Stability Analysis of Scalar Advection-Diffusion Equation

M. Behr

http://manila.mems.rice.edu/developer/tn

Abstract

This note recounts detailed stability and accuracy analysis of a scalar advection-diffusion equation.

1. Introduction

Although stability and accuracy proofs are possible for most of the finite element formulations, they are often avoided because of their complexity. Two desirable properties—stability and consistency—result in convergent methods. Consistency, and to a lesser extent, stability, are typically easily determined for a given weak form, and are slightly more difficult for discrete (Galerkin or Petrov-Galerkin) forms. In contrast, the convergence behavior is much harder to analyze; yet this analysis is crucial for determining both the order of convergence and the optimal design of stabilization parameters.
In this note, we carry out the convergence analysis for a (simple!) scalar advection-diffusion equation to its bitter end.

2. Strong Form

Consider the scalar advection-diffusion equation:

$$\mathcal{L}u = f \qquad \text{on } \Omega \in \mathbb{R}^d, \tag{1}$$

$$u = 0 \qquad \text{on } \Gamma = \partial\Omega, \tag{2}$$

where

$$\mathcal{L}u \doteq \mathbf{a} \cdot \nabla u - \nabla \cdot \kappa \nabla u. \tag{3}$$

The following non-essential assumptions are made in order to simplify an already complex proof:

- the velocity field $\mathbf{a}$ satisfies $\nabla \cdot \mathbf{a} = 0$,
- κ is a positive constant.

We will also define the element Peclet number which quantifies the relative strength of advective or diffusive terms:

$$\alpha = \frac{h|\mathbf{a}|}{2\kappa}. \tag{4}$$

3. Galerkin Form

The Galerkin form is given as: Find $u^h \in \mathcal{V}^h \subset H_0^1$ such that:

$$B(w^h, u^h) = L(w^h) \qquad \forall w^h \in \mathcal{V}^h, \tag{5}$$

where

$$B(w^h, u^h) \doteq \int^{\Omega} \left(w^h \mathbf{a} \cdot \boldsymbol{\nabla} u^h + \boldsymbol{\nabla} w^h \cdot \kappa \boldsymbol{\nabla} u^h \right) d\Omega, \tag{6}$$

$$L(w^h) \doteq \int^{\Omega} w^h f d\Omega, \tag{7}$$

and $B(\cdot, \cdot)$ and $L(\cdot)$ are bilinear and linear forms, respectively, and H_0^1 denotes the Sobolev space of functions with square-integrable first derivatives and satisfying the homogenous boundary condition on Γ.

3.1. Analysis Outline

In obtaining the convergence estimates, it is necessary to first determine the *consistency* and *stability* of the discrete form. These two properties can be then applied to obtain bounds on the error of the discrete solution in terms of interpolation errors which are strictly related to element size only. At this final step, the proper form for the stabilization parameter, if any, should become apparent.

3.2. Consistency

The discrete form is consistent, if it is satisfied identically by the exact solution u. In other words:

$$B(w^h, u) = L(w^h) \qquad \forall w^h \in \mathcal{V}^h, \tag{8}$$

or

$$B(w^h, e) = 0 \qquad \forall w^h \in \mathcal{V}^h, \tag{9}$$

where $e = u^h - u$ is the *error*.

3.3. Stability

The discrete form is stable, if "small" deviations from the given "inputs"—boundary conditions, domain shape—results only in "small" deviations in the solution. A more precise statement of stability is: For each u^h, there exists w^h such that:

$$B(w^h, u^h) > 0 \qquad \text{if } \|u^h\| > 0, \tag{10}$$

That also means that any non-negligible component of the solution u^h will produce non-negligible change in the value of the bilinear form for at least one of the possible test functions w^h. The inequality in 10 must be strict.

Let's look at the stability of the Galerkin form. For each u^h we are looking for a corresponding w^h that would make the bilinear form non-zero. The best candidate is often u^h itself (or $-u^h$). But the Galerkin bilinear form gives:

$$\begin{aligned} B(u^h, u^h) &= \int_\Omega \left(u^h \mathbf{a} \cdot \nabla u^h + \nabla u^h \cdot \kappa \nabla u^h \right) d\Omega \\ &= -\int_\Omega \nabla \cdot \mathbf{a} \frac{(u^h)^2}{2} \, d\Omega + \int_\Gamma \frac{(u^h)^2}{2} \mathbf{a} \cdot \mathbf{n} \, d\Gamma \\ &\quad + \| \kappa^{\frac{1}{2}} \nabla u^h \|^2 \qquad \forall u^h \in \mathcal{V}^h. \end{aligned} \tag{11}$$

The first two terms on the right hand side are zero due to the assumptions we made (otherwise this reasoning can still proceed but with some extra effort). The problem with the Galerkin form is now clearly seen: as $\kappa \to 0$, its stability vanishes. For advection-dominated flows, the bilinear form can be arbitrarily small for non-zero u^h.
Note that this *instability* is not proven directly here; we are just *unable* to find a test function that would give us a lower bound on the form value. In some sense proving stability is easier than proving instability.

4. Artificial Diffusion

A crude way of gaining stability for advection-dominated flows has been the artificial diffusion (AD) method. The bilinear form becomes:

$$\begin{aligned} B_{\mathrm{AD}}(w^h, u^h) &\doteq \int_\Omega \left(w^h \mathbf{a} \cdot \nabla u^h + \nabla w^h \cdot \kappa \nabla u^h \right) d\Omega \\ &\quad + \int_\Omega \nabla w^h \cdot \bar{\kappa} \nabla u^h d\Omega, \end{aligned} \tag{12}$$

where $\bar{\kappa} = \mathcal{O}(h)$ is a positive parameter. Due to the presence of that parameter, the form is stable in the sense of the preceding section, even as $\kappa \to 0$, but it is not consistent. The convergence proof as shown below could not proceed.

5. Galerkin Least-Squares Form

Now let us see how the picture changes if a stabilized Galerkin Least-Squares (GLS) formulation is used. The bilinear form becomes:

$$\begin{aligned} B_{\mathrm{GLS}}(w^h, u^h) &\doteq \int_\Omega \left(w^h \mathbf{a} \cdot \nabla u^h + \nabla w^h \cdot \kappa \nabla u^h \right) d\Omega \\ &\quad + \sum_e \int_{\Omega^e} \underbrace{\left(\mathbf{a} \cdot \nabla w^h - \nabla \cdot \kappa \nabla w^h \right)}_{\mathcal{L} w^h} \tau \mathcal{L} u^h d\Omega \end{aligned} \tag{13}$$

and it clearly remains consistent. To determine stability, we look at:

$$\begin{aligned} B_{\text{GLS}}(u^h, u^h) &= \int_\Omega \left(u^h \mathbf{a} \cdot \boldsymbol{\nabla} u^h + \boldsymbol{\nabla} u^h \cdot \kappa \boldsymbol{\nabla} u^h \right) d\Omega \\ &+ \sum_e \int_{\Omega^e} \mathcal{L} u^h \tau \mathcal{L} u^h d\Omega \\ &= \| \kappa^{\frac{1}{2}} \boldsymbol{\nabla} u^h \|^2 + \| \tau^{\frac{1}{2}} \mathcal{L} u^h \|_h^2 \equiv |||u^h|||^2 \qquad \forall u^h \in \mathcal{V}^h. \end{aligned} \tag{14}$$

The expression $B_{\text{GLS}}(u^h, u^h)$ is thus seen to define a *norm*, which we will denote as $|||u^h|||^2$. This is equivalent with stability. Why do we only include element interiors in the integration in the stabilization term? If the integral was over the entire domain, we would have to require the first derivatives of the interpolation functions to be continuous across the element boundaries, so that second derivatives remain square-integrable even when the element boundaries are included.

6. Convergence

Armed with consistency and stability, we may proceed with the convergence analysis. We will split the solution error into two components:

$$e = \underbrace{u^h - \tilde{u}^h}_{e^h \in \mathcal{V}^h} + \underbrace{\tilde{u}^h - u}_{\eta \in \mathcal{H}_0^1}, \tag{15}$$

where e^h is the *discrete error*, η is the *interpolation error*, and $\tilde{u}^h$ is the *interpolant*, i.e. a function in the trial function space which coincides at the nodes with the exact solution. In one sense, interpolant is as close as we can expect to get to the real solution (although not necessarily in the sense of the norms we will use). Then interpolation error is an error that "cannot be avoided"; it is dependent only on the mesh size, and not on the discretization method we employ. The discrete error is tied to the discretization method itself, and we hope to obtain bounds for this error which are similar to the existing bounds on the interpolation error. The two bounds will combine to give us the following bound on the total error:

$$|||e||| = \begin{cases} \mathcal{O}(h^{k+\frac{1}{2}}) & \text{advective limit} \\ \mathcal{O}(h^k) & \text{diffusive limit} \end{cases} \tag{16}$$

This bound will only be possible if τ is appropriately defined. Since we are always dealing with the GLS form from now on, the GLS subscript will be dropped from the bilinear form.

$$
\begin{aligned}
|||e^h|||^2 &= B(e^h, e^h) \\
&= B(e^h, e) - B(e^h, \eta) \\
&\quad \text{(first term is zero by consistency)} \\
&= \left| \int_\Omega e^h \mathbf{a} \cdot \boldsymbol{\nabla} \eta d\Omega + \underbrace{\int_\Omega \boldsymbol{\nabla} e^h \cdot \kappa \boldsymbol{\nabla} \eta d\Omega}_{3} + \underbrace{\sum_e \int_{\Omega^e} \mathcal{L} e^h \tau \mathcal{L} \eta d\Omega}_{4} \right| \\
&\quad \text{(integrating by parts and adding and subtracting a term)} \\
&= \left| - \int_\Omega \mathbf{a} \cdot \boldsymbol{\nabla} e^h \eta d\Omega + \int_\Gamma \mathbf{a} \cdot \mathbf{n} e^h \eta d\Gamma \right. \\
&\quad + \sum_e \int_{\Omega^e} \left(\boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} e^h \right) \eta d\Omega - \sum_e \int_{\Omega^e} \left(\boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} e^h \right) \eta d\Omega \\
&\quad \left. + \quad 3 \quad + \quad 4 \quad \right| \\
&\quad \text{(boundary term is zero due to boundary conditions)} \\
&= \left| \underbrace{- \sum_e \int_{\Omega^e} \mathcal{L} e^h \eta d\Omega}_{1} \underbrace{- \sum_e \int_{\Omega^e} \left(\boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} e^h \right) \eta d\Omega}_{2} \right. \\
&\quad \left. + \quad 3 \quad + \quad 4 \quad \right| \\
&= \left| \quad 1 \quad + \quad 2 \quad + \quad 3 \quad + \quad 4 \quad \right|. \qquad (17)
\end{aligned}
$$

All four terms are "mixed" in the sense that they contain products of expressions in e^h and η. We will try to make them bounded by a sum of expressions in e^h alone and in η alone. Then the e^h terms will be absorbed by the left hand side $|||e^h|||^2$, and the bound will depend on the interpolation error alone! For terms 1 and 2, we will use this identity:

$$
\left(\frac{\tau^{\frac{1}{2}} a}{2} - \tau^{-\frac{1}{2}} b \right)^2 = \frac{\tau a^2}{4} + \frac{b^2}{\tau} - ab > 0, \qquad (18)
$$

and for terms 3 and 4, another identity:

$$\left(\frac{a}{2}-b\right)^2 = \frac{a^2}{4} + b^2 - ab > 0. \tag{19}$$

to obtain bounds on ab. Consequently,

$$1 \leq \frac{1}{4}\|\tau^{\frac{1}{2}}\mathcal{L}e^h\|_h^2 + \|\tau^{-\frac{1}{2}}\eta\|^2 \tag{20}$$

$$+$$

$$2 \leq \frac{1}{4}\|\tau^{\frac{1}{2}}\left(\boldsymbol{\nabla}\cdot\kappa\boldsymbol{\nabla}e^h\right)\|_h^2 + \|\tau^{-\frac{1}{2}}\eta\|^2 \tag{21}$$

$$+$$

$$3 \leq \frac{1}{4}\|\kappa^{\frac{1}{2}}\boldsymbol{\nabla}e^h\|^2 + \|\kappa^{\frac{1}{2}}\boldsymbol{\nabla}\eta\|^2 \tag{22}$$

$$+$$

$$4 \leq \underbrace{\frac{1}{4}\|\tau^{\frac{1}{2}}\mathcal{L}e^h\|_h^2}_{\text{absorbed by } |||e^h|||^2} + \|\tau^{\frac{1}{2}}\mathcal{L}\eta\|_h^2. \tag{23}$$

The only e^h term that does not have a matching term inside $|||e^h|||^2$ is the first of the two terms in (22); for an appropriate definition of τ (forthcoming), it can be replaced in the inequality using the *inverse estimate*:

$$\|\tau^{\frac{1}{2}}\left(\boldsymbol{\nabla}\cdot\kappa\boldsymbol{\nabla}e^h\right)\|_h^2 \leq \|\kappa^{\frac{1}{2}}\boldsymbol{\nabla}e^h\|^2. \tag{24}$$

The first set of terms on the right-hand side of the inequalities add up to $\frac{1}{2}|||e^h|||^2$, which can be subtracted from the left-hand side. The remaining right-hand side *depends only on the interpolation error*:

$$\frac{1}{2}|||e^h|||^2 \leq 2\|\tau^{-\frac{1}{2}}\eta\|^2 + \underbrace{\left(\|\kappa^{\frac{1}{2}}\boldsymbol{\nabla}\eta\|^2 + \|\tau^{\frac{1}{2}}\mathcal{L}\eta\|_h^2\right)}_{|||\eta|||^2}, \tag{25}$$

or

$$|||e^h|||^2 \leq 4\|\tau^{-\frac{1}{2}}\eta\|^2 + 2|||\eta|||^2. \tag{26}$$

Finally, the total error can be bounded by:

$$|||e|||^2 \leq 2\left(|||e^h|||^2 + |||\eta|||^2\right) \leq 8\|\tau^{-\frac{1}{2}}\eta\|^2 + 6|||\eta|||^2. \tag{27}$$

From interpolation theory it is known that interpolation errors using polynomials of order k obey:

$$|\eta|_m = \mathcal{O}(h^{k+1-m}), \tag{28}$$

as long as elements are *quasi-uniform*, or not too distorted. Therefore, the norms on the right-hand side of (27), are of the order:

$$|||\eta|||^2 = \kappa\mathcal{O}(h^{2k}) + \tau\kappa\mathcal{O}(h^{2(k-1)}) \qquad \text{diffusive limit,} \tag{29}$$

$$|||\eta|||^2 = \qquad |\mathbf{a}|\tau\mathcal{O}(h^{2k}) \qquad \text{advective limit,} \tag{30}$$

$$\|\tau^{-\frac{1}{2}}\eta\|^2 = \qquad \frac{1}{\tau}\mathcal{O}(h^{2(k+1)}). \tag{31}$$

We can see that τ should be a polynomial in h to improve the last, weak, bound in (29), but of order low enough not to destroy the bound in (31). The optimal convergence (16) is obtained with this τ design:

$$\tau = \mathcal{O}(\frac{h}{|\mathbf{a}|}), \qquad \text{advective limit,} \tag{32}$$

$$\tau = \mathcal{O}(\frac{h^2}{\kappa}), \qquad \text{diffusive limit.} \tag{33}$$

Serendipitously (finite element people like this word), with this choice, the inequality (24) also holds. Note that (16) does not indicate that if we use linear interpolation functions, the solution will converge e.g. linearly in the diffusive limit. The norm used involves derivatives, and it is the first derivatives that will converge linearly. The solution itself will then converge quadratically in the diffusive limit.

7. Inverse Estimate

For 1D linear element, the inverse estimate can be stated as:

$$\frac{1}{C_0}||q||_0^2 \;\geq\; \sum_K h_K^2 \left|\left|\nabla q\right|\right|_0^2\Big|_K, \tag{34}$$

$$\frac{1}{C_0}\left|\left|q\right|\right|_0^2\Big|_K \;\geq\; h_K^2 \left|\left|\nabla q\right|\right|_0^2\Big|_K. \tag{35}$$

Denoting by $\bar{q}$ the average value of q in the element (at midpoint), and by q' the range of q within that element, we can write:

$$\left|\left|q\right|\right|_0^2\Big|_K = \int_0^{h_K}\left(\bar{q} - \frac{q'}{2} + \frac{q'\xi}{h_K}\right)^2 d\xi = \left(\bar{q}^2 + \frac{q'^2}{12}\right)h_K, \tag{36}$$

$$\left|\left|\nabla q\right|\right|_0^2\Big|_K = \int_0^{h_K}\left(\frac{q'}{h_K}\right)^2 d\xi = \frac{q'^2}{h_K}, \tag{37}$$

$$\frac{1}{C_0} \left|\left| q \right|\right|_0^2 \Big|_K = \frac{1}{C_0} \left(\bar{q}^2 + \frac{q'^2}{12} \right) h_K \geq \frac{1}{C_0} \frac{q'^2}{12} h_K \geq \frac{q'^2}{h_K} h_K^2 = h_K^2 \left|\left| \nabla q \right|\right|_0^2 \Big|_K . \tag{38}$$

The second-to-last inequality is satisfied if $C_0 = \frac{1}{12}$.

History

May 11, 2001	Written based on the Tom Hughes' notes.
November 21, 2001	Added inverse estimate section.
June 6, 2004	Corrected Eq. (31).