FOSLoSophy

Informal assessment of
First-Order System Least Squares (FOSLS)

As with any new development that breaks from convention, the least-squares approach can be easily misunderstood. Both understatement and overstatement are natural tendencies. Newcomers are quick to recognize the subtlety of this methodology, which is true enough, but it is easy to misjudge its complexity and its potential. It seems at first glance to demand more computational resources and to require increased regularity and smoothness, but these impressions are usually dispelled on closer inspection. On the other hand, as promising as FOSLS seems to the relatively few who are working with it, this methodology is not about to completely replace any other approach. It is simply a new tool in the struggle for effective solution of partial differential equations. But it is compelling in its ability to yield uniformly fast and accurate solvers for some applications that involve high Reynolds flow, linear elasticity approaching the incompressible limit, and highly indefinite Helmholtz problems, for example.

In any event, the purpose of this note is to move towards a better understanding of where the least-squares approach fits into the field of computational mathematics. We assume that the reader is familiar with basic FOSLS concepts (see the end of this document for relevant sources), but not necessarily with its implications or basic philosophy. We begin with a general discussion and end with examples that illustrate various points of this discussion.

One of the main advantages of FOSLS is its formulation of a *minimization principle* for problems that have no other natural optimization form. This leads to several benefits: discretization can be done by Rayleigh-Ritz techniques, which are usually simpler than Galerkin methods because they need only one type of subspace and more powerful because they can exploit the optimization structure; analogously, the only basic choices needed for designing multigrid solvers for FOSLS discretizations are relaxation and interpolation (other formulations usually also require choices for restriction and coarse-grid operators); such 'variational' multigrid methods are typically more effective and better supported by theory than non-variational schemes; heuristics and analysis of discretization can be based on approximation theory instead of more cumbersome and often misleading approaches like truncation-error analysis; and the problem is imbued with a sense of optimality, which is important in developing robust discretizations, especially in the presence of adaptive refinement, singularities and discontinuities, nonlinearities, and widely varying coefficients. Another advantage, which reflects the primary goal of this methodology, is that FOSLS represents a top-down *systematic* approach, with the basic aim of formulating the original problem so well that the numerical processes (discretization and multigrid solution) becomes as straightforward and optimal as possible. In this vein, an attempt is made to formulate the functional so that the system variables (e.g., potentials and fluxes) are essentially *decoupled,* meaning that the homogeneous form is *elliptic and continuous*

(more generally, it is equivalent to a diagonal form of scalar functionals). This allows essentially separate consideration of each variable in the discretization and multigrid solution processes. For example, it avoids restrictions like the LBB condition and the need for staggered grids. When the functional is designed to be fully product H^1 equivalent, then standard discretization and multigrid solution techniques can be easily and effectively applied. For example, no special 'stabilization' techniques like artificial diffusion or upwinding are needed. Another important attribute is that the FOSLS functional provides a practical and sharp *a posteriori error measure* at no additional cost: the value of the functional provides a measure of both the absolute and the relative total error for any approximation, no matter how it was obtained.

Of course, these advantages must come at some cost. First, there are *more system variables*. For example, a diffusion-convection equation would require four dependent variables in three-dimensional geometry. On the other hand, this does not necessarily mean more total scalar unknowns or greater overall computational cost, especially if FOSLS strengthened accuracy properties lead to smaller or, generally, more easily solved algebraic systems. It may also be that these new variables are what must ultimately be computed in practice. A second potential disadvantage is that, to obtain full product H^1 equivalence (equivalence of the homogeneous form to a form of decoupled H^1 scalar norms), the problem must exhibit *more regularity and smoothness* than it may possess. However, the H^{-1} norm versions of FOSLS apply whenever, say, mixed methods do, and the L^2 norm versions generally require only the additional assumption that the source be in L^2; moreover, these difficulties are usually isolated phenomena, occurring near rough boundaries or internal interfaces that require special treatment in any case. A third possible disadvantage is that FOSLS formulations are generally *non-conservative*, although this can usually be corrected by simple changes to the basic approach. Finally, development of a fully product-H^1-equivalent FOSLS functional can be *tricky*: augmenting the system with appropriate equations and boundary conditions can require much understanding and a bit of luck.

FOSLS H^{-1} Norm Version

The basic H^{-1} norm versions of FOSLS involve the simplest reduction of the original equations to first-order form, an L^2 norm applied to the 'flux' equations that define the new variables, and an H^{-1} norm that in essence preconditions the 'divergence' equations (those equations that usually come from higher derivatives in the original equations) and admits less smoothness (e.g., non-integrable source terms). As such, this version is usually *simpler* to define and more *generally applicable.* In fact, it applies under the same general framework as do mixed methods. The aim of H^{-1} norm versions of FOSLS is to develop a functional whose homogeneous form is equivalent to a diagonal form, usually involving lower-order norms on the less smooth variables (e.g., fluxes).

One difficulty with this version of FOSLS stems from the use of these weaker norms: while they do allow for less smooth data, they also generally require *delicate*

balance between the different terms of the functional. This can be seen by realizing that the primary objective in establishing diagonal equivalence is to show that the off-diagonal terms are small in some sense: for basic H^{-1} norm versions of FOSLS, these terms are usually of the same order as the diagonal terms, so there is little leeway to scale the terms of the functional differently. For example, applied to the usual elliptic equation with a diffusion coefficient that varies possibly with jumps and/or large changes of scale over the region, straightforward application of this version of FOSLS does not give diagonal equivalence uniformly in coefficient variations, the symptoms of which are degrading approximation and solver performance. Another aspect of this sensitivity to scaling is the infeasibility of 'two-stage' schemes. For Poisson's equation using the L^2 version of FOSLS with curl conditions, for example, the flux equations can be weighted with arbitrarily small positive parameters, without loss of accuracy or solver speed, because this FOSLS system is almost lower triangular. In fact, the weights can be set to zero and a proper functional for the fluxes remains, allowing for a second-stage recovery of the potential. Unfortunately, the intimate coupling of the divergence and flux terms prevents this in the basic H^{-1} norm version of FOSLS. By itself this is not necessarily troublesome, but this sensitivity to scaling might be limiting in the design of parameter-independent formulations, such as for high Reynolds number fluid flow problems, where flexibility of scaling is the basic tool used to devise L^2-type FOSLS functionals that are uniformly effective over all Reynolds number ranges.

An augmented H^{-1} norm version of FOSLS can be used to overcome these difficulties. The central idea is that the simple system used in the basic H^{-1} norm version is augmented with additional but consistent equations (e.g., when the new variables are defined in terms of gradients, a 'curl' equation might be used to balance the 'divergence' equation). The basic functional is thus augmented by adding a term involving an appropriate H^{-1} norm of the additional equations, which tends to equalize its scales and embue it with the superior properties of the analogous but better known L^2 norm functional discussed next.

FOSLS L^2 Norm Version

As with the augmented H^{-1} norm functional, the L^2 norm version involves additional equations, but now all terms involve the L^2 norm. When it applies, this version is usually the most effective form of FOSLS. To be sure, H^2 regularity is usually needed theoretically to establish full product H^1 equivalence of special L^2 norm formulations, and this can be tricky to achieve. However, no additional regularity is really needed to apply the L^2 norm version: it is generally enough that the source be in L^2. The basic advantage of this version is the *stronger sense of well-posedness*: applied to scalar second-order equations, the homogeneous form of the L^2 norm functional exhibits $(H^{div} \bigcap H^{curl}) \times H^1$ equivalence in general, which means local vector H^1 equivalence, and it exhibits full vector H^1 equivalence for special forms when the original problem has increased regularity. This equivalence leads to several other benefits, including *optimal performance* of standard discretization and multigrid schemes,

some *scale insensitivity* that allows 'two-stage' methods and design of uniformly well-posed functionals for many parameter-dependent problems, and *lower-order variable coupling* that allows some leeway in the approximation of individual variables. We attempt to clarify some of these benefits below. Finally, as with most formulations of this type that admit efficient multigrid solvers, this approach is highly vectorizable and parallelizable, especially since the individual variables are essentially decoupled and can therefore be processed simultaneously.

FOSLL*

An alternative to the H^{-1} Norm Version

The L^2 norm version of FOSLS can be described abstractly as starting with a preferably first-order equation $Lp = f$ and transforming it to the least-squares equation $L^*Lp = L^*f$. Another way to yield a self-adjoint nonnegative-definite system is to form the equation $LL^*q = f$. For this to be useful, L and its adjoint L^* must be imbued with certain theoretical properties, but a compelling feature of such a FOSLL* approach is that the Rayleigh-Ritz solution of $LL^*q = f$ on any subspace is the exact minimizer of the L^2 norm. This follows from noting that the energy functional for this system differs from the L^2 norm of the error by a constant. To see this, let Q be the exact solution of $LL^*q = f$ and q some approximation to it. Then

$$\begin{aligned} < LL^*q, q > -2 < q, f > &= < L^*q, L^*q > \quad -2 < q, LL^*Q > \\ &= < L^*q, L^*q > \quad -2 < L^*q, L^*Q > \\ &\qquad\qquad\quad + < L^*Q, L^*Q > - < L^*Q, L^*Q > \\ &= \|L^*(q-Q)\|^2 \quad - < L^*Q, L^*Q > \\ &= \|L^*(q-Q)\|^2 \quad - \text{ constant.} \end{aligned}$$

Illustration

A scalar elliptic equation

For concreteness, consider the convection-diffusion equation

$$\bigtriangledown^* a \bigtriangledown p + \mathbf{b} \cdot \bigtriangledown p = f$$

with appropriate boundary conditions, which we ignore here in deference to a simpler, more formal discussion. Here, $\bigtriangledown^* = - \bigtriangledown \cdot$, $a \geq 1$ is a scalar function, and $\mathbf{b}$ is a vector function. A natural H^{-1} norm FOSLS functional is

$$F\ (\mathbf{u}, p; f) \equiv || \bigtriangledown^* \mathbf{u} + \mathbf{b} \cdot (\mathbf{u}/a) - f||^2_{-1} + ||\mathbf{u}/\sqrt{a} - \sqrt{a} \bigtriangledown p||^2,$$

where $|| \cdot ||$ is the vector or scalar L^2 norm and $|| \cdot ||_{-1}$ is meant in the usual sense to be the norm induced by the dual of the H^1 inner product. Now minimization of F makes sense under very general smoothness ($f \in H^{-1}$) and regularity (H^1) assumptions. Unfortunately, this form is too sensitive to variations in a. To see this,

first note that the *homogeneous form* $F(\mathbf{u}, p; 0)$ is generally equivalent to the *diagonal form*

$$G(\mathbf{u}, p) \equiv ||\mathbf{u}/\sqrt{a}||^2 + ||\sqrt{a} \bigtriangledown p||^2,$$

that is, $F(\mathbf{u}, p; 0) \sim G(\mathbf{u}, p)$. This follows easily in the simplest case $a = 1$ and $\mathbf{b} = \mathbf{0}$, where the Hessians of F and G are as follows, with I denoting the identity:

$$F'' = \begin{pmatrix} I + \bigtriangledown L^{-1} \bigtriangledown * & -\bigtriangledown \\ -\bigtriangledown^* & \bigtriangledown^* \bigtriangledown \end{pmatrix} \text{ and } G'' = \begin{pmatrix} I & O \\ O & \bigtriangledown^* \bigtriangledown \end{pmatrix}.$$

Here, the representation corresponds to $\begin{pmatrix} \mathbf{u} \\ p \end{pmatrix}$ and the operator L is defined by $L = I + \text{diag}(\bigtriangledown^* \bigtriangledown)$, which comes from the H^{-1} norm. It is easy to see that the off-diagonal term $-\bigtriangledown$ is subdominant because $\mathcal{E} \equiv (I + \bigtriangledown L^{-1} \bigtriangledown^*)^{-1/2}(-\bigtriangledown)(\bigtriangledown^* \bigtriangledown)^{-1/2}$ satisfies $\mathcal{E}^* \mathcal{E} << I$ (i.e., $||\mathcal{E}|| << 1$). What is important to note is that this bound is only a consequence of *scale*, not *order*: $\mathcal{E}^* \mathcal{E}$ is less than I in norm, not in order. For the L^2 norm versions of FOSLS introduced below, the corresponding scaled off-diagonal term $\mathcal{E}$ is such that $\mathcal{E}^* \mathcal{E}$ is of 'negative order,' meaning in a loose sense that the off-diagonal terms of F'' involve lower derivatives than do the diagonal terms. This yields a much stronger sense of diagonal equivalence, which in turn leads to several additional benefits that are suggested below.

Now, in general, the sharpness of the equivalence $F(\mathbf{u}, p; 0) \sim G(\mathbf{u}, p)$ depends intimately on the magnitude $||\sqrt{a}||_\infty$. For the case $a = 0(\frac{1}{\epsilon})$, $F(\mathbf{u}, 0; 0)$ is essentially $|| \bigtriangledown^* \mathbf{u}||^2_{-1} + \epsilon ||\mathbf{u}||^2$. For gradient fields ($\mathbf{u} = \bigtriangledown \theta$), we have $F(\mathbf{u}, 0; 0) \sim (1 + \epsilon)||\mathbf{u}||^2 \sim \frac{1+\epsilon}{\epsilon} G(\mathbf{u}, p)$. For curl fields ($\mathbf{u} = \bigtriangledown \times \mathbf{v}$), $F(\mathbf{u}, 0; 0) \sim \epsilon ||\mathbf{u}||^2 \sim G(\mathbf{u}, p)$. Thus, the equivalence of F to G must feel swings of at least $\frac{1+\epsilon}{\epsilon} = 1 + 0(||a||_\infty)$. This dependence can lead to degradation of discretization and algebraic solution techniques.

A choice that appears to avoid this scaling trouble is

$$F(\mathbf{u}, p; f) \equiv ||\frac{1}{\sqrt{a}}(\bigtriangledown^* \mathbf{u} + \mathbf{b} \cdot (\mathbf{u}/a) - f)||^2_{-1} + ||\mathbf{u}/\sqrt{a} - \sqrt{a} \bigtriangledown p||^2.$$

Generally, we can now expect $F(\mathbf{u}, p; 0) \sim G(\mathbf{u}, p)$, where this diagonal equivalence may be uniform in variations in a. However, there is no apparent way to insulate this equivalence from the deleterious effects of large $||\mathbf{b}||$. This limitation stems from the delicate balance needed between the two terms defining F: scaling up the divergence (first) term by a large coefficient would lead to an unfortunate dominance of this singular term that weakens the decoupling between the components of $\mathbf{u}$ and causes trouble analogous to the large $||\sqrt{a}||_\infty$ case, while scaling up the flux (second) term would severely weaken the decoupling between $\mathbf{u}$ and p. Either scaling therefore weakens the diagonal form equivalence. Other aspects of this weaker sense of diagonal equivalence are that care must be taken: in approximating the H^{-1} norm (i.e., in discretizing L^{-1}); in discretizing the individual variables (loss of approximation properties of p in the $||\sqrt{a} \bigtriangledown p||$ sense would directly degrade the approximation

properties of $\mathbf{u}$ in the $||\mathbf{u}/\sqrt{a}||$ sense); and in constructing the appropriate algebraic solver.

By introducing a curl equation into this formulation, we can rebalance the functional terms so that these limitations are easily avoided. Specifically, since $\mathbf{u}/a$ is a gradient at the solution, its curl must be zero. This allows us to redefine the functional as follows:

$$F(\mathbf{u}, p; f) \equiv ||a \bigtriangledown \times(\mathbf{u}/a)||^2_{-1} + || \bigtriangledown^* \mathbf{u} + \mathbf{b} \cdot (\mathbf{u}/a) - f||^2_{-1} + ||\mathbf{u}/\sqrt{a} - \sqrt{a} \bigtriangledown p||^2.$$

There is flexibility in the choice of the H^{-1} norm of the curl term: while the usual Laplacian involved in the operator L for the divergence term is appropriate, an operator involving a (e.g., $\bigtriangledown^* a^2 \bigtriangledown$) might be more appropriate for the curl term. At least for the case of mildly varying $a >> 0$, it is possible to make this choice to obtain the following equivalence property:

$$F(\mathbf{u}, p; 0) \sim G(\mathbf{u}, p) \equiv ||\mathbf{u}||^2 + ||\sqrt{a} \bigtriangledown p||^2,$$

This is especially interesting because it implies that simple iterative methods (like steepest descent) can obtain optimal convergence without appeal to multileveling (except, as before, multigrid would be used to evaluate the inverse norms that define the functional).

Now the only additional requirement for the L^2 norm version of FOSLS to be applied is that f be in L^2. The added advantages of this version usually make it the method of choice when this is the case. Thus, *with no additional regularity*, we can replace the H^{-1} norms in the functional by L^2 norms to obtain

$$F(\mathbf{u}, p; f) \equiv ||a \bigtriangledown \times(\mathbf{u}/a)||^2 + || \bigtriangledown^* \mathbf{u} + \mathbf{b} \cdot (\mathbf{u}/a) - f||^2 + ||\mathbf{u}/\sqrt{a} - \sqrt{a} \bigtriangledown p||^2,$$

and we can be assured of the equivalence property

$$F(\mathbf{u}, p; 0) \sim G(\mathbf{u}, p) \equiv ||a \bigtriangledown \times(\mathbf{u}/a)||^2 + || \bigtriangledown^* \mathbf{u}||^2 + ||\mathbf{u}/\sqrt{a}||^2 + ||\sqrt{a} \bigtriangledown p||^2,$$

which is generally uniform in variations in a. With the further assumption of H^2 regularity of the original elliptic equation *without convection* ($\mathbf{b} = \mathbf{0}$), we obtain the stronger vector H^1 uniform equivalence

$$F(\mathbf{u}, p; 0) \sim G(\mathbf{u}, p) \equiv ||\mathrm{diag}(\bigtriangledown^* \bigtriangledown)\mathbf{u}||^2 + ||\mathbf{u}/\sqrt{a}||^2 + ||\sqrt{a} \bigtriangledown p||^2.$$

To understand this equivalence from a heuristic standpoint, consider the simple case $a = 1$ and $\mathbf{b} = \mathbf{0}$. Then

$$F''(\mathbf{u}, p; 0) = \begin{pmatrix} I + \mathrm{diag}(\bigtriangledown^* \bigtriangledown) & -\bigtriangledown \\ -\bigtriangledown^* & \bigtriangledown^* \bigtriangledown \end{pmatrix} \text{ and } G''(\mathbf{u}, p) = \begin{pmatrix} I + \mathrm{diag}(\bigtriangledown^* \bigtriangledown) & 0 \\ 0 & \bigtriangledown^* \bigtriangledown \end{pmatrix}.$$

Now the variable coupling is dictated by the scaled off-diagonal term $\mathcal{E} = (I + \mathrm{diag}(\bigtriangledown^* \bigtriangledown))^{1/2}(-\bigtriangledown)(\bigtriangledown^* \bigtriangledown)^{-1/2}$, which is easily seen to satisfy $\mathcal{E}^* \mathcal{E} << I$ because

of scale *and* order ($\mathcal{E}^*\mathcal{E}$ is of 'negative order'). This low order coupling leads to much more flexibility in terms of fast solvers: for small h, the only coupling between variables is between smooth components; this implies that the variables are almost completely decoupled for relaxation purposes on fine grids; and this in turn allows for the simplest relaxation schemes that process the variables separately. This low-order coupling also allows for more leeway in the individual approximation of variables: $\mathbf{u}$ can be approximated on a grid whose mesh size is the square root of that for p without adversely effecting the approximation for p. Finally, it enables separate scaling of the individual terms in the functional.

Concerning flexibility in functional scaling, consider the following *two-stage* FOSLS scheme: start by minimizing the functional obtained by deleting the flux term in F (this yields the first-stage functional $||a \bigtriangledown \times(\mathbf{u}/a)||^2 + || \bigtriangledown^* \mathbf{u} + \mathbf{b} \cdot (\mathbf{u}/a) - f||^2$ that is well posed in $\mathbf{u}$ alone), then recover p by fixing $\mathbf{u}$ to be the resulting approximation and minimizing the flux term ($||\mathbf{u}/\sqrt{a} - \sqrt{a} \bigtriangledown p||^2$) alone for p. It is easy to see that this two-stage scheme can usually be done with no real degradation in accuracy or efficiency. However, a more important issue here is the implication that this scale flexibility has on designing functionals for large Reynolds number flow (i.e., large $||\mathbf{b}||$). To see this implication in its simplest setting, consider the model case

$$\bigtriangledown^* \bigtriangledown p + bp_x = f$$

We can reduce this equation by an exponential transformation to the diffusion-only-type case

$$-e^{bx}(e^{-bx}p_x)_x - p_{yy} = f.$$

Here we assume that $x \in [0, 1)$ and $b \geq 0$ so that $e^{-bx} \in (0, 1)$. With this important *damping* scale term in mind, then an appropriate FOSLS functional can be obtained in a straightforward way, but by carefully ensuring that the flux terms do not dominate the div-curl (first and second) terms. This leads to
$F(u, v, p; f) \equiv$

$$\int\int [e^{-bx}((e^{bx}u - p_x)^2 + (v - p_y)^2) + (e^{\frac{bx}{2}}u_x + e^{-\frac{bx}{2}}v_y + e^{-\frac{bx}{2}}f)^2 + (e^{\frac{bx}{2}}u_y - e^{-\frac{bx}{2}}v_x)^2]dxdy.$$

It is better now to absorb e^{bx} into u (replace $e^{bx}u$ by u), so that we can in essence extract out the original transformation, with the aim of making the resulting functional less sensitive to the precise form of the exponential transformation:
$F(u, v, p; f) =$

$$\int\int [(u - p_x)^2 + (v - p_y)^2 + (u_x - bu + v_y + f)^2 + (u_y - v_x)^2]e^{-bx}dxdy.$$

What we have done is: use a *temporary* transformation of scale to recast the original (convection-diffusion) problem to familiar (diffusion) form; then apply FOSLS while exploiting the flexibility in scale to maintain diagonal equivalence; and then transform back so that the scale appears only as an integral weighting. This approach is

important because it ensures, under fairly general conditions, that the functional is uniformly equivalent to its diagonal form:

$$F(u,v,p;0) \sim \int\int [u^2 + (u_x + bu)^2 + u_y^2 + v^2 + v_x^2 + v_y^2 + p_x^2 + p_y^2] e^{-bx} dxdy.$$

What is essential here is that the equivalence holds equally well for arbitrarily large Reynolds number b.

Note that the functional obtained by this approach represents a simple modification of the basic FOSLS functional

$$|| \bigtriangledown \times \mathbf{u}||^2 + || \bigtriangledown^* \mathbf{u} + \mathbf{b}^* \mathbf{u} - f||^2 + ||\mathbf{u} - \bigtriangledown p||^2.$$

The only difference is a weighting of the L^2 norm by a term that attenuates exponentially in the streamwise direction into what is typically the boundary layer. This amounts to damping the influence of errors in all variables in the boundary layer, which seems appropriate because it has the effect of protecting the rest of the region (free stream) from contamination caused by the general inability to approximate the rapidly changing variables there. The functional equivalence to the H^1 -like norm weighted in this way ensures that free-stream discretization accuracy is optimal in all variables in the unscaled H^1 sense, and that multigrid converges with the usual elliptic speed.

Convection-dominated applications are extremely challenging and subtle in all but the simplest cases, such as that considered here. It is therefore difficult to know if the simple approach developed above for constant-coefficient problems can be extended to those of more practical interest. How general this approach is and how sensitive it is to the proper choice of scale remains an open question. However, it seems likely that a reasonable and efficient strategy for local approximation of the weighting (based on evolving elementwise estimates of the stream direction) could be devised for a fairly general class of variable-coefficient problems. In fact, even nonlinear convection-diffusion problems may be amenable to this approach, especially if full multigrid schemes are used to obtain good *initial* estimates from coarser grids.

Note that uniform diagonal equivalence means that the overall accuracy and efficiency can be considered separately for each variable. Since the variable coupling is subdominant in scale and order, then *relative accuracy within each variable is assured in its associated scalar norm.*

Two additional benefits come with the least-squares approach: sharp a posteriori error estimates, enabled because the true minimum value of the functional is known (zero); and natural treatment of nonconforming finite elements, enabled because the functional is directly posed in weak form. For either version of FOSLS, the quantities

$$F(\mathbf{v}, q; f) \qquad \text{and} \qquad \frac{F(\mathbf{v}, q; f)}{F(\mathbf{0}, 0; f)}$$

serve as *precise* measures of the absolute and relative errors in the approximation $(\mathbf{v}, q)$, no matter how it was obtained. Although this assumes that $F(\cdot, \cdot; 0)$ itself is

acceptable as an error norm, in any case these quantities provide *sharp* error estimates in terms of the appropriate Sobolev norm on $(\mathbf{v}, q)$, depending on the norm equivalence established for $F(\cdot, \cdot; 0)$. The advantages in terms of nonconforming elements is illustrated by considering a space that violates whatever boundary conditions might be present. For example, for the convection-diffusion case considered above, the boundary condition $p = g$ would naturally be replaced by $p = g^h$, where g^h is the trace of some function in the discrete space for p. (A similar approximation for FOSLS would be made for $\mathbf{u}$ components as well.) Of course, any finite element method must account for the error induced directly by this approximation, but, for standard finite elements applied to the original scalar convection-diffusion problem, we would also have to account for the error in the weak form caused by the new boundary integral terms that this approximation generates. However, no such terms are generated in the FOSLS functional since the weak form occurs directly, without integration by parts, so the only issue here is how well the nonconforming subspace approximates the original solution. In words borrowed from Gil Strang, no variational crimes have been committed. Thus, said in finite element terms, Strang's lemma reduces to the much simpler Ceá's lemma for FOSLS functionals.

Illustration

Stokes

In this section, we develop a least-squares functional for the two- and three-dimensional Stokes equations, generalized by allowing a pressure term in the continuity equation. By introducing a *velocity flux* variable (i.e., the gradient of velocity) and associated curl and *trace* equations, ellipticity is guaranteed under full regularity assumptions in an H^1 product norm appropriately weighted by the Reynolds number. Moreover, the generalized Stokes equations allow us to develop an analogous result for the Dirichlet problem for linear elasticity, where we obtain the more substantive result that the estimates that are uniform in the Poisson ratio.

With Ω a bounded, open, connected domain in $\Re^n$ ($n = 2$ or 3) and $\partial\Omega$ its Lipschitz boundary, our stationary pressure-perturbed form of the *generalized Stokes equations* in dimensionless variables is given by

$$\left\{\begin{array}{rcll} -\nu\Delta\,\mathbf{u} + \nabla\, p &=& \mathbf{f}, & \text{in} \quad \Omega, \\ \nabla\cdot\,\mathbf{u} + \delta\, p &=& g, & \text{in} \quad \Omega, \\ \mathbf{u} &=& \mathbf{0} & \text{on} \quad \partial\Omega, \end{array}\right.$$

where: Δ, ∇, and $\nabla\cdot$ stand for the Laplacian, gradient, and divergence operators, respectively; $\Delta\,\mathbf{u}$ signifies the n-vector of components $\Delta\, u_i$ (i.e., Δ applies to $\mathbf{u}$ componentwise): ν is the reciprocal of the Reynolds number Re; $\mathbf{f}$ is a given vector function; g is a given scalar function; and δ is a fixed nonnegative constant ($\delta = 0$ for Stokes, $\delta = 1$ for linear elasticity, and δ is assumed to be bounded uniformly in ν for the general case). Assume the *consistency conditions*

$$\int_\Omega g\, dz = \int_\Omega p\, dz = 0.$$

(For $\delta = 0$, the generalized Stokes equations can have a solution only when g satisfies this condition, and we are then free to ask that p satisfy it as well. For $\delta > 0$, in general we have only that $\int_\Omega g\, dz = \delta \int_\Omega p\, dz$, but this can be reduced to the consistency conditions simply by replacing p by $p - \frac{g}{\delta}$ and g by 0.)

Our generalized Stokes equations can be applied to linear elasticity given by

$$\left\{\begin{array}{rcll} -\mu\Delta\,\mathbf{u} - (\lambda+\mu)\nabla\nabla\cdot\,\mathbf{u} &=& \mathbf{f}, & \text{in} \quad \Omega, \\ \mathbf{u} &=& \mathbf{0}, & \text{on} \quad \partial\Omega, \end{array}\right. \tag{0.1}$$

where $\mathbf{u}$ now represents displacements and μ and λ are the (positive) Lamé constants. We do this by introducing the pressure variable $p = -\nabla\cdot\,\mathbf{u}$, by rescaling $\mathbf{f}$, and by letting $g = 0$, $\delta = 1$, and $\nu = \frac{\mu}{\lambda+\mu}$. (A more physical choice for this artificial pressure would have been $p = -\frac{\lambda}{2\mu}\,\nabla\cdot\,\mathbf{u}$, since it then becomes the hydrostatic pressure in the incompressible limit. We chose our particular scaling because it most easily conforms to the generalized Stokes equations. Also, ν here should not be confused with the Poisson ratio since we use it in this discussion only in the fluid dynamics sense.)

Let **curl** $\equiv \nabla\times$ denote the curl operator. (Here and henceforth, we use notation for the case $n = 3$ and consider the special case $n = 2$ in the natural way by identifying $\Re^2$ with the (x_1, x_2)-plane in $\Re^3$. Thus, if $\mathbf{u}$ is two dimensional, then the curl of $\mathbf{u}$ means the scalar function

$$\nabla\times \mathbf{u} = \partial_1 u_2 - \partial_2 u_1,$$

where u_1 and u_2 are the components of $\mathbf{u}$.) The following identity is immediate:

$$\nabla\times (\nabla\times \mathbf{u}) = -\Delta\, \mathbf{u} + \nabla\, (\nabla\cdot\, \mathbf{u}).$$

(For $n = 2$, this dentity is interpreted as

$$\nabla^\perp\, (\nabla\times \mathbf{u}) = -\Delta\, \mathbf{u} + \nabla\, (\nabla\cdot\, \mathbf{u}),$$

where $\nabla^\perp$ is the formal adjoint of $\nabla\times$ defined by

$$\nabla^\perp q = \begin{pmatrix} \partial_2 q \\ -\partial_1 q \end{pmatrix}.)$$

Below, we define a new independent variable as the n^2-vector function of gradients of the u_i, $i = 1, 2, ..., n$. It is convenient to view the original n-vector functions as column vectors and the new n^2-vector functions as either block column vectors or matrices. Thus, given

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

and denoting $\mathbf{u}^t = (u_1, u_2, ..., u_n)$, then an operator G defined on scalar functions (e.g., $G = \nabla$) is extended to n-vectors componentwise:

$$G\mathbf{u}^t = (Gu_1, Gu_2, ..., Gu_n)$$

and

$$G\mathbf{u} = \begin{pmatrix} Gu_1 \\ Gu_2 \\ \vdots \\ Gu_n \end{pmatrix}.$$

If $\mathbf{U}_i \equiv G\mathbf{u}_i$ is a n-vector function, then we write the matrix

$$\begin{aligned} \underline{\mathbf{U}} \equiv G\mathbf{u}^t &= (\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_n) \\ &= \begin{pmatrix} U_{11} & U_{12} & \cdots & U_{1n} \\ U_{21} & U_{22} & \cdots & U_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ U_{n1} & U_{n2} & \cdots & U_{nn} \end{pmatrix}. \end{aligned}$$

We then define the *trace* operator tr according to

$$tr\underline{\mathbf{U}} = \sum_{i=1}^{n} U_{ii}.$$

If D is an operator on n-vector functions (e.g., $D = \nabla\times$), then its extension to matrices is defined by

$$D\underline{\mathbf{U}} = (D\mathbf{U}_1,\, D\mathbf{U}_2,\, \cdots,\, D\mathbf{U}_n).$$

When each $D\mathbf{U}_i$ is a scalar function (e.g., $D = \nabla\cdot$), we view the extension as a mapping to column vectors, so we use the convention

$$(D\underline{\mathbf{U}})^t = \begin{pmatrix} D\mathbf{U}_1 \\ D\mathbf{U}_2 \\ \vdots \\ D\mathbf{U}_n \end{pmatrix}.$$

We also extend the tangential operator $\mathbf{n}\times$ componentwise ($\mathbf{n}$ denotes the outward unit normal on $\partial\Omega$):

$$\mathbf{n}\times\underline{\mathbf{U}} = (\mathbf{n}\times\mathbf{U}_1,\, \mathbf{n}\times\mathbf{U}_2,\, \cdots,\, \mathbf{n}\times\mathbf{U}_n).$$

Finally, inner products and norms on the matrix functions are defined in the natural componentwise way, e.g.,

$$\|\underline{\mathbf{U}}\|^2 = \sum_{i=1}^{n} \|\mathbf{U}_i\|^2 = \sum_{i,\,j=1}^{n} \|U_{ij}\|^2.$$

Introducing the *velocity flux* variable

$$\underline{\mathbf{U}} = \nabla\mathbf{u}^t = (\nabla\, u_1,\, \nabla\, u_2,\, \cdots,\, \nabla\, u_n),$$

then the generalized Stokes system can be recast as the following equivalent first-order system:

$$\left\{\begin{array}{rcll} \underline{\mathbf{U}} - \nabla\,\mathbf{u}^t &=& \underline{\mathbf{0}}, & \text{in}\quad \Omega, \\ -\nu\,(\nabla\cdot\underline{\mathbf{U}})^t + \nabla\, p &=& \mathbf{f}, & \text{in}\quad \Omega, \\ \nabla\cdot\mathbf{u} + \delta\, p &=& g, & \text{in}\quad \Omega, \\ \mathbf{u} &=& \mathbf{0}, & \text{on}\quad \partial\Omega. \end{array}\right.$$

Note that the definition of $\underline{\mathbf{U}}$, the "continuity" condition $\nabla\cdot\mathbf{u} + \delta\, p = g$ in Ω, and the Dirichlet condition $\mathbf{u} = \mathbf{0}$ on $\partial\Omega$ imply the respective properties

$$\nabla\times\underline{\mathbf{U}} = \underline{\mathbf{0}} \quad \text{in } \Omega, \quad tr\,\underline{\mathbf{U}} + \delta\, p = g \quad \text{in } \Omega, \quad \text{and } \mathbf{n}\times\underline{\mathbf{U}} = \underline{\mathbf{0}} \quad \text{on } \partial\Omega.$$

Then an equivalent extended first-order system is given by

$$\left\{\begin{array}{rcll} \underline{\mathbf{U}} - \nabla\,\mathbf{u}^t &=& \mathbf{0}, & \text{in}\quad \Omega, \\ -\nu\,(\nabla\cdot\underline{\mathbf{U}})^t + \nabla\, p &=& \mathbf{f}, & \text{in}\quad \Omega, \\ \nabla\, tr\,\underline{\mathbf{U}} + \delta\nabla\, p &=& \nabla g, & \text{in}\quad \Omega, \\ \nabla\times\underline{\mathbf{U}} &=& \underline{\mathbf{0}}, & \text{in}\quad \Omega, \\ \mathbf{u} &=& \mathbf{0}, & \text{on}\quad \partial\Omega, \\ \mathbf{n}\times\underline{\mathbf{U}} &=& \underline{\mathbf{0}}, & \text{on}\quad \partial\Omega. \end{array}\right.$$

This extended system is what we treat with least squares, although in stages based on the observation that we can separate it into two essentially well-posed systems:

$$\text{Stage 1}\left\{\begin{array}{rcll} -\nu\,(\nabla\cdot\underline{\mathbf{U}})^t+\nabla\,p &=& \mathbf{f}, & \text{in } \Omega,\\ \nabla tr\,\underline{\mathbf{U}}+\delta\nabla\,p &=& \nabla g, & \text{in } \Omega,\\ \nabla\times\underline{\mathbf{U}} &=& \underline{\mathbf{0}}, & \text{in } \Omega,\\ \mathbf{n}\times\underline{\mathbf{U}} &=& \underline{\mathbf{0}}, & \text{on } \partial\Omega, \end{array}\right.$$

and

$$\text{Stage 2}\left\{\begin{array}{rcll} \nabla\,\mathbf{u}^t &=& \underline{\mathbf{U}}, & \text{in } \Omega,\\ \mathbf{u} &=& \mathbf{0}, & \text{on } \partial\Omega. \end{array}\right.$$

What we mean by this separation is that Stage 1 can be solved for $\underline{\mathbf{U}}$ and p, and then Stage 2, with $\underline{\mathbf{U}}$ and p now given, can be solved for $\mathbf{u}$. We are encouraged to proceed in this two-stage approach because the connections between the stages (i.e., $\underline{\mathbf{U}}$ and $\delta\,p$) are differentially of low order compared to the dominant order of the stages themselves. Besides, the legitimacy of this approach has been established theoretically and numerically.

To this end, define the respective Stage 1 and Stage 2 functionals by

$$F_1(\underline{\mathbf{U}},p;\,\mathbf{f},g)=\|\mathbf{f}+\nu\,(\nabla\cdot\underline{\mathbf{U}})^t-\nabla\,p\|^2+\nu^2\,\|\nabla\times\underline{\mathbf{U}}\|^2+\nu^2\,\|\nabla tr\,\underline{\mathbf{U}}+\delta\nabla\,p-\nabla g\|^2$$

and

$$F_2(\mathbf{u};\,\underline{\mathbf{U}},p,g)=\|\nabla\,\mathbf{u}^t-\underline{\mathbf{U}}\|^2.$$

We can easily show that $F_1(\underline{\mathbf{U}},p;\,\mathbf{0},0)$ is uniformly equivalent to $\nu^2\,\|\underline{\mathbf{U}}\|_1^2+\|p\|_1^2$, and that $F_2(\mathbf{u};\,\underline{\mathbf{0}})$ is uniformly equivalent to $\nu^2\,\|\mathbf{u}\|_1^2$. Indeed, $F_2(\mathbf{u};\,\underline{\mathbf{0}})$ is the squared H^1 semi-norm! The practical implication is that the generalized Stokes equations may be solved optimally and uniformly in a *two-stage* process that involves first minimizing $F_1(\underline{\mathbf{U}},p;\,\mathbf{f},g)$ over $(\underline{\mathbf{U}},p)\in\{\underline{\mathbf{V}}\in H^1(\Omega)^{n^2}\ :\ \mathbf{n}\times\underline{\mathbf{V}}=\underline{\mathbf{0}}\,on\ \partial\Omega\}\times(H^1(\Omega)/\Re)$, and then fixing $\underline{\mathbf{U}}$ and minimizing $F_2(\mathbf{u};\,\underline{\mathbf{U}})$ over $\mathbf{u}\in H_0^1(\Omega)^n$. It is clear that the accuracy obtained in the first stage for $(\underline{\mathbf{U}},p)$ is more than enough to achieve similar accuracy in the second stage for $\mathbf{u}$, and that the second stage can be avoided if velocities/displacements are not needed. These important practical advantages are a result of the more general property that the coupling between $(\underline{\mathbf{U}},p)$ and $\mathbf{u}$ is subdominant in the sense of order of the associated differential operators (i.e., the second-order normal equations associated with the least-squares principle for the full system have only first-order differential operators appearing in the off-diagonal blocks connecting $(\underline{\mathbf{U}},p)$ and $\mathbf{u}$).

One advantage of FOSLS is the freedom to incorporate equations and boundary conditions in the functional or to impose them on the space. For example, we can impose the grad-trace equation $\nabla tr\,\underline{\mathbf{U}}+\delta\nabla\,p-\nabla g$ by restricting the space to the subset of variables that satisfy this equation, namely, we just minimize

$$F_1(\underline{\mathbf{U}},\,p;\,\mathbf{f},\,g)=\|\mathbf{f}+\nu\,(\nabla\cdot\underline{\mathbf{U}})^t-\nabla\,p\|^2+\nu^2\,\|\nabla\times\underline{\mathbf{U}}\|^2$$

over $\{(\underline{\mathbf{U}},\, p) \in \{\underline{\mathbf{U}} \in H^1(\Omega)^{n^2} \,:\, \mathbf{n} \times \underline{\mathbf{U}} = \underline{\mathbf{0}}\, on\ \partial\Omega,\, U_{11} + U_{22} + \delta p = g\, in\, \Omega\}$. We can accomplish this simply by eliminating $U_{22} = g - \delta p - U_{11}$. Note that this formulation means that *conservation is exactly satisfied in the sense of the velocity flux variables.*

FOSLS Formal Normal

Informal analysis of Stokes

To understand why the FOSLS functionals are product H^1 equivalent, consider the simple two-dimensional standard Stokes equations with $\delta = 0$ and U_{22} eliminated. Then it is straightforward to see that the differential operator for the Euler-Lagrange equations associated with $F_1(\underline{\mathbf{U}},\, p;\, \mathbf{f},\, g) = \|\mathbf{f} + \nu\,(\nabla \cdot \underline{\mathbf{U}})^t - \nabla\, p\|^2 + \nu^2\, \|\nabla \times \underline{\mathbf{U}}\|^2$ is of the form L^*L, where * signifies the operator adjoint and L is given by

$$L = \begin{pmatrix} \nu\partial_x & \nu\partial_y & 0 & \partial_x \\ -\nu\partial_y & 0 & \nu\partial_x & \partial_y \\ \nu\partial_y & -\nu\partial_x & 0 & 0 \\ \nu\partial_x & 0 & \nu\partial_y & 0 \end{pmatrix}.$$

This is where the formal part comes in: We simply assume that the boundary conditions allow us to write the adjoint of L as $-L^T$, and we freely exchange the order of derivatives:

$$L^* = \begin{pmatrix} -\nu\partial_x & \nu\partial_y & -\nu\partial_y & -\nu\partial_x \\ -\nu\partial_y & 0 & \nu\partial_x & 0 \\ 0 & -\nu\partial_x & 0 & -\nu\partial_y \\ -\partial_x & -\partial_y & 0 & 0 \end{pmatrix}.$$

We thus have that

$$L^*L = \begin{pmatrix} 2\nu^2\mathcal{L} & 0 & 0 & \nu(-\partial_x^2 + \partial_y^2) \\ 0 & \nu^2\mathcal{L} & 0 & -\nu\partial_{xy} \\ 0 & 0 & \nu^2\mathcal{L} & -\nu\partial_{xy} \\ \nu(-\partial_x^2 + \partial_y^2) & -\nu\partial_{xy} & -\nu\partial_{xy} & \mathcal{L} \end{pmatrix},$$

where $\mathcal{L} = -(\partial_x^2 + \partial_y^2)$ is the negative Laplacian. This *formal normal* has the block structure

$$\begin{pmatrix} \nu^2\mathcal{D} & -\nu X \\ -\nu X & \mathcal{L} \end{pmatrix},$$

where

$$\mathcal{D} = \begin{pmatrix} 2\mathcal{L} & 0 & 0 \\ 0 & \mathcal{L} & 0 \\ 0 & 0 & \mathcal{L} \end{pmatrix} \text{ and } X = \begin{pmatrix} \partial_x^2 - \partial_y^2 \\ \partial_{xy} \\ \partial_{xy} \end{pmatrix}.$$

The issue now is whether this system is diagonally dominant, that is, whether νX is small relative to the diagonal. If so, this system operator would clearly be dominated by individual Laplacians, giving the minimization problem the product H^1 equivalence we seek. To this end, note that the formal normal is equivalent to

$$\begin{pmatrix} I & -\mathcal{X} \\ -\mathcal{X}^* & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} 0 & \mathcal{X} \\ \mathcal{X}^* & 0 \end{pmatrix},$$

where $\mathcal{X} = \mathcal{D}^{-\frac{1}{2}} X \mathcal{L}^{-\frac{1}{2}}$. Our aim now is to show that the spectral radius of the last *cross term* here is much less than one, but that spectral radius is just the square root of the largest eigenvalue of $\mathcal{X}^*\mathcal{X}$, which (continuing with our free reordering of derivatives) is the square root of the maximum eigenvalue for the generalized eigenvalue problem

$$\left(\frac{1}{2}(\partial_x^2 - \partial_y^2)^2 + 2\partial_{xy}^2\right) v = \lambda \mathcal{L} v,$$

which is just

$$\frac{1}{2}\mathcal{L}v = \lambda \mathcal{L} v.$$

The maximum eigenvalue of the cross term is thus $\frac{\sqrt{2}}{2}$, which is indeed less than one (uniformly so).

FOSLS Myth Demeaners

FOSLS requires too much smoothness. The L^2-norm version of FOSLS generally does require full H^2 regularity to achieve product H^1 ellipticity, but it requires no more than mixed methods do if you are willing to settle for $H^{div} \cap H^{curl}$ ellipticity. Moreover, the inverse-norm version of FOSLS achieves ellipticity (albeit in a lower product Sobolev norm) without requiring full H^2 regularity.

Inverse-norm FOSLS is better than the L^2-norm FOSLS. The inverse-norm approach is quite a bit more expensive than the L^2-norm approach when they both apply. The inverse-norm approach has much more complexity–not to mention its general sensitivity to local changes in the problem that the inverse norm does not account for. It can yield an $O(N)$ method, but with a much bigger constant. Just one inverse-norm-preconditioned step can cost about as much as a total solve using FMG for the L^2-norm approach, which typically solves problems to the level of discretization error in 10-20 total work units. The inverse-norm approach cannot compete with this usuallyand it also loses the sharp a posteriori error measures that L^2-norm FOSLS readily provides.

FOSLS needs too many variables. First, the FOSLS inverse-norm approach does not necessarily require extra variables: you can simply eliminate the new variables by restricting the space to the right subspace. This amounts to just applying an inverse norm to the original system. Second, these new variables are often the higher

derivative quantities needed in practice, and seldom do other approaches obtain anywhere near the accuracy for them that FOSLS does. Finally, because it is often more accurate, FOSLS can typically solves problems using much coarser grids that other methods require, more than outweighing the typical doubling or so of the cost that extra variables incur.

FOSLS changes the physical meaning. Applying a least-squares principle to a first-order system may seem like it converts a hyperbolic equation like $p(x) = f(x), p(0) = 0$ to an elliptic problem. In fact, if you ignore boundary conditions, then the Euler-Lagrange equations associated with minimizing $F(p) = \|p - f\|^2$ (that is, $\nabla F(p) = 0$) is the elliptic-type equation $-p'' = f'$. This relationship is probably a source of this misconception. The real point is that the boundary conditions cannot be ignored: translating $minF(p)$ to the Euler-Lagrange equation properly must incorporate the boundary terms that arise from integration by parts. The real point is that applying a norm to the defect $p - f$ simply means that you are not changing the solution or the character of the equation, but rather just articulating how you measure error in the approximation. A more subtle misconception comes from how FOSLS may be applied to a convection-diffusion equation. Consider the simple scalar problem $-\epsilon p''(x) + p(x) = f(x)$, where $\epsilon > 0$is small. You can obtain uniform performance of standard finite elements and standard multigrid by just rewriting this equation in the elliptic form $(e^{-\epsilon x}p')' = e^{-\epsilon x}f$. It is easy to think that this basic step and its realization in a FOSLS functional may lose the perception that the original flow is essentially convective outside the boundary layer. Admittedly, this new elliptic form might require reinterpretation of the roles of convection and diffusion, and this unfamiliarity could be compounded by reformulating it in functional form. However, the flow is convective only at a certain physical scale, that is, only when ϵ/h is negligible, and the exponential rescaling used here really exposes this aspect of the physics. Close connection to the physics should come naturally as one gets used to the new FOSLS formulations.

FOSLS uses the wrong norm. Most methods cannot even explicitly say what norm they are using. Finite differences attempts to use a max norm, but this connection is weakened dramatically by the use of overly conservative truncation error and stability arguments. The truncation error argument is really an attempt to keep control (almost never optimality) over the discrete residual, which is like a discrete H^1 error norm for first-order problems or discrete H^2 error norm for second-order problems. However, such estimates relate only very loosely to the L^2 error norm via the inverse of the discrete operator. In any case, it seems misleading to say that FOSLS is fixated on the H^1 norm (or whatever norm FOSLS equivalence is established in) because the computational process can simply focus on other objectives. Just as is commonly done with any other approach, one can ignore the FOSLS norm and attempt to control other error criteria. For example, an attempt can be made to reign in the L^2 error by refining where the derivatives of the emerging solution are large. Another possibility is to rescale the functional to emphasize local regions of importance. For example, FOSLS for convection-dominated flow has obtained uni-

form performance of the discretization and multigrid solver. This uniform optimality is obtained by a special functional scaling that achieves uniform equivalence to an H^1 norm that is damped in the boundary layer. Although this is a natural scaling because the solution tends to be oscillatory in the boundary layer, if more accuracy is needed there, then the functional can be carefully rescaled to reflect this goal. Magnifying the div-curl functional in the boundary layer by a large constant does not expose div-curl harmonic error (zero div-curl residuals in the layer) any more than they are present outside the layer. But one cannot expect to control harmonic error locally. However, the question that does remain is: How do div-curl and H^1 harmonic errors relate?

Further Information

See the papers (and the references that they cite) available at

http://amath-www.colorado.edu/appm/faculty/stevem/Home.html