrt{2}$  (this is simply explained by invoking conservation of mass).



To Do: smooth punch

#### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (no-slip)
- isothermal
- non-isoviscous
- nonlinear rheology

## 15 fieldstone: the annulus benchmark

This benchmark is based on Thieulot & Puckett [Subm.] in which an analytical solution to the isoviscous incompressible Stokes equations is derived in an annulus geometry. The velocity and pressure fields are as follows:

$$v_r(r, \theta) = g(r)k \sin(k\theta), \quad (70)$$

$$v_\theta(r, \theta) = f(r) \cos(k\theta), \quad (71)$$

$$p(r, \theta) = kh(r) \sin(k\theta), \quad (72)$$

$$\rho(r, \theta) = \aleph(r)k \sin(k\theta), \quad (73)$$

with

$$f(r) = Ar + B/r, \quad (74)$$

$$g(r) = \frac{A}{2}r + \frac{B}{r} \ln r + \frac{C}{r}, \quad (75)$$

$$h(r) = \frac{2g(r) - f(r)}{r}, \quad (76)$$

$$\aleph(r) = g'' - \frac{g'}{r} - \frac{g}{r^2}(k^2 - 1) + \frac{f}{r^2} + \frac{f'}{r}, \quad (77)$$

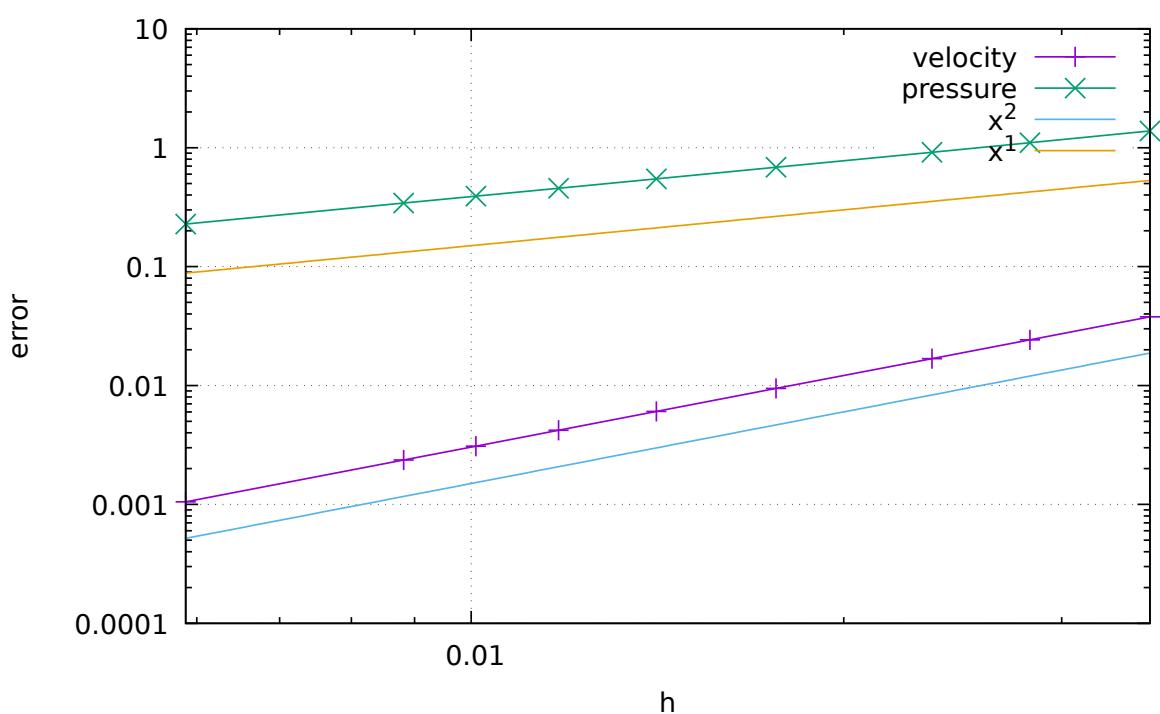
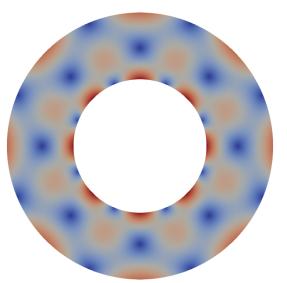
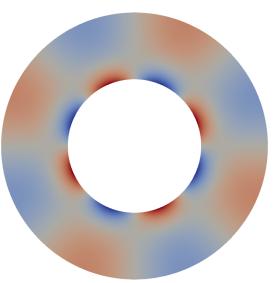
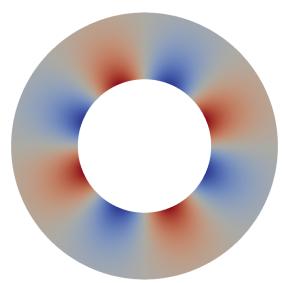
$$A = -C \frac{2(\ln R_1 - \ln R_2)}{R_2^2 \ln R_1 - R_1^2 \ln R_2}, \quad (78)$$

$$B = -C \frac{R_2^2 - R_1^2}{R_2^2 \ln R_1 - R_1^2 \ln R_2}. \quad (79)$$

The parameters  $A$  and  $B$  are chosen so that  $v_r(R_1) = v_r(R_2) = 0$ , i.e. the velocity is tangential to both inner and outer surfaces. The gravity vector is radial and of unit length. In the present case, we set  $R_1 = 1$ ,  $R_2 = 2$  and  $C = -1$ .

### features

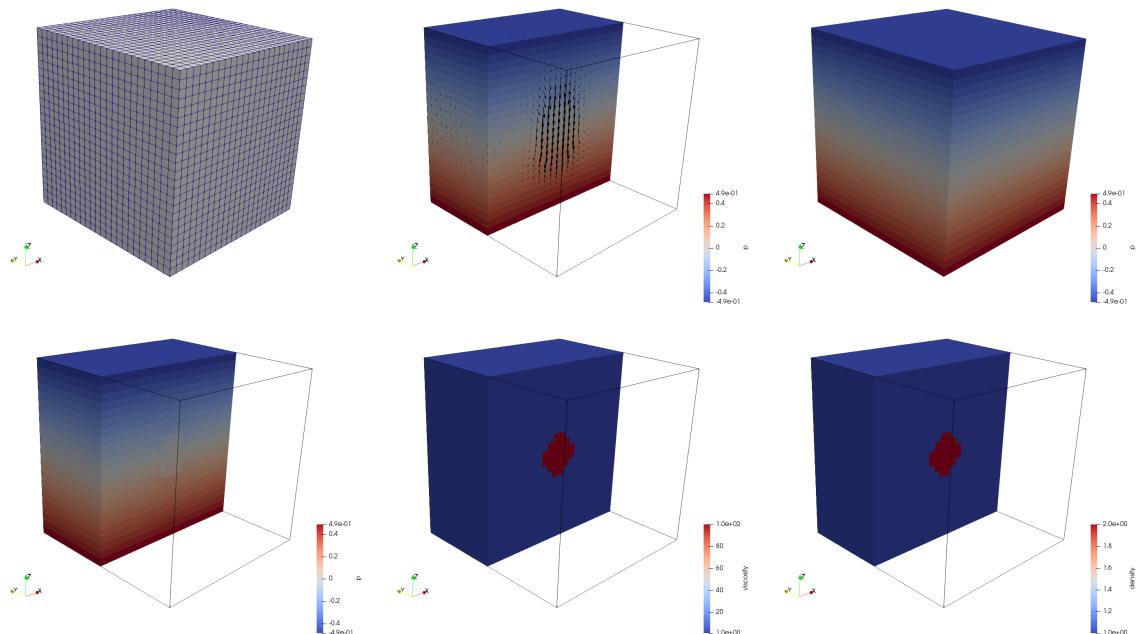
- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions
- direct solver
- isothermal
- isoviscous
- analytical solution
- annulus geometry
- elemental boundary conditions



## 16 fieldstone: stokes sphere (3D) - penalty

### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (free-slip)
- direct solver
- isothermal
- non-isoviscous
- 3D
- elemental b.c.
- buoyancy driven



## 17 fieldstone: stokes sphere (3D) - mixed formulation

This is the same setup as Section 17.

### features

- $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- Dirichlet boundary conditions (free-slip)
- direct solver
- isothermal
- non-isoviscous
- 3D
- elemental b.c.
- buoyancy driven

## 18 fieldstone: consistent pressure recovery

What follows is presented in [96]. The second part of their paper wishes to establish a simple and effective numerical method to calculate variables eliminated by the penalisation process. The method involves an additional finite element solution for the nodal pressures using the same finite element basis and numerical quadrature as used for the velocity.

Let us start with:

$$p = -\lambda \nabla \cdot \mathbf{v}$$

which lead to

$$(q, p) = -\lambda(q, \nabla \cdot \mathbf{v})$$

and then

$$\left( \int \mathbf{N} \mathbf{N} d\Omega \right) \cdot \mathbf{P} = - \left( \lambda \int \mathbf{N} \nabla \mathbf{N} d\Omega \right) \cdot \mathbf{V}$$

or,

$$\mathbf{M} \cdot \mathbf{P} = -\mathbf{D} \cdot \mathbf{V}$$

and finally

$$\mathbf{P} = -\mathbf{M}^{-1} \cdot \mathbf{D} \cdot \mathbf{V}$$

with  $\mathbf{M}$  of size  $(np \times np)$ ,  $\mathbf{D}$  of size  $(np * ndof \times np * ndof)$  and  $\mathbf{V}$  of size  $(np * ndof)$ . The vector  $\mathbf{P}$  contains the  $np$  nodal pressure values directly, with no need for a smoothing scheme. The mass matrix  $\mathbf{M}$  is to be evaluated at the full integration points, while the constraint part (the right hand side of the equation) is to be evaluated at the reduced integration point.

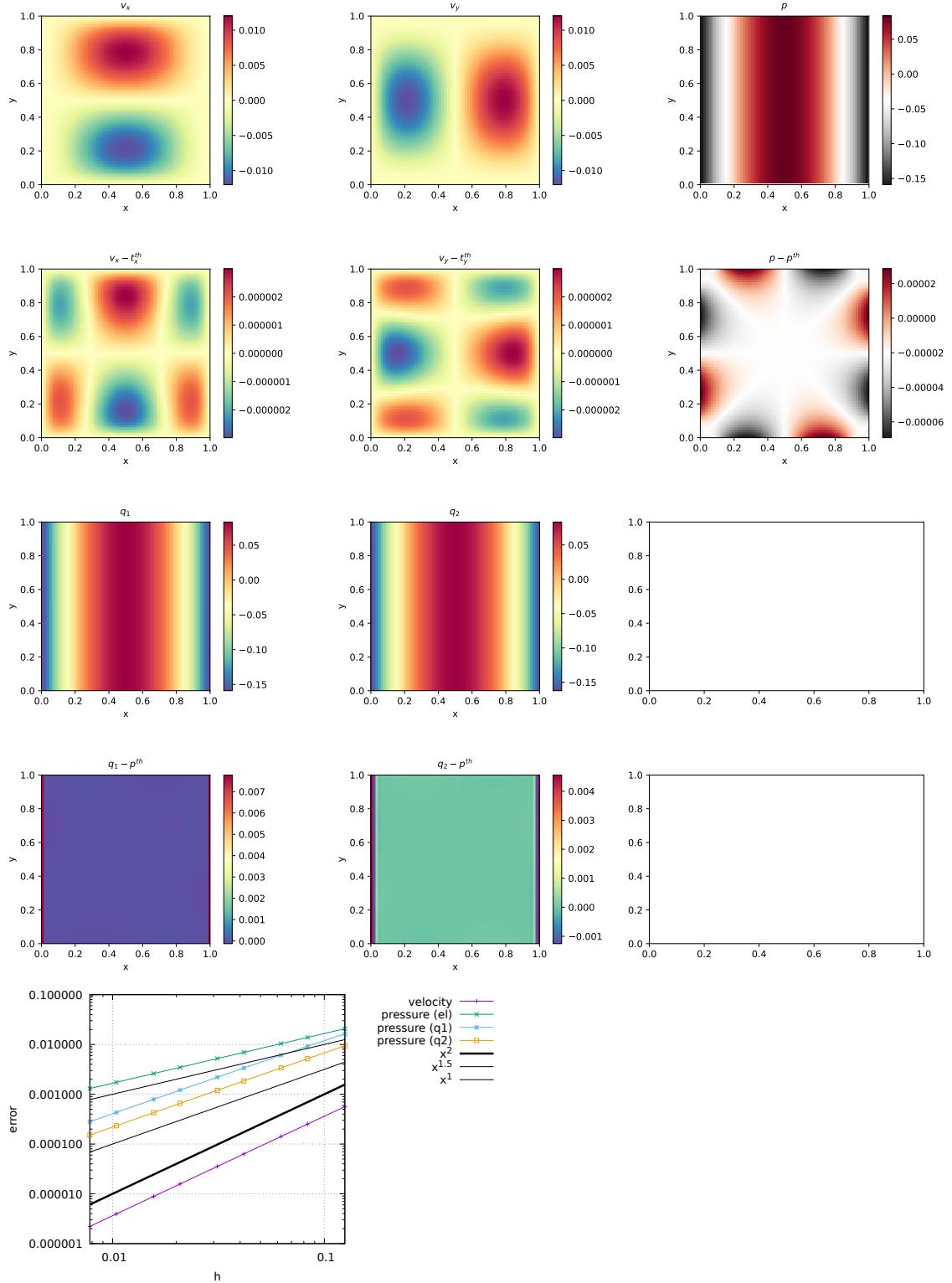
As noted by [96], it is interesting to note that when linear elements are used and the lumped matrices are used for the  $\mathbf{M}$  the resulting algebraic equation is identical to the smoothing scheme based on the averaging method only if the uniform square finite element mesh is used. In this respect this method is expected to yield different results when elements are not square or even rectangular.

---

$q_1$  is smoothed pressure obtained with the center-to-node approach.

$q_2$  is recovered pressure obtained with [96].

All three fulfill the zero average condition:  $\int p d\Omega = 0$ .



In terms of pressure error,  $q_2$  is better than  $q_1$  which is better than elemental.

QUESTION: why are the averages exactly zero ?!

TODO:

- add randomness to internal node positions.
- look at elefant algorithms

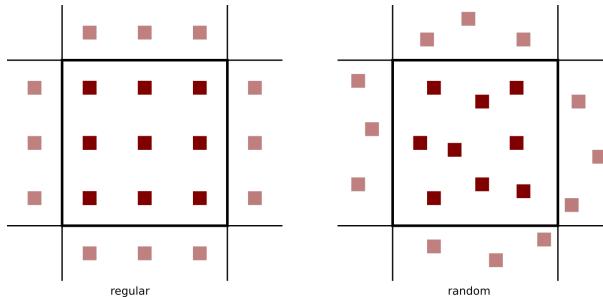
## 19 fieldstone: the Particle in Cell technique (1) - the effect of averaging

### features

- $Q_1 \times P_0$  element
- incompressible flow
- penalty formulation
- Dirichlet boundary conditions (no-slip)
- isothermal
- non-isoviscous
- particle-in-cell

After the initial setup of the grid, markers can then be generated and placed in the domain. One could simply randomly generate the marker positions in the whole domain but unless a *very* large number of markers is used, the chance that an element does not contain any marker exists and this will prove problematic. In order to get a better control over the markers spatial distribution, one usually generates the marker per element, so that the total number of markers in the domain is the product of the number of elements times the user-chosen initial number of markers per element.

Our next concern is how to actually place the markers inside an element. Two methods come to mind: on a regular grid, or in a random manner, as shown on the following figure:



In both cases we make use of the basis shape functions: we generate the positions of the markers (random or regular) in the reference element first ( $r_{im}, s_{im}$ ), and then map those out to the real element as follows:

$$x_{im} = \sum_i^m N_i(r_{im}, s_{im}) x_i \quad y_{im} = \sum_i^m N_i(r_{im}, s_{im}) y_i$$

where  $x_i, y_i$  are the coordinates of the vertices of the element.

When using *active* markers, one is faced with the problem of transferring the properties they carry to the mesh on which the PDEs are to be solved. As we have seen, building the FE matrix involves a loop over all elements, so one simple approach consists of assigning each element a single property computed as the average of the values carried by the markers in that element. Often in colloquial language "average" refers to the arithmetic mean:

$$\langle \phi \rangle_{am} = \frac{1}{n} \sum_k^n \phi_i$$

where  $\langle \phi \rangle_{am}$  is the arithmetic average of the  $n$  numbers  $\phi_i$ . However, in mathematics other means are commonly used, such as the geometric mean:

$$\langle \phi \rangle_{gm} = \left( \prod_i^n \phi_i \right)$$

and the harmonic mean:

$$\langle \phi \rangle_{hm} = \left( \frac{1}{n} \sum_i^n \frac{1}{\phi_i} \right)^{-1}$$

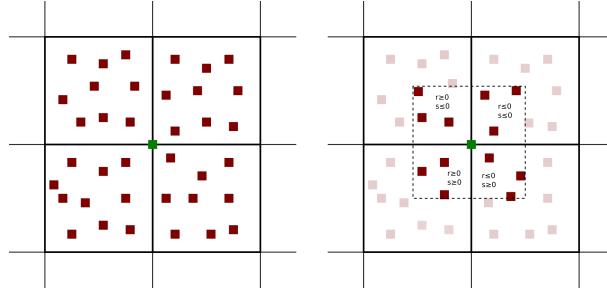
Furthermore, there is a well known inequality for any set of positive numbers,

$$\langle \phi \rangle_{am} \geq \langle \phi \rangle_{gm} \geq \langle \phi \rangle_{hm}$$

which will prove to be important later on.

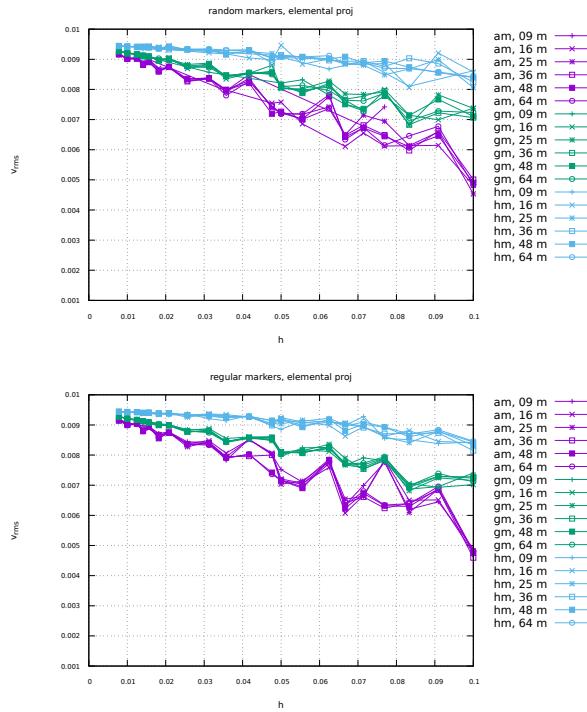
Let us now turn to a simple concrete example: the 2D Stokes sphere. There are two materials in the domain, so that markers carry the label "mat=1" or "mat=2". For each element an average density and viscosity need to be computed. The majority of elements contains markers with a single material label so that the choice of averaging does not matter (it is trivial to verify that if  $\phi_i = \phi_0$  then  $\langle \phi \rangle_{am} = \langle \phi \rangle_{gm} = \langle \phi \rangle_{hm} = \phi_0$ ). Remain the elements crossed by the interface between the two materials: they contain markers of both materials and the average density and viscosity inside those depends on 1) the total number of markers inside the element, 2) the ratio of markers 1 to markers 2, 3) the type of averaging.

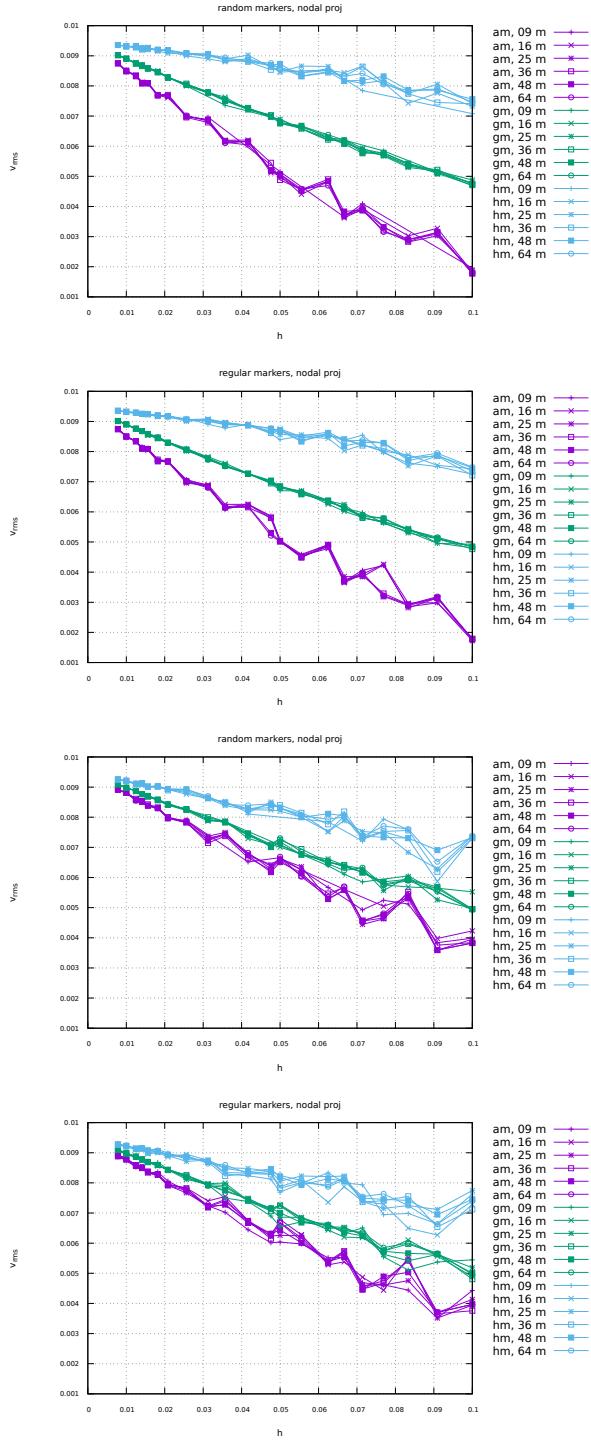
This averaging problem has been studied and documented in the literature [72, 23, 80, 64]



Nodal projection. Left: all markers inside elements to which the green node belongs to are taken into account. Right: only the markers closest to the green node count.

The setup is identical as the Stokes sphere experiment. The bash script `runall` runs the code for many resolutions, both initial marker distribution and all three averaging types. The viscosity of the sphere has been set to  $10^3$  while the viscosity of the surrounding fluid is 1. The average density is always computed with an arithmetic mean. Root mean square velocity results are shown hereunder:



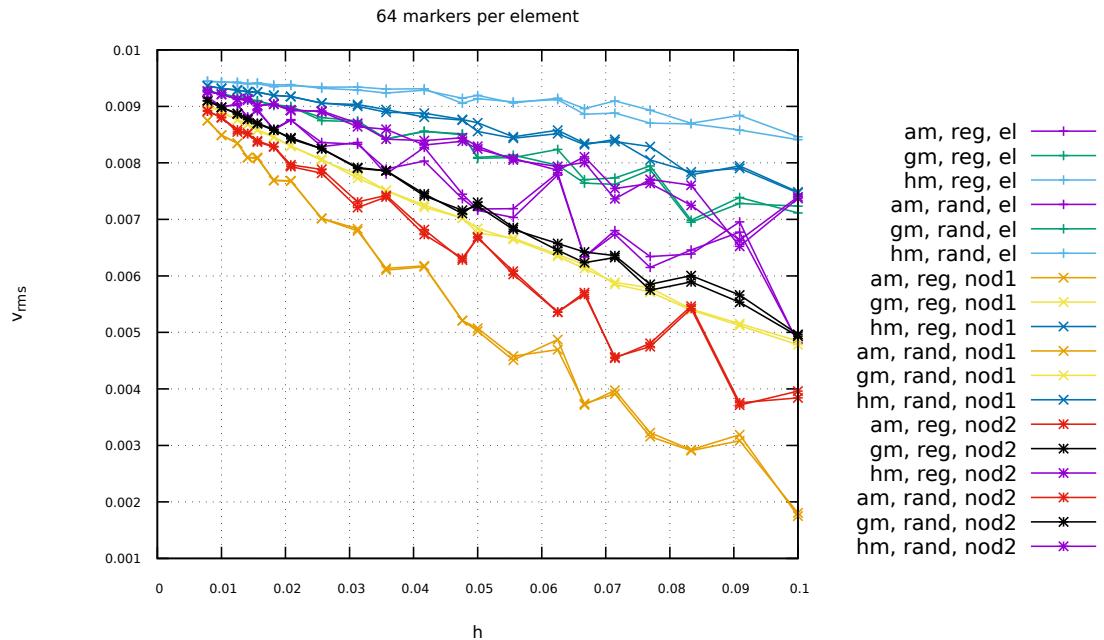


Left column: random markers. Right column: regular markers. Top row: elemental projection. Middle row: nodal 1 projection. Bottom row: nodal 2 projection.

Conclusions:

- With increasing resolution ( $h \rightarrow 0$ ) vrms values seem to converge towards a single value, irrespective of the number of markers.
- At low resolution, say 32x32 (i.e.  $h=0.03125$ ), vrms values for the three averagings differ by about 10%. At higher resolution, say 128x128, vrms values are still not converged.

- The number of markers per element plays a role at low resolution, but less and less with increasing resolution.
- Results for random and regular marker distributions are not identical but follow a similar trend and seem to converge to the same value.
- At low resolutions, elemental values yield better results.
- harmonic mean yields overall the best results



## 20 fieldstone: solving the full saddle point problem

The details of the numerical setup are presented in Section 7.

The main difference is that we no longer use the penalty formulation and therefore keep both velocity and pressure as unknowns. Therefore we end up having to solve the following system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} V \\ P \end{pmatrix} = \begin{pmatrix} f \\ h \end{pmatrix} \quad \text{or,} \quad \mathbb{A} \cdot X = rhs$$

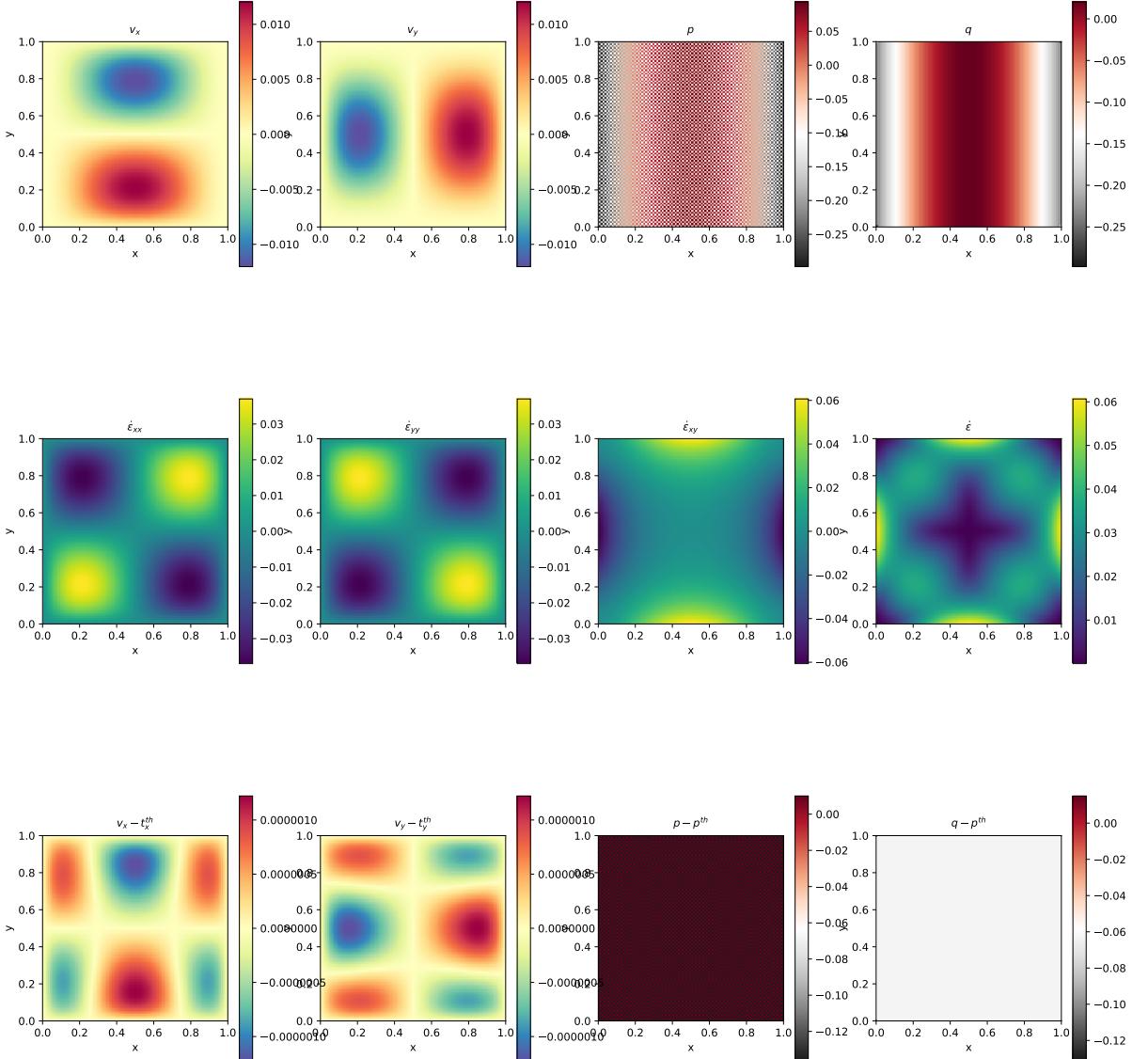
Each block  $\mathbb{K}$ ,  $\mathbb{G}$  and vector  $f$ ,  $h$  are built separately in the code and assembled into the matrix  $\mathbb{A}$  and vector  $rhs$  afterwards.  $\mathbb{A}$  and  $rhs$  are then passed to the solver. We will see later that there are alternatives to solve this approach which do not require to build the full Stokes matrix  $\mathbb{A}$ .

Each element has  $m = 4$  vertices so in total  $ndofV \times m = 8$  velocity dofs and a single pressure dof, commonly situated in the center of the element. The total number of velocity dofs is therefore  $NfemV = nnp \times ndofV$  while the total number of pressure dofs is  $NfemP = nel$ . The total number of dofs is then  $Nfem = NfemV + NfemP$ .

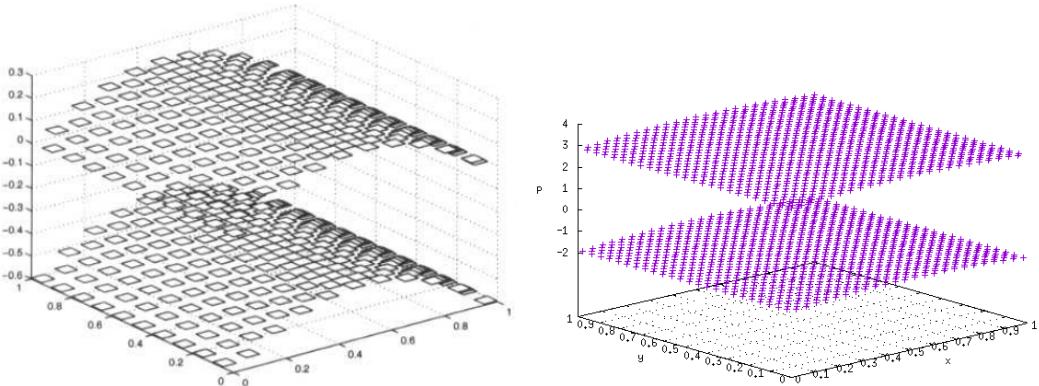
As a consequence, matrix  $\mathbb{K}$  has size  $NfemV, NfemV$  and matrix  $\mathbb{G}$  has size  $NfemV, NfemP$ . Vector  $f$  is of size  $NfemV$  and vector  $h$  is of size  $NfemP$ .

### features

- $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- Dirichlet boundary conditions (no-slip)
- direct solver (?)
- isothermal
- isoviscous
- analytical solution
- pressure smoothing



Unlike the results obtained with the penalty formualtion (see Section 7), the pressure showcases a very strong checkerboard pattern, similar to the one in [26].



Left: pressure solution as shown in [26]; Right: pressure solution obtained with fieldstone.

Rather interestingly, the nodal pressure (obtained with a simple center-to-node algorithm) fails to recover a correct pressure at the four corners.

## 21 fieldstone: solving the full saddle point problem in 3D

When using  $Q_1 \times P_0$  elements, this benchmark fails because of the Dirichlet b.c. on all 6 sides and all three components. However, as we will see, it does work well with  $Q_2 \times Q_1$  elements. .

This benchmark begins by postulating a polynomial solution to the 3D Stokes equation [24]:

$$\mathbf{v} = \begin{pmatrix} x + x^2 + xy + x^3y \\ y + xy + y^2 + x^2y^2 \\ -2z - 3xz - 3yz - 5x^2yz \end{pmatrix} \quad (80)$$

and

$$p = xyz + x^3y^3z - 5/32 \quad (81)$$

While it is then trivial to verify that this velocity field is divergence-free, the corresponding body force of the Stokes equation can be computed by inserting this solution into the momentum equation with a given viscosity  $\mu$  (constant or position/velocity/strain rate dependent). The domain is a unit cube and velocity boundary conditions simply use Eq. (80). Following [13], the viscosity is given by the smoothly varying function

$$\mu = \exp(1 - \beta(x(1-x) + y(1-y) + z(1-z))) \quad (82)$$

One can easily show that the ratio of viscosities  $\mu^*$  in the system follows  $\mu^* = \exp(-3\beta/4)$  so that choosing  $\beta = 10$  yields  $\mu^* \simeq 1808$  and  $\beta = 20$  yields  $\mu^* \simeq 3.269 \times 10^6$ .

We start from the momentum conservation equation:

$$-\nabla p + \nabla \cdot (2\mu \dot{\epsilon}) = \mathbf{f}$$

The  $x$ -component of this equation writes

$$f_x = -\frac{\partial p}{\partial x} + \frac{\partial}{\partial x}(2\mu \dot{\epsilon}_{xx}) + \frac{\partial}{\partial y}(2\mu \dot{\epsilon}_{xy}) + \frac{\partial}{\partial z}(2\mu \dot{\epsilon}_{xz}) \quad (83)$$

$$= -\frac{\partial p}{\partial x} + 2\mu \frac{\partial}{\partial x} \dot{\epsilon}_{xx} + 2\mu \frac{\partial}{\partial y} \dot{\epsilon}_{xy} + 2\mu \frac{\partial}{\partial z} \dot{\epsilon}_{xz} + 2 \frac{\partial \mu}{\partial x} \dot{\epsilon}_{xx} + 2 \frac{\partial \mu}{\partial y} \dot{\epsilon}_{xy} + 2 \frac{\partial \mu}{\partial z} \dot{\epsilon}_{xz} \quad (84)$$

Let us compute all the block separately:

$$\begin{aligned} \dot{\epsilon}_{xx} &= 1 + 2x + y + 3x^2y \\ \dot{\epsilon}_{yy} &= 1 + x + 2y + 2x^2y \\ \dot{\epsilon}_{zz} &= -2 - 3x - 3y - 5x^2y \\ 2\dot{\epsilon}_{xy} &= (x + x^3) + (y + 2xy^2) = x + y + 2xy^2 + x^3 \\ 2\dot{\epsilon}_{xz} &= (0) + (-3z - 10xyz) = -3z - 10xyz \\ 2\dot{\epsilon}_{yz} &= (0) + (-3z - 5x^2z) = -3z - 5x^2z \end{aligned}$$

In passing, one can verify that  $\dot{\epsilon}_{xx} + \dot{\epsilon}_{yy} + \dot{\epsilon}_{zz} = 0$ . We further have

$$\begin{aligned}\frac{\partial}{\partial x} 2\dot{\epsilon}_{xx} &= 2(2 + 6xy) \\ \frac{\partial}{\partial y} 2\dot{\epsilon}_{xy} &= 1 + 4xy \\ \frac{\partial}{\partial z} 2\dot{\epsilon}_{xz} &= -3 - 10xy \\ \frac{\partial}{\partial x} 2\dot{\epsilon}_{xy} &= 1 + 2y^2 + 3x^2 \\ \frac{\partial}{\partial y} 2\dot{\epsilon}_{yy} &= 2(2 + 2x^2) \\ \frac{\partial}{\partial z} 2\dot{\epsilon}_{yz} &= -3 - 5x^2 \\ \frac{\partial}{\partial x} 2\dot{\epsilon}_{xz} &= -10yz \\ \frac{\partial}{\partial y} 2\dot{\epsilon}_{yz} &= 0 \\ \frac{\partial}{\partial z} 2\dot{\epsilon}_{zz} &= 2(0)\end{aligned}$$

$$\frac{\partial p}{\partial x} = yz + 3x^2y^3z \quad (85)$$

$$\frac{\partial p}{\partial y} = xz + 3x^3y^2z \quad (86)$$

$$\frac{\partial p}{\partial z} = xy + x^3y^3 \quad (87)$$

**Pressure normalisation** Here again, because Dirichlet boundary conditions are prescribed on all sides the pressure is known up to an arbitrary constant. This constant can be determined by (arbitrarily) choosing to normalised the pressure field as follows:

$$\int_{\Omega} p \, d\Omega = 0 \quad (88)$$

This is a single constraint associated to a single Lagrange multiplier  $\lambda$  and the global Stokes system takes the form

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} & 0 \\ \mathbb{G}^T & 0 & \mathcal{C} \\ 0 & \mathcal{C}^T & 0 \end{pmatrix} \begin{pmatrix} V \\ P \\ \lambda \end{pmatrix}$$

In this particular case the constraint matrix  $\mathcal{C}$  is a vector and it only acts on the pressure degrees of freedom because of Eq.(88). Its exact expression is as follows:

$$\int_{\Omega} p \, d\Omega = \sum_e \int_{\Omega_e} p \, d\Omega = \sum_e \int_{\Omega_e} \sum_i N_i^p p_i \, d\Omega = \sum_e \sum_i \left( \int_{\Omega_e} N_i^p \, d\Omega \right) p_i = \sum_e \mathcal{C}_e \cdot \mathbf{p}_e$$

where  $\mathbf{p}_e$  is the list of pressure dofs of element  $e$ . The elemental constraint vector contains the corresponding pressure basis functions integrated over the element. These elemental constraints are then assembled into the vector  $\mathcal{C}$ .

### 21.0.1 Constant viscosity

Choosing  $\beta = 0$  yields a constant velocity  $\mu(x, y, z) = \exp(1) \simeq 2.718$  (and greatly simplifies the right-hand side) so that

$$\frac{\partial}{\partial x} \mu(x, y, z) = 0 \quad (89)$$

$$\frac{\partial}{\partial y} \mu(x, y, z) = 0 \quad (90)$$

$$\frac{\partial}{\partial z} \mu(x, y, z) = 0 \quad (91)$$

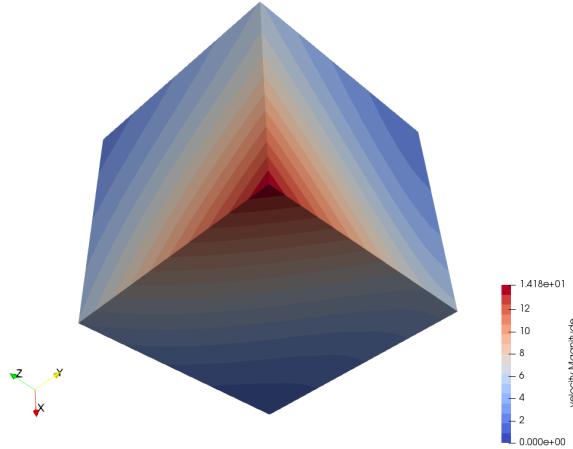
and

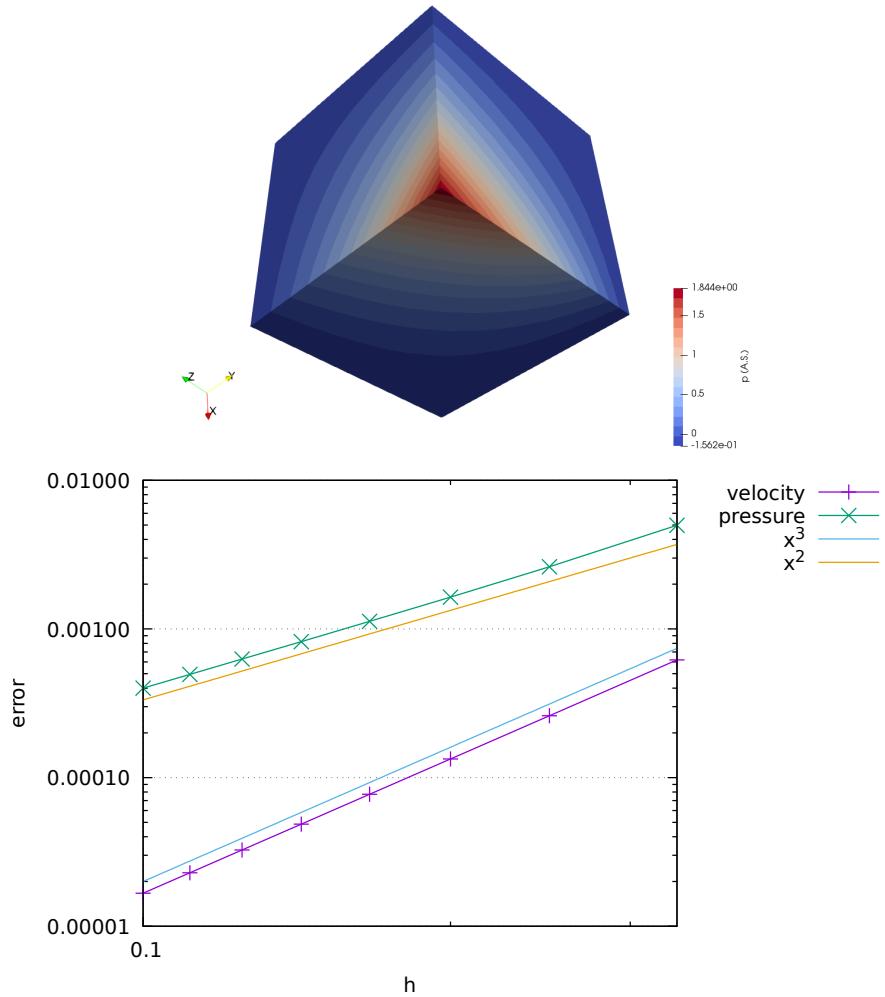
$$\begin{aligned} f_x &= -\frac{\partial p}{\partial x} + 2\mu \frac{\partial}{\partial x} \dot{\epsilon}_{xx} + 2\mu \frac{\partial}{\partial y} \dot{\epsilon}_{xy} + 2\mu \frac{\partial}{\partial z} \dot{\epsilon}_{xz} \\ &= -(yz + 3x^2y^3z) + 2(2 + 6xy) + (1 + 4xy) + (-3 - 10xy) \\ &= -(yz + 3x^2y^3z) + \mu(2 + 6xy) \\ f_y &= -\frac{\partial p}{\partial y} + 2\mu \frac{\partial}{\partial x} \dot{\epsilon}_{xy} + 2\mu \frac{\partial}{\partial y} \dot{\epsilon}_{yy} + 2\mu \frac{\partial}{\partial z} \dot{\epsilon}_{yz} \\ &= -(xz + 3x^3y^2z) + \mu(1 + 2y^2 + 3x^2) + \mu(2(2 + 2x^2) + \mu(-3 - 5x^2)) \\ &= -(xz + 3x^3y^2z) + \mu(2 + 2x^2 + 2y^2) \\ f_z &= -\frac{\partial p}{\partial z} + 2\mu \frac{\partial}{\partial x} \dot{\epsilon}_{xz} + 2\mu \frac{\partial}{\partial y} \dot{\epsilon}_{yz} + 2\mu \frac{\partial}{\partial z} \dot{\epsilon}_{zz} \\ &= -(xy + x^3y^3) + \mu(-10yz) + 0 + 0 \\ &= -(xy + x^3y^3) + \mu(-10yz) \end{aligned}$$

Finally

$$\mathbf{f} = - \begin{pmatrix} yz + 3x^2y^3z \\ xz + 3x^3y^2z \\ xy + x^3y^3 \end{pmatrix} + \mu \begin{pmatrix} 2 + 6xy \\ 2 + 2x^2 + 2y^2 \\ -10yz \end{pmatrix}$$

Note that there seems to be a sign problem with Eq.(26) in [13].





### 21.0.2 Variable viscosity

The spatial derivatives of the viscosity are then given by

$$\begin{aligned}
 \frac{\partial}{\partial x} \mu(x, y, z) &= -(1 - 2x)\beta\mu(x, y, z) \\
 \frac{\partial}{\partial y} \mu(x, y, z) &= -(1 - 2y)\beta\mu(x, y, z) \\
 \frac{\partial}{\partial z} \mu(x, y, z) &= -(1 - 2z)\beta\mu(x, y, z)
 \end{aligned}$$

and the right-hand side by

$$\begin{aligned}
\mathbf{f} &= - \begin{pmatrix} yz + 3x^2y^3z \\ xz + 3x^3y^2z \\ xy + x^3y^3 \end{pmatrix} + \mu \begin{pmatrix} 2 + 6xy \\ 2 + 2x^2 + 2y^2 \\ -10yz \end{pmatrix} \\
&\quad - (1 - 2x)\beta\mu(x, y, z) \begin{pmatrix} 2\dot{\epsilon}_{xx} \\ 2\dot{\epsilon}_{xy} \\ 2\dot{\epsilon}_{xz} \end{pmatrix} - (1 - 2y)\beta\mu(x, y, z) \begin{pmatrix} 2\dot{\epsilon}_{xy} \\ 2\dot{\epsilon}_{yy} \\ 2\dot{\epsilon}_{yz} \end{pmatrix} - (1 - 2z)\beta\mu(x, y, z) \begin{pmatrix} 2\dot{\epsilon}_{xz} \\ 2\dot{\epsilon}_{yz} \\ 2\dot{\epsilon}_{zz} \end{pmatrix} \\
&= - \begin{pmatrix} yz + 3x^2y^3z \\ xz + 3x^3y^2z \\ xy + x^3y^3 \end{pmatrix} + \mu \begin{pmatrix} 2 + 6xy \\ 2 + 2x^2 + 2y^2 \\ -10yz \end{pmatrix} \\
&\quad - (1 - 2x)\beta\mu \begin{pmatrix} 2 + 4x + 2y + 6x^2y \\ x + y + 2xy^2 + x^3 \\ -3z - 10xyz \end{pmatrix} - (1 - 2y)\beta\mu \begin{pmatrix} x + y + 2xy^2 + x^3 \\ 2 + 2x + 4y + 4x^2y \\ -3z - 5x^2z \end{pmatrix} - (1 - 2z)\beta\mu \begin{pmatrix} -3z - 10xyz \\ -3z - 5x^2z \\ -4 - 6x - 6y - 10xz \end{pmatrix}
\end{aligned}$$

Note that at  $(x, y, z) = (0, 0, 0)$ ,  $\mu = \exp(1)$ , and at  $(x, y, z) = (0.5, 0.5, 0.5)$ ,  $\mu = \exp(1 - 3\beta/4)$  so that the maximum viscosity ratio is given by

$$\mu^* = \frac{\exp(1 - 3\beta/4)}{\exp(1)} = \exp(-3\beta/4)$$

By varying  $\beta$  between 1 and 22 we can get up to 7 orders of magnitude viscosity difference.

#### features

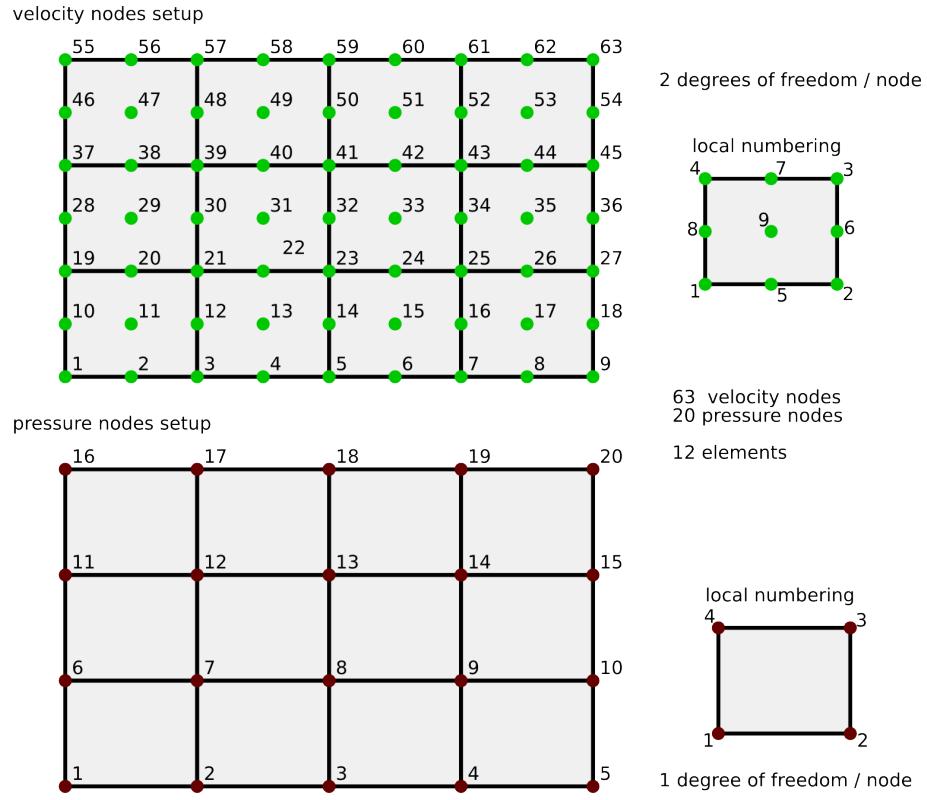
- $Q_1 \times P_0$  element
- incompressible flow
- saddle point system
- Dirichlet boundary conditions (free-slip)
- direct solver
- isothermal
- non-isoviscous
- 3D
- elemental b.c.
- analytical solution

## 22 fieldstone: solving the full saddle point problem with $Q_2 \times Q_1$ elements

The details of the numerical setup are presented in Section 7.

Each element has  $m_V = 9$  vertices so in total  $ndof_V \times m_V = 18$  velocity dofs and  $ndof_P \times m_P = 4$  pressure dofs. The total number of velocity dofs is therefore  $NfemV = nnp \times ndofV$  while the total number of pressure dofs is  $NfemP = nel$ . The total number of dofs is then  $Nfem = NfemV + NfemP$ .

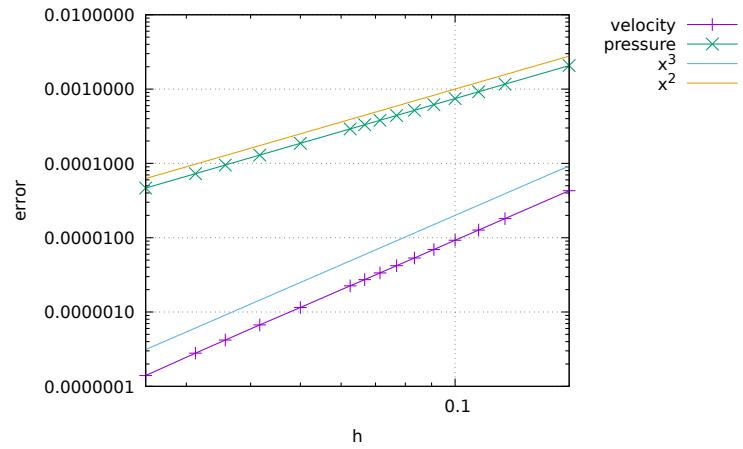
As a consequence, matrix  $\mathbb{K}$  has size  $NfemV, NfemV$  and matrix  $\mathbb{G}$  has size  $NfemV, NfemP$ . Vector  $f$  is of size  $NfemV$  and vector  $h$  is of size  $NfemP$ .



renumber all nodes to start at zero!! Also internal numbering does not work here

### features

- $Q_2 \times Q_1$  element
- incompressible flow
- mixed formulation
- Dirichlet boundary conditions (no-slip)
- isothermal
- isoviscous
- analytical solution



## 23 fieldstone: solving the full saddle point problem with $Q_3 \times Q_2$ elements

The details of the numerical setup are presented in Section 7.

Each element has  $m_V = 16$  vertices so in total  $ndof_V \times m_V = 32$  velocity dofs and  $ndof_P \times m_P = 9$  pressure dofs. The total number of velocity dofs is therefore  $NfemV = nnp \times ndofV$  while the total number of pressure dofs is  $NfemP = nel$ . The total number of dofs is then  $Nfem = NfemV + NfemP$ .

As a consequence, matrix  $\mathbb{K}$  has size  $NfemV, NfemV$  and matrix  $\mathbb{G}$  has size  $NfemV, NfemP$ . Vector  $f$  is of size  $NfemV$  and vector  $h$  is of size  $NfemP$ .

```

60====61====62====63====64====65====66====67====68====70
||          ||          ||          ||
50    51    52    53    54    55    56    57    58    59
||          ||          ||          ||
40    41    42    43    44    45    46    47    48    49
||          ||          ||          ||
30====31====32====33====34====35====36====37====38====39
||          ||          ||          ||
20    21    22    23    24    25    26    27    28    29
||          ||          ||          ||
10    11    12    13    14    15    16    17    18    19
||          ||          ||          ||
00====01====02====03====04====05====06====07====08====09

```

Example of 3x2 mesh.  $nmx=10$ ,  $nny=7$ ,  $nnp=70$ ,  $nelx=3$ ,  $nely=2$ ,  $nel=6$

```

12====13====14====15          06=====07=====08
||  ||  ||  ||          ||  ||  ||  ||
08====09====10====11          ||  ||  ||  ||
||  ||  ||  ||          03=====04=====05
04====05====06====07          ||  ||  ||  ||
||  ||  ||  ||          00=====01=====02
00====01====02====03

```

Velocity (Q3)

```

(r,s)_{00}=(-1,-1)          (r,s)_{00}=(-1,-1)
(r,s)_{01}=(-1/3,-1)         (r,s)_{01}=(0,-1)
(r,s)_{02}=(+1/3,-1)         (r,s)_{02}=(+1,-1)
(r,s)_{03}=(+1,-1)           (r,s)_{03}=(-1,0)
(r,s)_{04}=(-1,-1/3)         (r,s)_{04}=(0,0)
(r,s)_{05}=(-1/3,-1/3)       (r,s)_{05}=(+1,0)
(r,s)_{06}=(+1/3,-1/3)       (r,s)_{06}=(-1,+1)
(r,s)_{07}=(+1,-1/3)          (r,s)_{07}=(0,+1)
(r,s)_{08}=(-1,+1/3)          (r,s)_{08}=(+1,+1)
(r,s)_{09}=(-1/3,+1/3)
(r,s)_{10}=(+1/3,+1/3)
(r,s)_{11}=(+1,+1/3)
(r,s)_{12}=(-1,+1)
(r,s)_{13}=(-1/3,+1)
(r,s)_{14}=(+1/3,+1)
(r,s)_{15}=(+1,+1)

```

Pressure (Q2)

```

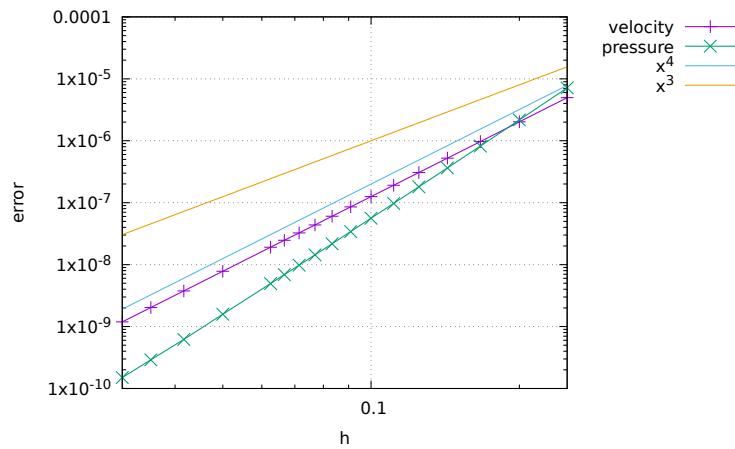
(r,s)_{00}=(-1,-1)
(r,s)_{01}=(0,-1)
(r,s)_{02}=(+1,-1)
(r,s)_{03}=(-1,0)
(r,s)_{04}=(0,0)
(r,s)_{05}=(+1,0)
(r,s)_{06}=(-1,+1)
(r,s)_{07}=(0,+1)
(r,s)_{08}=(+1,+1)

```

Write about 4 point quadrature.

### features

- $Q_3 \times Q_2$  element
- incompressible flow
- mixed formulation
- isothermal
- isoviscous
- analytical solution



velocity error rate is cubic, pressure superconvergent since the pressure field is quadratic and therefore lies into the  $Q_2$  space.

## 24 fieldstone: the Busse benchmark

This three-dimensional benchmark was first proposed by [14]. It has been subsequently presented in [77, 85, 1, 61, 20, 49]. We here focus on Case 1 of [14]: an isoviscous bimodal convection experiment at  $Ra = 3 \times 10^5$ .

The domain is of size  $a \times b \times h$  with  $a = 1.0079h$ ,  $b = 0.6283h$  with  $h = 2700\text{km}$ . It is filled with a Newtonian fluid characterised by  $\rho_0 = 3300\text{kg.m}^{-3}$ ,  $\alpha = 10^{-5}\text{K}^{-1}$ ,  $\mu = 8.0198 \times 10^{23}\text{Pa.s}$ ,  $k = 3.564\text{W.m}^{-1}\text{.K}^{-1}$ ,  $c_p = 1080\text{J.K}^{-1}\text{.kg}^{-1}$ . The gravity vector is set to  $\mathbf{g} = (0, 0, -10)^T$ . The temperature is imposed at the bottom ( $T = 3700^\circ\text{C}$ ) and at the top ( $T = 0^\circ\text{C}$ ).

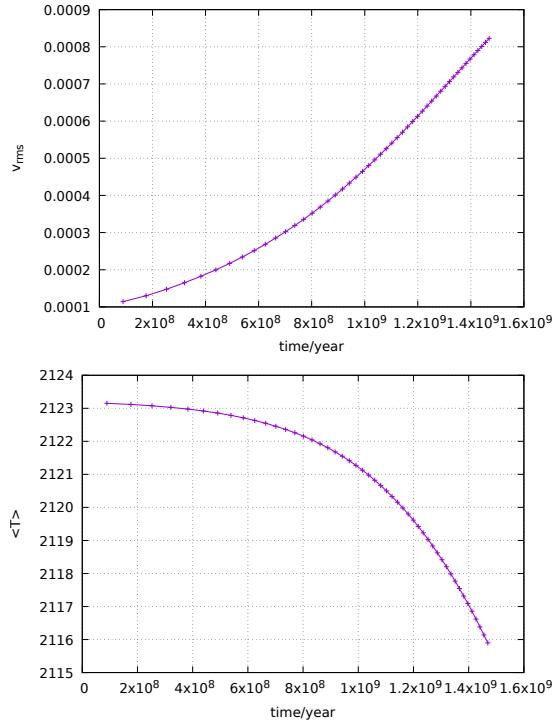
The various measurements presented in [14] are listed hereafter:

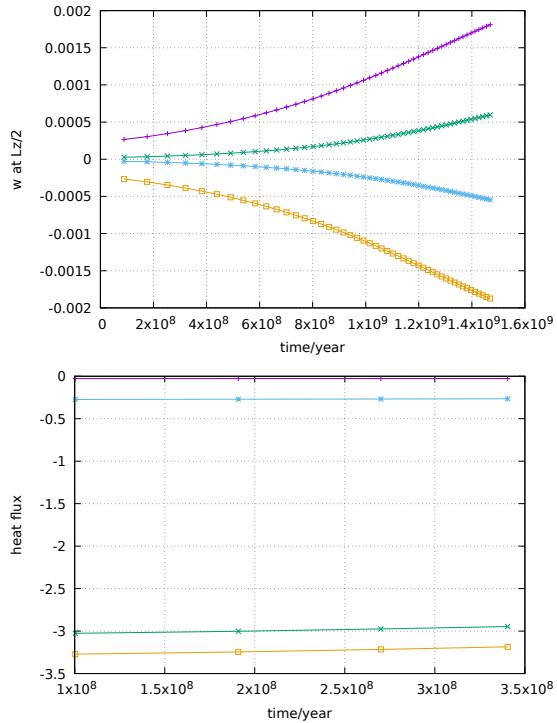
- The Nusselt number  $Nu$  computed at the top surface following Eq. (62):

$$Nu = L_z \frac{\int \int_{z=L_z} \frac{\partial T}{\partial y} dx dy}{\int \int_{z=0} T dx dy}$$

- the root mean square velocity  $v_{rms}$  and the temperature mean square velocity  $T_{rms}$
- The vertical velocity  $w$  and temperature  $T$  at points  $\mathbf{x}_1 = (0, 0, L_z/2)$ ,  $\mathbf{x}_2 = (L_x, 0, L_z/2)$ ,  $\mathbf{x}_3 = (0, L_y, L_z/2)$  and  $\mathbf{x}_4 = (L_x, L_y, L_z/2)$ ;
- the vertical component of the heat flux  $Q$  at the top surface at all four corners.

The values plotted hereunder are adimensionalised by means of a reference temperature (3700K), a reference lengthscale 2700km, and a reference time  $L_z^2/\kappa$ .





### features

- $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- Dirichlet boundary conditions (free-slip)
- direct solver
- isothermal
- non-isoviscous
- 3D
- elemental b.c.
- buoyancy driven

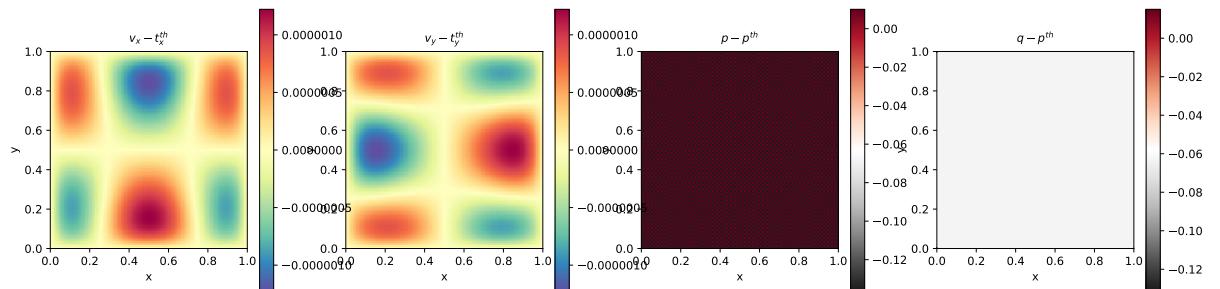
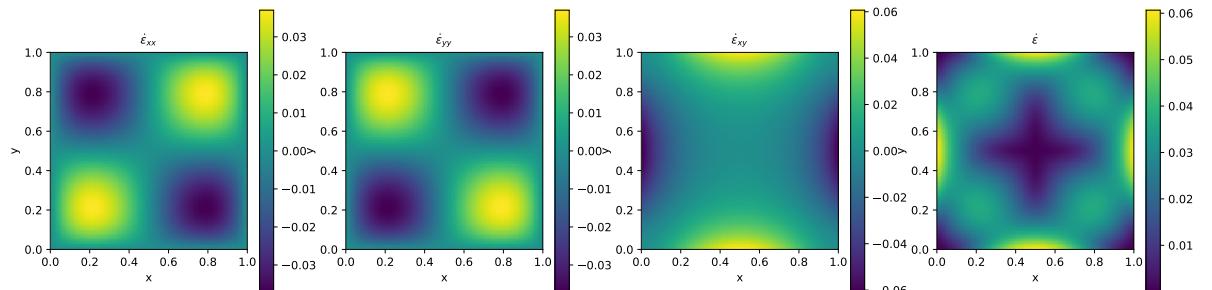
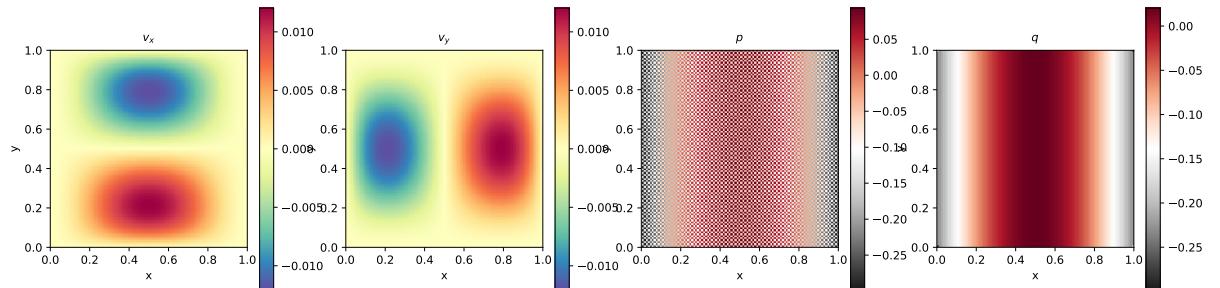
**ToDo:** look at energy conservation. run to steady state and make sure the expected values are retrieved.

## 25 fieldstone: The non-conforming $Q_1 \times P_0$ element

### features

- Non-conforming  $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- isothermal
- non-isoviscous
- analytical solution
- pressure smoothing

try Q1 mapping instead of isoparametric.



## **26 fieldstone: The stabilised $Q_1 \times Q_1$ element**

The details of the numerical setup are presented in Section 7.

## 27 fieldstone: compressible flow (1)

This work is part of the MSc thesis of T. Weir (2018).

We first start with an isothermal Stokes flow, so that we disregard the heat transport equation and the equations we wish to solve are simply:

$$-\nabla \cdot \left[ 2\eta \left( \dot{\epsilon}(\mathbf{v}) - \frac{1}{3}(\nabla \cdot \mathbf{v})\mathbf{1} \right) \right] + \nabla p = \rho \mathbf{g} \quad \text{in } \Omega, \quad (93)$$

$$\nabla \cdot (\rho \mathbf{v}) = 0 \quad \text{in } \Omega \quad (94)$$

The second equation can be rewritten  $\nabla \cdot (\rho \mathbf{v}) = \rho \nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla \rho = 0$  or,

$$\nabla \cdot \mathbf{v} + \frac{1}{\rho} \mathbf{v} \cdot \nabla \rho = 0$$

Note that this presupposes that the density is not zero anywhere in the domain.

We use a mixed formulation and therefore keep both velocity and pressure as unknowns. We end up having to solve the following system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} \\ \mathbb{G}^T + \mathbb{Z} & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} f \\ h \end{pmatrix} \quad \text{or,} \quad \mathbb{A} \cdot X = rhs$$

Where  $\mathbb{K}$  is the stiffness matrix,  $\mathbb{G}$  is the discrete gradient operator,  $\mathbb{G}^T$  is the discrete divergence operator,  $\mathcal{V}$  the velocity vector,  $\mathcal{P}$  the pressure vector. Note that the term  $\mathbb{Z}\mathcal{V}$  derives from term  $\mathbf{v} \cdot \nabla \rho$  in the continuity equation.

Each block  $\mathbb{K}$ ,  $\mathbb{G}$ ,  $\mathbb{Z}$  and vectors  $f$  and  $h$  are built separately in the code and assembled into the matrix  $\mathbb{A}$  and vector  $rhs$  afterwards.  $\mathbb{A}$  and  $rhs$  are then passed to the solver. We will see later that there are alternatives to solve this approach which do not require to build the full Stokes matrix  $\mathbb{A}$ .

*Remark:* the term  $\mathbb{Z}\mathcal{V}$  is often put in the rhs (i.e. added to  $h$ ) so that the matrix  $\mathbb{A}$  retains the same structure as in the incompressible case. This is indeed how it is implemented in ASPECT. This however requires more work since the rhs depends on the solution and some form of iterations is needed.

In the case of a compressible flow the strain rate tensor and the deviatoric strain rate tensor are no more equal (since  $\nabla \cdot \mathbf{v} \neq 0$ ). The deviatoric strainrate tensor is given by<sup>8</sup>

$$\dot{\epsilon}^d(\mathbf{v}) = \dot{\epsilon}(\mathbf{v}) - \frac{1}{3} Tr(\dot{\epsilon})\mathbf{1} = \dot{\epsilon}(\mathbf{v}) - \frac{1}{3}(\nabla \cdot \mathbf{v})\mathbf{1}$$

In that case:

$$\dot{\epsilon}_{xx}^d = \frac{\partial u}{\partial x} - \frac{1}{3} \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = \frac{2}{3} \frac{\partial u}{\partial x} - \frac{1}{3} \frac{\partial v}{\partial y} \quad (95)$$

$$\dot{\epsilon}_{yy}^d = \frac{\partial v}{\partial y} - \frac{1}{3} \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = -\frac{1}{3} \frac{\partial u}{\partial x} + \frac{2}{3} \frac{\partial v}{\partial y} \quad (96)$$

$$2\dot{\epsilon}_{xy}^d = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \quad (97)$$

and then

$$\dot{\epsilon}^d(\mathbf{v}) = \begin{pmatrix} \frac{2}{3} \frac{\partial u}{\partial x} - \frac{1}{3} \frac{\partial v}{\partial y} & \frac{1}{2} \frac{\partial u}{\partial y} + \frac{1}{2} \frac{\partial v}{\partial x} \\ \frac{1}{2} \frac{\partial u}{\partial y} + \frac{1}{2} \frac{\partial v}{\partial x} & -\frac{1}{3} \frac{\partial u}{\partial x} + \frac{2}{3} \frac{\partial v}{\partial y} \end{pmatrix}$$

From  $\vec{\tau} = 2\eta \vec{\epsilon}^d$  we arrive at:

$$\begin{pmatrix} \tau_{xx} \\ \tau_{yy} \\ \tau_{xy} \end{pmatrix} = 2\eta \begin{pmatrix} \dot{\epsilon}_{xx}^d \\ \dot{\epsilon}_{yy}^d \\ \dot{\epsilon}_{xy}^d \end{pmatrix} = 2\eta \begin{pmatrix} 2/3 & -1/3 & 0 \\ -1/3 & 2/3 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{pmatrix} = \eta \begin{pmatrix} 4/3 & -2/3 & 0 \\ -2/3 & 4/3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{pmatrix}$$

or,

$$\vec{\tau} = \mathbf{C}_\eta \mathbf{B} \mathbf{V}$$

---

<sup>8</sup>See the ASPECT manual for a justification of the 3 value in the denominator in 2D and 3D.

In order to test our implementation we have created a few manufactured solutions:

- benchmark #1 (ibench=1): Starting from a density profile of:

$$\rho(x, y) = xy \quad (98)$$

We derive a velocity given by:

$$v_x(x, y) = \frac{C_x}{x}, v_y(x, y) = \frac{C_y}{y} \quad (99)$$

With  $g_x(x, y) = \frac{1}{x}$  and  $g_y(x, y) = \frac{1}{y}$ , this leads us to a pressure profile:

$$p = -\eta \left( \frac{4C_x}{3x^2} + \frac{4C_y}{3y^2} \right) + xy + C_0 \quad (100)$$

This gives us a strain rate:

$$\dot{\epsilon}_{xx} = \frac{-C_x}{x^2} \quad \dot{\epsilon}_{yy} = \frac{-C_y}{y^2} \quad \dot{\epsilon}_{xy} = 0$$

In what follows, we choose  $\eta = 1$  and  $C_x = C_y = 1$  and for a unit square domain  $[1 : 2] \times [1 : 2]$  we compute  $C_0$  so that the pressure is normalised to zero over the whole domain and obtain  $C_0 = -1$ .

- benchmark #2 (ibench=2): Starting from a density profile of:

$$\rho = \cos(x) \cos(y) \quad (101)$$

We derive a velocity given by:

$$v_x = \frac{C_x}{\cos(x)}, v_y = \frac{C_y}{\cos(y)} \quad (102)$$

With  $g_x = \frac{1}{\cos(y)}$  and  $g_y = \frac{1}{\cos(x)}$ , this leads us to a pressure profile:

$$p = \eta \left( \frac{4C_x \sin(x)}{3 \cos^2(x)} + \frac{4C_y \sin(y)}{3 \cos^2(y)} \right) + (\sin(x) + \sin(y)) + C_0 \quad (103)$$

$$\dot{\epsilon}_{xx} = C_x \frac{\sin(x)}{\cos^2(x)} \quad \dot{\epsilon}_{yy} = C_y \frac{\sin(y)}{\cos^2(y)} \quad \dot{\epsilon}_{xy} = 0$$

We choose  $\eta = 1$  and  $C_x = C_y = 1$ . The domain is the unit square  $[0 : 1] \times [0 : 1]$  and we obtain  $C_0$  as before and obtain

$$C_0 = 2 - 2 \cos(1) + 8/3 \left( \frac{1}{\cos(1)} - 1 \right) \simeq 3.18823730$$

(thank you WolframAlpha)

- benchmark #3 (ibench=3)
- benchmark #4 (ibench=4)
- benchmark #5 (ibench=5)

## features

- $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- Dirichlet boundary conditions (no-slip)
- isothermal
- isoviscous
- analytical solution
- pressure smoothing

ToDo:

- pbs with odd vs even number of elements
- q is 'fine' everywhere except in the corners - revisit pressure smoothing paper?
- redo A v d Berg benchmark (see Tom Weir thesis)

## 28 fieldstone: compressible flow (2)

This work is part of the MSc thesis of T. Weir (2018).

### 28.1 The physics

Let us start with some thermodynamics. Every material has an equation of state. The equilibrium thermodynamic state of any material can be constrained if any two state variables are specified. Examples of state variables include the pressure  $p$  and specific volume  $\nu = 1/\rho$ , as well as the temperature  $T$ .

After linearisation, the density depends on temperature and pressure as follows:

$$\rho(T, p) = \rho_0 ((1 - \alpha(T - T_0) + \beta_T p)$$

where  $\alpha$  is the coefficient of thermal expansion, also called thermal expansivity:

$$\alpha = -\frac{1}{\rho} \left( \frac{\partial \rho}{\partial T} \right)_p$$

$\alpha$  is the percentage increase in volume of a material per degree of temperature increase; the subscript  $p$  means that the pressure is held fixed.

$\beta_T$  is the isothermal compressibility of the fluid, which is given by

$$\beta_T = \frac{1}{K} = \frac{1}{\rho} \left( \frac{\partial \rho}{\partial P} \right)_T$$

with  $K$  the bulk modulus. Values of  $\beta_T = 10^{-12} - 10^{-11}$  Pa $^{-1}$  are reasonable for Earth's mantle, with values decreasing by about a factor of 5 between the shallow lithosphere and core-mantle boundary. This is the percentage increase in density per unit change in pressure at constant temperature. Both the coefficient of thermal expansion and the isothermal compressibility can be obtained from the equation of state.

The full set of equations we wish to solve is given by

$$-\nabla \cdot [2\eta \dot{\epsilon}^d(\mathbf{v})] + \nabla p = \rho_0 ((1 - \alpha(T - T_0) + \beta_T p) \mathbf{g} \quad \text{in } \Omega \quad (104)$$

$$\nabla \cdot \mathbf{v} + \frac{1}{\rho} \mathbf{v} \cdot \nabla \rho = 0 \quad \text{in } \Omega \quad (105)$$

$$\rho C_p \left( \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right) - \nabla \cdot k \nabla T = \rho H + 2\eta \dot{\epsilon}^d : \dot{\epsilon}^d + \alpha T \left( \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) \quad \text{in } \Omega, \quad (106)$$

Note that this presupposes that the density is not zero anywhere in the domain.

### 28.2 The numerics

We use a mixed formulation and therefore keep both velocity and pressure as unknowns. We end up having to solve the following system:

$$\begin{pmatrix} \mathbb{K} & \mathbb{G} + \mathbb{W} \\ \mathbb{G}^T + \mathbb{Z} & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} f \\ h \end{pmatrix} \quad \text{or,} \quad \mathbb{A} \cdot X = rhs$$

Where  $\mathbb{K}$  is the stiffness matrix,  $\mathbb{G}$  is the discrete gradient operator,  $\mathbb{G}^T$  is the discrete divergence operator,  $\mathcal{V}$  the velocity vector,  $\mathcal{P}$  the pressure vector. Note that the term  $\mathbb{Z}\mathcal{V}$  derives from term  $\mathbf{v} \cdot \nabla \rho$  in the continuity equation.

As perfectly explained in the step 32 of deal.ii<sup>9</sup>, we need to scale the  $\mathbb{G}$  term since it is many orders of magnitude smaller than  $\mathbb{K}$ , which introduces large inaccuracies in the solving process to the point that the solution is nonsensical. This scaling coefficient is  $\eta/L$ . After building the  $\mathbb{G}$  block, it is then scaled as follows:  $\mathbb{G}' = \frac{\eta}{L} \mathbb{G}$  so that we now solve

<sup>9</sup>[https://www.dealii.org/9.0.0/doxygen/deal.II/step\\_32.html](https://www.dealii.org/9.0.0/doxygen/deal.II/step_32.html)

$$\begin{pmatrix} \mathbb{K} & \mathbb{G}' + \mathbb{W} \\ \mathbb{G}'^T + \mathbb{Z} & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V} \\ \mathcal{P}' \end{pmatrix} = \begin{pmatrix} f \\ h \end{pmatrix}$$

After the solve phase, we recover the real pressure with  $\mathcal{P} = \frac{\eta}{L}\mathcal{P}'$ .

adapt notes since I should scale  $\mathbb{W}$  and  $\mathbb{Z}$  too.  $h$  should be scaled too !!!!!!!

Each block  $\mathbb{K}$ ,  $\mathbb{G}$ ,  $\mathbb{Z}$  and vectors  $f$  and  $h$  are built separately in the code and assembled into the matrix  $\mathbb{A}$  and vector  $rhs$  afterwards.  $\mathbb{A}$  and  $rhs$  are then passed to the solver. We will see later that there are alternatives to solve this approach which do not require to build the full Stokes matrix  $\mathbb{A}$ .

**Remark 1:** the terms  $\mathbb{Z}\mathcal{V}$  and  $\mathbb{W}\mathcal{P}$  are often put in the rhs (i.e. added to  $h$ ) so that the matrix  $\mathbb{A}$  retains the same structure as in the incompressible case. This is indeed how it is implemented in ASPECT, see also appendix A of [51]. This however requires more work since the rhs depends on the solution and some form of iterations is needed.

**Remark 2:** Very often the adiabatic heating term  $\alpha T(\mathbf{v} \cdot \nabla p)$  is simplified as follows: If you assume the vertical component of the gradient of the dynamic pressure to be small compared to the gradient of the total pressure (in other words, the gradient is dominated by the gradient of the hydrostatic pressure), then  $-\rho\mathbf{g} \simeq \nabla p$  and then  $\alpha T(\mathbf{v} \cdot \nabla p) \simeq -\alpha\rho T\mathbf{v} \cdot \mathbf{g}$ . We will however not be using this approximation in what follows.

We have already established that

$$\vec{\tau} = \mathbf{C}_\eta \mathbf{B} V$$

The following measurements are carried out:

- The root mean square velocity (**vrms**):

$$v_{rms} = \sqrt{\frac{1}{V} \int_V v^2 dV}$$

- The average temperature (**Tavrg**):

$$\langle T \rangle = \frac{1}{V} \int_V T dV$$

- The total mass (**mass**):

$$M = \int_V \rho dV$$

- The Nusselt number (**Nu**):

$$Nu = -\frac{1}{Lx} \frac{1}{\Delta T} \int_0^{L_x} \frac{\partial T(x, y = L_y)}{\partial y} dx$$

- The kinetic energy (**EK**):

$$E_K = \int_V \frac{1}{2} \rho v^2 dV$$

- The work done against gravity

$$\langle W \rangle = - \int_V \rho g_y v_y dV$$

- The total viscous dissipation (**visc\_diss**)

$$\langle \Phi \rangle = \int \Phi dV = \frac{1}{V} \int 2\eta \dot{\epsilon} : \dot{\epsilon} dV$$

- The gravitational potential energy (**EG**)

$$E_G = \int_V \rho g_y (L_y - y) dV$$

- The internal thermal energy (ET)

$$E_T = \int_V \rho_{(0)} C_p T dV$$

**Remark 3:** Measuring the total mass can be misleading: indeed because  $\rho = \rho_0(1 - \alpha T)$ , then measuring the total mass amounts to measuring a constant minus the volume-integrated temperature, and there is no reason why the latter should be zero, so that there is no reason why the total mass should be zero...!

### 28.3 The experimental setup

The setup is as follows: the domain is  $Lx = Ly = 3000\text{km}$ . Free slip boundary conditions are imposed on all four sides. The initial temperature is given by:

$$T(x, y) = \left( \frac{L_y - y}{L_y} - 0.01 \cos\left(\frac{\pi x}{L_x}\right) \sin\left(\frac{\pi y}{L_y}\right) \right) \Delta T + T_{surf}$$

with  $\Delta T = 4000\text{K}$ ,  $T_{surf} = T_0 = 273.15\text{K}$ . The temperature is set to  $\Delta T + T_{surf}$  at the bottom and  $T_{surf}$  at the top. We also set  $k = 3$ ,  $C_p = 1250$ ,  $|g| = 10$ ,  $\rho_0 = 3000$  and we keep the Rayleigh number  $Ra$  and dissipation number  $Di$  as input parameters:

$$Ra = \frac{\alpha g \Delta T L^3 \rho_0^2 C_p}{\eta k} \quad Di = \frac{\alpha g L}{C_p}$$

From the second equation we get  $\alpha = \frac{Di C_p}{gL}$ , which we can insert in the first one:

$$Ra = \frac{Di C_p^2 \Delta T L^2 \rho_0^2}{\eta k} \quad \text{or}, \quad \eta = \frac{Di C_p^2 \Delta T L^2 \rho_0^2}{Ra k}$$

For instance, for  $Ra = 10^4$  and  $Di = 0.75$ , we obtain  $\alpha \simeq 3 \cdot 10^{-5}$  and  $\eta \simeq 10^{25}$  which are quite reasonable values.

### 28.4 Scaling

Following [47], we non-dimensionalize the equations using the reference values for density  $\rho_r$ , thermal expansivity  $\alpha_r$ , temperature contrast  $\Delta T_r$  (`refTemp`), thermal conductivity  $k_r$ , heat capacity  $C_p$ , depth of the fluid layer  $L$  and viscosity  $\eta_r$ . The non-dimensionalization for velocity,  $u_r$ , pressure  $p_r$  and time,  $t_r$  become

$$u_r = \frac{k_r}{\rho_r C_p L} \quad (\text{refvel})$$

$$p_r = \frac{\eta_r k_r}{\rho_r C_p L^2} \quad (\text{refpress})$$

$$t_r = \frac{\rho_r C_p L^2}{k_r} \quad (\text{reftime})$$

In the case of the setup described hereabove, and when choosing  $Ra = 10^4$  and  $Di = 0.5$ , we get:

```
alphaT 2.08333e-05
eta 8.437500e+24
reftime 1.125000e+19
refvel 2.666667e-13
refPress 7.500000e+05
```

## 28.5 Conservation of energy 1

### 28.5.1 under BA and EBA approximations

Following [51], we take the dot product of the momentum equation with the velocity  $\mathbf{v}$  and integrate over the whole volume<sup>10</sup>:

$$\int_V [-\nabla \cdot \boldsymbol{\tau} + \nabla p] \cdot \mathbf{v} dV = \int_V \rho \mathbf{g} \cdot \mathbf{v} dV$$

or,

$$-\int_V (\nabla \cdot \boldsymbol{\tau}) \cdot \mathbf{v} dV + \int_V \nabla p \cdot \mathbf{v} dV = \int_V \rho \mathbf{g} \cdot \mathbf{v} dV$$

Let us look at each block separately:

$$-\int_V (\nabla \cdot \boldsymbol{\tau}) \cdot \mathbf{v} dV = -\int_S \underbrace{\boldsymbol{\tau} \cdot \mathbf{n}}_{=0 \text{ (b.c.)}} dS + \int_V \boldsymbol{\tau} : \nabla \mathbf{v} dV = \int_V \boldsymbol{\tau} : \dot{\boldsymbol{\epsilon}} dV = \int_V \Phi dV$$

which is the volume integral of the shear heating. Then,

$$\int_V \nabla p \cdot \mathbf{v} dV = \int_S \underbrace{p \cdot \mathbf{n}}_{=0 \text{ (b.c.)}} dS - \int_V \underbrace{\nabla \cdot \mathbf{v}}_{=0 \text{ (incomp.)}} pdV = 0$$

which is then zero in the case of an incompressible flow. And finally

$$\int_V \rho \mathbf{g} \cdot \mathbf{v} dV = W$$

which is the work against gravity.

Conclusion for an *incompressible* fluid: we should have

$$\int_V \Phi dV = \int_V \rho \mathbf{g} \cdot \mathbf{v} dV \quad (107)$$

This formula is hugely problematic: indeed, the term  $\rho$  in the rhs is the full density. We know that to the value of  $\rho_0$  corresponds a lithostatic pressure gradient  $p_L = \rho_0 gy$ . In this case one can write  $\rho = \rho_0 + \rho'$  and  $p = p_L + p'$  so that we also have

$$\int_V [-\nabla \cdot \boldsymbol{\tau} + \nabla p'] \cdot \mathbf{v} dV = \int_V \rho' \mathbf{g} \cdot \mathbf{v} dV$$

which will ultimately yield

$$\int_V \Phi dV = \int_V \rho' \mathbf{g} \cdot \mathbf{v} dV = \int_V (\rho - \rho_0) \mathbf{g} \cdot \mathbf{v} dV \quad (108)$$

Obviously Eqs.(107) and (108) cannot be true at the same time. The problem comes from the nature of the (E)BA approximation:  $\rho = \rho_0$  in the mass conservation equation but it is not constant in the momentum conservation equation, which is of course inconsistent. Since the mass conservation equation is  $\nabla \cdot \mathbf{v} = 0$  under this approximation then the term  $\int_V \nabla p \cdot \mathbf{v} dV$  is always zero for any pressure (full pressure  $p$ , or overpressure  $p - p_L$ ), hence the paradox. This paradox will be lifted when a consistent set of equations will be used (compressible formulation). On a practical note, Eqs.(107) is not verified by the code, while (108) is.

In the end:

$$\boxed{\int_V \Phi dV = \underbrace{\int_V (\rho - \rho_0) \mathbf{g} \cdot \mathbf{v} dV}_{\text{visc.diss}} + \underbrace{\int_V (\rho - \rho_0) \mathbf{g} \cdot \mathbf{v} dV}_{\text{work_grav}}} \quad (109)$$

<sup>10</sup>Check: this is akin to looking at the power, force\*velocity, says Arie

### 28.5.2 under no approximation at all

$$\int_V \nabla p \cdot \mathbf{v} dV = \int_S p \underbrace{\mathbf{v} \cdot \mathbf{n}}_{=0 \text{ (b.c.)}} dS - \int_V \nabla \cdot \mathbf{v} pdV = 0 \quad (110)$$

$$= \int_V \frac{1}{\rho} \mathbf{v} \cdot \nabla \rho pdV = 0 \quad (111)$$

(112)

ToDo: see section 3 of [51] where this is carried out with the Adams-Williamson eos.

## 28.6 Conservation of energy 2

Also, following the Reynold's transport theorem [56], p210, we have for a property  $A$  (per unit mass)

$$\frac{d}{dt} \int_V A \rho dV = \int_V \frac{\partial}{\partial t} (A \rho) dV + \int_S A \rho \mathbf{v} \cdot \mathbf{n} dS$$

Let us apply to this to  $A = C_p T$  and compute the time derivative of the internal energy:

$$\frac{d}{dt} \int_V \rho C_p T dV = \int_V \frac{\partial}{\partial t} (\rho C_p T) dV + \int_S \rho \underbrace{\mathbf{v} \cdot \mathbf{n}}_{=0 \text{ (b.c.)}} dS = \underbrace{\int_V C_p T \frac{\partial \rho}{\partial t} dV}_I + \underbrace{\int_V \rho C_p \frac{\partial T}{\partial t} dV}_{II} \quad (113)$$

In order to expand  $I$ , the mass conservation equation will be used, while the heat transport equation will be used for  $II$ :

$$I = \int_V C_p T \frac{\partial \rho}{\partial t} dV = - \int_V C_p T \nabla \cdot (\rho \mathbf{v}) dV = - \int_V C_p T \rho \underbrace{\mathbf{v} \cdot \mathbf{n}}_{=0 \text{ (b.c.)}} dS + \int_V \rho C_p \nabla T \cdot \mathbf{v} dV \quad (114)$$

$$II = \int_V \rho C_p \frac{\partial T}{\partial t} dV = \int_V \left[ -\rho C_p \mathbf{v} \cdot \nabla T + \nabla \cdot k \nabla T + \rho H + \Phi + \alpha T \left( \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) \right] dV \quad (115)$$

$$= \int_V \left[ -\rho C_p \mathbf{v} \cdot \nabla T + \rho H + \Phi + \alpha T \left( \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) \right] dV + \int_V \nabla \cdot k \nabla T dV \quad (116)$$

$$= \int_V \left[ -\rho C_p \mathbf{v} \cdot \nabla T + \rho H + \Phi + \alpha T \left( \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) \right] dV + \int_S k \nabla T \cdot \mathbf{n} dS \quad (117)$$

$$= \int_V \left[ -\rho C_p \mathbf{v} \cdot \nabla T + \rho H + \Phi + \alpha T \left( \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) \right] dV - \int_S \mathbf{q} \cdot \mathbf{n} dS \quad (118)$$

Finally:

$$I + II = \frac{d}{dt} \underbrace{\int_V \rho C_p T dV}_{ET} = \int_V \left[ \rho H + \Phi + \alpha T \left( \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) \right] dV - \int_S \mathbf{q} \cdot \mathbf{n} dS \quad (119)$$

$$= \int_V \rho H dV + \underbrace{\int_V \Phi dV}_{visc\_diss} + \underbrace{\int_V \alpha T \frac{\partial p}{\partial t} dV}_{extra} + \underbrace{\int_V \alpha T \mathbf{v} \cdot \nabla p dV}_{adiab\_heating} - \underbrace{\int_S \mathbf{q} \cdot \mathbf{n} dS}_{heatflux\_boundary} \quad (120)$$

This was of course needlessly complicated as the term  $\partial \rho / \partial t$  is always taken to be zero, so that  $I = 0$  automatically. The mass conservation equation is then simply  $\nabla \cdot (\rho \mathbf{v}) = 0$ . Then it follows that

$$0 = \int_V C_p T \nabla \cdot (\rho \mathbf{v}) dV = - \int_V C_p T \rho \underbrace{\mathbf{v} \cdot \mathbf{n}}_{=0 \text{ (b.c.)}} dS + \int_V \rho C_p \nabla T \cdot \mathbf{v} dV \quad (121)$$

$$= \int_V \rho C_p \nabla T \cdot \mathbf{v} dV \quad (122)$$

so that the same term in Eq.(118) vanishes too, and then Eq.(120) is always valid, although one should be careful when computing  $E_T$  in the BA and EBA cases as it should use  $\rho_0$  and not  $\rho$ .

## 28.7 The problem of the onset of convection

[wiki] In geophysics, the Rayleigh number is of fundamental importance: it indicates the presence and strength of convection within a fluid body such as the Earth's mantle. The mantle is a solid that behaves as a fluid over geological time scales.

The Rayleigh number essentially is an indicator of the type of heat transport mechanism. At low Rayleigh numbers conduction processes dominate over convection ones. At high Rayleigh numbers it is the other way around. There is a so-called critical value of the number with delineates the transition from one regime to the other.

This problem has been studied and approached both theoretically and numerically [86, e.g.] and it was found that the critical Rayleigh number  $Ra_c$  is

$$Ra_c = (27/4)\pi^4 \simeq 657.5$$

in setups similar to ours.

### VERY BIG PROBLEM

The temperature setup is built as follows:  $T_{surf}$  is prescribed at the top,  $T_{surf} + \Delta T$  is prescribed at the bottom. The initial temperature profile is linear between these two values. In the case of BA, the actual value of  $T_{surf}$  is of no consequence. However, for the EBA the full temperature is present in the adiabatic heating term on the rhs of the hte, and the value of  $T_{surf}$  will therefore influence the solution greatly. This is very problematic as there is no real way to arrive at the surface temperature from the King paper. On top of this, the density uses a reference temperature  $T_0$  which too will influence the solution without being present in the controlling  $Ra$  and  $Di$  numbers!!

In light thereof, it will be very difficult to recover the values of King et al for EBA!

#### features

- $Q_1 \times P_0$  element
- compressible flow
- mixed formulation
- Dirichlet boundary conditions (no-slip)
- isoviscous
- analytical solution
- pressure smoothing

Relevant literature: [7, 44, 78, 51, 47, 52, 53, 40]

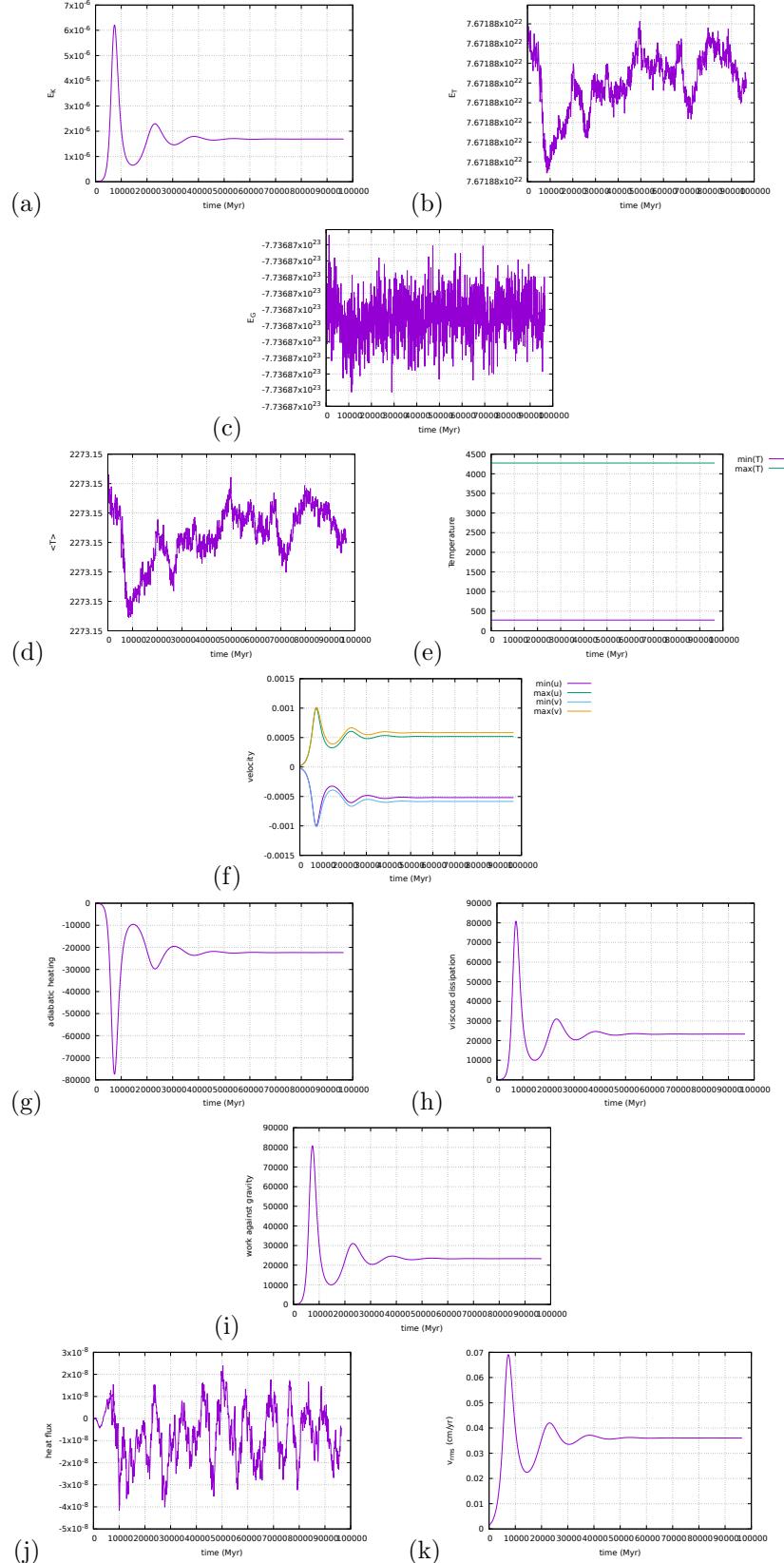
ToDo:

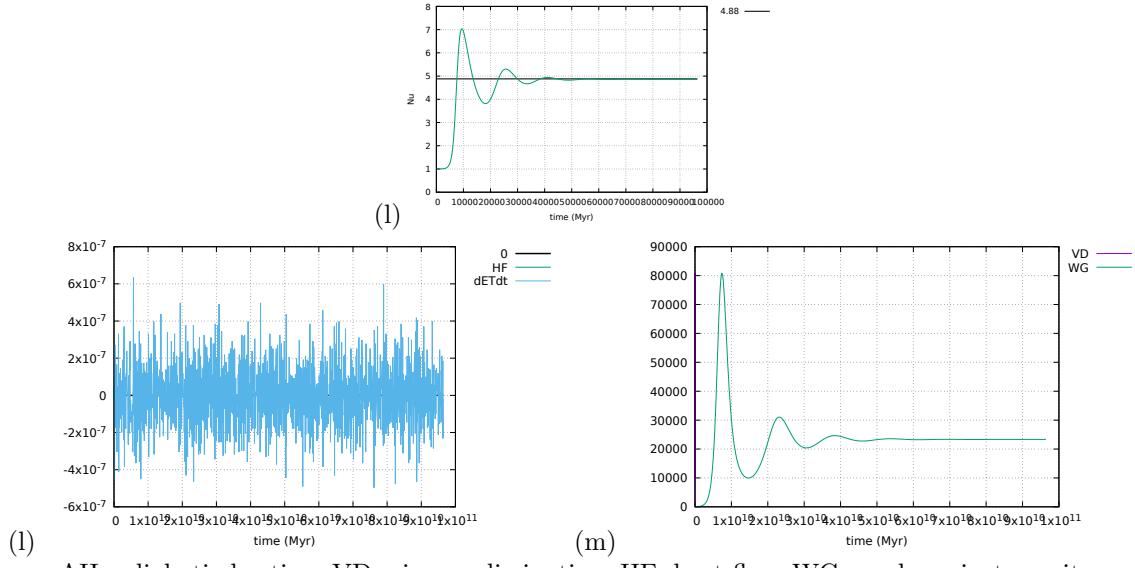
- heat flux is at the moment elemental, so Nusselt and heat flux on boundaries measurements not as accurate as could be.

- implement steady state detection
- do  $Ra = 10^5$  and  $Ra = 10^6$
- velocity average at surface
- non dimensional heat flux at corners [8]
- depth-dependent viscosity (case 2 of [8])

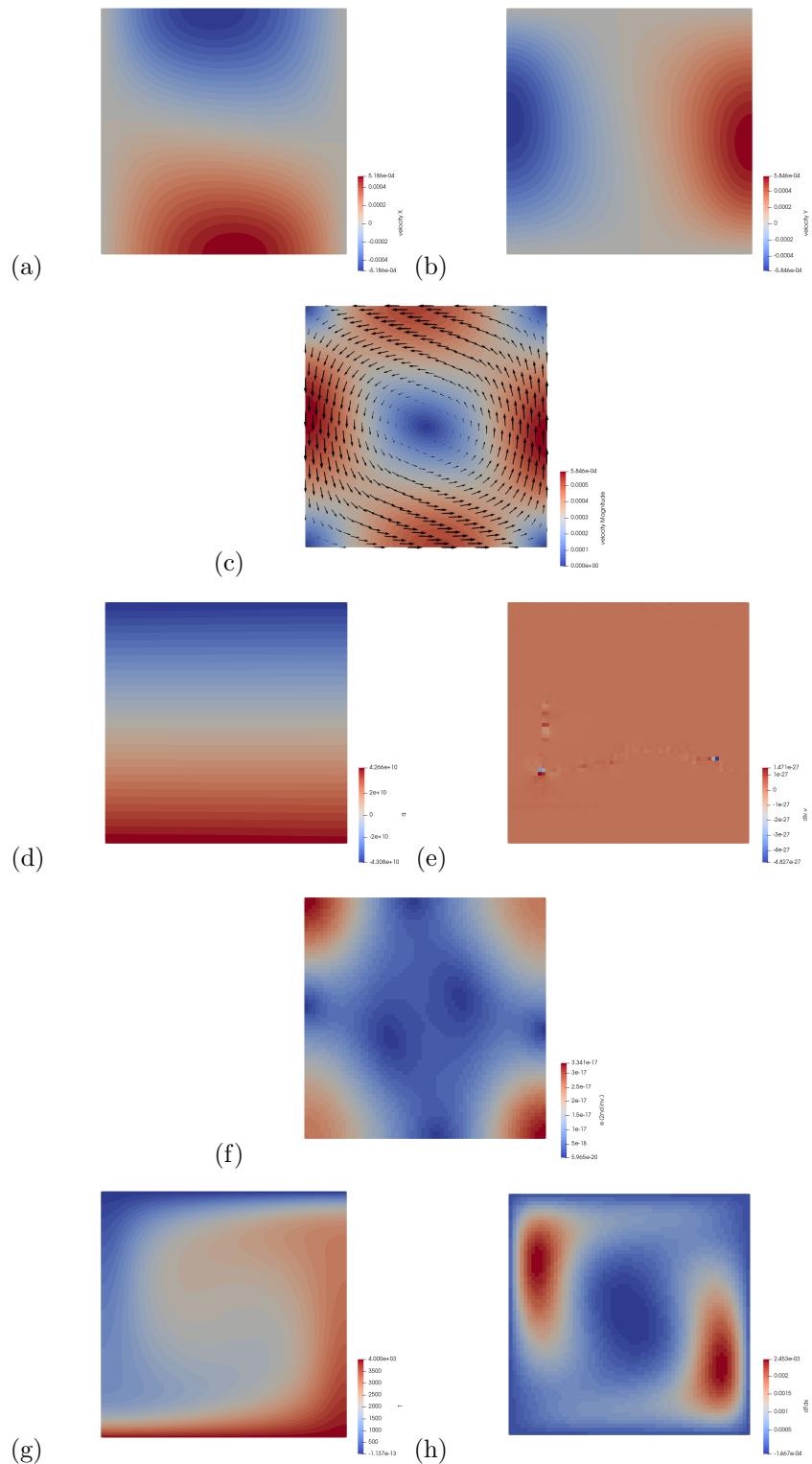
## 28.8 results - BA - $Ra = 10^4$

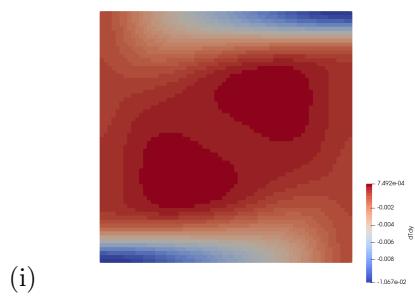
These results were obtained with a 64x64 resolution, and CFL number of 1. Steady state was reached after about 1250 timesteps.





Eq.(120) is verified by (l) and Eq.(109) is verified by (m).

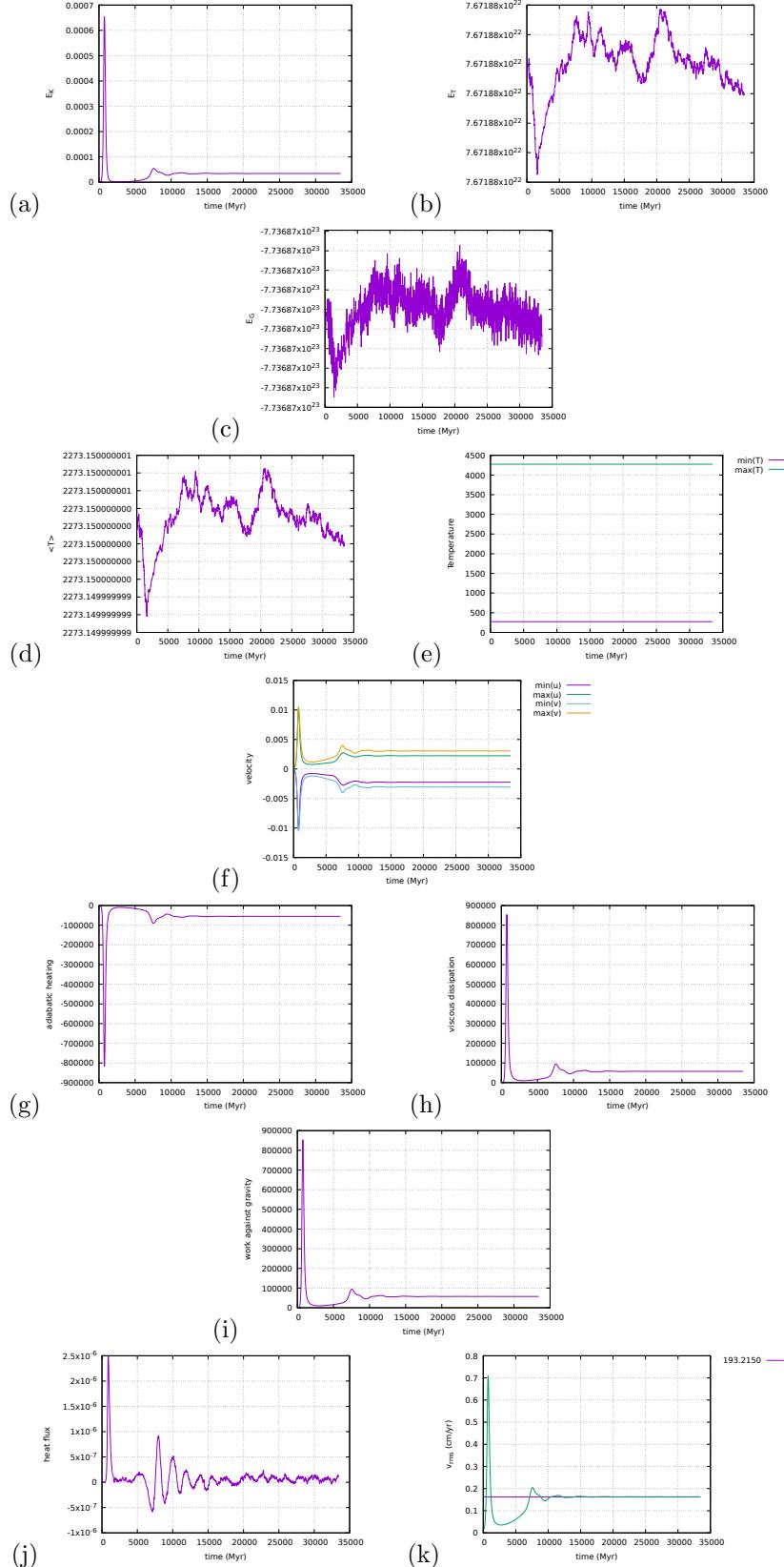


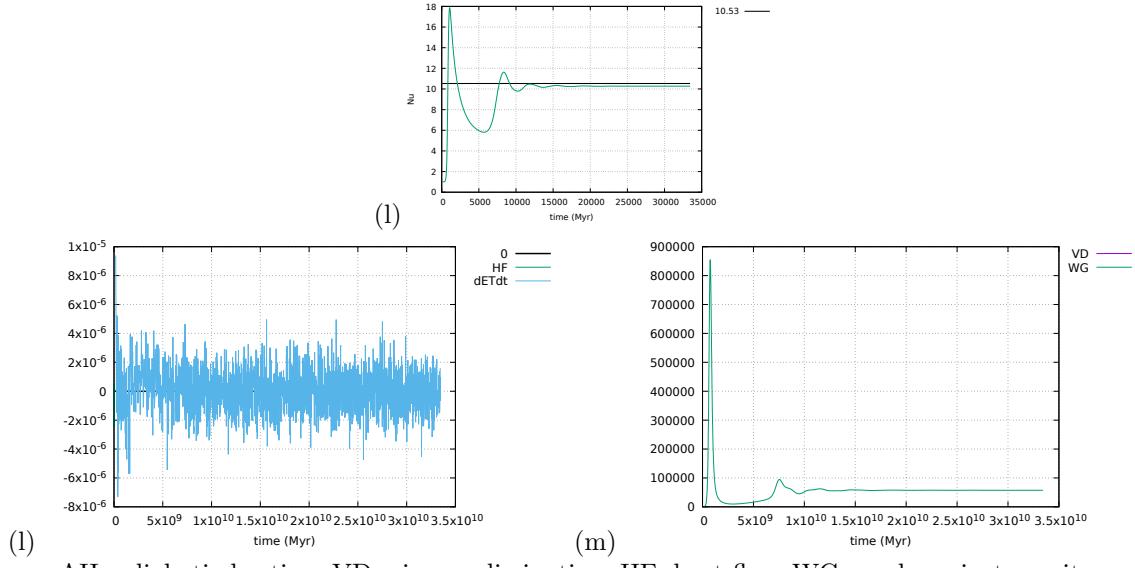


(i)

## 28.9 results - BA - $Ra = 10^5$

These results were obtained with a 64x64 resolution, and CFL number of 1. Steady state was reached after about 1250 timesteps.



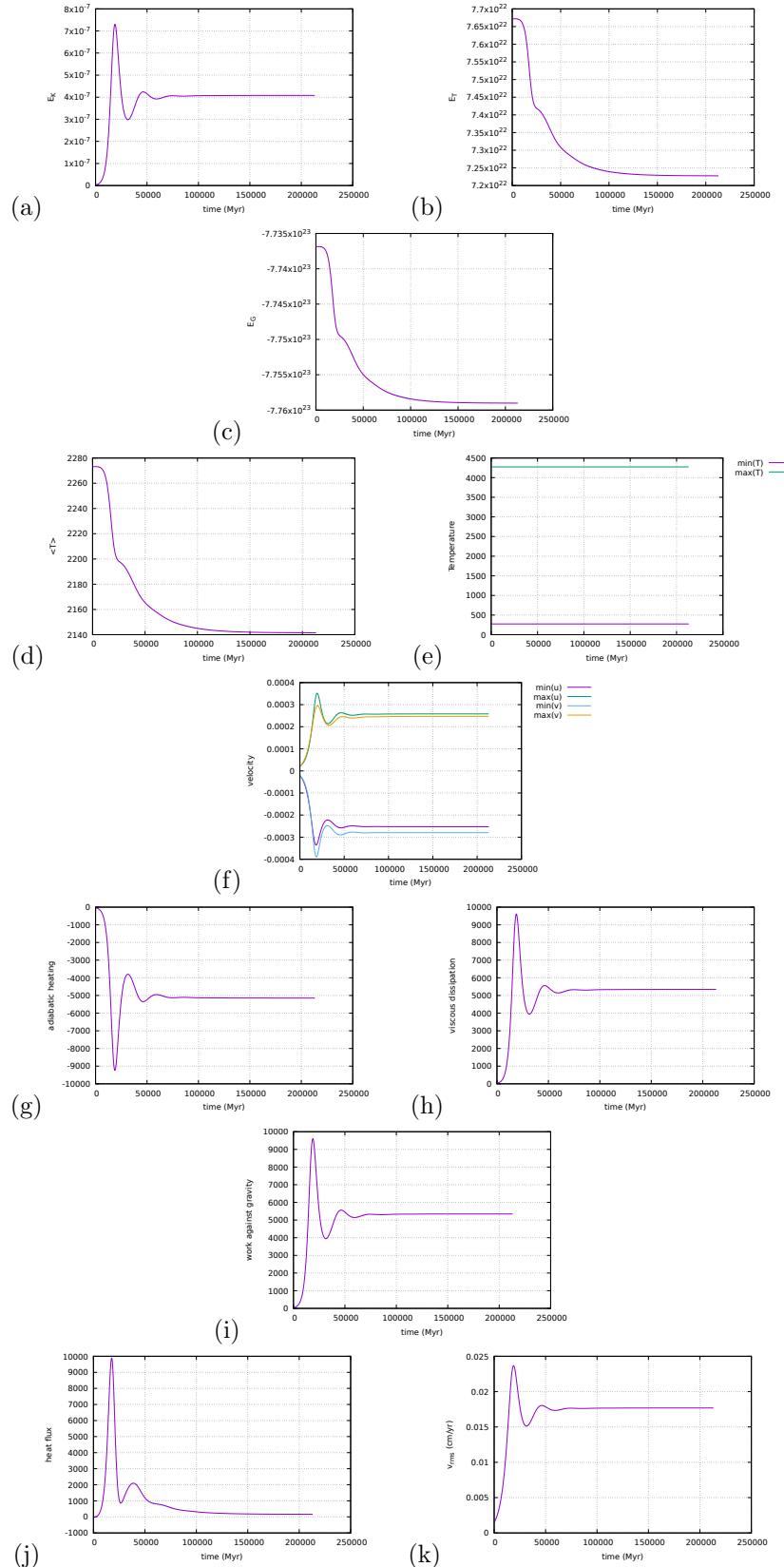


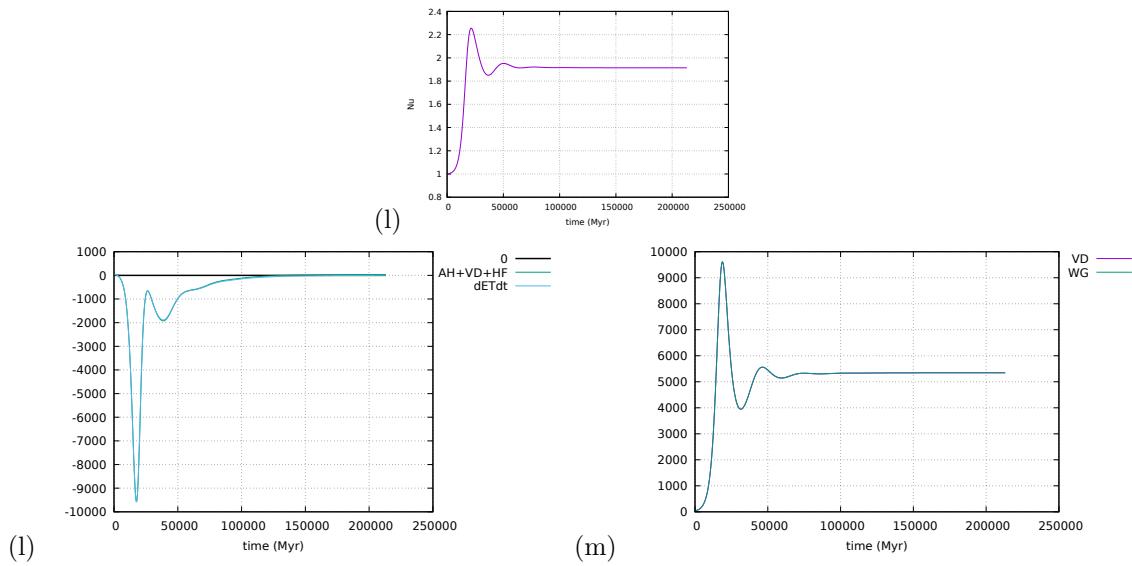
Eq.(120) is verified by (l) and Eq.(109) is verified by (m).

## **28.10 results - BA - $Ra = 10^6$**

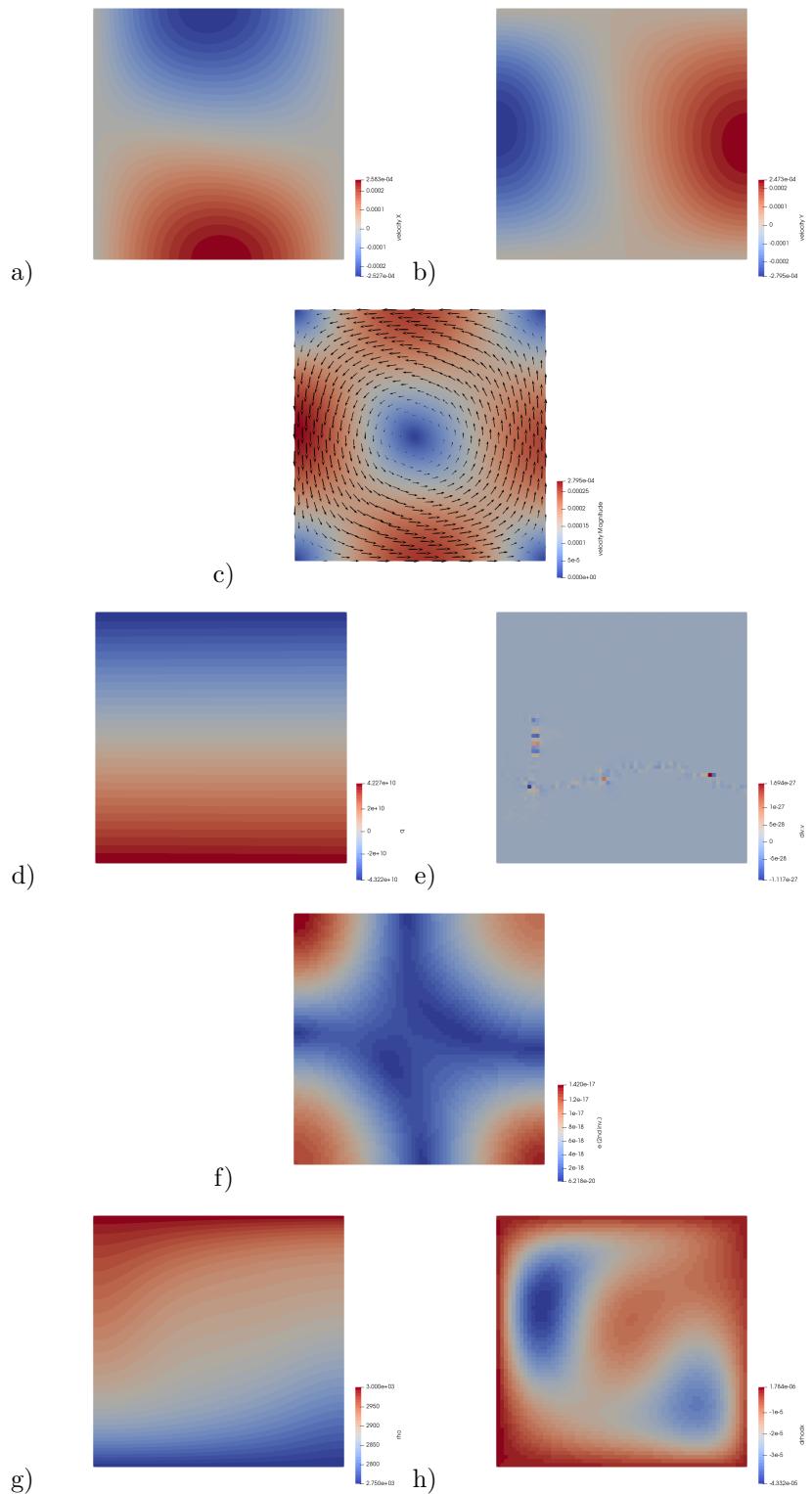
## 28.11 results - EBA - $Ra = 10^4$

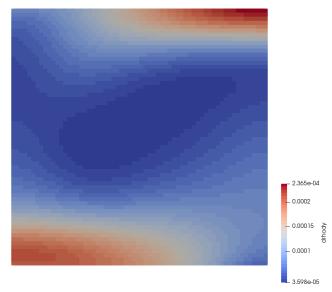
These results were obtained with a 64x64 resolution, and CFL number of 1. Steady state was reached after about 2500 timesteps



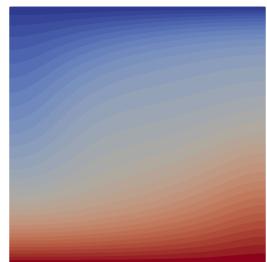


Eq.(120) is verified by (l) and Eq.(109) is verified by (m).

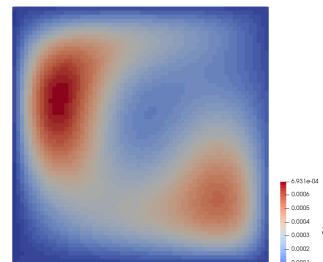




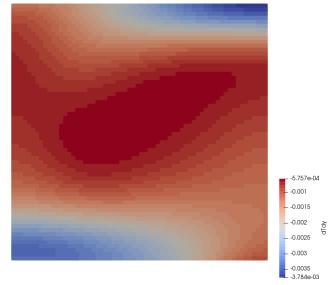
i)



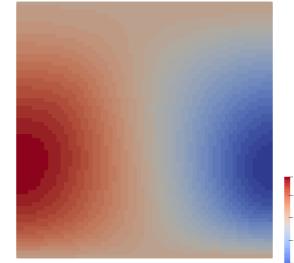
j)



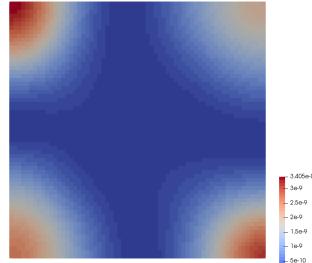
k)



l)



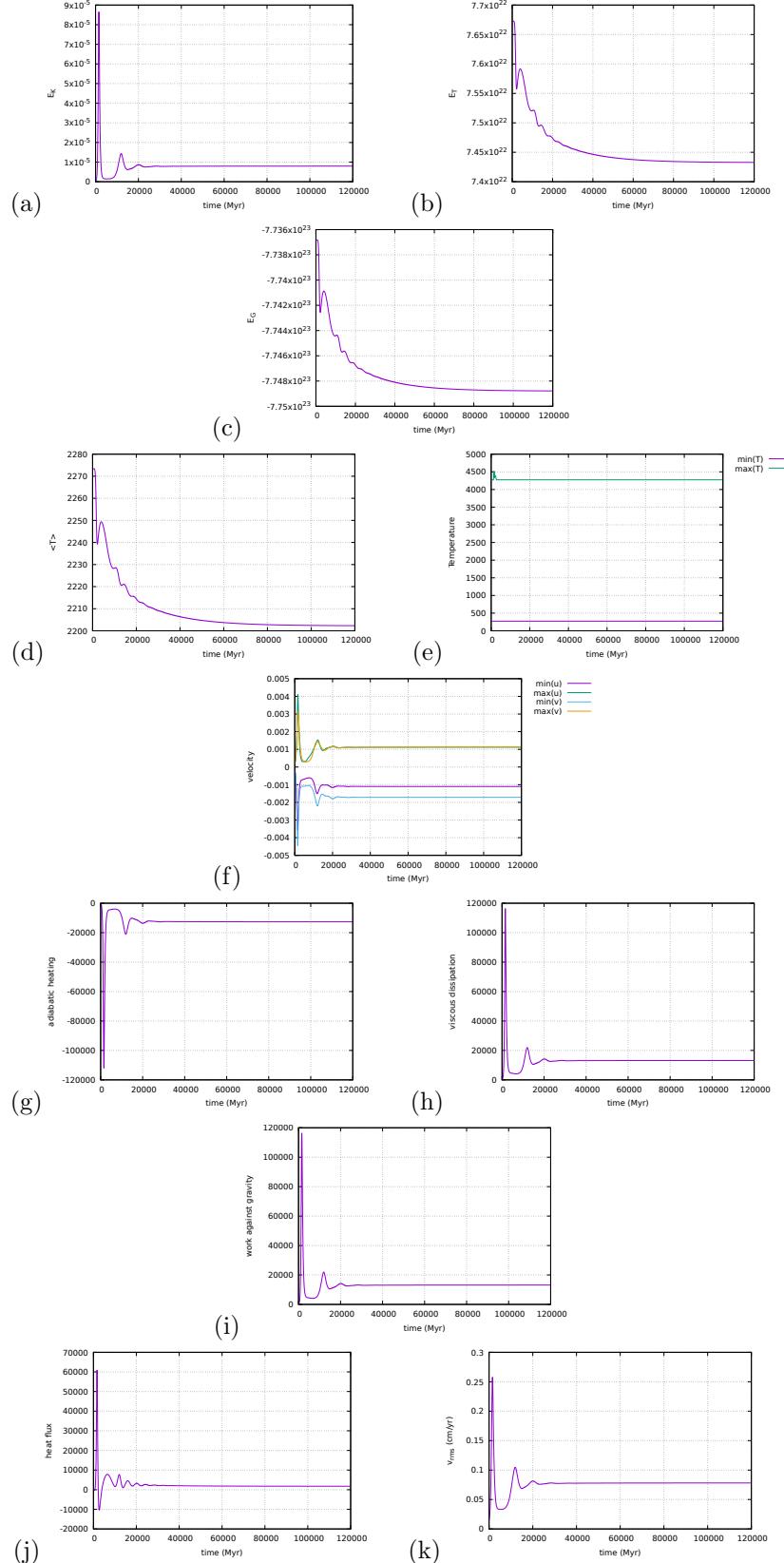
m)

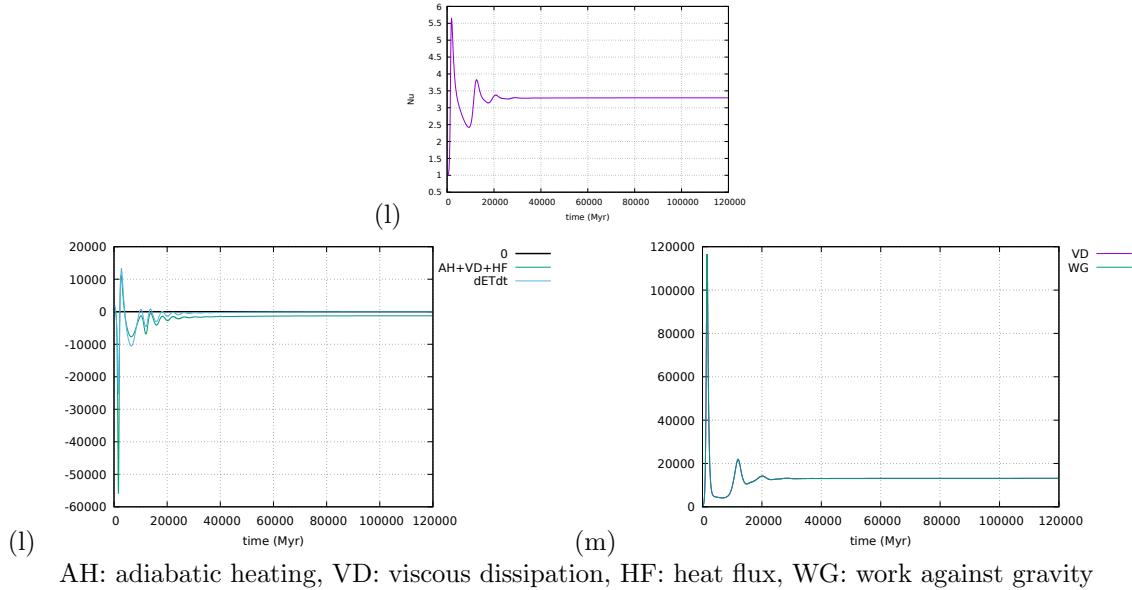


n)

## 28.12 results - EBA - $Ra = 10^5$

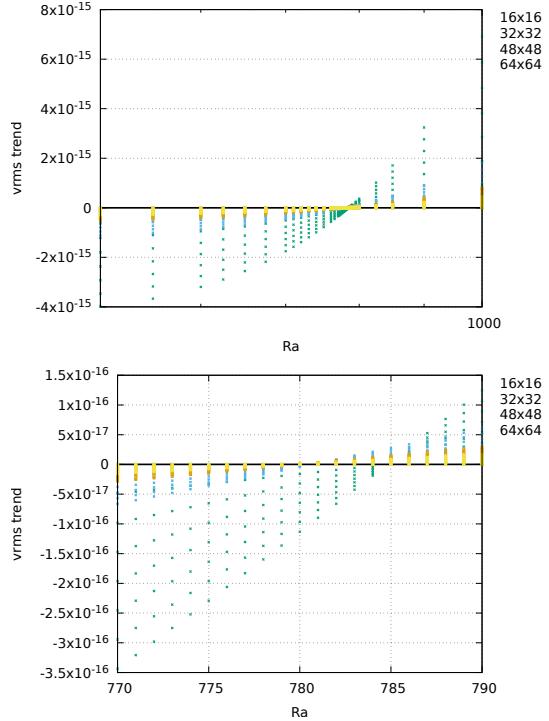
These results were obtained with a 64x64 resolution, and CFL number of 1. Simulation was stopped after about 4300 timesteps.





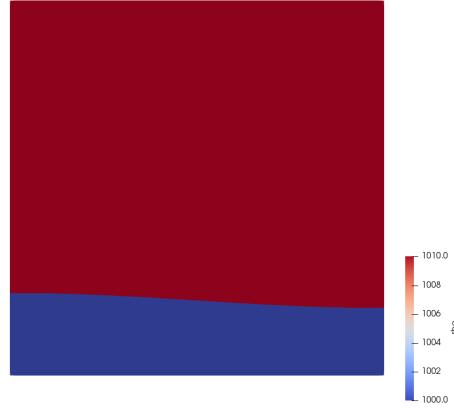
### 28.13 Onset of convection

The code can be run for values of Ra between 500 and 1000, at various resolutions for the BA formulation. The value  $v_{rms}(t) - v_{rms}(0)$  is plotted as a function of  $Ra$  and for the 10 first timesteps. If the  $v_{rms}$  is found to decrease, then the Rayleigh number is not high enough to allow for convection and the initial temperature perturbation relaxes by diffusion (and then  $v_{rms}(t) - v_{rms}(0) < 0$ ). If the  $v_{rms}$  is found to increase, then  $v_{rms}(t) - v_{rms}(0) > 0$  and the system is going to showcase convection. The zero value of  $v_{rms}(t) - v_{rms}(0)$  gives us the critical Rayleigh number, which is found between 775 and 790.



## 29 fieldstone: Rayleigh-Taylor instability (1)

This numerical experiment was first presented in [87]. It consists of an isothermal Rayleigh-Taylor instability in a two-dimensional box of size  $L_x = 0.9142$  and  $L_y = 1$ .

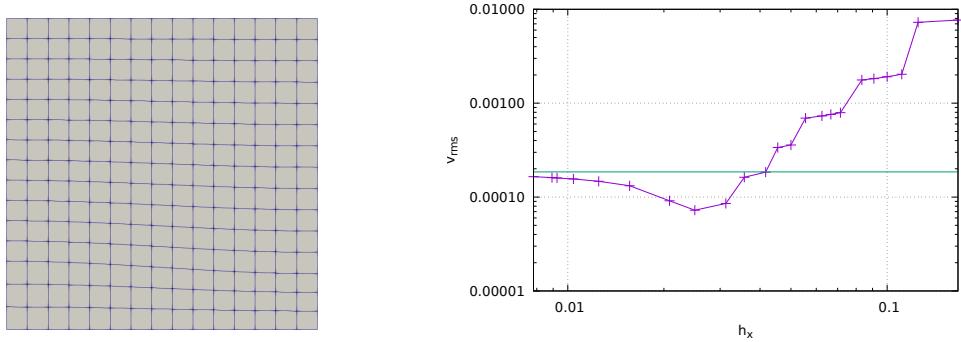


Two Newtonian fluids are present in the system: the buoyant layer is placed at the bottom of the box and the interface between both fluids is given by  $y(x) = 0.2 + 0.02 \cos\left(\frac{\pi x}{L_x}\right)$ . The bottom fluid is parametrised by its mass density  $\rho_1$  and its viscosity  $\mu_1$ , while the layer above is parametrised by  $\rho_2$  and  $\mu_2$ .

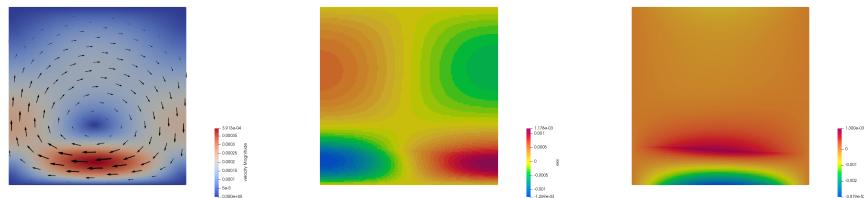
No-slip boundary conditions are applied at the bottom and at the top of the box while free-slip boundary conditions are applied on the sides.

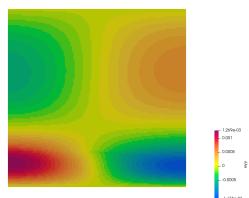
In the original benchmark the system is run over 2000 units of dimensionless time and the timing and position of various upwellings/downwellings is monitored. In this present experiment only the root mean square velocity is measured at  $t = 0$ : the code is indeed not yet foreseen of any algorithm capable of tracking deformation.

Another approach than the ones presented in the extensive literature which showcases results of this benchmark is taken. The mesh is initially fitted to the fluids interface and the resolution is progressively increased. This results in the following figure:



The green line indicates results obtained with my code ELEFANT with grids up to 2000x2000 with the exact same methodology.





### features

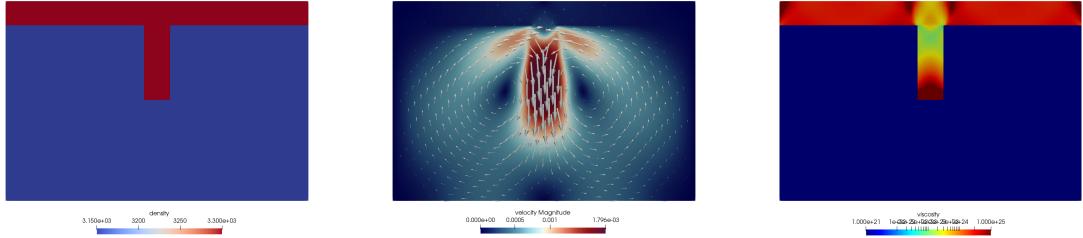
- $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- isothermal
- numerical benchmark

## 30 fieldstone: Slab detachment benchmark (1)

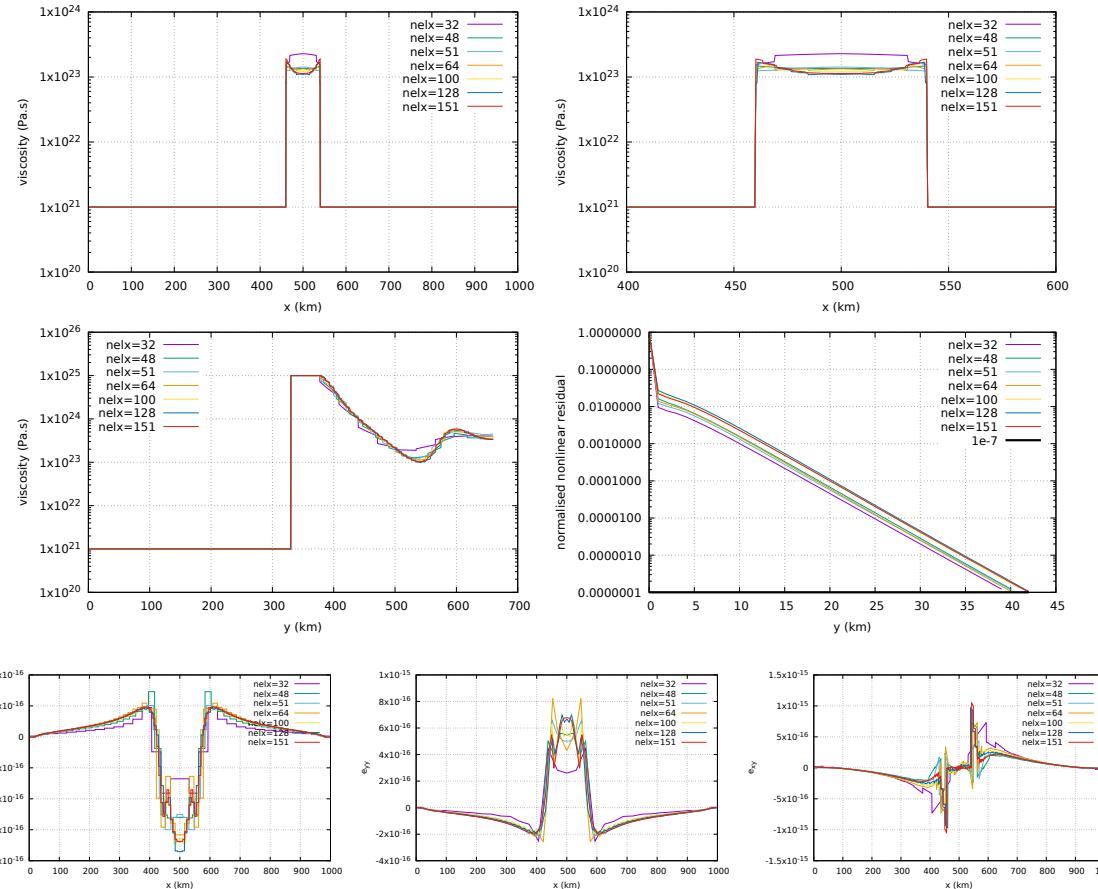
As in [71], the computational domain is  $1000\text{km} \times 660\text{km}$ . No-slip boundary conditions are imposed on the sides of the system while free-slip boundary conditions are imposed at the top and bottom. Two materials are present in the domain: the lithosphere (mat.1) and the mantle (mat.2). The overriding plate (mat.1) is  $80\text{km}$  thick and is placed at the top of the domain. An already subducted slab (mat.1) of  $250\text{km}$  length hangs vertically under this plate. The mantle occupies the rest of the domain.

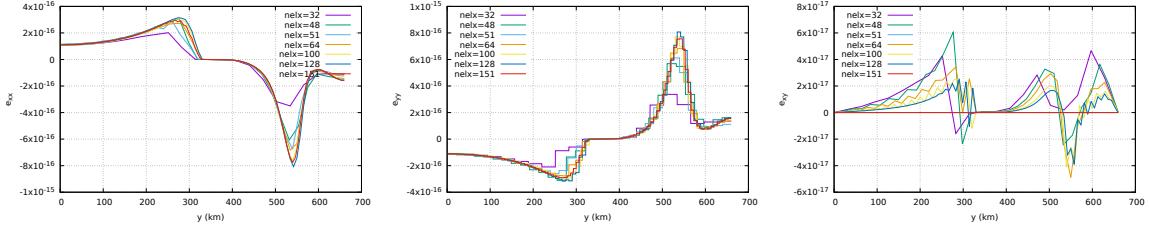
The mantle has a constant viscosity  $\mu_0 = 10^{21}\text{Pa.s}$  and a density  $\rho = 3150\text{kg/m}^3$ . The slab has a density  $\rho = 3300\text{kg/m}^3$  and is characterised by a power-law flow law so that its effective viscosity depends on the second invariant of the strainrate  $I_2$  as follows:  $n_s = 4$  and  $A = (2 \times 4.75 \times 10^{11})^{-n_s}$ , or  $\mu_0 = 4.75 \times 10^{11}$ .

$$\mu_{eff} = \frac{1}{2} A^{-1/n_s} I_2^{1/n_s - 1} = \frac{1}{2} [(2 \times 4.75 \times 10^{11})^{-n_s}]^{-1/n_s} I_2^{1/n_s - 1} = 4.75 \times 10^{11} I_2^{1/n_s - 1} = \mu_0 I_2^{1/n_s - 1} \quad (123)$$



Fields at convergence for  $151 \times 99$  grid.





## features

- $Q_1 \times P_0$  element
- incompressible flow
- mixed formulation
- isothermal
- nonlinear rheology

Todo: nonlinear mantle, pressure normalisation

## References

- [1] M. Albers. A local mesh refinement multigrid method for 3D convection problems with strongly variable viscosity. *J. Comp. Phys.*, 160:126–150, 2000.
- [2] V. Allken, R. Huismans, and C. Thieulot. Three dimensional numerical modelling of upper crustal extensional systems. *J. Geophys. Res.*, 116:B10409, 2011.
- [3] V. Allken, R. Huismans, and C. Thieulot. Factors controlling the mode of rift interaction in brittle-ductile coupled systems: a 3d numerical study. *Geochem. Geophys. Geosyst.*, 13(5):Q05010, 2012.
- [4] V. Allken, R.S. Huismans, H. Fossen, and C. Thieulot. 3D numerical modelling of graben interaction and linkage: a case study of the Canyonlands grabens, Utah. *Basin Research*, 25:1–14, 2013.
- [5] K.-J. Bathe. *Finite Element Procedures in Engineering Analysis*. Prentice-Hall, 1982.
- [6] M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.
- [7] D. Bercovici, G. Schubert, and G.A. Glatzmaier. Three-dimensional convection of an infinite Prandtl-number compressible fluid in a basally heated spherical shell. *J. Fluid Mech.*, 239:683–719, 1992.
- [8] B. Blankenbach, F. Busse, U. Christensen, L. Cserepes, D. Gunkel, U. Hansen, H. Harder, G. Jarvis, M. Koch, G. Marquart, D. Moore, P. Olson, H. Schmeling, and T. Schnaubelt. A benchmark comparison for mantle convection codes. *Geophys. J. Int.*, 98:23–38, 1989.
- [9] J. Braun, C. Thieulot, P. Fullsack, M. DeKool, and R.S. Huismans. DOUAR: a new three-dimensional creeping flow model for the solution of geological problems. *Phys. Earth. Planet. Inter.*, 171:76–91, 2008.
- [10] J. Braun and P. Yamato. Structural evolution of a three-dimensional, finite-width crustal wedge. *Tectonophysics*, 484:181–192, doi:10.1016/j.tecto.2009.08.032, 2009.
- [11] H.H. Bui, R. Fukugawa, K. Sako, and S. Ohno. Lagrangian meshfree particles method (SPH) for large deformation and failure flows of geomaterial using elasticplastic soil constitutive model. *Int. J. Numer. Anal. Geomech.*, 32(12):1537–1570, 2008.
- [12] P.S. Bullen. *Handbook of Means and Their Inequalities*. Springer; 2nd edition, 2003.
- [13] C. Burstedde, G. Stadler, L. Alisic, L.C. Wilcox, E. Tan, M. Gurnis, and O. Ghattas. Large-scale adaptive mantle convection simulation. *Geophys. J. Int.*, 192:889–906, 2013.
- [14] F.H. Busse, U. Christensen, R. Clever, L. Cserepes, C. Gable, E. Giannandrea, L. Guillou, G. Houseman, H.-C. Nataf, M. Ogawa, M. Parmentier, C. Sotin, and B. Travis. 3D convection at infinite Prandtl number in Cartesian geometry - a benchmark comparison. *Geophys. Astrophys. Fluid Dynamics*, 75:39–59, 1993.
- [15] J.S. Chen, C. Pan, and T.Y.P. Chang. On the control of pressure oscillation in bilinear-displacement constant-pressure element. *Comput. Methods Appl. Mech. Engrg.*, 128:137–152, 1995.
- [16] Edmund Christiansen and Knud D. Andersen. Computation of collapse states with von mises type yield condition. *International Journal for Numerical Methods in Engineering*, 46:1185–1202, 1999.
- [17] Edmund Christiansen and Ole S. Pedersen. Automatic mesh refinement in limit analysis. *International Journal for Numerical Methods in Engineering*, 50:1331–1346, 2001.
- [18] M. crouzeix and P.A. Raviart. Conforming and non-conforming finite element methods for solving the stationary Stokes equations. *RAIRO*, 7:33–76, 1973.
- [19] C. Cuvelier, A. Segal, and A.A. van Steenhoven. *Finite Element Methods and Navier-Stokes Equations*. D. Reidel Publishing Company, 1986.

- [20] D.R. Davies, C.R. Wilson, and S.C. Kramer. Fluidity: A fully unstructured anisotropic adaptive mesh computational modeling framework for geodynamics. *Geochem. Geophys. Geosyst.*, 12(6), 2011.
- [21] P. Davy and P. Cobbold. Indentation tectonics in nature and experiment. 1. experiments scaled for gravity. *Bulletin of the Geological Institutions of Uppsala*, 14:129–141, 1988.
- [22] J. de Frutos, V. John, and J. Novo. Projection methods for incompressible flow problems with WENO finite difference schemes. *J. Comp. Phys.*, 309:368–386, 2016.
- [23] Y. Deubelbeiss and B.J.P. Kaus. Comparison of Eulerian and Lagrangian numerical techniques for the Stokes equations in the presence of strongly varying viscosity. *Phys. Earth Planet. Interiors*, 171:92–111, 2008.
- [24] C.R. Dohrmann and P.B. Bochev. A stabilized finite element method for the Stokes problem based on polynomial pressure projections. *Int. J. Num. Meth. Fluids*, 46:183–201, 2004.
- [25] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*. 2003.
- [26] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*. John Wiley & Sons, 2003.
- [27] Jean Donea and Antonio Huerta. *Finite Element Methods for Flow Problems*. John Wiley & Sons, 2003.
- [28] T. Duretz, D.A. May, T.V. Gerya, and P.J. Tackley. Discretization errors and free surface stabilisation in the finite difference and marker-in-cell method for applied geodynamics: A numerical study. *Geochem. Geophys. Geosyst.*, 12(Q07004), 2011.
- [29] R. Eid. Higher order isoparametric finite element solution of Stokes flow . *Applied Mathematics and Computation*, 162:1083–1101, 2005.
- [30] E. Erturk. Discussions on Driven Cavity Flow. *Int. J. Num. Meth. Fluids*, 60:275–294, 2009.
- [31] P.J. Frey and P.-L. George. *Mesh generation*. Hermes Science, 2000.
- [32] P. Fullsack. An arbitrary Lagrangian-Eulerian formulation for creeping flows and its application in tectonic models. *Geophys. J. Int.*, 120:1–23, 1995.
- [33] M. Gerbault, A.N.B. Poliakov, and M. Daignieres. Prediction of faulting from the theories of elasticity and plasticity: what are the limits? *Journal of Structural Geology*, 20:301–320, 1998.
- [34] Taras Gerya. *Numerical Geodynamic Modelling*. Cambridge University Press, 2010.
- [35] T.V. Gerya, D.A. May, and T. Duretz. An adaptive staggered grid finite difference method for modeling geodynamic Stokes flows with strongly variable viscosity. *Geochem. Geophys. Geosyst.*, 14(4), 2013.
- [36] G.H. Golub and C.F. van Loan. *Matrix Computations, 4th edition*. John Hopkins University Press, 2013.
- [37] P.M. Gresho and R.L. Sani. *Incompressible flow and the Finite Element Method, vol II*. John Wiley and Sons, Ltd, 2000.
- [38] D. Griffiths and D. Silvester. Unstable modes of the q1-p0 element. Technical Report 257, University of Manchester/UMIST, 1994.
- [39] M. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice and Algorithms*. Academic, Boston, 1989.
- [40] T. Heister, J. Dannberg, R. Gassmöller, and W. Bangerth. High Accuracy Mantle Convection Simulation through Modern Numerical Methods. II: Realistic Models and Problems. *Geophys. J. Int.*, 210(2):833–851, 2017.

- [41] T.J.R. Hughes. *The Finite Element Method. Linear Static and Dynamic Finite Element Analysis*. Dover Publications, Inc., 2000.
- [42] T.J.R. Hughes, W.K. Liu, and A. Brooks. Finite element analysis of Incompressible viscous flows by the penalty function formulation. *J. Comp. Phys.*, 30:1–60, 1979.
- [43] Hoon Huh, Choong Ho Lee, and Wei H. Yang. A general algorithm for plastic flow simulation by finite element limit analysis. *International Journal of Solids and Structures*, 36:1193–1207, 1999.
- [44] J. Ita and S.D. King. Sensitivity of convection with an endothermic phase change to the form of governing equations, initial conditions, boundary conditions, and equation of state. *J. Geophys. Res.*, 99(B8):15,919–15,938, 1994.
- [45] L. Jolivet, P. Davy, and P. Cobbold. Right-lateral shear along the Northwest Pacific margin and the India-Eurasia collision. *Tectonics*, 9(6):1409–1419, 1990.
- [46] L.M. Kachanov. *Fundamentals of the Theory of Plasticity*. Dover Publications, Inc., 2004.
- [47] S. King, C. Lee, P. van Keeken, W. Leng, S. Zhong, E. Tan, N. Tosi, and M. Kameyama. A community benchmark for 2D Cartesian compressible convection in the Earths mantle. *Geophy. J. Int.*, 180:7387, 2010.
- [48] J.R. Koseff and R.L. Street. The Lid-Driven Cavity Flow: A Synthesis of Qualitative and Quantitative Observations. *J. Fluids Eng*, 106:390–398, 1984.
- [49] M. Kronbichler, T. Heister, and W. Bangerth. High accuracy mantle convection simulation through modern numerical methods . *Geophy. J. Int.*, 191:12–29, 2012.
- [50] R. Lee, P. Gresho, and R. Sani. Smoothing techniques for certain primitive variable solutions of the Navier-Stokes equations. . *Int. Journal for Numerical Methods in Engineering*, 14:1785–1804, 1979.
- [51] W. Leng and S. Zhong. Viscous heating, adiabatic heating and energetic consistency in compressible mantle convection. *Geophy. J. Int.*, 173:693–702, 2008.
- [52] W. Leng and S. Zhong. Implementation and application of adaptive mesh refinement for thermochemical mantle convection studies. *Geochem. Geophys. Geosyst.*, 12(4), 2011.
- [53] X. Liu and S. Zhong. Analyses of marginal stability, heat transfer and boundary layer properties for thermal convection in a compressible fluid with infinite Prandtl number. *Geophy. J. Int.*, 194:125–144, 2013.
- [54] C. Loiselet, J. Braun, L. Husson, C. Le Carlier de Veslud, C. Thieulot, P. Yamato, and D. Grujic. Subducting slabs: Jellyfishes in the Earth’s mantle. *Geochem. Geophys. Geosyst.*, 11(8):doi:10.1029/2010GC003172, 2010.
- [55] D.S. Malkus and T.J.R. Hughes. Mixed finite element methods - reduced and selective integration techniques: a unification of concepts. *Comput. Meth. Appl. Mech. Eng.*, 15:63–81, 1978.
- [56] L.E. Malvern. *Introduction to the mechanics of a continuous medium*. Prentice-Hall, Inc., 1969.
- [57] A. Mizukami. A mixed finite element method for boundary flux computation. *Computer Methods in Applied Mechanics and Engineering*, 57:239–243, 1986.
- [58] P. Molnar and P. Tapponnier. Relation of the tectonics of eastern China to the India-Eurasia collision: Application of the slip-line field theory to large-scale continental tectonics. *Geology*, 5:212–216, 1977.
- [59] L. Moresi, S. Quenette, V. Lemiale, C. Mériaux, B. Appelbe, and H.-B. Mühlhaus. Computational approaches to studying non-linear dynamics of the crust and mantle. *Phys. Earth. Planet. Inter.*, 163:69–82, 2007.
- [60] M. Nettlesheim, T.A. Ehlers, D.M. Whipp, and A. Koptev. The influence of upper-plate advance and erosion on overriding plate deformation in orogen syntaxes. *Solid Earth*, 9:1207–1224, 2018.

- [61] C. O'Neill, L. Moresi, D. Müller, R. Albert, and F. Dufour. Ellipsis 3D: a particle-in-cell finite element hybrid code for modelling mantle convection and lithospheric deformation. *Computers and Geosciences*, 32:1769–1779, 2006.
- [62] G. Peltzer and P. Tapponnier. Formation and evolution of strike-slip faults, rifts, and basins during the india-asia collision: an experimental approach. *J. Geophys. Res.*, 93(B12):15085–15177, 1988.
- [63] A. Pinelli and A. Vacca. Chebyshev collocation method and multidomain decomposition for the incompressible Navier-Stokes equations. *International Journal for numerical methods in fluids*, 18:781–799, 1994.
- [64] A.E. Pusok, B.J.P. Kaus, and A.A. Popov. On the Quality of Velocity Interpolation Schemes for Marker-in-Cell Method and Staggered Grids. *Pure and Applied Geophysics*, pages doi:10.1007/s00024-016-1431-8, 2016.
- [65] T. Rabczuk, P.M.A. Areias, and T. Belytschko. A simplified mesh-free method for shear bands with cohesive surfaces . *Int. J. Num. Meth. Eng.*, 69:993–1021, 2007.
- [66] J.N. Reddy. On penalty function methods in the finite element analysis of flow problems. *Int. J. Num. Meth. Fluids*, 2:151–171, 1982.
- [67] J. Revenaugh and B. Parsons. Dynamic topography and gravity anomalies for fluid layers whose viscosity varies exponentially with depth. *Geophysical Journal of the Royal Astronomical Society*, 90(2):349–368, 1987.
- [68] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.
- [69] R.L. Sani, P.M. Gresho, R.L. Lee, and D.F. Griffiths. The cause and cure (?) of the spurious pressures generated by certain FEM solutions of the incompressible Navier-Stokes equations: part 1. *Int. J. Num. Meth. Fluids*, 1:17–43, 1981.
- [70] R.L. Sani, P.M. Gresho, R.L. Lee, D.F. Griffiths, and M. Engelmann. The cause and cure (?) of the spurious pressures generated by certain fem solutions of the incompressible navier-stokes equations: part 2. *Int. J. Num. Meth. Fluids*, 1:171–204, 1981.
- [71] S.M. Schmalholz. A simple analytical solution for slab detachment. *Earth Planet. Sci. Lett.*, 304:45–54, 2011.
- [72] H. Schmeling, A.Y. Babeyko, A. Enns, C. Faccenna, F. Funiciello, T. Gerya, G.J. Golabek, S. Grigull, B.J.P. Kaus, G. Morra, S.M. Schmalholz, and J. van Hunen. A benchmark comparison of spontaneous subduction models - Towards a free surface. *Phys. Earth. Planet. Inter.*, 171:198–223, 2008.
- [73] D.W. Schmid and Y.Y. Podlachikov. Analytical solutions for deformable elliptical inclusions in general shear. *Geophys. J. Int.*, 155:269–288, 2003.
- [74] G. Schubert, D.L. Turcotte, and P. Olson. *Mantle Convection in the Earth and Planets*. Cambridge University Press, 2001.
- [75] M. Spiegelman, D.A. May, and C. Wilson. On the solvability of incompressible Stokes with viscoplastic rheologies in geodynamics. *Geochem. Geophys. Geosyst.*, 17:2213–2238, 2016.
- [76] J. Suckale, J.-C. Nave, and B.H. Hager. It takes three to tango: 1. Simulating buoyancy-driven flow in the presence of large viscosity contrasts. *J. Geophys. Res.*, 115(B07409), 2010.
- [77] P. Tackley. *Three-dimensional models of mantle convection: Influence of phase transitions and temperature-dependent viscosity*. PhD thesis, California Institute of Technology, 1994.
- [78] E. Tan and M. Gurnis. Compressible thermochemical convection and application to lower mantle structures. *J. Geophys. Res.*, 112(B06304), 2007.
- [79] Paul Tapponnier and Peter Molnar. Slip-line field theory and large-scale continental tectonics. *Nature*, 264:319–324, November 1976.

- [80] M. Thielmann, D.A. May, and B.J.P. Kaus. Discretization errors in the Hybrid Finite Element Particle-In-Cell Method. *Pure and Applied Geophysics*, 2014.
- [81] C. Thieulot. FANTOM: two- and three-dimensional numerical modelling of creeping flows for the solution of geological problems. *Phys. Earth. Planet. Inter.*, 188:47–68, 2011.
- [82] C. Thieulot. GHOST: Geoscientific Hollow Sphere Tessellation. *Solid Earth*, 9(1–9), 2018.
- [83] C. Thieulot, P. Fullsack, and J. Braun. Adaptive octree-based finite element analysis of two- and three-dimensional indentation problems. *J. Geophys. Res.*, 113:B12207, 2008.
- [84] J.F. Thompson, B.K. Soni, and N.P. Weatherill. *Handbook of grid generation*. CRC press, 1998.
- [85] R.A. Trompert and U. Hansen. On the Rayleigh number dependence of convection with a strongly temperature-dependent viscosity. *Physics of Fluids*, 10(2):351–360, 1998.
- [86] D.L. Turcotte and G. Schubert. *Geodynamics, 2nd edition*. Cambridge, 2012.
- [87] P.E. van Keken, S.D. King, H. Schmeling, U.R. Christensen, D. Neumeister, and M.-P. Doin. A comparison of methods for the modeling of thermochemical convection. *J. Geophys. Res.*, 102(B10):22,477–22,495, 1997.
- [88] D.M. Whipp, C. Beaumont, and J. Braun. Feeding the aneurysm: Orogen-parallel mass transport into Nanga Parbat and the western Himalayan syntaxis. *J. Geophys. Res.*, 119:doi:10.1002/2013JB010929, 2014.
- [89] S.D. Willett. Dynamic and kinematic growth and change of a coulomb wedge. In K.R. McClay, editor, *Thrust Tectonics*, pages 19–31. Chapman and Hall, 1992.
- [90] H. Xing, W. Yu, and J. Zhang. In *Advances in Geocomputing, Lecture Notes in Earth Sciences*. Springer-Verlag, Berlin Heidelberg, 2009.
- [91] P. Yamato, L. Husson, J. Braun, C. Loiselet, and C. Thieulot. Influence of surrounding plates on 3D subduction dynamics. *Geophys. Res. Lett.*, 36(L07303):doi:10.1029/2008GL036942, 2009.
- [92] X. Yu and F. Tin-Loi. A simple mixed finite element for static limit analysis. *Computers and Structures*, 84:1906–1917, 2006.
- [93] S. Zhong. Analytic solutions for Stokes flow with lateral variations in viscosity. *Geophys. J. Int.*, 124:18–28, 1996.
- [94] S. Zhong, M. Gurnis, and G. Hulbert. Accurate determination of surface normal stress in viscous flow from a consistent boundary flux method. *Phys. Earth. Planet. Inter.*, 78:1–8, 1993.
- [95] S. Zhong, A. McNamara, E. Tan, L. Moresi, and M. Gurnis. A benchmark study on mantle convection in a 3-D spherical shell using CITCOMS. *Geochem. Geophys. Geosyst.*, 9(10), 2008.
- [96] O. Zienkiewicz and S. Nakazawa. The penalty function method and its application to the numerical solution of boundary value problems. *The American Society of Mechanical Engineers*, 51, 1982.
- [97] O.C. Zienkiewicz, M. Huang, and M. Pastor. Localization problems in plasticity using finite elements with adaptive remeshing. *International Journal for Numerical and Analytical Methods in Geomechanics*, 19:127–148, 1995.
- [98] O.C. Zienkiewicz, C. Humpheson, and R.W. Lewis. Associated and non-associated visco-plasticity and plasticity in soil mechanics . *Géotechnique*, 25(4):671–689, 1975.

# Index

- $P_m \times P_n$ , 12
- $P_m \times P_{-n}$ , 12
- $Q_1 \times P_0$ , 53, 60, 62, 68, 75, 79, 95, 102, 119, 121
- $Q_1$ , 13, 17
- $Q_2 \times Q_1$ , 12, 81, 86
- $Q_2$ , 13, 19
- $Q_3 \times Q_2$ , 89
- $Q_3$ , 14, 20
- $Q_m \times P_{-n}$ , 12
- $Q_m \times Q_n$ , 12
- $Q_m \times Q_{-n}$ , 12
- analytical solution, 53, 62, 79, 86, 89, 92, 95, 102
- arithmetic mean, 75
- basis functions, 17
- Boussinesq, 8
- bubble function, 12
- bulk modulus, 97
- bulk viscosity, 7
- CG, 45
- checkerboard, 40
- cohesion, 39
- Compressed Sparse Column, 32
- Compressed Sparse Row, 32
- compressibility, 97
- compressible flow, 102
- conforming element, 12
- conjugate gradient, 45
- connectivity array, 36
- convex polygon, 35
- CSC, 32
- CSR, 32
- divergence free, 22
- Drucker-Prager, 39
- dynamic viscosity, 7
- Gauss quadrature, 10
- geometric mean, 75
- harmonic mean, 76
- incompressible flow, 62, 68, 75, 79, 86, 89, 92, 95, 119, 121
- isothermal, 53, 60, 62, 68, 75, 79, 86, 89, 92, 95, 119, 121
- isoviscous, 53, 60, 79, 86, 89, 95, 102
- Legendre polynomial, 10
- method of manufactured solutions, 28
- midpoint rule, 9
- mixed formulation, 79, 86, 89, 92, 95, 102, 119, 121
- MMS, 28
- Newton-Cotes, 10
- non-conforming element, 12
- non-isoviscous, 62, 68, 75, 92
- nonconforming  $Q_1 \times P_0$ , 92
- nonlinear, 121
- nonlinear rheology, 68
- numerical benchmark, 119
- particle-in-cell, 75
- penalty formulation, 22, 53, 60, 62, 68, 75
- piecewise, 12
- pressure normalisation, 82
- pressure scaling, 42
- pressure smoothing, 40, 79, 92, 95, 102
- quadrature, 10
- rectangle rule, 9
- Schur complement, 45
- second viscosity, 7
- SPD, 45
- static condensation, 49
- strong form, 22
- structured grid, 35
- tensor invariant, 38
- thermal expansion, 97
- trapezoidal rule, 9
- unstructured grid, 35
- Viscosity Rescaling Method, 39
- von Mises, 39
- VRM, 39
- weak form, 22, 23
- work against gravity, 100

## Notes

[ ] insert here figure . . . . .	9
[ ] insert here figure . . . . .	9
[ ] list codes which use this approach . . . . .	22
[ ] add more examples coming from geo . . . . .	35
[ ] produce drawing of node numbering . . . . .	37
[ ] insert here the rederivation 2.1.1 of spmw16 . . . . .	39
[ ] produce figure to explain this . . . . .	41
[ ] link to proto paper . . . . .	41
[ ] link to least square and nodal derivatives . . . . .	41