Chapter 5

Finite Element Methods (FEM)

5.1 Generalities

Remark 5.1 *Finite element methods.* Finite element methods were one of the main topics of Numerical Mathematics 3. The knowledge of the lecture notes of Numerical Mathematics 3 is assumed. Only a few issues, which are important for this course on finite element methods for convection-dominated problems, will be reminded here.

Let $\{\mathcal{T}^h\}$ be a family of regular triangulations consisting of mesh cells $\{K\}$. The triangulations are assumed to be quasi-uniform. The diameter of a mesh cell K is denoted by h_K and it is $h = \max_K\{h_K\}$. Parametric finite element spaces will be considered with affine maps between a reference cell $\hat{K}$ and all physical cells K. □

Theorem 5.2 Local interpolation error estimate. *Let $I_K : C^s(K) \to P(K)$ be an interpolation operator as defined in Numerical Mathematics 3, where $P(K)$ is a polynomial space defined on K. Let $p \in [1,\infty]$ and $(m+1-s)p > 1$. Then there is a constant c, which is independent of $v \in W^{m+1,p}(K)$, such that*

$$\left\|(v - I_K v)^{(k)}\right\|_{L^p(K)} \leq c h_K^{m+1-k} \left\|v^{(m+1)}\right\|_{L^p(K)}, \quad 0 \leq k \leq m+1, \tag{5.1}$$

for all $v \in W^{m+1,p}(K)$.

Proof: See lecture notes of Numerical Mathematics 3. ■

Theorem 5.3 Inverse estimate. *Let $0 \leq k \leq l$ be natural numbers and let $p, q \in [1,\infty]$. Then there is a constant C_{inv}, which depends only on $k, l, p, q, \hat{K}, \hat{P}(\hat{K})$ such that*

$$\left\|D^l v^h\right\|_{L^q(K)} \leq C_{\text{inv}} h_K^{(k-l)-d(p^{-1}-q^{-1})} \left\|D^k v^h\right\|_{L^p(K)} \quad \forall\, v^h \in P(K). \tag{5.2}$$

Proof: See lecture notes of Numerical Mathematics 3. ■

5.2 The Galerkin Method

Remark 5.4 *The Galerkin method.* The standard finite element method, which just replaces in the variational formulation (4.2) the space V by $V^h \subset V$, is called Galerkin method: Find $u^h \in V^h$, such that for all $v^h \in V^h$

$$\varepsilon\left(\nabla u^h, \nabla v^h\right) + \left(\mathbf{b}\cdot\nabla u^h + c u^h, v^h\right) = \left(f, v^h\right). \tag{5.3}$$

□

Figure 5.1: Boris Grigorievich Galerkin (1871 – 1945).

Theorem 5.5 Lemma of Cea. *Let $V^h \subset V$ and assume the conditions of the Theorem of Lax–Milgram, Theorem 4.10. Then there is a unique solution of the problem to find $u^h \in V^h$ such that*

$$a\left(u^h, v^h\right) = f\left(v^h\right) \quad \forall\, v^h \in V^h \tag{5.4}$$

and it holds the error estimate

$$\left\|u - u^h\right\|_V \leq \frac{M}{m} \inf_{v^h \in V^h} \left\|u - v^h\right\|_V, \tag{5.5}$$

where u is the unique solution of the continuous problem (4.5) *and the constants are defined in Definition 4.8.*

Proof: The existence of a unique solution of the discrete problem follows directly from the Theorem of Lax–Milgram, since the subspace of a Hilbert space is also a Hilbert space and the properties of the bilinear form carry over from V to V^h.

Computing the difference of the continuous equation (4.5) and the discrete equation (5.4) yields

$$a\left(u - u^h, v^h\right) = 0 \quad \forall\, v^h \in V^h.$$

With

$$m \left\|v\right\|_V^2 \leq a(v,v) \quad \text{and} \quad |a(u,v)| \leq M \left\|u\right\|_V \left\|v\right\|_V$$

if follows for all $v^h \in V^h$ that

$$\begin{aligned} \left\|u - u^h\right\|_V^2 &\leq \frac{1}{m} a\left(u - u^h, u - u^h\right) = \frac{1}{m} a\left(u - u^h, u - v^h\right) \\ &\leq \frac{M}{m} \left\|u - u^h\right\|_V \left\|u - v^h\right\|_V. \end{aligned}$$

This inequality is equivalent to the statement of the theorem. ■

Remark 5.6 *Application to the Galerkin finite element method.* The properties of the bilinear form from problems (4.2) and (5.3) were studied in Example 4.9. It was shown that with appropriate regularity assumptions and under condition (4.4), the bilinear form is bounded with a constant M of order $\max\{\left\|\mathbf{b}\right\|_\infty, \left\|c\right\|_\infty\}$ and

it is coercive with $m = \varepsilon$. In this case, one can apply the Lemma of Cea and one obtains the error estimate

$$\left\|u - u^h\right\|_V \leq C \frac{\max\{\|\mathbf{b}\|_\infty, \|c\|_\infty\}}{\varepsilon} \inf_{v^h \in V^h} \left\|u - v^h\right\|_V, \quad C \in \mathbb{R}.$$

In the singularly perturbed case $\varepsilon \ll \|\mathbf{b}\|_\infty$, the first factor of this estimate becomes very large.

Thus, from this error estimate one cannot expect that the Galerkin finite element solution is accurate unless the second factor, which is the best approximation error, is very small. On uniformly refined grids, the best approximation error becomes very small only if the dimension of V^h becomes very large. □

Example 5.7 *Galerkin method for P_1 finite elements in one dimension.* Let $\Omega = (0,1)$ and consider the case that the coefficients of the problem are constant, i.e., $b(x) = b$, $c(x) = c$ and $f(x) = f$. For the P_1 finite element on the reference cell $\hat{K} = [-1,1]$, the basis functions and their derivatives are given by

$$\hat{\phi}_1(\hat{x}) = \frac{1}{2}(-\hat{x}+1), \quad \hat{\phi}_1'(\hat{x}) = -\frac{1}{2}, \qquad \hat{\phi}_2(\hat{x}) = \frac{1}{2}(\hat{x}+1), \quad \hat{\phi}_2'(\hat{x}) = \frac{1}{2}.$$

Consider the matrix entry $a_{i,i+1}$, which is computed with the test function $\phi_i(x)$, which is transformed to $\hat{\phi}_1(\hat{x})$, and the ansatz function $\phi_{i+1}(x)$, which is transformed to $\hat{\phi}_2(\hat{x})$. The common support is the mesh cell $[x_i, x_{i+1}]$. Let h denote the length of this cell, then one obtains

$$\begin{aligned} a_{i,i+1} &= \frac{2\varepsilon}{h} \int_{-1}^{1} \frac{1}{2} \cdot \left(-\frac{1}{2}\right) \, d\hat{x} + b \int_{-1}^{1} \frac{1}{2} \cdot \frac{1}{2}(-\hat{x}+1) \, d\hat{x} \\ &\quad + \frac{ch}{2} \int_{-1}^{1} \frac{1}{2}(\hat{x}+1) \frac{1}{2}(-\hat{x}+1) \, d\hat{x} \\ &= -\frac{\varepsilon}{h} + \frac{b}{2} + \frac{ch}{6}. \end{aligned}$$

For the i-th component of the right-hand side, one gets

$$f_i = f \int_0^1 \phi_i(x) \, dx = hf.$$

The other matrix entries can be calculated in a similar way, *exercise*.

If one applies the trapezoidal rule for the approximation of the integrals, one obtains

$$\frac{ch}{2} \int_{-1}^{1} \frac{1}{2}(-\hat{x}+1) \frac{1}{2}(\hat{x}+1) \, d\hat{x} = \frac{ch}{2} 2 \frac{0+0}{2} = 0.$$

Then, it follows that

$$a_{i,i+1} = -\frac{\varepsilon}{h} + \frac{b}{2}.$$

For $c = 0$ (or with trapezoidal rule) these are, apart of a factor h, the same matrix entries as for the central difference scheme, see Remark 3.10. One gets for $c = 0$

$$-h\varepsilon D^+ D^- u_i + h b_i D^0 u_i = h f_i \quad \Longleftrightarrow \quad -\varepsilon D^+ D^- u_i + b_i D^0 u_i = f_i.$$

Hence, one obtains in this special case the same results with the Galerkin finite element method and the central finite difference scheme. For singularly perturbed problems, these results are very bad.

The equivalence of the central finite difference method and the Galerkin finite element method generally does not hold if the coefficients are not constant. In higher dimensions, there are differences even for constant coefficients. But still, these two methods give very similar results. At any rate, the Galerkin finite element method is not useful in the singularly perturbed case also in more general situations. □

5.3 Stabilized Finite Element Methods

Remark 5.8 *On the $H^1(\Omega)$ norm for the numerical analysis of singularly perturbed problems.* Consider the problem: Find $u \in V = H_0^1(\Omega)$ such that

$$a(u,v) = f(v) \quad \forall \ v \in V \tag{5.6}$$

with

$$\begin{aligned} a(u,v) &:= \int_\Omega \Big(\varepsilon \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) + \mathbf{b}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) v(\mathbf{x}) + c(\mathbf{x}) u(\mathbf{x}) v(\mathbf{x})\Big)\ d\mathbf{x}, \\ f(v) &:= \int_\Omega f(\mathbf{x}) v(\mathbf{x})\ d\mathbf{x}. \end{aligned}$$

Let the condition

$$-\frac{1}{2}\nabla \cdot \mathbf{b}(\mathbf{x}) + c(\mathbf{x}) \geq \mu_0 > 0 \quad \text{almost everywhere in } \Omega$$

be satisfied, which is stronger than condition (4.4). Then, an analogous calculation as in Example 4.9 shows that $a(\cdot,\cdot)$ is uniformly coercive with respect to the following norm, which depends on ε,

$$\|v\|_\varepsilon^2 := \varepsilon \, |v|_{1,2}^2 + \mu_0 \, \|v\|_0^2 = \varepsilon \, \|\nabla v\|_{L^2(\Omega)}^2 + \mu_0 \, \|v\|_{L^2(\Omega)}^2 ,$$

i.e., there is a constant m which does not depend on ε such that

$$a(v,v) \geq m \, \|v\|_\varepsilon^2 \quad \forall \, v \in V.$$

Applying integration by parts, *exercise*, shows that there is a constant M, which is also independent on ε, such that

$$|a(v,w)| \leq M \, \|v\|_\varepsilon \, \|w\|_{H^1(\Omega)} \quad \forall \, (v,w) \in V \times V.$$

However, there is no constant $\tilde{M}$, which is independent of ε, with

$$|a(v,w)| \leq \tilde{M} \, \|v\|_\varepsilon \, \|w\|_\varepsilon \quad \forall \, (v,w) \in V \times V.$$

Applying the estimates with constants that are independent of ε, one obtains in a similar way as in the proof of the Lemma of Cea that

$$\left\|u - u^h\right\|_\varepsilon \leq C \inf_{v^h \in V^h} \left\|u - v^h\right\|_{H^1(\Omega)} ,$$

with C independent of ε. If V^h is a standard finite element space (piecewise polynomial), then one can show that in layers it is

$$\inf_{v^h \in V^h} \left\|u - v^h\right\|_{H^1(\Omega)} \to \infty \quad \text{for} \quad \varepsilon \to 0,$$

for fixed h. Consequently, there is no uniform convergence $\left\|u - u^h\right\|_\varepsilon \to 0$ for $h \to 0$. The norm $\|\cdot\|_{H^1(\Omega)}$ is not suited for the investigation of numerical methods for convection-dominated problems. It turns out that the use of appropriate norms is important for the numerical analysis of finite element methods for convection-dominated problems. □

5.3.1 Petrov–Galerkin Methods and Upwind Methods

Remark 5.9 *Petrov–Galerkin method.* A finite element method, whose ansatz and test space are different, is called Petrov–Galerkin method. Let S^h be the ansatz space and T^h be the test space with $\dim(S^h) = \dim(T^h)$, then the Petrov–Galerkin method reads as follows: Find $u^h \in S^h$ such that

$$a\left(u^h, v^h\right) = f\left(v^h\right) \quad \forall\, v^h \in T^h.$$

□

Example 5.10 *Petrov–Galerkin method and upwind method.* Consider

$$-\varepsilon u''(x) + bu'(x) = 0$$

with $b \in \mathbb{R} \setminus \{0\}$ and homogeneous Dirichlet boundary conditions. Use as functions in the ansatz space continuous piecewise linear functions

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h & \text{for } x \in [x_{i-1}, x_i], \\ (x_{i+1} - x)/h & \text{for } x \in [x_i, x_{i+1}], \\ 0 & \text{else,} \end{cases} \quad , \quad i = 1, \ldots, N-1.$$

Define the bubble functions

$$\sigma_{i-1/2}(x) = \begin{cases} 4(x - x_{i-1})(x_i - x)/h^2 & \text{for } x \in [x_{i-1}, x_i], \\ 0 & \text{else.} \end{cases}$$

The test functions will be defined as piecewise quadratic functions

$$\psi_i(x) = \phi_i(x) + \frac{3}{2}\kappa\left(\sigma_{i-1/2}(x) - \sigma_{i+1/2}(x)\right), \quad i = 1, \ldots, N-1,$$

where κ is an upwind parameter which has to be chosen. A direct calculation shows, *exercise*, that one obtains the following method

$$-\varepsilon D^+ D^- u_i + b\left[\left(\frac{1}{2} - \kappa\right) D^+ u_i + \left(\frac{1}{2} + \kappa\right) D^- u_i\right] = 0.$$

For the choice $\kappa = \operatorname{sgn}(b)/2$, one obtains the simple upwind finite difference scheme, see Definition 3.33.

A test function $\psi_i(x)$, defined in the nodes $\{0, 0.5, 1\}$, for $\kappa = 1/2$ is presented in Figure 5.2.

□

Remark 5.11 *Fitted upwind schemes.* It is also possible to define fitted upwind methods with Petrov–Galerkin methods, even for non-constant coefficients. However, one does not obtain better results as for the Iljin–Allen–Southwell method, Theorem 3.57, i.e., one obtains not more than linear convergence. □

5.3.2 The Streamline-Upwind Petrov–Galerkin (SUPG) Method

Remark 5.12 *Goal.* The goal consists in the construction of a method that is more stable than the Galerkin finite element method and which can be used with finite elements of arbitrary order. The convergence of this method, in an appropriate norm, should be of higher order.

Consider the problem (4.1) and assume at the moment that condition (4.4) is satisfied. Later, an even stronger assumption will be made. □

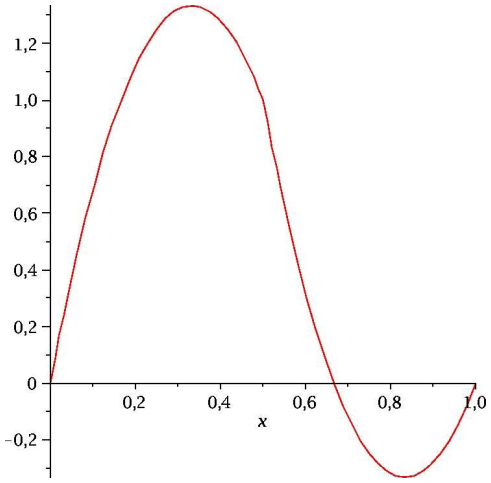

Figure 5.2: Piecewise quadratic test function for $\kappa = 1/2$.

Remark 5.13 *The basic idea.* The basic idea consists in a penalization of large values of the so-called strong residual. Such methods are called residual-based stabilizations.

Given a linear partial differential equation in strong form

$$A_{\text{str}} u_{\text{str}} = f, \quad f \in L^2(\Omega),$$

and its Galerkin finite element discretization: Find $u^h \in V^h$ such that

$$a^h\left(u^h, v^h\right) = \left(f, v^h\right) \quad \forall\, v^h \in V^h. \tag{5.7}$$

For residual-based stabilizations, a modification of A_{str} is needed which is well-defined for finite element functions. This modification should be also a linear operator and it is denoted by $A^h_{\text{str}} : V^h \to L^2(\Omega)$. The (strong) residual is now defined by

$$r^h\left(u^h\right) = A^h_{\text{str}} u^h - f \ \in L^2(\Omega).$$

In general, it holds $r^h\left(u^h\right) \neq 0$, but a good numerical approximation of the solution of the continuous problem should have in some sense a small residual. Now, instead of finding the solution of (5.7), the minimizer of the residual is searched, i.e, the following optimization problem is considered

$$\arg\min_{u^h \in V^h} \left\| r^h\left(u^h\right) \right\|^2_{L^2(\Omega)} = \arg\min_{u^h \in V^h} \left(r^h\left(u^h\right), r^h\left(u^h\right) \right). \tag{5.8}$$

The necessary condition for taking the minimum is the vanishing of the Gâteaux derivative. This derivative is computed by using the linearity of A^h_{str} and the bilinearity of the inner product in $L^2(\Omega)$

$$\begin{aligned} 0 &= \lim_{\varepsilon \to 0} \frac{\left(r^h\left(u^h + \varepsilon v^h\right), r^h\left(u^h + \varepsilon v^h\right)\right) - \left(r^h\left(u^h\right), r^h\left(u^h\right)\right)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{\left(r^h\left(u^h\right) + \varepsilon A^h_{\text{str}} v^h, r^h\left(u^h\right) + \varepsilon A^h_{\text{str}} v^h\right) - \left(r^h\left(u^h\right), r^h\left(u^h\right)\right)}{\varepsilon} \\ &= 2\left(r^h\left(u^h\right), A^h_{\text{str}} v^h\right) \quad \forall\, v^h \in V^h. \end{aligned}$$

It follows that the necessary condition for the solution of (5.8) is

$$\left(r^h\left(u^h\right), A^h_{\text{str}} v^h\right) = 0.$$

A generalization consists in considering the minimization problem

$$\arg\min_{u^h \in V^h} \left\| \delta^{1/2} r^h \left(u^h \right) \right\|^2_{L^2(\Omega)} = \arg\min_{u^h \in V^h} \left(\delta r^h \left(u^h \right), r^h \left(u^h \right) \right). \tag{5.9}$$

with the positive weighting function $\delta(\mathbf{x})$. Analogously to the derivation for the special case, one obtains as necessary condition for the minimum

$$\left(\delta r^h \left(u^h \right), A^h_{\text{str}} v^h \right) = 0. \tag{5.10}$$

The solutions of (5.8) or (5.9) will not be identical to the solution of the Galerkin discretization (5.7). It turns out that the reason for the Galerkin discretization to fail is that the solution possesses structures (scales) that are important but which are not resolved by the used finite element space (grid). For convection-diffusion problems, such structures are layers, particularly at boundaries. The numerical methods should also compute sharp layers. However the sharpness of layers in numerical solutions is restricted by the resolution, which is generally much coarser than the layer width. Hence, even for a numerical solution with sharp layers, the residual in the layer regions are very large. In particular, a numerical solution with sharp layers (with respect to the resolution of the finite element space) will not be the minimizer of (5.8) or (5.9), see Figure 5.3. The minimizers of (5.8) or (5.9) tend to possess strongly smeared layers and these solutions are useless in applications. For this reason, one considers in residual-based stabilizations a combination of the Galerkin discretization (5.7) and the minimization of the residual

$$a^h \left(u^h, v^h \right) + \left(\delta r^h \left(u^h \right), A^h_{\text{str}} v^h \right) = \left(f, v^h \right) \quad \forall\, v^h \in V^h. \tag{5.11}$$

The goal of numerical analysis consists in determining the weighting function δ optimally in an asymptotic sense. □

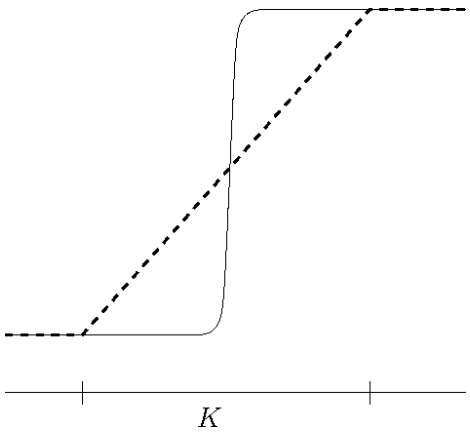

Figure 5.3: Function with sharp layer (solid line) and optimal piecewise linear approximation in a mesh cell K (dashed line). The equation which is fulfilled by the function in K is far from being satisfied by the piecewise linear approximation. Hence, despite the approximation is of the type considered to be optimal, the residual will be large.

Definition 5.14 Streamline-Upwind Petrov–Galerkin FEM, SUPG method, Streamline-Diffusion FEM, SDFEM. The Streamline-Upwind Petrov–Galerkin (SUPG) FEM or Streamline-Diffusion FEM (SDFEM) has the form: Find $u^h \in V^h$, such that

$$a^h \left(u^h, v^h \right) = f^h \left(v^h \right) \quad \forall\, v^h \in V^h \tag{5.12}$$

with

$$\begin{aligned} a^h(v,w) &:= a(v,w) \\ &\quad + \sum_{K\in\mathcal{T}^h} \int_K \delta_K \Big(-\varepsilon\Delta v(\mathbf{x}) + \mathbf{b}(\mathbf{x})\cdot\nabla v(\mathbf{x}) + c(\mathbf{x})v(\mathbf{x})\Big)\Big(\mathbf{b}(\mathbf{x})\cdot\nabla w(\mathbf{x})\Big)\, d\mathbf{x}, \\ f^h(w) &:= (f,w) + \sum_{K\in\mathcal{T}^h} \int_K \delta_K f(\mathbf{x})\Big(\mathbf{b}(\mathbf{x})\cdot\nabla w(\mathbf{x})\Big)\, d\mathbf{x}. \end{aligned} \tag{5.13}$$

Here, $\{\delta_K\}$ are user-chosen weights, which are called stabilization parameters or SUPG parameters. □

Remark 5.15 *Concerning the SUPG method.*

- The method was developed in Hughes and Brooks (1979); Brooks and Hughes (1982).
- The name "SUPG" comes from the fact that the method can be considered as a Petrov–Galerkin method with the test space

$$\operatorname{span}\left\{ w(\mathbf{x}) + \sum_{K\in\mathcal{T}^h} \delta_K \mathbf{b}(\mathbf{x})\cdot\nabla w(\mathbf{x}) \right\}.$$

- The SUPG method introduces artificial diffusion only in streamline direction $\mathbf{b}(\mathbf{x})\cdot\nabla w(x)$. From this property, the name "Streamline Diffusion FEM" originates.
- The operator A^h_{str} is given in the second part of the bilinear form (5.13). The second derivative for finite element functions is defined only piecewise.
- In the stabilization term of the SUPG method, not the strong operator A^h_{str} applied to the test function is used, as in (5.11), but only the first order term contained in this expression. However, for singularly perturbed problems, the first order term is the dominating term of the strong operator applied to the test function.
  It is also possible to define a method with the strong operator applied to the test function, the so-called Galerkin least squared (GLS) method. In general, one gets similar results with the SUPG and the GLS method, but the SUPG method is easier to implement.
- Generally, the SUPG parameter is a general function. However, in practice it is often chosen as a piecewise constant function. The goal of the finite element error analysis consists in proposing a good choice of this parameter.

□

Example 5.16 *SUPG in one dimension for P_1 finite elements.* Consider $\Omega = (0,1)$ and $V^h = P_1$ on an equidistant grid with $h_i = h$, $i = 1,\dots,N$. If all coefficients are constant, $c = 0$, and if one chooses the SUPG parameter also as a constant, then the left-hand side of the SUPG method reduces to

$$\begin{aligned} &\varepsilon\big(\big(u^h\big)',\big(v^h\big)'\big) + \big(b\big(u^h\big)', v^h\big) + \sum_{i=1}^{N} \delta \int_{x_{i-1}}^{x_i} \Big(-\varepsilon\cdot 0 + b\big(u^h\big)'(x)\Big)\Big(b\big(v^h\big)'(x)\Big)\, dx \\ &\qquad = \varepsilon\big(\big(u^h\big)',\big(v^h\big)'\big) + b\big(\big(u^h\big)', v^h\big) + \delta b^2\big(\big(u^h\big)',\big(v^h\big)'\big). \end{aligned}$$

This expression is of the form of a Galerkin finite element method for an equation with left-hand side

$$-\left(\varepsilon + \delta b^2\right) u''(x) + bu'(x).$$

It is known from Example 5.7 that the Galerkin finite element method is equivalent to a central finite difference scheme. The right-hand side of the SUPG method is

$$\begin{aligned}(f, v^h) + \sum_{i=1}^{N} \delta \int_{x_{i-1}}^{x_i} fb \left(v^h\right)'(x)\, dx &= (f, v^h) + \delta fb \underbrace{\sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \left(v^h\right)'(x)\, dx}_{=0} \\ &= (f, v^h) = hf_i.\end{aligned}$$

The sum vanishes, since each test function $\left(v^h\right)'(x)$ can be written as a linear combination of the basis functions $\{\phi_i(x)\}$ of P_1 and the integral of the derivative of each basis function vanishes. Alternatively, one can apply integration by parts to check this fact.

Altogether, the SUPG method with the conditions stated above is equivalent to the fitted finite difference scheme (3.10)

$$-\varepsilon \left(1 + \delta \frac{b^2}{\varepsilon}\right) D^+ D^- u_i + bD^0 u_i, = f_i,$$

i.e., $\sigma(q) = 1 + \delta b^2/\varepsilon$, $q = bh/(2\varepsilon)$. Choosing the SUPG parameter by

$$\delta(q) = \frac{h}{2b}\left(\coth(q) - \frac{1}{q}\right),$$

then it is

$$\sigma(q) = 1 + \frac{hb^2}{2b\varepsilon}\left(\coth(q) - \frac{1}{q}\right) = 1 + q\left(\coth(q) - \frac{1}{q}\right) = q\coth(q).$$

One obtains the Iljin–Allen–Southwell scheme. With $\delta = h/(2b)$, one gets the simple upwind scheme.

These simple connections do not hold in higher dimensions. □

Definition 5.17 Consistent finite element method. Let $u(\mathbf{x})$ be a sufficiently smooth solution of: Find $u \in V$ such that

$$a(u, v) = f(v) \quad \forall\, v \in V,$$

where $a(\cdot, \cdot)$ is an appropriate bilinear form and $f(\cdot)$ an appropriate functional. A finite element method related to this problem: Find $u^h \in V^h$ such that

$$a^h\left(u^h, v^h\right) = f^h\left(v^h\right) \quad \forall\, v^h \in V^h$$

is called consistent, if

$$a^h\left(u, v^h\right) = f^h\left(v^h\right) \quad \forall\, v^h \in V^h. \tag{5.14}$$

□

Remark 5.18 *Consistency.* Note that consistency of a finite element method is not the same as consistency of a finite difference method, see Definition 3.6. For finite element methods, consistency means that a sufficiently smooth solution satisfies also the discrete equation. □

Lemma 5.19 Galerkin orthogonality. *A consistent finite element method has the property of the Galerkin orthogonality*

$$a^h\left(u-u^h,v^h\right)=0 \quad \forall\, v^h\in V^h. \tag{5.15}$$

The error is "orthogonal" to the finite element space.

Proof: The statement of the lemma follows immediately by subtracting (5.12) and (5.14). ■

Lemma 5.20 Consistency of the SUPG method. *The SUPG method* (5.12) – (5.13) *is consistent.*

Proof: A sufficiently smooth solution $u(\mathbf{x})$ of (4.2) satisfies the strong form of the equation even pointwise. Hence, the residual is pointwise zero. Inserting this solution into the SUPG formulation (5.12) – (5.13) results in a vanishing of the stabilization term. It remains

$$a\left(u,v^h\right)=f\left(v^h\right) \quad \forall\, v^h\in V^h,$$

which is satisfied by any weak solution since $V^h\subset V$. That means, the smooth solution satisfies also the discrete equation. ■

Definition 5.21 SUGP norm. Let

$$-\frac{1}{2}\nabla\cdot\mathbf{b}(\mathbf{x})+c(\mathbf{x})\geq\omega>0. \tag{5.16}$$

In V^h, the SUPG norm is defined by

$$\left|\left|\left|v^h\right|\right|\right|_{\mathrm{SUPG}}:=\left(\varepsilon\left|v^h\right|_1^2+\omega\left\|v^h\right\|_0^2+\sum_{K\in\mathcal{T}^h}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla v^h\right)\right\|_{0,K}^2\right)^{1/2},$$

where $\|\cdot\|_{0,K}$ denotes the norm in $L^2(K)$. □

Theorem 5.22 Coercivity of the SUPG bilinear form. *Assume* $\mathbf{b}\in W^{1,\infty}(\Omega)$, $c\in L^\infty(\Omega)$, (5.16), *and let*

$$0<\delta_K\leq\frac{1}{2}\min\left\{\frac{h_K^2}{\varepsilon C_{\mathrm{inv}}^2},\frac{\omega}{\|c\|_{L^\infty(K)}^2}\right\}, \tag{5.17}$$

where C_{inv} *is the constant in the inverse estimate* (5.2) *Then, the SUPG bilinear form is coercive with respect to the SUPG norm, i.e., it is*

$$a^h\left(v^h,v^h\right)\geq\frac{1}{2}\left|\left|\left|v^h\right|\right|\right|_{\mathrm{SUPG}}^2 \quad \forall\, v^h\in V^h.$$

Proof: Integration by parts gives, see Example 4.9,

$$\left(\mathbf{b}\cdot\nabla v^h+cv^h,v^h\right)=\left(\left(-\frac{\nabla\cdot\mathbf{b}}{2}+c\right)v^h,v^h\right) \quad \forall\, v^h\in V^h.$$

With the definition of ω, one obtains

$$\begin{aligned}
a^h\left(v^h,v^h\right) &= \varepsilon\left|v^h\right|_1^2+\int_\Omega\underbrace{\left(c(\mathbf{x})-\frac{\nabla\cdot\mathbf{b}(\mathbf{x})}{2}\right)}_{\geq\omega>0}\left(v^h\right)^2(\mathbf{x})\,d\mathbf{x}+\sum_{K\in\mathcal{T}^h}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla v^h\right)\right\|_{0,K}^2\\
&\quad+\sum_{K\in\mathcal{T}^h}\int_K\delta_K\left(-\varepsilon\Delta v^h(\mathbf{x})+c(\mathbf{x})v^h(\mathbf{x})\right)\left(\mathbf{b}(\mathbf{x})\cdot\nabla v^h(\mathbf{x})\right)\,d\mathbf{x}\\
&\geq \left|\left|\left|v^h\right|\right|\right|_{\mathrm{SUPG}}^2-\left|\sum_{K\in\mathcal{T}^h}\int_K\delta_K\left(-\varepsilon\Delta v^h(\mathbf{x})+c(\mathbf{x})v^h(\mathbf{x})\right)\left(\mathbf{b}(\mathbf{x})\cdot\nabla v^h(\mathbf{x})\right)\,d\mathbf{x}\right|.
\end{aligned}$$

Now, the last term will be estimated from above. Then, one obtains altogether an estimate from below if the estimate of the last term is subtracted from the first term. In the following estimate, one uses the conditions (5.17) on the SUPG parameter. It is for each $K \in \mathcal{T}^h$

$$
\begin{aligned}
&\left| \int_K \delta_K \left(-\varepsilon \Delta v^h(\mathbf{x}) + c(\mathbf{x}) v^h(\mathbf{x}) \right) \left(\mathbf{b} \cdot \nabla v^h(\mathbf{x}) \right) \, d\mathbf{x} \right| \\
&\quad \leq \int_K \left(\delta_K^{1/2} \varepsilon \left| \Delta v^h(\mathbf{x}) \right| \right) \left(\delta_K^{1/2} \left| \mathbf{b} \cdot \nabla v^h(\mathbf{x}) \right| \right) d\mathbf{x} \\
&\qquad + \int_K \left(\delta_K^{1/2} \left| c(\mathbf{x}) \right| \left| v^h(\mathbf{x}) \right| \right) \left(\delta_K^{1/2} \left| \mathbf{b} \cdot \nabla v^h(\mathbf{x}) \right| \right) d\mathbf{x} \\
&\quad \overset{\text{CS}}{\leq} \left(\delta_K^{1/2} \varepsilon \left\| \Delta v^h \right\|_{0,K} + \delta_K^{1/2} \left\| c \right\|_{L^\infty(K)} \left\| v^h \right\|_{0,K} \right) \left\| \delta_K^{1/2} \left(\mathbf{b} \cdot \nabla v^h \right) \right\|_{0,K} \\
&\quad \overset{(5.2)}{\leq} \left(\delta_K^{1/2} \frac{\varepsilon C_{\text{inv}}}{h_K} \left\| \nabla v^h \right\|_{0,K} + \delta_K^{1/2} \left\| c \right\|_{L^\infty(K)} \left\| v^h \right\|_{0,K} \right) \left\| \delta_K^{1/2} \left(\mathbf{b} \cdot \nabla v^h \right) \right\|_{0,K} \\
&\quad \overset{(5.17)}{\leq} \left(\frac{h_K}{\sqrt{2\varepsilon} C_{\text{inv}}} \frac{\varepsilon C_{\text{inv}}}{h_K} \left\| \nabla v^h \right\|_{0,K} + \frac{\sqrt{\omega}}{\sqrt{2} \left\| c \right\|_{L^\infty(K)}} \left\| c \right\|_{L^\infty(K)} \left\| v^h \right\|_{0,K} \right) \left\| \delta_K^{1/2} \left(\mathbf{b} \cdot \nabla v^h \right) \right\|_{0,K} \\
&\quad = \left(\sqrt{\frac{\varepsilon}{2}} \left\| \nabla v^h \right\|_{0,K} + \sqrt{\frac{\omega}{2}} \left\| v^h \right\|_{0,K} \right) \left\| \delta_K^{1/2} \left(\mathbf{b} \cdot \nabla v^h \right) \right\|_{0,K} \\
&\quad \overset{\text{Young}}{\leq} \frac{\varepsilon}{2} \left\| \nabla v^h \right\|_{0,K}^2 + \frac{1}{4} \left\| \delta_K^{1/2} \left(\mathbf{b} \cdot \nabla v^h \right) \right\|_{0,K}^2 + \frac{\omega}{2} \left\| v^h \right\|_{0,K}^2 + \frac{1}{4} \left\| \delta_K^{1/2} \left(\mathbf{b} \cdot \nabla v^h \right) \right\|_{0,K}^2 \\
&\quad = \frac{1}{2} \left| \left| \left| v^h \right| \right| \right|_{\text{SUPG},K}^2 .
\end{aligned}
$$

Now, the proof is finished by summing over all mesh cells and inserting this estimate in the first estimate of the proof. ■

Corollary 5.23 Coercivity of the SUPG bilinear form for P_1 finite elements. *Let the assumptions of Theorem 5.22 with respect to the coefficients of the problem be valid. For piecewise linear finite elements, the SUPG bilinear form* (5.13) *is coercive with respect to the SUPG norm if*

$$
0 < \delta_K \leq \frac{\omega}{\left\| c \right\|_{L^\infty(K)}^2}. \tag{5.18}
$$

Proof: The proof is the same as for Theorem 5.22, where one uses that for piecewise linear finite elements $\Delta v^h(\mathbf{x})|_K = 0$ for all $K \in \mathcal{T}^h$. Thus, the corresponding terms do not appear in the proof. ■

Corollary 5.24 Existenz and uniqueness of a solution of the SUPG method. *Let the assumptions of Theorem 5.22 and Corollary 5.23, respectively, be valid. Then, the SUPG finite element method* (5.12) – (5.13) *has a unique solution.*

Proof: The statement is obtained by the application of the Theorem of Lax–Milgram, Theorem 4.10. The coercivity of the bilinear form was proved in Theorem 5.22 and Corollary 5.23, respectively. For the boundedness, one uses similar estimates as in the proof of Theorem 5.22 and in Example 4.9, *exercise*. ■

Remark 5.25 *On the coercivity of the SUPG bilinear form.*

- The proof of Theorem 5.22 is typical for the numerical analysis of stabilized finite element methods. One tries to get rid of the troubling terms by estimating them with the used norm. This approach works only if one uses an appropriate norm. In particular, the stabilization has to appear in the norm.
- Theorem 5.22 provides an upper bound for the SUPG parameter. This bound is generally not critical in applications.

- From Theorem 5.22 one obtains the stability of the SUPG method with respect to the SUPG norm. Stability means that an appropriate norm of the solution can be estimated with the data of the problem. It is

$$\begin{aligned}
&\left|\left|\left|u^h\right|\right|\right|_{\mathrm{SUPG}}^2 \\
&\quad\leq \quad 2a^h\left(u^h,u^h\right)=2f^h\left(u^h\right) \\
&\quad= \quad 2\left(f,u^h\right)+2\sum_{K\in\mathcal{T}^h}\int_K \delta_K f(\mathbf{x})\Big(\mathbf{b}(\mathbf{x})\cdot\nabla u^h(\mathbf{x})\Big)\,d\mathbf{x} \\
&\quad\overset{\mathrm{CS}}{\leq} \quad \frac{2}{\sqrt{\omega}}\left\|f\right\|_0\sqrt{\omega}\left\|u^h\right\|_0+2\sum_{K\in\mathcal{T}^h}\left\|\delta_K^{1/2}f\right\|_{0,K}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla u^h\right)\right\|_{0,K} \\
&\quad\overset{\mathrm{Young}}{\leq} \quad C\left\|f\right\|_0^2+\frac{1}{2}\left(\omega\left\|u^h\right\|_0^2+\sum_{K\in\mathcal{T}^h}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla u^h\right)\right\|_{0,K}^2\right).
\end{aligned}$$

  Now, the last terms on the right-hand side can be absorbed in the left-hand side and one has stability. The stability constant depends on ω and on the upper bound of δ_K.
- All $v^h\in V^h$ satisfy

$$\left|\left|\left|v^h\right|\right|\right|_{\mathrm{SUPG}}\geq\min\{1,\omega\}\left\|v^h\right\|_\varepsilon.$$

  Hence, the SUPG method is also stable with respect to the norm $\|\cdot\|_\varepsilon$. With respect to this norm, also the Galerkin finite element method is stable, however this method is not stable with respect to the SUPG norm. That means, the stability of the SUPG method is stronger than the stability of the Galerkin finite element method.

□

Theorem 5.26 Convergence of the SUPG method. *Let the solution of* (4.2) *satisfy* $u\in H^{k+1}(\Omega)$, $k\geq 1$, *let* $\mathbf{b}\in W^{1,\infty}(\Omega)$, $c\in L^\infty(\Omega)$, *and consider the SUPG method for* P_k *finite elements. Let the SUPG parameter be given as follows*

$$\delta_K=\begin{cases} C_0\dfrac{h_K^2}{\varepsilon} & \text{for } h_K<\varepsilon, \\ C_0h_K & \text{for } \varepsilon\leq h_K, \end{cases} \tag{5.19}$$

where the constant $C_0>0$ *is sufficiently small such that* (5.17) *is satisfied for* $k\geq 2$ *or* (5.18) *for* $k=1$, *respectively. Then, the solution* $u^h\in P_k$ *of the SUPG method* (5.12) *satisfies the following error estimate*

$$\left|\left|\left|u-u^h\right|\right|\right|_{\mathrm{SUPG}}\leq C\left(\varepsilon^{1/2}h^k+h^{k+1/2}\right)|u|_{k+1},$$

where the constant C *does not depend on* ε *and* h.

Proof: Let $u_I^h\in V^h$ be the Lagrange interpolant of $u(\mathbf{x})$. One obtains with the triangle inequality

$$\left|\left|\left|u-u^h\right|\right|\right|_{\mathrm{SUPG}}\leq\left|\left|\left|u-u_I^h\right|\right|\right|_{\mathrm{SUPG}}+\left|\left|\left|u_I^h-u^h\right|\right|\right|_{\mathrm{SUPG}}.$$

The first term on the right-hand side is the interpolation error. Using the interpolation error estimate (5.1), which is applied for each term of the SUPG norm individually, one

Figure 5.4: Joseph–Louis Lagrange (1736 – 1813).

gets

$$\begin{aligned}
&\left|\left|\left|u - u_I^h\right|\right|\right|_{\mathrm{SUPG}} \\
&\quad\leq \left(C\varepsilon h^{2k} \left|u\right|_{k+1}^2 + C\omega h^{2(k+1)} \left|u\right|_{k+1}^2 + C \sum_{K\in\mathcal{T}^h} \delta_K \left\|\mathbf{b}\right\|_{\infty,K}^2 h_K^{2k} \left|u\right|_{k+1,K}^2 \right)^{1/2} \\
&\quad\leq C\left(\varepsilon h^{2k} + h^{2(k+1)} + h^{2k+1}\right)^{1/2} \left|u\right|_{k+1} \\
&\quad\leq C\left(\varepsilon^{1/2} h^k + h^{k+1/2}\right) \left|u\right|_{k+1}.
\end{aligned}$$

Here, it was used that for both regimes it is $\delta_K \leq C_0 h_K \leq Ch$.

Consider now the second term on the right-hand side. The coercivity, Theorem 5.22, and the Galerkin orthogonality yield

$$\frac{1}{2}\left|\left|\left|u_I^h - u^h\right|\right|\right|_{\mathrm{SUPG}}^2 \leq a^h\left(u_I^h - u^h, u_I^h - u^h\right) = a^h\left(u_I^h - u, u_I^h - u^h\right).$$

Now, the triangle inequality is applied to $a^h\left(u_I^h - u, u_I^h - u^h\right)$ and every term is estimated individually. In these estimates, the interpolation estimate (5.1) plays an important role. Let $w^h = u_I^h - u^h$. One obtains for the diffusion term

$$\begin{aligned}
&\left|\varepsilon\left(\nabla\left(u_I^h - u\right), \nabla w^h\right)\right| \\
&\quad\overset{\mathrm{CS}}{\leq} \varepsilon\left\|\nabla\left(u_I^h - u\right)\right\|_0 \left\|\nabla w^h\right\|_0 = \varepsilon^{1/2}\left\|\nabla\left(u_I^h - u\right)\right\|_0 \varepsilon^{1/2}\left\|\nabla w^h\right\|_0 \\
&\quad\overset{(5.1)}{\leq} C\varepsilon^{1/2} h^k \left|u\right|_{k+1} \varepsilon^{1/2}\left\|\nabla w^h\right\|_0 \leq C\varepsilon^{1/2} h^k \left|u\right|_{k+1} \left|\left|\left|w^h\right|\right|\right|_{\mathrm{SUPG}}.
\end{aligned}$$

For the reactive term, one obtains in a similar way

$$\begin{aligned}
\left|\left(c(u_I^h - u), w^h\right)\right| &\overset{\mathrm{CS}}{\leq} \left\|c\right\|_\infty \left\|u_I^h - u\right\|_0 \left\|w^h\right\|_0 = \omega^{-1/2}\left\|c\right\|_\infty \left\|u_I^h - u\right\|_0 \omega^{1/2}\left\|w^h\right\|_0 \\
&\overset{(5.1)}{\leq} Ch^{k+1}\left|u\right|_{k+1} \left|\left|\left|w^h\right|\right|\right|_{\mathrm{SUPG}}.
\end{aligned}$$

Next, the terms are considered which come from the SUPG stabilization. Since for both

regimes it is $\varepsilon\delta_K \leq C_0 h_K^2$, one gets

$$\begin{aligned}
&\left| \sum_{K\in\mathcal{T}^h} \left(-\varepsilon\Delta\left(u_I^h - u\right), \delta_K \mathbf{b}\cdot\nabla w^h\right)_K \right| \\
&\overset{\text{CS}}{\leq} \sum_{K\in\mathcal{T}^h} \varepsilon^{1/2} \left\|\Delta\left(u_I^h - u\right)\right\|_{0,K} \varepsilon^{1/2}\delta_K^{1/2} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \\
&\leq C_0^{1/2} \sum_{K\in\mathcal{T}^h} h_K \varepsilon^{1/2} \left\|\Delta\left(u_I^h - u\right)\right\|_{0,K} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \\
&\overset{\text{CS}}{\leq} C_0^{1/2}\varepsilon^{1/2} h \left(\sum_{K\in\mathcal{T}^h} \left\|\Delta\left(u_I^h - u\right)\right\|_{0,K}^2\right)^{1/2} \left(\sum_{K\in\mathcal{T}^h} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}^2\right)^{1/2} \\
&\overset{(5.1)}{\leq} C\varepsilon^{1/2} h \left(\sum_{K\in\mathcal{T}^h} h_K^{2(k-1)} |u|_{k+1,K}^2\right)^{1/2} \left(\sum_{K\in\mathcal{T}^h} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}^2\right)^{1/2} \\
&\leq C\varepsilon^{1/2} h^k |u|_{k+1} \left|\left|\left|w^h\right|\right|\right|_{\text{SUPG}}.
\end{aligned}$$

For the other terms, one obtains with the relation $\delta_K \leq C_0 h_K$, which holds for both regimes,

$$\begin{aligned}
&\left| \sum_{K\in\mathcal{T}^h} \left(\mathbf{b}\cdot\nabla\left(u_I^h - u\right) + c\left(u_I^h - u\right), \delta_K\left(\mathbf{b}\cdot\nabla w^h\right)\right) \right| \\
&\overset{\text{CS}}{\leq} \sum_{K\in\mathcal{T}^h} \|\mathbf{b}\|_\infty \left\|\nabla\left(u_I^h - u\right)\right\|_{0,K} \delta_K^{1/2} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \\
&\quad + \sum_{K\in\mathcal{T}^h} \|c\|_\infty \left\|u_I^h - u\right\|_{0,K} \delta_K^{1/2} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \\
&\leq C \left(\sum_{K\in\mathcal{T}^h} h_K^{1/2} \left\|\nabla\left(u_I^h - u\right)\right\|_{0,K} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \right. \\
&\quad \left. + \sum_{K\in\mathcal{T}^h} h_K^{1/2} \left\|u_I^h - u\right\|_{0,K} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \right) \\
&\overset{\text{CS}}{\leq} Ch^{1/2} \left[\left(\sum_{K\in\mathcal{T}^h} \left\|\nabla\left(u_I^h - u\right)\right\|_{0,K}^2\right)^{1/2} + \left(\sum_{K\in\mathcal{T}^h} \left\|u_I^h - u\right\|_{0,K}^2\right)^{1/2} \right] \\
&\quad \times \left(\sum_{K\in\mathcal{T}^h} \left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}^2\right)^{1/2} \\
&\overset{(5.1)}{\leq} C\left(h^{k+1/2} + h^{k+3/2}\right) |u|_{k+1} \left|\left|\left|w^h\right|\right|\right|_{\text{SUPG}}.
\end{aligned}$$

To obtain an optimal estimate for the convective term, one has to apply first integration by parts

$$\begin{aligned}
\left(\mathbf{b}\cdot\nabla\left(u_I^h - u\right), w^h\right) &= \left(\nabla\left(u_I^h - u\right), \mathbf{b}w^h\right) = -\left(u_I^h - u, \nabla\cdot\left(\mathbf{b}w^h\right)\right) \\
&= -\left(u_I^h - u, (\nabla\cdot\mathbf{b})\, w^h\right) - \left(u_I^h - u, \mathbf{b}\cdot\nabla w^h\right).
\end{aligned}$$

Both terms on the right-hand side are estimated separately. Using the same tools as for

the other estimates, ones obtains

$$\begin{aligned}\left|\left(u_I^h - u, (\nabla\cdot\mathbf{b})\, w^h\right)\right| &\leq \omega^{-1/2}\left\|\nabla\cdot\mathbf{b}\right\|_\infty \left(\sum_{K\in\mathcal{T}^h}\left\|u_I^h - u\right\|_{0,K}^2\right)^{1/2} \omega^{1/2}\left\|w^h\right\|_0 \\ &\leq Ch^{k+1}\left|u\right|_{k+1}\left|\left|\left|w^h\right|\right|\right|_{\mathrm{SUPG}}.\end{aligned}$$

In the estimate of the other term, one has to distinguish if in the mesh cell K it is $\varepsilon \leq h_K$ or $\varepsilon > h_K$. One gets

$$\begin{aligned}&\left|\left(u_I^h - u, \mathbf{b}\cdot\nabla w^h\right)\right| \\ &\overset{\mathrm{CS}}{\leq} \sum_{\varepsilon\leq h_K}\delta_K^{-1/2}\left\|u_I^h - u\right\|_{0,K}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K} \\ &\quad + \sum_{\varepsilon> h_K}\left\|\mathbf{b}\right\|_\infty\left\|u_I^h - u\right\|_{0,K}\left\|\nabla w^h\right\|_{0,K} \\ &\overset{(5.1)}{\leq} C\left(\sum_{\varepsilon\leq h_K}\delta_K^{-1/2}h_K^{k+1}\left|u\right|_{k+1,K}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}\right. \\ &\quad \left. + \sum_{\varepsilon> h_K} h_K^{k+1}\left|u\right|_{k+1,K}\left\|\nabla w^h\right\|_{0,K}\right) \\ &\overset{C_0h_K\leq\delta_K,\varepsilon>h_K}{\leq} C\left(\sum_{\varepsilon\leq h_K} C_0^{-1/2}h_K^{-1/2}h_K^{k+1}\left|u\right|_{k+1,K}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}\right. \\ &\quad \left. + \sum_{\varepsilon> h_K} h_K^{k+1/2}\left|u\right|_{k+1,K}\varepsilon^{1/2}\left\|\nabla w^h\right\|_{0,K}\right) \\ &\overset{\mathrm{CS}}{\leq} Ch^{k+1/2}\left|u\right|_{k+1}\left[\left(\sum_{K\in\mathcal{T}^h}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}^2\right)^{1/2} + \varepsilon\left|w^h\right|_1\right] \\ &\leq Ch^{k+1/2}\left|u\right|_{k+1}\left|\left|\left|w^h\right|\right|\right|_{\mathrm{SUPG}}.\end{aligned}$$

Summarizing all estimates, the statement of the theorem is proved. ∎

Remark 5.27 *Concerning the error estimate.*

- In the convection-dominated regime $\varepsilon \ll h$, the order of convergence in the SUPG norm is $k + 1/2$ and in the diffusion-dominated case it is k. In the latter case, the SUPG norm is essentially the $H^1(\Omega)$ semi norm such that order k is optimal.
- It is essential for obtaining an estimate with a constant C which is independent of ε that the term
$$\left(\sum_{K\in\mathcal{T}^h}\left\|\delta_K^{1/2}\left(\mathbf{b}\cdot\nabla w^h\right)\right\|_{0,K}^2\right)^{1/2}$$
is part of the norm, which is used for estimating the error. Such an estimate does not hold for the norm $\left\|\cdot\right\|_\varepsilon$.
- For the interpretation of the results one has to take into account that different stabilization parameters by choosing different values of C_0 lead also to different norms on the left-hand side of the estimate.
- On the other hand, the value of a constant which is independent of ε is questionable since in general $\left|u\right|_{k+1}$ depends on ε.

- In numerical simulations, often one can observe even convergence of order h^{k+1} for the error in $L^2(\Omega)$, in particular on structured grids. However, in Zhou (1997) examples were constructed which show that the estimate of Theorem 5.26 is sharp also for the error in $L^2(\Omega)$.

□

Example 5.28 *SUPG in one dimension, the user-chosen constant.* The standard example

$$-\varepsilon u'' + u' = 1 \quad \text{on } (0,1), \quad u(0) = u(1) = 0,$$

does not fit into the theory of the SUPG method since $c(x) - \frac{b'(x)}{2} = 0$. Nevertheless, one can apply the SUPG method also for this example. An error estimate in the norm

$$\left(\varepsilon \left| v^h \right|_1^2 + \sum_{i=1}^{N} \left\| \delta_K^{1/2} b \left(v^h \right)' \right\|_{0,K}^2 \right)^{1/2}$$

can be proved. One looses the control on the error in the $L^2(0,1)$ norm.

A fundamental problem in the application of the SUPG method is the free constant C_0 in the definition of the parameter (5.19). At the present example, for $\varepsilon = 10^{-6}$, one can observe very well that one obtains for different constants rather different numerical results, see Figure 5.5. If C_0 is too large, then the layer is smeared, for an appropriate value of C_0 one obtains a solution which is almost exact in the nodes, and if C_0 is too small, then one can observe spurious (unphysical) oscillations in the layer.

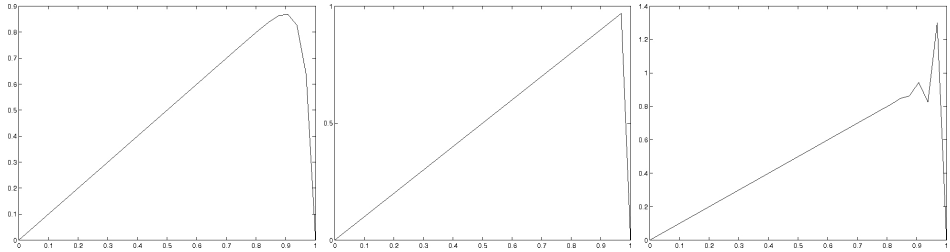

Figure 5.5: Results obtained with the SUPG method for the standard one-dimensional example, $C_0 = 1$, $C_0 = 0.5$, $C_0 = 0.25$ from left to right, $h = 1/32$, P_1 finite elements.

For general problems, it is difficult to choose C_0 appropriately. In higher dimensions, it is in general not possible to find C_0 such that the solution is (almost) exact in the nodes. Often, the solution computed with the SUPG method in higher dimensions exhibits spurious oscillations at the layers, e.g., see Example 5.30. □

Remark 5.29 *Different choices of the SUPG parameter.* In practice, one takes instead of (5.19) also the parameter

$$\delta_K = \frac{h_K}{2 \left\| \mathbf{b} \right\|_{L^\infty(K)}} \left(\coth(\mathrm{Pe}_K) - \frac{1}{\mathrm{Pe}_K} \right), \quad \mathrm{Pe}_K = \frac{\left\| \mathbf{b} \right\|_{L^\infty(K)} h_K}{2\varepsilon}, \tag{5.20}$$

where Pe_K is the local Péclet number, since in one dimensions one recovers under certain conditions the Iljin–Allen–Southwell scheme, see Example 5.16. There is no user-chosen constant in this parameter. Asymptotically, both parameters (5.19) and (5.20) have the same behavior. □

Example 5.30 *SUPG in two dimensions.* A standard test problem in two dimensions has the form

$$\begin{aligned} -\varepsilon\Delta u + (1,0)^T \cdot \nabla u &= 1 \quad \text{in } \Omega = (0,1)^2, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Besides the layer at the outflow boundary $x = 1$, there are also two layers at the boundaries $y = 0$ and $y = 1$. The layer at the outflow boundary is often called exponential layer and the layers parallel to the flow direction parabolic layers.

A numerical solution obtained for $\varepsilon = 10^{-8}$ with the Q_1 finite element method on a rather coarse grid and the SUPG parameter (5.20) is shown in Figure 5.6. One can see very well large spurious oscillations, in particular at the parabolic layers. These oscillations are a typical feature of solutions obtained with the SUPG method. They might become smaller with higher order elements or on finer grids. But they will generally vanish only if the layer is resolved.

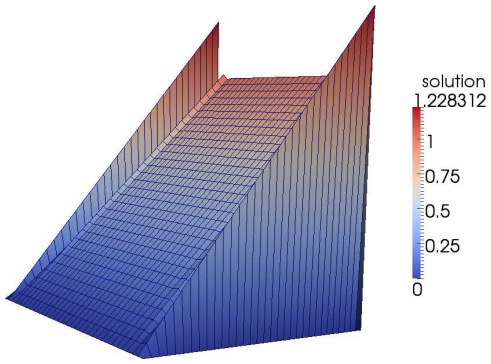

Figure 5.6: Result obtained with the SUPG method, Q_1 finite elements, and the SUPG parameter (5.20).

□