

Research Article

Error Analysis of Galerkin's Method for Semilinear Equations

Tadashi Kawanago

*Department of Mathematics, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku,
Tokyo 152-8551, Japan*

Correspondence should be addressed to Tadashi Kawanago, tadashi@math.titech.ac.jp

Received 7 January 2012; Revised 16 May 2012; Accepted 8 June 2012

Academic Editor: Hui-Shen Shen

Copyright © 2012 Tadashi Kawanago. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We establish a general existence result for Galerkin's approximate solutions of abstract semilinear equations and conduct an error analysis. Our results may be regarded as some extension of a precedent work (Schultz 1969). The derivation of our results is, however, different from the discussion in his paper and is essentially based on the convergence theorem of Newton's method and some techniques for deriving it. Some of our results may be applicable for investigating the quality of numerical verification methods for solutions of ordinary and partial differential equations.

1. Introduction

Let X be a real Hilbert space and $X_h \subset X$ be a closed subspace. Here, h is a positive parameter (which will tend to zero). We denote by P_h the orthogonal projection onto X_h . We assume that

$$\lim_{h \rightarrow 0} \|(I - P_h)u\| = 0 \quad \text{for any } u \in X, \quad (\text{H1})$$

where $I : X \rightarrow X$ is the identity operator. We are interested in studying error analysis of Galerkin's method for the following equation:

$$f(u) := u - \varphi(u) = 0. \quad (1.1)$$

Here, $\varphi : U \rightarrow X$ is a nonlinear map and U is a subset of X . We define $\tilde{f}_h : U \cap X_h \rightarrow X_h$ by

$$\tilde{f}_h(u) := P_h f(u) \quad \text{for } u \in U \cap X_h. \quad (1.2)$$

The equation $\tilde{f}_h(u) = 0$ is the Galerkin approximate equation of (1.1). A precedent work [1] by Schultz reads as follows.

Theorem 1.1 (see [1, Theorems 3.1 and 3.2]). *One assumes (H1). Let $R \in (0, \infty)$ be a constant and $U = \{u \in X; \|u\| \leq R\}$. One assumes that $\varphi : U \rightarrow X$ is a completely continuous map such that $\varphi(U) \subset U$. Then the following holds.*

- (i) *The equation $\tilde{f}_h(u) = 0$ has a solution u_h in $U \cap X_h$ for any h and there exists a monotone decreasing sequence $\{h_k\}_{k=1}^{\infty}$ with $\lim_{k \rightarrow \infty} h_k = 0$ and $u_{\infty} \in U$ such that $u_{h_k} \rightarrow u_{\infty}$ in X as $k \rightarrow \infty$ and u_{∞} is a solution of (1.1). Moreover, if u_{∞} is the unique solution of (1.1) in U , then one has $\lim_{h \rightarrow 0} u_h = u_{\infty}$ in X .*
- (ii) *Let $u_* \in U$ be a solution of (1.1). If φ has a Fréchet derivative, in a neighborhood, \mathcal{N} , of u_* and 0 is not in the spectrum of f' , then u_* is the unique solution of (1.1) in \mathcal{N} and $\tilde{f}_h(u) = 0$ has a solution $u_h \in X_h$ for any h , which is unique for sufficiently small h and*

$$\|u_* - u_h\| \simeq \|(I - P_h)u_*\|, \quad (1.3)$$

which means that $\|u_ - u_h\|$ and $\|(I - P_h)u_*\|$ are equivalent infinitesimals as $h \rightarrow 0$.*

In this paper, we always assume (H1) and the following (H2) in what follows:

$$\varphi \in C^1(U, X), \quad \text{and } \varphi'(u) \in \mathcal{L}(X) \text{ is compact for any } u \in U, \quad (H2)$$

where $U \subset X$ is an open set. Under the conditions (H1) and (H2), we obtain results similar to Theorem 1.1 (see Proposition 2.1 and Corollary 2.3). We also establish new other results on error analysis (see Theorems 2.4 and 2.5). Our results may be regarded as some extension of Theorem 1.1. The derivation of our results is, however, different from the proof of Theorem 1.1, which is based on the Brouwer fixed point theorem and the equality (4.10). Our proofs are essentially based on the convergence theorem of Newton's method (Theorem 3.2) and some techniques for deriving it. We remark that a version of the same theorem is applied in [2] to an ordinary periodic system for a purpose similar to ours.

Various ordinary and partial differential equations appearing in mathematical physics can be written in the form (1.1) with (H2) under an appropriate setting of the functional spaces. See Section 5 for some concrete examples.

We define $f_h : U \rightarrow X$ by

$$f_h(u) := u - P_h \varphi(u) \quad \text{for } u \in U. \quad (1.4)$$

The map f_h is a natural extension of \tilde{f}_h and is very useful in our analysis below. Obviously, u is a solution of $\tilde{f}_h(u) = 0$ if and only if u is a solution of $f_h(u) = 0$. We can treat the equation $f_h(u) = 0$ more easily than $\tilde{f}_h(u) = 0$ since f_h is defined globally.

One of our motivations for this study is to investigate the quality of a numerical verification method for solutions of differential equations. Some of our results in this paper may be applicable for such a purpose. See Remark 2.7 for further information.

The paper is organized as follows. In Section 2 we describe our main results. We prepare some preliminary abstract results in Section 3 and apply them to prove our main results in Section 4. In Section 5 we present some concrete examples on semilinear elliptic partial differential equations.

Notations. Let \mathcal{X} and \mathcal{Y} be Banach spaces.

- (1) We denote by $\|\cdot\|_{\mathcal{X}}$ the norm of \mathcal{X} . If \mathcal{X} is a Hilbert space, then $\|\cdot\|_{\mathcal{X}}$ stands for the norm induced by the inner product of \mathcal{X} . For $u \in \mathcal{X}$ and $r \in (0, \infty)$, we write $B_{\mathcal{X}}(u; r) := \{v \in \mathcal{X}; \|v - u\| < r\}$. The subscript will be often omitted if no possible confusion arises.
- (2) For an open set $V \subset \mathcal{X}$, $C^1(V, \mathcal{Y})$ denotes the space of continuously differentiable functions from V to \mathcal{Y} .
- (3) We denote by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ the space of bounded linear operators from \mathcal{X} to \mathcal{Y} and $\mathcal{L}(\mathcal{X})$ stands for $\mathcal{L}(\mathcal{X}, \mathcal{X})$. For $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $\|T\|_{\mathcal{X} \rightarrow \mathcal{Y}}$ denotes the operator norm of T . The subscript will be omitted if no possible confusion arises.
- (4) Let $\phi(h)$ and $\psi(h)$ be nonnegative functions. We write $\phi(h) \sim \psi(h)$ if $\phi(h)$ and $\psi(h)$ are infinitesimals of the same order as $h \rightarrow 0$, that is, $\phi(h) = O(1)\psi(h)$ and $\psi(h) = O(1)\phi(h)$ as $h \rightarrow 0$. We write $\phi(h) \simeq \psi(h)$ if $\phi(h)$ and $\psi(h)$ are equivalent infinitesimals as $h \rightarrow 0$, that is, $\phi(h) = \{1 + o(1)\}\psi(h)$ as $h \rightarrow 0$.
- (5) Let Ω be a bounded domain of \mathbf{R}^n . We denote Lebesgue spaces by $L^p(\Omega)$ ($1 \leq p \leq \infty$) with the norms $\|u\|_{L^p(\Omega)} = (\int_{\Omega} |u(x)|^p dx)^{1/p}$ for $1 \leq p < \infty$, $\|u\|_{L^\infty(\Omega)} = \text{ess. sup}\{|u(x)|; x \in \Omega\}$. We denote by $H_0^1(\Omega)$ the completion of $C_0^\infty(\Omega)$ (the space of C^∞ functions with compact support in Ω) in the Sobolev norm: $\|u\| = \|\nabla u\|_{L^2(\Omega)} := (\sum_{k=1}^n \|\partial u / \partial x_k\|_{L^2(\Omega)}^2)^{1/2}$. We denote by $H^{-1}(\Omega)$ the Sobolev space $\{F \in \mathfrak{D}'(\Omega); \exists C \in (0, \infty) \text{ such that } |F(\phi)| \leq C\|\phi\|_{H_0^1(\Omega)} \text{ for any } \phi \in C_0^\infty(\Omega)\}$ with the norm $\|F\|_{H^{-1}(\Omega)} = \sup\{|F(\phi)|; \phi \in C_0^\infty(\Omega) \text{ and } \|\phi\|_{H_0^1(\Omega)} \leq 1\}$. Here, $\mathfrak{D}'(\Omega)$ stands for the set of distributions on Ω .

2. Main Results

In this section we describe our main results. We assume (H1) and (H2). Let $u_* \in U$ be an isolated solution of (1.1), that is, u_* is a solution of (1.1) such that $f'(u_*) : X \rightarrow X$ is bijective. We set

$$T := \varphi'(u_*), \quad A := I - T = f'(u_*), \quad A_h := I - P_h T P_h, \quad (2.1)$$

for simplicity. The operator A_h is an almost diagonal operator introduced in [3]. First we have an existence theorem for Galerkin's approximate solutions of (1.1).

Proposition 2.1. *There exist $h_* > 0$ and $\{u_h\}_{h \in (0, h_*)} \subset U$ such that the following (i)–(iii) hold.*

- (i) *There exists $R_* > 0$ such that $u = u_h$ is the only solution of $f_h(u) = 0$ in $B(u_*; R_*)$ for any $h \in (0, h_*)$.*

(ii) $u = u_h$ is an isolated solution of $f_h(u) = 0$ for any $h \in (0, h_*)$.

(iii) $u_h \rightarrow u_*$ in X as $h \rightarrow 0$ with the estimate

$$\|u_* - u_h\| \leq C_h \|f_h(u_*)\| = C_h \|(I - P_h)u_*\| \quad \text{for any } h \in (0, h_*), \quad (2.2)$$

where $\{C_h\}_{h \in (0, h_*)} \subset (1, 2)$ and $C_h \rightarrow 1$ as $h \rightarrow 0$.

Remark 2.2. (i) Proposition 2.1(ii) is useful in our analysis below. Moreover, we immediately obtain from it that $u = u_h$ is an isolated solution of $\tilde{f}_h(u) = 0$ for any $h \in (0, h_*)$. This guarantees that we can always construct a Galerkin approximate solution u_h by Newton's method for small $h > 0$.

(ii) In various contexts in applications, X_h is finite-dimensional for any h . In such contexts the assumption (H1) implies that X is separable.

(iii) We do not assume $\dim X_h < \infty$. We briefly explain that it has some practical benefits. The case $\dim X_h = \infty$ appears, for example, in the following context. We are interested in the semi-discrete approximation to a periodic system described by a partial differential equation with a periodic forcing term. We may apply a Galerkin method only in space to the original system in order to construct a simpler approximate system described by ordinary differential equations. Then, for an isolated periodic solution of the original system, our Proposition 2.1 may guarantee that in a small neighborhood of it the approximate system has a periodic solution. For example, we can actually apply Proposition 2.1 to a semi-discrete approximation to a periodic system treated in [3]. See [4, Remark 3.4] for how to rewrite the system in [3] as (1.1).

In what follows in this section, $\{u_h\}_{h \in (0, h_*)}$ always denotes the sequence as described in Proposition 2.1. Since $u_* - u_h$ is decomposed into the X_h -component $P_h u_* - u_h$ and the X_h^\perp -component $(I - P_h)u_*$, we have $\|u_* - u_h\|^2 = \|P_h u_* - u_h\|^2 + \|(I - P_h)u_*\|^2$ and $\|(I - P_h)u_*\| \leq \|u_* - u_h\|$. So, the last inequality and (2.2) immediately imply (2.3) below.

Corollary 2.3. *We have*

$$\|u_* - u_h\| \simeq \|f_h(u_*)\| = \|(I - P_h)u_*\|, \quad (2.3)$$

$$\|P_h u_* - u_h\| = o(\|(I - P_h)u_*\|) \quad \text{as } h \rightarrow 0, \quad (2.4)$$

$$\|P_h u_* - u_h\| = o(\|u_* - u_h\|) \quad \text{as } h \rightarrow 0. \quad (2.5)$$

Actually, we easily verify that (2.3), (2.4) and (2.5) are mutually equivalent. They are very general features for the Galerkin method. The estimate (2.5) means that the X_h -component of the error $\|P_h u_* - u_h\|$ is an infinitesimal of a higher order of smallness with respect to the whole error $\|u_* - u_h\|$ as $h \rightarrow 0$.

The following two results are useful for applications (see Remark 2.7 below).

Theorem 2.4. *We have the following:*

$$\|u_* - u_h\| \simeq \|f(u_h)\| = \|(I - P_h)\varphi(u_h)\| \simeq \|A^{-1}f(u_h)\|, \quad (2.6)$$

$$\|P_h u_* - u_h\| \simeq \|A^{-1}f_h(P_h u_*)\| \simeq \|P_h A_h^{-1}f(P_h u_*)\| \sim \|P_h\{\varphi(u_*) - \varphi(P_h u_*)\}\|. \quad (2.7)$$

Theorem 2.5. (i) *We have*

$$\sup_{s \in [0,1]} \|\varphi'((1-s)u_* + sP_h u_*)(I - P_h)\| \longrightarrow 0 \quad \text{as } h \longrightarrow 0. \quad (2.8)$$

(ii) *Let ε_h be a positive constant for $h \in (0, h_*)$ such that*

$$\varepsilon_h \geq \sup_{s \in [0,1]} \|\varphi'((1-s)u_* + sP_h u_*)(I - P_h)\| \quad (2.9)$$

for any $h \in (0, h_)$. Then, there exist constants $h_1 \in (0, h_*)$ and $C_1 > 0$ such that*

$$\|P_h u_* - u_h\| \leq C_1 \varepsilon_h \|(I - P_h)u_*\| \quad \text{for any } h \in (0, h_1). \quad (2.10)$$

In view of Theorem 2.5 (i) and (ii), we can always take $\{\varepsilon_h\}_{h \in (0, h_*)}$ in (2.10) such that $\varepsilon_h \rightarrow 0$ as $h \rightarrow 0$. The following Remarks 2.6 and 5.3 below shows that our estimate (2.10) is in general sharper than an estimate which can be derived directly from the discussion in [1].

Remark 2.6. (i) In the same way as in the proof of [1, Theorem 3.2] we can obtain an estimate related to (2.10). We set $\eta_h := (2p_h + q_h + r_h)/(1 - (p_h + q_h + r_h))$, $p_h := \|A^{-1}(I - P_h)T\|$, $q_h := \|A^{-1}P_h T(I - P_h)\|$ and $r_h := \|A^{-1}\| \cdot \|\varphi(u_h) - \varphi(u_*) - T(u_h - u_*)\|/\|u_h - u_*\|$. It follows from Proposition 2.1 (iii) and Proposition 3.1 below that p_h , q_h and r_h converge to 0 as $h \rightarrow 0$. So, $\eta_h \rightarrow 0$ as $h \rightarrow 0$. Let $\hat{\varepsilon}_h$ be a positive constant for $h \in (0, h_*)$ such that $\hat{\varepsilon}_h \geq \sqrt{\eta_h^2 + 2\eta_h}$. Then we have

$$\|P_h u_* - u_h\| \leq \hat{\varepsilon}_h \|(I - P_h)u_*\| \quad \text{for any } h \in (0, h_*). \quad (2.11)$$

We can verify that

$$\hat{\varepsilon}_h \text{ is larger than } \|(I - P_h)T\|^{1/2} \text{ for sufficiently small } h > 0. \quad (2.12)$$

Indeed, we immediately obtain (2.12) from

$$p_h \simeq \|(I - P_h)T\|. \quad (2.13)$$

We derive (2.11) and (2.13) at the end of Section 4.

(ii) When we compute $\hat{\varepsilon}_h$ for concrete examples (e.g., examples in Section 5 below), it seems reasonable to estimate q_h as $q_h \leq C\|T(I - P_h)\|$. Here, C represents some positive

constant independent of h . Then, it is actually necessary to take $\hat{\varepsilon}_h$ such that $\hat{\varepsilon}_h \geq C\|T(I - P_h)\|^{1/2}$ for small $h > 0$. On the other hand, roughly speaking, (2.9) means that we can take $\varepsilon_h \approx \|T(I - P_h)\|$ for small $h > 0$ (See Remark 5.3 below). (We note that Proposition 3.1 below implies that $\|T(I - P_h)\| \rightarrow 0$ as $h \rightarrow 0$.)

(iii) We consider the case where T is self-adjoint (e.g., Example 5.1 below). In this case, we have $\|(I - P_h)T\| = \|T(I - P_h)\|$. So, by (2.12) $\hat{\varepsilon}_h$ is larger than $\|T(I - P_h)\|^{1/2}$ for small $h > 0$.

Remark 2.7. We mention applications of our results. Some of our results may be applicable for testing the quality of a numerical verification algorithm for solutions of differential equations. In general we obtain an upper bound of $\|u_* - u_h\|$ as output data from a numerical verification algorithm (See e.g., [5] and the references therein). By our Theorem 2.4 $\|u_* - u_h\|$ is sufficiently close to $\|f(u_h)\|$ for sufficiently small h . So, Theorem 2.4 shows that we can check the accuracy of the output upper bound of $\|u_* - u_h\|$ by finding the value of $\|f(u_h)\|$ when h is small. In [5] we proposed a numerical verification algorithm which also gives upper bounds of $\|P_h u_* - u_h\|$ as output data. Our Theorem 2.5 may be applicable for testing the accuracy of such upper bounds. See Remark 5.4 for more detailed information.

3. Preliminary Abstract Results

In this section, we prepare some abstract results in order to prove our main results in Section 2.

Proposition 3.1. *We assume (H1). Let $K : X \rightarrow X$ be a compact operator. Then we have the following:*

$$P_h K \rightarrow K, \quad K P_h \rightarrow K \quad \text{in } \mathcal{L}(X) \text{ as } h \rightarrow 0, \quad (3.1)$$

$$P_h K P_h \rightarrow K \quad \text{in } \mathcal{L}(X) \text{ as } h \rightarrow 0. \quad (3.2)$$

Proof. Though this result was proved in [6, Section 78], we give a simpler proof for the convenience of the reader. First we show that

$$\|K(I - P_h)\| \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad (3.3)$$

We proceed by contradiction. We assume that (3.3) does not hold. Then we have $\delta := \limsup_{h \rightarrow 0} \|K(I - P_h)\| > 0$. Therefore, there exist $\{h_n\}_{n=1}^{\infty}$ and $\{u_n\}_{n=1}^{\infty} \subset X$ such that $h_n \searrow 0$ as $n \rightarrow \infty$, $\|u_n\| = 1$ for $n \in \mathbf{N}$ and

$$\|K(I - P_{h_n})u_n\| \geq \frac{\delta}{2} \quad \text{for any } n \in \mathbf{N}. \quad (3.4)$$

Since K is compact and $(I - P_{h_n})u_n$ converges weakly to 0, we have $\|K(I - P_{h_n})u_n\| \rightarrow 0$ as $n \rightarrow \infty$. This contradicts (3.4). So, (3.3) holds. Since K^* is also compact, we obtain

$$\|(I - P_h)K\| = \|\{(I - P_h)K\}^*\| = \|K^*(I - P_h)\| \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad (3.5)$$

So, we have (3.1), which implies (3.2). \square

Next, we describe some results in a more general setting. In what follows in this section, let \mathcal{X} and \mathcal{Y} be Banach spaces and $U \subset \mathcal{X}$ be an open set. We assume $F \in C^1(U, \mathcal{Y})$.

Theorem 3.2. *Let $u_0 \in U$ and $L \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ be bijective. We define a map $g : U \rightarrow \mathcal{X}$ by*

$$g(u) := u - L^{-1}F(u). \quad (3.6)$$

Let $R > 0$ be a constant satisfying $\overline{B(u_0; R)} \subset U$ and $b : [0, R] \rightarrow [0, \infty)$ be a non-decreasing function such that

$$\sup \left\{ \|g'(u)\|; u \in \overline{B(u_0; r)} \right\} \leq b(r) \quad \text{for any } r \in [0, R]. \quad (3.7)$$

Let $\varepsilon_0 \geq 0$ be a constant such that

$$\|L^{-1}F(u_0)\| \leq \varepsilon_0. \quad (3.8)$$

We assume that there exist constants r_0 and r_1 such that $0 < r_0 \leq r_1 \leq R$,

$$\varepsilon_0 + \int_0^{r_0} b(r) dr \leq r_0, \quad (3.9)$$

$$b(r_1) < 1. \quad (3.10)$$

Then the equation $F(u) = 0$ has an isolated solution $u_* \in \overline{B(u_0; r_0)}$. Moreover, the solution of $F(u) = 0$ is unique in $\overline{B(u_0; r_1)}$.

Remark 3.3. (i) Theorem 3.2 is a new version of the convergence theorem of simplified Newton's method, which is a refinement of the classical versions such as [5, Theorem 0.1]. Actually, the former implies the latter.

(ii) The convergence theorem of simplified Newton's method is a very strong and general principle to verify the existence of isolated solutions. The reason is, roughly speaking, that the condition of the theorem is not only a sufficient condition to guarantee an isolated solution but also virtually a *necessary* condition for an isolated solution to exist. See [4, Remark 1.3] for the detail.

Proof of Theorem 3.2. Though we may consider Theorem 3.2 as a corollary of [5, Theorem 1.1], we describe the proof for completeness. We easily verify that u is a solution of $F(u) = 0$ if and only if u is a fixed point of $g(u)$. Let $u, v \in U$. We obtain

$$g(u) - g(v) = \int_0^1 \frac{d}{dt} g(v + t(u - v)) dt = \int_0^1 g'(v + t(u - v)) dt (u - v). \quad (3.11)$$

By (3.7) and (3.11) we have

$$\|g(u) - g(v)\| \leq b(r) \|u - v\| \quad \text{for any } r \in (0, R] \text{ and } u, v \in \overline{B(u_0; r)}. \quad (3.12)$$

We set $B_0 := \overline{B(u_0; r_0)}$. Let $u \in B_0$. In view of (3.7), (3.8), and (3.11) with $v := u_0$, we have

$$\|g(u_0) - u_0\| = \|L^{-1}F(u_0)\| \leq \varepsilon_0, \quad (3.13)$$

$$\|g(u) - g(u_0)\| \leq r_0 \int_0^1 b(r_0 t) dt = \int_0^{r_0} b(r) dr. \quad (3.14)$$

Combining (3.9), (3.13), and (3.14), we have $\|g(u) - u_0\| \leq r_0$, which implies $g(B_0) \subset B_0$. Therefore, in view of (3.10) and (3.12) g is a contraction on B_0 . By the contraction mapping principle there exists a unique solution $u = u_*$ on B_0 for the equation $F(u) = 0$. We immediately obtain from (3.10) and (3.12) that the solution of $F(u) = 0$ is unique on $\overline{B(u_0; r_1)}$. Finally, it suffices to show that

$$F'(u) : \mathcal{X} \longrightarrow \mathcal{Y} \text{ is bijective for any } u \in \overline{B(u_0; r_1)} \quad (3.15)$$

in order to prove that u_* is isolated. We denote by I the identity operator on \mathcal{X} . Let $u \in \overline{B(u_0; r_1)}$. Then, by (3.7) and (3.10) we have $\|g'(u)\| \leq b(r_1) < 1$. This implies that $I - g'(u) : \mathcal{X} \rightarrow \mathcal{X}$ is bijective. Since L is also bijective and $F'(u) = L\{I - g'(u)\}$, (3.15) holds. \square

The next result may be considered as a refinement of [7, Theorem 3.1 (3.14)] and [8, Theorem 3.1 (3.23)].

Proposition 3.4. *Let $u, v \in U$, $(1-s)u + sv \in U$ for any $s \in (0, 1)$ and $L \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ be bijective. We set $m := \max_{s \in [0, 1]} \|L - F'((1-s)u + sv)\|$. Then we have*

$$\frac{\|L^{-1}\{F(u) - F(v)\}\|}{1 + m\|L^{-1}\|} \leq \|u - v\|. \quad (3.16)$$

Moreover, if $m\|L^{-1}\| < 1$ then we also obtain

$$\|u - v\| \leq \frac{\|L^{-1}\{F(u) - F(v)\}\|}{1 - m\|L^{-1}\|}. \quad (3.17)$$

Proof. The proof is similar to that of Theorem 3.2. Let $g : U \rightarrow \mathcal{X}$ be a map defined by (3.6). We have

$$u - v = g(u) - g(v) + L^{-1}\{F(u) - F(v)\}. \quad (3.18)$$

It follows from (3.11) that $\|g(u) - g(v)\| \leq m\|L^{-1}\|\|u - v\|$. Combining this inequality and (3.18), we obtain (3.16) and (3.17). \square

Theorem 3.5. Let $u = u_* \in U$ be an isolated solution of the equation $F(u) = 0$. Let $h_0 > 0$ be a positive constant, $F_h \in C^1(U, \mathcal{Y})$ and $H_h \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ ($0 < h < h_0$). We set $H := F'(u_*)$. We assume that

$$F_h(u_*) \longrightarrow 0 \quad \text{in } \mathcal{Y} \text{ as } h \longrightarrow 0, \quad (3.19)$$

$$F'_h(u_*) \longrightarrow H, \quad H_h \longrightarrow H \quad \text{in } \mathcal{L}(\mathcal{X}, \mathcal{Y}) \text{ as } h \longrightarrow 0, \quad (3.20)$$

$$\lim_{r \searrow 0} d(r, u) = 0 \quad \text{for any } u \in U. \quad (3.21)$$

Here, $d(r, u) := \sup\{\|F'_h(u) - F'_h(v)\|; 0 < h < h_0 \text{ and } v \in U \cap \overline{B(u; r)}\}$. Then, there exist a constant $h_* \in (0, h_0)$ and sequences $\{c_h\}_{h \in (0, h_*)} \subset (1, 2)$, $\{u_h\}_{h \in (0, h_*)} \subset U$ such that the following (a)–(f) hold:

(a)

$$c_h \longrightarrow 1 \quad \text{as } h \longrightarrow 0, \quad (3.22)$$

(b) $u = u_h$ is an isolated solution of $F_h(u) = 0$ for any $h \in (0, h_*)$,

(c)

$$\|u_h - u_*\|_{\mathcal{X}} \leq c_h \|H_h^{-1} F_h(u_*)\|_{\mathcal{X}} \quad \text{for any } h \in (0, h_*), \quad (3.23)$$

(d) H_h is bijective with $\|H_h - F'_h(u_*)\| < 1/2 \|H_h^{-1}\|$ and $\|H_h^{-1}\| < 2 \|H^{-1}\|$ for any $h \in (0, h_*)$,

(e) the solution of $F_h(u) = 0$ is unique in $B(u_*; R_h)$ for any $h \in (0, h_*)$, where

$$R_h := \sup \left\{ R > 0; \overline{B(u_*; R)} \subset U, d(R, u_*) < \frac{1}{\|H_h^{-1}\|} - \|H_h - F'_h(u_*)\| \right\} > 0, \quad (3.24)$$

(f)

$$c_h \|H_h^{-1} F_h(u_*)\| < R_h \quad \text{for any } h \in (0, h_*). \quad (3.25)$$

Proof. By (3.20) and the stability property of linear operators (e.g., [3, Corollary 2.4.1]), $F'_h(u_*)$ and H_h are bijective for sufficiently small $h > 0$ and $F'_h(u_*)^{-1} \rightarrow H^{-1}$, $H_h^{-1} \rightarrow H^{-1}$ in $\mathcal{L}(\mathcal{Y}, \mathcal{X})$ as $h \rightarrow 0$. Let $\eta_h := \|H_h^{-1} F_h(u_*)\|$ and $g_h(u) := u - H_h^{-1} F_h(u)$. We set $d(r) := d(r, u_*)$ for $r > 0$ and define $b_h(r) := \|H_h^{-1}\| \{\|H_h - F'_h(u_*)\| + d(r)\}$. Let $c_h := 1/\{1 - b_h(2\eta_h)\}$, $r_h := c_h \eta_h$ and $\delta_h := (1/2 \|H_h^{-1}\|) - \|H_h - F'_h(u_*)\|$. Then, we easily verify that as $h \rightarrow 0$,

$$\eta_h \longrightarrow 0, \quad c_h \longrightarrow 1, \quad r_h \longrightarrow 0, \quad \delta_h \longrightarrow \frac{1}{2 \|H^{-1}\|} (> 0). \quad (3.26)$$

Therefore, there exist $h_* \in (0, h_0)$ and $\varepsilon \in (0, 1)$ such that for any $h \in (0, h_*)$, $F'_h(u_*)$ is bijective with (d), $1 \leq c_h < 2$, $d(r_h + \varepsilon) < \delta_h$ and $\overline{B(u_*; r_h + \varepsilon)} \subset U$. It follows that $\|g'_h(u)\| \leq b_h(r)$ for any $h \in (0, h_*)$, $r > 0$, and $u \in U \cap \overline{B(u_*; r)}$. We also have $r_h + \varepsilon \leq R_h$ and

$$\eta_h + \int_0^{r_h} b_h(r) dr \leq \eta_h + r_h b_h(r_h) \leq r_h. \quad (3.27)$$

Let $h \in (0, h_*)$ and $R \in (r_h, R_h)$. We apply Theorem 3.2 by setting $F := F_h$, $u_0 := u_*$, $L := H_h$, $b := b_h$, $r_0 := r_h$, $r_1 := R$ and $\varepsilon_0 := \eta_h$. Then, we obtain the desired conclusions. \square

Remark 3.6. Theorem 3.5 is related to [7, Theorem 3.1] and [8, Theorem 3.1]. Actually, their proofs are similar to ours. Our proof is based on the convergence theorem of simplified Newton's method, from which they may be derived similarly.

4. Proofs of Main Theorems

We prove the results in Section 2. We use the notation (2.1).

Proof of Proposition 2.1. We apply Theorem 3.5 by putting $\mathcal{X} = \mathcal{Y} := X$, $F := f$, $F_h := f_h$, $H := A$ and $H_h := A_h$. We show (3.19)–(3.21). By (H1) we have $f_h(u_*) = (I - P_h)u_* \rightarrow 0$ in X as $h \rightarrow 0$. Therefore, (3.19) holds. It follows from (H2) and Proposition 3.1 that

$$f'_h(u_*) = I - P_h T \rightarrow I - T = A, \quad A_h \rightarrow A \quad \text{in } \mathcal{L}(X) \text{ as } h \rightarrow 0. \quad (4.1)$$

So, (3.20) holds. Let $r > 0$, $u \in U$ and $v \in U \cap \overline{B(u; r)}$. Since φ' is continuous, we have $\|f'_h(u) - f'_h(v)\| \leq \|\varphi'(u) - \varphi'(v)\| \rightarrow 0$ as $r \searrow 0$, which implies (3.21). Therefore, by Theorem 3.5, there exist a small constant $h_* > 0$, $\{u_h\}_{h \in (0, h_*)}$ and $\{C_h\}_{h \in (0, h_*)}$ such that (a)–(f) with $c_h := C_h$ hold. So, we immediately obtain (ii) and $u_h \rightarrow u_*$ in X as $h \rightarrow 0$. Since $A_h^{-1} f_h(u_*) = f_h(u_*) = (I - P_h)u_*$, (a) and (c) imply (2.2). So, (iii) holds. In view of (d) and (e), we have (i) with

$$R_* := \sup \left\{ R > 0; \overline{B(u_*; R)} \subset U \text{ and } \hat{d}(R) < \frac{1}{4\|A^{-1}\|} \right\}, \quad (4.2)$$

where $\hat{d}(R) := \sup \{ \|\varphi'(v) - \varphi'(u_*)\|; v \in U \cap \overline{B(u_*; R)} \}$. The proof is complete. \square

Proof of Theorem 2.4. We set $u(s, h) := (1-s)u_h + su_*$ for simplicity. Proposition 2.1 (iii) implies $\max_{s \in [0, 1]} \|u_* - u(s, h)\| = \|u_* - u_h\| \rightarrow 0$ as $h \rightarrow 0$. First we show (2.6). We have $f(u_h) = -(I - P_h)\varphi(u_h) = (I - P_h)f(u_h)$, $A_h^{-1} \rightarrow A^{-1}$ in $\mathcal{L}(X)$ as $h \rightarrow 0$ and

$$\begin{aligned} m_h &:= \max_{s \in [0, 1]} \|A_h - f'(u(s, h))\| \\ &\leq \|A_h - A\| + \max_{s \in [0, 1]} \|f'(u_*) - f'(u(s, h))\| \rightarrow 0 \quad \text{as } h \rightarrow 0. \end{aligned} \quad (4.3)$$

Since $A_h f(u_h) = f(u_h)$, we have $A_h^{-1} f(u_h) = f(u_h)$. We apply Proposition 3.4 with $L := A_h$, $u := u_h$, $v := u_*$ and $F := f$ to obtain

$$\frac{\|f(u_h)\|}{1 + m_h \|A_h^{-1}\|} \leq \|u_h - u_*\| \leq \frac{\|f(u_h)\|}{1 - m_h \|A_h^{-1}\|} \quad \text{for small } h > 0, \quad (4.4)$$

which implies $\|u_h - u_*\| \simeq \|f(u_h)\|$. We also have $\|u_h - u_*\| \simeq \|A^{-1} f(u_h)\|$ by the above discussion with $L := A_h$ replaced by $L := A$. Next, we show (2.7). In the same way as above we apply Proposition 3.4 with $L := A$ (resp., $L := A_h$), $u := u_h$, $v := P_h u_*$ and $F := f_h$ to have

$$\|P_h u_* - u_h\| \simeq \left\| A^{-1} f_h(P_h u_*) \right\| \quad \left(\text{resp., } \|P_h u_* - u_h\| \simeq \left\| A_h^{-1} f_h(P_h u_*) \right\| \right). \quad (4.5)$$

Since $f_h(P_h u_*) = P_h f(P_h u_*)$ and A_h commutes with P_h , we have $\|P_h u_* - u_h\| \simeq \|P_h A_h^{-1} f(P_h u_*)\|$. Combining (4.5) and $f_h(P_h u_*) = P_h \{\varphi(u_*) - \varphi(P_h u_*)\}$, we obtain $\|P_h u_* - u_h\| \sim \|P_h \{\varphi(u_*) - \varphi(P_h u_*)\}\|$. \square

Proof of Theorem 2.5. We set $u_*(s, h) := (1 - s)u_* + sP_h u_*$ for simplicity.

(i) It follows from (H2) and Proposition 3.1 that

$$\|\varphi'(u_*)(I - P_h)\| \longrightarrow 0. \quad (4.6)$$

By (H1) and the continuity of $\varphi'(u)$ at $u = u_*$ we have

$$\sup_{s \in [0,1]} \|\varphi'(u_*) - \varphi'(u_*(s, h))\| \longrightarrow 0 \quad \text{as } h \longrightarrow 0. \quad (4.7)$$

We obtain (2.8) from (4.6) and (4.7).

(ii) In the same way as (3.11) we have

$$\varphi(u_*) - \varphi(P_h u_*) = \int_0^1 \varphi'(u_*(s, h)) ds (I - P_h)u_*. \quad (4.8)$$

By this equality, (2.7) and (2.9), we have (2.10). \square

Finally we derive (2.11) and (2.12).

Proof of (2.11) and (2.13). Without loss of generality we assume $\hat{\varepsilon}_h = \sqrt{\eta_h^2 + 2\eta_h}$. First we derive (2.11). This proof is essentially the same as that of [1, Theorem 3.2]. It suffices to prove

$$\|u_* - u_h\| \leq (1 + \eta_h) \|(I - P_h)u_*\|, \quad (4.9)$$

which implies (2.11) in view of $\|u_* - u_h\|^2 = \|P_h u_* - u_h\|^2 + \|(I - P_h)u_*\|^2$. We have

$$\begin{aligned} & A(u_h - u_*) + (I - P_h)T(u_h - u_*) - P_h T(I - P_h)(u_h - u_*) \\ & \quad - P_h \{ \varphi(u_h) - \varphi(u_*) - T(u_h - u_*) \} \\ & = -A(I - P_h)u_* - (I - P_h)T(I - P_h)u_* \end{aligned} \quad (4.10)$$

It follows that

$$\|u_h - u_*\| \leq (p_h + q_h + r_h)\|u_h - u_*\| + (1 + p_h)\|(I - P_h)u_*\|, \quad (4.11)$$

which implies (4.9). Next we derive (2.13). Since $A^{-1} = I + K$ with $K := T(I - T)^{-1}$, we obtain from Proposition 3.1 that

$$\begin{aligned} \left\| A^{-1}(I - P_h)T - (I - P_h)T \right\| &= \|K(I - P_h)T\| \leq \|K(I - P_h)\| \|(I - P_h)T\| \\ &= o(1)\|(I - P_h)T\| \quad \text{as } h \rightarrow 0. \end{aligned} \quad (4.12)$$

So, (2.13) holds. \square

5. Concrete Examples

In this section we consider the following semilinear elliptic boundary value problem:

$$-\Delta u = G(x, u, \nabla u) \quad \text{in } \Omega \text{ with } u = 0 \text{ on } \partial\Omega, \quad (5.1)$$

where Ω is a bounded convex domain in \mathbf{R}^N ($N \leq 3$) with piecewise smooth boundary $\partial\Omega$. We will rewrite (5.1) as the form (1.1) under the appropriate setting of functional spaces. We simply denote $G(u) := G(\cdot, u, \nabla u)$. We assume $G \in C^1(H_0^1(\Omega), H^{-1}(\Omega))$. Let $L \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ be the operator defined by $Lu := -\Delta u$. We set $X := H_0^1(\Omega)$ with the norm $\|u\|_X := \|\nabla u\|_{L^2(\Omega)}$ and $\varphi(u) := L^{-1}G(u)$. Then, we have $\varphi, f \in C^1(X, X)$. We can rewrite (5.1) as $f(u) = 0$. We choose X_h as an approximate finite element subspace of X with mesh size h .

In what follows, we concentrate on the cases: $\Omega = (0, 1) \subset \mathbf{R}$ and $\Omega = (0, 1) \times (0, 1) \subset \mathbf{R}^2$. We use finite element methods with piecewise linear and bilinear elements on the uniform (rectangular) mesh with mesh size $h = 1/n$ ($n \in \mathbf{N}$). Then, we have $\dim X_h = n - 1$ in the 1-dimensional case and $\dim X_h = (n-1)^2$ in the 2-dimensional case. In this context the following basic estimates hold:

$$\|(I - P_h)u\|_{L^2(\Omega)} \leq C_a h \|u\|_X \quad \text{for any } u \in X, \quad (5.2a)$$

$$\|(I - P_h)u\|_X \leq C_b h \|\Delta u\|_{L^2(\Omega)} \quad \text{for any } u \in X \cap H^2(\Omega), \quad (5.2b)$$

$$\|(I - P_h)u\|_{L^\infty(\Omega)} \leq C_c h \|\Delta u\|_{L^2(\Omega)} \quad \text{for any } u \in X \cap H^2(\Omega), \quad (5.2c)$$

where C_a , C_b , and C_c are some positive constants independent of h and u . As in previous sections, we denote by u_* an isolated solution of $f(u) = 0$ and by u_h a finite element solution of $f(u) = 0$ (i.e., a solution of $\tilde{f}_h(u) = 0$) in a small neighborhood of u_* . In view of Proposition 2.1, u_h exists uniquely in a small neighborhood of u_* for sufficiently small $h > 0$. In our examples below we show that the following error estimate holds:

$$\|P_h u_* - u_h\| = O(h^2) \quad \text{as } h \rightarrow 0. \quad (5.3)$$

For simplicity we denote $u_*(s, h) := (1 - s)u_* + sP_h u_*$. We will derive (5.3) from Theorem 2.5 and the duality

$$\|\varphi'(u_*(s, h))(I - P_h)\| = \|(I - P_h)\varphi'(u_*(s, h))^*\|. \quad (5.4)$$

We now present two examples.

Example 5.1. We consider the following Burgers equation:

$$-\Delta u + uu_x = g \quad \text{in } \Omega := (0, 1) \times (0, 1) \text{ with } u = 0 \text{ on } \partial\Omega. \quad (5.5)$$

Here, $g(x, y)$ is a given function with $g \in L^2(\Omega)$. As mentioned above, we rewrite (5.5) as $f(u) := u - \varphi(u) = 0$. In the present case $\varphi : X \rightarrow X$ is a nonlinear map defined by $\varphi(u) := L^{-1}(-uu_x + g)$. By the elliptic regularity property we have $u_* \in H^2(\Omega)$ (see e.g., [9]). We will derive (5.3). Let $u, v \in X$. We easily verify that

$$\varphi'(v)u = -L^{-1}(vu_x + v_x u), \quad \varphi'(v)^*u = L^{-1}(vu_x) \quad \text{for any } u \in X. \quad (5.6)$$

By (5.2b) we have

$$\|(I - P_h)\varphi'(u_*(s, h))^*u\| \leq C_b h \|u_*(s, h)u_x\|_{L^2(\Omega)} \leq C_b h \|u_*(s, h)\|_{L^\infty(\Omega)} \|u\|_X. \quad (5.7)$$

It follows that

$$\|(I - P_h)\varphi'(u_*(s, h))^*\| \leq C_b h \|u_*(s, h)\|_{L^\infty(\Omega)}. \quad (5.8)$$

We obtain from (5.2c) that

$$\|u_*(s, h)\|_{L^\infty(\Omega)} \leq \|u_*\|_{L^\infty(\Omega)} + C_c h s \|\Delta u_*\|_{L^2(\Omega)}. \quad (5.9)$$

It follows from (5.4), (5.8), and (5.9) that

$$\sup_{s \in [0, 1]} \|\varphi'(u_*(s, h))(I - P_h)\| \leq C_b h \left(\|u_*\|_{L^\infty(\Omega)} + C_c h \|\Delta u_*\|_{L^2(\Omega)} \right) := \varepsilon_h. \quad (5.10)$$

By (5.10), (5.2b), and Theorem 2.5 we have (5.3).

Example 5.2. We consider the Emden equation

$$-\Delta u = u^2 \quad \text{in } \Omega := (0, 1) \times (0, 1) \text{ with } u = 0 \text{ on } \partial\Omega. \quad (5.11)$$

We omit the one-dimensional case since it is easier. We can treat the present case in a similar way to Example 5.1. We rewrite (5.11) as (1.1). In the present case $\varphi : X \rightarrow X$ is defined by $\varphi(u) := L^{-1}(u^2)$. Let $u, v \in X$. We verify that $\varphi'(v)u = 2L^{-1}(vu)$ and that $\varphi'(v)$ is self-adjoint. By (5.2b) and Sobolev's inequality we have

$$\begin{aligned} \|(I - P_h)\varphi'(u_*(s, h))u\| &\leq 2C_b h \|u_*(s, h)\|_{L^4(\Omega)} \|u\|_{L^4(\Omega)} \\ &\leq Ch \|u_*(s, h)\| \|u\| \leq Ch \|u_*\| \|u\| \end{aligned} \quad (5.12)$$

for any $s \in [0, 1]$ and $u \in X$. Here, $C > 0$ is a constant independent of s , h , and u . It follows from this inequality and (5.4) that

$$\sup_{s \in [0, 1]} \|\varphi'(u_*(s, h))(I - P_h)\| \leq Ch \|u_*\| := \varepsilon_h. \quad (5.13)$$

By (5.13), (5.2b), and Theorem 2.5 we have (5.3).

Remark 5.3. This remark is related to Remark 2.6.

(i) As mentioned in Remark 2.6 (ii), $\sup_{s \in [0, 1]} \|\varphi'((1-s)u_* + sP_h u_*)(I - P_h)\| \approx \|T(I - P_h)\|$ holds in general. Actually, in Example 5.2 (resp., Example 5.1) our best possible upper bound of $\sup_{s \in [0, 1]} \|\varphi'(u_*(s, h))(I - P_h)\|$ is the right-hand side of (5.13) (resp., (5.10)), which is just the same (resp., has the same order) as that of $\|T(I - P_h)\|$.

(ii) We pointed out that our estimate (2.10) is in general sharper than (2.11), which is directly derived from the discussion in [1]. In order to show it concretely, we apply (2.11) to the equations in Examples 5.1 and 5.2. In both cases our best possible error estimate is the following:

$$\|P_h u_* - u_h\| = O(h^{3/2}) \quad \text{as } h \rightarrow 0. \quad (5.14)$$

Compare (5.14) with (5.3), which is based on (2.10). Though we omit the detailed derivation of (5.14), we show here that we cannot obtain a better estimate than (5.14) if we use (2.11) as a basic estimate. By the same discussion in Examples 5.1 and 5.2 we have

$$\|(I - P_h)\varphi'(u_*)\| = O(h) \quad \text{as } h \rightarrow 0, \quad (5.15)$$

which is our best possible upper estimate of $\|(I - P_h)T\|$. So, in view of (2.13), it is necessary to take $\hat{\varepsilon}_h$ such that $\hat{\varepsilon}_h \geq Ch^{1/2}$ for small h . Here, $C > 0$ is a constant independent of h . (Compare this estimate with (5.10) and (5.13).) Therefore, we cannot improve (5.14) if we use (2.11) and (5.2b) as basic estimates.

Remark 5.4. Various numerical verification algorithms for solutions of differential equations were proposed up to now (see e.g., [10]). Some of them give upper bounds of $\|P_h u_* - u_h\|$ as

output data (see [5]). Theorem 2.5 may be applicable for checking the accuracy of such output upper bounds since we can apply it to given problems in order to compute the concrete order of $\|P_h u_* - u_h\|$ as $h \rightarrow 0$. For example, we treated problems (5.5) and (5.11) as concrete numerical examples in [5], where we proposed a numerical verification algorithm based on a convergence theorem of Newton's method. In these problems (5.3) is the theoretical estimate of $\|P_h u_* - u_h\|$ derived from our Theorem 2.5. The output data as upper bounds of $\|P_h u_* - u_h\|$ in [5, Section 3] seem to have just the order of h^2 as $h \rightarrow 0$. So, the accuracy of such output upper bounds in [5, Section 3] is satisfactory as long as we judge it by the theoretical estimate (5.3).

Acknowledgments

The author would like to express his sincere gratitude to Professor Takuya Tsuchiya and Professor Atsushi Yagi for their valuable comments and encouragement. He is grateful to the referee for constructive comments.

References

- [1] M. H. Schultz, "Error bounds for the Rayleigh-Ritz-Galerkin method," *Journal of Mathematical Analysis and Applications*, vol. 27, pp. 524–533, 1969.
- [2] M. Urabe, "Galerkin's procedure for nonlinear periodic systems," *Archive for Rational Mechanics and Analysis*, vol. 20, pp. 120–152, 1965.
- [3] T. Kawanago, "Computer assisted proof to symmetry-breaking bifurcation phenomena in nonlinear vibration," *Japan Journal of Industrial and Applied Mathematics*, vol. 21, no. 1, pp. 75–108, 2004.
- [4] T. Kawanago, "A symmetry-breaking bifurcation theorem and some related theorems applicable to maps having unbounded derivatives," *Japan Journal of Industrial and Applied Mathematics*, vol. 21, no. 1, pp. 57–74, 2004, Corrigendum to this paper: *Japan J. Indust. Appl. Math.* 22 (2005) 147.
- [5] T. Kawanago, "Improved convergence theorems of Newton's method designed for the numerical verification for solutions of differential equations," *Journal of Computational and Applied Mathematics*, vol. 199, no. 2, pp. 365–371, 2007.
- [6] S. G. Mikhlin, *Variational Methods in Mathematical Physics*, The Macmillan, New York, NY, USA, 1964.
- [7] M. Crouziex and J. Rappaz, *On Numerical Approximation in Bifurcation Theory*, Springer, 1990.
- [8] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, vol. 5 of *Springer Series in Computational Mathematics*, Springer, Berlin, Germany, 1986.
- [9] Y. Grisvard, *Elliptic Problems in Nonsmooth Domain*, Pitman, Boston, Mass, USA, 1985.
- [10] M. T. Nakao, "Numerical verification methods for solutions of ordinary and partial differential equations," *Numerical Functional Analysis and Optimization*, vol. 22, pp. 321–356, 2001.