# Chapter 2

# One-Dimensional Model Problem: Steady Case

## 2.1 The Steady Advection-Diffusion Equation

### 2.1.1 Problem Statement

The steady one-dimensional advection-diffusion equation is given by

$$u\phi_{,x} = \kappa\phi_{,xx} + f \tag{2.1}$$

where

$$\phi_{,x} \equiv \frac{\partial\phi}{\partial x} \quad \text{and} \quad \phi_{,xx} \equiv \frac{\partial^2\phi}{\partial x^2}\,. \tag{2.2}$$

The advective velocity $u$ and the diffusivity coefficient $\kappa$ are assumed to be given positive constants. The source term $f$ is a given function defined on $[0, L] \subset \Re$. The objective is to solve (2.1) for $\phi = \phi(x)$, where $x \in [0, L]$. The solution is not unique unless we specify appropriate boundary conditions. An example of well-posed boundary conditions is the Dirichlet case:

$$\left\{ \begin{array}{rcl} \phi(0) & = & g_0 \\ \phi(L) & = & g_L \end{array} \right. \tag{2.3}$$

where $g_0$ and $g_L$ are given constants.

### 2.1.2 Exact Solution

The exact solution of the one-dimensional advection-diffusion equation
without source term (i.e., $f \equiv 0$) is

$$\phi(x) = c_1 + c_2 \exp\left(\text{Pe}\,\frac{x}{L}\right) \tag{2.4}$$

where

- $c_1$ and $c_2$ are constants determined by the boundary conditions.
- $\mathrm{Pe} = uL/\kappa$ is the **Péclet number**.

The Péclet number represents the competition between advection and diffusion, i.e.,

- The flow is advection dominated when $\mathrm{Pe} > 1$.
- The flow is diffusion dominated when $\mathrm{Pe} < 1$.

Exact solutions are sketched in Figure 2.1 for the case $g_0 = 0$, $g_L = 1$, and $L = 1$.

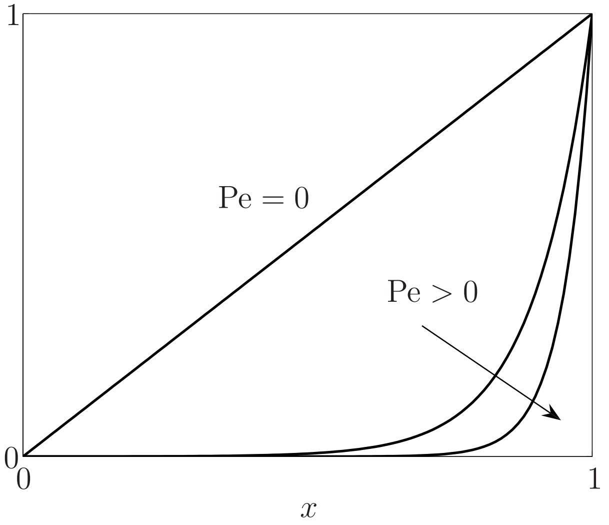


Figure 2.1: Sketches of the exact solution for steady advection-diffusion in one dimension.

## 2.2 Classical Finite Difference Methods

Consider a uniform mesh of $N$ segments of size $h$ (see Fig. 2.2).

### 2.2.1 Central Difference Approximation

At node $A$, the differential operators are approximated by difference quotients given by:

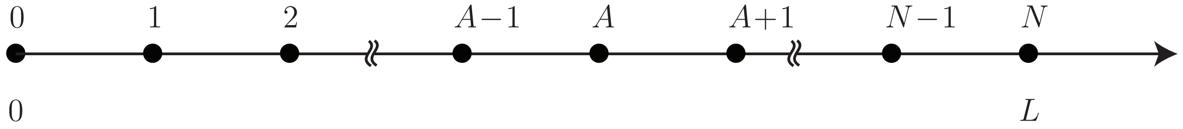


Figure 2.2: Uniform mesh for the one-dimensional advection-diffusion problem.

$$u\phi_{,x}(x_A) \approx u\frac{\phi_{A+1} - \phi_{A-1}}{2h} \tag{2.5}$$

and

$$\kappa\phi_{,xx}(x_A) \approx \kappa\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} \tag{2.6}$$

where $\phi_A$ is the approximation of $\phi(x_A)$. Thus, the problem to solve is:

$$u\frac{\phi_{A+1} - \phi_{A-1}}{2h} = \kappa\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} + f_A \tag{2.7}$$

for $A = 1, 2, ..., N-1$ with $f_A = f(x_A)$. This is a system of $N-1$ equations in $N-1$ unknowns.

### 2.2.2 Upwind Difference Approximation

At node $A$, the first-order differential operator is approximated by

$$u\phi_{,x}(x_A) \approx u\frac{\phi_A - \phi_{A-1}}{h} \tag{2.8}$$

The difference quotient is taken "upwind" of the point of reference, $x_A$. The second-order differential operator is the same as for central differences. Thus, the problem to solve is:

$$u\frac{\phi_A - \phi_{A-1}}{h} = \kappa\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} + f_A \tag{2.9}$$

for $A = 1, 2, ..., N-1$.

### 2.2.3 Solution of the Approximate Problems

Both of these approximations can be written in the form

$$a\phi_{A+1} - 2b\phi_A + c\phi_{A-1} = f_A \tag{2.10}$$

where $b = \frac{1}{2}(a+c)$ (we will verify this momentarily). Consider the homogeneous case (i.e., $f \equiv 0$) and assume that the solution of (2.10) has the form

$$\phi_A \sim \zeta^A \tag{2.11}$$

where $\zeta$ is to be found. Substitution of (2.11) into (2.10) leads to

$$a\zeta^{A+1} - 2b\zeta^{A} + c\zeta^{A-1} = 0 \;\Leftrightarrow\; a\zeta^2 - 2b\zeta + c = 0 \tag{2.12}$$

Thus,

$$\zeta = \frac{b \pm \sqrt{b^2 - ac}}{a} \tag{2.13}$$

Replacing $b$ by $\frac{1}{2}(a+c)$ in (2.13) yields

$$\zeta = \frac{(a+c) \pm (a-c)}{2a}. \tag{2.14}$$

Consequently, $\zeta_1 = 1$ and $\zeta_2 = c/a$, and thus the general solution is

$$\phi_A = c_3 \cdot 1 + c_4 \left(\frac{c}{a}\right)^A \tag{2.15}$$

where $c_3$ and $c_4$ are to be determined by the two Dirichlet boundary conditions.

This solution is applicable to the central and upwind difference approximations:

- In central differences, we have

$$\left.\begin{aligned} a &= \frac{u}{2h} - \frac{\kappa}{h^2} \\ b &= -\frac{\kappa}{h^2} \\ c &= -\frac{u}{2h} - \frac{\kappa}{h^2} \end{aligned}\right\} \tag{2.16}$$

(Observe that the constraint $b = \frac{1}{2}(a+c)$ is satisfied.)
Thus,

$$\begin{aligned} \frac{c}{a} &= \left(-\frac{u}{2h} - \frac{\kappa}{h^2}\right) / \left(\frac{u}{2h} - \frac{\kappa}{h^2}\right) \\ &= \frac{1+\alpha}{1-\alpha} \end{aligned} \tag{2.17}$$

where $\alpha \equiv \dfrac{uh}{2\kappa}$ is the ***mesh*** (or ***"element"***) ***Péclet number***. The solution of the approximate problem is then

$$\phi_A = c_3 + c_4 \left(\frac{1+\alpha}{1-\alpha}\right)^A \tag{2.18}$$

**Remark 2.1** *If $\alpha > 1$ ("numerically advection dominated"), the numerical solution (2.18) oscillates. In fact, for $\alpha \gg 1$, the term in parentheses in (2.18) is approximately $-1$. This is qualitatively inconsistent with the exact solution (cf. (2.4)).*

- In upwind differences, we have

$$\left.\begin{aligned} a &= -\frac{\kappa}{h^2} \\ b &= -\frac{u}{2h} - \frac{\kappa}{h^2} \\ c &= -\frac{u}{h} - \frac{\kappa}{h^2} \end{aligned}\right\} \tag{2.19}$$

(Again, it is apparent that the condition $b = \frac{1}{2}(a + c)$ holds.) This time

$$\begin{aligned} \frac{c}{a} &= \left(-\frac{u}{h} - \frac{\kappa}{h^2}\right) / \left(-\frac{\kappa}{h^2}\right) \\ &= (1 + 2\alpha) \end{aligned} \tag{2.20}$$

The solution of the approximate problem is then

$$\phi_A = c_3 + c_4(1 + 2\alpha)^A \tag{2.21}$$

**Remark 2.2** *$\phi_A$ is **strictly monotone** for all $\alpha > 0$. Unfortunately, this method converges very slowly.*

Figures 2.3(a) and 2.3(b) show a comparison of the difference approximations with the exact solution for the source-free case with $N = 10$, $g_0 = 0$ and $g_L = 1$. The exact solutions and difference approximations are linearly interpolated between nodes. One can see that upwind differences are overly diffusive whereas central differences are overly anti-diffusive.

### 2.2.4 Artificial Diffusion Interpretation of Upwind Differences

The upwind difference stencil for the first-order differential operator can be written in the form

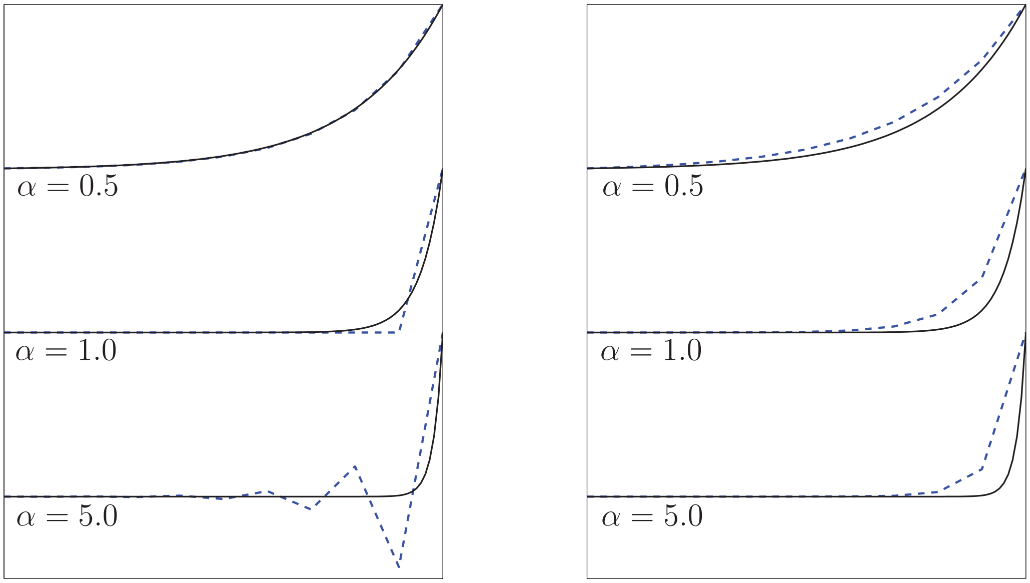
$$u\frac{\phi_A - \phi_{A-1}}{h} = u\frac{\phi_{A+1} - \phi_{A-1}}{2h} - \frac{uh}{2}\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} \tag{2.22}$$

Thus, the upwind difference method becomes:

$$u\frac{\phi_{A+1} - \phi_{A-1}}{2h} = \left(\kappa + \frac{uh}{2}\right)\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} + f_A \tag{2.23}$$

The "effective diffusion" $\kappa + \frac{uh}{2} = \kappa(1 + \alpha)$ is the sum of the physical diffusion $\kappa$ and an ***artificial diffusion*** $\frac{uh}{2}$.

When advection dominates (i.e. $\alpha \gg 1$), the artificial diffusion is much greater than physical diffusion. This does not mean the method is a disaster for the model problem because the artificial diffusion is *relative* to the central difference method, which is very anti-diffusive with respect to the exact solution.



(a) Central differences (overly anti-diffusive).

(b) Upwind differences (overly diffusive).

Figure 2.3: Steady one-dimensional advection-diffusion.

### 2.2.5 Exact Artificial Diffusion Method

We consider a new finite difference scheme in the form:

$$u\frac{\phi_{A+1} - \phi_{A-1}}{2h} = (\kappa + \tilde{\kappa})\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} \tag{2.24}$$

where $\tilde{\kappa}$ is to be determined.

**Remark 2.3** *This is the central difference method with $\kappa$ replaced by $\kappa + \tilde{\kappa}$.*

Therefore,

$$\left.\begin{aligned} a &= \frac{u}{2h} - \frac{\kappa + \tilde{\kappa}}{h^2} \\ b &= -\frac{\kappa + \tilde{\kappa}}{h^2} \\ c &= -\frac{u}{2h} - \frac{\kappa + \tilde{\kappa}}{h^2} \end{aligned}\right\} \tag{2.25}$$

and one can easily check again that $b = \frac{1}{2}(a + c)$. Hence,

$$\phi_A = c_3 + c_4 \left(\frac{1+\tilde{\alpha}}{1-\tilde{\alpha}}\right)^A \tag{2.26}$$

where $\tilde{\alpha}$ is given by the expression for $\alpha$ with $\kappa$ being replaced by $\kappa + \tilde{\kappa}$.

Let us now equate $\phi_A$ and the exact solution:

$$c_3 + c_4 \left(\frac{1+\tilde{\alpha}}{1-\tilde{\alpha}}\right)^A = c_1 + c_2 \exp(\mathrm{Pe}\,\frac{x_A}{L}) \tag{2.27}$$

Recall that for our mesh and node numbering (see Fig. 2.2), we have

$$x_A = Ah \tag{2.28}$$

Let us substitute (2.28) into (2.27) to get:

$$c_3 + c_4 \left(\frac{1+\tilde{\alpha}}{1-\tilde{\alpha}}\right)^A = c_1 + c_2 \left(\exp(\mathrm{Pe}\,\frac{h}{L})\right)^A \tag{2.29}$$

Thus, for equality to hold,

$$\begin{aligned} c_3 &= c_1 \\ c_4 &= c_2 \\ \left(\frac{1+\tilde{\alpha}}{1-\tilde{\alpha}}\right)^A &= \left(\exp(\mathrm{Pe}\,\frac{h}{L})\right)^A \end{aligned} \tag{2.30}$$

Recalling the definitions of the Péclet number Pe and the element Péclet number $\alpha$, from (2.30) we get

$$1 + \tilde{\alpha} = (1 - \tilde{\alpha})e^{2\alpha} \tag{2.31}$$

which results in

$$\begin{aligned} \tilde{\alpha} &= \frac{e^{2\alpha} - 1}{e^{2\alpha} + 1} \\ &= \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}} \\ &= \tanh\alpha \end{aligned} \tag{2.32}$$

By definition, $\tilde{\alpha}$ can be written as

$$\tilde{\alpha} \equiv \frac{uh}{2(\kappa + \tilde{\kappa})} = 1/\left(\frac{2\kappa}{uh} + \frac{2\tilde{\kappa}}{uh}\right) = 1/\left(\frac{1}{\alpha} + \tilde{\xi}\right) \tag{2.33}$$

where $\tilde{\xi}$ is a dimensionless parameter defined by

$$\tilde{\xi} \equiv \frac{2\tilde{\kappa}}{uh} \tag{2.34}$$

From (2.32) and (2.33), we obtain

$$\boxed{\tilde{\xi} = \coth\alpha - \frac{1}{\alpha}} \tag{2.35}$$

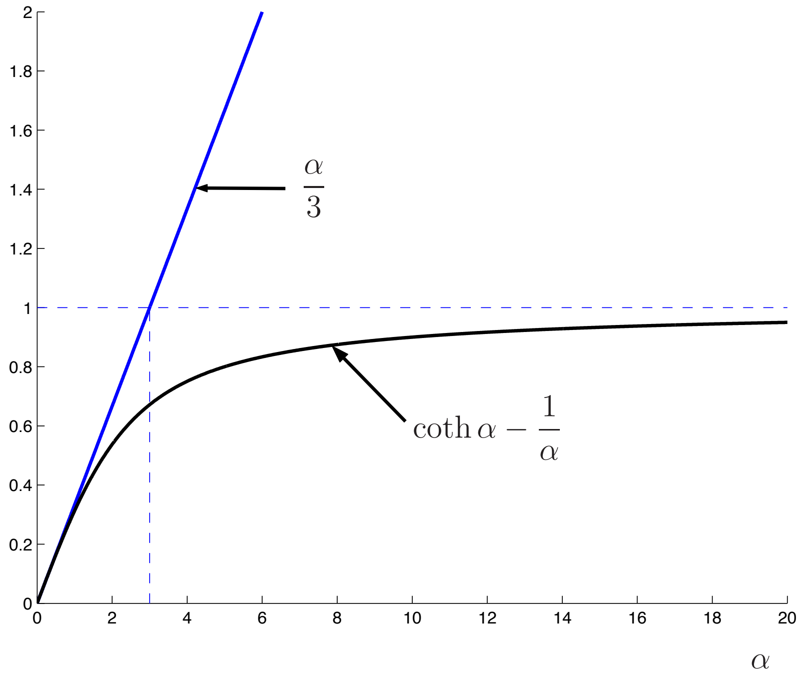


Figure 2.4: Optimal $\tilde{\xi}$.

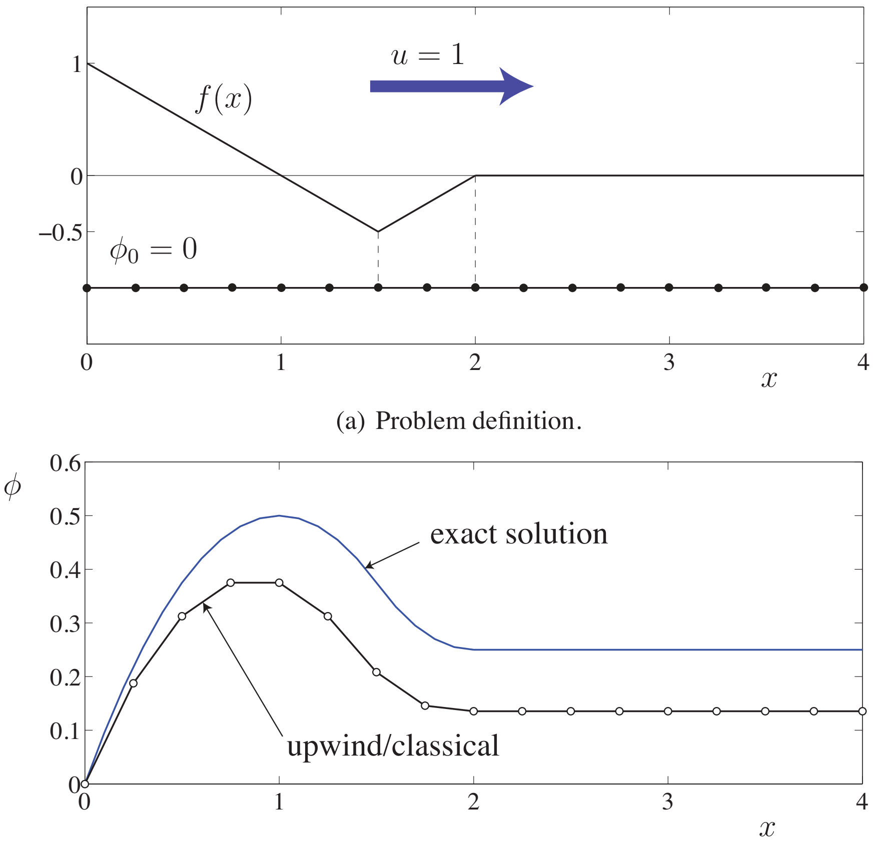
In Figure 2.4, $\tilde{\xi}$ is plotted versus $\alpha$. We see that $\tilde{\xi} \sim \alpha/3$ as $\alpha \to 0$.

**Remark 2.4** *Central differences correspond to* $\tilde{\kappa} = 0 \Leftrightarrow \tilde{\xi} = 0$.

**Remark 2.5** *Upwind differences correspond to* $\tilde{\kappa} = \frac{1}{2}uh \Leftrightarrow \tilde{\xi} = 1$.

**Remark 2.6** *The effective diffusion for upwind differences was observed to be* $\kappa(1+\alpha)$. *For the exact artificial diffusion method, the effective diffusion becomes:*

$$\kappa(1+\frac{\tilde{\kappa}}{\kappa}) = \kappa(1+\frac{uh}{2\kappa}\tilde{\xi}) = \kappa(1+\alpha\tilde{\xi}) \tag{2.36}$$



(a) Problem definition.

(b) Comparison of results.

Figure 2.5: Pure advection with a non-constant source term.

**Remark 2.7** *We have worked so far with the restriction that $u$ was a positive number. We can relax this restriction by letting,*

$$\alpha = \frac{|u|h}{2\kappa} \qquad , \qquad \tilde{\kappa} = \frac{|u|h}{2}\tilde{\xi} \tag{2.37}$$

The exact artificial diffusion method is an example of a ***modified equation approach***. Consider the following modified differential equation,

$$u\phi_{,x} = (\kappa + \tilde{\kappa})\phi_{,xx} \tag{2.38}$$

If we apply the central difference method to (2.38), the result is the exact artificial diffusion method given by eq. (2.24).

The method is only exact for $f = 0$. For the case $f \neq 0$, accuracy may degrade dramatically (see Fig. 2.5). This is often the case with methods based on modified equations. We shall see that the same thing happens later when we consider the time-dependent case.

## 2.3 Finite Element Methods

### 2.3.1 Variational Formulation

Let $\mathcal{S}$ be the set of ***trial solutions***: a collection of functions satisfying the Dirichlet boundary conditions present (i.e., $\phi(0) = g_0,\ \phi(L) = g_L$), and certain technical differentiability conditions that we will address later on. Let $\mathcal{V}$ be the set of ***weighting functions***: a collection of functions satisfying the homogeneous counterparts of the Dirichlet boundary conditions (i.e., $w(0) = 0,\ w(L) = 0$), and likewise certain technical differentiability conditions.

#### Weak Form of the Boundary Value Problem

A ***weak form*** of (2.1) can be stated as: Find $\phi \in \mathcal{S}$, such that for all $w \in \mathcal{V}$, we have

$$\int_0^L (-w_{,x} u\phi + w_{,x}\kappa\phi_{,x})\, dx = \int_0^L wf\, dx \tag{2.39}$$

(2.39) is also called a ***variational equation***.

The Dirichlet boundary conditions require:

$$\begin{array}{rclcrcl} \phi(0) & = & g_0 & , & \phi(L) & = & g_L \\ w(0) & = & 0 & , & w(L) & = & 0 \end{array} \tag{2.40}$$

#### Implications of the Weak Form

Integrating (2.39) by parts leads to

$$\int_0^L \left(wu\phi_{,x} - w\kappa\phi_{,xx}\right)\, dx = \int_0^L wf\, dx \tag{2.41}$$

Rearranging,

$$\int_0^L w\left(u\phi_{,x} - \kappa\phi_{,xx} - f\right)\, dx = 0 \tag{2.42}$$

Since $w$ is arbitrary by design, we must have

$$u\phi_{,x} - \kappa\phi_{,xx} - f = 0 \tag{2.43}$$

Thus, the advection-diffusion equation is satisfied by $\phi$, and by definition, $\phi \in \mathcal{S}$ satisfies the Dirichlet boundary condition.

**Remark 2.8** *The weak form is an alternative statement of the boundary value problem. Finite element methods are constructed from the weak form by replacing $\mathcal{S}$ and $\mathcal{V}$ by **finite-dimensional collections of functions**, an example of which will be described shortly.*

**Toward Finite Elements**

Let $\mathcal{S}^h \subset \mathcal{S}$ be a finite-dimensional collection of trial solutions, satisfying the Dirichlet boundary conditions. Let $\mathcal{V}^h \subset \mathcal{V}$ be a finite-dimensional collection of weighting functions, satisfying the homogeneous counterparts of the Dirichlet boundary conditions. The statement of the problem becomes: Find $\phi^h \in \mathcal{S}^h$, such that for all $w^h \in \mathcal{V}^h$, we have

$$\int_0^L (-w^h_{,x} u \phi^h + w^h_{,x} \kappa \phi^h_{,x})\, dx = \int_0^L w^h f\, dx \tag{2.44}$$

### 2.3.2 Galerkin Method

The Galerkin method is the most widely used and known basis of finite element discretizations. Once again, we consider the mesh of Figure 2.2, except that now the spacing between nodes need *not* be uniform. Let,

$$\begin{aligned} w^h &= \textstyle\sum_{A=0}^N N_A(x) w_A \\ \phi^h &= \textstyle\sum_{A=0}^N N_A(x) \phi_A \end{aligned} \tag{2.45}$$

where the $N_A$'s are called ***shape functions*** or ***interpolation functions***. There are many possibilities for shape functions; we will consider for the moment only the case of piecewise linear shape functions, shown in Figure 2.6.

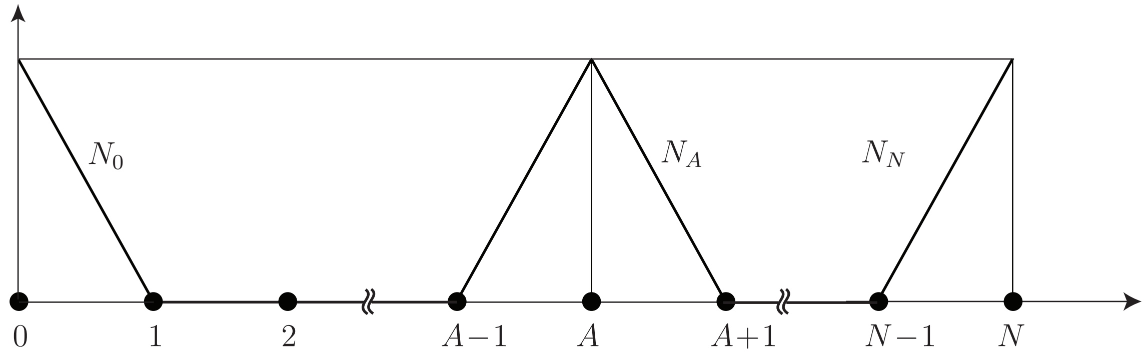


Figure 2.6: Basis functions for the piecewise linear finite element space.

The shape functions have the ***interpolation property***

$$N_A(x_B) = \delta_{AB} = \begin{cases} 1, & \text{if} \quad A = B \\ 0, & \quad\text{otherwise} \end{cases} \tag{2.46}$$

At a node $B$, we have

$$\phi^h(x_B) = \sum_A N_A(x_B)\phi_A = \sum_A \delta_{AB}\phi_A = \phi_B \tag{2.47}$$

The Dirichlet boundary conditions require

$$\begin{array}{rclcrcl} \phi_0 &=& g_0 &,& \phi_N &=& g_L \\ w_0 &=& 0 &,& w_N &=& 0 \end{array} \tag{2.48}$$

A typical function $w^h$ is shown in Figure 2.7.

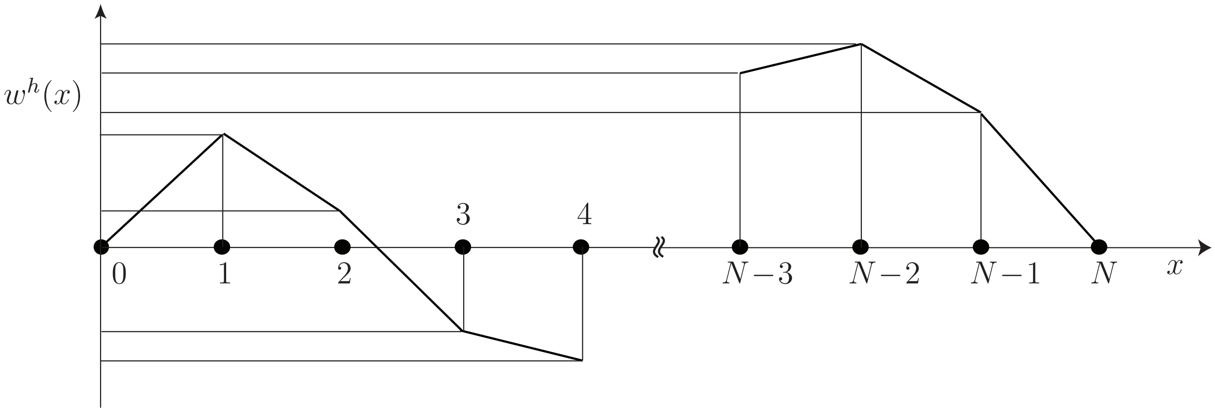


Figure 2.7: A typical member $w^h \in \mathcal{V}^h$.

The functions $\phi^h$ and $w^h$ are continuous (i.e., $C^0$), but they are *not* continuously differentiable (i.e., *not* $C^1$). This is the simplest example of a continuous finite element space. There are many other possibilities (see Hughes [117, Chapter 3]).

Substituting $\phi^h$ and $w^h$ given in (2.45) into (2.44) leads to

$$\sum_{A=1}^{N-1} w_A G_A = 0 \qquad \forall\; w^h = \sum_{A=1}^{N-1} w_A N_A \;\in \mathcal{V}^h \tag{2.49}$$

where

$$G_A = \sum_{B=0}^{N} \left\{ \int_0^L \left( -N_{A,x} u N_B + N_{A,x} \kappa N_{B,x} \right)\, dx \right\} \phi_B - \int_0^L N_A f \; dx \tag{2.50}$$

Due to the arbitrariness of the $w_A$'s, (2.49) implies

$$G_A = 0 \qquad A = 1, \ldots, N-1 \tag{2.51}$$

Let

$$\left\{ \begin{array}{rcl} K_{AB} &=& \int_0^L \left( -N_{A,x} u N_B + N_{A,x} \kappa N_{B,x} \right)\, dx \\ \\ F_A &=& \int_0^L N_A f \; dx \end{array} \right. \tag{2.52}$$

Thus,

$$\sum_{B=0}^{N} K_{AB} \phi_B = F_A \qquad A = 1, \ldots, N-1 \tag{2.53}$$

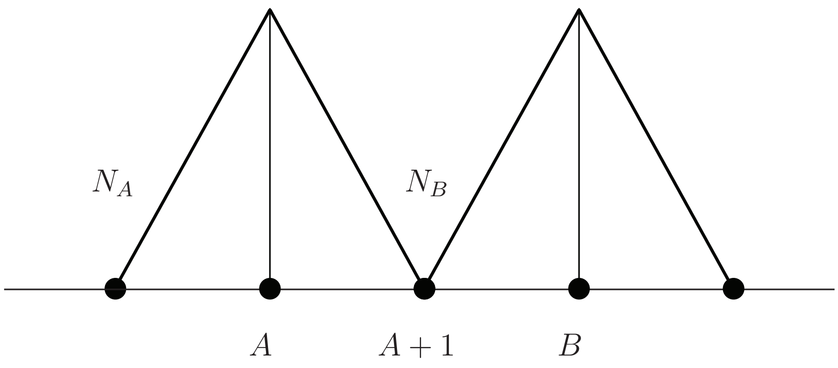


Figure 2.8: Tridiagonal matrix structure and shape function overlap. If $B > A+1$, the non-zero portions of $N_B$ and $N_A$ do not overlap.

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & 0 & & \cdots & 0 \\ K_{21} & K_{22} & K_{22} & & & \\ 0 & K_{32} & K_{33} & \ddots & & \vdots \\ \vdots & & \ddots & K_{N-3,N-3} & K_{N-3,N-2} & 0 \\ & & & K_{N-2,N-3} & K_{N-2,N-2} & K_{N-2,N-1} \\ 0 & & \ldots & 0 & K_{N-1,N-2} & K_{N-1,N-1} \end{bmatrix}$$

Figure 2.9: Schematic of the band structure of the tridiagonal matrix $\mathbf{K}$ obtained from piecewise linear finite elements.

Taking into account the Dirichlet boundary conditions gives

$$\sum_{B=1}^{N-1} K_{AB}\phi_B = R_A \qquad A = 1, \dots, N-1 \tag{2.54}$$

where

$$\left\{ \begin{array}{rcl} R_A & = & F_A - K_{A0}\phi_0 - K_{AN}\phi_N \\ & = & F_A - K_{A0}g_0 - K_{AN}g_L \end{array} \right. \tag{2.55}$$

This is a system of $N-1$ equations in $N-1$ unknowns which can be written in the following matrix form:

$$\mathbf{K}\boldsymbol{\phi} = \mathbf{R} \tag{2.56}$$

where $\mathbf{K} = [K_{AB}]$, $\boldsymbol{\phi} = \{\phi_B\}$ and $\mathbf{R} = \{R_A\}$. $\mathbf{R}$ is called the ***residual*** or ***resultant force vector***. Sometimes $\mathbf{K}$ is called the ***stiffness matrix***, a terminology which derives from the structural mechanics literature.

**Remark 2.9** $\mathbf{K}$ *is tridiagonal for the piecewise linear finite element space (see Figs. 2.8 and 2.9). This is a consequence of the "local support" of the shape functions and the degree-of-freedom numbering adopted (see Hughes [117, Chapter 1], for a detailed explanation).*

**Remark 2.10** $\mathbf{K}$ *is not symmetric, unless* $u = 0$.

**Remark 2.11** *Define* $a(w, \phi) = \displaystyle\int_0^L w_{,x}\kappa\phi_{,x}\,dx$. *This bilinear form is* ***symmetric****, i.e.,* $a(w, \phi) = a(\phi, w)$.

**Remark 2.12** *Define* $b(w, \phi) = -\displaystyle\int_0^L w_{,x}u\phi\,dx$. *An integration by parts reveals that for constant* $u$

$$\begin{array}{rcl} b(w, \phi) & = & \int_0^L wu\phi_{,x}\,dx \;-\; (wu\phi)|_0^L \\ \\ & = & -b(\phi, w) \end{array} \tag{2.57}$$

*Consequently,* $b(\cdot, \cdot)$ *is said to be* ***skew-symmetric***.

Assume $u = 0$ (diffusion dominated case). Then, the Galerkin finite element method with piecewise linear elements is known to be nodally exact for arbitrary $f = f(x)$ and non-uniform mesh (see Hughes [117, Chapter 1]). On the other hand, for the case $u \neq 0$, the solution is not exact. It behaves like the central difference method, for reasons which are apparent upon examining the resulting finite difference equations of the Galerkin finite element method obtained for constant $u$ and $h$:

$$\left\{ \begin{array}{rcl} h\left\{u\dfrac{\phi_{A+1} - \phi_{A-1}}{2h} - \kappa\dfrac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} - \dfrac{1}{h}\int_{x_{A-1}}^{x_{A+1}} N_A f\,dx\right\} & = & 0 \\ \\ A = 1, \dots, \quad N-1 & & \end{array} \right. \tag{2.58}$$

where $\phi_0 = g_0$ and $\phi_N = g_L$. These are the same as central differences except for the treatment of the source term. Specifically, the weighted integral $(1/h)\int_{x_{A-1}}^{x_{A+1}} N_A f\ dx$ appears in (2.58) instead of $f_A = f(x_A)$. This engenders some slight improvement, but it is not really very significant. Since we know the central difference method does not work for the case $\alpha > 1$, the inevitable conclusion is that the Galerkin finite element method is also ineffective in this case.

Integrating (2.44) by parts yields its ***Euler-Lagrange form***:

$$\begin{aligned}&\textstyle\sum_{A=1}^{N}\int_{x_{A-1}}^{x_A} w^h\left(u\phi^h_{,x} - \kappa\phi^h_{,xx} - f\right)\ dx \\ &-\textstyle\sum_{A=1}^{N-1} w^h(x_A)\left(\kappa\phi^h_{,x}(x_A^+) - \kappa\phi^h_{,x}(x_A^-)\right) = 0\end{aligned} \tag{2.59}$$

where $\phi^h_{,x}(x_A^{\pm}) = \lim\limits_{\epsilon\to 0^+}\phi^h_{,x}(x_A \pm \epsilon)$. See Figure 2.10.

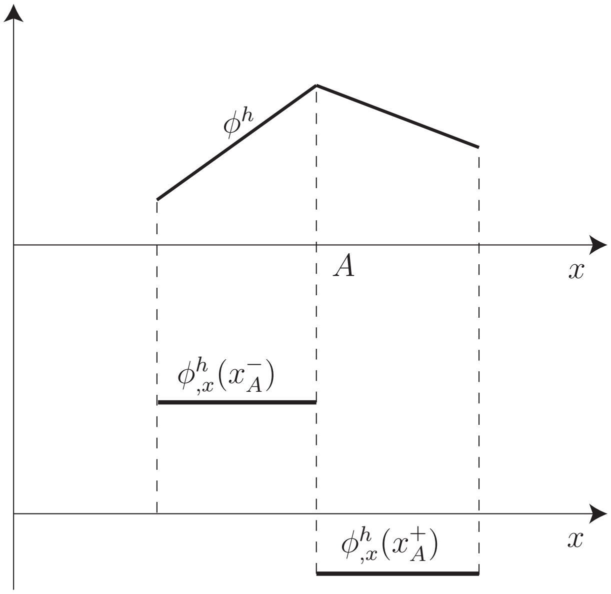


Figure 2.10: Finite element function and derivative near a node $A$.

The integral term corresponds to the (weak) satisfaction of the advection-diffusion equation on element interiors, whereas the sum term corresponds to the (weak) continuity of flux across element interfaces. By virtue of the continuity assumptions on trial solutions, the diffusive flux will not generally be continuous at the nodes.

**Exercise 2.1** *Derive (2.58) from (2.52) – (2.53).*

---

**Exercise 2.2** *Derive (2.59) from (2.44).*

---

### 2.3.3 Galerkin Method for the Modified Equation

The Galerkin method applied to the modified equation (i.e., $\kappa \leftarrow \kappa + \tilde{\kappa}$) results in

$$\begin{aligned} K_{AB} &= \textstyle\int_0^L \left(-N_{A,x} u N_B + N_{A,x}(\kappa + \tilde{\kappa}) N_{B,x}\right)\, dx \\ F_A &= \textstyle\int_0^L N_A f \, dx \end{aligned} \tag{2.60}$$

In the case of a uniform mesh and constant coefficients $u$ and $\kappa$, this leads to

$$\begin{aligned} h \left\{ u \frac{\phi_{A+1} - \phi_{A-1}}{2h} - (\kappa + \tilde{\kappa}) \frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} - \frac{1}{h} \textstyle\int_{x_{A-1}}^{x_{A+1}} N_A f \, dx \right\} &= 0 \\ A = 1, \ldots, \quad & N-1 \end{aligned} \tag{2.61}$$

**Remark 2.13** *If $f = 0$, (2.61) are the difference equations of the exact artificial diffusion method.*

**Remark 2.14** *When $f \neq 0$, the weighted integral treatment of $f$ does not circumvent the difficulties previously noted (see Fig. 2.5).*

### 2.3.4 Petrov-Galerkin Methods: SUPG

#### Historical Background

The terminology "Petrov-Galerkin" is used nowadays to refer to *any* weighted residual method other than the classical Galerkin method. The use of Petrov's name seems to emanate from a book by S. G. Mikhlin [185]. In a short historical introduction, Mikhlin refers to a paper of Petrov's in which a method other than the classical Galerkin method was employed. Based on this single contribution, it seems inappropriate to give Petrov the credit (and likewise the blame!) for *every* method not specifically Galerkin's. Nevertheless for reasons of common usage, we will retain the terminology "Petrov-Galerkin", although we no longer favor it.

#### Petrov-Galerkin Formulation

Consider the following variational method: Find $\phi^h \in \mathcal{S}^h$ such that $\forall\, w^h \in \mathcal{V}^h$,

$$\begin{aligned} & \textstyle\int_0^L \left(-w^h_{,x} u \phi^h + w^h_{,x} \kappa \phi^h_{,x}\right) dx \\ & \boxed{+ \sum_{A=1}^{A_{\max}} \int_{x_{A-1}}^{x_A} p^h \left(u \phi^h_{,x} - \kappa \phi^h_{,xx} - f\right) dx} \\ & = \textstyle\int_0^L w^h f \, dx \end{aligned} \tag{2.62}$$

where $p^h$ is a linear functional of $w^h$.

**Remark 2.15** *If $p^h = 0$, this is the Galerkin method.*

**Remark 2.16** *Integrating (2.62) by parts yields*

$$\begin{aligned}&\textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} (w^h + p^h) \left( u\phi^h_{,x} - \kappa \phi^h_{,xx} - f \right) dx \\ &- \textstyle\sum_{A=1}^{N-1} w^h(x_A) \left( \kappa \phi^h_{,x}(x_A^+) - \kappa \phi^h_{,x}(x_A^-) \right) = 0\end{aligned} \tag{2.63}$$

*In the first term of (2.63), we have changed the weighting of the advection-diffusion equation residual on the element interiors from $w^h$ to $w^h + p^h$. The second term is the same as for Galerkin (see (2.59) ).*

### Specification of $p^h$'s

We want to allow *discontinuous* $p^h$'s, as presented schematically in Figure 2.11. In this case, the term $p^h(u\phi^h_{,x} - \kappa\phi^h_{,xx} - f)$ seems, at first, not to be well defined. If $\phi^h$ is piecewise linear, its first derivative $\phi^h_{,x}$ is a sum of Heaviside functions, and its second derivative $\phi^h_{,xx}$ is a sum of Dirac generalized functions located at the nodes (see Fig. 2.12). A problem ostensibly arises from the fact that the action of a Dirac generalized function on a function which has a discontinuity at the location of the Dirac function is not defined. In fact, this problem is avoided by integrating $p^h(u\phi^h_{,x} - \kappa\phi^h_{,xx} - f)$ *only* over element interiors, where all functions are smooth. This is a key point, the correctness of which will be verified when we perform a mathematical analysis subsequently.

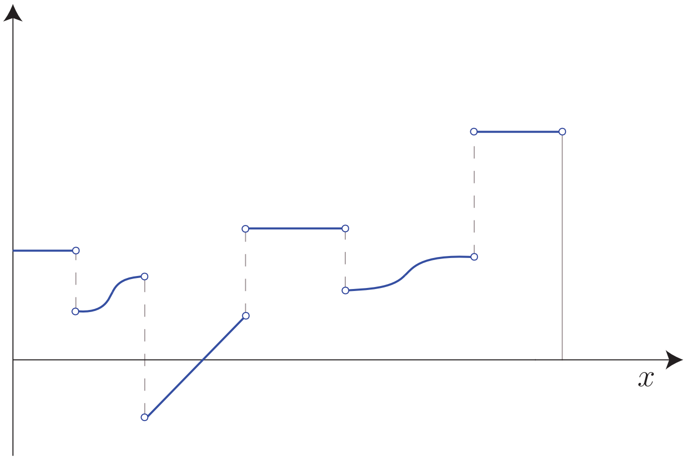
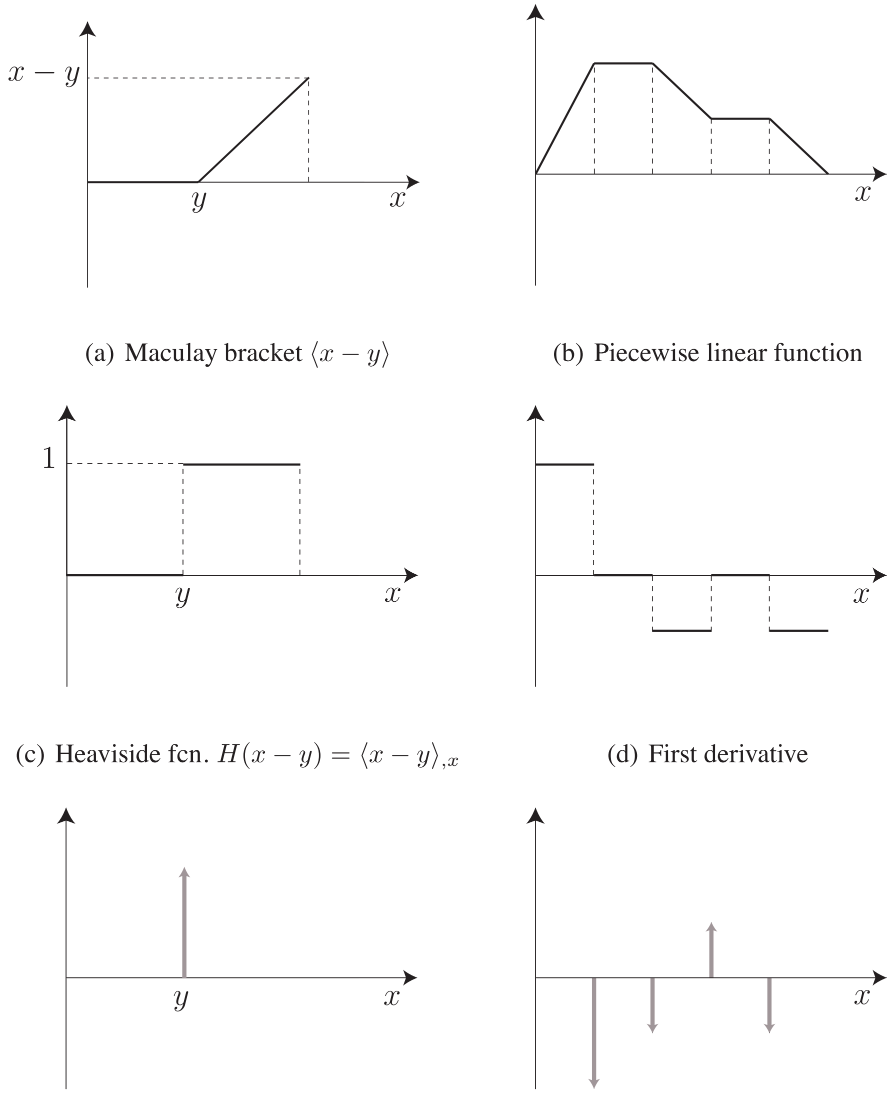


Figure 2.11: Schematic of $p^h$.



(a) Maculay bracket $\langle x - y \rangle$

(b) Piecewise linear function

(c) Heaviside fcn. $H(x - y) = \langle x - y \rangle_{,x}$

(d) First derivative

(e) Dirac delta fcn. $\delta(x - y) = H(x - y)_{,x}$

(f) Second derivative

Figure 2.12: Functions, distributions, and their derivatives. Elementary generalized functions (distributions) are presented on the left. On the right, from top to bottom, a piecewise continuous function and its first and second derivatives.

We want to develop a Petrov-Galerkin method similar to, and actually based on, the "exact" artificial diffusion method. Let us define $p^h$ as follows:

$$p^h \equiv \tau u w^h_{,x} \tag{2.64}$$

where $\tau$ is a parameter. We want to obtain the "exact" artificial diffusion method when $f = 0$ on a uniform mesh. For that purpose, we will show that $\tau$ must be chosen to be:

$$\tau = \frac{\tilde{\kappa}}{|u|^2} \tag{2.65}$$

Let

$$\begin{aligned} \tilde{w}^h &\equiv w^h + p^h \\ &= w^h + \tau u w^h_{,x} \end{aligned} \tag{2.66}$$

The dimensional equation corresponding to (2.66) is

$$\begin{aligned} [\tilde{w}^h] &= [w^h] + [\tau] \cdot [u] \cdot [w^h_{,x}] \\ &= [w^h] + [\tau] \cdot \frac{\text{length}}{\text{time}} \cdot \frac{[w^h]}{\text{length}} \\ &= [w^h] + \frac{[\tau]}{\text{time}} \cdot [w^h] \end{aligned} \tag{2.67}$$

Dimensional consistency requires that $\tau$ must have dimensions of time; it represents an ***element intrinsic time scale***. $\tilde{w}^h$ can be interpreted as a modified weighting function:

$$\tilde{w}^h = \sum_{A=1}^{N-1} (N_A + \tau u N_{A,x}) w_A = \sum_{A=1}^{N-1} \tilde{N}_A w_A \tag{2.68}$$

To understand the difference between $N_A$ and $\tilde{N}_A$, consider the advection-dominated case (i.e. $\alpha \gg 1$), in which $\tilde{\xi}(\alpha) \to 1$, yielding

$$\begin{aligned} \tau &= \frac{\tilde{\kappa}}{|u|^2} \\ &= \tilde{\xi} \frac{|u|h}{2|u|^2} \\ &= \frac{h}{2|u|} \end{aligned} \tag{2.69}$$

Making use of (2.69), we have

$$\tau u N_{A,x} = \frac{h}{2|u|} u N_{A,x} = (\text{sign } u) \frac{h}{2} N_{A,x} \tag{2.70}$$

Figure 2.13 shows the effect of this additional term. Note that $\int_{x_A}^{x_{A+1}} \tilde{N}_A \, dx = 0$ and $\int_{x_{A-1}}^{x_A} \tilde{N}_A \, dx = h$, i.e., all the "weight" of $\tilde{N}_A$ is confined to the element upwind of node $A$, and thus we view $\tilde{N}_A$ as representing an upwind-weighted shape function.

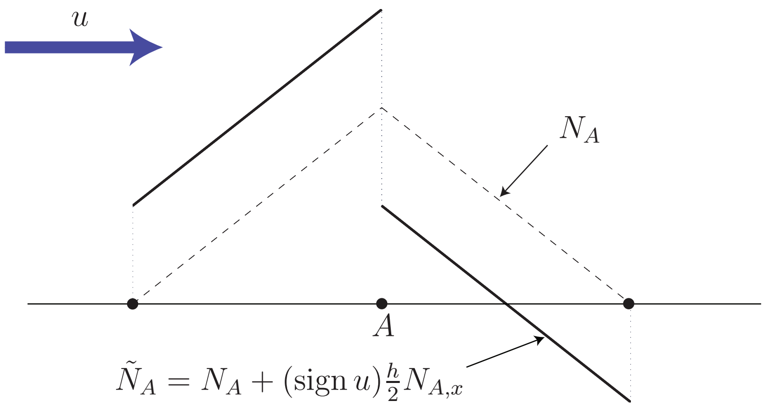

$u$
$N_A$
$A$
$\tilde{N}_A = N_A + (\operatorname{sign} u)\frac{h}{2}N_{A,x}$

Figure 2.13: Galerkin and Petrov-Galerkin weighting functions for node $A$.

**Remark 2.17** *In the case of piecewise linear finite element spaces,*

$$\kappa\phi^h_{,xx} = 0 \tag{2.71}$$

*on element interiors. Thus, the additional term is*

$$\begin{aligned}\int_{x_{A-1}}^{x_A} p^h(u\phi^h_{,x} - \kappa\phi^h_{,xx})\,dx &= \int_{x_{A-1}}^{x_A} \tau u w^h_{,x} u\phi^h_{,x}\,dx \\ &= \int_{x_{A-1}}^{x_A} \frac{\tilde{\kappa}}{|u|^2}|u|^2 w^h_{,x}\phi^h_{,x}\,dx \\ &= \int_{x_{A-1}}^{x_A} w^h_{,x}\tilde{\kappa}\phi^h_{,x}\,dx,\end{aligned} \tag{2.72}$$

*which is identical to the artificial diffusion term of the "exact" artificial diffusion method. On the other hand, if we work with higher-order elements, $\kappa\phi^h_{,xx} \neq 0$ on element interiors. Nevertheless, for appropriately chosen $\tau$, the method will still "work", as we shall verify when we analyze it mathematically. (Higher-order finite element functions are described in Hughes [117, Chapter 3]. An example, $C^0$ piecewise quadratics, is contrasted with piecewise linears in Fig. 2.14.)*

**Remark 2.18** *For piecewise linear finite elements and $f = 0$, the method is clearly identical to the "exact" artificial diffusion method. If $f \neq 0$, the method is different. This effect is* **crucial**. *The source term in (2.62) becomes:*

$$\sum_{A=1}^{N}\int_{x_{A-1}}^{x_A}(w^h + \tau u w^h_{,x})f\,dx = \sum_{A=1}^{N}\int_{x_{A-1}}^{x_A}\tilde{w}^h f\,dx \tag{2.73}$$

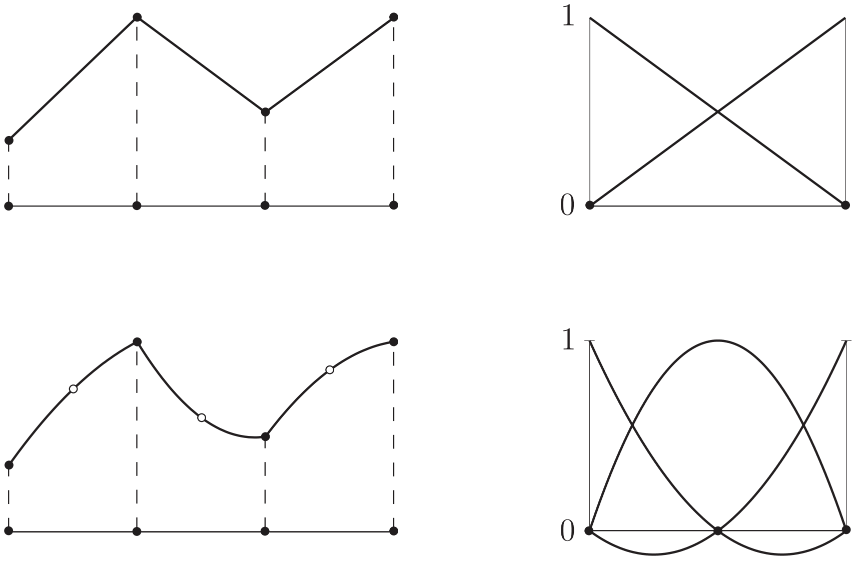


Figure 2.14: $C^0$ piecewise linears (above) and quadratics (below). On the left we have the global representation corresponding to these elements. On the right we have the Lagrange polynomial basis functions of an individual element.

### Matrix form

Assuming constant coefficients and linear elements, the arrays for this method are defined by:

$$\begin{aligned} K_{AB} &= \int_0^L (-N_{A,x} u N_B + N_{A,x}(\kappa + \tilde{\kappa}) N_{B,x})\, dx \\ F_A &= \int_0^L \tilde{N}_A f\, dx \end{aligned} \tag{2.74}$$

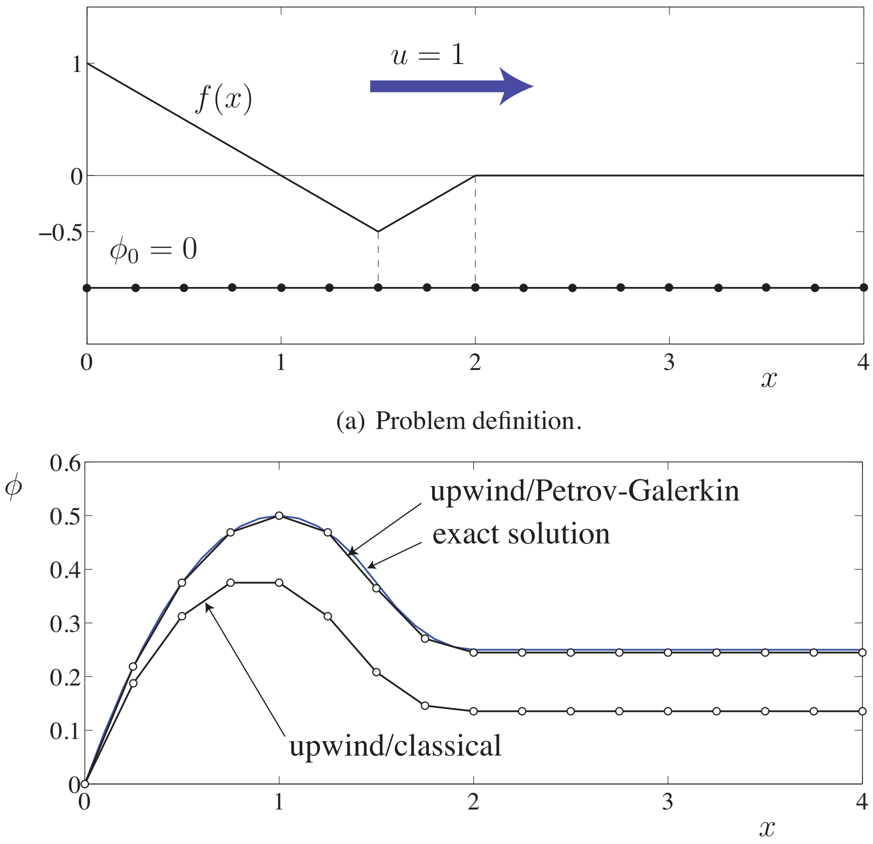
We notice that the treatment of the source term is different from that of the Galerkin method for the modified equation; cf. (2.60). The source is seen to be upwind-weighted in the present situation. In addition, on a uniform mesh, the difference equations are:

$$\begin{aligned} h\left[u\frac{\phi_{A+1} - \phi_{A-1}}{2h} - (\kappa + \tilde{\kappa})\frac{\phi_{A+1} - 2\phi_A + \phi_{A-1}}{h^2} - \frac{1}{h}\int_{x_{A-1}}^{x_{A+1}} \tilde{N}_A f\, dx\right] &= 0 \\ A = 1, \ldots, \quad N-1 & \end{aligned} \tag{2.75}$$

with

$$\phi_0 = g_0 \qquad , \qquad \phi_N = g_L \tag{2.76}$$

The effect of the upwind-weighting of the source term can be seen from the problem described in Figure 2.15(a). This problem amounts to pure advection, meaning $\kappa = 0$. The absence of the second-derivative term means that we can only prescribe one boundary condition. We specify $\phi_0 = 0$. Results are presented in Figure 2.15(b) for the method defined by the modified equation (2.61), in this case classical upwind differences, and the method defined by (2.74), referred to as



(a) Problem definition.

(b) Comparison of results.

Figure 2.15: Pure advection with a non-constant source term.

the Petrov-Galerkin method. *Et voilà*, now the latter method works (SUPG level in Fig. 2.15)! The important observation is that the upwind influence, when introduced via the weighting function, not only produces the stable, non-oscillatory behavior of upwind differences, but also attains excellent accuracy. This illustrates in the simplest setting that stability and accuracy can be simultaneously achieved within one approach.

**Exercise 2.3** *In the case where $f = 0$, show that the difference equations associated with this Petrov-Galerkin method reduce to central differences when $\alpha \to 0$, and to upwind differences when $\alpha \to \infty$.*

---

**Exercise 2.4** *Consider the Galerkin finite element method with piecewise linear interpolation and assume a uniform mesh. Rather than exactly integrating (2.52), devise a one-point quadrature rule for each element subdomain such that the difference equations are identical to (2.75) when $f = 0$. (See Hughes [115] for further discussion of this approach to the development of "upwind finite elements".)*

**Remark 2.19** *The Petrov-Galerkin variational method (2.62) can be generalized to variable coefficients and mesh spacing by defining:*

$$\begin{aligned} \tilde{\kappa}(x) &= \tilde{\xi}(x)u(x)h(x)/2 \\ \tilde{\xi}(x) &= \coth\left(\alpha(x)\right) - \frac{1}{\alpha(x)} \\ \alpha(x) &= \frac{|u(x)|h(x)}{2\kappa(x)} \\ h(x) &= x_A - x_{A-1} \qquad \text{if } x \in \,]x_{A-1}, x_A[ \end{aligned} \tag{2.77}$$

*and by replacing the expression $u\phi^h_{,x} - \kappa\phi^h_{,xx} - f$ in the boxed term in (2.62) by $(u\phi^h)_{,x} - (\kappa\phi^h_{,x})_{,x} - f$.*

### Weighted Residual Methods

Recall the Euler-Lagrange form of the variational equation, that is (2.63):

$$\begin{aligned} &\textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} (w^h + p^h)\left(u\phi^h_{,x} - \kappa\phi^h_{,xx} - f\right) dx \\ &- \textstyle\sum_{A=1}^{N-1} w^h(x_A)\left(\kappa\phi^h_{,x}(x_A^+) - \kappa\phi^h_{,x}(x_A^-)\right) = 0 \end{aligned} \tag{2.78}$$

The terms $[u\phi^h_{,x} - \kappa\phi^h_{,xx} - f]$ and $[\kappa\phi^h_{,x}(x_A^+) - \kappa\phi^h_{,x}(x_A^-)]$ are usually not zero. However, they are zero if $\phi^h$ is replaced by $\phi$, the exact solution of the advection-diffusion equation. For these reasons, these terms are referred to as ***residuals***. Variational methods, such as (2.78), in which all terms have the form of a residual multiplied by a weighting function, are referred to as ***weighted residual methods***. Thus, the Galerkin and Petrov-Galerkin methods are weighted residual methods. The property of being a weighted residual method has very important consequences in convergence analysis. In essence, it guarantees a sort of "consistency" of the approximate variational method with the exact problem. How this works will be seen subsequently.

Now, recall the variational equation of the "exact" artificial diffusion method:

$$\int_0^L \left(-w^h_{,x}u\phi^h + w^h_{,x}(\kappa + \tilde{\kappa})\phi^h_{,x}\right)\, dx - \int_0^L w^h f \, dx = 0 \tag{2.79}$$

Integrating by parts leads to (assuming constant $u$)

$$\begin{aligned} &\textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} w^h \left(u\phi^h_{,x} - \left((\kappa + \tilde{\kappa})\phi^h_{,x}\right)_{,x} - f\right)\, dx \\ - \;&\textstyle\sum_{A=1}^{N-1} w^h(x_A)\left(\left((\kappa + \tilde{\kappa})\phi^h_{,x}\right)(x_A^+) - \left((\kappa + \tilde{\kappa})\phi^h_{,x}\right)(x_A^-)\right) = 0 \end{aligned} \tag{2.80}$$

One can see that the two terms $[u\phi^h_{,x} - \left((\kappa + \tilde{\kappa})\phi^h_{,x}\right)_{,x} - f]$ and $[\left((\kappa + \tilde{\kappa})\phi^h_{,x}\right)(x_A^+) - \left((\kappa + \tilde{\kappa})\phi^h_{,x}\right)(x_A^-)]$ will usually *not* vanish when $\phi^h$ is replaced by $\phi$. Thus, this method is not a residual method. It is not consistent in the important sense alluded to above.

**SUPG Method**

The specific Petrov-Galerkin method studied employed $p^h = \tau u w^h_{,x}$ where $\tau = \tilde{\kappa}/|u|^2$. This choice of $p^h$ is referred to as an ***SUPG method***. SUPG stands for Streamline Upwind Petrov-Galerkin. The motivation for the word "streamline" can only be appreciated when we generalize to the multidimensional case. The use of the word "upwind" should be understandable by virtue of the upwind bias of the weighting function.

The SUPG term added to the Galerkin method can be written as

$$\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau u w^h_{,x} \left( \mathcal{L}\phi^h - f \right) \, dx \tag{2.81}$$

where $\mathcal{L}\phi^h = u\phi^h_{,x} - \kappa\phi^h_{,xx}$. The linear operator $\mathcal{L}$ can be split into two parts, $\mathcal{L} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{diff}}$ where $\mathcal{L}_{\text{adv}} = u\dfrac{\partial}{\partial x}$ and $\mathcal{L}_{\text{diff}} = -\kappa\dfrac{\partial^2}{\partial x^2}$. Thus, the SUPG term can also be written as

$$\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau \left( \mathcal{L}_{\text{adv}} w^h \right) \left( \mathcal{L}\phi^h - f \right) \, dx \tag{2.82}$$

Unfortunately, a given $\mathcal{L}$ may not come canonically decomposable into $\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{diff}}$. The following more general idea is one way to circumvent this shortcoming.

### 2.3.5 Galerkin/Least-Squares (GLS) Method

The ***Galerkin/least-squares method*** is defined by adding to the Galerkin formulation the term

$$\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau \mathcal{L} w^h \left( \mathcal{L}\phi^h - f \right) \, dx \tag{2.83}$$

In this case, the entire operator $\mathcal{L}$ acts on the weighting function. We see that within our general Petrov-Galerkin framework, Galerkin/least-squares corresponds to defining $p^h = \tau\mathcal{L}w^h$. To justify the *least-squares* terminology, consider the quadratic potential

$$\mathcal{P}(\phi^h) = \sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \frac{\tau}{2} \left( \mathcal{L}\phi^h - f \right)^2 dx \tag{2.84}$$

The variational derivative of $\mathcal{P}$ is

$$\begin{aligned} \delta\mathcal{P} = D\mathcal{P}(\phi^h) \cdot w^h \; &\equiv \frac{d}{d\epsilon}\mathcal{P}(\phi^h + \epsilon w^h)\Big|_{\epsilon=0} \\ &= \textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau \left( \mathcal{L}\phi^h - f \right) \mathcal{L} w^h \, dx \end{aligned} \tag{2.85}$$

Thus, Galerkin/least-squares is the sum of the Galerkin variational equation and the derivative of a least-squares potential.

Table 2.1: Properties of methods described in the present chapter.

| | Central Differences, Galerkin FEM | Upwind Differences | SUPG, GLS |
|---|---|---|---|
| Accuracy | Yes | No | Yes |
| Stability | No | Yes | Yes |

**Remark 2.20** *The Galerkin/least-squares method represents a general conception (see Franca and Hughes [80] and Hughes, Franca and Hulbert [126]. Its applicability is greater than SUPG, due to the restriction mentioned after (2.82).*

**Remark 2.21** *For piecewise linear elements, we have $\mathcal{L}w^h = uw^h_{,x}$ on element interiors. In this case, Galerkin/least-squares is identical to SUPG.*

**Remark 2.22** *Artificial diffusivity in the Galerkin/least-squares method emanates from lack of vanishing of the residual.*

**Remark 2.23** *The definition of $\tau$ derived for piecewise linear elements is applicable to both SUPG and Galerkin/least-squares. For higher-order elements, the definition is basically the same. We shall make further remarks on this point later on.*

**Remark 2.24** *The situation for methods studied in this chapter is summarized in Table 2.1.*

### 2.3.6 Abstract Notation

The Galerkin, SUPG and Galerkin/least-squares methods can be stated in compact, abstract notation as follows:

- ***Galerkin method***: Find $\phi^h \in \mathcal{S}^h$ such that

$$B(w^h, \phi^h) = L(w^h) \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.86}$$

where

$$\begin{aligned} B(w, \phi) &\equiv \textstyle\int_0^L \left(-w_{,x} u\phi + w_{,x}\kappa\phi_{,x}\right)\, dx \\ L(w) &\equiv \textstyle\int_0^L wf\, dx \end{aligned} \tag{2.87}$$

- ***SUPG method***: Find $\phi^h \in \mathcal{S}^h$ such that

$$B_{\text{SUPG}}(w^h, \phi^h) = L_{\text{SUPG}}(w^h) \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.88}$$

where

$$\begin{aligned} B_{\text{SUPG}}(w, \phi) &= B(w, \phi) + \textstyle\sum_{A=1}^N \int_{x_{A-1}}^{x_A} \tau\left(\mathcal{L}_{\text{adv}} w\right)\left(\mathcal{L}\phi\right)\, dx \\ L_{\text{SUPG}}(w) &= L(w) + \textstyle\int_0^L \tau\left(\mathcal{L}_{\text{adv}} w\right) f\, dx \end{aligned} \tag{2.89}$$

- ***Galerkin/least-squares method***: Find $\phi^h \in \mathcal{S}^h$ such that

  $$B_{\mathrm{GLS}}(w^h, \phi^h) = L_{\mathrm{GLS}}(w^h) \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.90}$$

  where

  $$\begin{aligned} B_{\mathrm{GLS}}(w, \phi) &= B(w, \phi) + \textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau\, (\mathcal{L}w)\, (\mathcal{L}\phi)\ dx \\ L_{\mathrm{GLS}}(w) &= L(w) + \textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau\, (\mathcal{L}w)\, f\ dx \end{aligned} \tag{2.91}$$

All the $B_{\ldots}(\cdot,\cdot)$'s are bilinear forms, i.e., for any functions $w$, $w_1$ and $w_2$ and arbitrary constants $c_1$ and $c_2$, we have

$$B_{\ldots}(c_1 w_1 + c_2 w_2, w) = c_1 B_{\ldots}(w_1, w) + c_2 B_{\ldots}(w_2, w) \tag{2.92}$$

and

$$B_{\ldots}(w, c_1 w_1 + c_2 w_2) = c_1 B_{\ldots}(w, w_1) + c_2 B_{\ldots}(w, w_2) \tag{2.93}$$

Moreover, all the $L_{\ldots}(\cdot)$'s are linear forms, i.e.,

$$L_{\ldots}(c_1 w_1 + c_2 w_2) = c_1 L_{\ldots}(w_1) + c_2 L_{\ldots}(w_2) \tag{2.94}$$

These properties facilitate a rapid derivation of the matrix equations. Let $K_{AB} = B_{\ldots}(N_A, N_B)$ and $F_A = L_{\ldots}(N_A)$. Then

$$\begin{aligned} 0 &= B_{\ldots}(w^h, \phi^h) - L_{\ldots}(w^h) \\ &= B_{\ldots}\left(\textstyle\sum_A w_A N_A, \sum_B \phi_B N_B\right) - L_{\ldots}\left(\sum_A w_A N_A\right) \\ &= \textstyle\sum_A w_A \left(\sum_B B_{\ldots}(N_A, N_B)\phi_B\right) - \sum_A w_A L_{\ldots}(N_A) \\ &= \textstyle\sum_A w_A \left(\sum_B K_{AB}\phi_B - F_A\right) \end{aligned} \tag{2.95}$$

which implies $\sum_{B=0}^{A_N} K_{AB}\phi_B = F_A$, $A = 1, \ldots, N-1$, due to the arbitrariness of the $w_A$'s. Taking account of Dirichlet boundary conditions, as before, all these methods yield algebraic equations of the form

$$\mathbf{K}\boldsymbol{\phi} = \mathbf{R} \tag{2.96}$$

## 2.4 Introduction to Functional Analysis

At this point it is necessary to review some elementary concepts and results from functional analysis. These play an important role in the exposition to follow. The reader who is expert in this subject may wish to skim over this material.

### 2.4.1 Linear Spaces

In our discussion of linear spaces, a good example to keep in mind is $\mathcal{V}$, the space of weighting functions. After some basic results are presented, we will show that $\mathcal{V}$ *really* is a linear space in the sense that it satisfies the technical conditions to follow. Throughout, we will use $v$'s and $w$'s to denote members of the linear space under consideration (think of these as functions), and $c$'s to denote constants.

A linear space is a collection which is closed under the operations of addition and scalar multiplication. This means that if $\mathcal{V}$ is a linear space and if $w_1, w_2 \in \mathcal{V}$ and $c_1, c_2$ are constants, then

$$c_1 w_1 + c_2 w_2 \in \mathcal{V}. \tag{2.97}$$

### 2.4.2 Inner Products

An inner product defined on a linear space $\mathcal{V}$ is a function $(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ with the following properties:

$$\begin{array}{lrcll}
\text{i.} & (v, w) & = & (w, v) & \text{(symmetry)} \\
\text{ii.} & (c_1 v_1 + c_2 v_2, w) & = & c_1 (v_1, w) + c_2 (v_2, w) & \text{(bilinearity)} \\
 & (v, c_1 w_1 + c_2 w_2) & = & c_1 (v, w_1) + c_2 (v, w_2) & \\
\text{iii.} & (v, v) & \geq & 0, \text{ and } (v, v) = 0 \text{ if and only if } v = 0 & \text{(positive-definiteness)}
\end{array} \tag{2.98}$$

An example of an inner product is given by

$$(v, w) = \int_0^L vw \, dx \tag{2.99}$$

Our $B_{\ldots}(\cdot, \cdot)$'s are not symmetric; consequently, they are *not* inner products.

### 2.4.3 Norms

A norm defined on a linear space $\mathcal{V}$ is a function $|| \cdot || : \mathcal{V} \to \mathbb{R}$ with the following properties:

$$\begin{array}{lrcll}
\text{i.} & ||v|| & \geq & 0, \text{ and } ||v|| = 0 \text{ if and only if } v = 0 & \text{(positive-definiteness)} \\
\text{ii.} & ||cv|| & = & |c| \, ||v||, \text{ where } |c| \text{ is the absolute value of } c & \\
\text{iii.} & ||v + w|| & \leq & ||v|| + ||w|| & \text{(triangle inequality)}
\end{array} \tag{2.100}$$

The triangle inequality can be visualized if the norm is interpreted as the length of a vector (see Fig. 2.16). An example of a norm is given by

$$||v|| = \left( \int_0^L v^2 \, dx \right)^{1/2} . \tag{2.101}$$

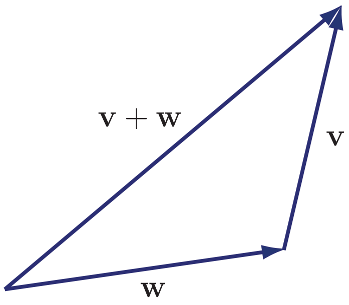


Figure 2.16: Triangle inequality.

### 2.4.4 Seminorms

A seminorm $|\cdot|$ defined on a linear space $\mathcal{V}$ is a function $|\cdot| : \mathcal{V} \to \mathbb{R}$ with the following properties:

$$\begin{array}{rrcll} \text{i.} & |v| & \geq & 0 & (\mathrm{positive-semidefiniteness}) \\ & & & & \text{(Property i. of a norm is weakened,} \\ & & & & \text{ i.e., } |v| \text{ could be zero for some } v \neq 0.) \\ \text{ii.} & |cv| & = & |c|\,|v| & \text{where again } |c| \text{ is the absolute value of } c, \\ & & & & \text{however, } |v| \text{ denotes the seminorm of } v. \\ \text{iii.} & |v+w| & \leq & |v| + |w| & \text{(triangle inequality)} \end{array} \tag{2.102}$$

### 2.4.5 Inner Product and Normed Spaces

An inner product (resp., a normed) space is a linear space with an inner product (resp., a norm) defined on it. An inner product space possesses a natural norm defined by:

$$||v|| = (v, v)^{1/2} \tag{2.103}$$

**Cauchy-Schwarz Inequality:**

$$\begin{aligned} |(v, w)|^2 &\leq (v, v)(w, w) \\ &= ||v||^2 \, ||w||^2 \end{aligned} \tag{2.104}$$

Thus,

$$\boxed{|(v,w)| \leq ||v||\,||w||} \tag{2.105}$$

A proof of the Cauchy-Schwarz inequality can be found in Hughes [117, p.265].

**Exercise 2.5** *Verify that the natural norm really satisfies the definition of a norm.*

**Solution**:

We must verify properties i–iii of a norm:

1. The natural norm is positive-definite by definition.

2. $$||cv|| \equiv (cv,cv)^{1/2} = \left(c^2(v,v)\right)^{1/2} = |c|\,||v|| \tag{2.106}$$

3. $$\begin{aligned} ||v+w||^2 &= (v+w, v+w) \\ &= (v,v) + (v,w) + (w,v) + (w,w) && \text{(by bilinearity)} \\ &= ||v||^2 + ||w||^2 + 2(v,w) && \text{(by definition of the natural norm)} \\ &\leq ||v||^2 + ||w||^2 + 2||v||\,||w|| && \text{(by Cauchy-Schwarz)} \\ &= (||v|| + ||w||)^2 \end{aligned} \tag{2.107}$$

Therefore, taking square roots,

$$||v+w|| \leq ||v|| + ||w|| \tag{2.108}$$

## 2.4.6 Hilbert and Sobolev Spaces

Hilbert spaces are inner product spaces which are complete in their natural norm. (Roughly speaking, this means that every convergent sequence in the space converges to a member of the space). Hilbert spaces are particular Sobolev spaces. Some examples follow.

$L_2 = L_2(]0, L[)$

Perhaps, the simplest Sobolev space is the space of square-integrable functions: $L_2$. The $L_2$-inner product is defined as

$$(v, w) = \int_0^L vw \, dx \tag{2.109}$$

The corresponding natural norm is

$$||v|| = (v, v)^{1/2} = \Big( \int_0^L v^2 \, dx \Big)^{1/2} < \infty \tag{2.110}$$

It is well-defined, since all functions in $L_2$ are by definition square-integrable.

$H^1 = H^1(]0, L[)$

$H^1$ is the space of $L_2$-functions with $L_2$-derivatives; that is, functions that are square-integrable and that possess square-integrable generalized derivatives. The $H^1$-inner product is defined by,

$$(v, w)_1 = \int_0^L (vw + L^2 v_{,x} w_{,x}) \, dx \tag{2.111}$$

The $H^1$-norm is

$$||v||_1 = (v, v)_1^{1/2} \tag{2.112}$$

The $H^1$-seminorm is

$$|v|_1 = \Big( \int_0^L L^2 (v_{,x})^2 \, dx \Big)^{1/2} \tag{2.113}$$

Clearly, $| \cdot |_1$ is positive-semidefinite. However, it is *not* positive-definite because, for $v =$ const $\neq 0$, $|v|_1 \equiv 0$.

**Exercise 2.6** *Show that*

$$||v||_1^2 = ||v||^2 + L^2 ||v_{,x}||^2 \tag{2.114}$$

---

$H^s = H^s(]0, L[)$

$H^s$ is the space of square-integrable functions having square-integrable derivatives through order $s$. The $H^s$-inner product is

$$(v, w)_s = \int_0^L (vw + L^2 v_{,x} w_{,x} + \cdots + L^{2s} v_{,\underbrace{x \cdots x}_{s \text{ times}}} w_{,\underbrace{x \cdots x}_{s \text{ times}}}) \, dx \tag{2.115}$$

The $H^s$-norm is

$$||v||_s = (v,v)_s{}^{1/2} \tag{2.116}$$

The $H^s$-seminorm is

$$|v|_s = \left( \int_0^L L^{2s} (v_{,\underbrace{x \cdots x}_{s \text{ times}}})^2 \, dx \right)^{1/2} \tag{2.117}$$

**Remark 2.25** $H^0 = L_2$.

**Remark 2.26** $H^s \subset H^r$ for $r \leq s$ *(see Fig. 2.17).*

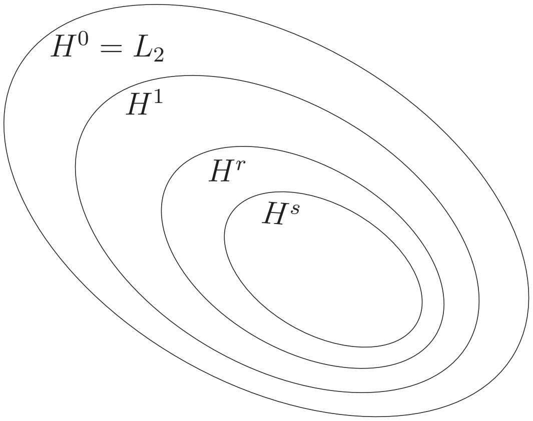


Figure 2.17: Imbedding of Sobolev spaces.

### 2.4.7 Finite Element Spaces

Our main examples of linear spaces are $\mathcal{V}$ and $\mathcal{V}^h$. The finite element space $\mathcal{V}^h$ is a subset of $\mathcal{V}$. We require, in turn, that $\mathcal{V}$ is a subset of $H^1$, viz.,

$$\mathcal{V}^h \subset \mathcal{V} \equiv \{ w \mid w \in H^1(]0,L[), w(0) = w(L) = 0 \} \tag{2.118}$$

That $\mathcal{V}$ is a subset of $H^1$ is the "technical condition" we alluded to at the beginning of Section 2.3. Our objective here is to rigorously establish that $\mathcal{V}$ and $\mathcal{V}^h$ are linear spaces. This requires a few preliminary results which are interesting in their own right.

A particular case of the ***Sobolev imbedding theorem*** states that if $w \in H^1(]0,L[)$, then $w \in C^0([0,L])$, the space of continuous functions. That is, any $H^1(]0,L[)$-function is continuous on $[0,L]$. In particular, if $w \in \mathcal{V}$, then the boundary values , $w(0)$ and $w(L)$, are well-defined by virtue of the fact that $w$ is a continuous function. (*Warning:* This result is not true for $H^1$-functions defined on a multidimensional domain. For specific information on this point, see Hughes [117, p.268]. A comprehensive study of Sobolev spaces is presented in Adams [1].)

**Proposition 2.1** *The $H^1$-seminorm is a norm on $\mathcal{V}$.*

*Proof*

We will show that $|\cdot|_1$ is positive-definite on $\mathcal{V}$. Suppose that $|v|_1 = 0$ for some $v \in \mathcal{V}$. From (2.113), we conclude that $v$ must be a *constant*. Since $v \in \mathcal{V}$, $v(0) = v(L) = 0$. Therefore, $v \equiv 0$. □

**Proposition 2.2** $||\cdot||_1$ *and* $|\cdot|_1$ *are* ***equivalent*** *norms on $\mathcal{V}$, i.e., there exist positive constants $c_1$, $c_2$, such that $\forall\ v \in \mathcal{V}$*

$$c_1||v||_1 \leq |v|_1 \leq c_2||v||_1 \tag{2.119}$$

(If two norms are equivalent, then a sequence which converges to zero in one of the norms, converges to zero in the other.)

In order to establish the left-hand inequality, we first need to prove the following lemma.

**Lemma 2.1 (Poincaré-Friedrichs inequality)** *There exists a positive constant c, such that $\forall\ v \in \mathcal{V}$,*

$$||v|| \leq c\ ||v_{,x}|| \tag{2.120}$$

*Proof of the Lemma*

For all $x$ in $[0, L]$:

$$\begin{aligned}
\Big(v(x) - v(0)\Big)^2 &= \left(\int_0^x v_{,x}\ dx\right)^2 && \text{(Fundamental Theorem of Calculus)}\\
&= \left(\int_0^x 1 \cdot v_{,x}\ dx\right)^2 \\
&\leq \left(\int_0^x 1^2\ dx\right)\left(\int_0^x v_{,x}^2\ dx\right) && \text{(Cauchy-Schwarz inequality)}\\
&\leq \left(\int_0^L 1^2\ dx\right)\left(\int_0^L v_{,x}^2\ dx\right)\\
&= L||v_{,x}||^2
\end{aligned} \tag{2.121}$$

Since $v \in \mathcal{V}$, $v(0) = 0$, and integrating over $]0, L[$:

$$||v||^2 \leq L^2||v_{,x}||^2 \tag{2.122}$$

Therefore,

$$\boxed{||v|| \leq L||v_{,x}||} \tag{2.123}$$

The lemma is established with $c = L$. □

Now, we return to the proof of the Proposition.

*Proof of Proposition*

We proceed as follows to prove the left-hand inequality in (2.119):

$$\begin{aligned}
|v|_1^2 \ &\equiv L^2 \, \|v_{,x}\|^2 \\
&= \frac{L^2}{2}\|v_{,x}\|^2 + \frac{L^2}{2}\|v_{,x}\|^2 \\
&\geq \frac{1}{2}\|v\|^2 + \frac{L^2}{2}\|v_{,x}\|^2 \qquad \text{(Poincaré-Friedrichs)} \\
&\geq \frac{1}{2}\left(\|v\|^2 + L^2\|v_{,x}\|^2\right) \\
&= \frac{1}{2}\|v\|_1^2
\end{aligned} \tag{2.124}$$

So, $c_1 = \dfrac{1}{\sqrt{2}}$. The right-hand inequality is trivial to establish, viz.,

$$\begin{aligned}
|v|_1^2 \ &\equiv L^2\|v_{,x}\|^2 \\
&\leq \|v\|^2 + L^2\|v_{,x}\|^2 \\
&= ||v||_1^2
\end{aligned} \tag{2.125}$$

Thus, $c_2 = 1$. This completes the proof of equivalence. □

Now, we return to the issue of rigorously establishing that $\mathcal{V}$ is a linear space. We need to show that if $v_1, v_2 \in \mathcal{V}$, and $c_1, c_2$ are constants, then $v \equiv c_1v_1 + c_2v_2 \in \mathcal{V}$. By definition of $\mathcal{V}$, i.e., (2.118), we require that $v \in H^1$ and $v(0) = v(L) = 0$. Satisfaction of the boundary conditions follows from

$$\begin{aligned}
v(0) \ &= c_1v_1(0) + c_2v_2(0) = 0 \\
v(L) \ &= c_1v_1(L) + c_2v_2(L) = 0
\end{aligned} \tag{2.126}$$

To show that $v \in H^1$, that is $||v||_1 < \infty$, we employ the triangle inequality and the properties of a norm:

$$\begin{aligned}
\|v\|_1 \ &= \|c_1v_1 + c_2v_2\|_1 \\
&\leq \|c_1v_1\|_1 + \|c_2v_2\|_1 \\
&= |c_1|\|v_1\|_1 + |c_2|\|v_2\|_1 \\
&< \infty
\end{aligned} \tag{2.127}$$

Assume that $\mathcal{V}^h$ consists of continuous piecewise polynomials of order $k$ on each element interval, e.g., $k = 1$ corresponds to the linear case studied previously, $k = 2$ to quadratics, etc.

The proof that $\mathcal{V}^h$ is a linear space is similar to that for $\mathcal{V}$, but we have to verify the additional property that $v^h \equiv c_1v_1^h + c_2v_2^h$ is a polynomial of degree less than or equal to $k$ on each element. This is obvious by virtue of the fact that $v_1^h$ and $v_2^h$ are polynomials of degree less than or equal to $k$. $\mathcal{V}^h$ is said to be a ***linear subspace*** of $\mathcal{V}$.

We have avoided mentioning $\mathcal{S}$ and $\mathcal{S}^h$ in our discussion of linear spaces. The reason for this is that $\mathcal{S}$ and $\mathcal{S}^h$ fail to be linear spaces because of the *inhomogeneous* boundary conditions built into their definitions. For example, if $\phi_1, \phi_2 \in \mathcal{S}$, then $\phi_1(0) = \phi_2(0) = g_0$, but then $\phi \equiv \phi_1 + \phi_2 \notin \mathcal{S}$ because $\phi(0) = \phi_1(0) + \phi_2(0) = 2g_0$.

## 2.5 Analysis of the Finite Element Methods

### 2.5.1 Mathematical Properties

#### Consistency

Each of our methods can be written in the form $B_{\dots}(w^h, \phi^h) = L_{\dots}(w^h)$. The property of being a weighted residual method entails $B_{\dots}(w^h, \phi) = L_{\dots}(w^h)$. Subtracting yields $B_{\dots}(w^h, \phi^h - \phi) = 0$, by bilinearity. Thus, the error $e = \phi^h - \phi$ is ***orthogonal*** to all $w^h \in \mathcal{V}^h$ with respect to $B_{\dots}(\cdot, \cdot)$. This may be thought of as an accuracy condition for the method. Sometimes it is referred to as ***strong consistency***.

#### Stability

The relevant stability condition is $B_{\dots}(w^h, w^h) \geq c\,|||w^h|||^2 \;\; \forall\, w^h \in \mathcal{V}^h$, where $c$ is a positive constant and $|||\cdot|||$ is some norm defined on $\mathcal{V}$. $B_{\dots}(\cdot, \cdot)$ is said to be ***coercive***. As an example, we can calculate the bound for the Galerkin method $\forall\, w \in \mathcal{V}$:

$$\begin{aligned} B(w, w) &\equiv \textstyle\int_0^L \left(-w_{,x} u w + \kappa \left(w_{,x}\right)^2\right)\, dx \\ &= \textstyle\int_0^L \left(-\frac{u}{2}\left((w)^2\right)_{,x} + \kappa \left(w_{,x}\right)^2\right)\, dx \\ &= -\frac{u}{2}\left((w)^2\right)\big|_0^L + \kappa \|w_{,x}\|^2 \\ &= \kappa \|w_{,x}\|^2 \end{aligned} \tag{2.128}$$

In general, we only need to establish stability for all $w^h \in \mathcal{V}^h$. However, for the Galerkin method it holds for all $w \in \mathcal{V}$, hence all $w^h \in \mathcal{V}^h$. The Galerkin method is then stable for all $\kappa > 0$, but $\kappa$ may be very small in a nondimensional sense ($\alpha \to \infty$). This turns out to be a manifestation of a serious deficiency of the Galerkin method.

To prove a method converges, we need to first establish properties of consistency and stability. In addition, two other properties of the usual finite element functions are important in the analyses of the methods under consideration:

- ***Interpolation estimates***.

- ***Inverse estimates***.

**Remark 2.27** *Inverse estimates are important in the analysis of the Petrov-Galerkin methods, but are not needed for the analysis of the Galerkin method.*

**Remark 2.28** *Interpolation estimates describe the accuracy of the finite element functions with respect to the functions that are being approximated. Accuracy and consistency are intimately linked concepts. Inverse estimates describe the stability of the functions. The dual notions of accuracy/consistency and stability appear for both the variational method and the finite element functions, as connoted in Table 2.2.*

Table 2.2: The notions of accuracy/consistency and stability.

| | Method | Finite element functions |
|---|---|---|
| Consistency/accuracy | Error orthogonality | Interpolation estimates |
| Stability | Coercivity | Inverse estimates |

**Remark 2.29** *There are times when either or both error orthogonality and coercivity fail, but convergence may still occur. These cases are described by somewhat weakened notions of consistency and stability of the method. For example, an artificial diffusion term may be added to a method, and if its coefficient is sufficiently small, convergence can still be proved, despite the error orthogonality condition being lost. This is sometimes referred to as* ***weak consistency****. Likewise, (see, e.g., Brenner and Scott [27], Brezzi and Fortin [33]) coercivity may fail but the inf-sup condition may be satisfied. There are methods that are coercive for a restricted range of the stabilization parameter, but are inf-sup stable for a larger range (see Bochev, Dohrman and Gunzburger [20] and Bochev* et al. *[21]). For almost all of this book it will be sufficient to only consider error orthogonality and coercivity properties.*

In the next section we will present some basic results on interpolation estimates.

### 2.5.2 Interpolation Estimates

The interpolation estimates that we shall be concerned with describe the approximation properties of finite element functions in Sobolev norms.

Let $h_A = x_A - x_{A-1}$ be the length of element $A$ and let $h = \max_A h_A$ be the maximum element length. (For a quadratic element the definition of the element length needs to be modified. See Figure 2.18. Likewise, for other higher-order elements.) Assume the exact solution $\phi \in H^r(]0, L[)$. Recall that this means $\phi$ possesses $r$ square-integrable derivatives ($r$ is said to be the ***degree of regularity*** of $\phi$).

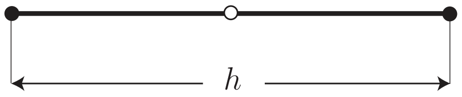


Figure 2.18: Length of a three-node quadratic element.

If $f = 0$, we know that $\phi$ is an exponential function. Thus, $\phi$ is infinitely differentiable, i.e., $\phi \in C^\infty \subset H^\infty$. But, if $f \in H^{r-2}$, then from the advection-diffusion equation, $\phi_{,xx} \in H^{r-2}$ and thus, $\phi \in H^r$. We need at least that $\phi \in H^1$ for our variational formulations to be well-defined, i.e., we require $r \geq 1$. If $r = 1$, $f \in H^{-1}$, which means that $f$ may contain Dirac delta functions. Even in this case, $\phi$ is no worse than continuous by the Sobolev imbedding theorem.

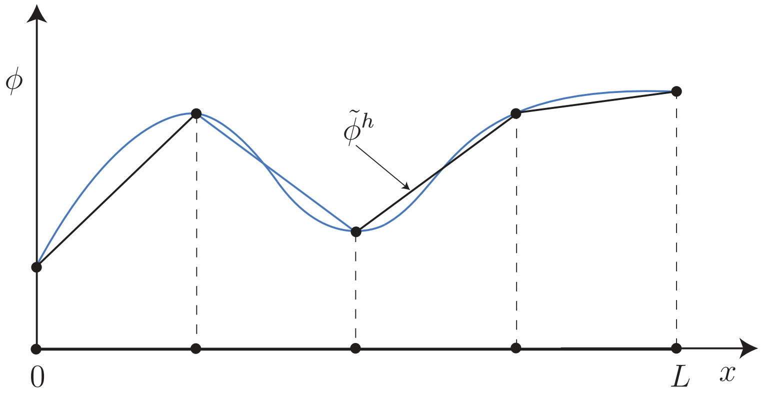


Figure 2.19: Linear interpolation of a function.

We want to know how well we can approximate $\phi \in H^r \cap \mathcal{S}$ by a function in $\mathcal{S}^h$. To answer this question, we examine the interpolant of $\phi$, namely $\tilde{\phi}^h \in \mathcal{S}^h$. (See Fig. 2.19 for a sketch of the piecewise linear interpolate of a function.) The ***interpolation error*** $\eta \equiv \tilde{\phi}^h - \phi$. The following estimate quantifies the interpolation error in the $s^{\text{th}}$ Sobolev norm:

$$\|\eta\|_s \leq c \left(\frac{h}{L}\right)^\alpha \|\phi\|_r \tag{2.129}$$

where

- $c$ is a non-dimensional constant *independent* of $h$ and $\phi$, but it can depend on $L$, $s$, $r$ and $\mathcal{S}^h$.

- $\alpha = \min\{k+1-s, r-s\}$, $k$ being the degree of the element polynomials. $\alpha$ is called the ***rate of convergence***. We can see that the higher the derivative being approximated, the poorer the estimate. However, the higher the order of the polynomials, the better the estimate, up to the point when $k+1 > r$.

(2.129) can be rewritten as

$$\boxed{\|\eta\|_s \leq c\left(\frac{h}{L}\right)^{\alpha} \|\phi\|_r} \tag{2.130}$$

We see that for $\alpha > 0$, $\|\eta\|_s \to 0$ as $h \to 0$ (mesh refinement). When $r$ is sufficiently large, i.e., $r > k+1$, we have

$$\boxed{\|\eta\|_s \leq c\left(\frac{h}{L}\right)^{k+1-s} \|\phi\|_{k+1}} \tag{2.131}$$

Indeed, if $r > k+1$, then $\phi \in H^r \subset H^{k+1}$. So, (2.129) can be applied to $\phi \in H^{k+1}$ which results in (2.130). This is often referred to as the ***standard interpolation estimate***.

***Examples***:

1. For linear elements ($k = 1$):
$$\begin{aligned} \|\eta\|_0 &\sim h^2 \\ \|\eta\|_1 &\sim h \end{aligned} \tag{2.132}$$

2. For quadratic elements ($k = 2$):
$$\begin{aligned} \|\eta\|_0 &\sim h^3 \\ \|\eta\|_1 &\sim h^2 \end{aligned} \tag{2.133}$$
In these cases, $w^h \in C^0$ (see Fig. 2.20), $w^h_{,x} \in C^{-1}$ is a discontinuous function, and $w^h_{,xx} \in C^{-2}$ includes Dirac delta functions. Consequently, the $H^2$-norm of $w^h$ is not well-defined and thus neither is the $H^2$-norm of $\eta$. Thus we cannot estimate the $H^s$-norm of $\eta$ for $s \geq 2$ in this case.

3. For cubic elements ($k = 3$):
$$\begin{aligned} \|\eta\|_0 &\sim h^4 \\ \|\eta\|_1 &\sim h^3 \end{aligned} \tag{2.134}$$
Piecewise cubics can be defined in various ways. The two most common are illustrated in Figure 2.21. Note that Hermite cubics are smoother than the ordinary Lagrange cubics, i.e., Hermite cubics have continuous first derivatives whereas Lagrange cubics are merely continuous. The generalized second derivative of a Hermite cubic is a piecewise discontinuous function. Consequently, Hermite cubics possess a well-defined $H^2$-norm. However, Lagrange cubics, which possess slope discontinuities at the end nodes, give rise to Dirac delta functions in their second derivatives and thus are not members of $H^2$. The upshot is that the interpolation estimate in the $H^2$-norm, viz.,
$$\|\eta\|_2 \sim h^2 \tag{2.135}$$
is valid for Hermite cubics but not Lagrange cubics.

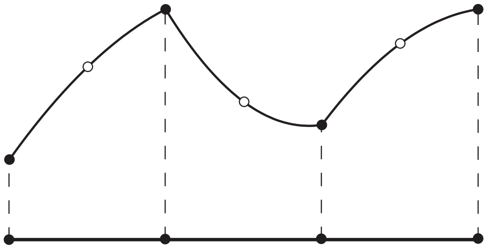


Figure 2.20: Illustration of a piecewise quadratic finite element interpolation function. Notice the slope discontinuities at element boundaries (solid nodes).

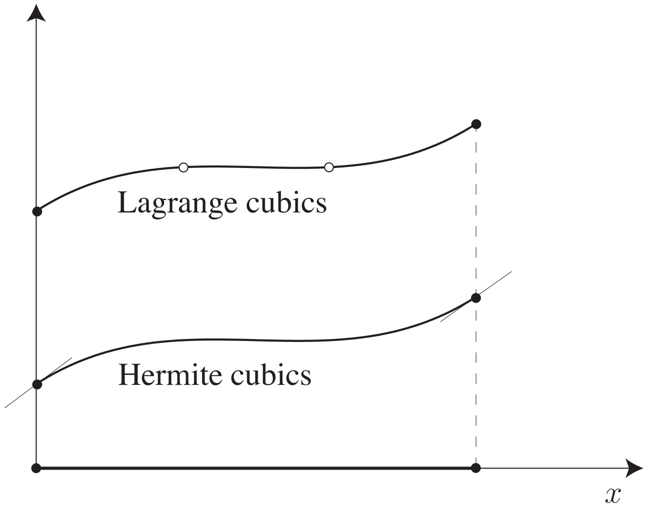


Figure 2.21: Piecewise cubic finite elements. Above: Lagrange cubics. Below: Hermite cubics.

Proofs of interpolation estimates can be found in Ciarlet [49], which is a standard reference for the general theory. Some interpolation estimates may be found in Johnson [164, pp. 24–25 and 84–91].

## 2.5.3 Some Useful Inequalities

### Two Elementary Inequalities

Let $z, z_1, z_2 \in \mathbb{R}$, then

$$z = z_1 + z_2 \Rightarrow |z| \leq |z_1| + |z_2| \tag{2.136}$$

**Remark 2.30** *This inequality is simply the triangular inequality on the linear space $\mathbb{R}$ equipped with $|\cdot|$ as norm.*

We can generalize (2.136) to a sum of $n$ real numbers, viz.,

$$z = \sum_{i=1}^{n} z_i \Rightarrow |z| \leq \sum_{i=1}^{n} |z_i| \tag{2.137}$$

**Inequality 1**

Let $a, b \in \mathbb{R}$. Then,

$$\boxed{|ab| \leq \frac{1}{2}(a^2 + b^2)} \tag{2.138}$$

*Proof*

**(i)** If $ab \geq 0$, then $0 \leq (a-b)^2 = a^2 - 2ab + b^2$. Thus, $ab = |ab| \leq \frac{1}{2}(a^2 + b^2)$.

**(ii)** If $ab < 0$, then $0 \leq (a+b)^2 = a^2 + 2ab + b^2$. Thus, $-ab = |ab| \leq \frac{1}{2}(a^2 + b^2)$.

□

**Inequality 2**

Let $a, b, \epsilon \in \mathbb{R}, \epsilon > 0$. Then,

$$\boxed{|ab| \leq \frac{1}{2}(\frac{a^2}{\epsilon} + \epsilon b^2)} \tag{2.139}$$

*Proof* Define $a' = a/\sqrt{\epsilon}$ and $b' = b\sqrt{\epsilon}$. Then, by Inequality 1, we have

$$\begin{aligned} |ab| = |a'b'| &\leq \tfrac{1}{2}(a'^2 + b'^2) \\ &= \frac{1}{2}(\frac{a^2}{\epsilon} + \epsilon b^2) \end{aligned} \tag{2.140}$$

□

**Inequality 3**

Let $v, w \in \mathcal{V}$, an inner product space. Then

$$\boxed{(v, w) \leq |(v, w)| \leq \frac{1}{2}\left(\frac{||v||^2}{\epsilon} + \epsilon||w||^2\right)} \tag{2.141}$$

where $|| \cdot ||$ is the natural norm on $\mathcal{V}$.

*Proof*

$$\begin{aligned} |(v, w)| &\leq ||v|| \; ||w|| && \text{(Cauchy-Schwarz)} \\ &\leq \tfrac{1}{2}\left(\frac{||v||^2}{\epsilon} + \epsilon||w||^2\right) && \text{(Inequality 2)} \end{aligned} \tag{2.142}$$

□

**Remark 2.31** *Inequality 3 is a very important result which is used repeatedly.*

### 2.5.4 Analysis of the Galerkin Method

We have already established

- The consistency of the method: $B(w^h, e) = 0 \;\; \forall \, w^h \in \mathcal{V}^h$.
- The stability of the method: $B(w^h, w^h) = \kappa \|w^h_{,x}\|^2 \;\; \forall \, w^h \in \mathcal{V}^h$.
- The interpolation estimate of the finite element space $\mathcal{S}^h$: We shall employ the interpolation estimate for the $H^1$-norm, namely, $\|\eta\|_1 \leq c\left(\frac{h}{L}\right)^k \|\phi\|_{k+1}$.

A basic step in finite element error analysis is to decompose the error, $e = \phi^h - \phi$, as follows:

$$e = \underbrace{\phi^h - \tilde{\phi}^h}_{e^h} + \underbrace{\tilde{\phi}^h - \phi}_{\eta} \tag{2.143}$$

where $e^h \in \mathcal{V}^h$ is the part of the error in the finite element space and $\eta \notin \mathcal{V}^h$ is the interpolation error. We now want to show that $e$ converges to zero in some norm, e.g., the $H^1$-norm. Typically, this is done in two steps:

- Estimate $e^h$. This is the difficult step.
- Assume that in Step 1 we have obtained an estimate of $e^h$ in some norm denoted by $|||\cdot|||$. Then, by the triangle inequality,

$$\begin{aligned} |||e||| &= |||e^h + \eta||| \\ &\leq |||e^h||| + |||\eta||| \end{aligned} \tag{2.144}$$

  Step 1 provides an estimate for $|||e^h|||$, and some type of interpolation estimate will take care of $|||\eta|||$. This completes the analysis.

The first part always involves the same steps. One starts with the stability result (2.128), using the fact that $e^h \in \mathcal{V}^h$, and proceeds as follows:

$$\begin{aligned} \kappa||e^h_{,x}||^2 &= B(e^h, e^h) \\ &= B(e^h, e - \eta) && \text{(definition of } e\text{)} \\ &= B(e^h, e) - B(e^h, \eta) && \text{(bilinearity)} \\ &= -B(e^h, \eta) && \text{(consistency)} \\ &= |B(e^h, \eta)| \\ &= \left| \int_0^L (-e^h_{,x} u\eta + e^h_{,x}\kappa\eta_{,x})\, dx \right| \\ &= |-u(e^h_{,x}, \eta) + \kappa(e^h_{,x}, \eta_{,x})| \\ &\leq u|(e^h_{,x}, \eta)| + \kappa|(e^h_{,x}, \eta_{,x})| \end{aligned} \tag{2.145}$$

The last inequality is one of the elementary inequalities. Applying Inequality 3 to the last line of (2.145), we get

$$\kappa||e^h_{,x}||^2 \leq u\frac{1}{2}\left(\frac{1}{\epsilon_1}||e^h_{,x}||^2 + \epsilon_1||\eta||^2\right) + \kappa\frac{1}{2}\left(\frac{1}{\epsilon_2}||e^h_{,x}||^2 + \epsilon_2||\eta_{,x}||^2\right) \tag{2.146}$$

where $\epsilon_1, \epsilon_2$ are positive, but otherwise arbitrary.

The idea is now to select $\epsilon_1$ and $\epsilon_2$ such that we have dimensional consistency, and such that we are able to "hide" the $||e^h_{,x}||^2$ term on the left-hand side. The meaning of this should become clear from the following analysis:

We will select $\epsilon_1$ and $\epsilon_2$ so that the $||e^h_{,x}||^2$-terms on the right-hand side sum to something smaller than the coefficient of $||e^h_{,x}||^2$ on the left-hand side, i.e., something smaller than $\kappa$, say $\kappa/2$. This can be achieved as follows:

$$\begin{aligned} u\frac{1}{2\epsilon_1} &\equiv \frac{\kappa}{4} \quad \Rightarrow \qquad \epsilon_1 = \frac{2u}{\kappa} \\ \kappa\frac{1}{2\epsilon_2} &\equiv \frac{\kappa}{4} \quad \Rightarrow \qquad \epsilon_2 = 2 \end{aligned} \tag{2.147}$$

Substituting these values of $\epsilon_1$ and $\epsilon_2$ into (2.146) yields

$$\kappa||e^h_{,x}||^2 \leq \frac{\kappa}{2}||e^h_{,x}||^2 + \frac{u^2}{\kappa}||\eta||^2 + \kappa||\eta_{,x}||^2 \tag{2.148}$$

We can now "hide" the $||e^h_{,x}||^2$ term of the right-hand side on the left-hand side:

$$\frac{\kappa}{2}||e^h_{,x}||^2 \leq \frac{u^2}{\kappa}||\eta||^2 + \kappa||\eta_{,x}||^2 \tag{2.149}$$

Rearranging,

$$\begin{aligned} ||e^h_{,x}||^2 &\leq 2\left(\frac{u^2}{\kappa^2}||\eta||^2 + ||\eta_{,x}||^2\right) \\ &= 2\left(\left(\frac{uh}{2\kappa}\right)^2\frac{4}{h^2}||\eta||^2 + ||\eta_{,x}||^2\right) \\ &= 2\left(\left(\frac{4\alpha^2}{h^2}\right)||\eta||^2 + ||\eta_{,x}||^2\right) \end{aligned} \tag{2.150}$$

where $\alpha = uh/(2\kappa)$ is the mesh Péclet number.

Interpolation estimates may be employed to bound the right-hand side. This result achieves the objective of Step 1, $e^h$ is estimated in a norm. (Recall that the $H^1$-seminorm constitutes a norm on $\mathcal{V}$. Consequently, it also serves as a norm on $\mathcal{V}^h \subset \mathcal{V}$.) Now that Step 1 is under control, we proceed to the simpler part, Step 2.

$$|e|_1 = |e^h + \eta|_1 \leq |e^h|_1 + |\eta|_1 \tag{2.151}$$

We work with the square of (2.151):

$$\begin{aligned}
|e|_1^2 &\le |e^h|_1^2 + 2|e^h|_1|\eta|_1 + |\eta|_1^2 \\
&\le |e^h|_1^2 + 2\tfrac{1}{2}\left(|e^h|_1^2 + |\eta|_1^2\right) + |\eta|_1^2 && \text{(Inequality 2)} \\
&= 2\left(|e^h|_1^2 + |\eta|_1^2\right) \\
&\le 2\big(2\left(\frac{4\alpha^2 L^2}{h^2}||\eta||^2 + |\eta|_1^2\right) + |\eta|_1^2\big) && \text{(by (2.150))} \\
&= 2\left(\frac{8\alpha^2 L^2}{h^2}||\eta||^2 + 3|\eta|_1^2\right) \\
&\le 2\left(\max\{3, 8\alpha^2\}\left(\frac{L^2}{h^2}||\eta||^2 + |\eta|_1^2\right)\right)
\end{aligned} \tag{2.152}$$

Taking square roots, and employing the interpolation estimate results in

$$\boxed{|e|_1 \le \max\{\sqrt{6}, 4\alpha\} c \left(\frac{h}{L}\right)^k ||\phi||_{k+1}} \tag{2.153}$$

This is the fundamental error estimate for the Galerkin finite element method. The rate of convergence (i.e., $k$) is optimal in the $|\cdot|_1$-norm. However, there is cause for concern due to the appearance of the mesh Péclet number on the right-hand side of (2.153). This suggests trouble in the advection-dominated case, as we will see in a moment.

***Discussion***

The error estimate (2.153) can be written in the form:

$$|e|_1 \le \begin{cases} \sqrt{6} c h^k ||\phi||_{k+1}, & 4\alpha \le \sqrt{6} \quad \text{(diffusion-dominated case)} \\ 4\alpha c h^k ||\phi||_{k+1}, & 4\alpha > \sqrt{6} \quad \text{(advection-dominated case)} \end{cases} \tag{2.154}$$

For linear finite elements, (2.154) specializes to

$$|e|_1 \le \begin{cases} \sqrt{6}\dfrac{h}{L} c ||\phi||_2, & \alpha\text{“small”} \\ \\ 4\alpha \dfrac{h}{L} c ||\phi||_2, & \alpha\text{“large”(advection-dominated case)} \end{cases} \tag{2.155}$$

When $\alpha$ is small (i.e., when $4\alpha \le \sqrt{6}$), everything is fine for Galerkin. But suppose $\alpha$ is large (i.e., $4\alpha \gg \sqrt{6}$). In this case, to reduce the factor $\alpha$ to $O(1)$ may require many elements. For example, assume we have a Péclet number of $20{,}000$. In this case, the number of elements will be

$$n_{\text{el}} = \frac{L}{h} = \frac{\text{Pe}}{2\alpha} = \frac{10{,}000}{\alpha} \tag{2.156}$$

Thus, $\alpha = O(1)$ requires $n_{\text{el}} = O(10{,}000)$ elements! This observation is consistent with the results obtained for the central difference method (see Section 2.2.3). Unless $h$ is very small, the Galerkin method is ineffective.

**Remark 2.32** *The term $||\phi||_2$ can be very large if $\text{Pe}$ is large. This remark is made precise in the following exercise.*

**Exercise 2.7** *Assume that $\text{Pe} \gg 1$. Estimate $||\phi||_2$ where $\phi$ is the solution of the advection-diffusion equation, assuming $f = 0$, $g_0 = 0$ and $g_L = 1$.*

**Solution**: $||\phi||_2 \sim L^{1/2}\text{Pe}^{3/2}$.

However, one must keep in mind that the interpolation estimates reported in Section 2.5.2 can be very pessimistic. They account for equally-spaced meshes rather than for meshes refined appropriately in regions of steep gradients (see Fig. 2.22).

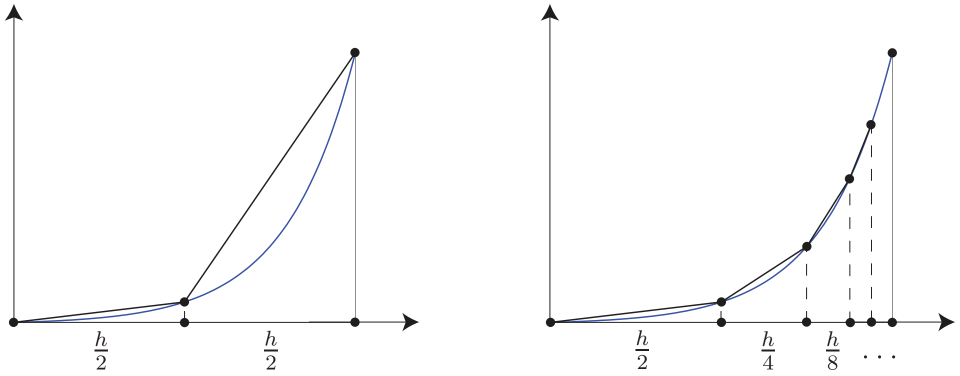


Figure 2.22: Interpolation with uniform and graded meshes.

**Remark 2.33** *For higher-order elements ($k > 1$), the interpolation estimate reads*

$$||\eta||_s \leq ch^{k+1-s}||\phi||_{k+1}. \tag{2.157}$$

*As $k$ is increased, $h^{k+1-s}$ decreases, whereas $||\phi||_{k+1}$ increases. The reader may wish to verify this statement as an exercise. For a sufficiently fine mesh, the situation is that $h^{k+1-s}$ will die off quickly enough to more than compensate for the larger $||\phi||_{k+1}$. However, for a coarse mesh it is not clear whether higher-order elements offer advantages. This may be observed from Figure 2.23 which depicts the linear interpolant on a six-element mesh and the quadratic interpolant on a corresponding three-element mesh. Near $x = 1$, the quadratic interpolant gives a better approximation of the slope, but it also exhibits an undershoot between nodes 4 and 5. Which interpolant is "better" depends how we define "better". This is a somewhat subjective matter. This remark has nothing in particular to do with the Galerkin method; it is based solely on interpolation properties.*

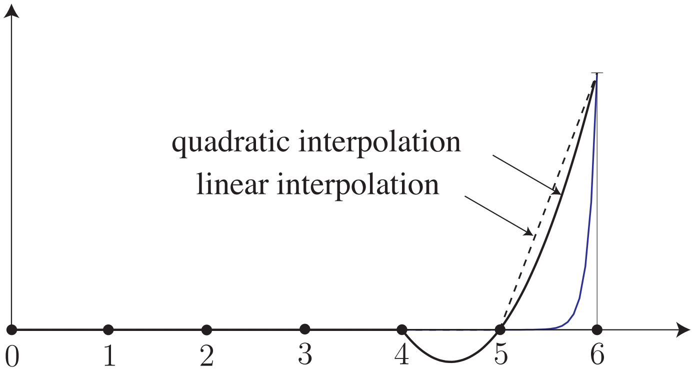


Figure 2.23: Linear versus quadratic interpolation.

**Exercise 2.8** *Assume*

$$\mathrm{Pe} = 2000 \quad , \quad g_0 = 0 \quad , \quad g_L = 1 \quad , \quad f = 0 \tag{2.158}$$

*The exact solution has been given previously. Likewise, for* $h = \text{const}$ *and* $k = 1$*, we have the solutions for the Galerkin and SUPG methods. (Observe that, under the stated assumptions, SUPG and Galerkin/least-squares are identical. Furthermore, the solution for these cases is the interpolate.) Write a computer program to plot for each method on a log-log scale, the* $L_2$*-norm of the error,* $||e||$*, the* $H^1$*-seminorm,* $|e|_1$*, and the* $H^1$*-norm,* $||e||_1$*, vs. the number of elements,* $L/h = 1, 2, 5, 10, 10^2, 10^3, 10^4, \text{and } 10^5$*. Note that if*

$$|||e||| \sim \beta \left(\frac{h}{L}\right)^{\gamma} \tag{2.159}$$

*where* $\beta$ *and* $\gamma$ *are two constants, then (see Fig. 2.24):*

$$\begin{aligned} \log |||e||| &\sim \log \beta + \gamma \log \left(\tfrac{h}{L}\right) \\ &= \log \beta - \gamma \log n_{\mathrm{el}} \end{aligned} \tag{2.160}$$

- *Determine* $\gamma$ *for each discernable branch.*
- *Comment on* $\gamma$ *and also on absolute error.*
- *Interpret the different branches.*

---

**Exercise 2.9** *(Some simple interpolation estimates in the "max-norm.")*

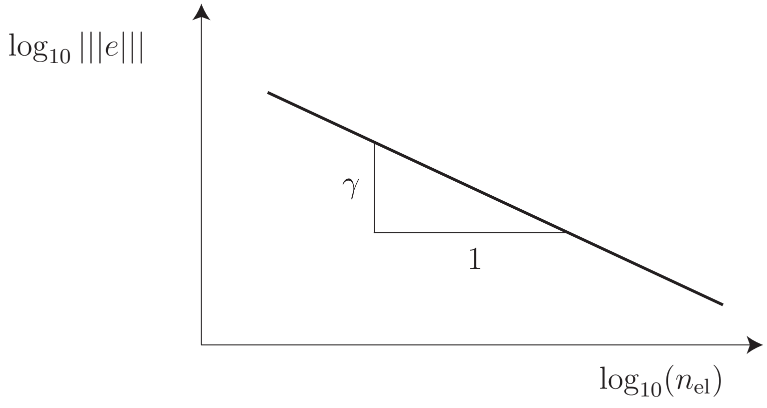


Figure 2.24: $|||e|||$ versus $n_{\text{el}}$.

*Consider piecewise linear finite element spaces. Given* $\phi \in C^2\,(]0, L[)$*, show that the interpolation error* $\eta = \tilde{\phi}^h - \phi$ *satisfies*

$$\begin{aligned} |\eta(x)| &\leq c_1 h^2 \max\nolimits_{y\in[0,L]} |\phi_{,xx}(y)| \\ |\eta_{,x}(x)| &\leq c_2 h \max\nolimits_{y\in[0,L]} |\phi_{,xx}(y)| \end{aligned} \tag{2.161}$$

*If we define the max-norm by*

$$|\eta|_{\max} = \max_{x\in[0,L]} |\eta(x)| \tag{2.162}$$

*the above inequalities become*

$$\begin{aligned} |\eta|_{\max} &\leq c_1 h^2 |\phi_{,xx}|_{\max} \\ |\eta_{,x}|_{\max} &\leq c_2 h |\phi_{,xx}|_{\max} \end{aligned} \tag{2.163}$$

*Determine the smallest values of* $c_1$ *and* $c_2$*. Note the similarities with Sobolev norm estimates.*

*Hint: Use finite Taylor expansions with derivative form of the remainder (see Hughes [117, p.27–30] for related calculations).*

### 2.5.5 Inverse Estimates

The inverse estimates enable us to evaluate the stability of the finite element functions. They are based on estimates for ***individual*** elements. The implications of this observation are crucial to the analysis of the Petrov-Galerkin methods. The objective is to bound the higher derivatives of finite element functions by lower derivatives, up to powers of the mesh parameter $h$. Let $\Omega^e$ denote the ***interior of element number e*** where $e = 1, 2, \ldots, n_{\text{el}}$, in which $n_{\text{el}}$ is the ***number of elements***. In our mesh consisting of piecewise linear elements, $n_{\text{el}} = A_{\max}$, and $\Omega^e$ is the interior of the open interval $]x_{A-1}, x_A[$, where $e = A$.

We decompose the $H^r(\Omega)$ seminorm into element contributions as follows:

$$\begin{aligned}|w|^2_{H^r(\Omega)} &\equiv |w|^2_r \\ &= \textstyle\int_\Omega L^{2r}(w_{,\underbrace{x\ldots x}_{r\text{ times}}})^2\, d\Omega \\ &= \textstyle\sum_{e=1}^{n_{\text{el}}}\int_{\Omega^e} L^{2r}(w_{,\underbrace{x\ldots x}_{r\text{ times}}})^2\, d\Omega \\ &= \textstyle\sum_{e=1}^{n_{\text{el}}} |w|^2_{H^r(\Omega^e)}\end{aligned} \tag{2.164}$$

that is,

$$|w|_{H^r(\Omega^e)} = \left(\int_{\Omega^e} L^{2r}(w_{,\underbrace{x\ldots x}_{r\text{ times}}})^2\, d\Omega\right)^{1/2} \tag{2.165}$$

It can be shown that, for all finite element functions $w^h$,

$$\boxed{|w^h|_{H^s(\Omega^e)} \le c_I\left(\frac{h}{L}\right)^{r-s}|w^h|_{H^r(\Omega^e)} \qquad \text{for } r \le s} \tag{2.166}$$

where $c_I$ is a non-dimensional constant independent of $h$. Higher-order elements have the effect of increasing $c_I$ (as $k \to \infty$, $c_I \to \infty$). It is convenient to express (2.166) in a slightly different form. A specific example that we will have occasion to use in our analyses of Petrov-Galerkin methods is developed as follows:

$$\begin{aligned}||w^h_{,xx}||_{\Omega^e} &= L^{-2}|w^h|_{H^2(\Omega^e)} \\ &\le L^{-2}c_I\left(\frac{h}{L}\right)^{-1}|w^h|_{H^1(\Omega^e)} \\ &= c_I h^{-1}||w^h_{,x}||_{\Omega^e}\end{aligned} \tag{2.167}$$

Thus,

$$\boxed{||w^h_{,xx}||_{\Omega^e} \le c_I h^{-1}||w^h_{,x}||_{\Omega^e}} \tag{2.168}$$

In general, (2.166) leads to

$$\boxed{||w^h_{,\underbrace{x\ldots x}_{s\text{ times}}}||_{\Omega^e} \le c_I h^{r-s}||w^h_{,\underbrace{x\ldots x}_{r\text{ times}}}||_{\Omega^e}} \tag{2.169}$$

Summing the squares over element interiors yields

$$\sum_{e=1}^{n_{\text{el}}}||w^h_{,xx}||^2_{\Omega^e} \le c_I^2 h^{-2}\sum_{e=1}^{n_{\text{el}}}||w^h_{,x}||^2_{\Omega^e} \tag{2.170}$$

$$\Rightarrow \qquad \boxed{||w^h_{,xx}||_{\Omega'} \leq c_I h^{-1} ||w^h_{,x}||_{\Omega'} = c_I h^{-1} ||w^h_{,x}||_{\Omega}} \tag{2.171}$$

where $\Omega' = \bigcup_e \Omega^e$ is the union of the element interiors, and $|| \cdot ||_{\mathcal{D}}$ is the $L_2(\mathcal{D})$-norm, $\mathcal{D}$ indicating an arbitrary domain.

**Remark 2.34** *Note that $||w^h_{,xx}||_{\Omega}$ is meaningless for typical $C^0(\Omega)$ finite element functions because $w^h \notin H^2(\Omega)$ due to the appearance of Dirac delta functions in $w^h_{,xx}$.*

**Remark 2.35** *For $k = 1$, $c_I = 0$ because $||w^h_{,xx}||_{\Omega'} \equiv 0$.*

**Exercise 2.10** *Some simple inverse estimates.*

**a.** *Show that*

$$||w^h_{,x}||_{\Omega^e} \leq c_I h^{-1} ||w^h||_{\Omega^e} \tag{2.172}$$

*for*

**i)** *the two-node linear element ($k = 1$); and*

**ii)** *the three-node quadratic element ($k = 2$). Assume the nodes are equally spaced; see Figure 2.25.*

*Determine the smallest $c_I$ in each case.*

**b.** *Show that*

$$||w^h_{,xx}||_{\Omega^e} \leq c_I h^{-1} ||w^h_{,x}||_{\Omega^e} \tag{2.173}$$

*for the three-node element and determine the smallest $c_I$.*

**c.** *Show that*

$$||w^h_{,xx}||_{\Omega^e} \leq c_I h^{-2} ||w^h||_{\Omega^e} \tag{2.174}$$

*for the three-node element and determine the smallest $c_I$.*

---

**Remark 2.36** *Some inverse estimates are discussed in Johnson [164, pp.141–145] and Ciarlet [49, pp.140–143] is the standard reference for the general theory.*

### 2.5.6 Analysis of the SUPG Method

In order to prove convergence of the SUPG method,

$$B_{\text{SUPG}}\left(w^h, \phi^h\right) = L_{\text{SUPG}}\left(w^h\right) \qquad \forall\, w^h \in \mathcal{V}^h, \tag{2.175}$$

we must establish the appropriate stability and consistency results. The four required ingredients are:

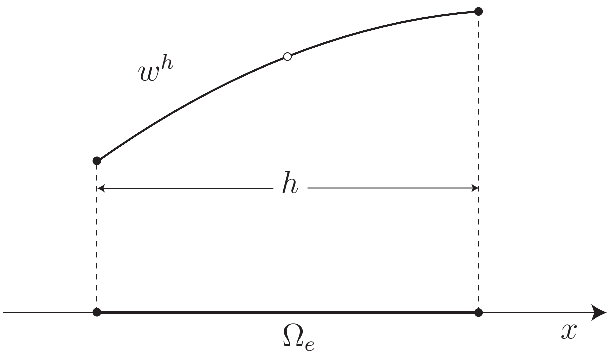


Figure 2.25: Three-node quadratic element.

- Stability of the method;
- Inverse estimate;
- Consistency of the method (weighted residual method);
- Interpolation estimate.

**Stability of the SUPG Method**

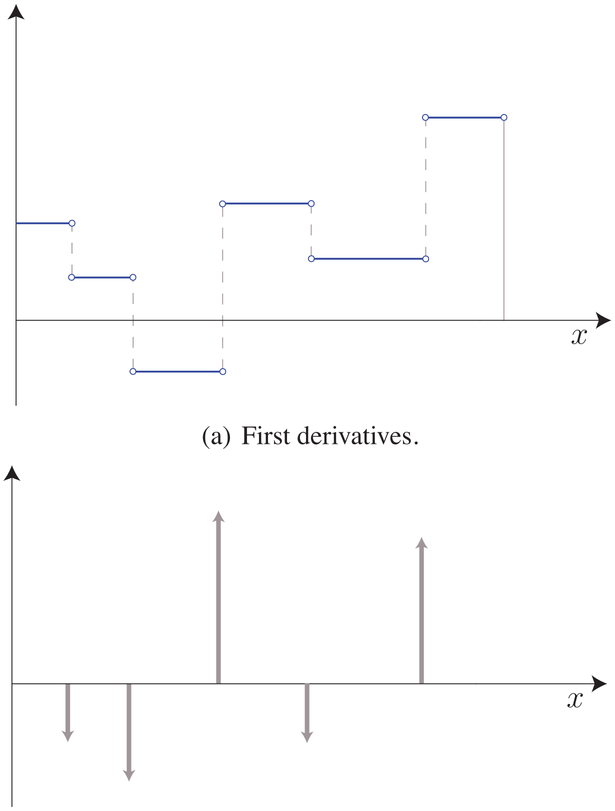
By definition (see Section 2.3.6), $\forall\ w^h \in \mathcal{V}^h$

$$B_{\text{SUPG}}\left(w^h, w^h\right) = B\left(w^h, w^h\right) + \sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \tau u w^h_{,x}\left(u w^h_{,x} - \kappa w^h_{,xx}\right) dx \tag{2.176}$$

From the analysis of the Galerkin method, we know that

$$B\left(w^h, w^h\right) = \kappa ||w^h_{,x}||^2 \tag{2.177}$$

We will now make the simplifying assumption that $h = \text{const}$. The case $h \neq \text{const}$ is essentially the same; the analysis would however be slightly encumbered. In addition, we will employ the following notational simplification: we replace $\sum\limits_{A=1}^{N} \int_{x_{A-1}}^{x_A}$ by $\sum\limits_{e=1}^{n_{\text{el}}} \int_{\Omega^e}$, where recall $\Omega^e$ denotes the interior of the $e^{\text{th}}$ element domain. Furthermore, the $L_2(\mathcal{D})$-inner product over a domain $\mathcal{D}$



(a) First derivatives.

(b) Second derivatives (Dirac delta functions).

Figure 2.26: First and second derivatives of a piecewise linear function.

will be noted $(\cdot,\cdot)_{\mathcal{D}}$. With these conventions, we write:

$$\begin{aligned}
\tau u^2 \textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} \left(w^h_{,x}\right)^2 dx &= \tau u^2 \textstyle\int_0^L \left(w^h_{,x}\right)^2 dx \qquad \text{(since } w^h_{,x} \in H^1\text{)} \\
&= \tau u^2 ||w_{,x}||^2 \\
-\tau u \kappa \textstyle\sum_{A=1}^{N} \int_{x_{A-1}}^{x_A} w^h_{,x} w^h_{,xx}\, dx &= -\tau u \kappa \textstyle\sum_{e=1}^{n_{\mathrm{el}}} \int_{\Omega^e} w^h_{,x} w^h_{,xx}\, dx \\
&= -\tau u \kappa \left(w^h_{,x}, w^h_{,xx}\right)_{\Omega'}
\end{aligned} \tag{2.178}$$

where $\Omega'$ denotes the union of element interiors, i.e.,

$$\Omega' \equiv \bigcup_{e=1}^{n_{\mathrm{el}}} \Omega^e \tag{2.179}$$

Thus,

$$\begin{aligned} B_{\mathrm{SUPG}}(w^h, w^h) &= \kappa ||w^h_{,x}||^2 + \tau u^2 ||w^h_{,x}||^2 - \tau u \kappa \left( w^h_{,x}, w^h_{,xx} \right)_{\Omega'} \\ &\geq (\kappa + \tau u^2) \, ||w^h_{,x}||^2 - \frac{1}{2} \tau u \kappa \left( \frac{1}{\epsilon} ||w^h_{,x}||^2 + \epsilon ||w^h_{,xx}||^2_{\Omega'} \right) \\ &= \left( \kappa + \tau u^2 - \frac{\tau u \kappa}{2\epsilon} \right) ||w^h_{,x}||^2 - \frac{1}{2} \tau u \kappa \epsilon ||w^h_{,xx}||^2_{\Omega'} \end{aligned} \tag{2.180}$$

(Note, we have used the fact that $w^h \in H^1$ implies that $||w^h_{,x}||_{\Omega'} = ||w^h_{,x}||$. Figure 2.26 shows that the first derivative of a piecewise linear function is a piecewise constant function which is square-integrable, whereas its generalized second derivative is a collection of Dirac delta functions whose squares are not even defined. Hence the integration of the first derivatives squared over the element interiors can be extented to the entire domain.)

Thus, using the inverse estimate (2.171), we have $\forall \; w^h \in \mathcal{V}^h$

$$B_{\mathrm{SUPG}}(w^h, w^h) \geq \left( \kappa + \tau u^2 - \frac{\tau u \kappa}{2\epsilon} - \frac{\tau u \kappa}{2} \epsilon c_I^2 h^{-2} \right) ||w^h_{,x}||^2 \tag{2.181}$$

Recall that $\epsilon$ is positive, but otherwise arbitrary. We assume that our formula for $\tau$ has the correct attributes. This turns out to be the case. (If we had no insight into $\tau$, we could let the analysis determine its necessary properties.) Looking ahead a bit, let us try $\epsilon = \tau u$. This leads to

$$\begin{aligned} B_{\mathrm{SUPG}}(w^h, w^h) &\geq \left( \frac{\kappa}{2} + \tau u^2 - \frac{(\tau u)^2}{2} \kappa c_I^2 h^{-2} \right) ||w^h_{,x}||^2 \\ &= \left( \frac{\kappa}{2} + \tau u^2 \left( 1 - \frac{\tau \kappa c_I^2 h^{-2}}{2} \right) \right) ||w^h_{,x}||^2 \end{aligned} \tag{2.182}$$

The appearance of the $\tau$ terms in (2.182) provides a device for improving stability in the advection-dominated limit, in contrast with the Galerkin method in which case $\tau = 0$. To ensure that the coefficient of $\tau u^2$ is sufficiently large, let us insist that

$$1 - \frac{\tau \kappa c_I^2 h^{-2}}{2} \geq \frac{1}{2} \tag{2.183}$$

This places a restriction on $\tau$, namely

$$\tau \leq \frac{h^2}{\kappa c_I^2} \tag{2.184}$$

But, recall that

$$\tau = \frac{\tilde{\kappa}}{u^2} = \frac{1}{u^2} \frac{|u| h}{2} \tilde{\xi} = \frac{h}{2|u|} \tilde{\xi} \tag{2.185}$$

where $\tilde{\xi}$ is shown in Figure 2.4. Thus, (2.184) requires

$$\frac{h}{2|u|} \tilde{\xi} \leq \frac{h^2}{\kappa c_I^2} \tag{2.186}$$

which in turn restricts $\tilde{\xi}$ to satisfy

$$\tilde{\xi} \leq \frac{2|u|h}{\kappa c_I^2} = \frac{4\alpha}{c_I^2} \tag{2.187}$$

Recall that, for $k = 1, c_I = 0$. Therefore, the definition of $\tilde{\xi}$ presented in (2.35) satisfies (2.187) for this case. For higher-order elements, $c_I$ increases. Thus, $\tilde{\xi}$ must have a smaller slope at the origin and, from (2.187) it can grow no more than linearly with $\alpha$ (see Fig. 2.27).

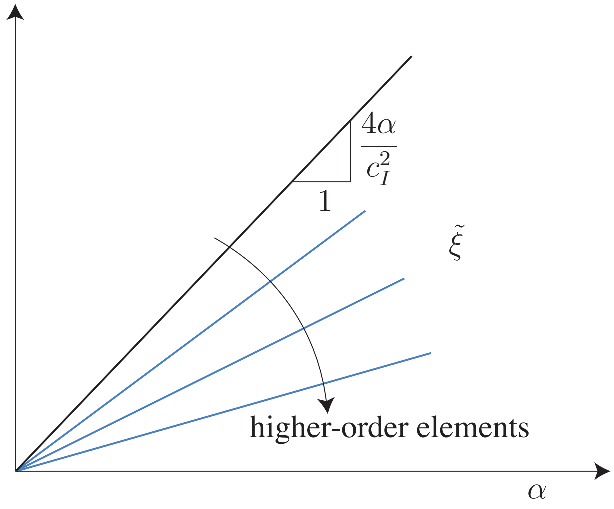


Figure 2.27: Stability region for $\tilde{\xi}$, in the case of higher-order elements.

**Remark 2.37** *The stability condition is primarily a condition on the diffusive end of the spectrum. It requires that the SUPG terms vanish sufficiently fast as* $\alpha \rightarrow 0$.

**Remark 2.38** *In the advection-dominated limit, we have*

$$(\tau u w^h_{,x}, (u w^h_{,x} - \kappa w^h_{,xx}))_{\Omega'} \rightarrow \tau u^2 ||w^h_{,x}||^2 \tag{2.188}$$

*It is clear in this case that the additional term is stabilizing.*

**Remark 2.39** *We can always adjust the slope of* $\tilde{\xi}$ *at the origin for any type of element to ensure the stability condition (2.187) is satisfied.*

According to the previous results, we can write $\forall\ w^h \in \mathcal{V}^h$

$$\begin{aligned} B_{\mathrm{SUPG}}(w^h, w^h) &\geq \frac{1}{2}\left(\kappa + \tau u^2\right)||w^h_{,x}||^2 \\ &\geq \frac{1}{2}\left(\kappa + \tilde{\kappa}\right)||w^h_{,x}||^2 \\ &\geq \frac{1}{2}\left(\kappa + \frac{|u|h}{2}\tilde{\xi}\right)||w^h_{,x}||^2 \\ &\geq \frac{1}{2}\kappa\left(1 + \alpha\tilde{\xi}\right)||w^h_{,x}||^2 \end{aligned} \tag{2.189}$$

This result should be contrasted with the Galerkin analog (2.128). The extra term here will be seen to provide the required stability in the advection-dominated situation.

### Convergence of the SUPG Method

The error is again expressed as $e = \phi^h - \phi = e^h + \eta$. Beginning with the stability result above, we proceed in similar fashion to the analysis of the Galerkin method (2.145):

$$\begin{aligned} \tfrac{1}{2}\kappa\left(1 + \alpha\tilde{\xi}\right)||e^h_{,x}||^2 &\leq B_{\mathrm{SUPG}}(e^h, e^h) \\ &= B_{\mathrm{SUPG}}(e^h, e - \eta) \\ &= B_{\mathrm{SUPG}}(e^h, e) - B_{\mathrm{SUPG}}(e^h, \eta) \\ &= |B_{\mathrm{SUPG}}(e^h, \eta)| \\ &= \left|B(e^h, \eta) + \textstyle\sum_{e=1}^{n_{\mathrm{el}}}\int_{\Omega^e}\tau u e^h_{,x}(u\eta_{,x} - \kappa\eta_{,xx})\,dx\right| \\ &= \left|B(e^h, \eta) + \tau u^2\left(e^h_{,x}, \eta_{,x}\right) - \tau u\kappa\left(e^h_{,x}, \eta_{,xx}\right)_{\Omega'}\right| \\ &\leq \left|B(e^h, \eta)\right| + \tau u^2\left|\left(e^h_{,x}, \eta_{,x}\right)\right| + \tau u\kappa\left|\left(e^h_{,x}, \eta_{,xx}\right)_{\Omega'}\right| \end{aligned} \tag{2.190}$$

$|B(e^h, \eta)|$ was analyzed in the Galerkin convergence proof. Reviewing (2.145)–(2.148), we have the estimate

$$|B(e^h, \eta)| \leq \frac{\kappa}{2}||e^h_{,x}||^2 + \left(\frac{u^2}{\kappa}||\eta||^2 + \kappa||\eta_{,x}||^2\right) \tag{2.191}$$

The $u^2/\kappa$-term led to the $(\mathrm{Pe})^2$-term in the Galerkin analysis, which was bad. The analysis which led to (2.191) can be redone replacing $\epsilon_1$ and $\epsilon_2$ in (2.146) by $\epsilon\epsilon_1$ and $\epsilon\epsilon_2$, respectively, leading to

$$|B(e^h, \eta)| \leq \frac{1}{\epsilon}\frac{\kappa}{2}||e^h_{,x}||^2 + \epsilon\left(\frac{u^2}{\kappa}||\eta||^2 + \kappa||\eta_{,x}||^2\right) \tag{2.192}$$

Looking forward again, it turns out that a good choice for $\epsilon$ is

$$\epsilon = \frac{2}{1+\frac{1}{2}\alpha\tilde{\xi}} \tag{2.193}$$

Then, (2.192) becomes

$$|B(e^h,\eta)| \leq \left(1+\frac{1}{2}\alpha\tilde{\xi}\right)\frac{1}{2}\frac{\kappa}{2}||e^h_{,x}||^2 + \frac{2}{\left(1+\frac{1}{2}\alpha\tilde{\xi}\right)}\left(\frac{u^2}{\kappa}||\eta||^2+\kappa||\eta_{,x}||^2\right) \tag{2.194}$$

We can now substitute (2.194) into (2.190), and bring the term $\left(1+\frac{1}{2}\alpha\tilde{\xi}\right)\frac{1}{2}\frac{\kappa}{2}||e^h_{,x}||^2$ to the left-hand side:

$$\begin{aligned}\frac{\kappa}{4}\left(1+\frac{3}{2}\alpha\tilde{\xi}\right)||e^h_{,x}||^2 \leq \frac{2}{1+\dfrac{1}{2}\alpha\tilde{\xi}} & \left(\frac{u^2}{\kappa}||\eta||^2+\kappa||\eta_{,x}||^2\right) \\ & +\tau u^2\left|\left(e^h_{,x},\eta_{,x}\right)\right| \\ & +\tau u\kappa\left|\left(e^h_{,x},\eta_{,xx}\right)_{\Omega'}\right|\end{aligned} \tag{2.195}$$

Dividing through by $\kappa$, and employing $\alpha = uh/2\kappa$ and Inequality 3, we obtain

$$\begin{aligned}\frac{1}{4}\left(1+\frac{3}{2}\alpha\tilde{\xi}\right)||e^h_{,x}||^2 \leq \frac{2}{\left(1+\dfrac{1}{2}\alpha\tilde{\xi}\right)} & \left(\frac{4\alpha^2}{h^2}||\eta||^2+||\eta_{,x}||^2\right) \\ & +\frac{\tau u^2}{\kappa}\frac{1}{2}\left(\frac{1}{\epsilon_1}||e^h_{,x}||^2+\epsilon_1||\eta_{,x}||^2\right) \\ & +\tau u\frac{1}{2}\left(\frac{1}{\epsilon_2}||e^h_{,x}||^2+\epsilon_2||\eta_{,xx}||^2_{\Omega'}\right)\end{aligned} \tag{2.196}$$

In order to hide the $||e^h_{,x}||^2$-terms of the right-hand side in the left-hand side, we choose

$$\begin{aligned}\frac{\tau u^2}{\kappa}\frac{1}{2}\frac{1}{\epsilon_1} = \frac{1}{16}\alpha\tilde{\xi} \quad &\Rightarrow \qquad \epsilon_1 = \frac{8\tau u^2}{\kappa\alpha\tilde{\xi}} \\ \tau u\frac{1}{2}\frac{1}{\epsilon_2} = \frac{1}{16}\alpha\tilde{\xi} \quad &\Rightarrow \qquad \epsilon_2 = \frac{8\tau u}{\alpha\tilde{\xi}}\end{aligned} \tag{2.197}$$

With these, (2.196) becomes:

$$\begin{aligned}\frac{1}{4}\left(1+\alpha\tilde{\xi}\right)||e^h_{,x}||^2 \leq \frac{2}{\left(1+\dfrac{1}{2}\alpha\tilde{\xi}\right)} & \left(\frac{4\alpha^2}{h^2}||\eta||^2+||\eta_{,x}||^2\right) \\ & +\frac{4}{\alpha\tilde{\xi}}\left(\frac{\tau u^2}{\kappa}\right)^2||\eta_{,x}||^2 \\ & +\frac{4}{\alpha\tilde{\xi}}\left(\tau u\right)^2||\eta_{,xx}||^2_{\Omega'}\end{aligned} \tag{2.198}$$

Let us simplify the coefficients of the second and third terms on the right-hand side of (2.198). Recalling the definition of $\tau$, we have

$$\begin{aligned} \frac{4}{\alpha\tilde{\xi}}\left(\frac{\tau u^2}{\kappa}\right)^2 &= \frac{4}{\alpha\tilde{\xi}}\left(\frac{h}{2|u|}\tilde{\xi}\right)^2\left(\frac{u^2}{\kappa}\right)^2 \\ &= \frac{4}{\alpha\tilde{\xi}}(\alpha\tilde{\xi})^2 = 4\alpha\tilde{\xi} \end{aligned} \tag{2.199}$$

and

$$\begin{aligned} \frac{4}{\alpha\tilde{\xi}}(\tau u)^2 &= \frac{4}{\alpha\tilde{\xi}}\left(\frac{h}{2|u|}\tilde{\xi}u\right)^2 \\ &= \frac{4}{\alpha\tilde{\xi}}\frac{h^2}{4}\tilde{\xi}^2 = h^2\frac{\tilde{\xi}}{\alpha} \end{aligned} \tag{2.200}$$

This last term may appear to be of some concern. However, recall from Figure 2.27 that

$$\tilde{\xi} \leq \frac{\alpha}{3} \tag{2.201}$$

Therefore,

$$h^2\frac{\tilde{\xi}}{\alpha} \leq \frac{h^2}{3} \tag{2.202}$$

Employing these in (2.198) we have,

$$\begin{aligned} \frac{1}{4}\left(1+\alpha\tilde{\xi}\right)||e^h_{,x}||^2 &\leq \Bigg\{\frac{2}{\left(1+\frac{1}{2}\alpha\tilde{\xi}\right)}\left(\frac{4\alpha^2}{h^2}||\eta||^2 + ||\eta_{,x}||^2\right) + 4\alpha\tilde{\xi}||\eta_{,x}||^2 \\ &\quad + \frac{1}{3}h^2||\eta_{,xx}||^2_{\Omega'}\Bigg\} \end{aligned} \tag{2.203}$$

Each coefficient of the right-hand side, when divided by the coefficient of the left-hand side, $\frac{1}{4}\left(1+\alpha\tilde{\xi}\right)$, can be bounded by a constant:

$$\begin{aligned} \frac{4}{\left(1+\alpha\tilde{\xi}\right)}\frac{2}{\left(1+\frac{1}{2}\alpha\tilde{\xi}\right)}4\alpha^2 &\leq c_1 = 64 \\ \frac{4}{\left(1+\alpha\tilde{\xi}\right)}\frac{2}{\left(1+\frac{1}{2}\alpha\tilde{\xi}\right)} &\leq c_2 = 8 \\ \frac{4}{\left(1+\alpha\tilde{\xi}\right)}4\alpha\tilde{\xi} &\leq c_3 = 16 \\ \frac{4}{\left(1+\alpha\tilde{\xi}\right)}\frac{1}{3} &\leq c_4 = \frac{4}{3} \end{aligned} \tag{2.204}$$

The reader may wish to verify the values of the $c$'s as an exercise. The precise values are not too important. The crucial point is that $c_1, c_2, c_3$, and $c_4$ are *independent* of $\alpha$ and $h$. Let $\bar{c} = \max\{c_1, c_2, c_3, c_4\}$. Then we get:

$$L^{-2}|e^h|_1^2 \equiv ||e^h_{,x}||^2 \leq \bar{c}\left(\frac{1}{h^2}||\eta||^2 + ||\eta_{,x}||^2 + h^2||\eta_{,xx}||^2_{\Omega'}\right) \tag{2.205}$$

The interpolation estimates for the terms on the right-hand side of (2.205) are

$$\begin{aligned} ||\eta|| &\leq c\left(\frac{h}{L}\right)^{k+1} ||\phi||_{k+1} \\ L||\eta_{,x}|| = |\eta|_1 &\leq c\left(\frac{h}{L}\right)^{k} ||\phi||_{k+1} \end{aligned} \tag{2.206}$$

and

$$L^2||\eta_{,xx}||_{\Omega'} = |\eta|_{H^2(\Omega')} \leq c\left(\frac{h}{L}\right)^{k-1} ||\phi||_{k+1} \tag{2.207}$$

Employing (2.206)–(2.207) in (2.205) yields:

$$\begin{aligned} |e^h|_1^2 &\leq \bar{c}c^2\left[\left(\frac{h}{L}\right)^{2k} + \left(\frac{h}{L}\right)^{2k} + \left(\frac{h}{L}\right)^{2k}\right] ||\phi||^2_{k+1} \\ &= 3\bar{c}c^2\left(\frac{h}{L}\right)^{2k} ||\phi||^2_{k+1} \end{aligned} \tag{2.208}$$

Letting $\hat{c}^2 = 3\bar{c}c^2$ and taking the square-roots leads to the following ***uniform*** estimate, valid for all values of $\alpha$,

$$\boxed{|e^h|_1 \leq \hat{c}\left(\frac{h}{L}\right)^{k} ||\phi||_{k+1}} \tag{2.209}$$

Use of the triangle inequality and the interpolation estimate (2.206) yields

$$\begin{aligned} |e|_1 &= |e^h + \eta|_1 \\ &\leq |e^h|_1 + |\eta|_1 \\ &\leq |e^h|_1 + c\left(\frac{h}{L}\right)^{k} ||\phi||_{k+1} \end{aligned} \tag{2.210}$$

Finally, combining (2.209) and (2.210), replacing $\hat{c} + c$ with $c$, completes the convergence proof for SUPG:

$$\boxed{|e|_1 \leq c\left(\frac{h}{L}\right)^{k} ||\phi||_{k+1}} \tag{2.211}$$

This estimate is ***uniform*** in $\alpha$.

□

We summarize the ingredients used in proving this result:

- Stability of the method:

$$B_{\mathrm{SUPG}}(w,w) \geq \frac{1}{2}\kappa(1+\alpha\tilde{\xi})||w_{,x}||^2 \qquad \forall\, w \in \mathcal{V}^h \tag{2.212}$$

- The inverse estimate:

$$||w_{,xx}||_{\Omega'} \leq c_I h^{-1}||w_{,x}|| \qquad \forall\, w \in \mathcal{V}^h \tag{2.213}$$

- Consistency of the method:

$$B_{\mathrm{SUPG}}(w,e) = 0 \qquad \forall\, w \in \mathcal{V}^h \tag{2.214}$$

- Interpolation estimate:

$$\left(\frac{h}{L}\right)^{-2}||\eta||^2 + |\eta|_1^2 + \left(\frac{h}{L}\right)^2 |\eta|^2_{H^2(\Omega')} \leq 3c^2 \left(\frac{h}{L}\right)^{2k} ||\phi||_{k+1} \tag{2.215}$$

The analysis also imposed restrictions on $\tau = \dfrac{h}{2|u|}\tilde{\xi}$, namely

1. $\tilde{\xi} \leq \dfrac{4\alpha}{c_I^2}$. This inequality was used in establishing the stability result. It is primarily a condition on the slope of $\tilde{\xi}$ at $\alpha = 0$.
2. $\dfrac{\tilde{\xi}}{\alpha} \leq \mathrm{const} = \dfrac{1}{3}$. The choice of the constant is actually arbitrary.

The first restriction implies the second one as long as $c_I \neq 0$.

**Remark 2.40** *Analyses of related situations (e.g., the hyperbolic limit $\kappa \to 0$) reveal that $\tilde{\xi}$ should approach a constant value (e.g., $1$ as in (2.35)) as $\alpha \to \infty$.*

**Remark 2.41** *The asymptotic properties of the optimal $\tau$, determined for linear elements by*

*(2.35), are given as follows:*

*For $\alpha$ small,*

$$\begin{aligned}\tau &\leq \frac{h}{2|u|}\frac{\alpha}{3} \\ &= \frac{h}{2|u|}\frac{1}{3}\frac{|u|h}{2\kappa} \\ &= \frac{h^2}{12\kappa} \\ &= O\left(\frac{h^2}{\kappa}\right) \qquad (2.216)\end{aligned}$$

*For $\alpha$ large,*

$$\tau \leq \frac{h}{2|u|} \cdot 1 = O\left(\frac{h}{|u|}\right) \qquad (2.217)$$

*A doubly asymptotic approximation is a computationally convenient alternative to the optimal $\tilde{\xi}$ (see Fig. 2.28). The analysis of SUPG reveals that a doubly asymptotic approximation is sufficient to maintain the optimal convergence rate of the error in the $H^1$-norm.*

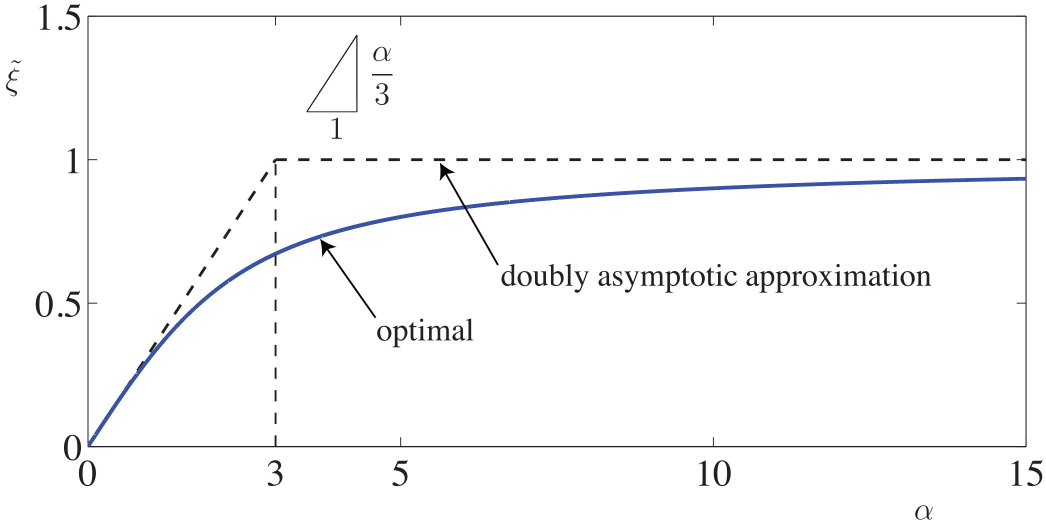


Figure 2.28: Doubly asymptotic approximation of optimal $\tilde{\xi}$.

**Remark 2.42** *The crucial difference in the analysis of SUPG is the additional term in the stability estimate. Without this term, we would run into the same deficiency noted in the analysis of the Galerkin method, namely coefficients on the right-hand side of order the mesh Péclet* *number. Note from (2.204) that, without the additional $\alpha\tilde{\xi}$-terms appearing in the denominator, $c_1$ would be $O(\alpha)$ in the advection-dominated case, giving rise to the aforementioned problem. The additional stability provided by SUPG changes this situation dramatically. At the same time, accuracy is not degraded: the error estimate (2.211) exhibits the optimal rate of convergence.*

**Remark 2.43** *For this model, it is possible to establish a 'local' convergence analysis, i.e., an analysis predicting the convergence of the solution in smooth parts of the flow, away from boundary and discontinuity layers. We refer to Johnson, Nävert and Pitkäranta [165] for a discussion.*

### 2.5.7 Analysis of the Galerkin/Least-Squares Method

Recall the variational equation of the Galerkin/least-squares method:

$$B_{\mathrm{GLS}}(w^h, \phi^h) = L_{\mathrm{GLS}}(w^h) \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.218}$$

where

$$\begin{aligned} B_{\mathrm{GLS}}(w^h, \phi^h) &= B(w^h, \phi^h) + \tau(\mathcal{L}w^h, \mathcal{L}\phi^h)_{\Omega'} && (2.219)\\ L_{\mathrm{GLS}}(w^h) &= L(w^h) + \tau(\mathcal{L}w^h, f)_{\Omega'} && (2.220) \end{aligned}$$

We have already established the consistency of the method, i.e.,

$$B_{\mathrm{GLS}}(w^h, e) = 0 \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.221}$$

#### Stability of the Galerkin/Least-Squares Method

Using the definition of $B_{\mathrm{GLS}}$ and the assumptions of the proof for SUPG (e.g., $h = \mathrm{const.}$), we have $\forall\, w^h \in \mathcal{V}^h$:

$$B_{\mathrm{GLS}}(w^h, w^h) = \kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} \geq 0 \qquad \forall\, \tau \geq 0 \tag{2.222}$$

The objective is to prove that $\kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} \geq c_s\kappa(1 + \alpha\tilde{\xi})||w^h_{,x}||^2$, where $c_s$ is a nondimensional stability constant, analogous to the stability condition for SUPG.

Employing Inequality 3, we have

$$\begin{aligned} \kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} &= \kappa||w^h_{,x}||^2 + \tau(u^2||w^h_{,x}||^2 + \kappa^2||w^h_{,xx}||^2_{\Omega'} \\ &\qquad -2u\kappa(w^h_{,x}, w^h_{,xx})_{\Omega'}) \\ &\geq \kappa||w^h_{,x}||^2 + \tau u^2||w^h_{,x}||^2 + \tau\kappa^2||w^h_{,xx}||^2_{\Omega'} \\ &\qquad -\tau u\kappa\left(\frac{1}{\epsilon}||w^h_{,x}||^2 + \epsilon||w^h_{,xx}||^2_{\Omega'}\right) \\ &= \left(\kappa + \tau u^2 - \frac{u\kappa\tau}{\epsilon}\right)||w^h_{,x}||^2 + \tau\kappa(\kappa - u\epsilon)||w^h_{,xx}||^2_{\Omega'} \end{aligned} \tag{2.223}$$

If we choose $\epsilon = \kappa/u$ to cancel the second term the first term becomes $\kappa||w^h_{,x}||^2$. This trivializes (2.223). We need a better choice. Let $\epsilon = \delta\kappa/u$, where $\delta > 1$. With this choice (2.223) becomes

$$\kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} \geq \left(\kappa + (1-\frac{1}{\delta})\tau u^2\right)||w^h_{,x}||^2 - (\delta-1)\tau\kappa^2||w^h_{,xx}||^2_{\Omega'} \qquad (2.224)$$

Employing the inverse estimate $||w^h_{,xx}||_{\Omega'} \leq c_I h^{-1}||w^h_{,x}||$ in (2.224) leads to

$$\kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} \geq \left(\kappa + (1-\frac{1}{\delta})\tau u^2 - (\delta-1)\tau\kappa^2 c_I^2 h^{-2}\right)||w^h_{,x}||^2 \qquad (2.225)$$

But

$$\begin{aligned} \tau\kappa^2 c_I^2 h^{-2} &= \frac{h}{2|u|}\tilde{\xi}\frac{\kappa^2 c_I^2}{h^2} \\ &= \frac{\kappa}{4}\left(\frac{2\kappa}{|u|h}\right)\tilde{\xi}c_I^2 \\ &= \kappa\left(\frac{\tilde{\xi}c_I^2}{4\alpha}\right) \end{aligned} \qquad (2.226)$$

We assume $\tilde{\xi}/\alpha \leq 4/{c_I}^2$, or equivalently, $\frac{\tilde{\xi}{c_I}^2}{4\alpha} \leq 1$. (This restriction on $\tilde{\xi}$ was required in the analysis of SUPG.) Thus (2.225) becomes

$$\begin{aligned} \kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} &\geq \left(\kappa + (1-\frac{1}{\delta})\tau u^2 - (\delta-1)\kappa\right)||w^h_{,x}||^2 \\ &\geq \left(\kappa(2-\delta) + (1-\frac{1}{\delta})\tau u^2\right)||w^h_{,x}||^2 \end{aligned} \qquad (2.227)$$

Let us choose $\delta$ such that $2-\delta = 1-\frac{1}{\delta}$. We find $\delta = (1+\sqrt{5})/2 \approx 1.618$ (the Golden Ratio!). Thus,

$$\kappa||w^h_{,x}||^2 + \tau||\mathcal{L}w^h||^2_{\Omega'} \geq (2-\delta)\kappa(1+\alpha\tilde{\xi})||w^h_{,x}||^2 \qquad (2.228)$$

The constant $c_s = (2-\delta) \approx 0.3819$.

**Convergence of the Galerkin/Least-Squares Method**

The steps are similar to the previous cases:

$$\begin{aligned}
\kappa||e^h_{,x}||^2 + \tau||\mathcal{L}e^h||^2_{\Omega'} &= B_{\mathrm{GLS}}(e^h, e^h) \\
&= B_{\mathrm{GLS}}(e^h, e-\eta) \\
&= B_{\mathrm{GLS}}(e^h, e) - B_{\mathrm{GLS}}(e^h, \eta) \\
&= |B_{\mathrm{GLS}}(e^h, \eta)| \\
&\leq |B(e^h, \eta)| + \tau|(\mathcal{L}e^h, \mathcal{L}\eta)_{\Omega'}| \\
&= |-(ue^h_{,x}, \eta) + \kappa(e^h_{,x}, \eta_{,x}) + \kappa(e^h_{,xx}, \eta)_{\Omega'} - \kappa(e^h_{,xx}, \eta)_{\Omega'}| \\
&\qquad +\tau|(\mathcal{L}e^h, \mathcal{L}\eta)_{\Omega'}| \\
&\leq |(-ue^h_{,x} + \kappa e^h_{,xx}, \eta)_{\Omega'}| + |\kappa(e^h_{,x}, \eta_{,x})| + \tau|(\mathcal{L}e^h, \mathcal{L}\eta)_{\Omega'}| \\
&\qquad +|(\kappa e^h_{,xx}, \eta)_{\Omega'}| \qquad (2.229)
\end{aligned}$$

Let us estimate these four terms:

$$\begin{aligned}
|(-ue^h_{,x} + \kappa e^h_{,xx}, \eta)_{\Omega'}| = |(\mathcal{L}e^h, \eta)_{\Omega'}| &\leq \frac{1}{2}\left(\frac{1}{\epsilon_1}||\mathcal{L}e^h||^2_{\Omega'} + \epsilon_1||\eta||^2\right) \\
|(\kappa e^h_{,xx}, \eta)_{\Omega'}| &\leq \frac{\kappa}{2}\left(\frac{1}{\epsilon_2}||e^h_{,xx}||^2_{\Omega'} + \epsilon_2||\eta||^2\right) \\
|\kappa(e^h_{,x}, \eta_{,x})| &\leq \frac{\kappa}{2}\left(\frac{1}{\epsilon_3}||e^h_{,x}||^2 + \epsilon_3||\eta_{,x}||^2\right) \\
\tau|(\mathcal{L}e^h, \mathcal{L}\eta)_{\Omega'}| &\leq \frac{\tau}{2}\left(\frac{1}{\epsilon_4}||\mathcal{L}e^h||^2_{\Omega'} + \epsilon_4||\mathcal{L}\eta||^2_{\Omega'}\right) \qquad (2.230)
\end{aligned}$$

We want to hide the first terms on the right-hand sides in the above four expressions. For that purpose, we choose

$$\epsilon_1 = \frac{2}{\tau} \quad , \quad \epsilon_3 = 2 \quad , \quad \epsilon_4 = 2 \qquad (2.231)$$

In order to determine a suitable expression for $\epsilon_2$, we need to use an inverse estimate:

$$\frac{\kappa}{2}\frac{1}{\epsilon_2}||e^h_{,xx}||^2_{\Omega'} \leq \frac{\kappa}{2}\frac{1}{\epsilon_2}c_I^2 h^{-2}||e^h_{,x}||^2 \qquad (2.232)$$

We want to choose $\epsilon_2$ such that $\frac{\kappa}{2}\frac{1}{\epsilon_2}c_I^2h^{-2} = \frac{\kappa}{4}$. This requires

$$\epsilon_2 = \frac{2c_I^2}{h^2} \tag{2.233}$$

Combining (2.229) through (2.233) results in

$$\begin{aligned} \frac{1}{2}\left(\kappa||e^h_{,x}||^2 + \tau||\mathcal{L}e^h||^2_{\Omega'}\right) &\leq \frac{1}{2}\epsilon_1||\eta||^2 + \frac{\kappa}{2}\epsilon_2||\eta||^2 + \frac{\kappa}{2}\epsilon_3||\eta_{,x}||^2 + \frac{\tau}{2}\epsilon_4||\mathcal{L}\eta||^2_{\Omega'} \\ &= \left(\frac{1}{\tau} + \frac{\kappa c_I^2}{h^2}\right)||\eta||^2 + \kappa||\eta_{,x}||^2 + \tau||\mathcal{L}\eta||^2_{\Omega'} \end{aligned} \tag{2.234}$$

We will now pause in the derivation, and examine terms in the right-hand side of (2.234). Every term seems under control, but is the coefficient $1/\tau$ of any concern? No, since $\tau = \frac{h}{2|u|}\tilde{\xi}$ and $\tilde{\xi}$ has the behavior shown in Figure 2.4. Therefore,

$$\tau = \begin{cases} O\left(\frac{h^2}{\kappa}\right) & \text{when } \alpha \text{ is small} \\ O\left(\frac{h}{|u|}\right) & \text{when } \alpha \text{ is large} \end{cases} \tag{2.235}$$

and there is no problem ("*so, don't worry, be happy*!"). In fact, it would even work if $\tilde{\xi}$ was of the form shown in Figure 2.29. For this form of $\tilde{\xi}$, the method reduces to Galerkin in the diffusion-dominated regime, which is effective (cf. (2.153) and discussion thereafter).

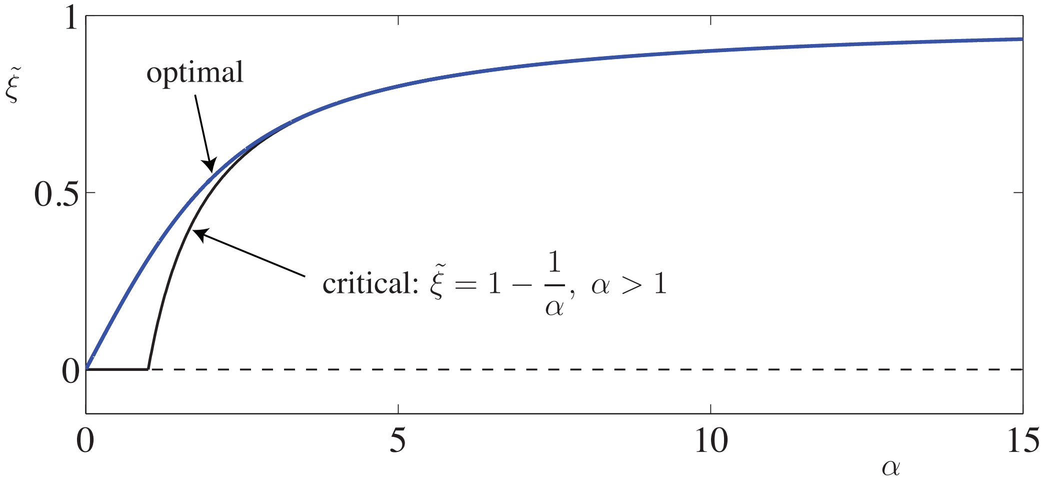


Figure 2.29: The so-called "critical" form of $\tilde{\xi}$ (see Christie *et al*. [48]).

We need to estimate the term $\tau||\mathcal{L}\eta||^2_{\Omega'}$:

$$\begin{aligned}
\tau||\mathcal{L}\eta||^2_{\Omega'} &= \tau\left(||u\eta_{,x}||^2 - 2(u\eta_{,x},\kappa\eta_{,xx})_{\Omega'} + ||\kappa\eta_{,xx}||^2_{\Omega'}\right) \\
&\leq \tau\left(||u\eta_{,x}||^2 + 2|(u\eta_{,x},\kappa\eta_{,xx})_{\Omega'}| + ||\kappa\eta_{,xx}||^2_{\Omega'}\right) \\
&\leq \tau\left(||u\eta_{,x}||^2 + 2\frac{1}{2}\left(||u\eta_{,x}||^2 + ||\kappa\eta_{,xx}||^2_{\Omega'}\right) + ||\kappa\eta_{,xx}||^2_{\Omega'}\right) \\
&= 2\tau\left(||u\eta_{,x}||^2 + ||\kappa\eta_{,xx}||^2_{\Omega'}\right) \\
&= \frac{h}{|u|}\tilde{\xi}\left(u^2||\eta_{,x}||^2 + \kappa^2||\eta_{,xx}||^2_{\Omega'}\right) \\
&= \tilde{\xi}\left(h|u|\,||\eta_{,x}||^2 + \frac{h\kappa^2}{|u|}||\eta_{,xx}||^2_{\Omega'}\right) \qquad (2.236)
\end{aligned}$$

We now return to the convergence proof. The stability result written for $w^h = e^h$ reads

$$\frac{1}{2}\left(c_s\kappa(1+\alpha\tilde{\xi})||e^h_{,x}||^2\right) \leq \frac{1}{2}\left(\kappa||e^h_{,x}||^2 + \tau||\mathcal{L}e^h||^2_{\Omega'}\right) \qquad (2.237)$$

Dividing through by $\kappa$, and combining with (2.234) and (2.236), yields

$$\begin{aligned}
\frac{1}{2}c_s(1+\alpha\tilde{\xi})||e^h_{,x}||^2 &\leq \left(\frac{2|u|}{\kappa h\tilde{\xi}} + \frac{c_I^2}{h^2}\right)||\eta||^2 + ||\eta_{,x}||^2 \\
&+ \tilde{\xi}\left(\frac{h|u|}{\kappa}||\eta_{,x}||^2 + \frac{h\kappa}{|u|}||\eta_{,xx}||^2_{\Omega'}\right) \qquad (2.238)
\end{aligned}$$

The coefficient $\frac{2|u|}{\kappa h\tilde{\xi}}$ needs to be estimated. The $\tilde{\xi}$-term in the denominator attains its smallest values in the diffusion-dominated limit. The form of $\tilde{\xi}$ provides the estimate $\tilde{\xi} \geq \beta\alpha$, where $\beta$ is a positive constant, for sufficiently small $\alpha$ (see Fig. 2.28). Thus

$$\begin{aligned}
\frac{2|u|}{\kappa h\tilde{\xi}} &\leq \frac{2|u|}{\kappa h\beta\alpha} \\
&= \frac{2|u|}{\kappa h\beta\frac{|u|h}{2\kappa}} \\
&= \frac{4}{\beta h^2} \qquad (2.239)
\end{aligned}$$

Substituting (2.239) into (2.238) yields:

$$\begin{aligned}\frac{1}{2}c_s(1+\alpha\tilde{\xi})||e^h_{,x}||^2 &\leq \left(\frac{4}{\beta}h^{-2}+c_I^2h^{-2}\right)||\eta||^2+||\eta_{,x}||^2+2\alpha\tilde{\xi}||\eta_{,x}||^2+\frac{h^2}{2\alpha}\tilde{\xi}||\eta_{,xx}||^2_{\Omega'}\\ &= \left(\frac{4}{\beta}+c_I^2\right)h^{-2}||\eta||^2+(1+2\alpha\tilde{\xi})||\eta_{,x}||^2+\frac{h^2}{2\alpha}\tilde{\xi}||\eta_{,xx}||^2_{\Omega'} \qquad (2.240)\end{aligned}$$

We now divide (2.240) through by $\frac{1}{2}c_s(1+\alpha\tilde{\xi})$ and examine the coefficients on the right-hand side. It may be verified from the properties of $\tilde{\xi}$ that there exist *constants* $c_1, c_2, c_3$, independent of $\alpha$ and $h$, such that

$$\begin{aligned}\frac{8+2\beta c_I^2}{\beta c_s(1+\alpha\tilde{\xi})} &\leq c_1\\ \frac{2(1+2\alpha\tilde{\xi})}{c_s\left(1+\alpha\tilde{\xi}\right)} &\leq c_2\\ \frac{2}{c_s\left(1+\alpha\tilde{\xi}\right)}\frac{\tilde{\xi}}{2\alpha} &\leq c_3 \qquad (2.241)\end{aligned}$$

Note that (2.241) requires that $\tilde{\xi}$ has a finite slope as $\alpha \to 0$. In other words, $\tilde{\xi}/\alpha \leq \text{const}$. This condition was also encountered in the analysis of SUPG. Let $\bar{c} = \max\{c_1, c_2, c_3\}$; then,

$$||e^h_{,x}||^2 \leq \bar{c}\Big(h^{-2}||\eta||^2+||\eta_{,x}||^2+h^2||\eta_{,xx}||^2_{\Omega'}\Big) \qquad (2.242)$$

Employing standard interpolation estimates, yields

$$\begin{aligned}L^{-2}|e^h|_1^2 \equiv ||e^h_{,x}||^2 &= \bar{c}\left(h^{-2}||\eta||^2+L^{-2}|\eta|_1^2+h^2L^{-4}|\eta|^2_{H^2(\Omega')}\right)\\ &\leq 3\,\bar{c}\,c^2\left(\frac{h}{L}\right)^{2k}L^{-2}||\phi||^2_{k+1} \qquad (2.243)\end{aligned}$$

Thus

$$\boxed{|e^h|_1 \leq \hat{c}\left(\frac{h}{L}\right)^k||\phi||_{k+1}} \qquad (2.244)$$

where $\hat{c} = \sqrt{3\bar{c}}c$ is independent of $\alpha$ and $h$. Finally,

$$\begin{aligned}|e|_1 &= |e^h+\eta|_1\\ &\leq |e^h|_1+|\eta|_1\\ &\leq |e^h|_1+c\left(\frac{h}{L}\right)^k||\phi||_{k+1} \qquad (2.245)\end{aligned}$$

Using (2.244) in (2.245) yields

$$\boxed{|e|_1 \leq c\left(\frac{h}{L}\right)^k||\phi||_{k+1}} \qquad (2.246)$$

**Remark 2.44** *Eq. (2.246) has the same form as the error estimate for SUPG.*

**Remark 2.45** *In conclusion, we may say that the hypotheses were essentially the same as for SUPG.*

We will now summarize the ingredients employed in the analysis of Galerkin/least-squares:

- Stability of the method:

$$c_s\kappa(1+\alpha\tilde{\xi})||w^h_{,x}||^2 \leq B_{\text{GLS}}(w^h,w^h) \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.247}$$

- Inverse Estimate:

$$||w^h_{,xx}||_{\Omega'} \leq c_I h^{-1}||w^h_{,x}|| \tag{2.248}$$

- Consistency of the method:

$$B_{\text{GLS}}(w^h,e) = 0 \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.249}$$

- Interpolation estimate:

$$\left(\frac{h}{L}\right)^{-2}||\eta||^2 + |\eta|_1^2 + \left(\frac{h}{L}\right)^2 |\eta|^2_{H^2(\Omega')} \leq 3c^2\left(\frac{h}{L}\right)^{2k}||\phi||^2_{k+1} \tag{2.250}$$

The restrictions imposed on $\tau = \dfrac{h}{2|u|}\tilde{\xi}$ by the analysis were:

1. $\tilde{\xi} \leq \dfrac{4\alpha}{c_I^2}$. As in SUPG, this condition appeared in the stability proof.
2. $\dfrac{\tilde{\xi}}{\alpha} \leq \text{const}$. As in SUPG, this condition was used in the convergence proof. (It is implied by the previous one if $c_I \neq 0$.)
3. We used the fact that for small $\alpha, \tilde{\xi} \geq \beta\alpha$, where $\beta$ is a constant. This restriction was a convenience rather than a necessity as we could have proceeded in an alternative way. See Figure 2.29 and remarks following (2.235).

**Exercise 2.11** *Consider the following method:*

$$B_{\text{MS}}(w^h,\phi^h) = L_{\text{MS}}(w^h) \qquad \forall\, w^h \in \mathcal{V}^h \tag{2.251}$$

*where*

$$B_{\text{MS}}(w^h,\phi^h) = B(w^h,\phi^h) + \tau(-\mathcal{L}^*w^h,\mathcal{L}\phi^h)_{\Omega'} \tag{2.252}$$

$$L_{\text{MS}}(w^h) = L(w^h) + \tau(-\mathcal{L}^*w^h,f)_{\Omega'} \tag{2.253}$$

*where $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$, that is,*

$$\mathcal{L}^* w^h = -u\, w^h_{,x} - \kappa\, w^h_{,xx} \tag{2.254}$$

*The subscript $MS$ stands for "multiscale." We will have a lot to say about methods with this structure in later chapters. Suffice it to say for now that it provides a glimpse of a "variational multiscale method." Carefully study the convergence proofs and error estimates for SUPG and Galerkin/Least-Squares in Sections 2.5.6 and 2.5.7, respectively, and, utilizing similar hypotheses, perform an analysis for (2.251)–(2.253). Summarize all ingredients required in a box, such as was done for SUPG and Galerkin/Least-Squares.*