Chapter 2
Indicators of Errors for Approximate Solutions of Differential Equations

Abstract Error indicators play an important role in mesh-adaptive numerical algorithms, which currently dominate in mathematical and numerical modeling of various models in physics, chemistry, biology, economics, and other sciences. Their goal is to present a comparative measure of errors related to different parts of the computational domain, which could suggest a reasonable way of improving the finite dimensional space used to compute the approximate solution. An "ideal" error indicator must possess several properties: efficiency, computability, and universality. In other words, it must correctly reproduce the distribution of errors, be indeed computable, and be applicable to a wide set of approximations. In practice, it is very difficult to satisfy all these requirements simultaneously so that different error indicators are focused on different aims and stress some properties at the sacrifice of others. We discuss the mathematical origins and algorithmic implementation of the most frequently used error indicators. Our goal is twofold: to discuss the main types of error indicators, which have already gained high popularity in numerical practice, and to suggest a unified conception, which covers practically all methods used in error indication.

For differential equations, we discuss indicators of two types. Indicators of the first type show the distribution of errors in the whole computational domain. Another group of indicators is focused on the so-called goal-oriented error functionals typically associated with some subdomains ("zones of interest"), where the accuracy of an approximate solution is especially important. Usually, the indicators of the latter type use solutions of adjoint boundary value problems. We discuss some new forms of these indicators, which do not exploit extra regularity of solutions and special properties of respective approximations (such as, e.g., superconvergence). Indicators that follow from a posteriori error majorants of the functional type are discussed in Chap. 3.

2.1 Error Indicators and Adaptive Numerical Methods

Adaptive numerical methods are based on the conception that efficient approximations should be constructed by means of a sequence of consequently refined finite dimensional spaces $\{V_k\}$, $k = 1, 2, \ldots$ such that the amount of linearly independent trial functions in V_{k+1} is larger than in V_k (i.e., $\dim V_{k+1} > \dim V_k$). Typically,

O. Mali et al., *Accuracy Verification Methods*,
Computational Methods in Applied Sciences 32, DOI 10.1007/978-94-007-7581-7_2,

the structure of these spaces is a priori unknown. Within the framework of the adaptive modeling conception, the generation of V_{k+1} is based upon the information encompassed in the approximation u_k associated with V_k. For this reason, it is necessary to have computable quantities that furnish information on the error e presented in terms of a certain error measure (e.g., in terms of the energy norm). Such quantities are called *Error Indicators*. Throughout the book, we denote them by the symbol $\mathbb{E}$ (which is generated by the initial letters E and I). Error indicators play an important role in mesh-adaptive numerical algorithms, which follow the formal scheme

$$V_1 \xrightarrow{\mathbb{E}(u_1)} V_2 \xrightarrow{\mathbb{E}(u_2)} \cdots V_k \xrightarrow{\mathbb{E}(u_k)} V_{k+1}.$$

A "good" error indicator must be easily computable and must correctly reproduce the distribution of errors. It is also desirable that an indicator be applicable to a wide set of approximations and imply quantities that provide a realistic presentation on the overall (global) error. In practice, it is very difficult to satisfy all these requirements simultaneously, so that different error indicators are focused on different aims and stress some properties at the expense of the others.

In this chapter, we discuss the general principles of error indication and examples of error indicators with the paradigm of finite element approximations of elliptic partial differential equations.

2.1.1 Error Indicators for FEM Solutions

Let T_s, $s = 1, 2, \ldots, N$ be elements (subdomains) associated with the mesh $\mathfrak{T}_h$ (with characteristic size h), and let u_h be an approximate solution computed on this mesh. Henceforth, the corresponding finite dimensional space is denoted by V_h, so that $u_h \in V_h$. Then, the true error is $e = u - u_h$. Denote by $m_s(e)$ the value of the error measure m associated with T_s. Usually, the error measure $m_s(e)$ is defined as a certain integral of $u - u_h$ related to T_s. For example, local error measures of approximate solutions to linear elliptic problems are often presented by the integrals

$$\left(\int_{T_s} |u - u_h|^2 \, \mathrm{d}x\right)^{1/2} \quad \text{or} \quad \left(\int_{T_s} \left|\nabla(u - u_h)\right|^2 \mathrm{d}x\right)^{1/2}.$$

The components of the vector

$$\mathbf{m}(e) = \left\{m_1(e), m_2(e), \ldots, m_N(e)\right\}$$

are nonnegative numbers, which may be rather different.

If the overall error encompassed in u_h is too big, then a new approximate solution should be computed on a new (refined) mesh $\mathfrak{T}_{h_{\mathrm{ref}}}$. Comparative analysis of $m_s(e)$ suggests where to add new degrees of freedom (new trial functions). However, in real life computations the vector $\mathbf{m}(e)$ is not known and, therefore, an error indicator

$\mathbb{E}(u_h)$ is used. The corresponding approximate values of errors $\mathbb{E}_s$ associated with the elements form the vector

$$\mathbb{E}(u_h) = \{\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_N\},$$

which is used instead of $\mathbf{m}(e)$. If the vector $\mathbb{E}(u_h)$ is close to $\mathbf{m}(e)$, i.e.,

$$\mathbf{m}(e) \approx \mathbb{E}(u_h), \tag{2.1}$$

then a new mesh $\mathfrak{T}_{h_{\text{ref}}}$ can be efficiently constructed on the basis of comparative analysis of $\mathbb{E}_s$. However, the fact that the adaptive procedure is efficient depends on how accurately the condition (2.1) is satisfied and how efficiently the information encompassed in $\mathbb{E}(u_h)$ is used to improve approximations.

2.1.2 Accuracy of Error Indicators

Certainly, the condition (2.1) looks vague unless a formal definition of the sign $\approx$ is given. Despite the huge amount of publications focused on error indication, to the best of our knowledge no commonly used definition has yet been accepted. Different authors may claim (explicitly or implicitly) different things, so the words "good error indicator" may take on a variety of meanings.

Below we suggest definitions, which can be used for a reasonable qualification of error indicators. They define "strong" and "weak" meanings of $\approx$, respectively.

Definition 2.1 The indicator $\mathbb{E}(u_h)$ is ε-accurate on the mesh $\mathfrak{T}_h$ if

$$\mathcal{M}\big(\mathbb{E}(u_h)\big) := \frac{|\mathbf{m}(e) - \mathbb{E}(u_h)|}{|\mathbf{m}(e)|} \le \varepsilon. \tag{2.2}$$

The value of $\mathcal{M}(\mathbb{E}(u_h))$ is the strongest quantitative measure of the accuracy of $\mathbb{E}(u_h)$.

This definition imposes strong requirements on $\mathbb{E}(u_h)$. Indeed, (2.2) guarantees that inaccuracies in the error distribution computed by $\mathbb{E}(u_h)$ are much smaller (provided that ε is a small number) than the overall error. Therefore, an indicator should be regarded as "accurate" if it meets (2.2) with relatively coarse ε.

From (2.2) it follows that the so-called efficiency index

$$I_{\text{eff}}\big(\mathbb{E}(u_h)\big) := \frac{|\mathbb{E}(u_h)|}{|\mathbf{m}(e)|} \le 1 + \mathcal{M}\big(\mathbb{E}(u_h)\big) \tag{2.3}$$

is close to 1, which means that $|\mathbb{E}(u_h)|$ provides a good evaluation of the overall error $|\mathbf{m}(e)|$.

The efficiency of $\mathbb{E}(u_h)$ may be different for different meshes and approximate solutions. It is desirable that the indicator is accurate for a sufficiently wide class

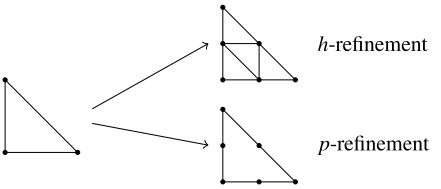

Fig. 2.1 Typical h-refinement and p-refinement

of approximations and meshes. The wider the class of approximations served by an indicator, the better it is from the computational point of view.

The majority of indicators suggested for finite element approximations are applicable only to Galerkin approximations (or to approximations that are very close to Galerkin solutions). Properties of the mesh used are also very important, and theoretical estimates of the quality of error indicators usually involve constants dependent on the aspect ratio of finite elements.

2.1.2.1 Marking Procedures

In adaptive finite element schemes, subsequent approximations are often constructed on nested meshes, where a refined mesh is obtained by "splitting" elements (h-refinement) or by increasing the amount and order of basis functions (p-refinement) of the current mesh. In Fig. 2.1, we depict typical refinement strategies for a linear triangular element, the degrees of freedom of which are function values at nodes. A detailed discussion on refinement methods can be found in, e.g., [BGP89, Dem07]. Alternative procedures intended to increase the set of basis functions lead to nonconforming methods (cf. Appendix B).

Typical adaptive schemes consists of solving the problem several times on a sequence of improving subspaces. In this type of practice, error indicators are used together with a certain *marker* that marks elements (subdomains) where errors are excessively high. A new subspace $V_{h_{\mathrm{ref}}}$ is constructed in such a way that these errors are diminished.

Let $\mathbf{B}$ denote the Boolean set $\{0, 1\}$ (we can assign the meaning "NO" to 0 and "YES" to 1). By $\mathbf{B}^N$ we denote the set of Boolean valued arrays (associated with one-, two- or multidimensional meshes) of total length N. If $\mathbf{b} = \{\flat_1, \flat_2, \ldots, \flat_N\} \in \mathbf{B}^N$, then $\flat_s \in \mathbf{B}$ for any $s = 1, 2, \ldots, N$. It is assumed that in the new mesh the elements (subdomains) marked by 1 should be refined, while those marked by 0 should be preserved (see Fig. 2.2). Note that the refined mesh in Fig. 2.2 contains the so-called "hanging nodes". In order to avoid them it is often necessary to refine also some neighboring subdomains marked by 0.

Remark 2.1 Modern mesh adaptation algorithms often make coarsening of a mesh in subdomains where local errors are insignificant (see, e.g., [BNP10, BS12, KM10,

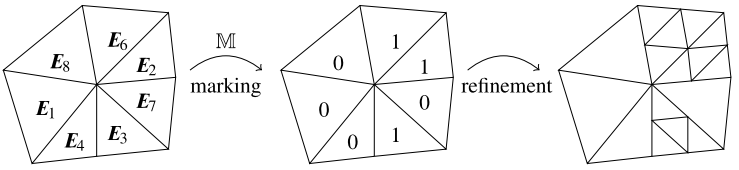

Fig. 2.2 Marking procedure and a refined mesh

Algorithm 2.1 Marking based on comparison with the average value

Input: $\boldsymbol{E}(u_h) \in \mathbb{R}^N$ {vector of errors indicated by $\boldsymbol{E}$}, N {number of elements}
$\widetilde{\boldsymbol{E}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{E}_i$ {Averaged value of the error on mesh elements}
for $i = 1 : N$ **do**
  if $\boldsymbol{E}_i \geq \widetilde{\boldsymbol{E}}$ **then**
    $\flat_i = 1$
  else
    $\flat_i = 0$
  end if
end for
Output: b {Marking of elements}

PPB12, Rhe80, SDW$^+$10, SMGG12] and the references cited therein). In this case, elements of $\mathbf{B}^N$ may attain three values: $\{-1, 0, 1\}$. The elements marked by -1 should be further aggregated in larger blocks.

From the mathematical point of view, marking is an operation performed by a special operator.

Definition 2.2 Marker $\mathbb{M}$ is a mapping (operator) acting from the set $\mathbb{R}^N_+$ (which contains estimated values of local errors) to the set $\mathbf{B}^N$.

Different markers generate different selection procedures, which are applied to the array of errors evaluated by an indicator $\boldsymbol{E}(u_h)$ in order to obtain a boolean array **b**. Further refinement is performed with the help of data encompassed in **b**.

Example 2.1 Algorithm 2.1 determines the simplest marker, which classifies the components of **e** into two groups by comparing with the average value.

Example 2.2 As before, $\boldsymbol{E}(u_h)$ is a vector with nonnegative components containing indicated errors and $\theta \in (0, 1)$ is a parameter (which determines the percentage of refined elements). Algorithm 2.2 ranks the values of $\boldsymbol{E}_i$ (from minimal to maximal

Algorithm 2.2 Marking based on a predefined amount of elements to be refined

Input: $\mathbb{E} \in \mathbb{R}^N$ {vector of errors}, N {number of elements}, $\theta \in (0,1)$
$i_{cut} = \mathbf{floor}((1-\theta)N)$
$\{\mathbb{E}_{\text{sorted}}, \mathbf{I}\} = \mathbf{sort}(\mathbb{E})$
for $i = 1$ **to** N **do**
  if $i < i_{cut}$ **then**
    $\flat(\mathbf{I}(i)) = 1$
  else
    $\flat(\mathbf{I}(i)) = 0$
  end if
end for
Output: b {Marking of elements}

values) and assigns 1 to the largest θN values. All other elements are marked by 0. In the formal description of the algorithm, we use a "sorting procedure" **sort**, which input is the array $\mathbb{E}$ and output is the array $\mathbb{E}_{\text{sorted}}$ containing local errors sorted in the descending order (i.e., $\mathbb{E}_{\text{sorted}}(j) \geq \mathbb{E}_{\text{sorted}}(j+1)$), and the array $\mathbf{I}$, which contains natural numbers (indexes of sorted elements) in the original vector, i.e., for any $j = 1, 2, \ldots, N$, $\mathbb{E}_{\text{sorted}}(j) = \mathbb{E}(\mathbf{I}(j))$. Algoritmization of such a procedure is a technical task, which we are not focused on. The procedure **floor**(z) selects the largest integer not greater than z.

Example 2.3 In the literature related to adaptive procedures, a selection method called the "bulk criterion" is often used. In it, we select by a certain method a set of elements for which the summed indicated error is greater than some "bulk" of the total indicated error (one of the first papers related to this method is [Dör96]; see also [BCH09]). Algorithm 2.3 forms the subset of elements which contains the highest indicated errors. The process stops when the error accumulated on previous steps exceeds the "bulk" level. This is sometimes referred to a "greedy" algorithm.

In order to demonstrate the performance of the above-discussed marking procedures, we consider the following diffusion problem:

$$\begin{aligned} -\Delta u &= 1, \quad \text{in } \Omega := (0,1)^2 \setminus \big([0.5,1] \times [0,0.5]\big), \\ u &= 0, \quad \text{on } \Gamma. \end{aligned}$$

We compute u_h by the finite element method using piecewise affine approximations (Courant elements), and use the indicator $\mathbb{E}(u_h)$ generated by the gradient-averaging method (see Sect. 2.2.2.1). We apply both Algorithms 2.1 and 2.2. In Fig. 2.3 the mesh and elements marked by a certain method are depicted (above) and the histogram of indicated errors and the marked elements are presented (below). In general, Algorithms 2.1 and 2.2 may suggest to refine rather different amount of elements.

Algorithm 2.3 Marking based on the bulk criterion

Input: $\mathbb{E}(u_h)$ {vector of errors}
$\theta \in (0, 1)$ {bulk factor}
$\{\mathbb{E}_{\text{sorted}}, \mathbf{I}\} = \mathbf{sort}(\mathbb{E})$
$\mathbb{E}_{tot} = \sum_{i=1}^{N} \mathbb{E}_i$ {total error}
$\mathbb{E}_{\text{bulk}} = \theta \mathbb{E}_{tot}$ {value of the "bulk" error}
$i = 1$
$\mathbb{E}_{tmp} = 0$ {temporary value of accumulated error}
while $\mathbb{E}_{\text{bulk}} \geq \mathbb{E}_{tmp}$ **do**
  $\flat(\mathbf{I}(i)) = 1$
  $\mathbb{E}_{tmp} = \mathbb{E}_{tmp} + \mathbb{E}_{\text{sorted}}(i)$
  $i = i + 1$
end while
Output: b {Marking of elements}

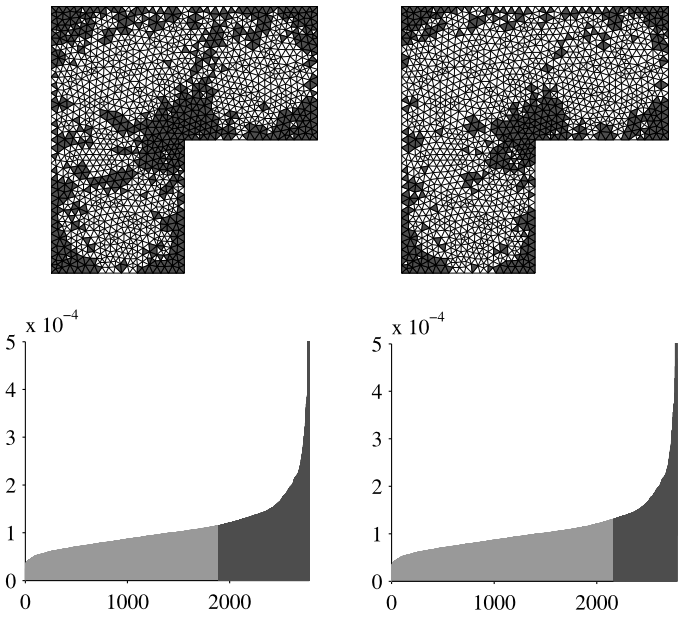

Fig. 2.3 Marking by Algorithms 2.1 (*left*) and 2.2 (*right*), marked elements ($\flat_i = 1$) are *colored darker*. *Above* are meshes and *below* the histograms of element-wise errors

Remark 2.2 We note that the marking of elements with the highest errors makes sense only if the errors differ significantly. If they have close values, then any ranking is not really motivated. For example, consider an almost uniform error distribution and two markings presented in Fig. 2.4. It is obvious that in this the case refining only the shadowed elements mesh is a rather disputable strategy because

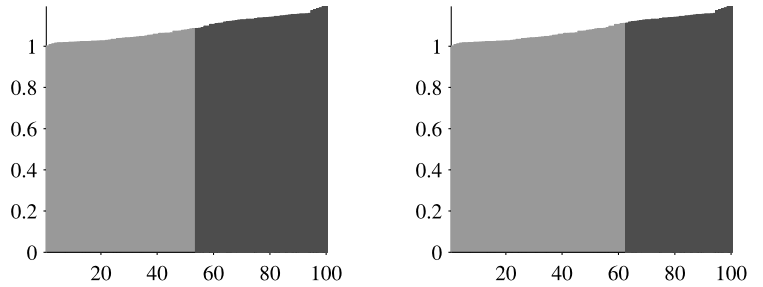

Fig. 2.4 Algorithms 2.1 (*left*) and 2.2 (*right*) are applied to mark elements of almost uniform error distribution, elements to be refined are *darker*

Table 2.1 Logical operation $\equiv$ in Definition 2.3

a	b	$a \equiv b$
0	0	1
1	0	0
0	1	0
1	1	1

$$\begin{aligned}
\mathbb{M}(\mathbf{m}(e)) &= [1\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 0\ \ 0\ \ 1\ \ 1] \\
\mathbb{M}(\mathbb{E})(u_h) &= [0\ \ 1\ \ 1\ \ 0\ \ 1\ \ 1\ \ 1\ \ 0\ \ 1\ \ 0] \\
(\mathbb{M}(\mathbf{m}(e)) \equiv \mathbb{M}(\mathbb{E})(u_h)) &= [0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 1\ \ 0\ \ 1\ \ 1\ \ 0]
\end{aligned}$$

every element makes almost equal contribution to the overall error. In this situation, the uniform refinement of all elements would be more adequate.

Remark 2.3 In principle, one can use the information provided by an indicator without any ranking procedure and construct a completely new mesh where element sizes are related to respective errors. Moreover, in adaptive hp-FEM, the element size and the order of basis functions can be varied simultaneously (see, e.g., [AS99, Dem07]).

To compare different error indicators in the context of element-wise marking, we introduce two operations with Boolean valued arrays. Let $\mathbf{a} = \{a_i\}$ and $\mathbf{b} = \{b_i\}$ be elements of $\mathbf{B}^N$. By $[]\mathbf{a}[]$ we denote the sum $\sum_{i=1}^{N} a_i$ and $\equiv$ denotes the logical equivalence rule (see Table 2.1, left).

Definition 2.3 An indicator $\mathbb{E}(u_h)$ is ε-accurate on the mesh $\mathfrak{T}_h$ with respect to the marker $\mathbb{M}$ if

$$\mathcal{M}\big(\mathbb{E}(u_h), \mathbb{M}\big) := 1 - \frac{[]\mathbb{M}(\mathbf{m}(e)) \equiv \mathbb{M}(\mathbb{E}(u_h))[]}{N} \le \varepsilon. \tag{2.4}$$

This definition is illustrated by Table 2.1 (right). We see that the operation $\equiv$ counts the cases in which markings based on the true error measure and on $\mathbb{E}(u_h)$ coincide. In the array of $N = 10$ elements the number of coincides is 5. Hence, in this example $\mathcal{M}(\mathbb{E}(u_h), \mathbb{M}) = 1 - \frac{5}{10} = 0.5$. This quantity shows that the indicator

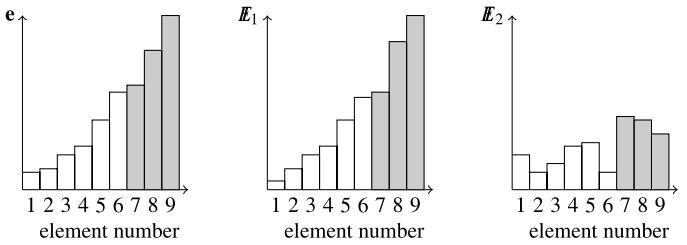

Fig. 2.5 True error distribution **e** for a set of nine elements and local errors generated by two indicators $\mathbb{E}_1$ and $\mathbb{E}_2$

is unacceptably coarse. If in another example we have an array containing, e.g., 10000 elements and the number of inconveniences (with respect to $\mathbb{M}(\mathbf{m}(e))$) is 8, then $\mathcal{M}(\mathbb{E}(u_h), \mathbb{M}) = 1 - \frac{9992}{10000} = 0.0008$. This shows the high accuracy of the indicator.

It is easy to see that the accuracy measure $\mathcal{M}(\mathbb{E}(u_h), \mathbb{M})$ is much weaker than the measure introduced in Definition 2.1. For example, in Fig. 2.5 we depict the exact distribution of local errors (left) and two distributions generated by two indicators (which are rather different). However, a marker designed to select three elements with the highest errors would select the same elements (shadowed). This example shows the difference between the accuracy measures (2.2) and (2.4). We see that the indicator $\mathbb{E}_2$ may be accurate in the sense of (2.4), but do not provide a true idea of the values of errors. This situation is quite typical. Often error indicators are based on heuristic argumentation and have no mathematical justification (in the best case they can be justified only in the above weak sense). Nevertheless, numerical analysts and engineers use them. Customarily they motivate this by saying that in some tests performed with the help of a marking procedure the indicator manages to properly mark the elements. In general, these arguments are not convincing because there is no guarantee that similar results will be obtained in other computations.

If $\mathbb{E}(u_h)$ is not accurate in the strong sense (i.e., it does not show actual values of the error), then the quality of marking may be good for one marker (mesh) and quite bad for another. Therefore, we believe that the indicators suggested for reliable numerical experiments should satisfy Definition 2.1.

It is clear that direct accuracy verification for an error indicator can be performed only in test examples where the exact solutions are known (so that we can find e). In other cases, the validity of an indicator is usually motivated by some indirect arguments (e.g., by those based on a priori regularity and asymptotic analysis). Some of the most popular motivations are considered below.

2.2 Error Indicators for the Energy Norm

To present various error indicators related to energy norms of linear elliptic equations within the framework of a unified scheme, we consider the classical Poisson's problem

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.5}$$

$$u = 0 \quad \text{on } \Gamma, \tag{2.6}$$

where Ω is an open bounded connected subset in $\mathbb{R}^d$ with Lipschitz continuous boundary Γ and $f \in L^2(\Omega)$.

The generalized solution (see Sect. B.1) satisfies the relation

$$\int_\Omega \nabla u \cdot \nabla w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x, \quad \forall w \in V_0 := \mathring{H}^1(\Omega). \tag{2.7}$$

Let $v \in V_0$ be an approximation of u. We are interested in evaluation of the global error norm $\|\nabla e\| = \|\nabla(u - v)\|$ and local errors $m_s(e)$ associated with subdomains (elements).

Note that

$$\sup_{w \in V_0} \left\{ \int_\Omega \big(\nabla(u - v) \cdot \nabla w\big) \, \mathrm{d}x - \frac{1}{2}\|\nabla w\|^2 \right\}$$

$$\leq \sup_{\tau \in L^2(\Omega, \mathbb{R}^d)} \int_\Omega \left(\nabla(u - v) \cdot \tau - \frac{1}{2}|\tau|^2 \right) \mathrm{d}x = \frac{1}{2}\big\|\nabla(u - v)\big\|^2.$$

On the other hand,

$$\sup_{w \in V_0} \int_\Omega \left(\nabla(u - v) \cdot \nabla w - \frac{1}{2}|\nabla w|^2 \right) \mathrm{d}x \geq \frac{1}{2}\big\|\nabla(u - v)\big\|^2.$$

Thus,

$$\big\|\nabla(u - v)\big\|^2 = \sup_{w \in V_0} \int_\Omega \big(2\nabla(u - v) \cdot \nabla w - |\nabla w|^2\big) \, \mathrm{d}x$$

$$= \sup_{w \in V_0} \left\{ -\|\nabla w\|^2 - 2 \int_\Omega (\nabla v \cdot \nabla w - f w) \, \mathrm{d}x \right\},$$

and we conclude that

$$\big\|\nabla(u - v)\big\|^2 = \sup_{w \in V_0} \big\{ -\|\nabla w\|^2 - 2\ell_v(w) \big\}, \tag{2.8}$$

where

$$\ell_v(w) := \int_\Omega (\nabla v \cdot \nabla w - f w)\,\mathrm{d}x$$

is the *residual functional*. This relation serves as a basis for various error estimation methods.

It is easy to show that the variational problem on the right-hand side of (2.8) has a unique solution and this solution is $w = u - v$. Indeed,

$$\ell_v(u-v) = \int_\Omega \big(\nabla v \cdot \nabla(u-v) - \nabla u \cdot \nabla(u-v)\big)\,\mathrm{d}x = -\big\|\nabla(u-v)\big\|^2,$$

and we see that the right-hand side coincides with the left-hand one. Hence, (2.8) implies

$$\big|\ell_v(u-v)\big| = \big\|\nabla(u-v)\big\|^2. \tag{2.9}$$

We can use (2.9) to indicate the error $\|\nabla(u-v)\|$ and classify the following three principal ways:

A: Estimate $\ell_v(u-v)$ in (2.8) from the above, and use the computable part(s) of the estimate as error indicator(s).
B: Replace ℓ_v in (2.8) by a close functional, which leads to a directly computable estimator.
C: Solve the problem (2.8) numerically.

Below we discuss several error indicators, which are based on the approaches (A), (B), or (C).

2.2.1 Error Indicators Based on Interpolation Estimates

Error estimators of this type can be referred to the group (A). They originate from the papers [BR78b, BR78a]. In the literature, they are often called "residual type a posteriori error estimators". Various modifications and advanced forms have been discussed in numerous publications (see, e.g., [AO92, AO00, BS01, BWS11, Car99, CV99, DR98, EJ88, JH92, Ver96]). Let the approximate solution $v = u_h$ be the Galerkin approximation computed on $V_{0h} \subset V_0$, i.e.,

$$\int_\Omega \nabla u_h \cdot \nabla w_h\,\mathrm{d}x = \int_\Omega f w_h\,\mathrm{d}x, \quad \forall w_h \in V_{0h}. \tag{2.10}$$

With the help of (2.10), we can deduce an upper bound of the residual functional and suggest error indicators associated with computable parts of the estimate.

We represent the residual functional in the form

$$\ell_{u_h}(w) = \int_\Omega \big(\nabla u_h \cdot \nabla(w - \pi_h w) - f(w - \pi_h w)\big)\,\mathrm{d}x,$$

where $\pi_h : V_0 \to V_{0h}$ denotes an interpolation operator. Assume that Ω consists of subdomains (e.g., simplexes T_k, which form the mesh $\mathfrak{T}_h$), and u_h is sufficiently regular on each subdomain. Then, we integrate by parts and obtain

$$\ell_{u_h}(w) = \sum_{k=1}^{N} \int_{T_k} (\Delta u_h + f)(\pi_h w - w)\,\mathrm{d}x + \sum_{\substack{l,s=1 \\ l>s}}^{N} \int_{E_{ls}} [\nabla u_h \cdot n_{ls}](w - \pi_h w)\,\mathrm{d}s, \tag{2.11}$$

where [ ] denotes the jump, E_{ls} is the common boundary (edge) of T_l and T_s (boundary edges do not have this term), and n_{ls} is the unit normal vector to E_{ls} outward to T_l if $l > s$ (we recall that the integral over E_{ls} is assumed to be equal to zero if the elements l and s have no common edge).

It is easy to see that

$$\int_{T_k} (\Delta u_h + f)(\pi_h w - w)\,\mathrm{d}x \le \|\Delta u_h + f\|_{T_k} \|w - \pi_h w\|_{T_k},$$

$$\int_{E_{ls}} \left[\frac{\partial u_h}{\partial n}\right](w - \pi_h w)\,\mathrm{d}s \le \left\|\left[\frac{\partial u_h}{\partial n}\right]\right\|_{E_{ls}} \|w - \pi_h w\|_{E_{ls}}.$$

Now, we need to bound $\|w - \pi_h w\|_{T_k}$ and $\|w - \pi_h w\|_{E_{ls}}$ by $\|\nabla w\|$, i.e., we need *interpolation estimates* associated with the operator π_h. The derivation of such estimates is more difficult than for the operator Π_h considered in Sect. C.2. It is clear that the estimates must rely on geometrical features of T_k and properties of V_{0h}. In the case of piecewise affine continuous approximations, a polygonal $\Omega \subset \mathbb{R}^2$, and a simplicial mesh, the corresponding interpolation operator $\pi_h : H^1(\Omega) \to V_{0h}$ has been studied in [Clé75].

Let $v \in \mathring{H}^1(\Omega)$ and X_j be an inner node of the triangulation $\mathfrak{T}_h$. We define the set

$$\omega_j := \{x \in T_t \mid X_j \in \overline{T}_t, t = 1, 2, \ldots, N\},$$

which contains all the elements having common node X_j. Define $p_j(x) \in P^1(\omega_j)$ by the relation

$$\int_{\omega_j} (v - p_j) q\,\mathrm{d}x = 0, \quad \forall q \in P^1(\omega_j). \tag{2.12}$$

This definition means that p_j is the L^2-projection of v on ω_j. Now, π_h is defined by setting

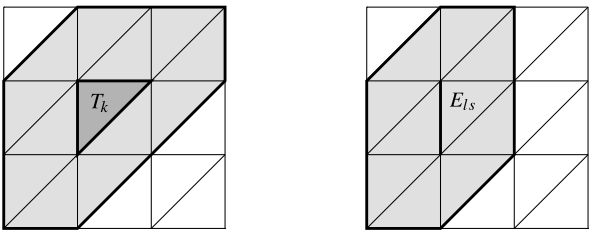

Fig. 2.6 The sets $\varpi(T_k)$ and $\varpi(E_{ls})$ on a regular mesh

$$\pi_h v(X_j) = p(X_j), \quad \forall X_j \in \operatorname{int}\Omega, \tag{2.13}$$

$$\pi_h v(X_j) = 0, \quad \forall X_j \in \Gamma. \tag{2.14}$$

This mapping is linear, continuous, and is subject to the relations (see, e.g., [Ver96])

$$\|v - \pi_h v\|_{2,T_k} \le C_{1k}^{int} \operatorname{diam} T_k \|v\|_{1,2,\varpi(T_k)}, \tag{2.15}$$

$$\|v - \pi_h v\|_{2,E_{ls}} \le C_{2ls}^{int} |E_{ls}|^{1/2} \|v\|_{1,2,\varpi(E_{ls})}, \tag{2.16}$$

where the sets (patches) associated with T_k and E_{ls} are defined as follows:

$$\varpi(T_k) := \{x \in \overline{T}_t \mid \overline{T}_t \cap \overline{T}_k \neq \emptyset, t = 1, 2, \ldots, N\},$$
$$\varpi(E_{ls}) := \{x \in \overline{T}_t \mid \overline{T}_t \cap \overline{E}_{ls} \neq \emptyset, t = 1, 2, \ldots, N\}.$$

See Fig. 2.6 for a clarifying illustration.

The constants C_{1k}^{int} and C_{2ls}^{int} depend on the structure of the mesh, and the factors $\operatorname{diam}(T_k)$ and $|E_{ls}|^{1/2}$ depend on the mesh size parameter h. We have

$$\begin{aligned}
&\sum_{k=1}^{N} \int_{T_k} (\Delta u_h + f)(w - \pi_h w)\,\mathrm{d}x \\
&\quad \le \sum_{k=1}^{N} \|\Delta u_h + f\|_{2,T_k} \|w - \pi_h w\|_{2,T_k} \\
&\quad \le \sum_{k=1}^{N} \|\Delta u_h + f\|_{2,T_k} C_{1k}^{int} \operatorname{diam} T_k \|w\|_{1,2,\varpi(T_k)} \\
&\quad \le \left(\sum_{k=1}^{N} \left(C_{1k}^{int}\right)^2 (\operatorname{diam} T_k)^2 \|\Delta u_h + f\|_{2,T_k}^2 \right)^{1/2} \sqrt{\iota_T(w)},
\end{aligned} \tag{2.17}$$

where $\iota_T(w) = \sum_{k=1}^{N} \|w\|_{1,2,\varpi(T_k)}^2$. It is easy to see that

$$\iota_T(w) \le C_T^2(\mathfrak{T}_h) \|w\|_{1,2,\Omega}^2, \tag{2.18}$$

where $C_T(\mathfrak{T}_h)$ depends on the topological structure of the mesh. We note that since one and the same element T_k occurs in several different patches ϖ, the constant is greater than one (it depends on the maximal amount of elements in a patch).

Analogously,

$$\begin{aligned}
&\sum_{\substack{l,s=1\\ l>s}}^{N} \int_{E_{ls}} [\nabla u_h \cdot n_{ls}](w - \pi_h w)\,\mathrm{d}s \\
&\quad \le \sum_{\substack{l,s=1\\ l>s}}^{N} \left\| [\nabla u_h \cdot n_{ls}] \right\|_{2,E_{ls}} C_{2ls}^{int} |E_{ls}|^{1/2} \|w\|_{1,2,\varpi(E_{ls})} \\
&\quad \le \left(\sum_{\substack{l,s=1\\ l>s}}^{N} \left(C_{2ls}^{int}\right)^2 |E_{ls}| \left\| [\nabla u_h \cdot n_{ls}] \right\|_{2,E_{ls}}^2 \right)^{1/2} \sqrt{\iota_E(w)},
\end{aligned} \tag{2.19}$$

where

$$\iota_E(w) = \sum_{\substack{l,s=1\\ l>s}}^{N} \|w\|_{1,2,\varpi(E_{ls})}^2.$$

We have

$$\iota_E(w) \le C_E^2(\mathfrak{T}_h) \|w\|_{1,2,\Omega}^2, \tag{2.20}$$

where $C_E(\mathfrak{T}_h)$ also depends on the mesh.

By (2.17) and (2.19), we find that

$$\begin{aligned}
\left|\ell_{u_h}(w)\right| \le \Bigg(& C_T \left(\sum_{k=1}^{N} \left(C_{1k}^{int}\right)^2 (\operatorname{diam} T_k)^2 \|\Delta u_h + f\|_{2,T_k}^2 \right)^{1/2} \\
& + C_E \left(\sum_{\substack{l,s=1\\ l>s}}^{m} \left(C_{2ls}^{int}\right)^2 |E_{ls}| \left\| [\nabla u_h \cdot n_{ls}] \right\|_{2,E_{ls}}^2 \right)^{1/2} \Bigg) \|w\|_{1,2,\Omega}.
\end{aligned} \tag{2.21}$$

Let $C = \max\{C_T, C_E\}\sqrt{1 + C_{F\Omega}^2}$. Then,

$$\left|\ell_{u_h}(w)\right| \le C \mathbb{E}(u_h) \|\nabla w\|, \tag{2.22}$$

where

$$\mathbb{E}(u_h) = \left(\sum_{k=1}^{N} \left(C_{1k}^{int}\right)^2 (\operatorname{diam} T_k)^2 \|\Delta u_h + f\|_{2,T_k}^2 \right)^{1/2} + \left(\sum_{\substack{l,s=1 \\ l>s}}^{N} \left(C_{2ls}^{int}\right)^2 |E_{ls}| \left\| [\nabla u_h \cdot n_{ls}] \right\|_{2,E_{ls}}^2 \right)^{1/2}.$$

By (2.8), we obtain

$$\left\| \nabla(u - u_h) \right\|^2 \le \sup_{w \in V_0} \left\{ -\|\nabla w\|^2 + 2C\,\mathbb{E}(u_h) \|\nabla w\| \right\} \le C^2 \mathbb{E}^2(u_h).$$

Hence,

$$\left\| \nabla(u - u_h) \right\| \le C\,\mathbb{E}(u_h). \tag{2.23}$$

We can represent this estimate in a slightly different form

$$\left\| \nabla(u - u_h) \right\| \le \hat{C}\,\hat{\mathbb{E}}(u_h), \tag{2.24}$$

where the indicator

$$\hat{\mathbb{E}}(u_h) = \left(\sum_{k=1}^{N} \left(C_{1k}^{int}\right)^2 (\operatorname{diam} T_k)^2 \|\Delta u_h + f\|_{2,T_k}^2 + \sum_{\substack{l,s=1 \\ l>s}}^{N} \left(C_{2ls}^{int}\right)^2 |E_{ls}| \left\| [\nabla u_h \cdot n_{ls}] \right\|_{2,E_{ls}}^2 \right)^{1/2}$$

is a sum of locally defined quantities.

It is worth outlining that in the process of deriving (2.23) and (2.24), we several times considerably overestimated the right-hand side, so that the equality sign in (2.8) and (2.11) is irretrievably lost. For this reason, the estimates obtained with the help of the above mathematical arguments may overestimate the error even if we manage to find and use sharp values of the interpolation constants C_{1k}^{int} and C_{2ls}^{int}. However, the latter task is not easy (especially for nonuniform meshes, which arise in the process of mesh adaptation). Indeed, to find C_{1k}^{int} we must solve the problem

$$\sup_{w \in V_0} \frac{\|w - \pi_h w\|_{2,\varpi(T_k)}}{\|w\|_{1,2,\varpi(T_k)}}, \tag{2.25}$$

which is an infinite dimensional problem. In some publications, it is suggested to find the constant approximately (e.g., by using a finite dimensional space formed by low order polynomial functions w). In this case, the true value of sup in (2.25) may be not achieved and, therefore, the overall estimate looses reliability. Moreover, solving a large number of local problems (2.25) (even for finite dimensional spaces) requires considerable numerical efforts. The corresponding computational

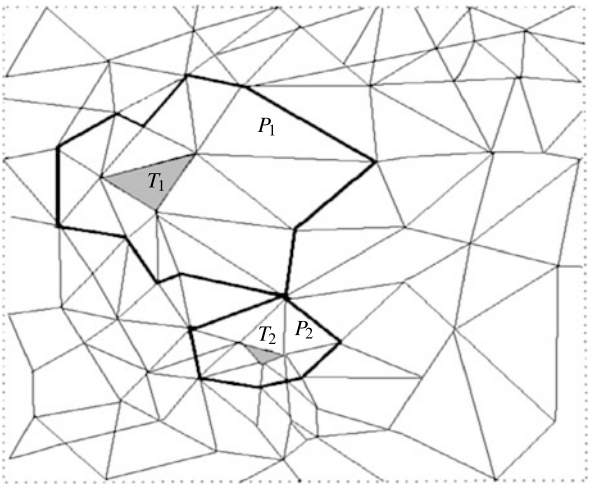

Fig. 2.7 Two patches of a nonuniform mesh

expenditures must be taken into account. After each mesh refinement, new constants associated with patches of the new mesh must be recomputed. Patches of highly nonuniform meshes may contain a different number of elements and complicated geometry (especially in 3D). For example, in Fig. 2.7 bold lines show boundaries of two patches P_1 and P_2 associated with two elements T_1 and T_2 of a nonuniform plane mesh. In real life computations, adaptive methods may generate meshes with much higher irregularities than those depicted in Fig. 2.7. In the case of highly irregular mesh, it is impossible to compute all the constants within the framework of a certain unified procedure similar to that we use for the constants in $H^2 \to C^0$ interpolation estimates, which can be fairly easily evaluated by interpolation estimates on the basic (etalon) simplex (see Sect. C.2). Thus, sharp computations of all the constants C_{1k}^{int} and C_{2ls}^{int} for thousands of different patches lead to high computational expenditures.

In view of these reasons, getting realistic and guaranteed error bounds with the help of (2.23) and (2.24) is rather challenging even for relatively simple elliptic equations (see, e.g., [CF00a], where these questions are systematically studied with the paradigm of boundary value problems in L-shaped domains).

A true meaning of the indicator $\hat{\mathbb{E}}$ is that it suggests easily computable quantities associated with elements, which can be used as error indicators. The standard argument for this is as follows. Assume that we use a quasi-uniform mesh. Then, we may assume that all (or almost all) constants C_{1k}^{int} have approximately the same value, and can be replaced by a single constant C_1^{int}. If the constants C_{2ls}^{int} are also replaced by a single constant C_2^{int}, then (2.21) implies an estimate

$$\hat{\mathbb{E}}(u_h) \approx \hat{C} \left(\sum_{k=1}^{N} \eta^2(T_k) \right)^{1/2}, \tag{2.26}$$

where

$$\eta^2(T_k) = \left(C_1^{int}\right)^2 (\operatorname{diam} T_k)^2 \|\Delta u_h + f\|^2_{2,T_k} + \frac{(C_2^{int})^2}{2} \sum_{E_{ls} \in \overline{T}_k} |E_{ls}| \left\| [\nabla u_h \cdot n_{ls}] \right\|^2_{2,E_{ls}}. \tag{2.27}$$

The multiplier $1/2$ arises in the second term because any interior edge is common for two elements.

Remark 2.4 Sometimes only the last term containing jumps is used as an efficient error indicator (in many cases it dominates, see, e.g., [CV99]).

2.2.2 Error Indicators Based on Approximation of the Error Functional

Assume that the functional ℓ_v in (2.8) can be efficiently approximated by another functional, i.e., $\ell_v \simeq \widetilde{\ell}_v$, and, moreover, for the new functional we have the estimate

$$\left|\widetilde{\ell}_v(w)\right| \le Q(v)\|\nabla w\|, \tag{2.28}$$

where $Q(v)$ is a computable nonnegative functional. Then, (cf. (2.8))

$$\begin{aligned} \left\|\nabla(u-v)\right\|^2 &= \sup_{w \in V_0} \left\{-\|\nabla w\|^2 - 2\ell_v(w)\right\} \simeq \sup_{w \in V_0} \left\{-\|\nabla w\|^2 - 2\widetilde{\ell}_v(w)\right\} \\ &\le \sup_{w \in V_0} \left\{-\|\nabla w\|^2 + 2Q(v)\|\nabla w\|\right\} = Q^2(v). \end{aligned} \tag{2.29}$$

This relation shows the general idea of generating indicators of the group (B) and motivates the indicator $Q(v)$. Certainly, the quality of such an error indicator depends on the closeness of ℓ_v and $\widetilde{\ell}_v$.[1] The functional $\widetilde{\ell}_v$ can be constructed by a certain post-processing procedure.

> Post-processing is a computational procedure that adjusts computed data to some a priori knowledge on properties of the exact solution. This procedure should be fairly simple, being compared with the expenditures required for computing the approximate solution.

Below, we describe several post-processing procedures.

[1] In general, the functionals must be close in the sense of $H^{-1}(\Omega)$.

2.2.2.1 Averaging of Gradients (Fluxes)

Gradient averaging procedures are often used to post-process gradients (fluxes, stresses) computed by finite element approximations of elliptic boundary value problems. Among first publications in this direction we mention the papers [ZZ87, ZZ88], which generated an interest in gradient recovery methods. Similar methods were investigated in numerous publications (see, e.g., [AO92, BC02, BR93, BS01, HTW02, Ver96, Wan00, WY02, ZBZ98, ZN05]). Mathematical justifications of the error indicators obtained in this way follow from the *superconvergence* phenomenon (see, e.g., [KN84, KNS98, Wah95]). Superconvergence arises on regular (quasiregular) meshes and, in simple terms, means that some components of approximate solutions obtained by inexpensive post-processing procedures converge to the corresponding components of the exact solution with a rate higher than the rate that can be predicted by standard a priori estimates. One of the most widely known results justified by superconvergence claims that a relatively simple averaging of ∇u_h yields a vector-valued function, which approximates ∇u much better than ∇u_h. Assume that in our problem this phenomenon takes place, and the gradient ∇u can be successfully represented by $G_h(\nabla u_h)$, where G_h is a certain post-processing operator. Then,

$$\int_\Omega (\nabla u_h \cdot \nabla w - f w)\,\mathrm{d}x \simeq \int_\Omega Z(u_h) \cdot \nabla w\,\mathrm{d}x,$$

where $Z(u_h) := \nabla u_h - G_h(\nabla u_h)$ (and (2.28) holds if we set $Q(u_h) = \|Z(u_h)\|$).

We recall (2.8) and deduce the relation

$$\|\nabla e\|^2 \simeq \sup_{w \in V_0} \left\{ -\|\nabla w\|^2 - 2\int_\Omega Z(u_h) \cdot \nabla w\,\mathrm{d}x \right\} \le \big\|Z(u_h)\big\|^2,$$

which means that

$$\|\nabla e\| \simeq \big\|Z(u_h)\big\|.$$

This relation suggests the idea to use the function $Z(u_h)$ as an error indicator and set

$$\mathbb{E}_s(u_h) = \big\|Z(u_h)\big\|_{T_s}.$$

So far we did not define particular forms of the operator G_h, which can be constructed by many different methods. Some of them are discussed below. At this point, we only note that

> Various post-processing procedures (averaging, smoothing, regularization) lead to various error indicators.

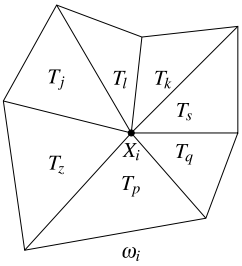

Fig. 2.8 A patch ω_i associated with the node X_i. $I_{\omega_i} = \{s, j, k, p, l, q, z\}$

2.2.2.2 Averaging of Fluxes in H^1

In the majority of cases, post-processing is performed by local averaging procedures. Consider the patch ω_i associated with the node X_i (see Fig. 2.8)

$$\overline{\omega}_i = \bigcup_{j \in I_{\omega_i}} \overline{T}_j,$$

where I_{ω_i} contains indexes of simplexes in ω_i.

Define $\mathbf{g}^{(i)}$ as the vector-valued function in $P^k(\omega_i, \mathbb{R}^d)$ solving the minimization problem:

$$\inf_{\mathbf{g} \in P^k(\omega_i, \mathbb{R}^d)} \int_{\omega_i} |\mathbf{g} - \nabla u_h|^2 \, \mathrm{d}x. \tag{2.30}$$

Using $\mathbf{g}^{(i)}$, we can define values of an averaged gradient at the node X_i.

Consider the simplest case $k = 0$ and assume that u_h is a piecewise affine continuous function. Then, the components of ∇u_h are constants on T_j. We denote them by $(\nabla u_h)_j$ and find $\mathbf{g}^{(i)} \in P^0(\omega_i, \mathbb{R}^d)$ such that

$$\int_{\omega_i} \left|\mathbf{g}^{(i)} - \nabla u_h\right|^2 \mathrm{d}x = \inf_{\mathbf{g} \in P^0(\omega_i, \mathbb{R}^d)} \int_{\omega_i} |\mathbf{g} - \nabla u_h|^2 \, \mathrm{d}x. \tag{2.31}$$

It is easy to see that

$$\mathbf{g}^{(i)} = \sum_{j \in I_{\omega_i}} \frac{|T_j|}{|\omega_i|} (\nabla u_h)_j. \tag{2.32}$$

We set $\mathrm{G}(\nabla u_h)(X_i) = \mathbf{g}^{(i)}$. Repeat this procedure for all nodes and define the vector-valued function $y_{\mathrm{G}} := \mathrm{G}(\nabla u_h)$ by the piecewise affine extrapolation of these values. This vector-valued function belongs to H^1 and in many cases approximates ∇u much better then the original (numerical) flux ∇u_h. This fact is justified by the *superconvergence phenomenon* (see, e.g., [KNS98, Wah95]).

Various averaging formulas of this type are represented in the form

$$\mathbf{g}^{(i)} = \sum_{j \in I_{\omega_i}} \lambda_j (\nabla u_h)_j, \qquad \sum_{j \in I_{\omega_i}} \lambda_j = 1, \tag{2.33}$$

where the quantities λ_j are weight factors. In (2.32), we set

$$\lambda_j = \frac{|T_j|}{|\omega_i|}.$$

If the mesh is regular and all the quantities $|T_{ij}|$ are equal, then (2.32) reads

$$\mathbf{g}^{(i)} = \frac{1}{M} \sum_{j \in I_{\omega_i}} (\nabla u_h)_j, \tag{2.34}$$

where M is the number of elements in ω_i. For internal nodes, the factors λ_{ij} may also be defined by the rule

$$\lambda_j = \frac{|\gamma_j|}{2\pi},$$

where $|\gamma_j|$ is the radian measure of the angle of T_j associated with the node X_i. However, if a node belongs to the boundary, then it is better to choose special weights. Their values depend on the mesh and on the boundary type (see, e.g., [HK87]).

Another way of defining $\mathbf{g}^{(i)}$ is to solve the problem

$$\inf_{g \in \mathbb{P}^k(\omega_i, \mathbb{R}^d)} \sum_{s=1}^{m_i} \left| g(x_s) - \nabla u_h(x_s) \right|^2,$$

where the points $x_s \in \overline{\omega}_i$ are so-called *superconvergent* points (see, e.g., [KN87, KNS98]).

If $k = 0$, then by similar arguments we obtain

$$\mathbf{g}^{(i)} = \frac{1}{m_i} \sum_{s=1}^{m_i} \nabla u_h(x_s). \tag{2.35}$$

As in the previous case, we define the vector-valued function $\mathrm{G}_h(\nabla u_h)$ by the piecewise affine extrapolation of these values.

2.2.2.3 Averaging of Fluxes in $H(\Omega, \mathrm{div})$

Post-processing operators for fluxes can be based on Raviart–Thomas elements of the lowest order (see, e.g., in [BF86, RT91]). The corresponding averaging operator G_{RT} generates an averaged flux in the space $H(\Omega, \mathrm{div})$ by averaging normal components of fluxes. Since the true flux belongs to this space (provided that $f \in L^2(\Omega)$), this way of averaging is quite natural.

Consider a patch formed by two elements T_i and T_j having a common edge E_{ij} (see Fig. 2.9). If u_h is constructed by P^1-approximations, then $(\nabla u_h)|_{T_i}$ and

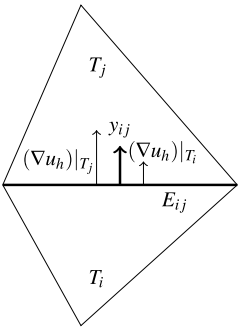

Fig. 2.9 Patch related to E_{ij} and averaged flux y_{ij}

$(\nabla u_h)|_{T_j}$ are constant vectors. In general, their normal components on E_{ij} are different. We define the (common) normal flux on E_{ij} as follows:

$$(y \cdot n_{ij})|_{E_{ij}} = \big(\kappa_{ij}(\nabla u_h)|_{T_i} + (1-\kappa_{ij})(\nabla u_h)|_{T_j}\big) \cdot n_{ij},$$

where $\kappa_{ij} \in (0,1)$ is the weight factor associated with E_{ij}. In the simplest case, $\kappa_{ij} = 1/2$. Another option (which takes into account the sizes of the elements) is

$$\kappa_{ij} = \frac{|T_i|}{|T_i| + |T_j|}.$$

For the boundary edges, we use the only one existing flux. Thus, three normal fluxes on three sides of each element are determined. The field inside the element is obtained by the standard RT^0-extension of normal fluxes. As a result, we have an averaged flux

$$y_{\mathrm{RT}} = \mathrm{G}_{\mathrm{RT}}(\nabla u_h) \in H(\Omega, \mathrm{div}).$$

Similar averaging procedures can be constructed in the case of 3D approximations, e.g., by averaging normal fluxes over the faces of a tetrahedron.

2.2.2.4 Averaging of Fluxes with Partial Equilibration

Since the exact flux p must satisfy the equilibrium (balance) equation $\mathrm{div}\, p + f = 0$, it is sensible to post-process it in such a way that the residual of this equation is minimal (e.g., in the integral sense). There are methods that produce equilibrated (or almost equilibrated) fluxes (see, e.g., [AO00, Bra07, LL83]). Sometimes these methods are rather sophisticated and use solutions of local Neumann type problems on patches. We have no space to properly discuss them here more systematically and, therefore, refer the reader to the above-mentioned and many other publications cited therein.

We conclude by describing a simple relaxation type algorithm, which allows to quasi-equilibrate y_{RT}.

Consider two neighboring elements with common edge E_{ij}. Our goal is to select the quantity $\gamma_{ij} = y \cdot n_{ij}$ in such a way that

$$\int_{T_i} \big((\operatorname{div} y)|_{T_i} + f\big)^2 \,\mathrm{d}x + \int_{T_j} \big((\operatorname{div} y)|_{T_j} + f\big)^2 \,\mathrm{d}x \to \min.$$

We use the identity $\int_{T_i} \operatorname{div} y \,\mathrm{d}x = \int_{\partial T_i} y \cdot n_i \,\mathrm{d}x$ and the fact that $(\operatorname{div} y)|_{T_i}$ and $(\operatorname{div} y)|_{T_j}$ are constant on T_i and T_j, respectively. Then, the corresponding value of γ_{ij} is explicitly defined by the relation (see [Rep08])

$$\gamma_{ij} = \frac{\mu_j |T_i| - \mu_i |T_j| + |T_i||T_j|(\{f\}_{T_j} - \{f\}_{T_i})}{|E_{ij}|(|T_i| + |T_j|)}, \tag{2.36}$$

where $\{f\}_{T_j}$ is the mean value of f on T_j.

Using the same idea, we recompute normal fluxes for all edges. At each step of this procedure the value of $\| \operatorname{div} y + f \|^2_\Omega$ decreases. After several cycles of minimization we obtain a vector-valued field, which is equilibrated much better than the original one.

2.2.2.5 Global Averaging

In many cases, an efficient averaging operator is obtained if local minimization problems on patches are replaced by a global problem (this method may generate essential computational expenditures). Consider the following problem: Find $\bar{\mathbf{g}}_h$ in a certain (global) set $U_h(\Omega)$, which minimizes the quantity $\sum_i \int_{T_i} |\mathbf{g}_h - \nabla u_h|^2 \,\mathrm{d}x$ among all $\mathbf{g}_h \in U_h(\Omega)$. Very often $\bar{\mathbf{g}}_h$ is a better image of ∇u than the functions obtained by local procedures. Moreover, mathematical justifications of the methods based on global averaging procedures can be performed under weaker assumptions, which makes them applicable to a wider class of problems (see, e.g., [CB02, CF00b, HTW02]).

2.2.2.6 Averaging by Least Squares Surface Fitting

In [Wan00], it was suggested a different recovery procedure, which is efficient for problems with sufficiently smooth solutions. The analysis is based on the representation

$$u - Q_\tau u_h = (u - Q_\tau u) + Q_\tau (u - u_h), \tag{2.37}$$

where u is the exact solution of a linear elliptic problem, u_h is the Galerkin approximation computed on a mesh $\mathfrak{T}_h$, and Q_τ is the L^2-projection operator on the finite dimensional space constructed on a mesh $\mathfrak{T}_\tau$ with the help of piecewise polynomial functions of the order $r \geq 0$. The key estimate is

$$\| Q_\tau u - Q_\tau u_h \| \leq C h^{s-1+\alpha \min\{0, 2-s\}} \| u - u_h \|_{H^1}, \tag{2.38}$$

where $\alpha \in (0, 1)$ is a parameter that connects h and τ in the way $\tau = h^\alpha$. The original problem is assumed to be H^s-regular with $1 \le s \le k+1$, and k is the degree of polynomials used in the Galerkin approximation. From (2.38) it follows that

$$\|u - Q_\tau u_h\| \le C h^{\beta(h,\tau,r,k)} \tag{2.39}$$

provided that $u \in H^{k+1}(\Omega) \cap H^{r+1}(\Omega_0) \cap V_0$. In (2.39), the rate β depends on h, τ, r, and k, and is greater than 2, provided that u is regular enough, and the space V_τ is selected appropriately (i.e., it is sufficiently rich). The constant C depends on the norm of u. Concrete values of the convergence rate for various k, r, and α are presented in the paper [Wan00].

2.2.2.7 Error Indicators Based on Solutions of Local Subproblems

The splitting of the error functional $\ell_{u_h}(w)$ into a number of functionals (defined by solutions of local subproblems (see, e.g., [Ain98, AO00] and further developments in [AR10]) generates another class of error indicators, which can be assigned to the group (B). Below we present a sketch of the underlying ideas. For a consequent study, we address the reader to the above-cited literature and many other publications cited therein.

Let Ω be a union of nonoverlapping domains (elements) Ω_i, $i = 1, 2, \dots, N$. Denote the common edge of Ω_i and Ω_j by Γ_{ij} and $\Gamma_{0i} := \partial\Omega_i \cap \Gamma$ and assume that for each Ω_i we know a function u_i such that

$$\ell_{u_h}(w) = \sum_{i=1}^{N} \int_{\Omega_i} \nabla u_i \cdot \nabla w \, \mathrm{d}x. \tag{2.40}$$

Consider a function $\bar{u} : \Omega \to \mathbb{R}$ that coincides with $u_i(x)$ if $x \in \Omega_i$. Assume that the functions u_i preserve continuity on the boundaries Γ_{ij} and the function $\bar{u}(x)$ belongs to $H^1(\Omega)$. Then, (2.40) reads

$$\int_\Omega \nabla(\bar{u} + u_h) \cdot \nabla w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x, \quad \forall w \in V_0(\Omega). \tag{2.41}$$

The relation (2.41) means that $u = u_i + u_h$ on Ω_i. Therefore, $u_i = u - u_h$, and we know the errors.

One way to determine u_i is to use solutions of local subproblems with Neumann (or Dirichlet–Neumann) type boundary conditions. For each Ω_i we solve the following problem: Find $u_i \in H^1(\Omega_i)$ such that $u_i = 0$ on Γ_{i0} and

$$\int_{\Omega_i} \nabla u_i \cdot \nabla w \, \mathrm{d}x \cong \int_{\Omega_i} f w \, \mathrm{d}x + \sum_{j=1}^{N} \int_{\Gamma_{ij}} \zeta_{ij} g w \, \mathrm{d}s - \int_{\Omega_i} \nabla u_h \cdot \nabla w \, \mathrm{d}x, \quad \forall w \in V_0(\Omega_i), \tag{2.42}$$

where g is a reconstruction of ∇u_h (in the simplest case this reconstruction can be performed by averaging of $\nabla u_h \cdot n_{ij}$ associated with two neighboring elements). The space $V_0(\Omega_i)$ is defined as follows. If $\Gamma_{0i} \neq \emptyset$, then $V_0(\Omega_i)$ is a subspace of $H^1(\Omega_i)$, which contains the functions vanishing on Γ_{i0}. If $\Gamma_{0i} = \emptyset$, then the local problem is considered with Neumann conditions and $V_0(\Omega_i)$ is the subspace of $H^1(\Omega_i)$ containing functions with zero mean. The weight ζ_{ij} is equal to zero if $i = j$. If $i > j$, then it is equal to 1, and $\zeta_{ij} = -1$ in the opposite case. It is easy to see that each internal boundary Γ_{ij} generates two integrals with equal absolute values and opposite signs. Therefore, the sum of all integrals does not contain such terms, and we obtain

$$\sum_{i=1}^{N} \int_{\Omega_i} \nabla u_i \cdot \nabla w \, \mathrm{d}x = \int_{\Omega} f w \, \mathrm{d}x - \int_{\Omega} \nabla u_h \cdot \nabla w \, \mathrm{d}x$$
$$= \ell_{u_h}(w), \quad \forall w \in V_0(\Omega), \tag{2.43}$$

which shows that the relation (2.40) holds.

This simple procedure may contain certain technical difficulties. One of them is that for internal domains the function g (which defines the Neumann type boundary conditions of the local subproblems) cannot be taken arbitrarily. This follows from the fact that the Neumann problem may be unsolvable if the external data do not satisfy an additional condition. For the problem (2.42) this condition is as follows:

$$\int_{\Omega_i} f \, \mathrm{d}x + \sum_{j=1}^{N} \int_{\Gamma_{ij}} \zeta_{ij} g \, \mathrm{d}s = 0. \tag{2.44}$$

Therefore, a special equilibration procedure that transforms g in order to satisfy (2.44) on each element is required. After that, exact solutions u_i of local problems must be found. Except special cases, this problem cannot be solved exactly and, therefore, instead of u_i some approximations $\widetilde{u}_i$ of local solutions are often used. Then, $\ell_{u_h}(w)$ is replaced by a directly computable functional

$$\widetilde{\ell}_{u_h}(w) = \sum_{i=1}^{N} \int_{\Omega_i} \nabla \widetilde{u}_i \cdot \nabla w \, \mathrm{d}x. \tag{2.45}$$

It generates the quantities $\mathbb{E}_i(u_h) = \|\nabla \widetilde{u}_i\|_{\Omega_i}$, which can be used to indicate local errors, and the quantity $|\mathbb{E}(u_h)| = (\sum_i (\mathbb{E}_i(u_h))^2)^{1/2}$, which serves as an indicator of the global error. Accuracy of such an estimate depends on the choice of g and on the accuracy of the computed approximations $\widetilde{u}_i$.

2.2.3 Error Indicators of the Runge Type

Consider again the case $v = u_h$, where u_h is the Galerkin approximation on $V_{0h} \subset V_0$. We can try to get an error indicator by solving the variational problem

in (2.8) numerically using a certain finite dimensional subspace $V_{0h_{\mathrm{ref}}}$ instead of V_0 (cf. (2.7)), i.e., by applying the relation

$$\left\|\nabla(u-u_h)\right\|^2 \geq \sup_{w\in V_{0h_{\mathrm{ref}}}} \left\{-\|\nabla w\|^2 - 2\ell_{u_h}(w)\right\}. \tag{2.46}$$

Thus, in our classification, estimators of this group belong to the class (C). It should be noted that this procedure makes sense only if the space $V_{0h_{\mathrm{ref}}}$ is essentially richer than V_{0h} (if $V_{0h_{\mathrm{ref}}} = V_{0h}$ then $\ell_{u_h}(w) = 0$, for any $w \in V_{0h_{\mathrm{ref}}}$ and, therefore, the value of sup in (2.46) is zero).

Assume that

$$V_{0h} \subset V_{0h_{\mathrm{ref}}}, \quad \dim V_{0h_{\mathrm{ref}}} > \dim V_{0h}. \tag{2.47}$$

The function $w_{h_{\mathrm{ref}}}$ maximizing the right-hand side of (2.46) satisfies the relation

$$\int_\Omega \nabla w_{h_{\mathrm{ref}}} \cdot \nabla w \,\mathrm{d}x = \int_\Omega (f w - \nabla u_h \cdot \nabla w)\,\mathrm{d}x, \quad \forall w \in V_{0h_{\mathrm{ref}}}, \tag{2.48}$$

which is equivalent to

$$\int_\Omega \nabla(w_{h_{\mathrm{ref}}} + u_h) \cdot \nabla w \,\mathrm{d}x = \int_\Omega f w \,\mathrm{d}x, \quad \forall w \in V_{0h_{\mathrm{ref}}}. \tag{2.49}$$

Hence, $u_{h_{\mathrm{ref}}} = w_{h_{\mathrm{ref}}} + u_h$, where $u_{h_{\mathrm{ref}}}$ is the Galerkin solution on $V_{0h_{\mathrm{ref}}}$. We have

$$\left\|\nabla(u-u_h)\right\|^2 \geq -\left\|\nabla(u_{h_{\mathrm{ref}}} - u_h)\right\|^2 - 2\ell_{u_h}(u_{h_{\mathrm{ref}}} - u_h).$$

Since

$$\begin{aligned}
\ell_{u_h}(u_{h_{\mathrm{ref}}} - u_h) &= \int_\Omega \left(\nabla u_h \cdot \nabla(u_{h_{\mathrm{ref}}} - u_h) - f(u_{h_{\mathrm{ref}}} - u_h)\right)\mathrm{d}x \\
&= \int_\Omega \left(\nabla u_h \cdot \nabla(u_{h_{\mathrm{ref}}} - u_h) - \nabla u_{h_{\mathrm{ref}}} \cdot \nabla(u_{h_{\mathrm{ref}}} - u_h)\right)\mathrm{d}x \\
&= -\left\|\nabla(u_h - u_{h_{\mathrm{ref}}})\right\|^2,
\end{aligned}$$

we conclude that the quantity $\|\nabla(u_h - u_{h_{\mathrm{ref}}})\|$ estimates $\|\nabla e\|$ from below. If $V_{0h_{\mathrm{ref}}}$ is much wider than V_{0h}, then $\|\nabla(u_h - u_{h_{\mathrm{ref}}})\|$ can be used to measure the global error, and the corresponding contributions $\mathbb{E}_s$ can be used for indication of element-wise errors. It is easy to see that this type error indicator always underestimates the error. In fact, it coincides with the indicator suggested by C. Runge at the beginning of the 20th century. In the simplest form, it reads as follows: *if the difference between two approximate solutions computed on a coarse mesh $\mathfrak{T}_h$ and on a certain refined mesh $\mathfrak{T}_{h_{\mathrm{ref}}}$ (e.g., $h_{\mathrm{ref}} = h/2$) has become small, then both $u_{h_{\mathrm{ref}}}$ and u_h are probably close to the exact solution u.*

In other words, this rule suggests the use of global or local norms of $u_h - u_{h_{\mathrm{ref}}}$ as error indicators. Henceforth, we denote it by $\mathbb{E}_{\mathrm{Runge}}(u_h)$. This indicator is simple

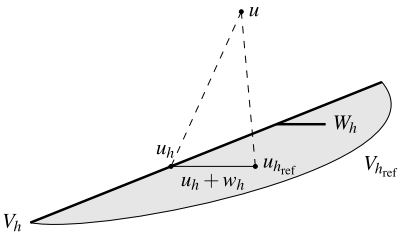

Fig. 2.10 The subspaces V_h, W_h, and $V_{h_{\mathrm{ref}}}$, the exact solution u and solutions u_h and $u_{h_{\mathrm{ref}}}$, from the respective subspaces

and looks very natural. For these reasons, it was easily accepted by engineers, who often consider it as a self-evident criterion. However, it is not difficult to find examples showing that this heuristic rule may be wrong. In particular, $\mathbb{E}_{\mathrm{Runge}}(u_h)$ may lead to misleading conclusions if the space V_h has been refined "improperly", i.e., if new (appended) trial functions do not really improve the approximation. In that case, u_h and $u_{h_{\mathrm{ref}}}$ may be quite close to each other but not close to u. We note that a correct form of the Runge's rule, which indeed provides guaranteed upper bounds of approximation errors, follows from error majorants of the functional type (see Sect. 3.6 of [Rep08] and Sect. 3.5.1 of this book).

Below we discuss *hierarchically based error indication methods*, where error indicators are constructed with the help of auxiliary problems on enriched finite dimensional subspaces (local or global) (see, e.g., [Ago02, DLY89, DMR91, DN02] and the references therein). Thus, in principle they invoke the same idea as does the Runge indicator, but in a more economical way.

Assume that the spaces V_h and $V_{h_{\mathrm{ref}}}$ are constructed in such a way that

$$V_{h_{\mathrm{ref}}} = V_h \oplus W_h.$$

In Fig. 2.10, we schematically depict the space V, the subspaces V_h, W_h, and $V_{h_{\mathrm{ref}}}$ and the corresponding approximate solutions u_h and $u_{h_{\mathrm{ref}}}$. It is easy to see that

$$\int_\Omega \big|\nabla(u-u_h)\big|^2\,\mathrm{d}x = \int_\Omega \big|\nabla(u-u_{h_{\mathrm{ref}}})\big|^2\,\mathrm{d}x + \int_\Omega \big|\nabla(u_h-u_{h_{\mathrm{ref}}})\big|^2\,\mathrm{d}x + 2\int_\Omega \nabla(u-u_{h_{\mathrm{ref}}})\cdot\nabla(u_{h_{\mathrm{ref}}}-u_h)\,\mathrm{d}x,$$

where

$$\int_\Omega \nabla(u-u_{h_{\mathrm{ref}}})\cdot\nabla(u_{h_{\mathrm{ref}}}-u_h)\,\mathrm{d}x = \int_\Omega f(u_{h_{\mathrm{ref}}}-u_h)\,\mathrm{d}x - \int_\Omega f u_{h_{\mathrm{ref}}}\,\mathrm{d}x + \int_\Omega f u_h\,\mathrm{d}x = 0.$$

Hence,

$$\big\|\nabla(u-u_h)\big\|^2 = \big\|\nabla(u-u_{h_{\mathrm{ref}}})\big\|^2 + \big\|\nabla(u_h-u_{h_{\mathrm{ref}}})\big\|^2 = \big\|\nabla(u-u_{h_{\mathrm{ref}}})\big\|^2 + \big\|\mathbb{E}_{\mathrm{Runge}}(u_h)\big\|^2.$$

Further analysis is based on the so-called *saturation assumption*

$$\left\|\nabla(u-u_{h_{\mathrm{ref}}})\right\| \le \lambda \left\|\nabla(u-u_h)\right\|, \quad \lambda \le 1, \tag{2.50}$$

which formalizes a rather natural condition: $u_{h_{\mathrm{ref}}}$ is closer to u than u_h. Usually, the space W_h is constructed by locally based approximations of higher order (e.g., by "bubble-functions"). In this case, the asymptotic relation $\lambda \sim h^q$ is often considered as a justification of the saturation property. However, in general, proving this inequality (with an explicit $\lambda < 1$) is a difficult task.

With the help of (2.50), we obtain

$$\left(1-\lambda^2\right)\left\|\nabla(u-u_h)\right\|^2 = \left\|\mathbb{E}_{\mathrm{Runge}}(u_h)\right\|^2 \le \left\|\nabla(u-u_h)\right\|^2. \tag{2.51}$$

This inequality can be used for error control, provided that λ is known, but even in that case, the computation of $u_{h_{\mathrm{ref}}}$ may be too expensive. Since $V_{h_{\mathrm{ref}}}$ differs from V_h only by the orthogonal complement W_h, the difference $u_{h_{\mathrm{ref}}} - u_h = \widehat{w}_h$ belongs to this subspace. This fact suggests the idea to compute the correction function with the help of a subsidiary problem defined on W_h (instead of $V_{h_{\mathrm{ref}}}$). However, in general, the projection of $u_{h_{\mathrm{ref}}}$ onto V_h does not coincide with u_h and the true projection $\widehat{u}_h$ is unknown. Instead, an approximation of $u_{h_{\mathrm{ref}}}$ is sought in the form $u_h + w_h$, where w_h is defined as an element minimizing the distance from $u_h + \widetilde{w}_h$ to u, which leads to the problem

$$\inf_{w_h \in W_h} \frac{1}{2} \int_\Omega \left|\nabla(u-u_h-w_h)\right|^2 \mathrm{d}x.$$

It is easy to see that the latter problem is equivalent to

$$\inf_{w_h \in W_h} \left\{ \frac{1}{2}\|\nabla w_h\|^2 - \int_\Omega \nabla(u-u_h)\cdot\nabla w_h \,\mathrm{d}x \right\}$$

or

$$\inf_{w_h \in W_h} \left\{ \frac{1}{2}\|\nabla w_h\|^2 - \int_\Omega f w_h \,\mathrm{d}x + \int_\Omega \nabla u_h \cdot \nabla w_h \,\mathrm{d}x \right\}.$$

We arrive at the following problem: Find $\widetilde{w}_h \in W_h$ such that

$$\int_\Omega \nabla \widetilde{w}_h \cdot \nabla w_h \,\mathrm{d}x = \int_\Omega f w_h \,\mathrm{d}x - \int_\Omega \nabla u_h \cdot \nabla w_h \,\mathrm{d}x, \quad \forall w_h \in W_h. \tag{2.52}$$

The following questions rise: how large is the difference between $\widetilde{w}_h$ and $\widehat{w}_h$, and when $\widetilde{w}_h$ can be used instead of $\widehat{w}_h$ (we recall that $u_{h_{\mathrm{ref}}} = \widehat{u}_h + \widehat{w}_h$). To answer them, we first recall that u satisfies the integral relation

$$\int_\Omega \nabla u \cdot \nabla w \,\mathrm{d}x = \int_\Omega f w \,\mathrm{d}x, \quad \forall w \in V, \tag{2.53}$$

u_h and $u_{h_{\mathrm{ref}}}$ are Galerkin solutions, i.e.,

$$\int_\Omega \nabla u_h \cdot \nabla w_h \,\mathrm{d}x = \int_\Omega f w_h \,\mathrm{d}x, \qquad \forall w_h \in V_h,$$
$$\int_\Omega \nabla u_{h_{\mathrm{ref}}} \cdot \nabla w_{h_{\mathrm{ref}}} \,\mathrm{d}x = \int_\Omega f w_{h_{\mathrm{ref}}} \,\mathrm{d}x, \quad \forall w_{h_{\mathrm{ref}}} \in V_{h_{h_{\mathrm{ref}}}} \subset V,$$

and

$$\int_\Omega (\nabla u_{h_{\mathrm{ref}}} - u_h) \cdot \nabla w_h \,\mathrm{d}x = 0, \quad \forall w_h \in V_h. \tag{2.54}$$

Also, we assume that the spaces V_h and W_h are such that the *strengthened Cauchy inequality*

$$\left| \int_\Omega \nabla v_h \cdot \nabla w_h \,\mathrm{d}x \right| \le \gamma \left(\int_\Omega \nabla v_h \cdot \nabla v_h \,\mathrm{d}x \right)^{1/2} \left(\int_\Omega \nabla w_h \cdot \nabla w_h \,\mathrm{d}x \right)^{1/2} \tag{2.55}$$

holds, where $\gamma \in (0, 1)$ is a constant independent of h. In this case,

$$\left\| \nabla (u - u_h) \right\| \le C_{\lambda\gamma} \| \nabla \widetilde{w}_h \|. \tag{2.56}$$

To prove this fact, we argue as follows. By the Galerkin orthogonality (cf. (2.54)), we have

$$\int_\Omega \nabla (u_{h_{\mathrm{ref}}} - u_h) \cdot \nabla (\widehat{u}_h - u_h) \,\mathrm{d}x = 0. \tag{2.57}$$

In view of (2.52),

$$\int_\Omega \nabla \widetilde{w}_h \cdot \nabla \widehat{w}_h \,\mathrm{d}x = \int_\Omega f \widehat{w}_h \,\mathrm{d}x - \int_\Omega \nabla u_h \cdot \nabla \widehat{w}_h \,\mathrm{d}x = \int_\Omega \nabla (u_{h_{\mathrm{ref}}} - u_h) \cdot \nabla \widehat{w}_h \,\mathrm{d}x,$$

whence

$$\int_\Omega \nabla (u_{h_{\mathrm{ref}}} - u_h - \widetilde{w}_h) \cdot (\nabla \widehat{w}_h) \,\mathrm{d}x = 0. \tag{2.58}$$

From (2.57) and (2.58), we conclude that

$$\begin{aligned} 0 &= \int_\Omega \nabla (u_{h_{\mathrm{ref}}} - u_h - \widetilde{w}_h) \cdot \nabla \widehat{w}_h \,\mathrm{d}x + \int_\Omega \nabla (u_{h_{\mathrm{ref}}} - u_h) \cdot \nabla (\widehat{u}_h - u_h) \,\mathrm{d}x \\ &= \int_\Omega \nabla (u_{h_{\mathrm{ref}}} - u_h) \cdot \nabla (\widehat{w}_h + \widehat{u}_h - u_h) \,\mathrm{d}x - \int_\Omega \nabla \widetilde{w}_h \cdot \nabla \widehat{w}_h \,\mathrm{d}x \\ &= \| u_{h_{\mathrm{ref}}} - u_h \|^2 - \int_\Omega \nabla \widetilde{w}_h \cdot \nabla \widehat{w}_h \,\mathrm{d}x. \end{aligned}$$

Thus,

$$\left\| \nabla (u_{h_{\mathrm{ref}}} - u_h) \right\|^2 = \int_\Omega \nabla \widetilde{w}_h \cdot \nabla \widehat{w}_h \,\mathrm{d}x. \tag{2.59}$$

Note that

$$\left\|\nabla(u_{h_{\mathrm{ref}}}-u_h)\right\|^2=\left\|\nabla(u_{h_{\mathrm{ref}}}-\widehat{u}_h)\right\|^2+\left\|\nabla(\widehat{u}_h-u_h)\right\|^2 +2\int_\Omega \nabla(u_{h_{\mathrm{ref}}}-\widehat{u}_h)\cdot\nabla(\widehat{u}_h-u_h)\,\mathrm{d}x.$$

Here $\widehat{u}_h-u_h\in V_h$ and $u_{h_{\mathrm{ref}}}-\widehat{u}_h=\widehat{w}_h\in W_h$, so that we use (2.55) and obtain

$$\begin{aligned}\left\|\nabla(u_{h_{\mathrm{ref}}}-u_h)\right\|^2&\ge\|\nabla\widehat{w}_h\|^2+\left\|\nabla(\widehat{u}_h-u_h)\right\|^2-2\gamma\|\nabla\widehat{w}_h\|\left\|\nabla(\widehat{u}_h-u_h)\right\|\\&\ge\left(1-\gamma^2\right)\|\nabla\widehat{w}_h\|^2.\end{aligned}$$

From this relation and (2.59), we find that

$$\|\nabla\widehat{w}_h\|^2\le\frac{1}{1-\gamma^2}\left\|\nabla(u_{h_{\mathrm{ref}}}-u_h)\right\|^2=\frac{1}{1-\gamma^2}\int_\Omega\nabla\widetilde{w}_h\cdot\nabla\widehat{w}_h\,\mathrm{d}x. \tag{2.60}$$

Thus, we see that the true correction function $\widehat{w}_h$ is subject to $\widetilde{w}_h$:

$$\|\nabla\widehat{w}_h\|\le\frac{1}{1-\gamma^2}\|\nabla\widetilde{w}_h\|. \tag{2.61}$$

Now, we recall that $\|\nabla(u-u_h)\|^2=\|\nabla(u-u_{h_{\mathrm{ref}}})\|^2+\|\nabla(u_h-u_{h_{\mathrm{ref}}})\|^2$ and use (2.59). We have

$$\begin{aligned}\left\|\nabla(u-u_h)\right\|^2&=\left\|\nabla(u-u_{h_{\mathrm{ref}}})\right\|^2+\int_\Omega\nabla\widetilde{w}_h\cdot\widehat{w}_h\,\mathrm{d}x\\&\le\lambda^2\left\|\nabla(u-u_h)\right\|^2+\|\nabla\widetilde{w}_h\|\|\nabla\widehat{w}_h\|\\&\le\lambda^2\left\|\nabla(u-u_h)\right\|^2+\frac{1}{1-\gamma^2}\|\nabla\widetilde{w}_h\|^2.\end{aligned}$$

From here, we conclude that

$$\left\|\nabla(u-u_h)\right\|^2\le\frac{1}{(1-\lambda^2)(1-\gamma^2)}\|\nabla\widetilde{w}_h\|^2, \tag{2.62}$$

which shows that $\|\nabla e\|\simeq\|\nabla\widetilde{w}_h\|$ and motivates using $\|\nabla\widetilde{w}_h\|$ as an error indicator.

2.3 Error Indicators for Goal-Oriented Quantities

Evaluation of approximation errors in terms of special "goal-oriented" quantities is very popular in engineering computations. A consequent exposition can be found in [BR03] and in numerous publications devoted to *goal-oriented* a posteriori error estimates and applications of them to various problems (see, e.g, [BR12, BR96, HRS00, KM10, MS09, OP01, PP98, Ran00, RV10, SO97, SRO07]). In this method,

estimates are derived for the quantity $\langle \ell, u - u_h \rangle$, where ℓ is a given linear functional and u_h is a conforming approximation. In general, ℓ belongs to the dual energy space V_0^*. Typically, ℓ is focused on some special properties of approximate solutions. For example, if ℓ is an integral type functional (e.g., $\ell \in L^2(\Omega)$) localized in a certain subdomain $\omega \subset \Omega$, then $|\langle \ell, u - u_h \rangle|$ characterizes the quality of u_h in ω. A way of evaluating this quantity is based on the following idea, which we discuss with the example of the basic elliptic problem: Find $u \in V_0 := \mathring{H}^1(\Omega)$ such that

$$\int_\Omega A \nabla u \cdot \nabla w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x, \quad \forall w \in V_0, \tag{2.63}$$

where A is a positive definite matrix with bounded coefficients.

Let $A^\star$ be the matrix adjoint to A and u_ℓ the solution of the respective *adjoint* problem

$$\int_\Omega A^\star \nabla u_\ell \cdot \nabla w \, \mathrm{d}x = \langle \ell, w \rangle, \quad \forall w \in V_0. \tag{2.64}$$

From (2.63) and (2.64), it follows that

$$\langle \ell, u - u_h \rangle = \int_\Omega A^\star \nabla u_\ell \cdot \nabla (u - u_h) \, \mathrm{d}x \tag{2.65}$$

$$= \int_\Omega (f u_\ell - A \nabla u_h \cdot \nabla u_\ell) \, \mathrm{d}x =: I_\ell(u_\ell, u_h). \tag{2.66}$$

Hence, $\langle \ell, u - u_h \rangle$ is equal to the functional $I_\ell(u_\ell, u_h)$ and can be easily estimated, provided that u_ℓ is known (we note that finding u_ℓ amounts to solving another boundary value problem having the same complexity as (2.63)). In the majority of cases, u_ℓ is unknown and, therefore, it is replaced by an approximation $u_{\ell\tau}$ computed on an *adjoint mesh* $\mathcal{T}_\tau$ (which does not necessarily coincide with $\mathcal{T}_h$). Then, the non-computable quantity $I_\ell(u_\ell, u_h)$ is approximated by the computable quantity $I_\ell(u_{\ell\tau}, u_h)$.

If $u_{\ell\tau}$ is a sharp approximation of u_ℓ (in general, it should be sharper than u_h), then the quantity $|\mathbb{E}_\ell(u_{\ell\tau}, u_h)| := |I_\ell(u_{\ell\tau}, u_h)|$ serves as an indicator of the goal-oriented error $|\langle \ell, u - u_h \rangle|$. However, getting a sharp approximation of u_ℓ may lead to essential additional expenditures. In order to minimize them, one can apply different modifications (generalizations) of (2.65), which the reader can find in the publications mentioned at the beginning of Sect. 2.3.

2.3.1 Error Indicators Relying on the Superconvergence of Averaged Fluxes in the Primal and Adjoint Problems

Henceforth, for the sake of simplicity we assume that A is a symmetric matrix. We rewrite I_ℓ in the form

$$I_\ell(u_h, u_{\ell\tau}) = I_{\ell 1}(u_h, u_{\ell\tau}) + I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell), \tag{2.67}$$

where

$$I_{\ell 1}(u_h, u_{\ell\tau}) := \int_\Omega (f u_{\ell\tau} - A\nabla u_h \cdot \nabla u_{\ell\tau})\,\mathrm{d}x$$

is a directly computable functional and

$$I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell) := \int_\Omega A(\nabla u - \nabla u_h)\cdot(\nabla u_\ell - \nabla u_{\ell\tau})\,\mathrm{d}x$$

involves unknown u and u_ℓ, i.e., the exact solutions of (2.63) and (2.64), respectively. Note that if u_h is a Galerkin approximation and $\mathcal{T}_\tau$ coincides with $\mathcal{T}_h$, then $I_{\ell 1}(u_h, u_{\ell\tau}) = 0$.

Estimate (2.67) is a source of various indicators. One of them is based on the idea of replacing unknown fluxes

$$p := A\nabla u \quad \text{and} \quad p_\ell := \nabla u_\ell$$

by $G_h p_h$ and $G_\tau p_{\ell\tau}$, where $p_h := A\nabla u_h$, $p_{\ell\tau} := A\nabla u_{\ell\tau}$, and G_h and G_τ are some suitable averaging operators associated with the primal and adjoint meshes, respectively. In [KNR03, NR04], it is proved that under the standard assumptions (which guarantee superconvergence of averaged fluxes computed for the primal and adjoint problems) such a replacement generates errors of a higher order (with respect to h and τ). In view of this fact, the quantity

$$\mathbb{E}_\ell(u_h, u_{\ell\tau}) := I_{\ell 1}(u_h, u_{\ell\tau}) + \mathbb{E}_{\ell 2}(u_h, u_{\ell\tau}), \tag{2.68}$$

where

$$\mathbb{E}_{\ell 2}(u_h, u_{\ell\tau}) := \int_\Omega A^{-1}(G_h p_h - p_h)\cdot(G_\tau p_{\ell\tau} - p_{\ell\tau})\,\mathrm{d}x$$

is used instead of $I_\ell(u_h, u_{\ell\tau})$. However, such an indicator is justified only if both problems (primal and adjoint) are sufficiently regular, so that u_h and $u_{\ell\tau}$ possess superconvergent fluxes. This fact imposes rather obligatory conditions on $\mathcal{T}_\tau$, which may be difficult to satisfy. Typically, the mesh $\mathcal{T}_h$ generated by commonly used solvers is sufficiently regular (so that one can await the superconvergence of p_h, at least in the major part of Ω). For the adjoint mesh $\mathcal{T}_\tau$, such a regularity is difficult to guarantee. Indeed, this mesh should satisfy two conditions, which in fact contradict each other. On the one hand, $\dim V_\tau$ should not significantly exceed $\dim V_h$ (otherwise the adjoint problem is computationally much more expensive than the primal one). On the other hand, $\mathcal{T}_\tau$ should be "sufficiently dense" in the vicinity of ω. This observation motivates attempts at finding other error indicators which are not based on the superconvergence of adjoint fluxes.

2.3.2 Error Indicators Using the Superconvergence of Approximations in the Primal Problem

An error indicator that does not attract the superconvergence of averaged gradients in the adjoint problem was suggested in [NRT08]. The idea behind is to represent the term $I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell)$ in a new form, namely:

$$
\begin{aligned}
&I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell)\\
&\quad := \sum_{T_i \in \mathcal{T}_\tau} \int_{T_i} (\nabla u - \nabla u_h) \cdot (p_\ell - p_{\ell\tau})\, \mathrm{d}x\\
&\quad = \sum_{T_i \in \mathcal{T}_\tau} \left(\int_{T_i} (u_h - u) \mathcal{R}(p_{\ell\tau})\, \mathrm{d}x + \int_{\partial T_i} (u - u_h)(p_\ell - p_{\ell\tau}) \cdot \nu_i\, \mathrm{d}s \right)\\
&\quad = I_{\ell 21}(u_h, p_{\ell\tau}; u) + I_{\ell 22}(u_h, p_{\ell\tau}; u, p_\ell),
\end{aligned}
$$

where ν_i is a unit outward normal to ∂T_i and

$$
\mathcal{R}(p_{\ell\tau}) := \operatorname{div} p_{\ell\tau} + \ell.
$$

Since u, u_h, and p_ℓ are continuous on interelement boundaries, we find that

$$
\begin{aligned}
I_{\ell 22}(u_h, p_{\ell\tau}; u, p_\ell) &= \sum_{T_i \in \mathcal{T}_\tau} \int_{\partial T_i} (u - u_h)(p_\ell - p_{\ell\tau}) \cdot \nu_i\, \mathrm{d}s\\
&= \sum_{E_{ij} \in \mathcal{E}_\tau} \int_{E_{ij}} (u_h - u)[p_{\ell\tau} \cdot \nu_{ij}]_{E_{ij}}\, \mathrm{d}s.
\end{aligned}
$$

Here, $\mathcal{E}_\tau$ is the set of edges in the adjoint mesh, ν_{ij} is the unit normal to the edge E_{ij} (common for T_i and T_j), which is external to T_i if $i < j$. Since u_h and u satisfy the same Dirichlet boundary conditions, $\mathcal{E}_\tau$ contains only internal edges. In this functional, the exact solution of the adjoint problem is *completely excluded*. Therefore, the justification of the estimator is not connected with superconvergence in the adjoint problem, and we may hope that it is insensitive with respect to adjoint mesh structure. To obtain a computable error indicator, in [NRT08] the *superconvergent post-processing* of the function u_h (by the operator Q_τ; see (2.37)–(2.39)) and a *regularization* of the adjoint flux $p_{\ell\tau}$ (which eliminates the jumps $[p_{\ell\tau} \cdot \nu_{ij}]_{E_{ij}}$) were used. Below, the corresponding regularization operator is denoted by G_τ and the Wang projection operator by $\mathcal{W}$. In particular, such an operator can be constructed with the help of Hsieh–Clough–Tocher finite element approximations (see, e.g., [BH81, Cia78b]). Then, $I_{\ell 22} = 0$ and $I_{\ell 21}$ is replaced by

$$
\mathbb{I}_{\ell 21}(u_h, u_{\ell\tau}; u) := \int_\Omega \big(u_h - \mathcal{W}(u_h)\big) \mathcal{R}\big(G_\tau(p_{\ell\tau})\big)\, \mathrm{d}x,
$$

and we arrive at the indicator

$$\langle \ell, u - u_h \rangle \approx \mathbb{E}_\ell(u_h) := I_{\ell 1}(u_h, u_{\ell\tau}) + \int_\Omega \big(u_h - \mathcal{W}(u_h)\big)\mathcal{R}\big(G_\tau(p_{\ell\tau})\big)\,\mathrm{d}x. \quad (2.69)$$

Another representation of $I_{\ell 2}$ leads to a somewhat different error indicator. Let q be a vector-valued function in $H(\Omega, \mathrm{div})$. Then,

$$\begin{aligned} &I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell) \\ &\quad := \int_\Omega (\nabla u - \nabla u_h)(p_\ell - p_{\ell\tau})\,\mathrm{d}x \\ &\quad = \int_\Omega (\nabla u - \nabla u_h)(q - p_{\ell\tau})\,\mathrm{d}x + \int_\Omega (u - u_h)(\mathrm{div}\, q + \ell)\,\mathrm{d}x \\ &\quad = \int_\Omega A^{-1}(p - p_h)\cdot(q - p_{\ell\tau})\,\mathrm{d}x + \int_\Omega (u - u_h)(\mathrm{div}\, q + \ell)\,\mathrm{d}x. \end{aligned} \quad (2.70)$$

In this relation, u_ℓ is excluded from the right-hand side without a regularization of $q_{\ell\tau}$. This relation implies an error indicator if one reconstructs p and u with the help of the recovery operators G_h and $\mathcal{W}$, respectively.

We have

$$\begin{aligned} \langle \ell, u - u_h \rangle \;&\approx\; \mathbb{E}_\ell(u_h, p_{\ell\tau}) \\ &:= I_{\ell 1}(u_h, u_{\ell\tau}) \\ &\quad + \int_\Omega A^{-1}\big(G_h(p_h) - p_h\big)\cdot(q - p_{\ell\tau})\,\mathrm{d}x \\ &\quad + \int_\Omega \big(u_h - \mathcal{W}(u_h)\big)(\mathrm{div}\, q + \ell)\,\mathrm{d}x, \end{aligned} \quad (2.71)$$

where q is an arbitrary vector valued function. If q is equilibrated (or almost equilibrated), then the last term can be ignored and we obtain a simpler indicator

$$\mathbb{E}_\ell(u_h, p_{\ell\tau}) := I_{\ell 1}(u_h, u_{\ell\tau}) + \int_\Omega A^{-1}\big(G_h(p_h) - p_h\big)\cdot(q - p_{\ell\tau})\,\mathrm{d}x. \quad (2.72)$$

It is clear that properties of $\mathbb{E}_\ell(u_h, p_{\ell\tau})$ depend on superconvergence properties of averaged fluxes in the primal problem and on the difference between q and $p_{\ell\tau}$. Numerical examples and asymptotic exactness of the above-introduced indicators are discussed in [NRT08]. One of the examples is presented below.

Example 2.4 We consider the following elliptic type problem:

$$\Delta u + 1 = 0 \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \quad (2.73)$$

and define

$$\langle \ell, u - u_h \rangle = \int_\Omega \ell_\omega(u - u_h)\,\mathrm{d}x, \quad (2.74)$$

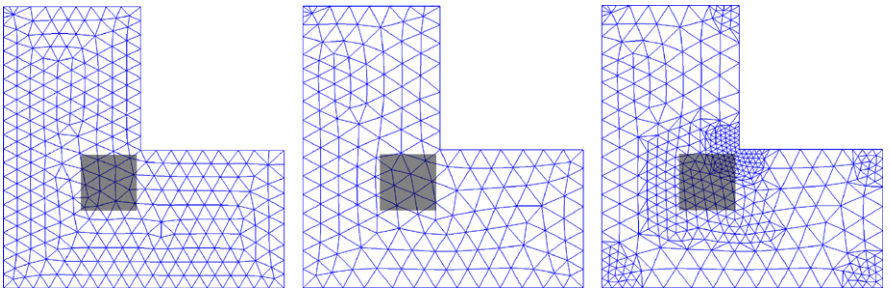

Fig. 2.11 The meshes $\mathcal{T}_1$ (315 nodes) (*left*), $\mathcal{T}_2$ (193 nodes) (*middle*), and $\mathcal{T}_3$ (451 nodes) (*right*) used in the test; the region of interest ω is *shadowed*

where

$$\ell_\omega(x) = \begin{cases} 1, & \text{if } x \in \omega \subset \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{2.75}$$

Both primal and adjoint problems are solved with the help of piecewise linear finite element approximations. As usual, the efficiency index is defined by the relation

$$i_{\text{eff}} := \frac{\mathbb{E}_\ell(u_h)}{|\langle \ell, u - u_h \rangle|}.$$

The primal problem is solved on the mesh $\mathcal{T}_1$ (see Fig. 2.11). It is known that the corresponding exact solution u has singularity in the re-entrant corner. The adjoint problem was solved on $\mathcal{T}_1$, on a rather coarse regular mesh $\mathcal{T}_2$ and on the mesh $\mathcal{T}_3$ adapted to the configuration of the domain ω (shadowed). Numerical results are summarized in Table 2.2, where we compare the indicators (2.68), (2.69), and (2.71). We see that error indicators based on (2.69) and (2.71) demonstrate better performance than (2.68). Other tests in [NRT08] for problems with regular and rather irregular solutions confirm advantages of (2.69) and especially of (2.71).

2.3.3 Error Indicators Based on Partial Equilibration of Fluxes in the Original Problem

First, we prove one principal result, which yields another (in a sense more convenient) form of the functional $I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell)$.

Proposition 2.1 *The term $I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell)$ is equal to the quantity*

$$\int_\Omega A^{-1}\big(\mathsf{P}_{Q_f}(p_h) - p_h\big) \cdot (\eta_\ell - A\nabla u_{\ell\tau})\,\mathrm{d}x := I_{\ell 2}(p_h, u_{\ell\tau}, \eta_\ell), \tag{2.76}$$

where η_ℓ is an arbitrary function in the set

Table 2.2 Efficiency of the estimators in Example 2.4

Indicator	N_{nod}	$\mathcal{T}_\tau$	$I_{\ell 1}$	$\mathbb{E}_{\ell 2}$	$\mathbb{E}_\ell$	i_{eff}
(2.68)	315	$\mathcal{T}_1$	0.00000	0.00264	0.00264	1.58
	193	$\mathcal{T}_2$	0.00119	0.00138	0.00257	1.54
	451	$\mathcal{T}_3$	0.00184	0.00040	0.00223	1.34
Indicator	N_{nod}	$\mathcal{T}_\tau$	$I_{\ell 1}$	$\mathbb{E}_{\ell 21}$	$\mathbb{E}_\ell$	i_{eff}
(2.69)	315	$\mathcal{T}_1$	0.00163	0.00051	0.00213	1.28
	193	$\mathcal{T}_2$	0.00189	0.00064	0.00253	1.51
	451	$\mathcal{T}_3$	0.00181	0.00013	0.00193	1.16
Indicator	N_{nod}	$\mathcal{T}_\tau$	$I_{\ell 1}$	$\mathbb{E}_{\ell 21}$	$\mathbb{E}_\ell$	i_{eff}
(2.71)	315	$\mathcal{T}_1$	0.00108	0.00055	0.00163	0.98
	193	$\mathcal{T}_2$	0.00126	0.00053	0.00179	1.07
	451	$\mathcal{T}_3$	0.00178	0.00000	0.00178	1.06

$$Q_\ell(\Omega) := \big\{q \in H(\Omega, \mathrm{div}) \mid \mathrm{div}\, q + \ell = 0\big\},$$

and the operator $P_{Q_f} : Q \to Q_f$ *is defined by the relation*

$$\big\|q - \mathsf{P}_{Q_f}(q)\big\|_{A^{-1}} \le \|q - q_f\|_{A^{-1}}, \quad \forall q_f \in Q_f. \tag{2.77}$$

Proof Let η_0 be a solenoidal vector-valued function. Then,

$$I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell) = \int_\Omega (\nabla u - \nabla u_h) \cdot (A\nabla u_\ell + \eta_0 - A\nabla u_{\ell\tau})\, \mathrm{d}x.$$

Since $A\nabla u_\ell \in Q_\ell$, we conclude that

$$I_{\ell 2}(u_h, u_{\ell\tau}; u, u_\ell) = \int_\Omega A^{-1}(p - p_h) \cdot (\eta_\ell - A\nabla u_{\ell\tau})\, \mathrm{d}x,$$

where η_ℓ is an arbitrary element of Q_ℓ. From (2.77) with $q = p_h$, it follows that

$$\int_\Omega A^{-1}\big(p_h - \mathsf{P}_{Q_f}(p_h)\big) \cdot \eta_0\, \mathrm{d}x = 0, \quad \forall \eta_0 \in Q_0. \tag{2.78}$$

Since p and $\mathsf{P}_{Q_f}(p_h)$ belong to $Q_f(\Omega)$, we conclude that $(p - \mathsf{P}_{Q_f}(p_h)) \in Q_0$. In view of (2.78), we obtain

$$\begin{aligned} 0 &= \int_\Omega A^{-1}\big(p_h - \mathsf{P}_{Q_f}(p_h)\big) \cdot \big(p - \mathsf{P}_{Q_f}(p_h)\big)\, \mathrm{d}x \\ &= \int_\Omega A^{-1}\big(p_h - p + p - \mathsf{P}_{Q_f}(p_h)\big) \cdot \big(p - \mathsf{P}_{Q_f}(p_h)\big)\, \mathrm{d}x \end{aligned}$$

$$= \int_\Omega (\nabla u_h - \nabla u) \cdot \big(p - \mathsf{P}_{Q_f}(p_h)\big)\,\mathrm{d}x + \big\|p - \mathsf{P}_{Q_f}(p_h)\big\|^2_{A^{-1}}$$
$$= \big\|p - \mathsf{P}_{Q_f}(p_h)\big\|^2_{A^{-1}},$$

and the relation (2.76) follows. □

We note that the term $I_{\ell 2}(p_h, u_{\ell\tau}, \eta_\ell)$ does not contain the exact solution of the adjoint problem. The only difficulty in computing $I_{\ell 2}(p_h, u_{\ell\tau}, \eta_\ell)$ consists of the projection to Q_f. A computable error indicator arises if the exact projection $\mathsf{P}_{Q_f}(p_h)$ is replaced by an approximate $\widetilde{p}_h$ (which can be constructed with the help of a certain quasi-equilibration procedure). Then, we replace $I_{\ell 2}(p_h, u_{\ell\tau}, \eta_\ell)$ by the term

$$\mathbb{I}_{\ell 2}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell) := \int_\Omega A^{-1}(\widetilde{p}_h - p_h) \cdot (\eta_\ell - A\nabla u_{\ell\tau})\,\mathrm{d}x \tag{2.79}$$

and find that

$$\langle \ell, u - u_h\rangle = I_{\ell 1}(u_h, u_{\ell\tau}) + \mathbb{I}_{\ell 2}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell) + \mathcal{R}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell), \tag{2.80}$$

where the first two terms are explicitly computable and the remainder term is defined by the relation

$$\mathcal{R}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell) := \int_\Omega A^{-1}\big(\mathsf{P}_{Q_f}(p_h) - \widetilde{p}_h\big) \cdot (\eta_\ell - A\nabla u_{\ell\tau})\,\mathrm{d}x.$$

An upper bound of this term can be explicitly evaluated.

Proposition 2.2 *The remainder term is subject to the estimate*

$$\big|\mathcal{R}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell)\big|$$
$$\le \left(\|p_h - \widetilde{p}_h\|_{A^{-1}} + \frac{C_{F\Omega}}{c_1}\|\operatorname{div}\widetilde{p}_h + f\|\right)\|\eta_\ell - A\nabla u_{\ell\tau}\|_{A^{-1}} := \mu_{h\tau}. \tag{2.81}$$

Proof We have

$$\big|\mathcal{R}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell)\big| \le \big\|\mathsf{P}_{Q_f}(p_h) - \widetilde{p}_h\big\|_{A^{-1}}\|\eta_\ell - A\nabla u_{\ell\tau}\|_{A^{-1}}.$$

It is easy to see that

$$\big\|\mathsf{P}_{Q_f}(\widetilde{p}_h) - \mathsf{P}_{Q_f}(p_h)\big\|_{A^{-1}} \le \|\widetilde{p}_h - p_h\|_{A^{-1}}.$$

This fact follows from the relation

$$\int_\Omega A^{-1}\big(p_h - \widetilde{p}_h - \mathsf{P}_{Q_f}(p_h) + \mathsf{P}_{Q_f}(\widetilde{p}_h)\big) \cdot \eta_0\,\mathrm{d}x = 0, \quad \forall \eta_0 \in Q_0,$$

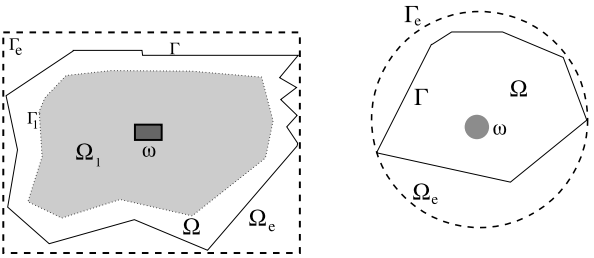

Fig. 2.12 Actual domains Ω and sample domains Ω_e

if we set $\eta_0 = \mathsf{P}_{Q_f}(\widetilde{p}_h) - \mathsf{P}_{Q_f}(p_h) \in Q_0$. Hence,

$$\left\|\mathsf{P}_{Q_f}(p_h) - \widetilde{p}_h\right\|_{A^{-1}} \le \|p_h - \widetilde{p}_h\|_{A^{-1}} + \left\|\mathsf{P}_{Q_f}(\widetilde{p}_h) - \widetilde{p}_h\right\|_{A^{-1}}.$$

Since

$$\left\|\mathsf{P}_{Q_f}(\widetilde{p}_h) - \widetilde{p}_h\right\|_{A^{-1}} = \inf_{q_f \in Q_f} \|\widetilde{p}_h - q_f\|_{A^{-1}} \le \frac{C_{F\Omega}}{c_1} \|\operatorname{div} \widetilde{p}_h + f\|,$$

we arrive at (2.81). □

Remark 2.5 From (2.80) and (2.81), it follows that

$$\begin{aligned} &I_{\ell 1}(u_h, u_{\ell\tau}) + \mathbb{E}_{\ell 2}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell) - \mu_{h\tau} \\ &\quad \le \langle \ell, u - u_h \rangle \le I_{\ell 1}(u_h, u_{\ell\tau}) + \mathbb{E}_{\ell 2}(p_h, \widetilde{p}_h, u_{\ell\tau}, \eta_\ell) + \mu_{h\tau}, \end{aligned}$$

which yields guaranteed error bounds. Certainly these bounds are sensible only if the quantity $\mu_{h\tau}$ is small compared to the first two terms. Since $\mu_{h\tau}$ is directly computable, this requirement can be verified in practical computations.

Finally, we discuss a particular form of the above-introduced error indicator based on solutions of specially constructed *sample problems*. In (2.80), the function $u_{\ell\tau}$ can be replaced by any conforming approximation v_ℓ of u_ℓ (in the derivation of this relation the Galerkin orthogonality of $u_{\ell\tau}$ was not used). Therefore,

$$\langle \ell, u - u_h \rangle = I_{\ell 1}(u_h, v_\ell) + \mathbb{E}_{\ell 2}(p_h, \widetilde{p}_h, v_\ell, \eta_\ell) + \mathcal{R}(p_h, \widetilde{p}_h, v_\ell, \eta_\ell). \tag{2.82}$$

A way of constructing v_ℓ and η_ℓ is to use the exact solution of an adjoint problem for a close domain Ω_e having a simple geometric form. In Fig. 2.12 (left), this domain is presented by a dashed rectangular and ω is the domain (zone) of interest, in which ℓ is nonzero. In Fig. 2.12 (right), this domain is a circle. In the simplest form, the idea of the method is as follows (see [NR09] for more details). Consider the problem (2.63) with the boundary condition $u_0 = 0$. Let $\Omega \subset \Omega_e$. Assume that we know the functions $p_e \in H(\Omega_e, \operatorname{div})$ and $u_e \in V_0(\Omega_e)$ such that

$$\int_{\Omega_e} p_e \cdot \nabla w \, dx = \int_{\Omega_e} \ell w \, dx, \quad \forall w \in V_0(\Omega_e), \tag{2.83}$$

and

$$\int_{\Omega_e} (p_e - A\nabla u_e) \cdot \eta \, \mathrm{d}x = 0, \quad \forall \eta \in Q(\Omega_e). \tag{2.84}$$

It is easy to see that u_e and p_e represent the solution of the adjoint problem in Ω_e and the respective flux. If Ω_e has a simple form (e.g., it is a rectangular, a cube or a sphere) then these functions can be found either analytically or numerically with a high accuracy (since Ω_e has a simple form, sharp approximations can be constructed with the help of, e.g., spectral methods or other methods adapted to such type domains).

Let ϕ be a continuous function such that

$$\phi = 0 \quad \text{on } \Gamma, \qquad 0 \le \phi(x) \le 1 \quad \text{in } \Omega,$$
$$\phi(x) = 1 \quad \text{in } \Omega_1, \qquad \nabla\phi \in L^\infty(\Omega, \mathbb{R}^d).$$

Set $\eta_\ell = p_e$ and $v_\ell = \phi u_e$. Since $\phi u_e \in V_0(\Omega)$, we can use it in the indicator. Then, $A\nabla v_\ell = \phi A\nabla u_e + u_e A\nabla\phi$, $\eta_\ell = A\nabla u_e$ and the remainder term has the following form:

$$\mathcal{R}(p_h, \widetilde{p}_h, v_\ell, \eta_\ell) := \int_{\Omega\setminus\Omega_1} A^{-1}\big(\mathsf{P}_{Q_f}(p_h) - \widetilde{p}_h\big) \cdot \big((1-\phi)p_e - u_e A\nabla\phi\big) \, \mathrm{d}x.$$

If the flux $\widetilde{p}_h$ is almost equilibrated in the boundary strip $\Omega \setminus \Omega_1$, then the remainder term is very small so that the two first computable terms in (2.82) dominate and represent the major part of $\langle \ell, u - u_h \rangle$. Therefore, the quality of the error indicator

$$\langle \ell, u - u_h \rangle \approx \mathbb{E}_\ell(u_h, \widetilde{p}_h, v_\ell, \phi, \Omega_e) := I_{\ell 1}(u_h, v_\ell) + \mathbb{E}_{\ell 2}(p_h, \widetilde{p}_h, v_\ell, \eta_\ell)$$

depends mainly on the equilibration properties of $\widetilde{p}_h$ in the boundary strip.

http://www.springer.com/978-94-007-7580-0

Accuracy Verification Methods
Theory and Algorithms
Mali, O.; Neittaanmäki, P.; Repin, S.
2014, XIII, 355 p. 75 illus., Hardcover
ISBN: 978-94-007-7580-0