First order system least squares (FOSLS) explained

A. L. Codd
Department of Applied Mathematics
University of Colorado,
Boulder, CO 80309

December 8, 1999

introduction

Have you every wondered what FOSLS is? And when ***Steve*** and ***Tom*** start talking about FOSLS and Formal Normals and theoretical convergence rates do you just nod your head wisely and hope that they don't ask you any questions? Then this document is for you. It's incomplete, comprised only of the things that we have puzzled over, so there is no guarantee that it will answer any of your questions. For those unanswered questions you'll just have to read one of the many FOSLS papers or ask ***Tom*** or ***Steve***.

In this document, we explain how to apply FOSLS to a specific problem. We start with the standard Stokes equations, convert them to an appropriate first order system, then create the **FUNCTIONAL**. Using the **FUNCTIONAL**, we then demonstrate, but don't prove, (see FOSLS for the Stokes Equations with application to Linear elasticity) product H^1 ellipticity and explain what this means for finite elements and Multigrid. We follow this with a tiny spiel about norms but from the rest of the document it's easy to see that L^2 norms **RULE**.

It's at this point that we move to theoretical world with a discussion of the **FORMAL NORMAL** and **THEORETICAL CONVERGENCE RATES**. We briefly discuss adjoints and Gauss Seidel and show why ***Steve*** and ***Tom*** throw the convergence rate $\frac{1}{2}$ around.

After this small excursion to theoretical world, we return to the "real" world to discuss how to minimise the **FUNCTIONAL**. We create the **WEAK FORM**, which is actually what ends up being solved. It's at this point that the document stops. You will just have to do the subsequent steps all on your own.

The next step in the process is to select a finite element system and use it to discretise the **WEAK FORM**, to obtain a matrix system of equations. It is this system $\mathbf{Ax} = \mathbf{b}$ that is sent off to multigrid (or some other solver).

Stokes equations in 2D

We can use Stokes equations to model steady state, incompressible, low speed flow or linear elasticity, (with different meaning of the vector variables, velocity for the fluid case and displacement for the elasticity case).

dependent variables

There are three dependent variables, velocity in the x direction, u, velocity in the y direction, v, and pressure, p. We can write the velocity as a vector, $\mathbf{u} = \begin{pmatrix} u & v \end{pmatrix}^t$.

equations

The equations are derived using conservation laws, in particular, conservation of momentum and conservation of mass. We can either write the equations using vector or component notation. We initially display both.

Consider a region Ω, bounded, open, connected domain in $\mathbb{R}^2$ with Lipschitz boundary Γ. The domain equations are

$$\begin{array}{cc} \text{vector notation} & \text{component notation} \\ -\Delta \mathbf{u} + \nabla p = \mathbf{f}, & -(u_{xx} + u_{yy}) + p_x = f, \\ & -(v_{xx} + v_{yy}) + p_y = g, \\ \nabla \cdot \mathbf{u} = 0. & u_x + v_y = 0. \end{array} \tag{1}$$

boundary conditions

These equations are imposed on the boundary Γ. We assume Dirichlet boundary conditions.

$$\begin{array}{cc} \text{vector notation} & \text{component notation} \\ \mathbf{u} = \Phi. & u = \phi, \\ & v = \psi. \end{array} \tag{2}$$

To ensure uniqueness we impose an extra condition on p

$$\int_\Omega p dz = 0. \tag{3}$$

If p satisfies equations (1), (2) then so does $p + c$ where c is a constant. Condition (3) removes the constant.

first order system

The first step in the FOSLS process is to convert the second order system (1), (2) to a first order system.

dependent variables

We add four new variables to the three we already have:

$$\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} U & V \\ W & Z \end{pmatrix}, \qquad p.$$

equations

We add a definition equation for the four new dependent variables and the original second order equation is rewritten as a first order equation.

vector notation	component notation
$\mathbf{U} = \nabla \mathbf{u},$	$U = u_x,$
	$W = u_y,$
	$V = v_x,$
	$Z = v_y,$
$-\nabla \cdot \mathbf{U} + \nabla p = \mathbf{f},$	$-(U_x + W_y) + p_x = f,$
	$-(V_x + Z_y) + p_y = g,$
$\nabla \cdot \mathbf{u} = 0.$	$u_x + v_y = 0. \qquad (4)$

boundary conditions

The boundary conditions are unchanged.

vector notation	component notation
$\mathbf{u} = \Phi.$	$u = \phi,$
	$v = \psi. \qquad (5)$

What is the least squares functional?

The functional is created by summing the least squares norms of the residuals of equations (4) and (5).

$$G_0(\mathbf{u}, \mathbf{U}, p; \mathbf{f}, \Phi) = \|\mathbf{U} - \nabla \mathbf{u}\|_{0,\Omega}^2 + \| - \nabla \cdot \mathbf{U} + \nabla p - \mathbf{f}\|_{0,\Omega}^2 + \|\nabla \cdot \mathbf{u}\|_{0,\Omega}^2 + \|\mathbf{u} - \Phi\|_{\frac{1}{2},\Gamma}^2. \quad (6)$$

Why not minimise this functional?

Our mission is to

divide and conquer.

We try to set up a functional so that the variables are essentially decoupled and we basically have scalar Poisson problems (product H^1 ellipticity, to be explained later). This decoupling means we get optimal finite element approximation and optimal multigrid convergence. The functional $G_0(\mathbf{u}, \mathbf{U}, p; \mathbf{f}, \Phi)$, is not product H^1 elliptic.

Another way of thinking about the problem is that this system does not expose all the possible errors. Let's say $\mathbf{u}^*$, $\mathbf{U}^*$, and p^* are the actual solutions to our first order system. Let's consider that we add an error $\mathbf{E}$ to $\mathbf{U}^*$. Then the functional (6) is

$$\begin{aligned} G_0(\mathbf{u}^*, \mathbf{U}^* + \mathbf{E}, p^*; \mathbf{f}, \Phi) =& \|\mathbf{U}^* + \mathbf{E} - \nabla \mathbf{u}^*\|_{0,\Omega}^2 + \| - \nabla \cdot \mathbf{U}^* - \nabla \cdot \mathbf{E} + \nabla p^* - \mathbf{f}\|_{0,\Omega}^2 \\ &+ \|\nabla \cdot \mathbf{u}^*\|_{0,\Omega}^2 + \|\mathbf{u}^* - \Phi\|_{\frac{1}{2},\Gamma}^2 \\ =& \|\mathbf{E}\|_{0,\Omega}^2 + \|\nabla \cdot \mathbf{E}\|_{0,\Omega}^2. \end{aligned} \quad (7)$$

A good functional means that if there is error in the solution then the functional is large. If $\mathbf{E}$ is divergence free but oscillatory, i.e. $\nabla \cdot \mathbf{E} = \mathbf{0}$ and $\mathbf{E}$ has large derivative values, then this functional won't be large. Taking I or $\nabla \cdot$ of $\mathbf{E}$, as we have done above, doesn't give a large value to oscillatory $\mathbf{E}$ and the error $\mathbf{E}$ is hidden from relaxation. We need to add something to our functional that is consistent with our Stokes system yet exposes this error. Something that works is $\|\nabla \times \mathbf{U}\|_{0,\Omega}^2$.

And a final suggestion for why this system is just no good is that there is a continuity relationship between U and W, and V and Z, i.e., $U_y = W_x$ and $V_y = Z_x$ that we basically ignore in this formulation.

augmented first order system

What do we add to (4) and (5), to make a happy first order system? First of all, we note that $\mathrm{tr}\mathbf{U} = \nabla \cdot \mathbf{u}$, therefore we add a first order form of $\mathrm{tr}\mathbf{U} = 0$ to the system, i.e., $\nabla(\mathrm{tr}\mathbf{U}) = \mathbf{0}$. Secondly, we note that $\mathbf{U} = \nabla\mathbf{u} \Rightarrow \nabla \times \mathbf{U} = \mathbf{0}$ and we add this curl term to the system thus exposing those nasty divergence free errors. This also has the effect of imposing the continuity requirements for the $\mathbf{U}$ terms.

equations

The new system of equations for the domain Ω is

$$
\begin{array}{rr}
\text{vector notation} & \text{components notation} \\
\mathbf{U} = \nabla\mathbf{u}, & U = u_x, \\
 & W = u_y, \\
 & V = v_x, \\
 & Z = v_y, \\
-\nabla \cdot \mathbf{U} + \nabla p = \mathbf{f}, & -(U_x + W_y) + p_x = f, \\
 & -(V_x + Z_y) + p_y = g, \\
\nabla \cdot \mathbf{u} = 0, & u_x + v_y = 0, \\
\nabla \mathrm{tr}\mathbf{U} = \mathbf{0}, & U_x + Z_x = 0, \\
 & U_y + Z_y = 0, \\
\nabla \times \mathbf{U} = 0. & U_y - W_x = 0, \\
 & V_y - Z_x = 0.
\end{array} \tag{8}
$$

boundary conditions

Adding the extra equations for $\mathbf{U}$ to the system means that none of the relationships between the $\mathbf{U}$ terms and the other variables are applied second hand. This is like saying the $\mathbf{U}$ is just as important as $\mathbf{u}$ and p. It is just as important to make sure that the boundary conditions are not applied to $\mathbf{U}$ in a second hand way. We have Dirichlet boundary conditions, therefore we know the tangential components of $\mathbf{U}$ along the boundary Γ and we add this to the boundary conditions,

$$
\begin{array}{rr}
\text{vector notation} & \text{components notation} \\
\mathbf{u} = \Phi, & u = \phi, \\
 & v = \psi, \\
\mathbf{n} \times \mathbf{U} = \mathbf{n} \times \nabla\Phi. & -n_2 U + n_1 W = -n_2\phi_x + n_1\phi_y, \\
 & -n_2 V + n_1 Z = -n_2\psi_x + n_1\psi_y.
\end{array} \tag{9}
$$

In the case where Γ is horizontal, we obviously know the x derivatives of ϕ and ψ, i.e., $U = u_x = \frac{d\phi}{dx}$ and $V = v_x = \frac{d\psi}{dx}$.

functional

Now we can create the least squares functional using (8) and (9).

$$G(\mathbf{u},\mathbf{U},p;\mathbf{f},\Phi) = \|\mathbf{U}-\nabla\mathbf{u}\|_{0,\Omega}^2 + \|-\nabla\cdot\mathbf{U}+\nabla p-\mathbf{f}\|_{0,\Omega}^2 + \|a\nabla\cdot\mathbf{u}\|_{0,\Omega}^2 + \|a\nabla\mathrm{tr}\mathbf{U}\|_{0,\Omega}^2 + \|\nabla\times\mathbf{U}\|_{0,\Omega}^2 + \|\mathbf{u}-\Phi\|_{\frac{1}{2},\Gamma}^2 + \|\mathbf{n}\times\mathbf{U}-\mathbf{n}\times\nabla\Phi\|_{\frac{1}{2},\Gamma}^2. \tag{10}$$

A new and exciting possibility has just opened up now that we've added all those extra equations. We can actually think of this as two separate problems. We can do this **only** because $\mathbf{U}$ is not getting any information (relationships and boundary conditions) second hand and it can stand alone. In the first problem, we delete all the equations with $\mathbf{u}$ in them and solve just for $\mathbf{U}$ and p. In the second problem we use the solutions $\mathbf{U}$ and p from the first problem, delete all the equations that don't have $\mathbf{u}$ in them, and then solve for $\mathbf{u}$. This is called a two stage algorithm.

We've also added a scale factor, a, in front of the continuity equations. The reason for this will be revealed later.

norms L^2,H^1, H^{-1}, $H^{\frac{1}{2}}$?

The L^2 norm is

$$\|f\|_{0,\Omega} = \left(\int_\Omega |f|^2\right)^{\frac{1}{2}}.$$

The H^1 norm is

$$\|f\|_{1,\Omega} = \left(\|f\|_{0,\Omega}^2 + \|f_x\|_{0,\Omega}^2 + \|f_y\|_{0,\Omega}^2\right)^{\frac{1}{2}}.$$

The H^{-1} norm is

$$\|f\|_{-1,\Omega} = \sup_{0\neq g\in H^1} \frac{\int_\Omega fg}{\|g\|_{1,\Omega}}.$$

The $H^{-\frac{1}{2}}$ norm is

$$\|r\|_{\frac{1}{2},\Gamma_{\boldsymbol{\xi}}} = \inf\{\|g\|_{1,\Omega_{\boldsymbol{\xi}}} : g \in H^1(\Omega_{\boldsymbol{\xi}}),\ \text{trace } g = r \text{ on } \Gamma_{\boldsymbol{\xi}}\}.$$

In theory, we use L^2 norms for the domain and $H^{\frac{1}{2}}$ norms for the boundary. In practice we use L^2 norms for the domain and L^2 norms scaled by $\frac{1}{h}$, for the boundary.

In the Stokes paper, they create a functional using H^{-1} norms. Apparently that's like doing a L^2 norm on the equations after they have been multiplied by $(-\Delta)^{-\frac{1}{2}}$.

Let's just look at the second term in functional (10) and assume we want a H^{-1} norm. Then,

$$\|-\nabla\cdot\mathbf{U} + \nabla p - \mathbf{f}\|_{-1,\Omega}^2 \approx \|(-\Delta)^{-\frac{1}{2}}(-\nabla\cdot\mathbf{U} + \nabla p - \mathbf{f})\|_{0,\Omega} \tag{11}$$

This looks really quite bizarre. We can say (loosely) that that $(-\Delta)^{-1}$ is like anti-differentiating twice, that is, like integrating $\mathbf{f}$ twice. This means that $(-\Delta)^{-\frac{1}{2}}$ is like integrating $\mathbf{f}$ once. So, we **don't** need $\mathbf{f} \in L^2$ we just need $(-\Delta)^{-\frac{1}{2}}\mathbf{f} \in L^2$. This means that they, ***Tom*** and ***Steve***, can consider problems where $\mathbf{f} \in H^{-1}$. A bonus from using inverse norms is that they don't need H^2 regularity of the original problem to get the H^1 ellipticity result. See Stokes for details.

Product H^1 ellipticity

We define the norm

$$V_0 = \{\mathbf{V} \in H^1(\Omega)^{n^2} : \mathbf{n} \times \mathbf{V} = \mathbf{0} \text{ on } \Gamma\}. \tag{12}$$

We define the regularity bound

$$\|\mathbf{u}\|_{2,\Omega} + \|p\|_{1,\Omega} \leq C\|\mathbf{f}\|_{0,\Omega}. \tag{13}$$

In Stokes, they prove the H^1 ellipticity result

Theorem 1. *Assume that the domain Ω is a bounded convex polyhedron or has $C^{1,1}$ boundary and that the regularity bound (13) holds. Then there exists a constant C such that for any $(\mathbf{U}, \mathbf{u}, p) \in V_0 \times H_0^1(\Omega)^n \times (H^1(\Omega)/\mathbb{R})$, we have*

$$\frac{1}{C}\left(\|\mathbf{U}\|_{1,\Omega}^2 + \|\mathbf{u}\|_{1,\Omega}^2 + \|p\|_{1,\Omega}^2\right) \leq G(\mathbf{u}, \mathbf{U}, p; \mathbf{0}, \mathbf{0}) \tag{14}$$

and

$$G(\mathbf{u}, \mathbf{U}, p; \mathbf{0}, \mathbf{0}) \leq C\left(\|\mathbf{U}\|_{1,\Omega}^2 + \|\mathbf{u}\|_{1,\Omega}^2 + \|p\|_{1,\Omega}^2 t\right) \tag{15}$$

Why do we care?

Let's have a close look at our functional (10). Now let's say $\mathbf{u}^*$, $\mathbf{U}^*$ and p^* are the actual solutions, so they satisfy all the equations (8), (9). Then we can write

$$\begin{aligned}
G(\mathbf{u}, \mathbf{U}, p; \mathbf{f}, \Phi) =& \|\mathbf{U} - \nabla\mathbf{u} - (\mathbf{U}^* - \nabla\mathbf{u}^*)\|_{0,\Omega}^2 + \| - \nabla \cdot \mathbf{U} + \nabla p - \mathbf{f} - (\nabla \cdot \mathbf{U}^* + \nabla p^* - \mathbf{f})\|_{0,\Omega}^2 \\
&+ a^2\|\nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^*\|_{0,\Omega}^2 + a^2\|\nabla \mathrm{tr}\mathbf{U} - \nabla \mathrm{tr}\mathbf{U}^*\|_{0,\Omega}^2 + \|\nabla \times \mathbf{U} - \nabla \times \mathbf{U}^*\|_{0,\Omega}^2 \\
&+ \|\mathbf{u} - \Phi - (\mathbf{u}^* - \Phi)\|_{\frac{1}{2},\Gamma}^2 + \|\mathbf{n} \times \mathbf{U} - \mathbf{n} \times \nabla\Phi - (\|\mathbf{n} \times \mathbf{U}^* - \mathbf{n} \times \nabla\Phi)\|_{\frac{1}{2},\Gamma}^2 \\
=& \|\mathbf{U} - \mathbf{U}^* - \nabla(\mathbf{u} - \mathbf{u}^*)\|_{0,\Omega}^2 + \| - \nabla \cdot (\mathbf{U} - \mathbf{U}^*) + \nabla(p - p^*)\|_{0,\Omega}^2 \\
&+ a^2\|\nabla \cdot (\mathbf{u} - \mathbf{u}^*)\|_{0,\Omega}^2 + a^2\|\nabla \mathrm{tr}(\mathbf{U} - \mathbf{U}^*)\|_{0,\Omega}^2 \\
&+ \|\nabla \times (\mathbf{U} - \mathbf{U}^*)\|_{0,\Omega}^2 + \|\mathbf{u} - \mathbf{u}^*\|_{\frac{1}{2},\Gamma}^2 + \|\mathbf{n} \times (\mathbf{U} - \mathbf{U}^*)\|_{\frac{1}{2},\Gamma}^2 \\
=& G(\mathbf{u} - \mathbf{u}^*, \mathbf{U} - \mathbf{U}^*, p - p^*; \mathbf{0}, \mathbf{0})
\end{aligned} \tag{16}$$

So this means that $G(\mathbf{errors}; \mathbf{0})$ is equivalent to a norm that we like. We like the H^1 norm because it includes values and derivatives.

What else does this mean?

It means we get **optimal finite element approximation** and **optimal MG convergence**. It means that the individual variables are essentially decoupled.

optimal finite element approximation

Let's consider the finite element approximation where the finite element space is composed of piecewise polynomials of degree 1. Say we are approximating $g(x)$ by $g_h(x) = g(x_0) + xg'(x_0)$. We also know that $g(x) = g(x_0) + xg'(x_0) + \frac{1}{2}x^2 g''(\xi)$. Consequently, we get the bound

$$\begin{aligned}\|g(x) - g_h(x)\|_{0,\Omega} =& \|g(x_0) + xg'(x_0) + \tfrac{1}{2}x^2 g''(\xi) - g(x_0) - xg'(x_0)\|_{0,\Omega} \\ =& \|\tfrac{1}{2}x^2 g''(\xi)\|_{0,\Omega} \\ \leq& Ch^2 \|g''(x)\|_{0,\Omega}\end{aligned} \tag{17}$$

so

$$\|g(x) - g_h(x)\|_{1,\Omega} \leq Ch\|g\|_{2,\Omega}. \tag{18}$$

The complete finite element bound they get in various FOSLS papers is

$$\|\mathbf{U} - \mathbf{U}_h\|_{1,\Omega} + \|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega} + \|p - p_h\|_{1,\Omega} \leq Ch\left(\|\mathbf{U}\|_{2,\Omega} + \|\mathbf{u}\|_{2,\Omega} + \|p\|_{2,\Omega}\right). \tag{19}$$

We don't show here **optimal MG convergence**. It can be found and deciphered from a number of the FOSLS papers.

formal normal

The formal normal is a **theoretical** tool that can be used to estimate the **ideal** convergence factors of the algorithm. It can be used to determine the best functional, i.e. one with the most decoupling between the variables and the best **ideal** convergence rate. The whole system can be analysed in one go, but it adds to the complexity. We just show the general idea by considering the first stage of the two stage algorithm.

The formal normal assumes imposed zero Dirichlet boundary conditions. Therefore ignoring the boundary conditions, the functional we minimise,

$$F(\mathbf{U}, p; \mathbf{f}) = \| -\nabla \cdot \mathbf{U} + \nabla p - \mathbf{f}\|^2_{0,\Omega} + \|a\nabla \mathrm{tr}\mathbf{U}\|^2_{0,\Omega} + \|\nabla \times \mathbf{U}\|^2_{0,\Omega}, \tag{20}$$

which can be written

$$F(\mathbf{w}; \mathbf{f}) = \|\mathcal{L}\mathbf{w} - \mathbf{f}\|^2_{0,\Omega}, \tag{21}$$

where $\mathbf{w} = \begin{pmatrix} U & W & V & Z & p \end{pmatrix}^t$ and

$$\mathcal{L} = \begin{pmatrix} -\partial_x & -\partial_y & 0 & 0 & \partial_x \\ 0 & 0 & -\partial_x & -\partial_y & \partial_y \\ a\partial_x & 0 & 0 & a\partial_x & 0 \\ a\partial_y & 0 & 0 & a\partial_y & 0 \\ \partial_y & -\partial_x & 0 & 0 & 0 \\ 0 & 0 & \partial_y & -\partial_x & 0 \end{pmatrix} \tag{22}$$

We assume we have zero right hand sides, i.e., $\mathbf{f} = \mathbf{0}$. We also assume $\mathbf{w}$ has compact support on Ω, i.e. $\mathbf{w} = \mathbf{0}$ on Γ. To determine the formal normal we need to understand the formal adjoint. The formal adjoint of ∂_x is ∂_x^* and it satisfies $(\partial_x \mathbf{v}, \mathbf{w}) = (\mathbf{v}, \partial_x^* \mathbf{w})$. We use integration by parts, remembering that $\mathbf{w} = \mathbf{0}$ on Γ:

$$(\partial_x \mathbf{v}, \mathbf{w}) = \int_\Omega \frac{\partial \mathbf{v}}{\partial x} \mathbf{w} = [\mathbf{v}\mathbf{w}]_\Gamma - \int_\Omega \mathbf{v} \frac{\partial \mathbf{w}}{\partial x} = (\mathbf{v}, -\partial_x \mathbf{w}) \tag{23}$$

so $\partial_x^* = -\partial_x$, and in a similar way we get $\partial_y^* = -\partial_y$.

We can write our functional

$$\begin{aligned} F(\mathbf{w}; \mathbf{0}) = \|\mathcal{L}\mathbf{w}\|^2_{0,\Omega} &= (\mathcal{L}\mathbf{w}, \mathcal{L}\mathbf{w}) \qquad (24) \\ &= (\mathcal{L}^*\mathcal{L}\mathbf{w}, \mathbf{w}). \qquad (25) \end{aligned}$$

The formal normal is

$$\mathcal{L}^*\mathcal{L} = \begin{pmatrix} -(1+a^2)\Delta & 0 & 0 & -a^2\Delta & \partial_{xx} \\ 0 & -\Delta & 0 & 0 & \partial_{xy} \\ 0 & 0 & -\Delta & 0 & \partial_{xy} \\ -a^2\Delta & 0 & 0 & -(1+a^2)\Delta & \partial_{yy} \\ \partial_{xx} & \partial_{xy} & \partial_{xy} & \partial_{yy} & -\Delta \end{pmatrix}. \tag{26}$$

decoupling the variables

Remember, our mission is to

divide and conquer,

which implies that the optimal form of $\mathfrak{L}^*\mathfrak{L}$ would be a diagonal matrix.

Let's look at our $\mathfrak{L}^*\mathfrak{L}$. There's always going to be the coupling between $\mathbf{U}$ and p represented by the fifth column and fifth row of $\mathfrak{L}^*\mathfrak{L}$. However, it is possible to reduce the coupling between the $\mathbf{U}$ terms, described by the leading 4×4 block. It's immediately obvious that the pesky $-a^2\Delta$ terms, representing the coupling between U and Z, need to go away. This suggests that we should rotate the variables,

$$U_{new} = \frac{1}{\sqrt{2}}(U + Z), \tag{27}$$

$$Z_{new} = \frac{1}{\sqrt{2}}(U - Z). \tag{28}$$

Equlvalently we can write

$$(\mathfrak{L}^*\mathfrak{L})_r = \mathbf{Q}^*\mathfrak{L}^*\mathfrak{L}\mathbf{Q} = \begin{pmatrix} -(1+2a^2)\Delta & 0 & 0 & 0 & \frac{1}{\sqrt{2}}\Delta \\ 0 & -\Delta & 0 & 0 & \partial_{xy} \\ 0 & 0 & -\Delta & 0 & \partial_{xy} \\ 0 & 0 & 0 & -(1+2a^2)\Delta & \partial_{xx} - \partial_{yy} \\ \frac{1}{\sqrt{2}}\Delta & \partial_{xy} & \partial_{xy} & \partial_{xx} - \partial_{yy} & -\Delta \end{pmatrix} \tag{29}$$

where

$$\mathbf{Q} = \mathbf{Q}^* = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{30}$$

What about the ideal convergence rate?

We consider the rotated system because it's easier. The ideal convergence rate depends on the intervariable interaction so it is smaller for the rotated system because of the reduced coupling between the $\mathbf{U}$ terms. To determine the rate we first rewrite $(\mathcal{L}^*\mathcal{L})_r$, as the product of three matrices as below

$$\begin{aligned}(\mathcal{L}^*\mathcal{L})_r &= \begin{pmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{x}^* & B \end{pmatrix} = \begin{pmatrix} \mathbf{A}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0}^* & B^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}^{-\frac{1}{2}}\mathbf{x}B^{-\frac{1}{2}} \\ B^{-\frac{1}{2}}\mathbf{x}^*\mathbf{A}^{-\frac{1}{2}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0}^* & B^{\frac{1}{2}} \end{pmatrix} \\ &= \mathbf{D}^* \begin{pmatrix} \mathbf{I} & \mathbf{E} \\ \mathbf{E}^* & \mathbf{I} \end{pmatrix} \mathbf{D}, \end{aligned} \tag{31}$$

where $\mathbf{x}^* = \begin{pmatrix} \frac{1}{\sqrt{2}}\Delta & \partial_{xy} & \partial_{xy} & \partial_{xx} - \partial_{yy} \end{pmatrix}$. Since $\mathbf{D}$ is block diagonal, we can view $\mathbf{E}$ as the vector that couples $\mathbf{U}$ and p and the intervariable coupling can be measured in terms of the norm of $\mathbf{E}$

$$\|\mathbf{E}\| = k = \sqrt{\rho(\mathbf{E}^*\mathbf{E})} = \sqrt{\rho(B^{-\frac{1}{2}}\mathbf{x}^*\mathbf{A}^{-1}\mathbf{x}B^{-\frac{1}{2}})}. \tag{32}$$

This can be calculated,

$$k^2 = (-\Delta)^{\frac{1}{2}} \begin{pmatrix} \frac{1}{\sqrt{2}}\Delta \\ \partial_{xy} \\ \partial_{xy} \\ \partial_{xx} - \partial_{yy} \end{pmatrix}^t (-\Delta)^{-1} \begin{pmatrix} \frac{1}{1+2a^2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{1+2a^2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}}\Delta \\ \partial_{xy} \\ \partial_{xy} \\ \partial_{xx} - \partial_{yy} \end{pmatrix} (-\Delta)^{\frac{1}{2}}.$$

To work out an actual real number for this we assume $(-\Delta)^{-1}$ commutes with everything and we get

$$\mathbf{E}^*\mathbf{E} = \frac{1+a^2}{1+2a^2} \tag{33}$$

which has a value of $\frac{2}{3}$ when $a = 1$ and $\frac{1}{2}$ when $a \to \infty$. What does this tell us? It tells us that if we impose the continuity equation by enforcing $\|\nabla \mathrm{tr}\mathbf{U}\| = 0$ then we reduce the convergence rate. This also has the advantage of reducing the number of equations and the number of dependent variables.

So?

To show how k^2 relates to the convergence rate we consider Gauss-Seidel iterations. We can use expansion (31), to write the functional in a new way

$$F(\mathbf{w};\mathbf{0}) = (\mathcal{L}^*\mathcal{L}\mathbf{w},\mathbf{w}) = \left(\mathbf{D}^* \begin{pmatrix} \mathbf{I} & \mathbf{E} \\ \mathbf{E}^* & \mathbf{I} \end{pmatrix} \mathbf{D}\mathbf{w},\mathbf{w}\right) = \left(\begin{pmatrix} \mathbf{I} & \mathbf{E} \\ \mathbf{E}^* & \mathbf{I} \end{pmatrix} \mathbf{D}\mathbf{w},\mathbf{D}\mathbf{w}\right). \tag{34}$$

The minimum of this functional is going to occur when

$$\begin{pmatrix} \mathbf{I} & \mathbf{E} \\ \mathbf{E}^* & \mathbf{I} \end{pmatrix} \mathbf{D}\mathbf{w} = \mathbf{0}. \tag{35}$$

A Gauss Seidel iteration is

$$(\mathbf{D}\mathbf{w})^{n+1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{E}^* & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} & -\mathbf{E} \\ \mathbf{0}^* & \mathbf{0} \end{pmatrix} (\mathbf{D}\mathbf{w})^n = \begin{pmatrix} \mathbf{0} & -\mathbf{E} \\ \mathbf{0}^* & \mathbf{E}^*\mathbf{E} \end{pmatrix} (\mathbf{D}\mathbf{w})^n \tag{36}$$

which has a convergence rate of k^2.

weak form

What we really need to do is minimise the functional (10), which can be written as

$$F(\mathbf{w};\mathbf{f},\Phi) = \|\mathcal{L}\mathbf{w}-\mathbf{f}\|^2_{0,\Omega} + \|\mathcal{B}\mathbf{w}-\Phi\|^2_{\frac{1}{2},\Gamma} = (\mathcal{L}\mathbf{w}-\mathbf{f},\mathcal{L}\mathbf{w}-\mathbf{f}) + (\mathcal{B}\mathbf{w}-\Phi,\mathcal{B}\mathbf{w}-\Phi). \quad (37)$$

What we want to do is find a matrix system such that the solution of this matrix system is the minimum of the functional. We need to find the weak form.

Let's assume that $\hat{\mathbf{w}}$ is the $\mathbf{w}$ that gives the minimum of F. Then we can write

$$\begin{aligned} F(\hat{\mathbf{w}}+s\mathbf{v};\mathbf{f},\Phi) =& (\mathcal{L}\hat{\mathbf{w}}-\mathbf{f},\mathcal{L}\hat{\mathbf{w}}-\mathbf{f}) + (\mathcal{B}\hat{\mathbf{w}}-\Phi,\mathcal{B}\hat{\mathbf{w}}-\Phi) \\ &+ 2s\left[(\mathcal{L}\hat{\mathbf{w}}-\mathbf{f},\mathcal{L}\mathbf{v}-\mathbf{f}) + (\mathcal{B}\hat{\mathbf{w}}-\Phi,\mathcal{B}\mathbf{v}-\Phi)\right] \\ &+ s^2\left[(\mathcal{L}\mathbf{v},\mathcal{L}\mathbf{v}) + (\mathcal{B}\mathbf{v},\mathcal{B}\mathbf{v})\right] \\ =& F(\hat{\mathbf{w}};\mathbf{f},\Phi) \\ &+ 2s\left[(\mathcal{L}\hat{\mathbf{w}}-\mathbf{f},\mathcal{L}\mathbf{v}-\mathbf{f}) + (\mathcal{B}\hat{\mathbf{w}}-\Phi,\mathcal{B}\mathbf{v}-\Phi)\right] \\ &+ s^2\left[(\mathcal{L}\mathbf{v},\mathcal{L}\mathbf{v}) + (\mathcal{B}\mathbf{v},\mathcal{B}\mathbf{v})\right]. \end{aligned} \quad (38)$$

Now if $\hat{\mathbf{w}}$ is the minimum then this implies

$$(\mathcal{L}\hat{\mathbf{w}},\mathcal{L}\mathbf{v}-\mathbf{f}) + (\mathcal{B}\hat{\mathbf{w}},\mathcal{B}\mathbf{v}-\Phi) = (\mathbf{f},\mathcal{L}\mathbf{v}-\mathbf{f}) + (\Phi,\mathcal{B}\mathbf{v}-\Phi) \quad (39)$$

for every $\mathcal{L}\mathbf{v}$ in the space. This is the weak form.

In finite dimensional world, we approximate $\hat{\mathbf{w}}$ by $\hat{\mathbf{w}}_h$ and then say

$$(\mathcal{L}\hat{\mathbf{w}}_h,\mathcal{L}\mathbf{v}_h-\mathbf{f}) + (\mathcal{B}\hat{\mathbf{w}}_h,\mathcal{B}\mathbf{v}_h-\Phi) = (\mathbf{f},\mathcal{L}\mathbf{v}_h-\mathbf{f}) + (\Phi,\mathcal{B}\mathbf{v}_h-\Phi) \quad (40)$$

is true for every $\mathbf{v}_h$ in the finite element subspace.

conclusions

We've shown that not any old first order system can be used in FOSLS. The first order system we choose needs to yield H^1 ellipticity. This results in optimal finite element discretisation and optimal multigrid convergence.

The formal normal is a useful tool. It can be used to demonstrate the intervariable coupling and as we saw, it may also suggest rotations of the variables that will reduce this coupling. It also can be used to determine ideal convergence rates.

We've also shown how to obtain the weak form from the functional and how to go from this infinite dimensional weak form to the finite dimensional problem. It is the finite dimensional problem that is sent off to multigrid or some other solver.

exercises

1. Work out the formal normal for original first order system (4),(5).
2. Work out the formal normal for three-dimensional Stokes.
3. See if you can get ***Steve*** and ***Tom***'s H^1 ellipticity result in Theorem 1. (Hint: see their Stokes paper.)
4. Read all the FOSLS papers.