ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2008, Vol. 48, No. 12, pp. 2211–2216. © Pleiades Publishing, Ltd., 2008.
Original Russian Text © A.A. Panin, 2008, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2008, Vol. 48, No. 12, pp. 2180–2185.

On the Problem of Superconvergence of Finite Element Method Algorithms

A. A. Panin

Faculty of Physics, Moscow State University, Moscow, 119992 Russia
e-mail: a-panin@yandex.ru

Received February 26, 2008

Abstract—The coincidence of an approximate solution to the boundary value problem for an ordinary differential equation with the exact solution at mesh nodes is proved for a certain class of the generalized finite element methods.

DOI: 10.1134/S0965542508120105

Keywords: superconvergence, finite element method, generalized finite element method, Galerkin method, boundary value problem for a second-order ordinary differential equation.

INTRODUCTION

The finite element method (FEM) has been widely used for the numerical analysis of differential equations over the last 3–4 decades. A priori and a posteriori bounds have been obtained for approximate solutions found by the FEM. In particular, there are integral bounds representing an integral norm of the deviation of an approximate solution v from the exact solution u. If the bilinear form of the problem is coercive and the exact solution belongs to the class W_2^2, then a priori bounds have the form $\|u - v\|_{L^2} \leqslant Ch^2\|f\|_{L^2}$ or $\|u - v\|_{W_2^1} \leqslant Ch\|f\|_{L^2}$ (h is the maximum mesh size and f is the right-hand side of the equation), where the constant C is independent of u, v, f, and h (see [1]). Many bounds and a detailed exposition of the theory of piecewise polynomial approximation can be found in [2]. As for a posteriori bounds, they use an approximate solution itself in addition to the data of the problem. The so-called superconvergence phenomenon is also of interest. According to it, in the domain under consideration, there exist points x_i such that the difference $u(x_i) - v(x_i)$ (or $\nabla u(x_i) - \nabla v(x_i)$) tends to zero faster then the corresponding integral norm of the deviation. It is pointed out in [3] that this phenomenon was probably first mentioned in [4].

At the same time, the cases when an approximate and the exact solutions coincide at certain points are of great interest. A simple example is the one-dimensional Dirichlet problem for the equation $u'' = -f$. If an approximate solution is sought in the space of piecewise affine functions, then its values at mesh nodes coincide with the exact solution. However, if we consider piecewise cubic functions and require that their first derivatives are continuous, then this phenomenon is not observed. On the contrary, we can change the equation instead of the space of finite elements. For example, for a one-dimensional Helmholtz equation, the difference between these values can be significant because of the pollution effect (see [5]) occurring when solving this equation by the ordinary FEM. This effect implies that an approximate solution obtained by the Galerkin FEM does not coincide with the interpolation of the exact solution in the same finite-dimensional space (see [6]).

The generalized FEM (GFEM) was proposed in [7] and further developed in [8–10]. The main idea behind this method is to take into account the structure of an equation when constructing a finite-dimensional space whose shape functions (also called interpolation functions) can differ from polynomials. (Note that this approach is a kind of return to Galerkin methods with nonpolynomial basis functions.) Some general results on superconvergence are obtained for this method in [3], where a nonhomogeneous Neumann problem for the Poisson equation is theoretically studied in a two-dimensional domain. The results are illustrated by numerical experiments for a one-dimensional problem. Below, we deal with a different case.

1. DIFFERENTIAL PROBLEM

Consider the Dirichlet problem for the ordinary differential equation

$$\begin{gathered} Lu \equiv -u'' + b(x)u' + c(x)u = f, \quad x \in (0; 1), \\ u(0) = u(1) = 0. \end{gathered} \tag{1}$$

Assume that $f \in L^2(0; 1)$, $c \in L^\infty(0; 1)$, and $b(x)$ is W^1_∞-piecewise, which means that the interval (0, 1) can be divided into a finite number of intervals (x_k, x_{k+1}) with $b(x)$ belonging to $W^1_\infty(x_k; x_{k+1})$ on each interval. (b is not necessarily continuous on the entire (0, 1).) In addition, no constraints are imposed on the signs of b and c; that is, the corresponding bilinear form can be non-coercive. In this case, problem (1) can be unsolvable or have an infinite number of solutions. However, in many practical cases, the problem of invertibility of the operator L (if it has not been solved yet) can be studied experimentally while seeking a numerical solution (see [11, 12]); therefore, this problem is not important in this case.

The generalized statement of the problem is

$$a(u, w) \equiv \int\limits_0^1 [u'w' + b(x)u'w + c(x)uw]dx = \int\limits_0^1 f(x)wdx. \tag{2}$$

Problem (1) will be solved approximately using the Galerkin method. Take M trial functions $\{\varphi_m(x)\}_{m=1}^M$ and the same number of test functions $\{\psi_m(x)\}_{m=1}^M$, each belonging to the space $\overset{\circ}{W}{}^1_2(0; 1)$. (It is well known that the functions are continuous in this case.). An approximate solution is sought in the form $v(x) = \Sigma_{m=1}^M v_m\varphi_m(x)$.

In this paper, we do not consider convergence problems associated with the Galerkin method for the class of non-self-adjoint noncoercive bilinear forms $a(u, w)$. We are interested in finding conditions under which the resulting solution coincides with the exact solution at certain prescribed points. Assume that the set of points $0 = x_0 < \ldots < x_{N-1} < x_N = 1$ is given. Here and in what follows, we assume that this set contains all the discontinuity points of the function $b(x)$. In addition, we require that the trial functions contain a system of functions $\{\varphi_m\}_{m=1}^{N-1}$ (for the sake of convenience, they are placed at the beginning of the list) such that their linear combinations satisfy the conditions

$$\sum_{m=1}^{N-1} \alpha_m^{(i)}\varphi_m(x_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots, N-1,$$

which is equivalent to the fact that the square matrix $\{\varphi_m(x_n)\}_{m,n=1}^{N-1}$ is invertible. Then, if the system of linear algebraic equations (SLAE) of the Galerkin method is solvable, the exact solution at the given points can be found by an appropriate choice of test functions.

2. THE MAIN RESULT

Lemma. *Replacing the linearly independent system of trial functions* $\{\varphi_m\}$ *with another basis of the space generated by it does not affect the solvability of the SLAE of the Galerkin method or the form of an approximate solution.*

Proof. Let $V \equiv \{v_1, \ldots, v_M\}^{\mathrm{T}}$ be the coordinates of the function v in the initial basis $\{\varphi_m\}$. Then, the SLAE of the Galerkin method has the form $a(\Sigma_{m=1}^M v_m\varphi_m, \psi_l) = f_l$, where a is the bilinear form of the problem under consideration (see (2)) and $f_l = \int_0^1 f(x)\psi_l(x)dx$. Consider the matrix $A = \{A_{lm}\} = \{a(\varphi_m, \psi_l)\}$ and the vector $F = \{f_l\}^{\mathrm{T}}$. Then, the SLAE has the form $AV = F$. Change the basis $\{\varphi_m\}$ to the basis $\{\widetilde{\varphi_m}\}$. Denote the row vector $\{\varphi_m(x)\}$ of the old basis by φ, the row vector $\{\widetilde{\varphi_m}(x)\}$ of the new basis by $\tilde{\varphi}$, and the corresponding transformation matrix by Φ so that $\tilde{\varphi} = \varphi\Phi^{-1}$. Then, the matrix of the bilinear form in the new basis has the form $\tilde{A} = (\tilde{A}_{lm}) = a(\widetilde{\varphi_m}, \psi_l) = A\Phi^{-1}$ and the new coordinate vector is $\tilde{V} = \Phi V$. As we can see, since the transformation matrix is invertible, the SLAE in the expansion coefficients of the approximate solution in terms of the old basis remains the same: $A\Phi^{-1}\Phi V \equiv AV = F$. This completes the proof.

To construct test functions, consider the operator L^* that is adjoint of L in the Lagrangian sense (e.g., see [13, p. 64]): $L^*w \equiv -w'' - (b(x)w)' + c(x)w$.

Assume that, on each interval $(x_{n-1}; x_n)$ $(i = 1, 2, \dots, N)$, the homogeneous Dirichlet problem for the operator L^*

$$L^*\psi(x) = 0, \quad \psi(x_{n-1}) = \psi(x_n) = 0 \tag{3}$$

has only trivial solutions; that is, zero is not an eigenvalue of these operators on any of these intervals. (However, it is well known that this is equivalent to the unique solvability of the problem for the operator L.) Then, the boundary value problems

$$L^*\psi_n^{(1)}(x) = 0, \quad \psi_n^{(1)}(x_{n-1}) = 1, \quad \psi_n^{(1)}(x_n) = 0, \tag{4}$$

and

$$L^*\psi_n^{(2)}(x) = 0, \quad \psi_n^{(2)}(x_{n-1}) = 0, \quad \psi_n^{(2)}(x_n) = 1, \tag{5}$$

are uniquely solvable.

The test functions are chosen using an approach generalizing piecewise affine functions that are usual for the conventional FEM. Assume that

$$\psi_n(x) = \begin{cases} \psi_n^{(2)}(x) \quad \text{at} \quad x \in [x_{n-1}; x_n], \\ \psi_{n+1}^{(1)}(x) \quad \text{at} \quad x \in [x_n; x_{n+1}], \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

for $n = 1, 2, \dots, N$. It is clear that, by conditions (4) and (5), the functions $\{\psi_m(x)\}_{m=1}^{N-1}$ are continuous on the interval [0, 1] and form a Lagrange basis in the corresponding subspace of the space $\overset{\circ}{W}{}_2^1(0; 1)$; that is, $\psi_m(x_n) = \delta_{mn}$ $(m, n = 1, 2, \dots, N-1)$.

First, consider the case $M = N - 1$; that is, assume that the entire system of test functions is defined by formulas (6) and the number of trial functions $\varphi_m(x)$ is the same.

Theorem 1. *Assume the following:*

(1) *The differential problem in generalized statement* (2) *is uniquely solvable.*

(2) *Problems* (3) *for all intervals* $(x_{n-1}; x_n)$ $(n = 1, 2, \dots, N)$ *have only trivial solutions.*

(3) *The trial functions belong to* $\overset{\circ}{W}{}_2^1(0; 1)$ (*note that the test functions belong to this space by construction*).

(4) *There exist linear combinations* $\widetilde{\varphi_m}(x) = \Sigma_{n=1}^{N-1}\alpha_n^{(m)}\varphi_n(x)$ *of the trial functions such that* $\widetilde{\varphi_m}(x_n) = \delta_{mn}$ $(m, n = 1, 2, \dots, N-1)$ (*this implies, in particular, the linear independence of both the system of the above linear combinations and the original system of functions* $\{\varphi_m(x)\}$).

(5) *The SLAE of the Galerkin method with the functions* $\{\varphi_m(x)\}_{m=1}^{N-1}$ and $\{\psi_m(x)\}_{m=1}^{N-1}$ *constructed by formula* (6) *is uniquely solvable.*

Then, it holds that

$$v(x_n) = u(x_n), \quad n = 0, 1, \dots, N.$$

Proof. First, we note that the boundary conditions are fulfilled by condition 3 of the theorem. Take $\widetilde{\varphi_m}(x)$ as trial functions. By the lemma proved above (by the way, $\{\Phi_{mn}\} = \{\varphi_n(x_m)\}$ in it), this does not affect the finite-dimensional problem solvability and the approximate solution $v(x)$. These functions are denoted by $\varphi_m(x)$ as above. In this case, the coefficient v_m in the expansion $v(x) = \Sigma_{m=1}^{N-1}v_m\varphi_m(x)$ of the approximate solution is equal to $v(x_m)$.

Then, the SLAE of the Galerkin method is written as

$$a(v, \psi_l) = \int_0^1 f(x)\psi_l(x)dx, \quad l = 1, 2, \ldots, N-1. \tag{7}$$

Since the exact solution u also satisfies the generalized statement of the problem, we have

$$a(u, \psi_l) = \int_0^1 f(x)\psi_l(x)dx, \quad l = 1, 2, \ldots, N-1,$$

and (7) can be rewritten in the form

$$a(v, \psi_l) = a(u, \psi_l), \quad l = 1, 2, \ldots, N-1,$$

or in the expanded form

$$\int_0^1 [v'\psi_l' + b(x)v'\psi_l + cv\psi_l]dx = \int_0^1 [u'\psi_l' + b(x)u'\psi_l + cu\psi_l]dx.$$

Recall that $b(x) \in W_\infty^1(x_{n-1}; x_n)$ on each interval. Therefore, ψ_l belong to W_2^2 as solutions to the adjoint problems of type (4) on each interval $(x_{n-1}; x_n)$ (see [14]; however, they only belong to $\overset{\circ}{W}{}_2^1$ on the entire interval [0, 1]); consequently, we can integrate by parts, which implies the equations

$$\begin{gathered}
\sum_{n=1}^{N}\int_{x_{n-1}}^{x_n}[-v\psi_l'' - v(b(x)\psi_l)' + cv\psi_l]dx \\
+ \sum_{n=1}^{N}\{v(x_n)[\psi_l'(x_n-0) - b(x_n-0)\psi_l(x_n)] - v(x_{n-1})[\psi_l'(x_{n-1}+0) - b(x_{n-1}+0)\psi_l(x_{n-1})]\} \\
= \sum_{n=1}^{N}\int_{x_{n-1}}^{x_n}[-u\psi_l'' - u(b(x)\psi_l)' + cu\psi_l]dx \\
+ \sum_{n=1}^{N}\{u(x_n)[\psi_l'(x_n-0) - b(x_n-0)\psi_l(x_n)] - u(x_{n-1})[\psi_l'(x_{n-1}+0) - b(x_{n-1}+0)\psi_l(x_{n-1})]\}.
\end{gathered} \tag{8}$$

Here, the integrals are taken over individual intervals in order to point out that the functions ψ_l are twice differentiable only within each interval rather than on the entire interval because the first derivatives are discontinuous at the points x_n. The function $b(x)$ can also have discontinuities at those points. For the same reason, we use the symbols of limiting values for ψ_l' and b. The functions u, v, and ψ_l are continuous everywhere in [0, 1] because $u, v \in \overset{\circ}{W}{}_2^1(0; 1)$ and ψ_l is continuous by construction. Taking into account that the functions $\psi_l(x)$ satisfy the homogeneous equation $L^*\psi_l = 0$ on each interval $(x_{n-1}; x_n)$ by construction, we rewrite system (8) in the form

$$\begin{gathered}
\sum_{n=1}^{N}\{v(x_n)[\psi_l'(x_n-0) - b(x_n-0)\psi_l(x_n)] - v(x_{n-1})[\psi_l'(x_{n-1}+0) - b(x_{n-1}+0)\psi_l(x_{n-1})]\} \\
= \sum_{n=1}^{N}\{u(x_n)[\psi_l'(x_n-0) - b(x_n-0)\psi_l(x_n)] - u(x_{n-1})[\psi_l'(x_{n-1}+0) - b(x_{n-1}+0)\psi_l(x_{n-1})]\}.
\end{gathered} \tag{9}$$

In addition, take into consideration that, on the one hand, $\{v(x_m)\}_{m=1}^{N-1}$ are the coefficients v_m in the expansion of the approximate solution $v(x)$ in terms of $\varphi_m(x)$ (due to the choice of $\varphi_m(x)$, which was pointed out at the beginning of the proof) and, on the other hand, system (9) was derived using identical transformations of original system (7), which is uniquely solvable by condition. Hence, system (9) considered as a system in $v_m \equiv v(x_m)$ is also uniquely solvable. (Note that $v(x_0) = v(x_N) = 0$; therefore, they are known.) Note also that $u(x_0) = u(x_N) = 0$). It is obvious that the numbers $v(x_m) = u(x_m)$ are solutions to the system. Since the solution is unique, $v_m \equiv v(x_m)$ necessarily coincide with $u(x_m)$. The theorem is proved.

This result can easily be generalized by supplementing the chosen functions $\{\varphi_m(x)\}_{m=1}^{N-1}$ and $\{\psi_m(x)\}_{m=1}^{N-1}$ with the functions $\{\varphi_m(x)\}_{m=N}^{M}$ and $\{\psi_m(x)\}_{m=N}^{M}$ ($M > N$) that also belong to $\overset{\circ}{W}{}_2^1(0; 1)$.

Theorem 2. *Assume that problem* (1) *and the functions* $\{\varphi_m(x)\}_{m=1}^{N-1}$ *and* $\{\psi_m(x)\}_{m=1}^{N-1}$ *satisfy all the conditions of Theorem* 1. *Also, assume that* $\{\varphi_m(x)\}_{m=N}^{M}$ *and* $\{\psi_m(x)\}_{m=N}^{M}$ *belong to* $\overset{\circ}{W}{}_2^1(0; 1)$ and the SLAE *of the Galerkin method for the new systems of functions* $\{\varphi_m(x)\}_{m=1}^{M}$ *and* $\{\psi_m(x)\}_{m=1}^{M}$ *is uniquely solvable. Then, the function* $v(x) = \Sigma_{m=1}^{M} v_m\varphi_m(x)$ *satisfies the equalities* $v(x_n) = u(x_n)$ *for* $n = 0, 1, \dots, N$.

It should be emphasized that we do not increase the number of points at which the equality is valid. It is important that supplementing the systems of trial and test functions, which have already been considered, with new functions cannot violate the equalities $v(x_n) = u(x_n)$ fulfilled for these systems of functions.

Proof. It should be noted that, for the numbers $v(x_n)$, we can similarly obtain system (9), which is not the SLAE of the Galerkin method in the new systems of functions and in which $v(x_n)$ do not coincide with v_n ($n = 1, 2, \dots, N - 1$). However, the coefficients of this system determined only by the functions $\{\psi_l(x)\}_{l=1}^{N-1}$ coincide with those of the SLAE in Theorem 1, which is uniquely solvable by condition. Therefore, $v(x_n)$ ($n = 1, 2, \dots, N - 1$) are uniquely determined as in Theorem 1 and, since $v(x_n) = u(x_n)$ satisfy the system, this completes the proof.

Remark. Note once again that, to obtain the results described above, it is necessary to include all the discontinuity points of the function $b(x)$ in the set $\{x_n\}$.

3. SOME REMARKS CONCERNING THE APPLICATION OF THE RESULTS

Theorems 1 and 2 show that, using the proposed trial and test functions, in addition to a priori and a posteriori estimates of convergence of approximate solutions to the exact solution in integral and uniform norms established earlier, we can find exact values of the solution at any points of interest if analytical solutions to problems (4) and (5) with homogeneous equations are known.

This result is applicable, in particular, to solving problem (1) by the FEM, and it can be considered as a generalization of the well-known facts on superconvergence to one class of generalized finite elements and as a generalization of the similar fact concerning piecewise polynomial interpolation (see [15] or [16]). On the other hand, this is an improvement of the results concerning the FEM convergence considered in [5] and in the example in [17]. From the viewpoint of the finite difference method, we can say that we proposed a method for constructing an exact difference scheme.

REFERENCES

1. G. I. Marchuk and V. I. Agoshkov, *Introduction to Mesh Projection Methods* (Nauka, Moscow, 1981) [in Russian].
2. Ph. Ciarlet, *The Finite Element Method for Elliptic Problems*, (North-Holland, Amsterdam, 1977; Mir, Moscow, 1980).
3. I. Babuška, U. Banerjee, and J. E. Osborn, “Superconvergence in the Generalized Finite Element Method,” Techn. Report no. 0545, Univ. of Texas, Austin, 2005; http://www.ices.utexas.edu/research/reports/2005/0545.pdf.
4. L. A. Oganesyan and L. A. Rukhovets, “The Rate of Convergence of Varitional-Difference Schemes for Second-Order Elliptic Equations in a Two-Dimensional Domain with Smooth Border,” Zh. Vychisl. Mat. Mat. Fiz. **9**, 1102–1120 (1969).
5. I. M. Babuška and S. A. Sauter, “Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers?,” SIAM J. Numer. Anal. **34**, 2392–2423 (1997).
6. J. T. Oden, S. Prudhomme, and L. Demkowicz, “A Posteriori Error Estimation for Acoustic Wave Propagation Problems,” Techn. Report no. 0432, Univ. of Texas, Austin, 2004; http://www.ices.utexas.edu/research/reports/2004/0432.pdf.
7. I. Babuška, G. Caloz, and J. Osborn, “Special Finite Element Methods for a Class of Second Order Elliptic Problems with Rough Coefficients,” SIAM J. Numer. Anal. **31**, 945–981 (1994).

8. J. M. Melenk, *On Generalized Finite Element Methods*, PhD Thesis (Univ. of Maryland at College Park, Maryland, 1995).
9. J. M. Melenk and I. Babuška, "The Partition of Unity Finite Element Method: Basic Theory and Applications," Comput. Meth. Appl. Mech. Eng. **139**, 289–314 (1996).
10. I. Babuška and J. M. Melenk, "The Partition of Unity Finite Element Method," Int. J. Numer. Meth. Eng. **40**, 727–758 (1997).
11. M. T. Nakao, K. Hashimoto, and Y. Watanabe, "A Numerical Method to Verify the Invertibility of Linear Elliptic Operators with Applications to Nonlinear Problems," Computing **75** (1), 1–14 (2005).
12. M. T. Nakao and K. Hashimoto, "Constructive Error Estimates of Finite Element Approximations for Non-Coercive Elliptic Problems and Its Applications: MHF Preprint Series, no. MHF 2007-5, Kyushi Univ., Fukuoka, Japan, pp. 1-12; http://hdl.handle.net/2324/3405.
13. S. G. Mikhlin, *Direct Methods in Mathematical Physics* (Gostekhteorizdat, Moscow, 1950) [in Russian].
14. O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics* (Nauka, Moscow, 1973; Springer, New York, 1985).
15. M. T. Nakao, N. Yamamoto, and S. Kimura, "On the Best Constant in the Error Bound for the H_0^1-Projection into Piecewise Polynomial Spaces," J. Approx. Theory **93**, 491–500 (1998).
16. M. Schultz, *Spline Analysis* (Prentice Hall, London, 1973).
17. S. Sauter, "A Refined Finite Element Convergence Theory for Highly Indefinite Helmholtz Problems," Computing **78** (2), 101–115 (2006).