On computability of the Galerkin procedure

By Atsushi YOSHIKAWA*)

Professor Emeritus, Kyushu University

(Communicated by Heisuke HIRONAKA, M.J.A., May 14, 2007)

Abstract: It is shown that the Galerkin approximation procedure is an effective representation of the solution of a computable coercive variational problem in a computable Hilbert space.

Key words: computable Hilbert space, Galerkin approximation, Lax-Milgram theorem.

1. Introduction. In numerical analysis of a variational problem, a standard and basic method is the Galerkin approximation. It is true that numerical analysis and computable analysis, although both are deeply concerned with computers, do not necessarily share the motivations. However, similitude, if any exists, should be expected in such an approximating procedure. Here we explain how the Galerkin procedure in variational problems should be interpreted in the context of computable analysis.

In fact, we show that, in an environment which is reasonable in the sense of computable analysis, a computable solution of a computable variational problem is effectively realized by a computable sequence of Galerkin approximations. That is, in a computable Hilbert space, the unique computable solution of a computable version of the Lax-Milgram theorem is in fact realized as the effective limit of the computable sequence of the Galerkin approximations.

2. The Galerkin approximation and Céa's estimate. Let $\mathbf{X}$ be a separable Hilbert space over the real $\mathbf{R}$ (with the inner product $\langle\,,\,\rangle$ and the norm $\|\,\|$). Let $B(x,y)$ be a bounded coercive (or strongly accretive) bilinear form on $\mathbf{X}$. Thus, $B(x,y)$ is a bilinear form in $\mathbf{X}$ which satisfies $\mu\,\|x\| \le B(x,x)$, $x \in \mathbf{X}$ for some $\mu > 0$ and $|B(x,y)| \le M\,\|x\|\,\|y\|$, $x,\, y \in \mathbf{X}$ for some $M > 0$. The optimal values of μ and M are called the coercivity constant and the bound of the form B, respectively. Let F be a given bounded linear functional on $\mathbf{X}$. Note that there is an $f \in \mathbf{X}$ such that

$$F(v) = \langle f,\, v\rangle \quad \textit{for all} \quad v \in \mathbf{X} \tag{2.1}$$

by the Riesz-Fréchet theorem.

Consider the following variational problem: *Find* $u \in \mathbf{X}$ *such that*

$$B(u,v) = F(v) \quad \textit{for all} \quad v \in \mathbf{X}. \tag{2.2}$$

Actually (2.2) is uniquely solvable by the Lax-Milgram theorem [4] (and already by the Riesz-Fréchet theorem when B is symmetric). In fact, by the Lax-Milgram theorem, we have a linear isomorphism $T : \mathbf{X} \to \mathbf{X}$ such that

$$B(x,y) = \langle Tx, y\rangle \quad \textit{for all} \quad x, y \in \mathbf{X}. \tag{2.3}$$

Therefore, $u = T^{-1} f$ will do since (2.2) is equivalent to the equation $T\,u = f$.

Now suppose that $s_1, \cdots, s_N$ are linearly independent vectors in $\mathbf{X}$ and the corresponding subspace they generate is denoted by $\mathbf{X}_N$. Let P_N be the orthogonal projection from $\mathbf{X}$ onto $\mathbf{X}_N$. Let B_N and F_N be respectively the restrictions of B and F on $\mathbf{X}_N$. (2.2) then is reduced to the following : *Find* $u_N \in \mathbf{X}_N$ *such that*

$$B_N(u_N, v_N) = F_N(v_N) \quad \textit{for all} \quad v_N \in \mathbf{X}_N \tag{2.4}$$

or equivalently

$$(P_N\,T\,P_N)\;u_N = P_N\,f \tag{2.5}$$

with the operator $P_N\,T\,P_N$ acting on $\mathbf{X}_N$.

Now using the basis $s_1, \cdots, s_N$, we represent $\mathbf{X}_N$ as the linear space $\mathbf{R}^N$. Put $\mathbf{u}_N = {}^t(x_1, \cdots, x_N)$ and $\mathbf{v}_N = {}^t(y_1, \cdots, y_N)$ where $u_N = P_N\,u = \sum_{j=1}^N x_j\,s_j$ and $v_N = P_N\,v = \sum_{k=1}^N y_k\,s_k$. Then the bilinear form B_N is expressed in the form of an $N \times N$ matrix $\mathscr{B}_N = {}^t\big(B(s_j, s_k)\big)$, which is non-degenerate because of coercivity of B. Then (2.5) turns out to be a system of N inhomogeneous linear equations:

$$\mathscr{B}_N\,\mathbf{u}_N = \mathbf{f}_N$$

2000 Mathematics Subject Classiffcation. MSC: 03D80; 65J10; 46C05; 41A65.

*) Corresponding address: Faculty of Mathematics, Kyushu University, 6–10–1 Hakozaki, Higashi-ku, Fukuoka 812–8581, Japan.

where $\mathbf{f}_N = {}^t(F(s_1), \cdots, F(s_N))$. $\mathbf{u}_N$ is thus uniquely solved by the standard linear algebra. Hence u_N in (2.5) is determined with a certain explicitness by transporting $\mathbf{u}_N$ to the coefficients of $s_1, \cdots, s_N$.

In particular, note

$$B(u - u_N, v_N) = 0 \quad \textit{for all} \quad v_N \in \mathbf{X}_N. \tag{2.6}$$

In fact, from (2.2) and (2.4), the left-hand side equals to $\langle f, v_N\rangle - \langle f_N, v_N\rangle$, which vanishes since $v_N = P_N v_N$.

We call u_N the Galerkin approximation of u in $\mathbf{X}_N$.

Remark 1. Put $w_N = T^{-1} P_N f \in \mathbf{X}$. Since $P_N\, u_N = u_N \in \mathbf{X}_N$, we have

$$P_N\, T P_N(u_N - P_N w_N) = -P_N(I - P_N)w_N.$$

Then

$$\begin{aligned}\mu\, \|u_N - P_N w_N\| &\le \|(I - P_N)w_N\| \\ &\le \|(I - P_N)T^{-1}(I - P_N)f\| \\ &\quad + \|(I - P_N)T^{-1} f\|\end{aligned}$$

by (2.3) and the coercivity of B.

We have the following Céa's estimate ([3]. See also [2]):

Proposition 1. *Let $\mu > 0$ be the coercivity constant of the form B and M its bound. Then we have*

$$\begin{aligned}\|u - u_N\| &\le \frac{M}{\mu} \inf_{v_N \in \mathbf{X}_N} \|u - v_N\| \\ &= \frac{M}{\mu} \min_{v_N \in \mathbf{X}_N} \|u - v_N\|.\end{aligned} \tag{2.7}$$

In fact, by the coercivity,

$$\mu\, \|u - u_N\|^2 \le B(u - u_N, u - v_N) + B(u - u_N, u_N - v_N)$$

and the second term on the right-hand side vanishes because of (2.6). Recall then

$$|B(u - u_N, u - v_N)| \le M\, \|u - u_N\|\, \|u - v_N\|$$

which yields the estimate (2.7).

Remark 2. Let $\mathrm{dist}\,(u, \mathbf{X}_N)$ denote the distance of $u \in \mathbf{X}$ to the closed subspace $\mathbf{X}_N$ and $u_N^{\perp} \in \mathbf{X}_N$ the orthogonal projection of u to $\mathbf{X}_N$. Then

$$\mathrm{dist}\,(u, \mathbf{X}_N) = \|u - u_N^{\perp}\| = \min_{v_N \in \mathbf{X}_N} \|u - v_N\|.$$

Now let $\mathcal{S} = \{s_j\}_{j=1,2,\cdots}$ be a system of linearly independent elements of $\mathbf{X}$ such that its linear span is dense in $\mathbf{X}$. If, for each n, $\mathbf{X}_n$ is a subspace generated by $s_1, \cdots, s_n$, then $\{\mathbf{X}_n\}_{n=1,2,\cdots}$ is an increasing sequence of subspaces such that the union $\bigcup_{n=1}^{\infty} \mathbf{X}_n$ is dense in $\mathbf{X}$. Therefore, for each $x \in \mathbf{X}$, $\lim_{n\to\infty} \mathrm{dist}\,(x, \mathbf{X}_n) = 0$. Then the Galerkin approximations u_n in $\mathbf{X}_n$ coverges to the solution $u \in \mathbf{X}$ as $n \to \infty$, but the rate of convergence depends on $\mathcal{S}$ and u (i.e., B and F).

3. Computability of the Galerkin procedure. Now suppose furthermore that $\mathbf{X}$ is an effective separable Hilbert space (See [5, 6]). Recall that a Hilbert space is computable if it is endowed with a computability structure and is effective separable if it in addition has an effectively generating set [5]. We first review in a very sketchy manner the notion of computability in a Hilbert space.

A computability structure in $\mathbf{X}$ is by definition a non-empty set $\mathscr{S}$ of *computable* sequences, which are specified by the three axioms for a computability structure ([5; Chapter 2. p.81], in particular. See also the discussions below). An element $x \in \mathbf{X}$ is computable if the sequence $\{x, x, \cdots\}$ lies in $\mathscr{S}$ or is computable. A computable sequence $\mathcal{S} = \{s_n\}_{n=1,2,\cdots}$, that is, $\mathcal{S} \in \mathscr{S}$, is an effectively generating set of $\mathbf{X}$ if its rational linear span $\mathcal{D}$ is dense in $\mathbf{X}$. If $\mathbf{X}$ is an effective separable Hilbert space, all the sequences that are *effectively* obtained actually from $\mathcal{S}$ form a set of sequences which satisfy the axioms for a computability structure in $\mathbf{X}$ (This construction will be explained shortly). This set coincides with the original computability structure $\mathscr{S}$ in $\mathbf{X}$. That is, the computablitiy structure $\mathscr{S}$ is determined as the set of all the sequences effectively derived from $\mathcal{S}$. In this sense, for any given countable system $\mathcal{S}$ such as discussed in Remark 2, we can talk of the computability structure it determines.

There is a computable complete orthonormal system $\mathcal{E} = \{e_n\}_{n=1,2,\cdots}$ such that each s_j is a computable linear combination of $e_1, \cdots, e_j$. Note that if $\mathcal{S}$ is a system of linearly independent elements, then each e_j can be chosen as a linear combination of $s_1, \cdots, s_j$ (The Gram-Schmidt procedure. See [5, pp. 139–140]). $\mathcal{E}$ is also an effectively generating set and determines the same computability structure $\mathscr{S}$ as $\mathcal{S}$ does.

Now we explain more explicitly how the computability structure $\mathscr{S}$ is constructed from the effectively generating set $\mathcal{E}$ (and just in a similar manner if $\mathcal{E}$ is replaced by $\mathcal{S}$). A sequence $\{y_k\}_{k=1,2,\cdots}$ is computable if $y_k = \sum_{n=1}^{d(k)} c_{nk}\, e_n$ for each k. Here

$d : \mathbf{N} \to \mathbf{N}$ is a recursive function and $\{c_{nk}\}$ is a computable double sequence of rational numbers. Any effective limit $\{x_k\}_{k=1,2,\cdots}$ of a computable double sequences $\{y_{m,k}\}_{k,m=1,2,\cdots}$, is computable. Here we mean by an effective limt that there is a recursive function $r : \mathbf{N}^2 \to \mathbf{N}$ such that, for all N, k, $m > r(N,k)$ implies $\|x_k - y_{m,k}\| < 2^{-N}$. Finally, an element $x \in \mathbf{X}$ is computable if there is a computable sequence $\{x_k\}$ such that $\|x - x_k\| < 2^{-N}$ for $k \geq e(N)$ with an appropriate recursive function $e : \mathbf{N} \to \mathbf{N}$ (effective convergence. See [5]). We may here renumber $\{x_k\}$ so that $e(k) = k$ (fast convergence. [5]) unless the recursive numbering of $\{x_k\}$ itself is at stake.

Remark 3. If $\mathbf{X}$ is infinite dimensional, it admits infinitely many computably non-equivalent computable structures (See [5]. cf. also [7]). In concrete spaces such as the Sobolev spaces, we have the canonical computable structures, which are compatible with most of classical calculus.

Now consider $\mathcal{S}$ as in Remark 2. Recall that each subspace $\mathbf{X}_N$ is generated by $s_1, \cdots, s_N$ from the above $\mathcal{S}$ (or $e_1, \cdots, e_N$ from $\mathcal{E}$ in the present case). Therefore, any element of $\mathbf{X}_N$ is a linear combination of $s_1, \cdots, s_N$ or of $e_1, \cdots, e_N$.

Let $P_N : \mathbf{X} \to \mathbf{X}_N$ be the orthogonal projection. Then we have the following (see [5; Lemma 1, p.136]).

Lemma 1. *Let $x \in \mathbf{X}$. Let $x_N = P_N x$ $(N = 1, 2, \cdots)$ be the orthogonal projection of x on $\mathbf{X}_N$. Then x is a computable element of $\mathbf{X}$ if and only if $\{x_N\}_{N=1,2,\cdots}$ is a computable sequence in $\mathbf{X}$. Moreover, then $\{x_N\}$ converges effectively to x, that is, $\|x_N - x\| < 2^{-k}$ for $N > e(k)$ with an appropriate recursive function $e : \mathbf{N} \to \mathbf{N}$.*

Let us return to our original problem (2.2). Take a bilinear form B on $\mathbf{X}$. We may consider the conditions :

$$\text{(3.1)} \qquad \begin{array}{l} \textit{the double sequence} \\ \big\{B(e_j, e_k)\big\}_{j,k=1.2.\cdots} \quad \textit{is computable} \end{array}$$

and also

$$\text{(3.2)} \qquad \begin{array}{l} \textit{the sequence} \\ \left\{\displaystyle\sum_{k=1}^{\infty} |B(e_j, e_k)|^2\right\}_{j=1,2,\cdots} \quad \textit{is computable.} \end{array}$$

The meaning of these conditions is fully discussed in Brattka-Yoshikawa [1]. In particular, in case when B is coercive, the coercivity constant μ and the bound M are *computable* reals.

As for a bounded linear functional F on $\mathbf{X}$ (see (2.1)), we may consider the conditions

$$\text{(3.3)} \qquad \begin{array}{l} \textit{the sequence} \\ \{F(e_n)\}_{n=1,2,\cdots} \quad \textit{is computable} \end{array}$$

and

$$\text{(3.4)} \qquad \begin{array}{l} \textit{the functional norm} \\ \|F\|_* = \displaystyle\sup_{\|v\|=1} |F(v)| \quad \textit{is a computable real.} \end{array}$$

Remark 4. The conditions (3.1) and (3.3) are of the same nature and so are the conditions (3.2) and (3.4). This is seen by taking the tensor product $\mathbf{X}\widehat{\otimes}\mathbf{X}$ into account.

Now a version of the Lax-Milgram theorem in computable analysis reads as follows:

Theorem 1. *Suppose $\mathbf{X}$ is an effective separable Hilbert space. Let B be a coercive bilinear form on $\mathbf{X}$ satisfying* (3.1) *and* (3.2). *Then* (2.3) *is valid with T such that both T and T^{-1} map computable sequences in $\mathbf{X}$ into computable sequences in $\mathbf{X}$. For any bounded linear functional F on $\mathbf{X}$ which satisfy* (3.3) *and* (3.4), *there are uniquely determined computable elements $u \in \mathbf{X}$ and $u^1 \in \mathbf{X}$ such that*

$$F(v) = B(u, v) = B(v, u^1) \quad \textit{for any} \quad v \in \mathbf{X}.$$

For the proof, we refer to Brattka-Yoshikawa [1].

Now we reconsider Proposition 1 in the situation of Theorem 1. Since $s_1, \cdots, s_N$ are *computable* elements of $\mathbf{X}$ generated by $e_1, \ldots, e_N$, the system of equations (2.5) consists of a matrix $\mathscr{B}_N$ with *computable* entries and a right-hand side $\mathbf{f}_N$ with *computable* components because of the assumptions of Theorem 1. Hence, $\mathbf{u}_N$ is of computable components and thus the corresponding vector u_N in $\mathbf{X}_N$, the Galerkin approximation of u in $\mathbf{X}_N$, is a *computable* element of $\mathbf{X}$.

We have to show that the sequence $\{u_N\}_{N=1,2,\cdots}$ is a computable sequence in $\mathbf{X}$, and that this sequence effectively converges to u.

We first show the latter part. It is enough to pick up an effectively determined subsequence of $\{u_N\}$ which effectively converges to u.

Now since u is an computable element in $\mathbf{X}$, there is a computable sequence $\{\tilde{u}_n\}$ in $\mathbf{X}$ which effectively converges to u. More precisely, we have $\|u - \tilde{u}_k\| < 2^{-k}$, where

$$\tilde{u}_k = \sum_{j=1}^{\ell(k)} c_{jk}\, e_j$$

with a computable double sequence $\{c_{jk}\}$ and a recursive function $\ell : \mathbf{N} \to \mathbf{N}$, which can be strictly increasing. Note that the sequence $\{\tilde{u}_n\}$ is in general not related to the Galerkin approximations of u.

Observe that we may write

$$\tilde{u}_k = \sum_{j=1}^{\ell(k)} c'_{jk}\, s_j \in \mathbf{X}_{\ell(k)}$$

with a computable double sequence $\{c'_{jk}\}$ because of the Gram-Schmidt procedure to obtain $\mathcal{E}$ from $\mathcal{S}$. Here we recall that, for each N, $\mathbf{X}_N$ is a linear subspace of $\mathbf{X}$ generated by $s_1, \cdots, s_N$. Thus, in particular, $\mathrm{dist}\,(u, \mathbf{X}_{\ell(k)}) \le \|u - \tilde{u}_k\|$.

Consider now the sequence $\{u_{\ell(k)}\}$, which is an effectively determined subsequence of the Galerkin approximations $\{u_N\}$. For each Galerkin approximation $u_{\ell(k)} \in \mathbf{X}_{\ell(k)}$, we have

$$\begin{aligned}\|u - u_{\ell(k)}\| &\le \frac{M}{\mu}\mathrm{dist}\,(u, \mathbf{X}_{\ell(k)}) \\ &\le \frac{M}{\mu}\|u - \tilde{u}_k\| < \frac{M}{\mu}\, 2^{-k}\end{aligned}$$

by Proposition 1. Note $\frac{M}{\mu}$ is a computable real. Hence, the sequence $\{u_{\ell(k)}\}$ effectively converges to u and so does the full sequence $\{u_N\}$ of Galerkin approximations since the sequence of the subspaces $\mathbf{X}_N$ is increasing.

The above argument is still not enough to ensure computability of the sequence $\{u_N\}_{N=1,2,\cdots}$. Recall Lemma 1 together with Remark 1. Observe that $\{P_N f\}$ is a computable sequence in $\mathbf{X}$ since $f \in \mathbf{X}$ is a computable element. Hence, $\{w_N\}_{N=1,2,\cdots}$ and also $\{P_N\, w_N\}_{N=1,2,\cdots}$ are computable sequences in $\mathbf{X}$. By Lemma 1, the computable sequence of reals $\{\|(I - P_N)w_N\|\}_{N=1,2,\cdots}$ converges to 0 effectively. Therefore, $\{u_N\}$ is a computable sequence in $\mathbf{X}$ since it is the effective limit of $\{P_N\, w_N\}$.

Summarizing the above arguments, we have shown the following

Theorem 2. *Suppose that* $\mathbf{X}$ *is an effectively separable Hilbert space with the effective generating set* $\mathcal{S}$ *consisting of linearly independent elements. Let* B *and* F *be as in Theorem* 1. *Then the variational problem* (2.2) *has a uniquely determined solution* $u \in \mathbf{X}$ *which is computable. The Galerkin approximations* u_k *of* u *form a computable sequence in* $\mathbf{X}$*, which effectively converges to* u.

Remark 5. The proof of Theorem 1 in [1] is actually done in the context of the TTE or Type 2 Theory of Effectivity (see [6, pp.2–10], in particular). The TTE provides a more differentiated formulation based on a variant of the Turing machine (Type 2 machine), and is particularly valid in handling several different computability notions at the same time and thus in the discussions of computational complexity. However, as for the computability issue discussed in the present note, the TTE and the Pour-El & Richards approach [5] are equivalent since both rely on the same dense set $\mathcal{D}$ (see §3) of $\mathbf{X}$ as the reference to computability. We here adopt the Pour-El & Richards approach, which provides a formulation closer to the way in classical analysis.

References

[1] V. Brattka and A. Yoshikawa, Towards computability of elliptic boundary value problems in variational formulation, J. Complexity **22** (2006), no. 6, 858–880.

[2] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, Springer, New York, 1994.

[3] J. Céa, Approximation variationnelle des problèmes aux limites, Ann. Inst. Fourier (Grenoble) **14** (1964), fasc. 2, 345–444.

[4] P. D. Lax and A. N. Milgram, Parabolic equations, in *Contributions to the theory of partial differential equations*, 167–190, Ann. of Math. Stud., 33, Princeton Univ. Press, Princeton, N. J., 1954.

[5] M. B. Pour-El and J. I. Richards, *Computability in analysis and physics*, Springer, Berlin, 1989.

[6] K. Weihrauch, *Computable analysis*, Springer, Berlin, 2000.

[7] A. Yoshikawa, On an ad hoc computability structure in a Hilbert space, Proc. Japan Acad. Ser. A Math. Sci. **79** (2003), no. 3, 65–70.