Chapter

5 The finite element method in dimension two

It should already be clear that there is no difference between dimensions from the variational viewpoint. The same goes for the abstract part of variational approximations. The difference lies in the description of the finite dimensional approximation spaces. The FEM in any dimension is based on the same principle as in one dimension, that is to say, we consider spaces of piecewise polynomials of low degree, with lots of pieces for accuracy. Now things are right away quite different, and actually considerably more complicated, since polynomials have several variables, and open sets are much more varied than in dimension one.

5.1 Meshes in 2d

Let Ω be an open connected subset of $\mathbb{R}^2$. The idea is to cover Ω with a finite number of sets T_k of simple shape, $\Omega = \bigcup_{k=1}^{N_{\mathscr{T}}}$, with $\mathscr{T} = \{T_k\}$ and $N_{\mathscr{T}} = \operatorname{card} \mathscr{T}$. This decomposition will be used to decompose integrals into sums, thus we impose that $\operatorname{meas}(T_k \cap T_{k'}) = 0$ for $k \neq k'$. We will only consider two cases :

- The T_k are rectangles (and thus Ω is a union of rectangles). For definiteness, the sides of the rectangles will be parallel to the coordinate axes, without loss of generality.
- The T_k are triangles (and Ω is a polygon).

Such a structure will be called a *triangulation* (even in the case of rectangles...) or *mesh* on Ω. The T_k are called the elements, their sides the edges and their vertices are mesh nodes[1]. The fact that Ω must be a union of rectangles or more generally a polygon can appear to be unduly restrictive. There are however ways of going around this restriction and to cover very general domains. From now on, Ω will always be a polygon in $\mathbb{R}^2$.

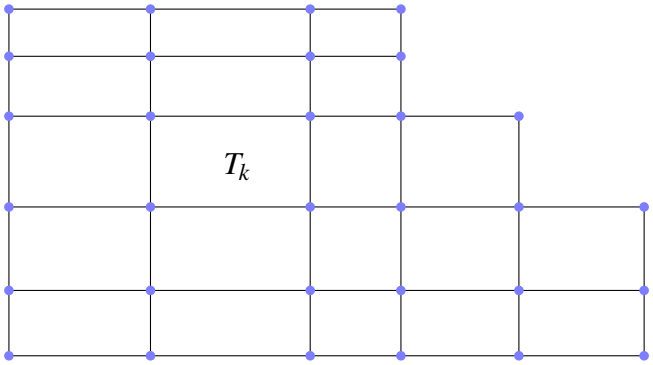

Figure 1. A rectangular mesh.

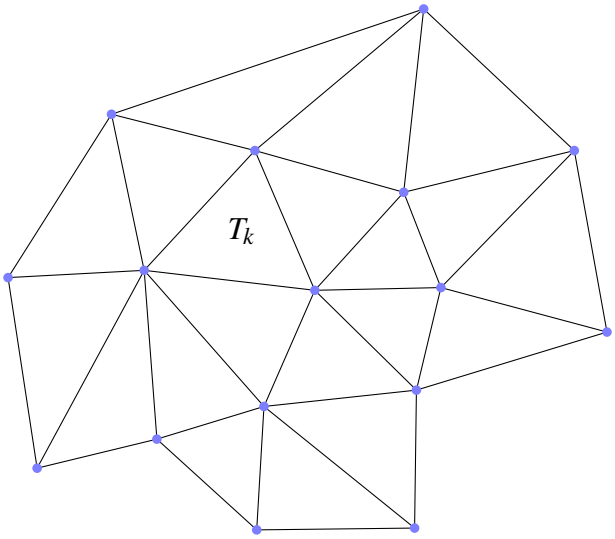

Figure 2. A triangular mesh.

Given a mesh $\mathscr{T}$, we let $h(T_k) = \operatorname{diam} T_k = \sup_{x,y\in T_k} \|x-y\|$ and

$$h = \max_{T_k\in\mathscr{T}} h(T_k).$$

The scalar h is called the mesh size. The approximation spaces will thus be of the form

$$V_h = \{v \in V; v_{|T_k} \text{ is a low degree polynomial}\},$$

[1]There are often additional mesh nodes, as we will see later.

to be made more precise later. This is case of bad traditional notation, since V_h does not depend solely on h, but on the whole mesh, of which h is but one characteristic length. Accordingly, when we say $h \to 0$, this means that we are given a sequence $\mathscr{T}_n$ of meshes whose mesh size tends to 0 when $n \to +\infty$. Naturally in practice, computer calculations are made on one or a small number of meshes. The convergence $h \to 0$ is only for theoretical purposes.

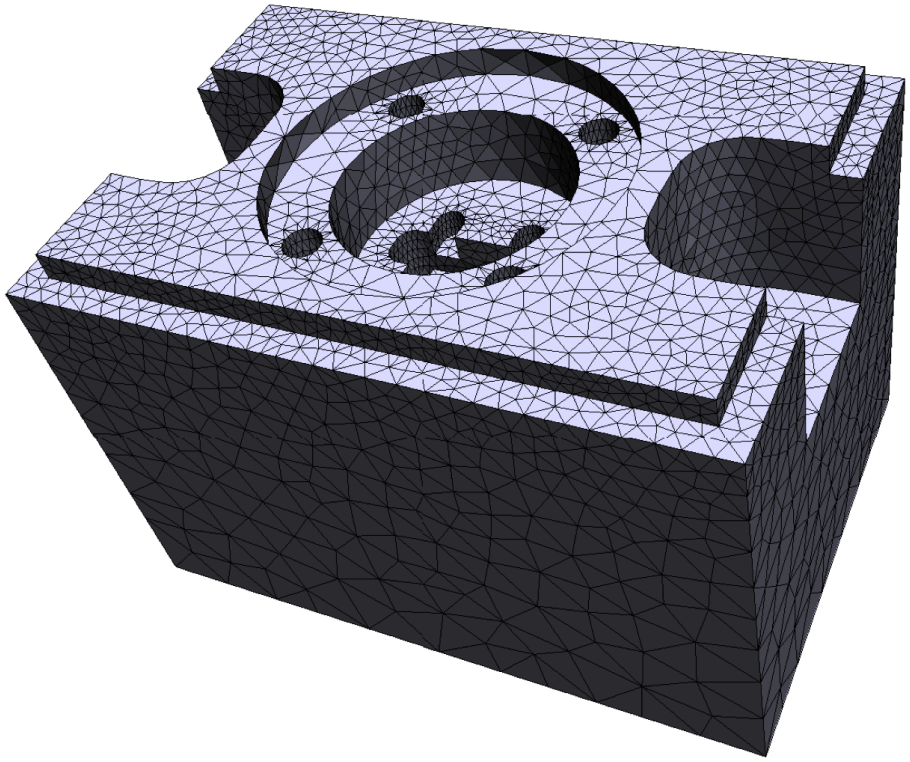

Figure 3. A real life mesh in 3d. The elements are tetrahedra that fill the volume, we just see triangular faces of those tetrahedra the touch the boundary. We will not talk about 3d problems in these notes, but as you may have noticed, most real life problems occur in 3d.

In order to be of use, a triangulation must satisfy a certain number of properties.

Definition 5.1.1 *A mesh is said to be* admissible *if*

i) For all $k \neq k'$, $T_k \cap T_{k'}$ is either empty, or consists of exactly one node or of one entire edge.

ii) No T_k is of zero measure.

Condition ii) means that no triangle or rectangle is degenerated, that is to say that its vertices are not aligned. Condition i) is easier to understand in terms of

which situations it precludes. For instance, any one of the following three cases is forbidden:

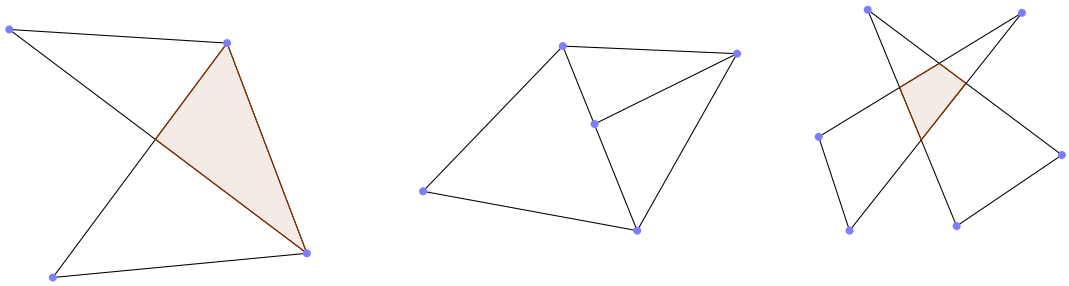

Figure 4. Forbidden meshes according to rule i).

For any triangle T, let $\rho(T)$ be the diameter of the inscribed circle (the center of the inscribed circle is called the incenter and is located at the intersection of the three internal angle bissectors, see Figure 5 below).

Definition 5.1.2 *Let $\mathcal{T}_h$ be a sequence of triangular meshes whose mesh size tends to 0. We say that the sequence is* regular family *is there exists a constant $C > 0$ such that for all h,*

$$\max_{T \in \mathcal{T}_h} \frac{h(T)}{\rho(T)} \leq C.$$

For a sequence of meshes not to be regular means that there are smaller and smaller triangles that become arbitrarily flat. Of course, the definition needs an infinite sequence of meshes to make sense. A similar condition for rectangular meshes is that the ratio of the longer side by the smaller side of each rectangle remains bounded from above. This definition is needed for convergence results.

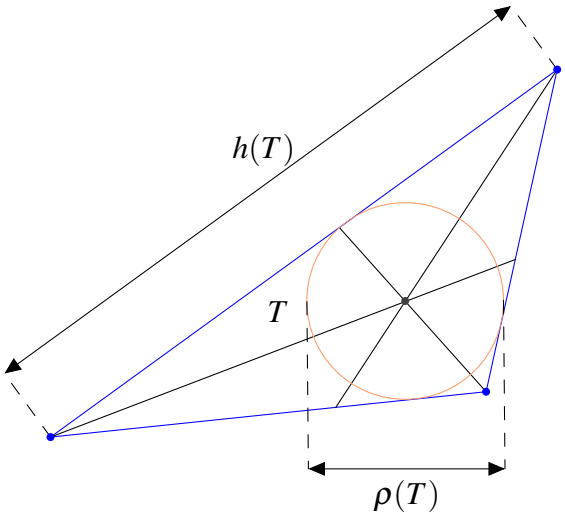

Figure 5. Triangle incircle and diameter.

Let us now give a general purpose proposition on piecewise regular functions on a mesh.

Proposition 5.1.1 *Let $\mathscr{T}$ be an admissible mesh on Ω. Define*

$$X_h = \{v \in C^0(\bar{\Omega}); v_{|T_k} \in C^1(\bar{T}_k) \text{ for all } T_k \in \mathscr{T}\}.$$

Then we have $X_h \subset H^1(\Omega)$ and $\partial_i v = \sum_{k=1}^{N_\mathscr{T}} \partial_i(v_{|T_k})\mathbf{1}_{T_k}$ for all $v \in X_h$.

Proof. Let $v \in X_h$. Clearly, $v \in L^2(\Omega)$ and we just need to compute its partial derivatives in the sense of distributions. Let us thus take an arbitrary function $\varphi \in \mathscr{D}(\Omega)$. We have

$$\langle \partial_i v, \varphi \rangle = -\langle v, \partial_i \varphi \rangle = -\int_\Omega v \partial_i \varphi \, dx = -\sum_{k=1}^{N_\mathscr{T}} \int_{T_k} v \partial_i \varphi \, dx.$$

Now v is C^1 on each $\bar{T}_k$, therefore we can use the integration by parts formula to obtain

$$\begin{aligned} -\int_{T_k} v \partial_i \varphi \, dx &= \int_{T_k} \partial_i(v_{|T_k}) \varphi \, dx - \int_{\partial T_k} v n_{k,i} \varphi \, d\Gamma \\ &= \int_\Omega \partial_i(v_{|T_k}) \mathbf{1}_{T_k} \varphi \, dx - \int_{\partial T_k} v n_{k,i} \varphi \, d\Gamma, \end{aligned}$$

where n_k denotes the unit exterior normal vector to ∂T_k. Note that since $v \in C^0(\bar{\Omega})$ there is no need to take the restriction of v to T_k in the boundary term. Summing on all triangles or rectangles, we obtain

$$\langle \partial_i v, \varphi \rangle = \int_\Omega \Big(\sum_{k=1}^{N_\mathscr{T}} \partial_i(v_{|T_k}) \mathbf{1}_{T_k} \Big) \varphi \, dx - \sum_{k=1}^{N_\mathscr{T}} \int_{\partial T_k} v n_{k,i} \varphi \, d\Gamma.$$

Now, each ∂T_k is composed of three or four edges and there are two cases:

1. Either the edge is included in $\partial\Omega$ and in this case $\varphi = 0$, hence the integral vanishes.

2. Or the edge is included in Ω (except possibly one node) and in this case, by condition i) of mesh admissibility, this edge is the intersection of exactly two elements T_k and $T_{k'}$. The two integrals corresponding to this edge cancel out each other, since $v\varphi$ is continuous, it takes the same value on $\partial T_k \cap \partial T_{k'}$ as seen from either side, and $n_k = -n_{k'}$, see Figure 6 below.

Finally, we see that

$$\sum_{k=1}^{N_\mathscr{T}} \int_{\partial T_k} v n_{k,i} \varphi \, d\Gamma = 0$$

and since the function $\sum_{k=1}^{N_\mathscr{T}} \partial_i(v_{|T_k}) \mathbf{1}_{T_k}$ is bounded, it is also in $L^2(\Omega)$. □

Remark 5.1.1 The above proof shows that in fact $X_h \subset W^{1,\infty}(\Omega)$. □

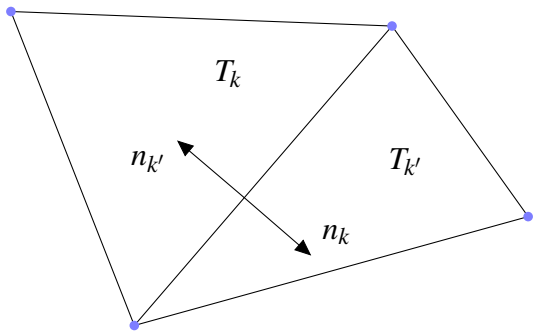

Figure 6. Pairwise cancellation of boundary integrals.

5.2 Rectangular Q_1 finite elements

We start over with the model problem

$$\begin{cases} -\Delta u + cu = f \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{5.1}$$

with $f \in L^2(\Omega)$, $c \in L^\infty(\Omega)$, $c \geq 0$ and $\Omega =]0,1[\times]0,1[$. The variational formulation is of course $V = H_0^1(\Omega)$, $a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + cuv)\,dx$ and $\ell(v) = \int_\Omega fv\,dx$.

Let us be given two integers N_1 and N_2, define $h_1 = \frac{1}{N_1+1}$ and $h_2 = \frac{1}{N_2+1}$. We define a rectangular mesh on Ω by letting

$$R_k = \{(x_1,x_2); ih_1 \leq x_1 \leq (i+1)h_1, jh_2 \leq x_2 \leq (j+1)h_2, \\ i = 0,\dots,N_1, j = 0,\dots,N_2\}.$$

The elements are rectangles of sides h_1 and h_2, parallel to the coordinate axes. There are $N_{\mathscr{T}} = (N_1+1)(N_2+1)$ elements. The mesh size is $h = \sqrt{h_1^2+h_2^2}$. Actually, since $\max(h_1,h_2) \leq h \leq \sqrt{2}\max(h_1,h_2)$, we may as well take $h = \max(h_1,h_2)$. The inscribed circle has diameter $\min(h_1,h_2)$, so the regularity requirement for a family of such meshes would be that $\frac{\max(h_1,h_2)}{\min(h_1,h_2)} \leq C$, or roughly speaking that N_1 and N_2 be of the same order of magnitude.

The mesh nodes are the points (ih_1, jh_2), $i = 0,\dots,N_1+1$, , $j = 0,\dots,N_2+1$. There is a total of $N_t = (N_1+2)(N_2+2)$ nodes, including $2(N_1+1)+2(N_2+1) = 2(N_1+N_2)+4 = N_b$ boundary nodes located on $\partial\Omega$ and $N_i = N_1N_2$ internal nodes located in Ω. Of course, $N_t = N_i + N_b$. We will talk about numbering issues later (numbering of nodes, numbering of elements).

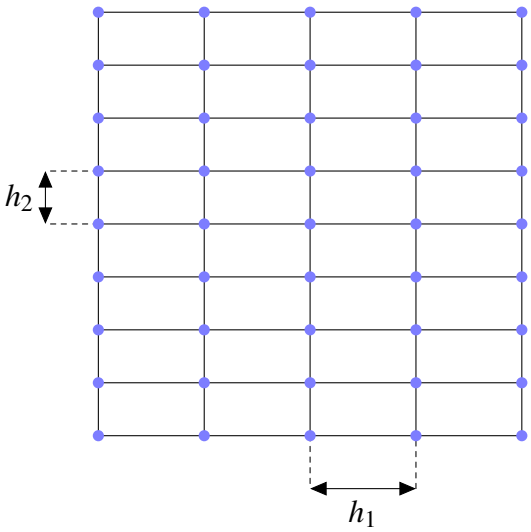

Figure 7. A rectangular mesh on $\Omega =]0,1[^2$, 32 elements, 45 nodes.

Let us now talk about the discrete approximation space. We first need to recall a few facts about the algebra of polynomials in several variables. First of all, there are several notions of degree for such polynomials. The *total degree* of a nonzero monomial in two variables aX^nY^m is $n+m$ (and the obvious generalization for more variables, that we will not use here). The total degree of a polynomial is the maximum total degree of its monomials. The *partial degree* of the same monomial is $\max(n,m)$. The partial degree of a polynomial is the maximum partial degree of its monomials. Since we are working on an infinite number field, $\mathbb{R}$, we can identify polynomials and polynomial functions on an open set of $\mathbb{R}^2$. We will perform this identification freely.

There are two families of spaces of polynomials that will be of interest to us.

Definition 5.2.1 *For each $k \in \mathbb{N}$, we denote by P_k the space of polynomials of total degree less or equal to k and by Q_k the space of polynomials of partial degree less or equal to k.*

Both spaces obviously are vector spaces. It is a fun(?) exercise in algebra to establish that $\dim P_k = \frac{(k+1)(k+2)}{2}$ and $\dim Q_k = (k+1)^2$.

Since the total degree of a polynomial as always larger than its partial degree, it follows that $P_k \subset Q_k$. Moreover, as the Q_k monomial X^kY^k is clearly of the highest possible total degree, we also have $Q_k \subset P_{2k}$. The only value of k for which these spaces coincide is thus $k = 0$, with only constant polynomials. The space P_1 is the space of affine functions

$$P_1 = \{p; p(x) = a_0 + a_1x_1 + a_2x_2\}$$

and the space Q_1 is described in terms of its canonical basis

$$Q_1 = \{p; p(x) = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2\}.$$

We can now introduce the approximation spaces. We start with a version without boundary conditions

$$W_h = \{v_h \in C^0(\bar{\Omega}); \forall R_k \in \mathscr{T}, v_{h|R_k} \in Q_1\}, \tag{5.2}$$

and a subspace thereof that includes homogeneous Dirichlet conditions

$$V_h = \{v_h \in W_h; v_h = 0 \text{ on } \partial\Omega\}. \tag{5.3}$$

The space W_h thus consists of globally continuous functions the restriction of which to each element coincides with one Q_1 polynomial per element. It is the same idea as in dimension 1. Since Q_1 polynomials are of course of class C^1, Proposition 5.1.1 immediately implies

Proposition 5.2.1 *We have $W_h \subset H^1(\Omega)$ and $V_h \subset H^1_0(\Omega)$.*

We now establish interpolation results for Q_1 polynomials and piecewise Q_1 functions. We start with a uniqueness result.

Proposition 5.2.2 *A function of W_h is uniquely determined by its values at the nodes of the mesh.*

Proof. A function v_h in W_h is uniquely determined by the values it takes in each rectangular element, that is to say by the collection of Q_1 polynomials that give its values in each element. It is thus sufficient to argue element by element. Let R be such an element and $S^i = (x_1^i, x_2^i)$ be its four vertices numbered counterclockwise starting from the lower left corner. We have $h_1 = x_1^i - x_1^1$ for $i = 2,3$ and $h_2 = x_2^i - x_2^1$ for $i = 3,4$. Since v_h is equal to a Q_1 polynomial in R, there exists four constants α_j, $j = 1,\ldots,4$ such that

$$v_h(x) = \alpha_1 + \alpha_2(x_1 - x_1^1) + \alpha_3(x_2 - x_2^1) + \alpha_4(x_1 - x_1^1)(x_2 - x_2^1).$$

Let us express the values of v_h at the four vertices.

$$\begin{aligned} v_h(S^1) &= \alpha_1 \\ v_h(S^2) &= \alpha_1 + \alpha_2(x_1^2 - x_1^1) + \alpha_3(x_2^2 - x_2^1) + \alpha_4(x_1^2 - x_1^1)(x_2^2 - x_2^1) \\ &= \alpha_1 + \alpha_2 h_1 \end{aligned}$$

since $x_2^2 = x_2^1$,

$$v_h(S^3) = \alpha_1 + \alpha_3 h_2$$
$$v_h(S^4) = \alpha_1 + \alpha_2 h_1 + \alpha_3 h_2 + \alpha_4 h_1 h_2.$$

This is a 4×4 linear system in the four unknowns α_j which we can rewrite in matrix form

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & h_1 & 0 & 0 \\ 1 & 0 & h_2 & 0 \\ 1 & h_1 & h_2 & h_1 h_2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} v_h(S^1) \\ v_h(S^2) \\ v_h(S^3) \\ v_h(S^4) \end{pmatrix}.$$

The determinant of the triangular matrix above is $h_1^2 h_2^2 \neq 0$, hence the system has one and only one solution for any given vertex values for v_h. Therefore, we have the announced uniqueness. □

We also have an existence result.

Proposition 5.2.3 *For any set of values assigned to the nodes of the mesh, there exists an element v_h of W_h that takes these values at the nodes.*

Proof. The previous proof shows that four values for the four vertices of an element determine one and only one Q_1 polynomial that interpolates these values at the element vertices. Therefore, if we are given a set of values for each node in the mesh, this set determines one Q_1 polynomial per element. The only thing to be checked is that these polynomials combine into a globally C^0 function. Indeed, discontinuities could arise at internal edges, those that are common to two elements. We have to see that this is not the case.

Let us thus consider the following situation, without loss of generality:

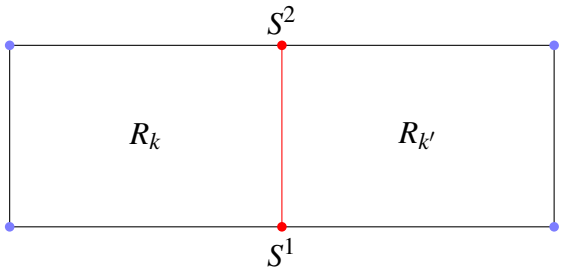

Figure 8. Continuity across an internal edge $[S^1, S^2]$.

We have two Q_1 polynomials q and q' such that $q(S^1) = q'(S^1)$ and $q(S^2) = q'(S^2)$. Writing

$$q(x) = \alpha_1 + \alpha_2(x_1 - x_1^1) + \alpha_3(x_2 - x_2^1) + \alpha_4(x_1 - x_1^1)(x_2 - x_2^1)$$

and

$$q'(x) = \beta_1 + \beta_2(x_1 - x_1^1) + \beta_3(x_2 - x_2^1) + \beta_4(x_1 - x_1^1)(x_2 - x_2^1)$$

as before, we obtain two equations

$$\alpha_1 = \beta_1, \quad \alpha_1 + \alpha_3 h_2 = \beta_1 + \beta_3 h_2,$$

from which it immediately follows that

$$\alpha_1 = \beta_1, \quad \alpha_3 = \beta_3.$$

Now, any point on the segment $[S^1, S^2]$ may be written as $\lambda S^1 + (1-\lambda)S^2$, with $\lambda \in [0,1]$. Therefore, on this segment, we have

$$q(x) = \alpha_1 + \alpha_3 h_2(1-\lambda) = \beta_1 + \beta_3 h_2(1-\lambda) = q'(x).$$

Consequently, the function defined by $q(x)$ if $x \in R_k$, $q'(x)$ if $x \in R_{k'}$ is continuous on $R_k \cup R_{k'}$. □

Remark 5.2.1 The global continuity follows from the fact that Q_1 polynomials are affine on any segment that is parallel to the coordinate axes. If two such polynomials coincide at two points of such a segment, they then coincide on the whole strait line going through the two points. Of course, they are not affine on any segment that is not parallel to the coordinate axes. □

Corollary 5.2.1 *Let S^j, $j = 1, \ldots, N_t$, be a numbering of the mesh nodes. There exists a unique family $(w_h^i)_{i=1,\ldots,N_t}$ such that $w_h^i \in W_h$ and $w_h^i(S^j) = \delta_{ij}$. This family is a basis of W_h, which is of dimension N_t, and for all $v_h \in W_h$, we have*

$$v_h = \sum_{i=1}^{N_t} v_h(S^i) w_h^i. \tag{5.4}$$

Proof. The existence and uniqueness of w_h^i follow readily from Propositions 5.2.2 and 5.2.3, since δ_{ij} for $1 \leq i, j \leq N_t$ is a set of values for all the nodes.

These Propositions also show that the linear mapping $W_h \to \mathbb{R}^{N_t}$, $v_h \mapsto (v_h(S^i))$ is an isomorphism, hence $\dim W_h = N_t$. The family $(w_h^i)_{i=1,\ldots,N_t}$ is the inverse image of the canonical basis of $\mathbb{R}^{N_t}$ by this isomorphism, therefore it is a basis of W_h. Finally, every element v_h of W_h is decomposed on this basis as $v_h = \sum_{i=1}^{N_t} \lambda_i w_h^i$, so that taking $x = S^j$, we obtain

$$v_h(S^j) = \sum_{i=1}^{N_t} \lambda_i w_h^i(S^j) = \sum_{i=1}^{N_t} \lambda_i \delta_{ij} = \lambda_j$$

which proves equation (5.4). □

We can now characterize the elements of V_h, *i.e.*, those functions of W_h that vanish on $\partial\Omega$.

Corollary 5.2.2 *Assume, for convenience only, that the nodes S^j, $j = 1,\dots,N_i$ are the interior nodes. Then the family $(w_h^i)_{i=1,\dots,N_i}$ is a basis of V_h, which is of dimension N_i.*

Proof. If a function is in V_h, then $v_h(S^j) = 0$ for $j > N_i$. Therefore, we necessarily have

$$v_h = \sum_{i=1}^{N_i} v_h(S^i) w_h^i.$$

It remains to be seen that $w_h^i \in V_h$ for $i \leq N_i$. This is clear, since these functions vanish on all boundary nodes. Hence by the same token as before, they vanish on all segments joining boundary nodes, and the whole boundary $\partial\Omega$ is composed of such segments. □

Remark 5.2.2 The functions w_h^i are called the basis functions for Q_1 Lagrange interpolation. The linear mappings $v_h \mapsto v_h(S^j)$ are again called the *degrees of freedom*.

It is easy to see that the support of w_h^i is composed of the four elements surrounding S^i when S^i is an interior node, two elements when it is a boundary node, but not a vertex of Ω, and just one element when it is one of the four vertices of Ω.

The graph of a basis function corresponding to an interior node over its support, is made of four hyperbolic paraboloid pieces that look like a tent,[2] see Figure 9 below. □

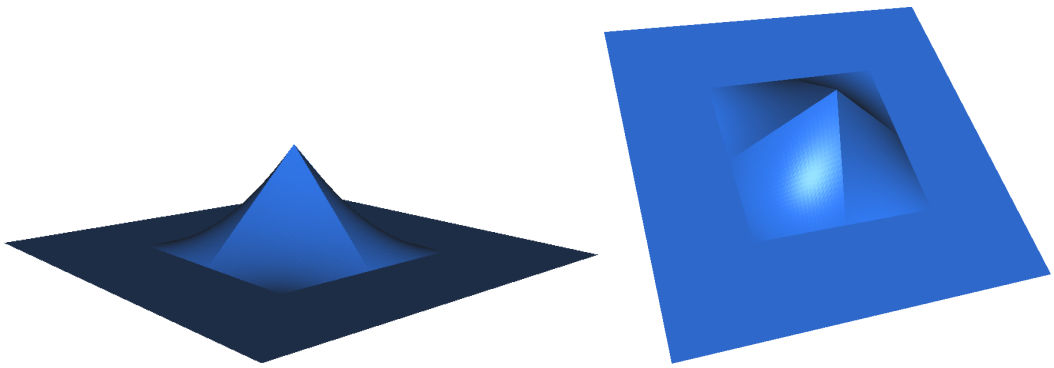

Figure 9. Two views of a Q_1 basis function for an interior node.

[2]Which is why they are sometimes called tent-functions.

Let us now show pictures of elements of V_h.

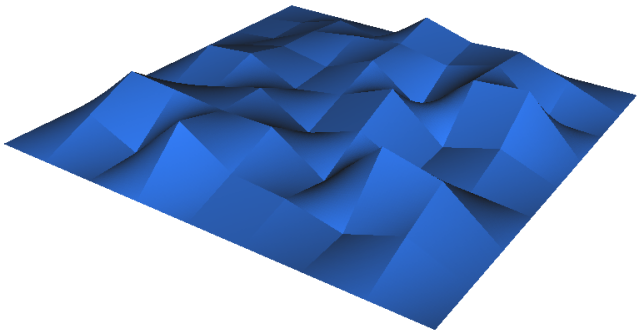

Figure 10. The graph of a random element of V_h. The fact that functions in V_h are piecewise affine on segments parallel to the coordinate axes is apparent, see Section 5.4.

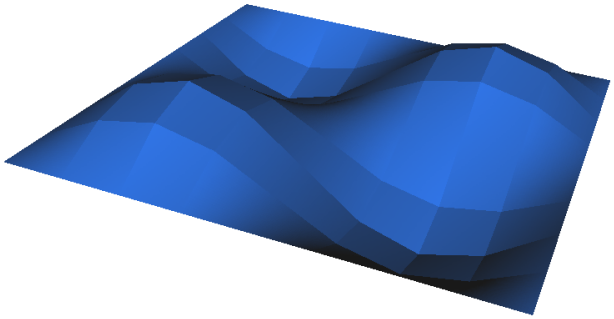

Figure 11. The graph of the V_h-interpolate of a simple $\sin(\pi x_1)\sin(\pi x_2)$ function.

5.3 Convergence and error estimate for the Q_1 FEM

The approximation space V_h is finite-dimensional, therefore closed, hence Céa's lemma applies. We thus need to estimate such quantities as $\|u - \Pi_h u\|_{H^1(\Omega)}$ where Π_h is some interpolation operator with values in V_h in order to obtain an error estimate and prove convergence. We now encounter a new difficulty, which is that H^1 functions are not continuous in two dimensions, therefore, the nodal values of u a priori do not make any sense and it is not possible to perform Lagrange interpolation on H^1.

We will thus need to make regularity hypotheses. We will admit the following particular case of the Sobolev embedding theorems, which is valid in dimension two.

Theorem 5.3.1 *We have* $H^2(\Omega) \hookrightarrow C^0(\bar{\Omega})$.

With this theorem at hand, we can V_h-interpolate H^2 functions.

Let us thus be given a regular family of meshes, that we index by $h = \max(h_1, h_2)$, regularity meaning here that there exists a constant C such that $\frac{\max(h_1,h_2)}{\min(h_1,h_2)} \leq C$. Let u be the solution of problem (5.1) in variational form and u_h its variational approximation on V_h. We will prove the following convergence and error estimate theorem.

Theorem 5.3.2 *There exists a constant C such that, if $u \in H^2(\Omega)$, we have*

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega))}. \tag{5.5}$$

The constant C is naturally not the same constant as a couple of lines higher. Actually, the proof of Theorem 5.3.2 will be broken into a series of lemmas, and constants C will come up that generally vary from line to line. This is what is called a generic constant... The important thing is not their value, but that they do not depend on any of the other quantities that appear, in this specific case, h and u.

Let $\widehat{R} = [0,1] \times [0,1]$ be the *reference rectangle*[3] or reference element. We let $\widehat{\Pi}$ denote the Q_1 interpolation operator on the four vertices of $\widehat{R}$. Let us begin our series of lemmas.

Lemma 5.3.1 *There exists a constant C such that, for all $\hat{v} \in H^2(\widehat{R})$*

$$\|\hat{v} - \widehat{\Pi}\hat{v}\|_{H^1(\widehat{R})} \leq C\|\nabla^2 \hat{v}\|_{L^2(\widehat{R})}. \tag{5.6}$$

Proof. We note that $P_1 \subset Q_1$, thus for all $p \in P_1$, we have $\widehat{\Pi}p = p$. Therefore

$$\|\hat{v} - \widehat{\Pi}\hat{v}\|_{H^1(\widehat{R})} = \|\hat{v} - p - \widehat{\Pi}(\hat{v} - p)\|_{H^1(\widehat{R})} \leq \|I - \widehat{\Pi}\|_{\mathscr{L}(H^2;H^1)}\|\hat{v} - p\|_{H^2(\widehat{R})}$$

for all $\hat{v} \in H^2(\widehat{R})$, $p \in P_1$. Consequently

$$\|\hat{v} - \widehat{\Pi}\hat{v}\|_{H^1(\widehat{R})} \leq C \inf_{p \in P_1} \|\hat{v} - p\|_{H^2(\widehat{R})} = C\|\hat{v} - P\hat{v}\|_{H^2(\widehat{R})}$$

where P denotes the H^2 orthogonal projection onto P_1.

Let us now show that there is a constant C such that

$$\|\hat{v} - P\hat{v}\|_{H^2(\widehat{R})} \leq C\|\nabla^2 \hat{v}\|_{L^2(\widehat{R})},$$

which will complete the proof of the Lemma. We argue by contradiction and assume there is no such constant C. In this case, there exists a sequence $\hat{v}_n \in H^2(\widehat{R})$ such that

$$\|\hat{v}_n - P\hat{v}_n\|_{H^2(\widehat{R})} = 1 \quad \text{and} \quad \|\nabla^2 \hat{v}_n\|_{L^2(\widehat{R})} \to 0,$$

[3] Ok, it's a square, and unluckily it happens to look a lot like Ω, although there is no conceptual connection between the two.

when $n \to +\infty$. Let us set $\hat{w}_n = \hat{v}_n - P\hat{v}_n$, which belongs to $P_1^\perp$. The second derivatives of a P_1 polynomial vanish, so that $\nabla^2 \hat{w}_n = \nabla^2 \hat{v}_n$. By Rellich's compact embedding theorem, there exists a sequence, still denoted $\hat{w}_n$ and a $\hat{w} \in H^1(\widehat{R})$ such that $\hat{w}_n \to \hat{w}$ in $H^1(\widehat{R})$. Then, the condition $\|\nabla^2 \hat{w}_n\|_{L^2(\widehat{R})} \to 0$ shows that $\hat{w}_n$ is a Cauchy sequence in $H^2(\widehat{R})$, which is complete. Hence, $\hat{w} \in H^2(\widehat{R})$ and $\hat{w}_n \to \hat{w}$ in $H^2(\widehat{R})$ as well. Now, the space $P_1^\perp$ is a H^2 orthogonal, hence is closed in $H^2(\widehat{R})$, from which is follows that $\hat{w} \in P_1^\perp$. On the other hand, we have $\nabla^2 \hat{w} = 0$, so that $\hat{w} \in P_1$. Consequently, $\hat{w} = 0$ and $\hat{w}_n \to 0$ in $H^2(\widehat{R})$, which contradicts $\|\hat{w}_n\|_{H^2(\widehat{R})} = 1$. The proof is complete. □

Remark 5.3.1 The above proof does not really use Q_1-interpolation, but only P_1-interpolation. It would thus equally apply for triangular elements, which we will discuss later. □

We now perform a change of variable between the reference element $\widehat{R}$ and a generic element R_k of the mesh.

Lemma 5.3.2 *Let R_k be an element of the mesh. There exists a unique affine bijective mapping F_k such that $F_k(\widehat{R}) = R_k$ and that maps the vertices counted counterclockwise from the lower left corner to their counterparts in R_k.*

Proof. Consider the following figure:

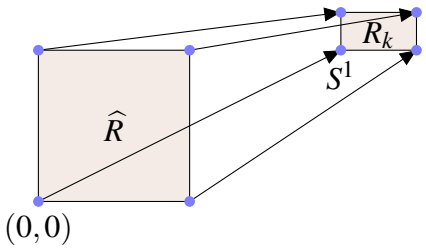

Figure 12. The affine change of variable from the reference element to the generic element.

In view of the figure, it is clearly enough to map the origin to point S^1 of coordinates $(x_1(R_k), x_2(R_k))$ and then to multiply abscissae by h_1 and ordinates by h_2. This yields

$$F_k(\hat{x}) = \begin{pmatrix} x_1(R_k) + h_1\hat{x}_1 \\ x_2(R_k) + h_2\hat{x}_2 \end{pmatrix}.$$

The inverse mapping is given by

$$F_k^{-1}(y) = \begin{pmatrix} \frac{y_1 - x_1(R_k)}{h_1} \\ \frac{y_2 - x_2(R_k)}{h_2} \end{pmatrix}.$$

It is also affine, naturally. □

Lemma 5.3.3 *There exists a constant C such that for all elements R_k and all $v \in H^1(R_k)$, setting $\hat{v}(\hat{x}) = v(F_k(\hat{x}))$, we have*

$$\int_{R_k} \|\nabla v\|^2 \, dx \leq C \int_{\widehat{R}} \|\nabla \hat{v}\|^2 \, d\hat{x}. \tag{5.7}$$

Proof. This is brute force computation. We have $v(x) = \hat{v}(F_k^{-1}(x))$ thus

$$\begin{aligned}\frac{\partial v}{\partial x_i}(x) &= \sum_{j=1}^{2} \frac{\partial \hat{v}}{\partial \hat{x}_j}(F_k^{-1}(x)) \frac{\partial (F_k^{-1})_j}{\partial x_i}(x) \\ &= h_i^{-1} \frac{\partial \hat{v}}{\partial \hat{x}_i}(F_k^{-1}(x)),\end{aligned}$$

by the multidimensional chain rule. We also need the Jacobian of the change of variables $x = F_k(\hat{x})$

$$dx = |\det DF_k(\hat{x})| \, d\hat{x} = h_1 h_2 \, d\hat{x}$$

to perform the change of variable in the integral. We obtain

$$\begin{aligned}\int_{R_k} \|\nabla v\|^2 \, dx &= \int_{\widehat{R}} \Big[h_1^{-2} \Big(\frac{\partial \hat{v}}{\partial \hat{x}_1} \Big)^2 + h_1^{-2} \Big(\frac{\partial \hat{v}}{\partial \hat{x}_1} \Big)^2 \Big] h_1 h_2 \, d\hat{x} \\ &\leq (\min(h_1, h_2))^{-2} h_1 h_2 \int_{\widehat{R}} \|\nabla \hat{v}\|^2 \, d\hat{x}.\end{aligned}$$

Now this is where the regularity of the mesh family intervenes. Letting $h = \max(h_1, h_2)$ and $\rho = \min(h_1, h_2)$, we have $\frac{1}{\rho} \leq \frac{C}{h}$. Therefore $(\min(h_1, h_2))^{-2} h_1 h_2 \leq \frac{C^2}{h^2} h^2 = C^2$, and the proof is complete. □

We now are in a position to prove Theorem 5.3.2.

Proof of Theorem 5.3.2. We use the H^1 semi-norm. Let Π_h be the V_h-interpolation operator and let $v_k = (u - \Pi_h u)_{|R_k}$ and $u_k = u_{|R_k}$. It is important to note that

$$\widehat{\Pi_h u_{|R_k}} = \widehat{\Pi} \widehat{u_k},$$

using the same hat notation as in Lemma 5.3.3 for the change of variables in functions. This is because affine change of variables of the form of F_k map Q_1 polynomials to Q_1 polynomials due to their special structure. Moreover, the two sides of the above equality satisfy the same interpolation conditions at the four vertices of the reference element, hence are equal everywhere. Therefore, we have

$$\widehat{v_k} = \widehat{u_k} - \widehat{\Pi} \widehat{u_k}.$$

We decompose the semi-norm squared as a sum over all elements

$$\|u - \Pi_h u\|^2_{H^1(\Omega)} = \sum_{k=1}^{N_{\mathscr{T}}} \int_{R_k} \|\nabla v_k\|^2 \, dx \leq C \sum_{k=1}^{N_{\mathscr{T}}} \int_{\widehat{R}} \|\nabla \widehat{v_k}\|^2 \, d\hat{x},$$

by Lemma 5.3.3.

By Lemma 5.3.1, we have

$$\begin{aligned}\int_{\widehat{R}} \|\nabla \widehat{v_k}\|^2 \, d\hat{x} &\leq C \int_{\widehat{R}} \|\nabla^2 \widehat{u_k}\|^2 \, d\hat{x} \\ &\leq C \sum_{i,j=1}^{2} \int_{\widehat{R}} \Big(\frac{\partial^2 \widehat{u_k}}{\partial \hat{x}_i \partial \hat{x}_j} \Big)^2 \, d\hat{x} \\ &= C \sum_{i,j=1}^{2} \int_{R_k} \Big(h_i h_j \frac{\partial^2 u_k}{\partial x_i \partial x_j} \Big)^2 \frac{1}{h_1 h_2} \, dx \\ &\leq Ch^2 \int_{R_k} \|\nabla^2 u_k\|^2 \, dx\end{aligned}$$

by performing the reverse change of variables, and using the regularity of the mesh family again. It follows that

$$\|u - \Pi_h u\|^2_{H^1(\Omega)} \leq Ch^2 \sum_{k=1}^{N_{\mathscr{T}}} \int_{R_k} \|\nabla^2 u_k\|^2 \, dx = Ch^2 \|\nabla^2 u\|^2_{L^2(\Omega)},$$

and the proof is complete. □

Remark 5.3.2 Under the hypothesis $u \in H^2(\Omega)$, which is satisfied in this particular case, due to elliptic regularity in a convex polygon, we thus have convergence of the Q_1 FEM when $h \to 0$, and we have an error estimate with a constant C that depends neither on h nor on u. The drawback however is that the proof does not tell us how large this constant is. □

5.4 Assembling the matrix

Let us assume that a numbering of the internal nodes, and thus of the basis functions of V_h, has been chosen: S^j and w_h^j, $j = 1, \ldots, N_i$. We have, by Q_1 interpolation

$$u_h = \sum_{j=1}^{N_i} u_h(S^j) w_h^j$$

and the matrix A has coefficients

$$A_{ij} = a(w_h^j, w_h^i) = \int_\Omega (\nabla w_h^j \cdot \nabla w_h^i + c w_h^j w_h^i) \, dx.$$

If we set

$$A_{ij}(R_k) = \int_{R_k} (\nabla w_h^j \cdot \nabla w_h^i + c w_h^j w_h^i) \, dx,$$

we see that

$$A_{ij} = \sum_{k=1}^{N_{\mathscr{T}}} A_{ij}(R_k),$$

and the coefficients can thus be computed element-wise. The idea is that many of the numbers $A_{ij}(R_k)$ do not need to be computed, since it is known that they vanish as soon as the intersection of the supports of w_h^j and w_h^i does not meet R_k. This vastly reduces the computer load.

Likewise, the right-hand side of the linear system can be written as

$$B_i = \int_\Omega f w_h^i \, dx = \sum_{k=1}^{N_{\mathscr{T}}} \int_{R_k} f w_h^i \, dx = \sum_{k=1}^{N_{\mathscr{T}}} B_i(R_k),$$

with only four nonzero terms.

Now the restriction of w_h^j to R_k is either zero, or one of the four Q_1 interpolation basis polynomials on R_k, which we denote p_i^k, $i = 1, \ldots, 4$. Here again, the reference element $\widehat{R}$ can be used with profit to compute the coefficients of the matrix. Let us recall the Q_1 Lagrange interpolation basis polynomials, or shape functions, on the reference rectangle

$$\hat{p}_1(\hat{x}) = (1-\hat{x}_1)(1-\hat{x}_2),\ \hat{p}_2(\hat{x}) = \hat{x}_1(1-\hat{x}_2),\ \hat{p}_3(\hat{x}) = \hat{x}_1\hat{x}_2,\ \hat{p}_4(\hat{x}) = (1-\hat{x}_1)\hat{x}_2. \tag{5.8}$$

We have already noticed that $p_i^k(x) = \hat{p}_i(F_k^{-1}(x))$ because both sides are Q_1 and satisfy the same interpolation conditions at the vertices. Let us give an example of computation with $\hat{p}_3$. We thus have

$$p_3^k(x) = \hat{p}_3(F_k^{-1}(x)) = \Big(\frac{x_1 - x_1(R_k)}{h_1}\Big)\Big(\frac{x_2 - x_2(R_k)}{h_2}\Big).$$

Therefore

$$\|\nabla p_3^k(x)\|^2 = \frac{1}{h_1^2 h_2^2}\big((x_1 - x_1(R_k))^2 + (x_2 - x_2(R_k))^2\big),$$

and assuming S^i is the upper right corner of R_k, we obtain by computing the integrals on R_k

$$A_{ii}(R_k) = \frac{h_1 h_2}{3}\Big(\frac{1}{h_1^2} + \frac{1}{h_2^2}\Big) + c_0 \frac{h_1 h_2}{9}$$

in the case when $c = c_0$ is a constant. Now there are four such contributions to A_{ii} coming from the four rectangles that surround S^i, hence

$$A_{ii} = \frac{4h_1 h_2}{3}\Big(\frac{1}{h_1^2} + \frac{1}{h_2^2}\Big) + c_0 \frac{4h_1 h_2}{9}.$$

The diagonal coefficients do not depend on the node numbering, but the off-diagonal ones do depend completely on it. So we have to talk about numbering, since in the 2d case, as opposed to the 1d case, no natural numbering appears at the onset.

We first note that there is a connection between Q_1 Lagrange approximation in two dimensions and P_1 Lagrange approximation in one dimension.

The four basis polynomials on $\widehat{R}$ are given by equation (5.8). In one dimension, the basis polynomials for P_1 Lagrange interpolation on $[0,1]$ are

$$\ell_1(x) = 1 - x, \quad \ell_2(x) = x.$$

Therefore, we see that

$$\hat{p}_1(x) = \ell_1(\hat{x}_1)\ell_1(\hat{x}_2), \hat{p}_2(x) = \ell_2(\hat{x}_1)\ell_1(\hat{x}_2),$$
$$\hat{p}_3(x) = \ell_2(\hat{x}_1)\ell_2(\hat{x}_2), \hat{p}_4(x) = \ell_1(\hat{x}_1)\ell_2(\hat{x}_2).$$

In this context, we introduce a notation: Let f and g be two functions in one variable. We define a function in two variables $f \otimes g$ by $f \otimes g(x_1, x_2) = f(x_1)g(x_2)$. This function is called the *tensor product* of f and g.[4] With this notation, we thus have $p_1 = \ell_1 \otimes \ell_1$ and so on.

This tensor product decomposition extends to the basis functions on Ω themselves. Let S^i be an interior node of coordinates $(i_1 h_1, i_2 h_2)$ and R_k^i, $k = 1, \ldots, 4$, the four elements surrounding it.

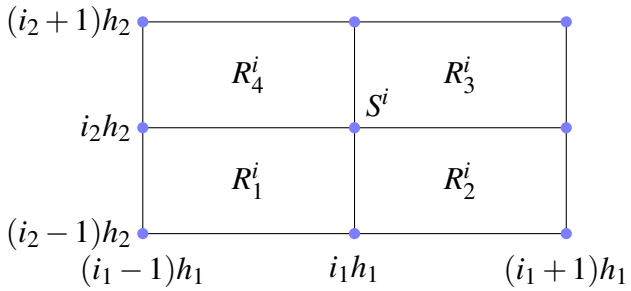

Figure 13. The support of w_h^i.

By direct verification of the interpolation relations, we easily check that

$$w_h^i(x) = \begin{cases} \ell_2\left(\frac{x_1}{h_1} - (i_1 - 1)\right)\ell_2\left(\frac{x_2}{h_2} - (i_2 - 1)\right) & \text{in } R_1^i, \\ \ell_1\left(\frac{x_1}{h_1} - i_1\right)\ell_2\left(\frac{x_2}{h_2} - (i_2 - 1)\right) & \text{in } R_2^i, \\ \ell_1\left(\frac{x_1}{h_1} - i_1\right)\ell_1\left(\frac{x_2}{h_2} - i_2\right) & \text{in } R_3^i, \\ \ell_2\left(\frac{x_1}{h_1} - (i_1 - 1)\right)\ell_1\left(\frac{x_2}{h_2} - i_2\right) & \text{in } R_4^i, \\ 0 & \text{elsewhere.} \end{cases}$$

[4] Consider this to be just vocabulary. We do not need to know anything about tensor products in general.

Therefore, if $w_{h_1}^{1,i_1}$ denotes the 1d hat function associated with node $i_1 h_1$ of the 1d mesh of $[0,1]$ of mesh size h_1, and likewise for $w_{h_2}^{2,i_2}$, we see that

$$w_h^i = w_{h_1}^{1,i_1} \otimes w_{h_2}^{2,i_2}.$$

Let us use this tensor product decomposition to number the basis functions. The idea is to use the indices i_1 and i_2 to sweep the rows and then the columns of the mesh[5]. We thus define a mapping $\{1,2,\dots,N_1\} \times \{1,2,\dots,N_2\} \to \{1,2,\dots,N_i\}$ by

$$(i_1, i_2) \mapsto i = i_1 + (i_2 - 1)N_1. \tag{5.9}$$

It is clearly a bijection (recall that $N_i = N_1 N_2$). To compute the inverse mapping, we note that $i_1 - 1$ is the remainder of the Euclidean division of $i-1$ by N_1, thus

$$i_1 = i - \left\lfloor \frac{i-1}{N_1} \right\rfloor N_1, \quad i_2 = \left\lfloor \frac{i-1}{N_1} \right\rfloor + 1. \tag{5.10}$$

Now, the support of a tensor product is the Cartesian product of the supports. Thus

$$\operatorname{supp} w_h^i = \operatorname{supp} w_{h_1}^{1,i_1} \times \operatorname{supp} w_{h_2}^{2,i_2} = [(i_1-1)h_1, (i_1+1)h_1] \times [(i_2-1)h_2, (i_2+1)h_2].$$

If an index j in the numbering corresponds to a couple (j_1, j_2), we thus see that $A_{ij} \neq 0$ if and only if the supports have non negligible intersection, that is to say

$$A_{ij} \neq 0 \Longleftrightarrow |j_1 - i_1| \leq 1 \text{ and } |j_2 - i_2| \leq 1.$$

In view of the numbering formulas above, saying the $|j_1 - i_1| \leq 1$ is equivalent to saying that $j - i = \alpha + kN_1$ with $\alpha = i_1 - j_1 = -1$, 0 or 1, and k an integer. Since we also have $i_2 = \frac{i-i_1}{N_1}$, it follows that $i_2 - j_2 = k = -1$, 0 or 1. Therefore, for a given i, that is a given row of A, there are at most nine values of j, that is nine columns, that contain a nonzero coefficient. Of course, not all rows contain nine nonzero coefficients. For example, the first row has four nonzero coefficients, the second row has six nonzero coefficients, and so on. Rows that correspond to index pairs (i_1, i_2) such that $2 \leq i_1, i_2 \leq N_1 - 1$ do have nine nonzero coefficients (they correspond to interior nodes with nine neighboring interior nodes, including themselves). Such a row looks like this:

$j =$ $i-N_1-1$ $i-N_1+1$ $i-1$ $i+1$ $i+N_1-1$ $i+N_1+1$
$i-N_1$ i $i+N_1$

We see three tridiagonal $N_1 \times N_1$ blocks emerging, that are themselves arranged block tridiagonally. The whole $(N_1 N_2) \times (N_1 N_2)$ matrix is thus composed of

[5] Or the other way around. But let's stick to this one here.

$N_2 \times N_2$ blocks, that are either $N_1 \times N_1$ zero or $N_1 \times N_1$ tridiagonal. Indeed, if we define the $N_1 \times N_1$ matrices A^{kl} to be the block of indices (k,l), which means it is comprised of lines $(k-1)N_1+1$ to kN_1 and columns $(l-1)N_1+1$ to lN_1, then using the inverse numbering (5.10), we see that

$$A_{ij} = a(w_h^j, w_h^i) = a(w_{h_1}^{1,j_1} \otimes w_{h_2}^{2,l}, w_{h_1}^{1,i_1} \otimes w_{h_2}^{2,k}),$$

for all (i,j) in this block. Therefore we have $(A^{kl})_{i_1 j_1} = a(w_{h_1}^{1,j_1} \otimes w_{h_2}^{2,l}, w_{h_1}^{1,i_1} \otimes w_{h_2}^{2,k})$, thus $A^{kl} = 0$ as soon as $|k-l| \geq 2$ and is tridiagonal for $|k-l| \leq 1$, for reasons of supports. We thus have

$$A = \begin{pmatrix} A^{11} & A^{12} & 0 & \cdots & 0 \\ A^{21} & A^{22} & A^{23} & \cdots & 0 \\ 0 & A^{32} & A^{33} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & A^{N_2-1,N_2} & A^{N_2N_2} \end{pmatrix},$$

where the tridiagonal block structure appears.

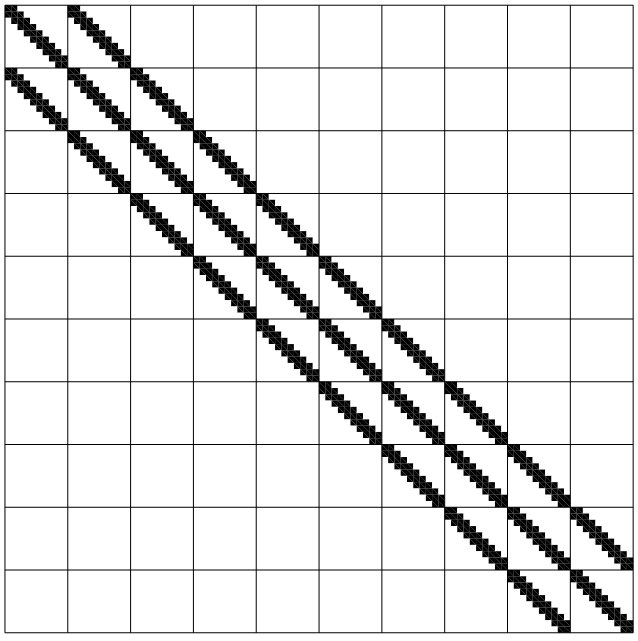

Figure 14. The block tridiagonal structure of A (here $N_1 = N_2 = 10$ so A is 100×100). Black squares indicate nonzero matrix coefficients, white areas zero. The coarse grid shows the 10×10 blocks.

The sweep columns then rows numbering thus gives rise to a well-structured matrix for which there exist efficient numerical methods. It is instructive to see what kind of matrix would result from other numberings that could be considered just as natural, such as the numbering used to prove that $\mathbb{N}^2$ is countable (although limited to a square here): start from the lower left node, go east one node, then north west, then north, then south east, etc.

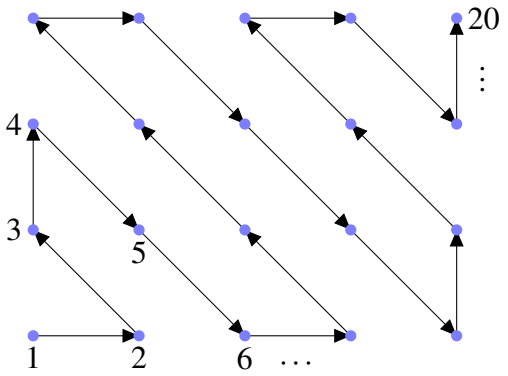

Figure 15. An alternate node numbering scheme.

For the same 100×100 case, we obtain a matrix structure[6] that looks like this

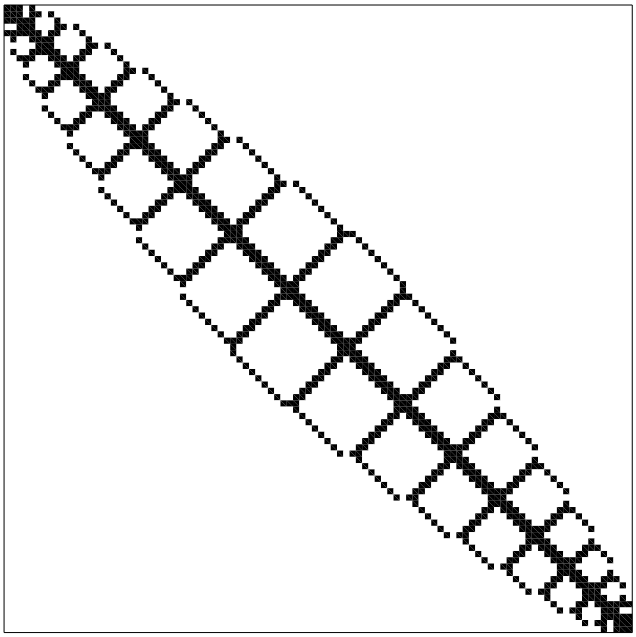

Figure 16. Structure of the alternate matrix.

[6] At least, if the Scilab script I wrote for it is correct...

which is quite pretty, but maybe not as good as the previous one for numerical purposes. Of course, the entries of the above matrix are the same as the previous ones, since both matrices are similar to each other via a permutation matrix.

5.5 General definition of a finite element

It is now time to look back and see what are the general characteristics of the finite elements we have seen so far, so as to finally define what a finite element is! We let $\mathbb{P} = \mathbb{R}[X,Y]$ denote the space of polynomials in two indeterminates.

Definition 5.5.1 *A two-dimensional* finite element *is a triple* $(T, P(T), \{\varphi_1, \ldots, \varphi_d\})$ *where*

1) T *is a compact, convex polygon.*

2) $P(T)$ *is a finite dimensional subspace of* $\mathbb{P}$*, considered as a function space on* T*.*

3) φ_i *are linear forms on* $\mathbb{P}$*, which are called the* degrees of freedom *of the finite element.*

Remark 5.5.1 In practice, T is either a triangle or a rectangle. The same definition applies in dimensions one and three (finite elements are rarely used in dimensions higher than four, although it happens). In dimension one, T is an interval. There is more variety in dimension three, starting with tetrahedra.

In the literature, in particular the engineering literature, finite elements are always presented this way, and not starting with the discrete space V_h and so on, as we have done up to know. Of course, starting from the top, *i.e.* the discrete space, down to the finite element is the logical way to proceed, instead of starting from the bottom. □

Definition 5.5.2 *We say that a finite element is* unisolvent *if for all* d*-uples of scalars* $(\alpha_1, \ldots, \alpha_d)$*, there exists one and only one polynomial* $p \in P(T)$ *such that* $\varphi_i(p) = \alpha_i$*,* $i = 1, \ldots, d$*.*

Unisolvence is a generalization of the interpolation property for all kinds of degrees of freedom.

Proposition 5.5.1 *If a finite element is unisolvent, then,* $d = \dim P(T)$*.*

Proof. This is fairly obvious. Assume we want to solve the d equations $\varphi_i(p) = \alpha_i$. Since φ_i are linear forms, these equations are linear equations in $\dim P(T)$ unknowns, once we choose a basis of $P(T)$. Hence if the number of equations and the number of unknowns are different, the system of equations certainly cannot be solved uniquely for all right-hand sides. □

Remark 5.5.2 So if we do not have the same number of degrees of freedom as the dimension of the finite element space, then the element in question is not unisolvent. Be careful that unisolvence is not just a question of dimensions, as the following example shows.

Take $T = \{|x_1| + |x_2| \leq 1\}$, $P(T) = Q_1$ and ϕ_i the values at the four vertices of T. We have $\dim Q_1 = 4$ but this element is not unisolvent, since $p(x) = x_1 x_2$ is in Q_1 and $\phi_i(p) = 0$ for all i even though $p \neq 0$.

Of course, $T = [0,1]^2$, $P(T) = Q_1$ and ϕ_i the values at the four vertices of T is unisolvent. This is the element we have been using so far in 2d.

Therefore, unisolvence somehow reflects the adequacy of the duality of the polynomial space and the degrees of freedom. □

In practice, unisolvence is checked using the following result.

Proposition 5.5.2 *A finite element is unisolvent if and only if* $d = \dim P(T)$ *and there exists a basis* (p_j) *of* $P(T)$ *such that* $\varphi_i(p_j) = \delta_{ij}$ *for all* i, j.

Proof. If the element is unisolvent, we already know that $d = \dim P(T)$. Moreover, choosing $\alpha_i = \delta_{ij}$ for $j = 1, \ldots, j$ yields the existence of p_j by the very definition. The family (p_j) is linearly independent, for if

$$\sum_{j=1}^{d} \lambda_j p_j = 0,$$

applying the linear form φ_i, we obtain

$$0 = \sum_{j=1}^{d} \lambda_j \varphi_i(p_j) = \sum_{j=1}^{d} \lambda_j \delta_{ij} = \lambda_i$$

for all i. Thus it is a basis of $P(T)$.

Conversely, assume that $d = \dim P(T)$ and that we have a basis p_j with the above property. Let us be given scalars α_i. Then the polynomial $p = \sum_{j=1}^{d} \alpha_j p_j$ is the only element of $P(T)$ such that $\varphi_i(p) = \alpha_i$ by the same argument. □

Remark 5.5.3 The polynomials p_j are called the *basis polynomials* or *shape functions* of the finite element. They are dual to the degrees of freedom. They are also used to construct the basis functions of the discrete approximation, as we have seen already in 1d with the P_1 Lagrange and P_3 Hermite approximations, and in 2d with the Q_1 Lagrange approximation. In the latter case, the shape functions were already given in equation (5.8).

This also indicates that unisolvence is far from being the end of the story in terms of finite elements. The basis polynomials must also be such that they can be combined into globally continuous functions so as to give rise to a conforming approximation. □

Speaking of duality, we also introduce $\Sigma(T) = \text{vect}\{\varphi_1, \ldots, \varphi_d\}$, the vector subspace of $\mathbb{P}^*$ spanned by the degrees of freedom. In a similar vein as Proposition 5.5.1, we also have

Proposition 5.5.3 *If a finite element is unisolvent, then,* $d = \dim \Sigma(T)$.

Proof. Clear. □

The basis polynomials and the degrees of freedom are obviously dual bases of their respective spanned spaces. In the counterexample shown above, the four linear forms are linearly independent as elements of $\mathbb{P}^*$, but not as elements of Q_1^*.

5.6 Q_2 and Q_3 finite elements

Let us briefly discuss what happens if we want to use higher degree polynomials. We start with Q_2 Lagrange elements for second order problems. The discrete approximation space is then

$$V_h = \{v_h \in C^0(\bar{\Omega}); \forall R_k \in \mathscr{T}, v_{h|R_k} \in Q_2, v_h = 0 \text{ on } \partial\Omega\}, \tag{5.11}$$

on the same rectangular mesh as before and the general approximation theory applies (note that this space is larger than the previous one). We concentrate on the description of the finite element first. We set $R = [0,1]^2$, $P(R) = Q_2$ and we need to describe the degrees of freedom. Since we are going to use Lagrange interpolation, these degrees of freedom are going to be values at some points of the element. The dimension of Q_2 is nine, therefore nine degrees of freedom are required to define a unisolvent finite element, *i.e.*, nine points or nodes in R. The choice of points must also be guided by the necessity of defining a global C^0 interpolation based on the nodal values on the mesh. The following set of points turns out to satisfy both requirements.

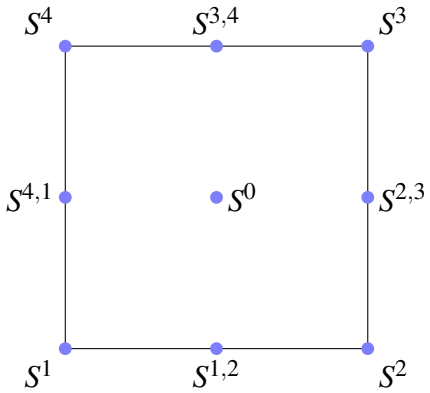

Figure 17. The nine nodes of the Q_2 Lagrange element.

We thus take the four vertices S^k as before, plus the four middles of the edges $S^{k,k+}$, where $k+ = k+1$ for $k = 1,2,3$ and $k+ = 1$ for $k = 4$, plus the center of gravity S^0.

Proposition 5.6.1 *The finite element* $(R, Q_2, \{p(S^k), p(S^{k,k+}), p(S^0), k = 1,\ldots,4\})$ *is unisolvent.*

Proof. The number of degrees of freedom matches the dimension of the space. It is thus sufficient to construct the basis polynomials. We will number them the same way as the node they correspond to. There are three polynomials to be constructed: p^1 from which the other p^k are deduced by symmetry, $p^{1,2}$ from which the other $p^{k,k+}$ are deduced by symmetry, and p^0.

Let us show how to compute p^0. The interpolation conditions to be satisfied are $p^0(S^0) = 1$ and $p^0 = 0$ on all other eight nodes. Now p^0 is zero at points $(0,0)$, $(0,\frac{1}{2})$ and $(0,1)$. The restriction of a Q_2 polynomial to the line $x_2 = 0$ is a second degree polynomial in the variable x_1, and we have just seen that this polynomial has three roots. Therefore it vanishes and $p^0 = 0$ on the straight line $x_2 = 0$. It follows that p^0 is divisible by x_2. The same argument shows that it is divisible by x_1, $(1-x_1)$ and $(1-x_2)$. These polynomials are relatively prime, thus p^0 is divisible by their product,

$$p^0(x) = q(x)x_1x_2(1-x_1)(1-x_2) = q(x)(x_1x_2 - x_1^2x_2 - x_1x_2^2 + x_1^2x_2^2).$$

Now $x_1x_2 - x_1^2x_2 - x_1x_2^2 + x_1^2x_2^2 \in Q_2$, thus the partial degree of q is less than 0, *i.e.*, q is a constant C. Evaluating now p^0 at point $S^0 = (\frac{1}{2},\frac{1}{2})$, we obtain

$$1 = C \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{C}{16}.$$

Finally, we find that

$$p^0(x) = 16x_1x_2(1-x_1)(1-x_2).$$

Conversely, it is clear that this particular polynomial is in Q_2 and satisfies the required interpolation conditions.

The same arguments, that we leave as an exercise, show that

$$p^1(x) = (1-x_1)(1-2x_1)(1-x_2)(1-2x_2),$$

and

$$p^{1,2}(x) = 4x_1(1-x_1)(1-x_2)(1-2x_2).$$

as we said before the remaining six polynomials are obtained by considerations of symmetry. □

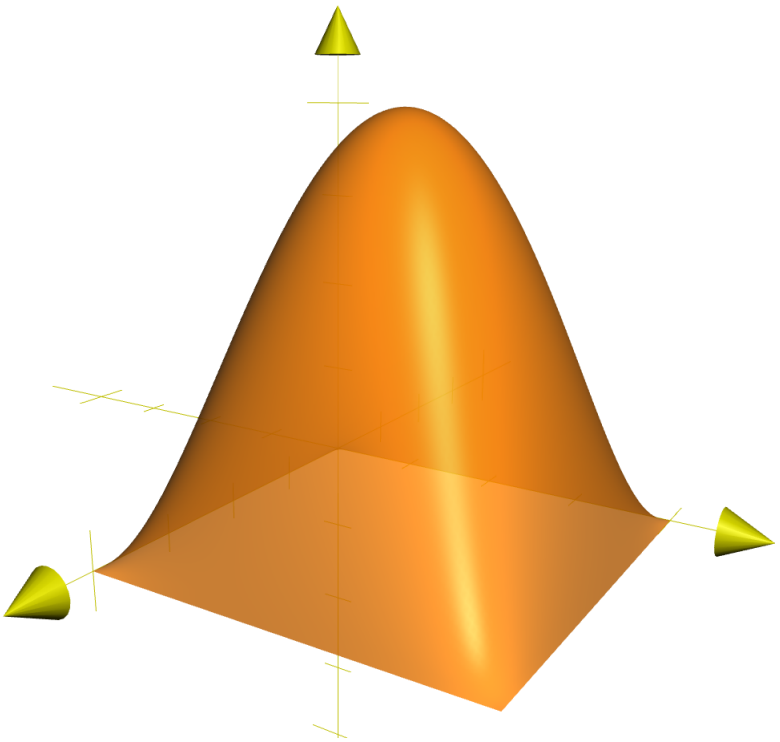

Figure 18. The graph of p^0.

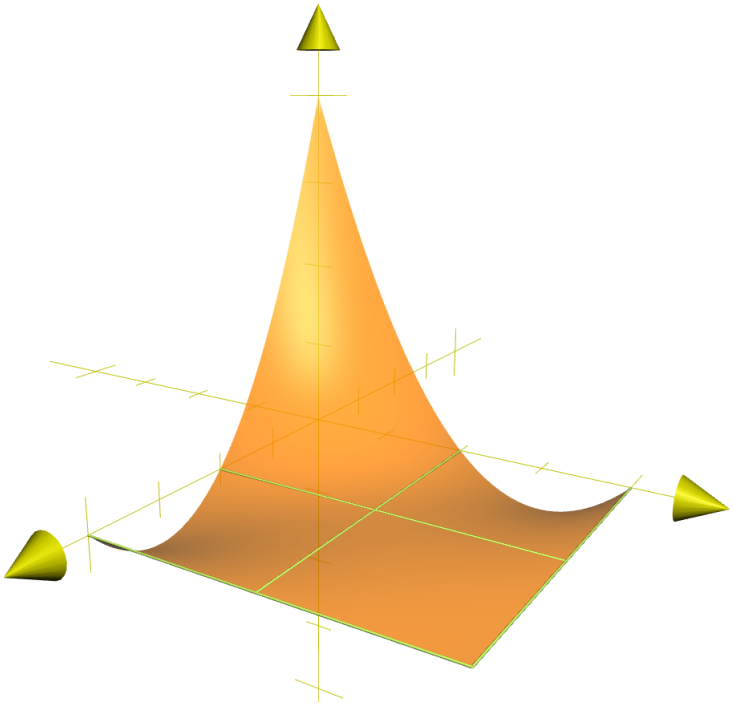

Figure 19. The graph of p^1, with the segments where p^1 vanishes.

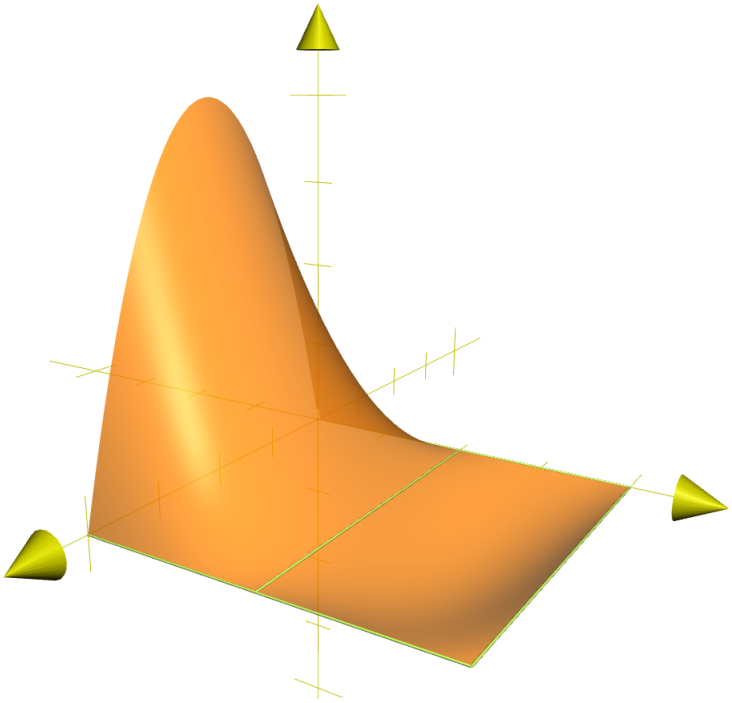

Figure 20. The graph of $p^{1,2}$, with the segments where $p^{1,2}$ vanishes.

Let us now consider the whole mesh. The nodes no longer are just the element vertices, but also the middles of the edges and centers of gravity of the elements.

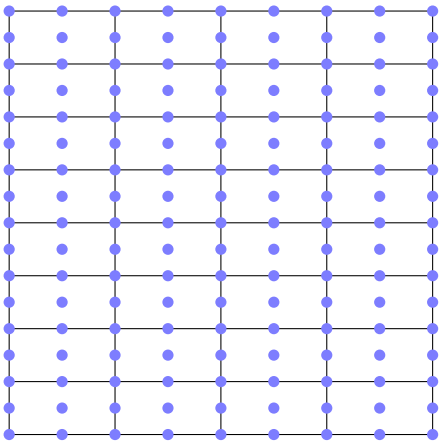

Figure 21. Same mesh as Figure 7, 32 elements, 153 nodes.

We then have the exact analog of Propositions 5.2.2 and 5.2.3.

Proposition 5.6.2 *A function of V_h is uniquely determined by its values at the internal nodes of the mesh and all sets of values are interpolated by an element of V_h.*

Proof. By unisolvence, nine values for the nine nodes of an element determine one and only one Q_2 polynomial that interpolates these nodal values (we take the value 0 for the nodes located on the boundary). Therefore, if we are given a set of values for each node in the mesh, this set determines one Q_2 polynomial per element. Let us check that they combine into a globally C^0 function.

Let us thus consider the following situation, without loss of generality:

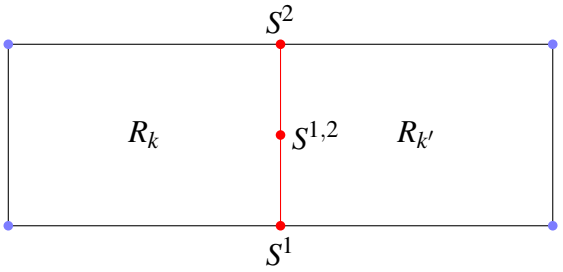

Figure 22. Continuity across an internal edge, Q_2 case.

We thus have two Q_2 polynomials q and q' such that $q(S^1) = q'(S^1)$, $q(S^{1,2}) = q'(S^{1,2})$ and $q(S^2) = q'(S^2)$. On the segment $[S^1, S^2]$, $q - q'$ is a second degree polynomial in the variable x_2 that has three roots. Therefore, $q - q' = 0$ on this segment, and the function defined by $q(x)$ if $x \in R_k$, $q'(x)$ if $x \in R_{k'}$ is continuous on $R_k \cup R_{k'}$. □

Corollary 5.6.1 *Let us be given a numbering of the nodes S^k, $k = 1, \ldots, N = (2N_1 + 1)(2N_2 + 1)$. There is a basis of V_h composed of the functions w_h^i defined by $w_h^i(S^j) = \delta_{ij}$ and and for all $v_h \in V_h$, we have*

$$v_h = \sum_{i=1}^{N} v_h(S^i) w_h^i. \tag{5.12}$$

Proof. Same as before. □

Below are pictures of the different types of basis functions, depending on which kind of nodes they are attached to. More pictures to be found on the course Web page.

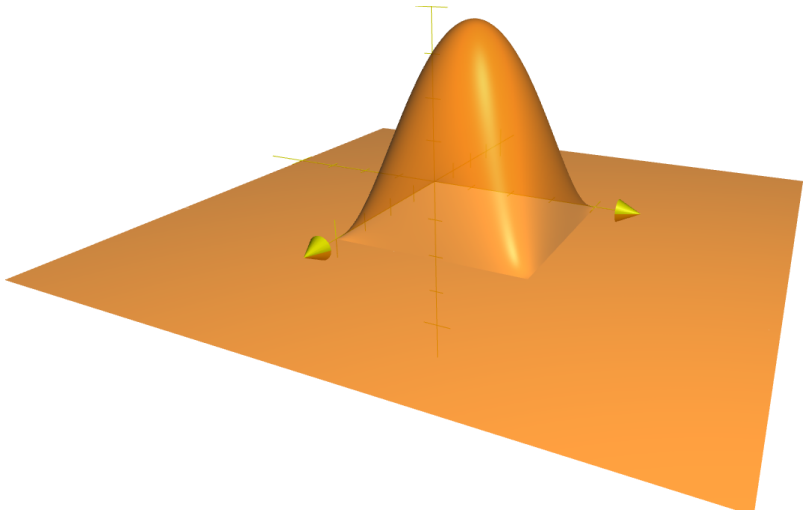

Figure 23. Basis function corresponding to an element center of gravity.

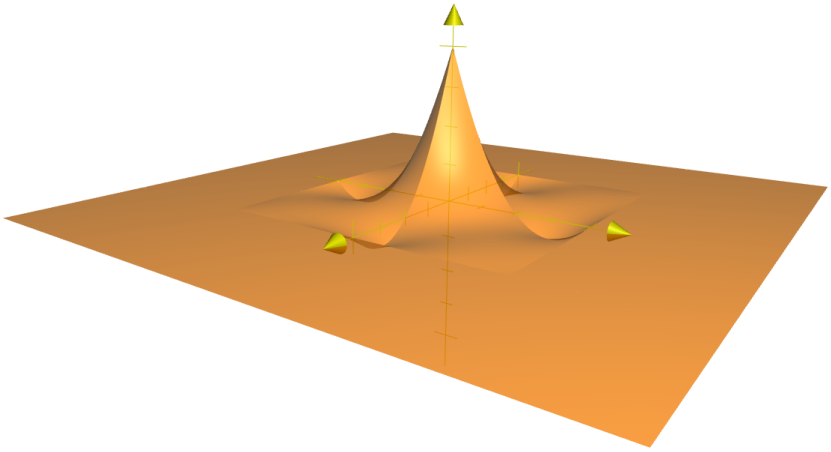

Figure 24. Basis function corresponding to an element vertex.

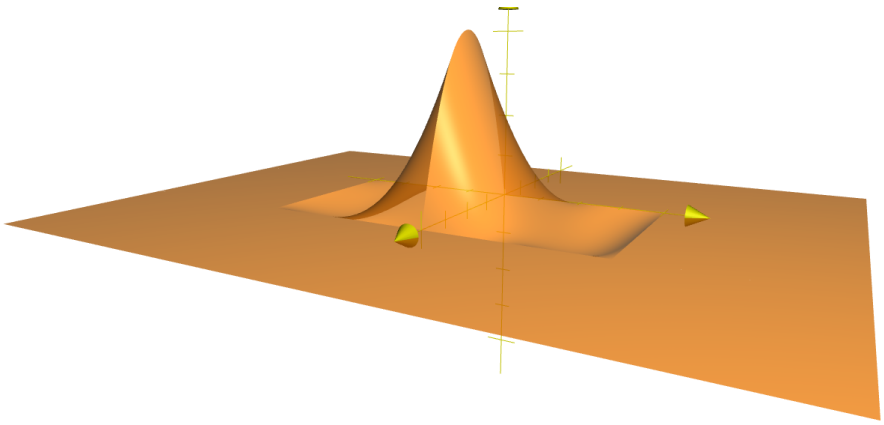

Figure 25. Basis function corresponding to an edge middle.

Note that the latter two change sign in Ω. This was not the case for Q_1 basis functions.

We do not pursue here matrix assembly and node numbering issues. It is to be expected that the structure of the matrix is more complicated than in the Q_1 case.

The question arises as to why introduce Q_2 elements and deal with the added complexity compared with the Q_1 case. One reason is that we thus obtain a higher order approximation method. Indeed, if $u \in H^3(\Omega)$, then we have (exercise) a better error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^2 \|u\|_{H^3(\Omega)},$$

than with Q_1 elements. The estimate is better in the sense that $h^2 \ll h$ when h is small, even though we do not have any idea of the order of magnitude of the constants and the norms. So the extra implementation and computational costs must be balanced against the increased accuracy that is expected from the higher degree finite element approximation. For instance, a cheaper computation may be achieved with the same accuracy by taking less elements.

Let us say a few words about Q_3 finite elements. We could define Q_3 Lagrange elements by taking 16 nodes per element, since $\dim Q_3 = 16$. We would need four nodes per edge to ensure global continuity, hence the four vertices plus two points on the thirds of each edge. Four more points must be chosen inside, with obvious simple possibilities.

We can also use Q_3 finite elements for Hermite interpolation that result in C^1 functions suitable for conforming approximation of fourth order problems. In this case, the degrees of freedom must also include partial derivative values. We would thus take as degrees of freedom the 4 vertex values and the 8 first partial derivatives values at the vertices. This would seem to be enough, as we recognize 1d P_3 Hermite interpolation on each edge, and there is a tensor product structure $Q_3[X,Y] = P_3[X] \otimes P_3[Y]$.

Surprisingly, this is not enough. Indeed, it would only give 12 degrees of freedom for a 16-dimensional space, and there would infinitely many different possible basis polynomials, in the sense that the interpolation relations would be satisfied, since there is infinitely many different ways of adding 4 more degrees of freedom. Moreover, it is not clear which choice would guarantee global C^1 regularity. Surprisingly again, if we complete the set of degrees of freedom with the 4 vertex values of the *second* derivatives $\frac{\partial^2 p}{\partial x_1 \partial x_2}$, we obtain a unisolvent element that is suitable for C^1 approximation. See the class Web page for pictures of the basis polynomials and C^1 basis functions constructed from them.

We have seen an interesting example of the same polynomial space used with two completely different sets of degrees of freedom and yielding two completely different approximation spaces, Q_3 Lagrange and Q_3 Hermite.