Preconditioned finite elements method

Let $\mathcal{V}$ be a Hilbert space, $(\cdot,\cdot)_{\mathcal{V}}$ an inner product on $\mathcal{V}$ and $\|\cdot\|_{\mathcal{V}}$ the corresponding induced norm. Let a be a coercive, continuous, bilinear form on $\mathcal{V}$, that is, $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and there exist m, M, $0 < m \leq M$ such that for all $u, v, w \in \mathcal{V}$, $\alpha, \beta \in \mathbb{R}$,

$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$, $a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$ (bilinearity),

$|a(u, v)| \leq M\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}}$ (continuity),

$a(v, v) \geq m\|v\|^2_{\mathcal{V}}$ (coercivity).

Let $\mathcal{V}'$ be the set of all continuous, linear forms on $\mathcal{V}$.

Then the following Lax-Milgran results hold:

(LM1) For any $\mathcal{F} \in \mathcal{V}'$, there exists a unique element $u = u_{\mathcal{F}} \in \mathcal{V}$ such that $a(u, v) = \mathcal{F}(v)$, $\forall\, v \in \mathcal{V}$.

(LM2) If, moreover, $\mathcal{V}_h$ is a finite-dimensional subspace of $\mathcal{V}$, then to $\mathcal{F}$ we can also associate an element $u_h \in \mathcal{V}_h$ such that $a(u_h, v_h) = \mathcal{F}(v_h)$, $\forall\, v_h \in \mathcal{V}_h$, which is uniquely defined too.

Approximating u by u_h

Intuitively, such element u_h can be used as an approximation of u if $\mathcal{V}_h$ belongs to a family of subspaces $\{\mathcal{V}_h\}_{h\to 0}$ of increasing dimension, such that the closure of $\cup_{h\to 0}\mathcal{V}_h$ coincides with $\mathcal{V}$. In fact, it can be shown that an hypothesis of *consistency* on $\{\mathcal{V}_h\}_{h\to 0}$ (implying the latter property) yields the result:

$$h \to 0 \;\Rightarrow\; \|u - u_h\|_{\mathcal{V}} \;\to\; 0. \tag{conv}$$

Consistency of $\{\mathcal{V}_h\}_{h\to 0}$ in $\mathcal{V}$. $\{\mathcal{V}_h\}_{h\to 0}$, $\mathcal{V}_h \subset \mathcal{V}$, is said to be consistent in $\mathcal{V}$ if there exist $\mathbb{V} \subset \mathcal{V}$ dense in $\mathcal{V}$ (with respect to $\|\cdot\|_{\mathcal{V}}$) and an operator $\mathcal{R}_h : \mathbb{V} \to \mathcal{V}_h$ such that for any $v \in \mathbb{V}$, $\|\mathcal{R}_h(v) - v\|_{\mathcal{V}} \to 0$ as $h \to 0$ ($\mathcal{R}_h$ might be an interpolation operator).

Let us show that the consistency of $\{\mathcal{V}_h\}_{h\to 0}$ implies (conv). First we prove that the error $\|u - u_h\|_{\mathcal{V}}$ is proportional to the minimal error we can have with $\mathcal{V}_h$. Note that

$$a(u, v_h) = F(v_h), \quad a(u_h, v_h) = F(v_h) \;\Rightarrow\; a(u - u_h, v_h) = 0 \;\forall\, v_h \in \mathcal{V}_h$$

and this remark implies

$$m\|u - u_h\|^2_{\mathcal{V}} \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq M\|u - u_h\|_{\mathcal{V}}\|u - v_h\|_{\mathcal{V}},$$

$$\|u - u_h\|_{\mathcal{V}} \leq \frac{M}{m} \inf_{v_h \in \mathcal{V}_h} \|u - v_h\|_{\mathcal{V}}. \tag{cea}$$

Now we can prove (conv). Let $v \in \mathbb{V}$ be such that $\|v - u\|_{\mathcal{V}} < \varepsilon$. By (cea) and the consistency hypothesis, if $h < h_\varepsilon$ (h is suitably small), then

$$\|u - u_h\|_{\mathcal{V}} \leq \frac{M}{m}\|u - \mathcal{R}_h(v)\|_{\mathcal{V}} \leq \frac{M}{m}(\|u - v\|_{\mathcal{V}} + \|v - \mathcal{R}_h(v)\|_{\mathcal{V}}) < \frac{M}{m}2\varepsilon.$$

How to compute u_h

Let N_h be the dimension of $\mathcal{V}_h$, and φ_i, $i = 1, \ldots, N_h$, a basis for $\mathcal{V}_h$. Then $u_h = \sum_{j=1}^{N_h} (u_h)_j \varphi_j$ and the condition $a(u_h, v_h) = F(v_h)$, $v_h \in \mathcal{V}_h$, can be rewritten as follows:

$$\sum_{j=1}^{N_h} (u_h)_j a(\varphi_j, \varphi_i) = F(\varphi_i), \quad i = 1, \ldots, N_h.$$

So the $(u_h)_j$ defining u_h can be obtained by solving a linear system $A\mathbf{x} = \mathbf{b}$, being $a_{ij} = a(\varphi_j, \varphi_i)$, $b_i = F(\varphi_i)$, $1 \le i, j \le N_h$. It is important to notice that the symmetric part of the coefficient matrix A is positive definite, that is $\mathbf{z}^T A\mathbf{z} > 0$, $\forall\, \mathbf{z} \in \mathbb{R}^n$ $\mathbf{z} \neq \mathbf{0}$. In fact, by the coercivity of a, we have

$$\mathbf{z}^T A\mathbf{z} = \sum_{ij} z_i a(\varphi_j, \varphi_i) z_j = a(\sum z_j \varphi_j, \sum z_i \varphi_i) \ge m \| \sum z_i \varphi_i \|_{\mathcal{V}}^2 > 0$$

unless the z_i are all null (the φ_i are assumed linearly independent).

Example: a differential problem solved by the finite element method

Assume that $u : \Omega \to \mathbb{R}$ is the unique solution of the differential problem

$$\begin{array}{l} -\nabla(\alpha \nabla u) + \beta \nabla u + \gamma u = f, \quad x \in \Omega, \\ u = \varphi, \quad x \in \Gamma_D, \\ \frac{\partial u}{\partial \mathbf{n}_c} = \psi, \quad x \in \Gamma_N. \end{array}$$

Here Ω is an open set in $\mathbb{R}^d$, Γ_D and Γ_N are open subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma_D} \cup \overline{\Gamma_N}$, $\alpha : \Omega \to \mathbb{R}^{d^2}$, $\beta : \Omega \to \mathbb{R}^d$, $\gamma, f : \Omega \to \mathbb{R}$.

Then, for all v, $v|_{\Gamma_D} = 0$,

$$a(u, v) := \int_\Omega \alpha \nabla u \nabla v + \int_\Omega \beta \nabla u v + \int_\Omega \gamma u v = \int_\Omega f v + \int_{\Gamma_N} \psi v d\sigma.$$

If we set $u = u_\varphi + w$ with $u_\varphi, w : \Omega \to \mathbb{R}$, $u_\varphi|_{\Gamma_D} = \varphi$ and $w|_{\Gamma_D} = 0$, then the latter equation becomes:

$$a(w, v) = \int_\Omega f v + \int_{\Gamma_N} \psi v d\sigma - a(u_\varphi, v) =: F(v).$$

So, we have the following

Problem. Find w, $w|_{\Gamma_D} = 0$ such that $a(w, v) = F(v)$, $\forall\, v$, $v|_{\Gamma_D} = 0$. Moreover, the functions w, v must be also such that $a(w, v)$ and $F(v)$ are well defined, that is we also require $w, v \in H^1(\Omega)$ where

$$\begin{array}{l} H^1(\Omega) = \{v \in L^2(\Omega) : \ D_i v \in L^2(\Omega)\} \\ (D_i v \in L^2(\Omega) \ \text{ iff } \ \exists g_i \in L^2(\Omega) \,|\, \int_\Omega g_i \varphi = -\int_\Omega v D_i \varphi \, \forall\, \varphi \in C_0^\infty(\Omega) \ (D_i v := g_i)). \end{array}$$

Briefly, *find* $w \in H^1_{0,\Gamma_D}(\Omega) \mid a(w, v) = F(v)$, $\forall\, v \in H^1_{0,\Gamma_D}(\Omega)$.

Under suitable conditions on the data $\partial\Omega, \alpha, \beta, \gamma, \varphi, \psi$, the space $\mathcal{V} = H^1_{0,\Gamma_D}(\Omega)$ is a Hilbert space with respect to the inner product $(u, v)_{\mathcal{V}} = (u, v)_{1,\Omega} = (u, v)_{L^2(\Omega)} + \sum_{i=1\ldots d} (D_i u, D_i v)_{L^2(\Omega)}$, and the forms a and F are well defined and satisfy the conditions required by the Lax-Milgran results (LM1) and (LM2) to hold. So, the w of the problem is well defined, and for any finite-dimensional

subspace $\mathcal{V}_h$ of $\mathcal{V} = H^1_{0,\Gamma_D}(\Omega)$ is well defined a function $w_h \in \mathcal{V}_h$ such that $a(w_h, v_h) = F(v_h)$, $\forall\, v_h \in \mathcal{V}_h$.

Definition of w_h convergent to w

In order to yield functions w_h convergent to w as $h \to 0$, we only have to define $\mathcal{V}_h$ such that $\{\mathcal{V}_h\}_{h\to 0}$ is consistent in $H^1_{0,\Gamma_D}(\Omega)$. Let us do this in the case $d = 2$, $\Omega =$ polygon, by using the finite element method.

Let τ_h be a triangulation of Ω of diameter h, that is a set of triangles T such that

- $T \in \tau_h \Rightarrow T = \overline{T} \subset \overline{\Omega}$ and $\text{diam}\,(T) =: h_T \le h := \max_{T\in\tau_h} h_T$,
- $\cup_{T\in\tau_h} T = \overline{\Omega}$,
- $T_1, T_2 \in \tau_h \Rightarrow T_1 \cap T_2$ is a common vertex, a common side, the whole triangle $T_1 = T_2$ or the empty set.

To any T in the following we need to associate also the number ρ_T, the diameter of the circle enclosed in T. Let S_h be the set of all functions $p : \Omega \to \mathbb{R}$ such that $p|_T$ is a degree-1 polynomial (in x_1 and x_2) and set $\mathcal{V}^*_h = S_h \cap C^0(\overline{\Omega})$. Let $i = 1, 2, \ldots, N^*_h$ be the nodes of the triangulation τ_h (the verteces of the triangles of τ_h) and denote by φ_i the elements of $\mathcal{V}^*_h$ satisfying the identities $\varphi_i(j) = \delta_{ij}$, $i, j = 1, 2, \ldots, N^*_h$. Obviously any element v of $\mathcal{V}^*_h$ can be expressed as $v = \sum_{i=1}^{N^*_h} v(i)\varphi_i$.

Choose $\mathcal{V}_h = \mathcal{V}^*_h \cap H^1_{0,\Gamma_D}(\Omega) = \text{Span}\,\{\varphi_1, \ldots, \varphi_{N_h}\}$ where $1, \ldots, N_h$ are the nodes of the triangulation τ_h which are not on $\overline{\Gamma_D}$. We want to show that $\{\mathcal{V}_h\}_{h\to 0}$ is consistent in $H^1_{0,\Gamma_D}(\Omega)$, so that the well defined functions $w_h = \sum_{j=1\ldots N_h}(w_h)_j\varphi_j \in \mathcal{V}_h$ such that $a(w_h, v_h) = F(v_h)$, $\forall\, v_h \in \mathcal{V}_h$, *strongly* converge to w as $h \to 0$, i.e. $\|w - w_h\|_{\mathcal{V}} \to 0$.

First we introduce the space $\mathbb{V} = H^1_{0,\Gamma_D}(\Omega) \cap H^2(\Omega)$, contained and dense in $\mathcal{V} = H^1_{0,\Gamma_D}(\Omega)$ ($H^2(\Omega) = \{v \in H^1(\Omega) : D_{ij}v \in L^2(\Omega)\}$, $(u, v)_{2,\Omega} = (u, v)_{1,\Omega} + \sum_{ij}(D_{ij}u, D_{ij}v)_{0,\Omega}$, $|v|^2_{2,\Omega} = \sum_i \|D_{ii}v\|^2_{0,\Omega}$). Now let v be an element of such $\mathbb{V}$. Notice that $v \in C^0(\overline{\Omega})$ $(\ldots)$, thus the function $\Pi_h v = \sum_{i=1}^{N_h} v(i)\varphi_i$ of $\mathcal{V}_h$, interpolating v in the nodes of the triangulation, is well defined. Moreover, $\Pi_h v$ is a function of $H^1(\Omega)$ and one can measure the interpolating error using the norm of $\mathcal{V}$:

$$\|v - \Pi_h v\|_{1,\Omega} \le c_e h|v|_{2,\Omega}, \quad c_e \text{ constant.}$$

The latter inequality holds if the family of triangulations $\{\tau_h\}_{h\to 0}$ is chosen regular, that is there exists a constant c_r such that $h_T/\rho_T \le c_r$ for all $T \in \tau_h$ and h. Thus we have the operator $\mathcal{R}_h : \mathbb{V} \to \mathcal{V}_h$ required by the consistency hypothesis, it is the interpolating operator Π_h.

Observe that in case the function w is in $H^2(\Omega)$ we can say more: $\|w - w_h\|_{\mathcal{V}} \to 0$ at least at the same rate of h. In fact,

$$\|w - w_h\|_{\mathcal{V}} \le \frac{M}{m}\|w - \Pi_h w\|_{\mathcal{V}} \le \frac{M}{m} c_e h|w|_{2,\Omega}.$$

Computation of w_h

In order to compute $w_h = \sum_{j=1}^{N_h}(w_h)_j\varphi_j$ one has to solve the linear system

$$A\mathbf{x} = \mathbf{b}, \quad a_{ij} = a(\varphi_j, \varphi_i), \quad b_i = F(\varphi_i).$$

In fact, $(w_h)_j = (A^{-1}\mathbf{b})_j$. More specifically, in our example, if $s(g)$ denotes the set $\text{supp}\,(g)$, then the entries of A and $\mathbf{b}$ are:

$$\begin{array}{l} a_{ij} = \int_{s(\varphi_j)\cap s(\varphi_i)} \alpha\nabla\varphi_j\nabla\varphi_i + \int_{s(\varphi_j)\cap s(\varphi_i)} \beta\nabla\varphi_j\varphi_i + \int_{s(\varphi_j)\cap s(\varphi_i)} \gamma\varphi_j\varphi_i, \\ b_i = \int_{s(\varphi_i)} f\varphi_i + \int_{\Gamma_N\cap s(\varphi_i)} \psi\varphi_i d\sigma - \int_{s(\varphi_i)\cap s(u_\varphi)} \alpha\nabla u_\varphi\nabla\varphi_i \\ \qquad - \int_{s(\varphi_i)\cap s(u_\varphi)} \beta\nabla u_\varphi\varphi_i - \int_{s(\varphi_i)\cap s(u_\varphi)} \gamma u_\varphi\varphi_i, \end{array}$$

$1 \le i, j \le N_h$. Here the φ_i are the Lagrange basis of $\mathcal{V}_h$ ($\varphi_i(j) = \delta_{ij}$). So, the $(w_h)_j$ are the values of w_h in the nodes j ($(w_h)_j = w_h(j)$) and the matrix A is sparse, in fact for any fixed i, the number of j such that the measure of $s(\varphi_j)\cap s(\varphi_i)$ is not zero is smaller than a constant (with respect to h) dependent upon the regularity parameter c_r of the triangulations (such constant is a bound for the number of nodes j linked directly to i). But these properties are far from to be essential: in particular, more important would be to know that the matrix A is well conditioned. Unfortunately, even in case the differential problem is simply the Poisson problem ($\alpha = I$, $\beta = \mathbf{0}$, $\gamma = 0$, $\Gamma_N = \emptyset$) the matrix A has a condition number growing as $(1/h)^2$, if the Lagrange functions are used to represent w_h. (This estimate of the condition number holds more in general for the convection-diffusion problem, if the triangulations are quasi-uniform (i.e. $h_T \ge c_u h$, $\forall\, T \in \tau_h\ \forall\, h$) and regular).

So, consider an arbitrary basis $\{\tilde\varphi_i\}$ of $\mathcal{V}_h$, and represent w_h in terms of this basis: $w_h = \sum_{j=1}^{N_h} \tilde{(w_h)}_j\tilde\varphi_j$. Then, $\tilde{(w_h)}_j = (\tilde A^{-1}\tilde{\mathbf{b}})_j$ where

$$\tilde a_{ij} = a(\tilde\varphi_j, \tilde\varphi_i), \quad \tilde b_i = F(\tilde\varphi_i),$$

$$\begin{array}{l} \tilde a_{ij} = \int_{s(\tilde\varphi_j)\cap s(\tilde\varphi_i)} \alpha\nabla\tilde\varphi_j\nabla\tilde\varphi_i + \int_{s(\tilde\varphi_j)\cap s(\tilde\varphi_i)} \beta\nabla\tilde\varphi_j\tilde\varphi_i + \int_{s(\tilde\varphi_j)\cap s(\tilde\varphi_i)} \gamma\tilde\varphi_j\tilde\varphi_i, \\ \tilde b_i = \int_{s(\tilde\varphi_i)} f\tilde\varphi_i + \int_{\Gamma_N\cap s(\tilde\varphi_i)} \psi\tilde\varphi_i d\sigma - \int_{s(\tilde\varphi_i)\cap s(u_\varphi)} \alpha\nabla u_\varphi\nabla\tilde\varphi_i \\ \qquad - \int_{s(\tilde\varphi_i)\cap s(u_\varphi)} \beta\nabla u_\varphi\tilde\varphi_i - \int_{s(\tilde\varphi_i)\cap s(u_\varphi)} \gamma u_\varphi\tilde\varphi_i. \end{array}$$

If the $\tilde\varphi_i$ are such that $\mu_2(\tilde A) < \mu_2(A)$ (...), then we can solve the system $\tilde A\mathbf{x} = \tilde{\mathbf{b}}$, better conditioned than $A\mathbf{x} = \mathbf{b}$, and then, if needed, recover $\mathbf{w}_h = (w_h(j))_{j=1}^{N_h} = A^{-1}\mathbf{b}$ solving the system $S\mathbf{w}_h = \tilde{\mathbf{w}}_h = (\tilde{(w_h)}_j)_{j=1}^{N_h} = \tilde A^{-1}\tilde{\mathbf{b}}$. (Of course, all this is convenient if S is a matrix of much lower complexity than A).

Let us prove this assertion in detail. Let v_h be a generic element of $\mathcal{V}_h$ and let S be the matrix such that $\tilde{\mathbf{v}}_h = S\mathbf{v}_h$, being $\mathbf{v}_h = [(v_h)_1 \cdots (v_h)_{N_h}]^T$ and $\tilde{\mathbf{v}}_h = [\tilde{(v_h)}_1 \cdots \tilde{(v_h)}_{N_h}]^T$ such that $v_h \in \mathcal{V}_h$

$$\sum_{j=1}^{N_h} (v_h)_j\varphi_j = v_h = \sum_{j=1}^{N_h} \tilde{(v_h)}_j\tilde\varphi_j.$$

Then we have

$$\varphi_s = \sum_{j=1}^{N_h} [S^T]_{sj}\tilde\varphi_j, \quad s = 1, \ldots, N_h,$$

and therefore

$$\begin{array}{rcl} a_{ij} & = & a(\sum_{r=1}^{N_h}[S^T]_{jr}\tilde\varphi_r, \sum_{m=1}^{N_h}[S^T]_{im}\tilde\varphi_m) = \sum_{r,m}[S^T]_{im}a(\tilde\varphi_r, \tilde\varphi_m)[S]_{rj} \\ & = & \sum_{r,m}[S^T]_{im}\tilde a_{mr}[S]_{rj} = [S^T\tilde A S]_{ij}, \end{array}$$

$$b_i = F(\sum_{j=1}^{N_h}[S^T]_{ij}\tilde{\varphi}_j) = \sum_{j=1}^{N_h}[S^T]_{ij}F(\tilde{\varphi}_j) = [S^T\tilde{\mathbf{b}}]_i.$$

Thus, the equalities $A = S^T\tilde{A}S$ and $\mathbf{b} = S^T\tilde{\mathbf{b}}$ must hold, and the thesis follows.

In the Poisson case, $-\Delta u = f$, $x \in \Omega$, $u = \varphi$, $x \in \partial\Omega$, a basis $\{\tilde{\varphi}_i\}$ for $\mathcal{V}_h$ can be introduced yielding a matrix $\tilde{A}$ whose condition number $\mu_2(\tilde{A})$ grows like $(\log_2(1/h))^2$. (An analogous result in the convection-diffusion case (a not symmetric) in 1995 was not known!). We now see (not in all details) that this is possible by using a particular family of triangulations τ_h.

Let τ_0 be a rada triangulation of Ω. Let us define τ_1. For each triangle T of τ_0 draw the triangle whose verteces are the middle points of the sides of T. The four triangles you see (similar to T) are the triangles of τ_1. Note that if h_0 is the diameter of τ_0 ($h_0 = \max\{h_T : T \in \tau_0\}$), then h_1, the diameter of τ_1, is equal to $2^{-1}h_0$. Note also that the nodes of τ_0 are nodes of τ_1; the new nodes of τ_1 are the middle points of the sides of τ_0. We can continue in this way, and define the triangulations τ_2, τ_3, ..., τ_j, ... (obviously, τ_j is an abbreviation for τ_{h_j}). The diameter of the generic τ_j is $h_j = 2^{-j}h_0$, and the family of triangulations $\{\tau_j\}_{j=0}^{+\infty}$ is regular and quasi-uniform.

To each τ_j we can associate the space $\mathcal{V}_j = \mathcal{V}_{h_j}^* \cap H_0^1(\Omega)$ of the functions which are continuous, null on $\partial\Omega$, and degree-1 polynomials in each $T \in \tau_j$. Note that $\mathcal{V}_j \subset \mathcal{V}_{j+1}$.

Let x_{jk}, $k \in \mathcal{I}_j = \{1, \ldots, N_j\} \subset \{1, \ldots, N_j^*\}$, denote the generic inner node of the triangulation τ_j, and $\{\varphi_{jk} : k \in \mathcal{I}_j\}$ the Lagrange basis of $\mathcal{V}_j$, $\varphi_{jk}(x_{jl}) = \delta_{kl}$, $k, l \in \mathcal{I}_j$. Obviously, any $v \in \mathcal{V}_j$ can be represented as $v = \sum_{k\in\mathcal{I}_j} v(x_{jk})\varphi_{jk}$, and, if Π_j is the interpolating operator, then

$$v \in \mathcal{C}^0(\overline{\Omega}) \;\rightarrow\; \Pi_j(v) = \sum_{k\in\mathcal{I}_j} v(x_{jk})\varphi_{jk}.$$

Instead of $v(x_{jk})$ we will write shortly v_{jk}.

Now that all is defined, consider a function $v \in \mathcal{V}_{j+1}$ and observe that

$$\sum_{k\in\mathcal{I}_{j+1}} v_{j+1,k}\varphi_{j+1,k} = v = \Pi_{j+1}v = \Pi_j v + (v - \Pi_j v) = \sum_{k\in\mathcal{I}_j} v_{jk}\varphi_{jk} + (v - \Pi_j v).$$

Now the question is: what must we add to $\mathcal{V}_j$ in order to obtain $\mathcal{V}_{j+1}$? This question can be reduced to: what elements of $\{\varphi_{j+1,k} : k \in \mathcal{I}_{j+1}\}$ are needed to represent $v - \Pi_j v$?

Let x be a point of $\overline{\Omega}$ and let T be a triangle of τ_j including x. Let us observe the above quantities and, in particular, the function $v - \Pi_j v$ on T. Call $x_{jk_1}, x_{jk_2}, x_{jk_3}$ ($k_1, k_2, k_3 \in \mathcal{I}_j$) the verteces of T. Note that they are nodes also of τ_{j+1}, thus $x_{jk_i} = x_{j+1,\rho(k_i)}$, for some $\rho(k_i) \in \mathcal{I}_{j+1}^o = \{k \in \mathcal{I}_{j+1} : x_{j+1,k}$ is a node of $\tau_j\}$. Call $x_{j+1,\sigma(k_1k_2)}$, $\sigma(k_1k_2) \in \mathcal{I}_{j+1}^n = \mathcal{I}_{j+1}\backslash\mathcal{I}_{j+1}^o$, the middle point of the side $\overline{x_{jk_1}\,x_{jk_2}}$ of T, which is a *new node*, a node of τ_{j+1}, but not of τ_j. Draw the restrictions to T of the functions v and $\Pi_j v$. Then it is clear that, on T,

$$\begin{aligned} v - \Pi_j v &= [v_{j+1,\sigma(k_1k_2)} - \tfrac{1}{2}(v_{j,k_1} + v_{j,k_2})]\varphi_{j+1,\sigma(k_1k_2)} \\ &\quad +[v_{j+1,\sigma(k_2k_3)} - \tfrac{1}{2}(v_{j,k_2} + v_{j,k_3})]\varphi_{j+1,\sigma(k_2k_3)} \\ &\quad +[v_{j+1,\sigma(k_3k_1)} - \tfrac{1}{2}(v_{j,k_3} + v_{j,k_1})]\varphi_{j+1,\sigma(k_3k_1)} \\ &= \tilde{v}_{j,\sigma(k_1k_2)}\varphi_{j+1,\sigma(k_1k_2)} + \tilde{v}_{j,\sigma(k_2k_3)}\varphi_{j+1,\sigma(k_2k_3)} + \tilde{v}_{j,\sigma(k_3k_1)}\varphi_{j+1,\sigma(k_3k_1)} \\ &= \textstyle\sum_{k\in\mathcal{I}_{j+1}^n} \tilde{v}_{jk}\varphi_{j+1,k} \end{aligned}$$

where

$$\tilde{v}_{jk} = v_{j+1,k} - \frac{1}{2}(v_{j,k'} + v_{j,k''}), \quad k \in \mathcal{I}_{j+1}^n,$$

and $k', k'' \in \mathcal{I}_j$ are such that $x_{jk'} = x_{j+1,\rho(k')}$, $x_{jk''} = x_{j+1,\rho(k'')}$ are the extreme points of the side of τ_j having $x_{j+1,k}$ as middle point. Thus, if $\psi_{jk} := \varphi_{j+1,k} = \varphi_{j+1,\sigma(k'k'')}$, $k \in \mathcal{I}_{j+1}^n$, then

$$v \in \mathcal{V}_{j+1} \;\Rightarrow\; v = \sum_{k\in\mathcal{I}_{j+1}} v_{j+1,k}\varphi_{j+1,k} = \sum_{k\in\mathcal{I}_j} v_{jk}\varphi_{jk} + \sum_{k\in\mathcal{I}_{j+1}^n} \tilde{v}_{jk}\psi_{j,k}.$$

It follows that $\mathcal{V}_{j+1} = \mathcal{V}_j + \mathcal{W}_j$, $\mathcal{W}_j := \mathrm{Span}\,\{\psi_{jk} : k \in \mathcal{I}_{j+1}^n\}$, and the set $\{\varphi_{jk} : k \in \mathcal{I}_j\} \cup \{\psi_{j,k} : k \in \mathcal{I}_{j+1}^n\}$ is an alternative basis of $\mathcal{V}_{j+1}$. So, the answer to the question is: the Lagrangian functions of $\mathcal{V}_{j+1}$ corresponding to the new nodes.

Observe that the v_{jk}, $k \in \mathcal{I}_j$, and the $\tilde{v}_{jk}$, $k \in \mathcal{I}_{j+1}^n$, can be computed from the $v_{j+1,k}$, $k \in \mathcal{I}_{j+1}$, via the formulas

$$\begin{array}{l} v_{jk} = v_{j+1,\rho(k)}, \quad k \in \mathcal{I}_j \\ \tilde{v}_{j,k} = v_{j+1,k} - \frac{1}{2}[v_{j+1,\rho(k')} + v_{j+1,\rho(k'')}], \quad k \in \mathcal{I}_{j+1}^n. \end{array}$$

Viceversa, the $v_{j+1,k}$, $k \in \mathcal{I}_{j+1}$, can be computed from the v_{jk}, $k \in \mathcal{I}_j$, an from the $\tilde{v}_{jk}$, $k \in \mathcal{I}_{j+1}^n$, via the formulas:

$$\begin{array}{l} v_{j+1,\rho(k)} = v_{jk}, \quad k \in \mathcal{I}_j \\ v_{j+1,k} = \tilde{v}_{j,k} + \frac{1}{2}[v_{j,k'} + v_{j,k''}], \quad k \in \mathcal{I}_{j+1}^n. \end{array}$$

These formulas can be written in matrix form:

$$\begin{bmatrix} v_{jk} \\ \quad k \in \mathcal{I}_j \\ \tilde{v}_{jk} \\ \quad k \in \mathcal{I}_{j+1}^n \end{bmatrix} = \begin{bmatrix} I_{|\mathcal{I}_j|} & 0 \\ B & I_{|\mathcal{I}_{j+1}^n|} \end{bmatrix} \begin{bmatrix} v_{j+1,\rho(k)} \\ \quad k \in \mathcal{I}_j \\ v_{j+1,k} \\ \quad k \in \mathcal{I}_{j+1}^n \end{bmatrix}, \quad \begin{bmatrix} v_{j+1,\rho(k)} \\ \quad k \in \mathcal{I}_j \\ v_{j+1,k} \\ \quad k \in \mathcal{I}_{j+1}^n \end{bmatrix} = \begin{bmatrix} I_{|\mathcal{I}_j|} & 0 \\ -B & I_{|\mathcal{I}_{j+1}^n|} \end{bmatrix} \begin{bmatrix} v_{jk} \\ \quad k \in \mathcal{I}_j \\ \tilde{v}_{jk} \\ \quad k \in \mathcal{I}_{j+1}^n \end{bmatrix}$$

where each row of B has only two nonzero elements, both equal to $-\frac{1}{2}$.

Now let J be a positive integer. We are ready to introduce a basis $\tilde{\varphi}_{J,k}$, $k \in \mathcal{I}_J$, of the space $\mathcal{V}_J$ yielding a matrix $\tilde{A}$, $\tilde{a}_{r,s} = a(\tilde{\varphi}_{J,s}, \tilde{\varphi}_{J,r})$, $r, s \in \mathcal{I}_J$, whose condition number in the Poisson case grows like $O((\log_2(1/h_J))^2) = O((J + \log_2(1/h_0))^2)$, and thus is smaller than the condition number of the matrix A, $a_{r,s} = a(\varphi_{J,s}, \varphi_{J,r})$, $r, s \in \mathcal{I}_J$, yielded by the Lagrange basis $\varphi_{J,k}$, $k \in \mathcal{I}_J$ of $\mathcal{V}_J$ (recall that $\mu_2(A) = O((1/h_J)^2) = O((2^J/h_0)^2)$).

Observe that if $v \in \mathcal{V}_J$ then

$$\begin{array}{rcl} \sum_{k\in\mathcal{I}_J} v_{Jk}\varphi_{Jk} & = & v = \Pi_J v = \Pi_0 v + \sum_{j=0}^{J-1}(\Pi_{j+1}v - \Pi_j v) \\ & = & \sum_{k\in\mathcal{I}_0} v_{0k}\varphi_{0k} + \sum_{j=0}^{J-1}\sum_{k\in\mathcal{I}_{j+1}^n} \tilde{v}_{jk}\psi_{j,k}, \end{array}$$

$$\psi_{jk} = \varphi_{j+1,k}, \quad k \in \mathcal{I}_{j+1}^n, \quad j = 0, \ldots, J-1.$$

It follows that $\mathcal{V}_J$ admits the representation

$$\mathcal{V}_J = \mathcal{V}_{J-1} + \mathcal{W}_{J-1} = \mathcal{V}_{J-2} + \mathcal{W}_{J-2} + \mathcal{W}_{J-1} = \mathcal{V}_0 + \mathcal{W}_0 + \ldots + \mathcal{W}_{J-1}$$

and the set $\{\tilde{\varphi}_{J,k} : k \in \mathcal{I}_J\} := \{\varphi_{0k} : k \in \mathcal{I}_0\} \cup \{\psi_{0,k} : k \in \mathcal{I}_1^n\} \cup \cdots \cup \{\psi_{J-1,k} : k \in \mathcal{I}_J^n\}$ is an alternative basis of $\mathcal{V}_J$.

Remark. Observe that if

$$\mathbf{v}_J = (v_{J,k})_{k\in\mathcal{I}_J},$$
$$\tilde{\mathbf{v}}_J = (\tilde{v}_{J,k})_{k\in\mathcal{I}_J} = ((v_{0,k})_{k\in\mathcal{I}_0}\ (\tilde{v}_{0,k})_{k\in\mathcal{I}_1^n}\ \cdots\ (\tilde{v}_{J-1,k})_{k\in\mathcal{I}_J^n}),$$

then $\tilde{\mathbf{v}}_J = S\mathbf{v}_J = E_0P_0E_1P_1\cdots E_{J-1}P_{J-1}\mathbf{v}_J$ where the P_k are permutation matrices, the E_k are matrices of the form

$$E_0 = \begin{bmatrix} I_{|\mathcal{I}_0|} & 0 & \\ B_0 & I_{|\mathcal{I}_1^n|} & \\ & & I \end{bmatrix},\ E_1 = \begin{bmatrix} I_{|\mathcal{I}_1|} & 0 & \\ B_1 & I_{|\mathcal{I}_2^n|} & \\ & & I \end{bmatrix},\ \ldots,\ E_{J-1} = \begin{bmatrix} I_{|\mathcal{I}_{J-1}|} & 0 \\ B_{J-1} & I_{|\mathcal{I}_J^n|} \end{bmatrix}$$

and the B_k, in the definition of the E_k, have only two nonzero elements for each row, both equal to $-\frac{1}{2}$. So, $\mathbf{v}_J$ can be computed from $\tilde{\mathbf{v}}_J$ (as well as $\tilde{\mathbf{v}}_J$ can be computed from $\mathbf{v}_J$) with $2(|\mathcal{I}_J| - |\mathcal{I}_0|)$ divisions by 2.

The transform of $\mathbf{v}_J$ into $\tilde{\mathbf{v}}_J$ is described in detail here below:

$$\mathbf{v}_J = \begin{bmatrix} v_{J,k} \\ k \in \mathcal{I}_J \end{bmatrix} \rightarrow P_{J-1}\mathbf{v}_J = \begin{bmatrix} v_{J,\rho(k)} \\ k \in \mathcal{I}_{J-1} \\ --- \\ v_{J,k} \\ k \in \mathcal{I}_J^n \end{bmatrix} \rightarrow E_{J-1}P_{J-1}\mathbf{v}_J = \begin{bmatrix} v_{J-1,k} \\ k \in \mathcal{I}_{J-1} \\ --- \\ \tilde{v}_{J-1,k} \\ k \in \mathcal{I}_J^n \end{bmatrix}$$

$$\rightarrow P_{J-2}E_{J-1}P_{J-1}\mathbf{v}_J = \begin{bmatrix} v_{J-1,\rho(k)} \\ k \in \mathcal{I}_{J-2} \\ --- \\ v_{J-1,k} \\ k \in \mathcal{I}_{J-1}^n \\ --- \\ \tilde{v}_{J-1,k} \\ k \in \mathcal{I}_J^n \end{bmatrix} \rightarrow E_{J-2}P_{J-2}E_{J-1}P_{J-1}\mathbf{v}_J = \begin{bmatrix} v_{J-2,k} \\ k \in \mathcal{I}_{J-2} \\ --- \\ \tilde{v}_{J-2,k} \\ k \in \mathcal{I}_{J-1}^n \\ --- \\ \tilde{v}_{J-1,k} \\ k \in \mathcal{I}_J^n \end{bmatrix}$$

$$\cdots \rightarrow E_0P_0\cdots E_{J-1}P_{J-1}\mathbf{v}_J = \begin{bmatrix} v_{0,k} \\ k \in \mathcal{I}_0 \\ --- \\ \tilde{v}_{0,k} \\ k \in \mathcal{I}_1^n \\ --- \\ \cdots \\ --- \\ \tilde{v}_{J-1,k} \\ k \in \mathcal{I}_J^n \end{bmatrix}.$$

Theorem. Let $\tau_j, \mathcal{V}_j, \Pi_j$, $j = 0, 1, \ldots, J$, be the triangulations of Ω, the subspaces of $\mathcal{V} = H_0^1(\Omega)$ and the interpolating operators $C^0(\overline{\Omega}) \rightarrow \mathcal{V}_j$ defined above. For any $v \in \mathcal{V}_J$ set

$$\hat{\|}v\hat{\|}^2 = |\Pi_0 v|_{1,\Omega}^2 + \sum_{j=0}^{J-1} \sum_{k\in\mathcal{I}_{j+1}^n} |(\Pi_{j+1}v - \Pi_j v)(x_{j+1,k})|^2 = |\Pi_0 v|_{1,\Omega}^2 + \sum_{j=0}^{J-1} \sum_{k\in\mathcal{I}_{j+1}^n} |\tilde{v}_{j,k}|^2.$$

Then there exist two positive constants c_1, c_2 (depending only on the angles of τ_0) such that

$$c_1 \frac{\hat{\|}v\hat{\|}^2}{J^2} \leq |v|^2_{1,\Omega} \leq c_2 \hat{\|}v\hat{\|}^2.$$

This inequality, involving the coefficients $\tilde{v}_{J,k}$ of $v \in \mathcal{V}_J$ with respect to the hierarchical basis $\tilde{\varphi}_{J,k}$, $k \in \mathcal{I}_J$, is due to Yserentant. It allows us to evaluate the condition number of $\tilde{A}$, $\tilde{a}_{r,s} = a(\tilde{\varphi}_{J,s}, \tilde{\varphi}_{J,r})$, $r, s \in \mathcal{I}_J$, in the Poisson case where $a(u,v) = \int_\Omega \nabla u \nabla v$.

First note that

$$\begin{array}{rcl} |\Pi_0 v|^2_{1,\Omega} & = & |\sum_{k\in\mathcal{I}_0} v_{0,k}\varphi_{0,k}|^2_{1,\Omega} = \int_\Omega \nabla(\sum_{k\in\mathcal{I}_0} v_{0,k}\varphi_{0,k}) \cdot \nabla(\sum_{k\in\mathcal{I}_0} v_{0,k}\varphi_{0,k}) \\ & = & \sum_{k,s\mathcal{I}_0} v_{0k} v_{0s} \int_\Omega \nabla\varphi_{0k} \nabla\varphi_{0s}, \end{array}$$

thus the Yserentant inequality can be rewritten as follows:

$$c_1 \frac{\tilde{\mathbf{v}}_J^T \begin{bmatrix} N & 0 \\ 0 & I \end{bmatrix} \tilde{\mathbf{v}}_J}{J^2} \leq \tilde{\mathbf{v}}_J^T M \tilde{\mathbf{v}}_J \leq c_2 \tilde{\mathbf{v}}_J^T \begin{bmatrix} N & 0 \\ 0 & I \end{bmatrix} \tilde{\mathbf{v}}_J \qquad \text{(Y)}$$

where $N = (\int_\Omega \nabla\varphi_{0r} \cdot \nabla\varphi_{0,s})_{r,s\in\mathcal{I}_0}$ and $M = (\int_\Omega \nabla\tilde{\varphi}_{J,r} \cdot \nabla\tilde{\varphi}_{J,s})_{r,s\in\mathcal{I}_J}$. Note that N and M are positive definite matrices. Note also that in case of the Poisson differential problem $-\Delta u = f$, $x \in \Omega$, $u = \varphi$, $x \in \partial\Omega$, the form a is simply $a(u,v) = \int_\Omega \nabla u \nabla v$, i.e. we have the continuous problem

$$w \in \mathcal{V} = H^1_{0,\Gamma_D}(\Omega) \mid \int_\Omega \nabla w \nabla v = \int_\Omega fv - \int_\Omega \nabla u_\varphi \nabla v, \quad \forall\, v \in \mathcal{V} = H^1_{0,\Gamma_D}(\Omega)$$

which is reduced first to the discrete problem

$$w_J \in \mathcal{V}_J \mid \int_\Omega \nabla w_J \nabla v_J = \int_\Omega f v_J - \int_\Omega \nabla u_\varphi \nabla v_J, \quad \forall\, v_J \in \mathcal{V}_J$$

and then, via the representation $w_J = \sum_{k\in\mathcal{I}_J} (\tilde{w_J})_k \tilde{\varphi}_{J,k}$, to the linear system

$$\begin{array}{l} \tilde{A}\mathbf{x} = \tilde{\mathbf{b}}, \quad \tilde{a}_{r,s} = \int_\Omega \nabla\tilde{\varphi}_{J,r} \nabla\tilde{\varphi}_{J,s}, \quad \tilde{b}_r = \int_\Omega f\tilde{\varphi}_{J,r} - \int_\Omega \nabla u_\varphi \nabla\tilde{\varphi}_{J,r}, \\ (\tilde{w_J})_k = (\tilde{A}^{-1}\tilde{\mathbf{b}})_k. \end{array}$$

Observe that the coefficient matrix $\tilde{A}$ of this system is exactly the matrix M in (Y). Now we prove that $\mu_2(M) = O((\log_2 \frac{1}{h_J})^2)$.

Consider the Cholesky factorization of N, $N = L_N L_N^T$, and note that

$$L := \begin{bmatrix} L_N & \\ & I \end{bmatrix} \;\Rightarrow\; LL^T = \begin{bmatrix} N & \\ & I \end{bmatrix}.$$

Set $\mathbf{z} = L^T\tilde{\mathbf{v}}_J$, $\tilde{\mathbf{v}}_J = L^{-T}\mathbf{z}$. By (Y), for all vectors $\mathbf{z} \neq \mathbf{0}$ we have

$$c_1 \frac{1}{J^2} \leq \frac{\mathbf{z}^T L^{-1} M L^{-T} \mathbf{z}}{\mathbf{z}^T\mathbf{z}} \leq c_2.$$

Thus, if λ is any eigenvalue of the positive definite matrix $L^{-1}ML^{-T}$, then

$$\frac{c_1}{J^2} \leq \lambda \leq c_2,$$

and this result implies that the condition number of $L^{-1}ML^{-T}$ is bounded by $\frac{c_2}{c_1}J^2$. Since L_N is a small matrix and its dimension does not depend on J, it follows that $\mu_2(M) \leq cJ^2 = c(\log_2 \frac{h_0}{h_J})^2$ for some constant c.