Journal of Computational and Applied Mathematics 394 (2021) 113523

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Efficiency of nonparametric finite elements for optimal-order enforcement of Dirichlet conditions on curvilinear boundaries

Vitoriano Ruas [a,*,1], Marco Antonio Silva Ramos [b]

[a] Sorbonne Université, CNRS, UMR 7190-Institut Jean Le Rond d'Alembert, Paris, France
[b] DCC, Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brazil

ARTICLE INFO

Article history:
Received 22 May 2020
Received in revised form 4 February 2021

Keywords:
Curved domain
Dirichlet conditions
Nonparametric finite-elements
Optimal order
Petrov-Galerkin formulation
Straight-edged simplex

ABSTRACT

In recent papers (see e.g. Ruas (2020a) and Ruas (2020b)) a nonparametric technique of the Petrov–Galerkin type was analyzed, whose aim is the accuracy enhancement of higher order finite element methods to solve boundary value problems with Dirichlet conditions, posed in smooth curved domains. In contrast to parametric elements, it employs straight-edged triangular or tetrahedral meshes fitting the domain. In order to attain best-possible orders greater than one, piecewise polynomial trial-functions are employed, which interpolate the Dirichlet conditions at points of the true boundary. The test-functions in turn are defined upon the standard degrees of freedom associated with the underlying method for polytopic domains. As a consequence, when the problem at hand is self-adjoint a non symmetric linear system has to be solved. This paper is primarily aimed at showing that in this case, an efficient symmetrization of the solution procedure can be achieved by means of a fast converging iterative method. In order to illustrate the great generality of our nonparametric approach, experimentation is presented with a finite element method having degrees of freedom other than nodal values. More specifically we consider a nonconforming quadratic element in the solution of the three-dimensional Poisson equation. The performance evaluation however is conducted as well for two versions of the classical conforming quadratic method, namely, the nonparametric Petrov–Galerkin formulation considered in Ruas (2020b) and the standard isoparametric one. The study of this symmetrization is completed by an optimal error estimation in the broken H^1-norm for the nonparametric version of the nonconforming method, which had not been addressed in previous work.

1. Introduction

In the past decades Petrov–Galerkin formulations of boundary value problems showed to be a powerful tool to overcome difficulties brought about by the space discretization of certain types of partial differential equations. A significant illustration is provided by the families of methods proposed by Franca and Hughes and collaborators in the late eighties for the finite-element modeling of various problems in Continuum Mechanics, in particular as a popular alternative to Galerkin methods for viscous incompressible flow (see e.g. [1]). The outstanding contributions in the seventies of Babuška (see e.g. [2]) and Brezzi [3], among other authors, were decisive to provide a theoretical background that allowed to formally justify the reliability of Petrov–Galerkin formulations, namely, the so-called *inf–sup* condition.

* Corresponding author.
E-mail address: vitoriano.ruas@upmc.fr (V. Ruas).

[1] This work was partially supported by CNPq, the National Research Council of Brazil.

https://doi.org/10.1016/j.cam.2021.113523
0377-0427/

In a series of papers published since 2017 (cf. Ruas [4,5] and Ruas and Silva Ramos [6]) a nonparametric technique of the Petrov–Galerkin type was introduced, in order to enhance the accuracy of higher order finite element methods to solve boundary value problems with Dirichlet conditions, posed in smooth curved domains. In contrast to parametric elements, it employs straight-edged triangular or tetrahedral meshes fitting the domain. In order to attain best-possible orders greater than one, piecewise polynomial trial-functions are employed, which interpolate the Dirichlet conditions at points of the true boundary. In the two-dimensional case this kind of trial-functions is similar to the one also employed as test-functions by the method known as *interpolated boundary conditions* studied in [7]. However, in spite of being very intuitive and known since the seventies (cf. [8] and [9]), the lack of an extension to three-dimensional problems seems to have inhibited its use among practitioners. In contrast, the test-functions for our method are defined upon the degrees of freedom associated with the underlying finite element method for the mesh forming a polytope equal to the union of straight-edged simplexes. This polytope fits the curved domain in such a manner that all of its vertexes lie on the boundary of the latter. In doing so the integration domain is restricted to this polytope, thereby rendering method's implementation straightforward in both two- and three-dimensional geometries. Moreover only polynomial algebra is necessary, while best-order approximations can be obtained for non-restrictive choices of boundary nodal points.

Generally speaking, the Petrov–Galerkin methodology studied in this work is designed to enforce Dirichlet conditions in the form of prescribed boundary degrees of freedom of various types, in connection with methods of order greater than one in problem's natural norm, for a wide spectrum of boundary value problems. According to numerous numerical experiments reported in previous papers, including those cited above, it showed to be fully reliable in different contexts. It also appeared to be superior to well known techniques to tackle the same kind of problem, in case they exist. For instance in [10] and [11] comparisons of this method with the isoparametric version of the finite element method for second order boundary value problems revealed that the former is more accurate than the latter. As a matter of fact, as far as the authors can see, the new method's only real demerit is the fact that non symmetric linear systems have to be solved, even when the problem at hand is self-adjoint. The primary aim of this paper is to show that, in such a case, an efficient symmetrization of the solution procedure can be achieved by means of a fast converging iterative method.

So far the nonparametric approach considered in this work was only studied as applied to finite elements, which are conforming in the case of polytopic domains. However our technique to handle Dirichlet conditions prescribed on curved boundaries has a wide scope of applicability. This feature is exemplified here by applying such a symmetrization procedure to the solution of the three-dimensional Poisson equation by a nonconforming quadratic finite element with degrees of freedom other than nodal values. This method is based on the same type of piecewise quadratic interpolation as the one introduced in [12], in order to represent the velocity in the framework of the stable solution of incompressible viscous flow problems. Actually the corresponding velocity representation enriched by the quartic bubble-functions of the tetrahedra combined with a discontinuous piecewise linear pressure in each tetrahedron, is a sort of nonconforming three-dimensional analog of the popular conforming Crouzeix–Raviart mixed finite element [13] for solving viscous flow problems in two-dimension space. After carrying out a numerical validation of the Petrov–Galerkin approach combined with the symmetrization procedure for a nonparametric version of this nonconforming quadratic method, its efficiency as compared to the isoparametric and nonparametric versions of the conforming quadratic element is examined. An error estimation in the broken H^1-norm for the nonconforming method in the case of a curved domain completes these studies.

An outline of the paper is as follows. Section 2 is devoted to some preliminaries, in which we first recall the model Poisson equation in a smooth three-dimensional domain and present some pertaining notations; several definitions, notations and assumptions related to the finite element meshes are also introduced therein. In Section 3 we describe our technique to handle the Dirichlet boundary conditions for the model problem, in connection with the nonconforming quadratic finite element method; the underlying approximate problem is posed and corresponding stability and well-posedness results are given. In Section 4 we address the symmetrization solution procedure and validate the resulting numerical scheme. In Section 5 the performance of such a scheme is compared with the one of the asymmetric solution procedure, both extended to the standard conforming quadratic Lagrange element. Error estimates for the nonconforming method in the Petrov–Galerkin formulation to treat curved boundaries are given in Section 6. Finally in Section 7 we draw some conclusions from the whole work.

2. Preliminaries

In this section we specify the model problem considered in this work and supply some material to be used in the sequel.

2.1. The model problem and pertaining notations

Let us consider as a model the Poisson equation with Dirichlet boundary conditions in a three-dimensional domain Ω with boundary Γ having suitable regularity properties, that is,

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u = g \text{ on } \Gamma, \end{cases} \tag{1}$$

where f and g are given functions defined in Ω and on Γ.

For quadratic finite element methods our technique is most effective in case $u \in H^3(\Omega)$. In order to make sure that u possesses such a regularity property we shall assume that $f \in H^1(\Omega)$ and $g \in H^{5/2}(\Gamma)$ (cf. [14]). We observe that, owing to the Sobolev Embedding Theorem [14], g is necessarily continuous. We must further assume that Γ is at least of the C^1-class. Actually, more than this, we make the assumption that the principal curvatures of Γ (cf. [15]) are uniquely defined almost everywhere. Notice that in doing so we are not requiring that Γ be of the C^2-class.

Throughout this article $\|\cdot\|_0$ stands for the standard norm of $L^2(\Omega)$. Furthermore $\|\cdot\|_{r,D}$ and $|\cdot|_{r,D}$ represent, respectively, the standard norm and semi-norm of Sobolev space $H^r(D)$ (cf. [14]), for $r \in \Re^+$ with $H^0(D) = L^2(D)$, D being any bounded subset of $\Re^3$. We also denote by $\|\cdot\|_{m,p,D}$ the usual norm of $W^{m,p}(D)$ for $m \in \mathbb{N}^*$ and $p \in [1, \infty] \setminus \{2\}$ with $W^{0,p}(D) = L^p(D)$. Whenever D is Ω the subscript , D is dropped.

2.2. Meshes and related notions

Let us be given a mesh $\mathcal{T}_h$ consisting of straight-edged tetrahedra satisfying the usual compatibility conditions (see e.g. [16]). Every element of $\mathcal{T}_h$ is to be viewed as a closed set. Moreover this mesh is assumed to fit Ω in such a way that all the vertexes of the polyhedron $\cup_{T\in\mathcal{T}_h} T$ lie on Γ. We denote the interior of this union set by Ω_h and define $\tilde{\Omega}_h := \Omega \cap \Omega_h$ together with $\Omega'_h := \Omega \cup \Omega_h$. The boundaries of Ω_h and $\tilde{\Omega}_h$ are respectively denoted by Γ_h and $\tilde{\Gamma}_h$ and moreover $\Gamma'_h := \bar{\Omega}_h \cap \Gamma$. $\mathcal{T}_h$ is assumed to belong to a regular family of partitions in the sense of [16], though not necessarily quasi-uniform. The boundary of every $\forall T \in \mathcal{T}_h$ is represented by ∂T, while h_T is the diameter of T and $h := \max_{T\in\mathcal{T}_h} h_T$. We make the non essential and yet reasonable assumption that any element in $\mathcal{T}_h$ have at most either one edge or one face contained in Γ_h.

Let $\mathcal{S}_h$ be the subset of $\mathcal{T}_h$ consisting of tetrahedra having one face on Γ_h and $\mathcal{R}_h$ be the subset of $\mathcal{T}_h \setminus \mathcal{S}_h$ consisting of tetrahedra having exactly one edge on Γ_h. We further set $\mathcal{O}_h := \mathcal{S}_h \cup \mathcal{R}_h$. Notice that, owing to our initial assumption, the interior of any tetrahedron in $\mathcal{T}_h \setminus \mathcal{O}_h$ has an empty intersection with Γ_h. For every $T \in \mathcal{S}_h$ we denote by O_T the vertex of T not belonging to Γ. Finally we introduce the notations $\|\cdot\|_{0,h}$ (resp. $\|\cdot\|_{0,\tilde{h}}$) for the standard norm of $L^2(\Omega_h)$ (resp. $L^2(\tilde{\Omega}_h)$).

Remark 1. Even though for practical purposes this is by no means necessary, in all the constructions and analyzes given hereafter, we shall assume that the mesh is sufficiently fine. We refer to [11] for a precise quantification of the assumed smallness of h. ■

We also need some definitions and auxiliary results regarding the set $(\Omega \setminus \Omega_h) \cup (\Omega_h \setminus \Omega)$.

With every edge e of the mesh contained in Γ_h we associate a closed plane set δ_e containing e, delimited by Γ and e itself. The plane of δ_e can be arbitrarily chosen about e. However for better results it should be close to the bisector of the faces of the pair of elements in $\mathcal{S}_h$ intersecting at e, which can eventually be a face shared by both. Such a choice will be assumed throughout this work. We also define $\tilde{\delta}_e := \delta_e \cap \Omega$. In Fig. 1 we illustrate one out of three plane sets δ_e corresponding to the edges of the faces F_T and $F_{T'}$ contained in Γ_h of tetrahedra T and T' belonging to $\mathcal{S}_h$. More precisely δ_e is depicted for the edge e common to F_T and $F_{T'}$.

Further, for every $T \in \mathcal{S}_h$, we define a closed set Δ_T delimited by Γ, ∂T and the plane sets $\tilde{\delta}_e$ associated with the edges of F_T, as illustrated in Fig. 1. In this manner we can assert that, if Ω is convex, Ω_h is a proper subset of Ω and $\tilde{\Omega}$ is the union of the disjoint sets Ω_h and $\cup_{T\in\mathcal{S}_h} \Delta_T$. Otherwise $\Omega_h \setminus \Omega$ is a nonempty set containing subsets of $T \in \mathcal{S}_h$ whose volume is an $O(h_T^4)$ and subsets of $T \in \mathcal{R}_h$ whose volume is an $O(h_T^5)$, both types of subsets corresponding to non-convex portions of Γ. Whatever the case, the above configurations are of merely academic interest and carry no practical meaning, as much as the sets $T_\Delta := T \cup \Delta_T \ \forall T \in \mathcal{S}_h$ or $T_\Delta := T \cup \delta_e \ \forall T \in \mathcal{R}_h$, $\tilde{T} := T \cap \Omega \ \forall T \in \mathcal{O}_h$ and $\Delta'_T := \Delta_T \setminus \Omega$.

Referring to Figs. 2 and 3 for illustrations in particular cases, $\mathcal{T}_h$ is supposed to fulfill the following reasonable conditions:

Assumption$^+$: h is small enough for the intersection P with Γ of the half line s with origin at O_T passing through any point $M \in F_T$ to be uniquely defined $\forall T \in \mathcal{S}_h$.

Assumption^{++}: h is small enough for the intersection $Q \in \delta_e$ with Γ of the half line r perpendicular to e with origin at any point $N \in e$ to be uniquely defined.

We recall a result formally established in [11], according to which there exists a mesh-independent constant C_Γ such that $length(\overline{MP}) \le C_\Gamma h_T^2$ and $length(\overline{NQ}) \le C_\Gamma h_T^2$.

3. A nonconforming method with mean-value degrees of freedom

In this section we apply our technique to handle Dirichlet conditions on curved boundaries to a nonconforming method with degrees of freedom other than function nodal values. Incidentally we note that for many well known nonconforming finite element methods the construction of an isoparametric counterpart brings no improvement. This does not prevent suitable parametric elements from being successfully employed in this case. However to the best of author's knowledge studies in this direction are incipient. This fact motivates us to show here that our technique for handling curvilinear boundaries can be optimally extended in a straightforward manner to finite element methods, which are nonconforming

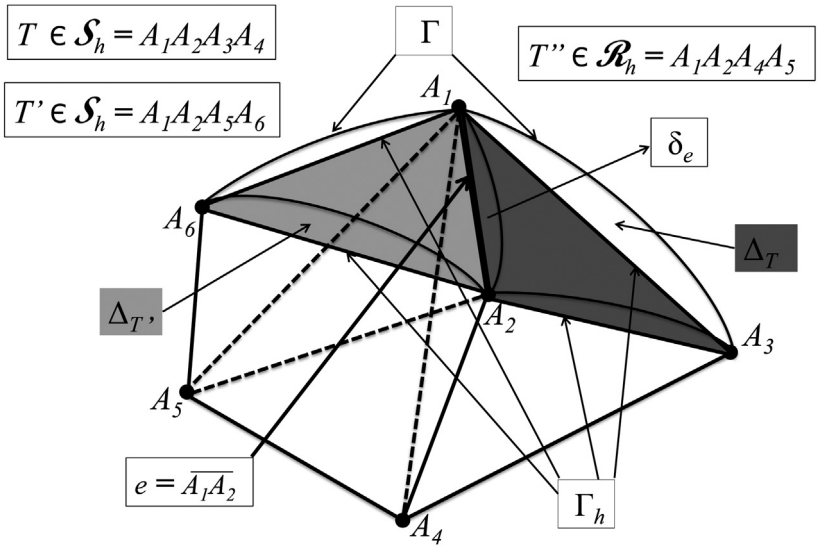

Fig. 1. Sets Δ_T, $\Delta_{T'}$, δ_e for $T, T' \in \mathcal{S}_h$ having a common edge e and a tetrahedron T'' in $\mathcal{R}_h$.

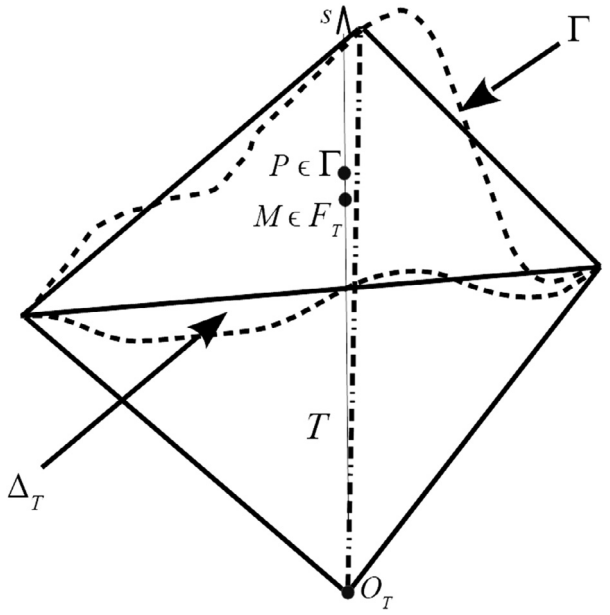

Fig. 2. $P :=$ intersection with Γ of the line joining vertex O_T to the centroid M of $F_T \subset \Gamma_h$.

even in the case of polytopes. We use such a nonconforming approach to solve the model problem (1), confining ourselves to the case of homogeneous boundary conditions for the sake of simplicity, though without any loss of essential aspects.

To begin with we recall the space V_h of test-functions defined in Ω_h, associated with the method under consideration.

F and e being a face and an edge of a tetrahedron $T \in \mathcal{T}_h$ respectively, we denote by M the centroid of F, by A and B the end-points of e and by N the mid-point of e. Now any function $v \in V_h$ restricted to every T is a polynomial of degree less than or equal to two, defined upon the following set of degrees of freedom:

- The four values $\mu_F(v)$ of v at the centroids M of F;
- The six mean values $\nu_e(v)$ along e, where $\nu_e(v) = 0.4v(N) + 0.3[v(A) + v(B)]$.

$\forall v \in V_h$ and $\forall F$ and e, we require that both $\mu_F(v)$ and $\nu_e(v)$ coincide for all tetrahedra of the mesh sharing the face F or the edge e; moreover we require that both $\mu_F(v)$ and $\nu_e(v)$ vanish whenever F or e is contained in Γ_h. Clearly enough these requirements are not sufficient to ensure the continuity in Ω_h of a function in V_h, and hence this space is not a subspace of $H_0^1(\Omega_h)$.

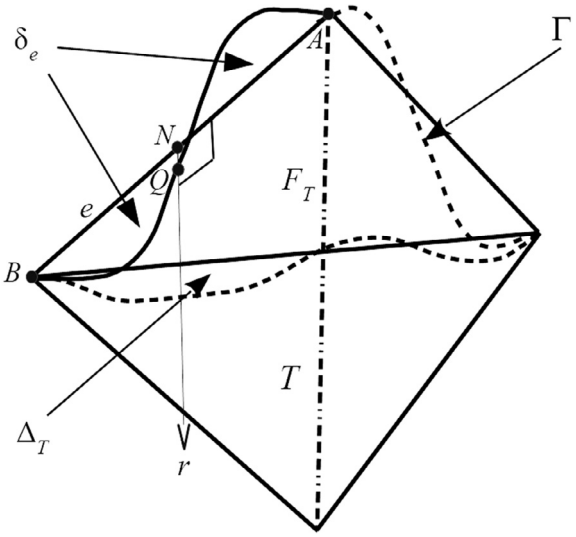

Fig. 3. $Q :=$ intersection with $\Gamma \cap \delta_e$ of the plane orthogonal to e containing its mid-point N.

The set of local canonical quadratic basis functions in a tetrahedron $T \in \mathcal{T}_h$ associated with the above degrees of freedom can be found in [12]. It is noteworthy that the gradients of all of them are an $O(h_T^{-1})$. This is a key property for the proof of Lemma 3.1 hereafter.

Similarly to the case of the standard Lagrangian piecewise quadratic elements, we define the trial-function space W_h in the same way as V_h, except for the fact that the degrees of freedom associated with faces F and edges e contained in Γ_h are modified as follows: For a given function $w \in W_h$, $\mu_F(w)$ is replaced by $\mu'_F(w)$ defined to be the value of w at the point P lying in the nearest intersection with Γ of the perpendicular to F passing through the centroid M of F as depicted in Fig. 2; referring to Fig. 3, $\nu_e(w)$ is replaced by $\nu'_e(w) := 0.4w(Q) + 0.3[w(A) + w(B)]$, where Q is the nearest intersection with Γ of the perpendicular to e in δ_e passing through N. $\forall w \in W_h$ we require that both $\mu'_F(w)$ and $\nu'_e(w)$ vanish for every face F or edge e contained in Γ_h.

It is convenient to extend to $\bar{\Omega} \setminus \bar{\Omega}_h$ any function $w \in W_h$ in such a way that its polynomial expression in $T \in \mathcal{O}_h$ also applies to points in $T_\Delta \setminus T$. In doing so the distinct expressions of w in δ_e are those in the tetrahedra belonging to $\mathcal{O}_h$ sharing the edge e.

Notice that the sets Δ_T, T_Δ enable the extension of $w \in W_h$ to $\bar{\Omega} \setminus \bar{\Omega}_h$, but play no role, neither in the definition, nor in the practical implementation of our method, as seen below.

Next we prove,

Lemma 3.1. *Provided h is small enough, $\forall T \in \mathcal{S}_h$ (resp. $\forall T \in \mathcal{R}_h$), given a set of m real values b_i, $1 \le i \le m$, with $m = 6$ (resp. $m = 9$), there exists a unique function $w_T \in \mathcal{P}_2(T)$ such that $\mu'_F(w_T) = 0$ and $\nu'_e(w_T) = 0$ if F and e are a face or an edge of T contained in Γ_h, and such that $\mu_F(w_T)$ and $\nu_e(w_T)$ take the assigned value b_i, if neither F nor e is a face or an edge of T contained in Γ_h.*

Proof. We first consider the case of an element in $T \in \mathcal{R}_h$. Referring to Fig. 3, let us extend the vector $\vec{b} := [b_1, b_2, \ldots, b_9]$ of $\Re^9$ into the vector of $\Re^{10}$ still denoted by $\vec{b}$, by adding the component $b_{10} = 0$ standing for the mean value ν'_e, $e \subset \Gamma_h$. If Q were replaced by the mid-point N of e, it is clear that the result would hold true, according to the properties of the interpolation under consideration (cf. [12]). The vector $\vec{a}$ of coefficients a_i for $i = 1, 2, \ldots, 10$ of the underlying canonical basis functions $\varphi_j \in \mathcal{P}_2(T)$ for $1 \le i \le 10$ would be precisely b_i for $1 \le i \le 10$. Denoting the associated degrees of freedom (for a straight-edged tetrahedron) by π_i, where π_i is some μ_F for $1 \le i \le 4$ and π_i is some ν_e for $5 \le i \le 10$, we assume that the corresponding canonical basis functions φ_i are numbered accordingly. This means that the matrix K whose entries are $k_{ij} := \pi_i(\varphi_j)$ is the identity matrix. Let $\pi'_i = \pi_i$ if $1 \le i \le 9$ and $\pi'_{10}(w)$ be given by $0.4w(Q) + 0.3[w(A) + w(B)]$ for any $w \in \mathcal{P}_2(T)$, where A and B are the end-points of $e \subset \Gamma_h$. We must establish that the 10×10 linear system of algebraic equations $K'\vec{a} = \vec{b}$ is uniquely solvable, where K' is the matrix with entries $k'_{ij} := \pi'_i(\varphi_j)$. Clearly we have $K' = K + E_K$, where the entries of E_K are $e_{ij} := \pi'_i(\varphi_j) - \pi_i(\varphi_j)$. At this point we recall the constant C_Γ depending only on Γ specified at the end of Section 2.2, such that the length of the segment $\overline{NQ}$ is bounded above by $C_\Gamma h_T^2$. From Rolle's Theorem it follows that $\forall\, i, j$, $|e_{ij}| \le 0.4 C_\Gamma h_T^2 \|\mathbf{grad}\ \varphi_j\|_{0,\infty,T_\Delta}$.

From Lemma 3.2 of [11] we know that $\|\mathbf{grad}\ \varphi_j\|_{0,\infty,T_\Delta} \le \mathcal{C}_\infty \|\mathbf{grad}\ \varphi_j\|_{0,\infty,T}$ for a suitable mesh-independent constant $\mathcal{C}_\infty$. Moreover from standard arguments we know that the latter norm in turn is bounded above by another mesh-independent constant times h_T^{-1}. In short we have $|e_{ij}| \le C_E h_T$ $\forall\, i, j$, where C_E is independent of T. Hence the matrix

K' equals the identity matrix plus an $O(h_T)$ matrix E_K. Therefore K' is an invertible matrix, as long as h is sufficiently small.

The case of an element $T \in \mathcal{S}_h$ can be dealt with as a mere variant of the above argument, and in this respect we also refer to Lemma 3.3 of [11]. ■

Lemma 3.1 allows us to assert that W_h is indeed a nonempty function space, whose dimension equals the one of V_h.

Before pursuing we introduce the broken gradient operator $\mathbf{grad}_h$ for any function w defined in Ω_h which is continuously differentiable in every $T \in \mathcal{T}_h$, given by $[\mathbf{grad}_h w]_{|T} \equiv \mathbf{grad}\ w_{|T}\ \forall T \in \mathcal{T}_h$.

Now if u is a function in $H^2(\Omega) \cap H_0^1(\Omega)$, we can define $I_h(u) \in W_h$ to be the function given by $\mu_F(I_h(u)) = \mu_F(u)$ and $\nu_e(I_h(u)) = \nu_e(u)$ for all the faces F and edges e of tetrahedra in $\mathcal{T}_h$ not contained in Γ_h. From standard interpolation results it is not difficult to establish that I_h enjoys the following property:

There exists a mesh-independent constant C_P such that $\forall u \in H^3(\Omega) \cap H_0^1(\Omega)$ it holds,

$$\|\mathbf{grad}_h(u - I_h(u))\|_{0,h} \leq C_P h^2 |u|_3. \tag{2}$$

Extending f by zero in $\Omega_h \setminus \Omega$ and still denoting the resulting function by f, the following problem is considered to approximate (1):

$$\begin{cases} \text{Find } u_h \in W_h \text{ such that } a_h(u_h, v) = L_h(v)\ \forall v \in V_h, \\ \text{where} \\ a_h(w, v) := \int_{\Omega_h} \mathbf{grad}_h w \cdot \mathbf{grad}_h v, \ \text{ for } w \in W_h + H^1(\Omega_h),\ v \in V_h \\ \text{and} \\ L_h(v) := \int_{\Omega_h} f v\ \forall v \in V_h. \end{cases} \tag{3}$$

The matrix associated with (3) is a sparse band matrix whose sparsity structure is the same as for the standard Galerkin FEM, in which the spaces of trial functions and test functions coincide. However here such a matrix is non symmetric, since the basis functions of W_h and V_h are the same only for nodes not belonging to elements in $\mathcal{O}_h$. Hence the stability and well-posedness of problem (3) are not trivial issues, which we next address.

Proposition 3.2. *If h is sufficiently small there exists a constant $\alpha > 0$ independent of h such that,*

$$\forall w \in W_h \neq 0, \ \sup_{v \in V_h \setminus \{0\}} \frac{a_h(w, v)}{\|\mathbf{grad}_h w\|_{0,h} \|\mathbf{grad}_h v\|_{0,h}} \geq \alpha. \tag{4}$$

Proof. Given $w \in W_h$, let v be the unique function in V_h such that all its degrees of freedom attached to a face or an edge of the mesh not contained in Γ_h coincide with those of w. Notice that by construction $\mu_F(v) = 0$ and $\nu_e(v) = 0$ as long as F or e is contained in Γ_h.

For a given $T \in \mathcal{O}_h$ we denote by m_T the number of degrees of freedom $\{\pi_i^T\}_{i=1}^{m_T}$ of V_h attached to a face F or an edge e contained in Γ_h. Clearly enough we have

$$a_h(w, v) = \sum_{T \in \mathcal{T}_h} \int_T |\mathbf{grad}\ w|^2 - \sum_{T \in \mathcal{O}_h} \int_T \mathbf{grad}\ w \cdot \mathbf{grad}\ r_T(w), \tag{5}$$

where $r_T(w) = \sum_{i=1}^{m_T} \pi_i^T(w) \varphi_i^T$, φ_i^T being the canonical basis function of the space $\mathcal{P}_2(T)$ associated with the degree of freedom π_i^T.

Now from standard results it holds $\|\mathbf{grad}\ \varphi_i^T\|_{0,T} \leq C_\varphi h_T^{1/2}$ where C_φ is a mesh independent constant. Referring to Figs. 2 and 3, since $w(P) = \mu_F'(w) = 0$ (resp. $0.4w(Q) + 0.3[w(A) + w(B)] = \nu'(w) = 0$), where F (resp. e) generically represent a face (resp. an edge) of T contained in Γ_h, in accordance with the definition of W_h, a simple Taylor expansion about P (resp. Q) allows us to conclude that $|w(M)|$ (resp. $|w(N)|$) are bounded above by $l\|\mathbf{grad}\ w\|_{0,\infty,T_\Delta}$, where $l = length(\overline{PM})$ (resp. $length(\overline{QN})$), or yet that $|w(M)|$ (resp. $|w(N)|$) is bounded above by $C_\Gamma h_T^2 \|\mathbf{grad}\ w\|_{0,\infty,T_\Delta}$. On the other hand from Lemma 2.2 of [11] it holds $\|\mathbf{grad}\ w\|_{0,\infty,T_\Delta} \leq C_J h_T^{-3/2} \|\mathbf{grad}\ w\|_{0,T}$ for a mesh-independent constant C_J. Plugging all those estimates into (5), since $m_T \leq 4$, we obtain:

$$a_h(w, v) \geq \int_{\Omega_h} |\mathbf{grad}_h w|^2 - 4C_\varphi C_J C_\Gamma h \sum_{T \in \mathcal{O}_h} \|\mathbf{grad}\ w\|_{0,T}^2. \tag{6}$$

Then it holds with

$$c := 4C_\varphi C_J C_\Gamma, \tag{7}$$

$$a_h(w, v) \geq (1 - ch)\|\mathbf{grad}_h w\|_{0,h}^2. \tag{8}$$

Now using arguments in all similar to those employed above, we easily conclude that

$$\|\mathbf{grad}_h v\|_{0,h} \leq \|\mathbf{grad}_h w\|_{0,h} + \|\mathbf{grad}_h v - \mathbf{grad}_h w\|_{0,h} \leq (1+ch)\|\mathbf{grad}_h w\|_{0,h}. \tag{9}$$

Combining (8) and (9), provided $h \leq (2c)^{-1}$ we establish (4) with $\alpha = 1/3$. ■

Proposition 3.3. *Provided h is sufficiently small, problem (3) has a unique solution.*

Proof. From well-known results (cf. [2,3] and [17]) this is an immediate consequence of Proposition 3.2 and of the fact that V_h and W_h have the same dimension. ■

4. Symmetrization of the solution procedure

Since (3) is not a symmetric problem we can use the following iterative procedure to solve it as a sequence of symmetric problems.

First of all let n_h be the dimension of both V_h and W_h, that is the total number of degrees of freedom of both spaces not assigned to zero beforehand. Let also $\|\cdot\|_{0,\infty,h}$ be the norm of either V_h or W_h defined to be the maximum absolute value of their n_h degrees of freedom. For every $v \in V_h$ we denote by $\Pi_W(v)$ the function of W_h whose degrees of freedom coincide with those of v. Similarly for every $w \in W_h$ we denote by $\Pi_V(w)$ the function of V_h whose degrees of freedom coincide with those of w.

Now we consider the following symmetric problem,

$$\begin{cases} \text{Find } \bar{u}_h^0 \in V_h \text{ such that} \\ a_h(\bar{u}_h^0, v) = L_h(v)\ \forall v \in V_h, \end{cases} \tag{10}$$

which is clearly uniquely solvable.

Defining

$$u_h^0 := \Pi_W(\bar{u}_h^0) \in W_h, \tag{11}$$

solve successively for $n = 1, 2, \ldots$ the problems,

$$\begin{cases} \text{Find } u_h^n \in W_h := \Pi_W(\bar{u}_h^n) \\ \text{where } \bar{u}_h^n \in V_h \text{ is the unique solution of} \\ a_h(\bar{u}_h^n, v) = a_h(\bar{u}_h^{n-1}, v) - a_h(u_h^{n-1}, v) + L_h(v)\ \forall v \in V_h, \end{cases} \tag{12}$$

until $\|u_h^n - u_h^{n-1}\|_{0,\infty,h}$ is less than a small tolerance ε.

Since the matrix associated with (12) is a symmetric positive definite matrix for the standard Galerkin FEM, the stability and well-posedness of (12) is guaranteed. It is also a band matrix with the same sparsity structure within its band as the matrix associated with (3).

Let us study the convergence of the above iterative procedure. With this aim we first set $\bar{u}_h = \Pi_V(u_h)$ and note that,

$$a_h(\bar{u}_h, v) = a_h(\bar{u}_h, v) - a_h(u_h, v) + L_h(v)\ \forall v \in V_h. \tag{13}$$

$\forall n \geq 0$, let $w_h^n := u_h^n - u_h \in W_h$ and $\bar{w}^n := \bar{u}_h^n - \bar{u}_h \in V_h$. Combining (12) with (13) we have:

$$a_h(\bar{w}_h^n, v) = a_h(\bar{w}_h^{n-1}, v) - a_h(w_h^{n-1}, v)\ \forall v \in V_h. \tag{14}$$

We next establish that, provided h is sufficiently small, $\|\mathbf{grad}_h w_h^n\|_{0,h}$ tends to zero roughly as fast as $b(h)(2ch)^n$ as n goes to infinity, where c fulfills $2ch \leq 1$ and $b(h)$ is an $O(h)$.

Since $\bar{w}_h^{n-1}$ only differs from w_h^{n-1} in elements in $\mathcal{O}_h$ we have,

$$a_h(\bar{w}_h^{n-1}, v) - a_h(w_h^{n-1}, v) = \sum_{T \in \mathcal{O}_h} \int_T \mathbf{grad}(\bar{w}_h^{n-1} - w_h^{n-1}) \cdot \mathbf{grad}\ v. \tag{15}$$

Using the same arguments leading to (6) together with (7), we obtain successively,

$$a_h(\bar{w}_h^{n-1}, v) - a_h(w_h^{n-1}, v) \leq ch \sum_{T \in \mathcal{O}_h} \|\mathbf{grad}\ w_h^{n-1}\|_{0,T} \|\mathbf{grad}\ v\|_{0,T}, \tag{16}$$

$$a_h(\bar{w}_h^{n-1}, v) - a_h(w_h^{n-1}, v) \leq ch \|\mathbf{grad}_h w_h^{n-1}\|_{0,h} \|\mathbf{grad}_h v\|_{0,h}. \tag{17}$$

Taking $v = \bar{w}_h^n$ in both (14) and (17) we come up with,

$$\|\mathbf{grad}_h \bar{w}_h^n\|_{0,h} \leq ch \|\mathbf{grad}_h w_h^{n-1}\|_{0,h} \tag{18}$$

Table 1
Number of iterations m such that $\|u_h^m - u_h^{m-1}\|0,\infty,h < 10^{-5}$ using Cholesky's method.

p	$\longrightarrow$	2	4	8	12	16
m	$\longrightarrow$	7	5	4	4	3
$\|u_h^m - u_h^{m-1}\|_{0,\infty,h}$	$\longrightarrow$	0.62431E−5	0.51154E−5	0.56915E−5	0.12739E−5	0.92735E−5

Table 2
Number of iterations m such that $\|u_h^m - u_h^{m-1}\|_{0,\infty,h} < 10^{-7}$ using the CG method.

p	$\longrightarrow$	4	8	12	16	24
m	$\longrightarrow$	9	13	19	37	95
$\|u_h^m - u_h^{m-1}\|_{0,\infty,h}$	$\longrightarrow$	0.14937E−7	0.42120E−7	0.24374E−7	0.93131E−7	0.65410E−7

Now noting that $\forall m \geq 0$ $\|\mathbf{grad}_h w_h^m\|_{0,h} \leq \|\mathbf{grad}_h \bar{w}_h^m\|_{0,h} + \|\mathbf{grad}_h(w_h^m - \bar{w}_h^m)\|_{0,h}$, similarly to (17), as long as h is less than $1/c$, we easily conclude that

$$\|\mathbf{grad}_h w_h^m\|_{0,h} \leq (1-ch)^{-1}\|\mathbf{grad}_h \bar{w}_h^m\|_{0,h} \ \forall m \geq 0. \tag{19}$$

Plugging (19) with $m = n-1$ into (18) we establish that,

$$\|\mathbf{grad}_h \bar{w}_h^n\|_{0,h} \leq \rho(h)\|\mathbf{grad}_h \bar{w}_h^{n-1}\|_{0,h} \ \forall n > 1 \text{ with } \rho(h) = ch/(1-ch). \tag{20}$$

Assuming that $h < 1/(2c)$ the fraction $\rho(h)$ will be less than one, and hence the quantity $\|\mathbf{grad}_h \bar{w}_h^n\|_{0,h}$ will decrease by a factor of $\rho(h)$ at every iteration. Actually using again (19), this time with $m = n$, and noting that $1 - ch \geq 1/2$ by assumption, we have,

$$\|\mathbf{grad}_h w_h^n\|_{0,h} \leq \sigma(h)(2ch)^n \text{ with } \sigma(h) := \|\mathbf{grad}_h(\bar{u}_h^0 - \bar{u}_h)\|_{0,h}/2 \ \forall n > 1. \tag{21}$$

Observing that $\bar{u}_h^0 - \bar{u}_h = (\bar{u}_h^0 - u) + (u - u_h) + (u_h - \bar{u}_h)$ and that the orders of magnitude of the norms $\|\mathbf{grad}_h \cdot\|_{0,h}$ of the terms in parentheses on the right hand side are respectively $O(h^{3/2})$, $O(h^2)$ and $O(h)$, we can assert that $\sigma(h)$ is bounded above by a coefficient $b(h)$, whose order of magnitude is at most an $O(h)$. All this advocates in favor of a faster convergence of the iterations (12), the smaller h.

Let us check the efficiency of the iterative symmetrization procedure (10)–(11)–(12) by solving a test-problem with successively refined meshes. The solution of the linear system resulting from (12) is computed by means of both Cholesky's method with BMS (band matrix storage) and the CG (conjugate gradient) method by storing only the non zero coefficients of the matrix, i.e., with VSMS (very sparse matrix storage). In the model problem Ω is the ellipsoid of equation $x^2/a^2 + y^2/b^2 + z^2 < 1$ in a cartesian coordinate system (x, y, z), whose origin is its center, with $a = 0.6$ and $b = 0.8$. We take an exact solution given by $u(x, y, z) = (1-x^2/a^2-y^2/b^2-z^2)(1-x^2/b^2-y^2/a^2-z^2)$, so that $f = -\Delta u$. The computations are carried out only for the octant corresponding to non negative values of the coordinates, with a family of quasi-uniform meshes consisting of $6p^3$ tetrahedra for an integer $p \geq 1$. For each value of p the mesh of the ellipsoid is the transformation of the uniform mesh of a unit cube with $6p^3$ tetrahedra having edges parallel to the line $x = y = z$, by suitably mapping the set of vertexes of the latter given in cartesian coordinates into the one of the actual mesh expressed in spherical coordinates. In this manner we have $h \simeq p^{-1}$.

In Tables 1 and 2 we show the number of iterations m necessary to satisfy tolerances of $\varepsilon = 10^{-5}$ and $\varepsilon = 10^{-7}$, for increasing values of p, using the Cholesky and the CG solver, respectively. The smaller value of ε in the latter case is due to the observation that this tolerance must be compatible with the necessarily small one in the convergence test of the CG method.

According to Table 1 the number of iterations necessary for convergence of the symmetrization procedure decreases indeed with the mesh size. Table 2 in turn points in the opposite direction, but this effect can be credited to the combination of two iterative procedures. Nevertheless such a behavior is far from being a drawback, as seen in the next section.

5. Comparative study

In this section we further investigate the iterative solution procedure of symmetric problems posed in the nonparametric Petrov–Galerkin variational form of the type (12). More particularly we carry out a comparative study thereof with the direct solution of the underlying non symmetric linear system. In this framework two approaches are assessed: The direct solution of the non symmetric system performed by either Crout's method with partial pivoting and BMS or the GMRES method with VSMS; the iterative solution with a symmetric positive definite matrix performed by either Cholesky's method with BMS or the CG method with VSMS, resp. Additionally we extend such numerical comparisons to the classical conforming quadratic finite element in both nonparametric Petrov–Galerkin and isoparametric form.

A Lenovo T440s laptop was employed in all the computations reported below.

Table 3
CPU time for solving (3): nr. of iterations (12) using Cholesky's method vs. Crout's method.

p	$\longrightarrow$	4	6	8	12	16
Iterative solution (Cholesky's method)	$\longrightarrow$	4.36 s	17.35 s	420.78 s	10,689.93 s	92,061.36 s
Direct solution (Crout's method)	$\longrightarrow$	19.00 s	136.92 s	1846.89 s	26,837.75 s	230,274.28 s

Table 4
CPU time for solving (3): nr. of iterations (12) using the CG method vs. GMRES method.

p	$\longrightarrow$	6	8	12	16	24
Iterative solution (CG method)	$\longrightarrow$	3.88 s	18.35 s	184.59 s	1076.87 s	13,794.24 s
Direct solution (GMRES method)	$\longrightarrow$	8.11 s	27.71 s	288.48 s	2453.89 s	44,973.59 s

Table 5
Errors, nr. of iterations using Cholesky's method and DOF count for the nonconforming FEM.

p	$\longrightarrow$	2	4	8	16
$\|u-u_h\|_{0,h}$	$\longrightarrow$	0.64013E−2	0.88793E−3	0.11467E−3	0.14585E−4
$\|\mathbf{grad}_h(u-u_h)\|_{0,h}$	$\longrightarrow$	0.11549E+0	0.34444E−1	0.91569E−2	0.23477E−2
$\|u-u_h\|_{0,\infty,h}$	$\longrightarrow$	0.27816E−1	0.39244E−2	0.62257E−3	0.85775E−4
m	$\longrightarrow$	7	5	4	3
M	$\longrightarrow$	218	1,468	10,712	81,712

Table 6
Errors, nr. of iterations using Cholesky's method and DOF count for the conforming FEM.

p	$\longrightarrow$	2	4	8	16
$\|u-\check{u}_h\|_{0,h}$	$\longrightarrow$	0.70568E−2	0.95648E−3	0.12203E−3	0.15445E−4
$\|\mathbf{grad}_h(u-\check{u}_h)\|_{0,h}$	$\longrightarrow$	0.11772E+0	0.353106E−1	0.94375E−2	0.24253E−2
$\|u-\check{u}_h\|_{0,\infty,h}$	$\longrightarrow$	0.36064E−1	0.69394E−2	0.10616E−2	0.14339E−3
$\check{m}$	$\longrightarrow$	6	5	4	3
$\check{M}$	$\longrightarrow$	125	729	4,913	35,937

5.1. Iterative vs. direct solution of (3)

First of all we compare the performance of the nonconforming quadratic method studied in Sections 3 and 4 to approximate (1) using both solution strategies. With this aim we take the same test-problem as in the previous section. Depending on whether the iterative symmetrization procedure is employed or not, in this comparison we use both direct solvers with BMS, namely, Cholesky's method and Crout's method, and the iterative solvers GC and GMRES with VSMS.

We supply in Table 3 the total processing (CPU) time in seconds for successively refined meshes, using direct solvers for both the iterative symmetrization procedure with a tolerance equal to 10^{-5} and the direct solution.

Similarly, we display in Table 4 the total CPU time in seconds for successively refined meshes for both the scheme (12) and the direct solution, using now the iterative solvers and a tolerance equal to 10^{-7} for all iterative procedures.

From the above results we infer the great superiority of the iterative symmetrization strategy over the direct solution, since the CPU time to run the former is much smaller than the CPU time required by the latter. This effect is even more noteworthy in case direct solvers are used, in which unreasonable processing times for barely intermediate meshes are reported (cf. Table 3). On the other hand, the fact that an increasing number of iterations is necessary for convergence of the procedure (12) as the mesh is refined (cf. Table 2), is probably the cause of a lesser discrepancy of CPU times in case iterative solvers are used, as one can see in Table 4. Notice however that in spite of these observations, the drastic reduction of matrix storage for iterative solution methods advocates in favor of them, as compared to direct ones.

5.2. Additional comparisons involving second order finite-element methods

Keeping the same test-problem as above, we pursue the performance evaluation of our iterative scheme as compared to the direct solution, taking also classical conforming quadratic finite elements. In order to have better insight on the merits of the nonparametric Petrov–Galerkin formulation, such comparisons are extended to isoparametric finite elements of the same order in standard Galerkin formulation.

Referring to [11], let $\check{u}_h$ represent the approximate solution of (1) related to mesh $\mathcal{T}_h$ obtained by the nonparametric Petrov–Galerkin approach, in connection with conforming Lagrange quadratic finite elements.

To begin with we illustrate the strength of the nonconforming approach, by displaying in Tables 5 through 8 data related to u_h and $\check{u}_h$, respectively, for different values of p, that is h. Besides the errors measured in three different manners, we supply a degree of freedom (DOF) count for both FEMs. The number of iterations necessary to satisfy the stop criterion for the iterative scheme of the type (12) is still denoted by m for the nonconforming method and by $\check{m}$ for the conforming method. The corresponding total number of DOFs are denoted by M and $\check{M}$ respectively. Similarly to the

Table 7
Errors, number of iterations using the CG method and DOF count for the nonconforming FEM.

p	$\longrightarrow$	3	6	12	24
$\|u-u_h\|_{0,h}$	$\longrightarrow$	0.20505E−2	0.26893E−3	0.34373E−4	0.56906E−5
$\|\mathbf{grad}_h(u-u_h)\|_{0,h}$	$\longrightarrow$	0.58245E−1	0.15976E−1	0.41403E−2	0.10525E−2
$\|u-u_h\|_{0,\infty,h}$	$\longrightarrow$	0.82123E−2	0.13700E−2	0.19521E−3	0.41366E−4
m	$\longrightarrow$	10	8	19	95
M	$\longrightarrow$	657	4,662	35,028	271,368

Table 8
Errors, nr. of iterations using the CG method and DOF count for the conforming FEM.

p	$\longrightarrow$	3	6	12	24
$\|u-\check{u}_h\|_{0,h}$	$\longrightarrow$	0.22262E−2	0.28733E−3	0.36439E−4	0.48589E−5
$\|\mathbf{grad}_h(u-\check{u}_h)\|_{0,h}$	$\longrightarrow$	0.59516E−1	0.16437E−1	0.42741E−2	0.10872E−2
$\|u-\check{u}_h\|_{0,\infty,h}$	$\longrightarrow$	0.13889E−1	0.23689E−2	0.33214E−3	0.43462E−4
$\check{m}$	$\longrightarrow$	7	6	6	5
$\check{M}$	$\longrightarrow$	343	2,197	15,625	117,649

Table 9
CPU time for solving (1) with the conforming FEM via direct solvers.

p	$\longrightarrow$	4	6	8	12	16
Iterative solution (Cholesky's method)	$\longrightarrow$	0.26 s	1.64 s	21.18 s	235.38 s	3013.57 s
Direct solution (Crout's method)	$\longrightarrow$	0.51 s	6.54 s	212.45 s	2617.61 s	40226.10 s

Table 10
CPU time for solving (1) with the conforming FEM via iterative solvers.

p	$\longrightarrow$	6	8	12	16	24
Iterative solution (CG method)	$\longrightarrow$	1.09 s	4.70 s	38.77 s	399.29 s	2189.46 s
Direct solution (GMRES method)	$\longrightarrow$	2.76 s	8.16 s	78.18 s	426.18 s	4234.10 s

previous subsection in the stop criterion the tolerance ε applies to the maximum absolute value of the difference between DOFs in two successive iterations.

The results given in Tables 5 and 6 were obtained with a Cholesky solver for $\varepsilon = 10^{-5}$.

As one infers from Tables 5 and 6 the methods under experimentation are both of the third order in $L^2(\Omega_h)$ and of the second order in the broken (semi)norm of $H^1(\Omega_h)$ as expected or predicted either in [11] or in Section 6 hereafter. Both methods are also fairly equivalent from the point of view of accuracy in these norms. On the other hand there is a clear advantage of the nonconforming method over the conforming method in terms of DOF ("pointwise") errors.

The results in Tables 7 and 8 were obtained by using a CG solver with VSMS. Here we took $\varepsilon = 10^{-7}$, which is also the tolerance employed in the stop criterion for the CG method.

Tables 7 and 8 confirm roughly the same orders of both FEM observed in Tables 5 and 6, and the slightly better accuracy of the nonconforming method except for the L^2-norm of the error for the finest mesh. Notice that the number of iterations necessary for convergence of the conforming method decreases smoothly as the mesh is refined, as expected, in contrast to the nonconforming method. This could explain the more significant deterioration of the accuracy in the L^2-norm observed for the latter method, as compared to the former. Whatever the case, such an effect advocates in favor of direct solvers, since in this case there is no need to adjust a tolerance to optimally fit the one of the iterative symmetrization scheme itself. However it turns out that iterative solvers are in principle less time consuming for a given mesh, while requiring much less storage.

It is also interesting to watch the behavior of both methods in terms of CPU time, when the direct and the iterative solving approaches are employed. Tables 9 and 10 supply the CPU times for the conforming quadratic method with successively refined meshes, similarly to Tables 3 and 4 respectively, for the nonconforming method.

It is noticeable here again the great superiority of the iterative approach from the point of view of processing time. A quick comparison of Tables 3 and 4 with Tables 9 and 10 also indicates that the nonconforming method is much more time consuming than the conforming method. However this is no surprise since there are more than twice as many degrees of freedom for the latter with respect to the former for the same mesh.

Next we compare the nonparametric Petrov–Galerkin formulation for the conforming quadratic element with the corresponding isoparametric formulation, whose optimal second order in the H^1-norm was established in [18]. We denote by $\tilde{u}_h$ the approximate solution to (1) determined by the isoparametric technique for the same mesh as $\check{u}_h$. Naturally enough $\tilde{u}_h$ is computed using the Cholesky's method with BMS and the CG method with VSMS. However for a more fair comparison with the nonparametric approach in terms CPU time, we also compute the isoparametric solution using Crout's method and the GMRES method, without taking into account symmetry. This is because in the case of non symmetric problems the use of both Cholesky's method and the CG method has to be discarded.

CPU times necessary to determine $\tilde{u}_h$ with direct solvers are displayed in Table 11.

Table 11
CPU time for solving (1) with the isoparametric quadratic FEM via direct solvers.

p	$\longrightarrow$	4	6	8	12	16
Cholesky's method	$\longrightarrow$	0.20 s	1.69 s	14.87 s	289.66 s	3552.16 s
Crout's method (with ∂ pivoting)	$\longrightarrow$	0.58 s	6.55 s	202.04 s	2248.96 s	39128.77 s

Table 12
CPU time for solving (1) with the isoparametric quadratic FEM via iterative solvers.

p	$\longrightarrow$	6	8	12	16	24
CG method	$\longrightarrow$	0.99 s	4.09 s	42.96 s	230.94 s	2560.18 s
GMRES method	$\longrightarrow$	1.87 s	8.96 s	82.78 s	500.15 s	4829.93 s

Table 13
Errors for the conforming quadratic FEM in isoparametric formulation.

p	$\longrightarrow$	2	4	8	16
$\|u-\tilde{u}_h\|_{0,h}$	$\longrightarrow$	0.75220E−2	0.10564E−2	0.13173E−3	0.16185E−4
$\|\mathbf{grad}_h(u-\tilde{u}_h)\|_{0,h}$	$\longrightarrow$	0.13931E+0	0.39089E−1	0.10015E−1	0.25061E−2
$\|u-\tilde{u}_h\|_{0,\infty,h}$	$\longrightarrow$	0.40980E−1	0.79148E−2	0.12384E−2	0.16897E−3

Table 14
Key storage data for the symmetric band matrices handled by Cholesky's method.

p	$\longrightarrow$	2	4	8	16
NC FE: HBW × NU (=TNE)	$\longrightarrow$	86 × 152	322 × 1, 216	1, 250 × 9, 728	4, 930 × 77,824
C FE: HBW × NU (=TNE)	$\longrightarrow$	43 × 64	147 × 512	547 × 4, 096	2, 115 × 32,768

Table 15
Key storage data for the symmetric sparse matrices handled by the CG method.

p	$\longrightarrow$	3	6	12	24
NC FE: TNE/NU ($\simeq \lambda_{NC}$)	$\longrightarrow$	4, 563/513	41,526/4, 104	353,268/32,832	2,912,328/262,656
C FE: TNE/NU ($\simeq \lambda_C$)	$\longrightarrow$	2, 125/216	21,052/1, 728	186,400/13,824	1, 566,856/110,592

An iterative-solver counterpart in terms of CPU time is supplied in Table 12.

It is no surprise that Tables 11 and 12 confirm the great superiority in terms of CPU time, of methods whose use is restricted to symmetric positive definite matrices, over methods applying to any regular matrix. In particular Cholesky's method is much better than Crout's method as shown in Table 11. Moreover, resorting to Tables 9 and 11, it turns out that both approaches are fairly equivalent in terms of CPU, with a slight advantage of isoparametric elements over nonparametric elements. This contradicts observations in the opposite sense in the two-dimensional case (cf. [4]). On the other hand, if one compares the solutions using Cholesky's method, isoparametric elements perform a little better only for the coarser meshes, while the contrary occurs in an increasingly significant manner as the mesh is refined. Such a behavior is noteworthy taking into account that iterations are necessary for the nonparametric approach, in contrast to the isoparametric approach. This seems to advocate in favor of the former, and could be due to its better matrix conditioning.

We push further our numerical study by comparing the solutions determined by the nonparametric and the isoparametric approaches in terms of accuracy. In Table 13 the errors for the isoparametric solution computed by Cholesky's method are given in three different measures.

Comparing the results displayed in Tables 6 and 13, we figure out that the nonparametric approach is a little more accurate than the isoparametric approach in all respects. Taking into account the previous observations, together with the two-dimensional experiments reported in [4] we are inclined to conclude that the former is superior to the latter.

To conclude we comment on the cost of storage in the experiments reported in this section.

First we note that the DOFs were numbered in a standard sequential manner for uniform meshes of a cube. In doing so the number of unknowns (NU) for the nonconforming method and the conforming method are $19p^3$ and $8p^3$, respectively. This also leads to band matrices for both methods, whose half band width (HBW) for large values of p is asymptotically equal to $19p^2$ for the nonconforming method and to $8p^2$ for the conforming method. It follows that the direct solvers handle arrays whose total number of entries (TNE) are roughly 19^2p^5 and 8^2p^5, respectively. This explains the growing discrepancy in CPU time to run direct solvers for both methods with the same mesh, as p increases (cf. Tables 7 and 8). On the other hand, in case iterative solvers are used, arrays with TNE asymptotically equal to $19\lambda_{NC}p^3$ and $8\lambda_C p^3$ are handled for the nonconforming method and the conforming method, where $\lambda_{NC} \simeq 11$ and $\lambda_C \simeq 14$ and the subscripts NC and C stand for nonconforming and conforming. This is the reason why the ratios between CPU times to run both methods with the same mesh using iterative solvers are smaller, as shown in Tables 9 and 10. Just to give an overview of the matrix storage required to run both finite element methods, we supply in self-explanatory Tables 14 and 15 the above key figures as p varies, for direct and iterative solvers, respectively.

6. Error estimates

In this section we establish error estimates for problem (3). Akin to [11] we distinguish the convex case from the non-convex case.

First we have:

Theorem 6.1. *Assume that $f \in H^1(\Omega)$ and $g \equiv 0$. As long as h is sufficiently small, if Ω is a convex domain smooth enough for the solution u of (1) to belong to $H^3(\Omega)$, there exists a constant $C(f)$ depending only on f such that the solution u_h of (3) satisfies:*

$$\|\mathbf{grad}_h(u - u_h)\|_{0,h} \leq C(f)h^2. \tag{22}$$

Proof. According to [17], using Proposition 3.2 we can write:

$$\|\mathbf{grad}_h(u - u_h)\|_{0,h} \leq \frac{1}{\alpha}\left[\|\mathbf{grad}_h(u - I_h(u))\|_{0,h} + \sup_{v \in V_h \setminus \{0\}} \frac{|a_h(u, v) - L_h(v)|}{\|\mathbf{grad}_h v\|_{0,h}}\right]. \tag{23}$$

Proof. Taking into account (2), all we have to do is to estimate the sup term on the right hand side of (23). As a matter of fact such an issue was basically addressed in [12]. More precisely the required estimate is a consequence of the fact that the L^2-projection of the trace on a face F of the mesh of any function $v \in V_h$ onto the space $\mathcal{P}_1(F)$, is a linear combination of the values $\mu_F(v)$ and $\nu_e(v)$, where e here generically represents the edges of F. This property implies the existence of a mesh-independent constant C_R such that,

$$|a_h(u, v) - L_h(v)| \leq C_R h^2 |u|_3 \|\mathbf{grad}_h v\|_{0,h}. \tag{24}$$

Then (22) directly follows from (23), (2) and (24). ■

Before pursuing we introduce Ω' as a smooth domain of $\Re^3$ close to Ω but strictly containing both Ω and Ω_h for all h small enough to conform to our assumptions on the meshes. According to Stein et al. [19] there exists an extension u' of u to Ω' such that $u' \in H^3(\Omega')$ and $u' \equiv u$ in Ω.

Now we prove

Theorem 6.2. *Assume that $u \in H^3(\Omega)$. Provided h is sufficiently small, there exists a mesh-independent constant $\tilde{C}$ such that the unique solution u_h to (3) satisfies:*

$$\|\mathbf{grad}_h(u - u_h)\|_{0,h}^{\sim} \leq \tilde{C} h^2 \|u'\|_{3,\Omega'}, \tag{25}$$

$u' \in H^3(\Omega')$ being the regular extension of u to Ω' constructed in accordance to Stein et al. [19].

Proof. First of all combining (3) with Proposition 3.2 we can write:

$$\|\mathbf{grad}_h(u_h - I_h(u'))\|_{0,h} \leq \frac{1}{\alpha} \sup_{v \in V_h \setminus \{0\}} \frac{|a_h(u', v) - L_h(v)| + |a_h(u' - I_h(u'), v)|}{\|\mathbf{grad}_h v\|_{0,h}}. \tag{26}$$

The first term in the numerator of (26) can be estimated in the following manner.

Following the same steps as in Theorem 5.9 of [11], we denote by $\mathcal{Q}_h$ the subset of $\mathcal{O}_h$ consisting of elements T such that $\tilde{T} \neq T$. Next we apply First Green's identity to $a_h(u', v)$. Noticing that v is not continuous across the inter-element boundaries, and recalling the notations $\Delta_T' = \Delta_T \setminus \Omega$ and ∂T for the boundary of $T \in \mathcal{T}_h$ and denoting by $\partial(\cdot)/\partial n_T$ the normal derivative on ∂T oriented outwards T we obtain:

$$\begin{cases} |a_h(u', v) - L_h(v)| = c_h(u', v) + d_h(u', v) \\ \text{where} \\ c_h(u', v) = \displaystyle\sum_{T \in \mathcal{T}_h} \int_{\partial T} v \frac{\partial u'}{\partial n_T} \\ \text{and} \\ d_h(u', v) = -\displaystyle\sum_{T \in \mathcal{Q}_h} \int_{\Delta_T'} \Delta u' v. \end{cases} \tag{27}$$

$c_h(u', v)$ can be estimated by means of standard arguments for nonconforming finite elements. More specifically in the case under study (cf. [12]) an estimate of the same nature as (24) applies to c_h, i.e.,

$$c_h(u', v) \leq C_R h^2 |u'|_{3,\Omega_h} \|\mathbf{grad}_h v\|_{0,h}. \tag{28}$$

As for bilinear form d_h first we observe that,

$$d_h(u', v) \leq \sum_{T \in \mathcal{Q}_h} [volume(\Delta'_T)]^{1/2} \|\Delta u'\|_{0,\Delta'_T} \|v\|_{0,\infty,\Delta'_T}. \tag{29}$$

Since $\mu_F(v) = 0$ for all faces F contained in Γ_h, there exists a mesh-independent constant C'_Γ such that

$$\|v\|_{0,\infty,\Delta'_T} \leq \|v\|_{0,\infty,T} \leq C'_\Gamma h_T \|\mathbf{grad}\ v\|_{0,\infty,T}. \tag{30}$$

Using the well-known inverse inequality (see e.g. [20]),

$$\|w\|_{0,\infty,T} \leq \mathcal{C}_I h_T^{-3/2} \|w\|_{0,T}\ \forall w \in \mathcal{P}_2(T), \tag{31}$$

where $\mathcal{C}_I$ is a mesh-independent constant, like in Theorem 5.8 of [11], the following result derives from (30),

$$\|v\|_{0,\infty,\Delta'_T} \leq C'_\Gamma \mathcal{C}_I h_T^{-1/2} \|\mathbf{grad}\ v\|_{0,T}. \tag{32}$$

Noticing that $volume(\Delta'_T)$ is bounded by h_T^4 multiplied by a constant C_Ω depending only on Ω, for both $T \in \mathcal{S}_h \cap \mathcal{Q}_h$ and $T \in \mathcal{R}_h \cap \mathcal{Q}_h$, from straightforward calculations it follows that,

$$\|\Delta u'\|_{0,\Delta'_T} \leq [C_\Omega]^{1/4} h_T \left[\int_{\Delta'_T} (\Delta u')^4\right]^{1/4} \quad \forall T \in \mathcal{Q}_h. \tag{33}$$

Then combining (29), (30), (32) and (33), applying the Cauchy–Schwarz inequality to the summation over T, and setting $C_S := [C_\Omega]^{3/4} C'_\Gamma \mathcal{C}_I$ we come up with,

$$d_h(u', v) \leq C_S h^2 \left\{\sum_{T \in \mathcal{Q}_h} h_T \left[\int_{\Delta'_T} (\Delta u')^4\right]^{1/2}\right\}^{1/2} \|\mathbf{grad}_h v\|_{0,h}. \tag{34}$$

Applying again the Cauchy–Schwarz inequality to the summation on the right hand side of (34) we readily obtain,

$$d_h(u', v) \leq C_S h^2 \left(\sum_{T \in \mathcal{Q}_h} h_T^2\right)^{1/4} \left[\sum_{T \in \mathcal{Q}_h} \int_{\Delta'_T} (\Delta u')^4\right]^{1/4} \|\mathbf{grad}_h v\|_{0,h}. \tag{35}$$

Noticing that there exists a constant $\hat{C}_\Gamma$ such that

$$\left[\sum_{T \in \mathcal{Q}_h} h_T^2\right]^{1/2} \leq \hat{C}_\Gamma \text{ independently of } h, \tag{36}$$

we come up with,

$$d_h(u', v) \leq C_S [\hat{C}_\Gamma]^{1/2} h^2 \|\Delta u'\|_{0,4,\Omega_h} \|\mathbf{grad}_h v\|_{0,h}. \tag{37}$$

Since $H^1(\Omega')$ is continuously embedded in $L^4(\Omega')$ (cf. [14]), from (37) we infer the existence of a mesh-independent constant C_R such that

$$d_h(u', v) \leq C_R h^2 \|\Delta u'\|_{1,\Omega'} \|\mathbf{grad}_h v\|_{0,h}, \tag{38}$$

Now we plug (28) and (38) into (27), and then the resulting inequality into (26). Finally using the trivial variant of (2) according to which

$$\|\mathbf{grad}_h(u' - I_h(u'))\|_{0,h} \leq C'_P h^2 |u'|_{3,\Omega'} \tag{39}$$

for a suitable mesh-independent constant C'_P together with the triangle inequality, the result follows. ■

7. Conclusions

The authors believe to have undoubtedly demonstrated that the nonparametric Petrov–Galerkin formulation studied in this work is a very efficient universal tool to solve boundary value problems posed in curved domains with Dirichlet boundary conditions. This assertion is supported by several evidences presented throughout the article.

The conclusions of the experimentation carried out in this work can be summarized as follows:

1. First of all we emphasize that, although the nonparametric formulation leads to non symmetric linear systems, even when the problem at hand is self-adjoint, in practical terms this fact is not a real demerit. Indeed, we saw that an easy-to-implement iterative procedure can be used to solve the system, thereby generating a fast-converging sequence of solutions of symmetric systems with a fixed matrix (to be factorized once for all before it starts, in the

case of a direct solver). It turns out that this solution procedure is much less time consuming than the direct solution. Moreover we observed that it can perform better with respect to methods whose system matrix is symmetric anyway, such as the isoparametric formulation of self-adjoint problems.

2. Error estimates for the nonparametric formulation can be proved using the well established theory of linear variational problems. We should emphasize that this is not at all restricted to the nonconforming method studied in Section 6. Indeed a similar analysis applies to many other classes of methods, such as Lagrange FEM of any order higher than one, as shown in [10] and [11], or yet Hermite FEM for biharmonic equations (cf. [10]).
3. The use of nonparametric shape and test functions allows for flexible constructions, in the sense that they are well adapted to several types of degrees of freedom, in contrast to classical formulations. In this work this property was exemplified more particularly for mean-value degrees of freedom associated with a nonconforming quadratic tetrahedral element, which adds to many other cases already addressed in [5,6,11] and [10].
4. The nonparametric Petrov–Galerkin formulation appeared to be more accurate than classical techniques for the same purpose, such as the isoparametric version of the finite element method, in case the latter exists.

Finally we note that some observations listed above had already been reported in the validation sections of previous publications such as [4–6,10] and [11]. However here the authors focused on a systematic efficiency study of the nonparametric formulation. Nevertheless they are aware of the fact that more experimentation with this new technique is necessary, in order to evaluate it in contexts other than those considered in this article. For this reason they intend to push further this kind of study in future work.

Remark 2. Besides direct methods known for roughly one hundred years or more, the numerical experimentation in this work was carried out by means of two iterative methods widely in use to solve linear systems, namely, the conjugate gradient method and the GMRES method. As a by-product of our studies, the globally great superiority of iterative methods over direct methods was highlighted once more. This is particularly due to the fact that, in principle, the former are significantly less time consuming than the latter, while enabling practitioners to work with much finer meshes. In the authors' view, both advantages largely make up for the eventual need to adjust numerical parameters or to call on side techniques for improving the convergence and/or the accuracy of iterative methods. Among them lies preconditioning, but we declined to use this technique here in order to avoid deviation from our main validation and comparison goals. This is also because preconditioning may fail, depending on the kind of technique and the FEM in use, or yet bring about little improvement of performance, owing to a substantial increase of computational effort. But nothing prevents one from testing and comparing countless techniques for enhanced linear system solving, focusing on special situations. For example, it might be interesting to check the performance of the modification of the conjugate gradient algorithm proposed in [21] for consecutive linear systems. Eventually this technique could further reduce CPU time, in the framework of the iterative solution procedure of the type (12) experimented here, as long as the problem to solve is self-adjoint and positive definite. ■

Acknowledgments

The first author gratefully acknowledges the financial support provided by CNPq, Brazil through grant 307996/2008-5. The authors are thankful to their colleague J. A. Cuminato for helpful discussions. ■

References

[1] L. Franca, T.J.R. Hughes, R. Stenberg, Stabilized finite element methods, in: M.D. Gunzburger, R.A. Nicolaides (Eds.), Incompressible Computational Fluid Dynamics, Cambridge University Press, 1994, pp. 87–107.
[2] I. Babuška, The finite element method with Lagrange multipliers, Numer. Math. 20 (1973) 170–192.
[3] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers, RAIRO Anal. Numér. 8 (2) (1974) 129–151.
[4] V. Ruas, Optimal simplex finite-element approximations of arbitrary order in curved domains circumventing the isoparametric technique, 2017, arXiv:1701.00663.
[5] V. Ruas, A simple alternative for accurate finite-element modeling in curved domains, in: Comptes-rendus du Congrès Franç, Lille, France, 2017.
[6] V. Ruas, M.A. Silva Ramos, A Hermite method for Maxwell's equations, Appl. Math. Inf. Sci. 12 (2) (2018) 271–283.
[7] S.C. Brenner, L.R. Scott, The Mathematical Theory of Finite Element Methods, in: Texts in Applied Mathematics, vol. 15, Springer, 2008.
[8] J. Nitsche, On Dirichlet problems using subspaces with nearly zero boundary conditions, in: A.K. Aziz (Ed.), The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, 1972.
[9] L.R. Scott, Finite Element Techniques for Curved Boundaries Ph.D. thesis, MIT, 1973.
[10] V. Ruas, Optimal Dirichlet-condition enforcement on curved boundaries for Lagrange and Hermite FEM with straight-edged simplexes, in: Zeitung für Angewandte Mathematik und Mechanik, 2020, http://dx.doi.org/10.1002/zamm.201900296.
[11] V. Ruas, Optimal-rate finite-element solution of Dirichlet problems in curved domains with straight-edged tetrahedra, IMA J. Numer. Anal. (2020) http://dx.doi.org/10.1093/imanum/draa029.
[12] V. Ruas, Finite element solution of 3D viscous flow problems using non standard degrees of freedom, Japan J. Ind. Appl. Math. 2 (2) (1985) 415–431.
[13] M. Crouzeix, P.A. Raviart, Conforming and nonconforming finite element methods for solving the stationary Stokes equations I, RAIRO Anal. Numér. 7 (R3) (1973) 33–75.
[14] R.A. Adams, Sobolev Spaces, Academic Press, 1975.
[15] H. Cartan, Formes différentielles, Hermann, 1967.

[16] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, North Holland, 1978.
[17] J.A. Cuminato, V. Ruas, Unification of distance inequalities for linear variational problems, Comput. Appl. Math. 34 (2015) 1009–1033.
[18] P.G. Ciarlet, P.A. Raviart, The combined effect of curved boundaries and numerical integration in isoparametric finite element methods, in: A.K. Aziz (Ed.), The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, 1972, pp. 409–474.
[19] D.B. Stein, R.D. Guy, B. Thomases, Immersed boundary smooth extension: A high-order method for solving PDE on arbitrary smooth domains using fourier spectral methods, J. Comput. Phys. 304 (2016) 252–274.
[20] R. Verfürth, A Posteriori Error Estimation Techniques for Finite Element Methods, Oxford Science Publication, 2013.
[21] J. Ehrel, F. Guyomarch, An augmented conjugate gradient method for solving consecutive symmetric positive definite linear systems, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1279–1299.