AN ASYMPTOTICALLY LINEAR FIXED POINT EXTENSION OF THE INF-SUP THEORY OF GALERKIN APPROXIMATION*

Joseph W. Jerome†
Department of Mathematics
Northwestern University
Evanston, Il 60208

Abstract

In 1972, Babuška and Aziz introduced a Galerkin approximation theory for saddle point formulations of linear partial differential equations (The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, 1972). It represented a powerful extension of the approximation theory for positive-definite, self-adjoint operators. Independently, a coherent theory for the approximation of fixed points of nonlinear mappings by numerical fixed points was devised by Krasnosel'skii and his coworkers (Approximate Solution of Operator Equations, Wolters-Noordhoff, 1972). In this paper, the Krasnosel'skii Calculus is shown to be a logical extension of the inf-sup theory constructed by Babuška and Aziz. In the process, we obtain sharp lower bounds, not emphasized by these authors. We also identify a novel fundamental approximation property of the nonlinear calculus, which we characterize as asymptotic linearity. This is adjoined to a robust estimate for replacement of numerical fixed points in the outer iteration by a single Newton inner iteration, beginning each time with the previously computed outer iteration numerical fixed point. This yields a tight closure to the property of asymptotic linearity.

*Dedicated to the memory of Farouk Odeh

†The author was supported by the National Science Foundation under grant DMS-9123208.

1 Introduction.

The mathematical foundation of the finite element method for positive-definite, self-adjoint linear equations rests upon two fundamental results, viz., the Bramble-Hilbert lemma ([3] and [6]) and the Aubin-Nitsche lemma ([1] and [13]). The first deals primarily with piecewise polynomial trial spaces, while the second is generic to Galerkin approximation. The purpose of the Bramble-Hilbert lemma is to utilize the energy norm best approximation property of the Galerkin approximation; the latter is elementary for positive-definite, self-adjoint equations. It achieves this by examining, via remainder analysis, interpolation or smoothing linear approximation operators, which reproduce polynomials locally, and hence are of optimal order. The net result of an application of this tool is order one convergence, in the energy norm, for second order equations in terms of piecewise polynomial Galerkin approximants. This presupposes the absence of geometric singular points, such as reentrant corners, or boundary condition transition point singularities. The Aubin-Nitsche lemma improves the order of approximation when measured in a ground space norm, specifically the L_2 norm. For example, for second order equations with Neumann boundary conditions, the convergence is of order two in the L_2 metric for Galerkin finite element approximants. There is a remarkable correlation with a theory introduced by Kolmogorov in 1936 ([11], [7]), called the theory of n-widths; the Galerkin approximation achieves the optimal L_2 order in terms of n, predicted by Kolmogorov's theory of best approximation by subspaces of dimension n. Altogether, the finite element Galerkin theory is quite tightly wound in this case; for an exposition, the reader is referred to [14].

Any generalization of the theory just described must begin with establishing a near best approximation estimate for the Galerkin approximation in energy type norms. In 1972, a landmark theory was introduced by Babuška and Aziz [2]. It represented a powerful extension of the approximation theory for positive-definite, self-adjoint operators, in the context of Galerkin approximation, or Petrov-Galerkin approximation. Particularly striking about this theory were both its rapid appearance, after the completion of the main features of the positive-definite, self-adjoint theory, as well as the tightness of the formulation of the *framework* and the *results*, in the linear, saddle point case. Though its principal application has been to finite elements, it is more comprehensive. Continuous bilinear forms, satisfying an inf-sup condition and some auxiliary conditions, were identified as properly defining an invertible operator framework, allowing for analysis and approximation theory. Without exploiting it directly, these authors created an underlying fixed point map, via the continuous inverse map. We shall summarize the essential features in the following section. It should be mentioned that [5] was developed almost simultaneously with [2], and was also influential. The linear theory is very nicely

described in the book of Brenner and Scott [4].

Independently, an operator calculus for approximation of the fixed points of nonlinear mappings by numerical fixed points, in general Banach spaces, was developed by Krasnosel'skii and his coworkers [12]. In the intervening years, no linking of these two theories has been made explicit in the literature to the author's knowledge. The Krasnosel'skii Calculus, a term coined by Kerkhoven and the author in their application of the theory to the semiconductor model [10], is intended for nonlinear equations and systems, and the fundamental estimates are masked by quantities expressing operator estimation. However, it was noticed by the author [9] that the inf-sup theory is implied by the Krasnosel'skii theory, so that the latter may be viewed as an appropriate generalization, to nonlinear analysis, of the Babuška-Aziz theory. No proofs were given of this fact in [9], though it was mentioned that the theory of [12] yields lower, as well as upper, bounds for the numerical approximation. This fact was not emphasized in [2]. It was originally the author's intention to leave the matter with the observations contained in [9]. Several participants at the 1992 Lancaster Summer School, at which the lectures of [9] were delivered, requested a more complete description, however, and a corresponding documentation of the precise sense in which the generalization holds. This paper, then, addresses such issues. In section §3, particularly Theorem 3.2, we shall summarize the theory of [12] as it applies here, and in §4 we shall draw the relevant correspondences, summarized in Corollary 4.1. It is important to note that only the Galerkin case of [2] is studied here, not the more general case of Petrov-Galerkin approximation, where the trial and test spaces may differ.

In the reduction of the nonlinear calculus to the inf-sup theory, the numerical fixed points of the approximate mapping $\mathbf{T}_n$ are designed to be the Galerkin approximations, with a similar relation holding between the analytical fixed point and the generalized boundary value problem solution. At the heart of the argument is the identification in the linear theory (cf. (3.22)) that the metric distance, between the numerical fixed point and the energy projection of the fixed point, is equal to the norm of the Newton increment, based upon the map, $\mathbf{I} - \mathbf{T}_n$, and starting at the projection of the fixed point. In §5, it is shown that this holds in an asymptotic sense for the general nonlinear calculus, under a Lipschitz condition on the derivative. This is what is meant by asymptotic linearity. We demonstrate more; for decreasing approximation errors the local linear approximation projected onto the nonlinear problem becomes progressively more accurate. This clarifies an issue raised by Ivo Babuška in 1987.

It is especially fitting that it was Farouk Odeh who originally suggested to the author and Kerkhoven that the operator calculus of [12] was the viable framework with which to derive the error estimates for the semiconductor device system in [10]. Although this paper does not directly deal with the

device equations, it does deal with the kind of fundamental mathematical issue which was the forté of Farouk's scientific agenda. It is dedicated, with fondness and gratitude, to his memory.

2 The Inf-Sup Theory

As is consistent with our objectives, we shall not present the most general formulation described in [2]. It is desired to solve the operator equation, $\mathbf{L}u = f$, approximately. In this theory, the solvability of the direct and adjoint problems is assured. If B denotes the bilinear form of the weak formulation on a Hilbert space E, with inner product, $(\cdot\,,\cdot\,)$, assume:

- continuity:
$$|B(v,w)| \leq C_1\|v\|\ \|w\|. \tag{2.1}$$

- sup condition:
$$\text{For}\ \ w \neq 0,\ \ \sup_v |B(v,w)| > 0. \tag{2.2}$$

- inf-sup condition:
$$\inf_{\|v\|=1} \sup_{\|w\|\leq 1} |B(v,w)| \geq C_2 > 0. \tag{2.3}$$

Assume also sup and inf-sup conditions on an approximation space, E_n:

$$\text{For}\ \ \psi \neq 0,\ \ \sup_\phi |B(\phi,\psi)| > 0. \tag{2.4}$$

$$\inf_{\|\phi\|=1} \sup_{\|\psi\|\leq 1} |B(\phi,\psi)| \geq c_2 > 0. \tag{2.5}$$

One concludes the Galerkin approximation, u_n, is well defined by the relation,

$$B(u_n,\psi) = (\mathbf{J}f,\psi),\ \ \forall\psi \in E_n, \tag{2.6}$$

where $\mathbf{J}$ denotes the Riesz map. Moreover, u_n is within a metric distance,

$$\delta_*\{1 + (C_1/c_2)\}, \tag{2.7}$$

of u, where

$$\delta_* := \|u - u_*\|, \tag{2.8}$$

and u_* is arbitrary in E_n (cf. [2, pp. 187-188]).

We note briefly the role of the hypotheses.

1. (2.1) $\Rightarrow$ $\mathbf{L}$ may be identified with a continuous linear map $\mathbf{R}$ on E (cf. (2.9)).

2. (2.3) $\Rightarrow$ $\mathbf{R}^{-1}$ exists on a closed domain of E.

3. (2.2) $\Rightarrow$ Domain and range of $\mathbf{R}$ are all of E.

In §4, we shall mention explicitly certain properties of this construction which are critical in linking the two theories under study. There is a fixed point formulation if 1 is not in the spectrum:

$$\mathbf{T}u = u, \;\; \mathbf{T}v := (\mathbf{I} - \mathbf{R})^{-1}(v - \mathbf{J}f), \;\; \mathbf{R} = \mathbf{JL}. \tag{2.9}$$

$\mathbf{T}$ is affine. Its domain independent derivative is defined by

$$\mathbf{T}' \equiv (\mathbf{I} - \mathbf{JL})^{-1}. \tag{2.10}$$

Note that 1 is an eigenvalue of $\mathbf{T}'$ if and only if 0 is an eigenvalue of $\mathbf{L}$. The latter is excluded, so the exclusion of 1 as a spectral value is a solvability hypothesis when $\mathbf{T}'$ is compact.

In the discussion to follow, the reader may visualize the solvability conditions as generalized by the uniform invertibility condition on the derivative of the fixed point map at the fixed point. In the present setting, this invertibility is simply an eigenvalue condition via the compactness properties. We now introduce this theory.

3 The Krasnosel'skii Calculus

Given a fixed point x_0 of a smooth mapping $\mathbf{T}$, a numerical approximation map $\mathbf{T}_n$, with numerical fixed point x_n, and a linear projection map $\mathbf{P}_n$, a theory is constructed to estimate $\|x_n - \mathbf{P}_n x_0\|$. In fact, the authors of [12] characterize the map $\mathbf{P}_n\mathbf{T}$ as the "Galerkin" approximate map, and $\mathbf{T}_n$ as a "perturbed Galerkin" map. Since x_0 is a fixed point of $\mathbf{T}$, the estimates represent the dispersion between these two methods. The mapping $\mathbf{P}_n\mathbf{T}$, while not actually implemented numerically, has a convergence rate which is readily estimated. Now, the manner in which the 'a priori' estimates are derived is to deduce a zero of the map $\mathbf{I}-\mathbf{T}_n$, in a ball centered at $\mathbf{P}_n x_0$, by constructing an equivalent contraction map: The methodology involves derivative inversion and a mean value calculus embodied in Lemma 3.1. The final result of that theory is stated as Theorem 3.1 below.

3.1 The Abstract Calculus

Let E be a Banach space and $\mathbf{T}$ a mapping from an open set Ω into E. We assume the existence of a fixed point x_0 for $\mathbf{T}$:

$$\mathbf{T}x_0 = x_0. \tag{3.1}$$

If $\{E_n\}$ denotes a sequence of subspaces of E of dimension $r(n) \geq n$, suppose that $\mathbf{T}_n : \Omega_n \mapsto E_n$, $\Omega_n := \Omega \cap E_n$, has a fixed point:

$$\mathbf{T}_n x_n = x_n. \tag{3.2}$$

Finally, let $\{\mathbf{P}_n\}$ be a family of linear projections onto E_n. We shall describe here the framework of the calculus developed by Krasnosel'skii et al. [12] for the convergence of the solutions of discretizations of fixed point equations (3.2) to the solutions of the original fixed point equation (3.1). The results depend fundamentally upon the following lemma, restated from [12, Lemma 19.1].

Lemma 3.1. *Let* $\mathbf{A}$ *be an operator in a Banach space* X *which is Fréchet differentiable in a closed ball centered at* x_*. *Suppose* $[\mathbf{A}'(x_*)]^{-1}$ *exists as a bounded linear operator on* X, *and that the following conditions hold:*

$$\sup_{\|x-x_*\| \leq \delta_0} \|[\mathbf{A}'(x_*)]^{-1}[\mathbf{A}'(x) - \mathbf{A}'(x_*)]\| \leq q, \tag{3.3}$$

$$\alpha := \|[\mathbf{A}'(x_*)]^{-1}\mathbf{A}(x_*)\| \leq \delta_0(1-q), \tag{3.4}$$

for some δ_0 *and* $0 < q < 1$. *Then the equation* $\mathbf{A}x = 0$ *has a unique solution* x_0 *in the ball,* $\|x_0 - x_*\| \leq \delta_0$, *and the estimate,*

$$\frac{\alpha}{1+q} \leq \|x_0 - x_*\| \leq \frac{\alpha}{1-q}, \tag{3.5}$$

holds.

The standard result of the calculus, to be quoted as Theorem 3.1, is a consequence of Lemma 3.1. Although we shall not directly make use of the theorem, in the form quoted, it will be advantageous to indicate its relationship to the core identities and inequalities which will establish the reduction to the inf-sup theory. Thus, we cite Theorem 19.1 in [12], and follow this with a discussion of the proof, containing the components of the reduction just cited.

Theorem 3.1. *Let the operators* $\mathbf{T}$ *and* $\mathbf{P}_n\mathbf{T}$ *be Fréchet-differentiable in* Ω, *and* $\mathbf{T}_n$ *Fréchet-differentiable in* Ω_n. *Assume that* (3.1) *has a solution* $x_0 \in \Omega$ *and the linear operator* $\mathbf{I} - \mathbf{T}'(x_0)$ *is continuously invertible in* E. *Let*

$$\|\mathbf{P}_n(x_0) - x_0\| \to 0, \tag{3.6}$$

$$\|\mathbf{P}_n\mathbf{T}\mathbf{P}_n x_0 - \mathbf{T}x_0\| \to 0, \tag{3.7}$$

$$\|\mathbf{P}_n\mathbf{T}'(\mathbf{P}_n x_0) - \mathbf{T}'(x_0)\| \to 0, \tag{3.8}$$

$$\|[\mathbf{T}_n - \mathbf{P}_n\mathbf{T}]\mathbf{P}_n x_0\| \to 0, \tag{3.9}$$

$$\|[\mathbf{T}_n' - (\mathbf{P}_n\mathbf{T})'](\mathbf{P}_n x_0)\| \to 0, \tag{3.10}$$

as $n \to \infty$. Finally, assume that for any $\epsilon > 0$ there exist n_ϵ and $\delta_\epsilon > 0$ such that

$$\|\mathbf{T}_n^{'}(x) - \mathbf{T}_n^{'}(\mathbf{P}_n x_0)\| \leq \epsilon \quad for \quad (n \geq n_\epsilon;\ \|x - \mathbf{P}_n x_0\| \leq \delta_\epsilon,\ x \in \Omega_n). \tag{3.11}$$

Then there exist n_0 and $\delta_0 > 0$ such that, when $n \geq n_0$, equation (3.2) has a unique solution x_n in the ball $\|x - x_0\| \leq \delta_0$. Moreover,

$$\|x_n - x_0\| \leq \|[\mathbf{I} - \mathbf{P}_n]x_0\| + \|x_n - \mathbf{P}_n x_0\| \to 0 \quad \text{as} \quad n \to \infty, \tag{3.12}$$

and $\|x_n - \mathbf{P}_n x_0\|$ satisfies the following two-sided estimate ($c_1, c_2 > 0$):

$$c_1\|\mathbf{P}_n\mathbf{T}x_0 - \mathbf{T}_n\mathbf{P}_n x_0\| \leq \|x_n - \mathbf{P}_n x_0\| \leq c_2\|\mathbf{P}_n\mathbf{T}x_0 - \mathbf{T}_n\mathbf{P}_n x_0\|. \tag{3.13}$$

The role of (3.8) and (3.10) is to obtain the uniform boundedness of certain inverse mappings. More precisely, when the invertibility hypothesis is joined with these relations, one concludes the existence of constants κ and κ' such that

$$\|[\mathbf{I} - \mathbf{T}_n'(\mathbf{P}_n x_0)]^{-1}\| \leq \kappa, \quad \|\mathbf{I} - \mathbf{T}_n'(\mathbf{P}_n x_0)\| \leq \kappa', \tag{3.14}$$

for $n \geq n_*$. Now the following statement is obtained by direct estimation. The numbers α_n, for α_n defined by

$$\alpha_n = \|[\mathbf{I} - \mathbf{T}_n'(\mathbf{P}_n x_0)]^{-1}(\mathbf{I} - \mathbf{T}_n)\mathbf{P}_n x_0\|, \tag{3.15}$$

satisfy $\alpha_n \to 0$, hence (3.12), via

$$\frac{1}{\kappa'}\|(\mathbf{P}_n\mathbf{T} - \mathbf{T}_n\mathbf{P}_n)x_0\| \leq \alpha_n \leq \kappa\|(\mathbf{P}_n\mathbf{T} - \mathbf{T}_n\mathbf{P}_n)x_0\| \tag{3.16}$$

for $n \geq n_*$. The theorem can now be proved as follows. The hypotheses of Lemma 3.1 can be verified via (3.11) and (3.14), with $x_* \mapsto \mathbf{P}_n x_0$, and $\mathbf{A} = \mathbf{I} - \mathbf{T}_n$. Here, the choice of $0 < q < 1$ is immaterial. The conclusion of the lemma, with $x_0 \mapsto x_n$, combined with (3.16), yields (3.13), with

$$c_1 = \frac{1}{\kappa'(1+q)},\ c_2 = \frac{\kappa}{1-q}.$$

Uniqueness follows from an adjustment of δ_0. On the basis of the previous discussion, we can state the following alternative formulation of Theorem 3.1. It also incorporates the important special case of affine mappings.

Theorem 3.2. *Let the operators* $\mathbf{T}$ *and* $\mathbf{P}_n\mathbf{T}$ *be Fréchet-differentiable in* Ω*, and* $\mathbf{T}_n$ *Fréchet-differentiable in* Ω_n*. Assume that* (3.1) *has a solution* $x_0 \in \Omega$ *and*

$$\|\mathbf{P}_n x_0 - x_0\| \to 0, \tag{3.17}$$

$$\|\mathbf{P}_n\mathbf{T}\mathbf{P}_n x_0 - \mathbf{T}x_0\| \to 0, \tag{3.18}$$

$$\|[\mathbf{T}_n - \mathbf{P}_n\mathbf{T}]\mathbf{P}_n x_0\| \to 0, \tag{3.19}$$

as $n \to \infty$*. Also, assume that for any* $\epsilon > 0$ *there exist* n_ϵ *and* $\delta_\epsilon > 0$ *such that*

$$\|\mathbf{T}_n'(x) - \mathbf{T}_n'(\mathbf{P}_n x_0)\| \le \epsilon \quad \textit{for} \quad (n \ge n_\epsilon;\ \|x - \mathbf{P}_n x_0\| \le \delta_\epsilon,\ x \in \Omega_n). \tag{3.20}$$

Suppose, finally, that the following bounds hold:

$$\|[\mathbf{I} - \mathbf{T}_n'(\mathbf{P}_n x_0)]^{-1}\| \le \kappa, \quad \|\mathbf{I} - \mathbf{T}_n'(\mathbf{P}_n x_0)\| \le \kappa', \tag{3.21}$$

for $n \ge n_*$*. Then the conclusions of Theorem* 3.1 *hold, including* (3.13)*. In the special case that* $\mathbf{T}$ *and* $\mathbf{T}_n$ *are affine, the choice* $q = 0$ *is possible, and the following relation holds:*

$$\|x_n - \mathbf{P}_n x_0\| = \alpha_n. \tag{3.22}$$

4 The Inf-Sup Theory as a Special Case

In this section, we shall draw the connections between the two preceding sections. Specifically, we shall show that the hypotheses of the Babuška-Aziz theory imply the hypotheses, and hence the conclusions, of the Krasnosel'skii theory as stated in Theorem 3.2. The latter conclusions are seen to imply those of the inf-sup theory. We begin by itemizing key properties of the construction of [2]. It is now understood that E is a Hilbert space. In applications, E is typically the Sobolev space, H^1.

- The fixed point map, $\mathbf{T}$, is defined by (2.9). The mapping $\mathbf{R}$ in the definition of $\mathbf{T}$ is determined by the relation,

$$B(u, v) = (\mathbf{R}u, v), \ \ \forall u, v \in E. \tag{4.1}$$

- If the numerical fixed point map on E_n is defined by

$$\mathbf{T}_n v = (\mathbf{I} - \mathbf{P}_n\mathbf{R})^{-1}(v - \mathbf{P}_n\mathbf{J}f),\ v \in E_n, \tag{4.2}$$

  it follows that the Galerkin approximation, written as u_n, is characterized as a fixed point of $\mathbf{T}_n$. Indeed, it is shown in ([2, pp. 187-188]) that $\mathbf{P}_n\mathbf{R}$ maps E_n into itself according to

$$B(\phi, \psi) = (\mathbf{P}_n\mathbf{R}\phi, \psi), \ \ \forall \phi, \psi \in E_n. \tag{4.3}$$

  The fixed point property then follows immediately from (2.6) and (4.3).

- The domain independent derivative of $\mathbf{T}_n$ is given by

$$\mathbf{T}_n' v = (\mathbf{I} - \mathbf{P}_n \mathbf{R})^{-1} v, \quad \forall v \in E_n. \tag{4.4}$$

- If v is an eigenvector of $\mathbf{T}'$, given by (2.10), corresponding to eigenvalue, 1, then v is also an eigenvector of $\mathbf{R}$, corresponding to eigenvalue, 0. This follows from the relation,

$$\mathbf{T}' v = (\mathbf{I} - \mathbf{R})^{-1} v = v.$$

  Since the latter contradicts both the sup condition (2.2) and the inf-sup condition (2.3), 1 is not an eigenvalue. This represents the principal nonsingularity hypothesis of §3.

Lemma 4.1. *Let u denote the unique solution of the operator equation, $\mathbf{L}u = f$. Set*

$$v_n = [\mathbf{I} - \mathbf{T}_n']^{-1} (\mathbf{I} - \mathbf{T}_n) \mathbf{P}_n u. \tag{4.5}$$

Then the relations,

$$\mathbf{P}_n \mathbf{R} v_n = \mathbf{P}_n \mathbf{R} \mathbf{P}_n u - \mathbf{P}_n \mathbf{J} f, \tag{4.6}$$

$$\mathbf{P}_n \mathbf{R} u = \mathbf{P}_n \mathbf{R} u_n = \mathbf{P}_n \mathbf{J} f, \tag{4.7}$$

hold. In particular, if $(\mathbf{P}_n \mathbf{R})^{-1}$ denotes inversion on E_n, it follows that

$$\alpha_n = \|v_n\| = \|(\mathbf{P}_n \mathbf{R})^{-1} [\mathbf{P}_n \mathbf{R} (\mathbf{P}_n u - u)]\|. \tag{4.8}$$

Proof. We begin with (4.5), apply the mapping, $[\mathbf{I} - \mathbf{T}_n']$, insert the representations given by (4.2) and (4.4), and simplify to obtain (4.6).

In order to obtain the second equality in (4.7), begin with the relation, $\mathbf{T}_n u_n = u_n$, substitute (4.2), and simplify. In a similar manner, the relation, $\mathbf{T} u = u$, implies

$$\mathbf{R} u = \mathbf{J} f,$$

so that both equalities in (4.7) are seen to hold. The identity (4.8) is now a simple consequence of (4.6) and (4.7), and the characterization (3.15) of α_n. □

The following theorem is the basis for the result described by (2.7).

Theorem 4.1. *We have the relations,*

$$C_2 \|\mathbf{P}_n u - u\| \le \|\mathbf{P}_n \mathbf{R} (\mathbf{P}_n u - u)\| = \sup_{\|v\|_{E_n} = 1} |B(\mathbf{P}_n u - u, v)| \le C_1 \|\mathbf{P}_n u - u\|, \tag{4.9}$$

$$\alpha_n \le \|(\mathbf{P}_n \mathbf{R})^{-1}\| \|\mathbf{P}_n \mathbf{R} (\mathbf{P}_n u - u)\| \le \frac{C_1}{c_2} \|\mathbf{P}_n u - u\|, \tag{4.10}$$

$$\alpha_n \ge \frac{1}{\|\mathbf{P}_n \mathbf{R}\|} \|\mathbf{P}_n \mathbf{R} (\mathbf{P}_n u - u)\| \ge \frac{C_2}{C_1} \|\mathbf{P}_n u - u\|. \tag{4.11}$$

Proof. The equality of (4.9) proceeds from the following. The identity,

$$\|\mathbf{P}_n\mathbf{R}(\mathbf{P}_nu-u)\| = \sup_{\|v\|_{E_n}=1} |\ (\mathbf{P}_n\mathbf{R}(\mathbf{P}_nu-u),v)\ | = \sup_{\|v\|_{E_n}=1} |\ (\mathbf{R}(\mathbf{P}_nu-u),v)\ |, \tag{4.12}$$

is a simple duality characterization, in tandem with an application of the projection. The equality follows when this is combined with (4.1). The two inequalities of (4.9) are immediate from the equality, in conjunction with the inf-sup property, (2.3), and the continuity property, (2.1), respectively. It is routine that

$$\|(\mathbf{P}_n\mathbf{R})^{-1}\| \leq \frac{1}{c_2}, \tag{4.13}$$

$$\|\mathbf{P}_n\mathbf{R}\| \leq C_1. \tag{4.14}$$

These inequalities, when joined with (4.9), lead directly to (4.11) and (4.10). □

We identify the critical implications in the following.

Corollary 4.1. *The relation* (2.7) *provides an upper bound for* $\|u-u_n\|$. *In addition, the lower bound,*

$$\|\mathbf{P}_nu-u_n\| \geq \frac{C_2}{C_1}\|\mathbf{P}_nu-u\|, \tag{4.15}$$

holds.

5 Asymptotic Linearity

The result expressed in Theorem 3.1, (3.13), depends upon the following inequality:

$$\frac{\alpha_n}{1+q} \leq \|x_n-\mathbf{P}_nx_0\| \leq \frac{\alpha_n}{1-q}, \tag{5.1}$$

which is a restatement of (3.5). The number q is fixed here. It is an interesting question whether a refined version of (5.1) holds, with a sequence q_n replacing q, $q_n \to 0$. In this case, we can call the approximations defined by x_n asymptotically linear, since

$$\alpha_n \sim \|x_n-\mathbf{P}_nx_0\| \tag{5.2}$$

in this case, with the usual meaning that the quotients tend to one. This constitutes a natural extension of the relation of equality, which holds in the affine case (cf. (3.22)), and accounts for the use of the description of asymptotic linearity. We recall that $\alpha_n \to 0$, as a consequence of (3.16) and the hypotheses (3.6), (3.7), and (3.9). We have the following.

Theorem 5.1. *Suppose that $\mathbf{T}_n'$ is Lipschitz continuous with constant $2C > 0$, independent of n, so that (3.11) holds with $\delta_\epsilon = \epsilon/(2C)$. Then the numbers q_n, defined for n such that $\alpha_n < \frac{1}{8\kappa C}$ by,*

$$q_n = \frac{1}{2}\left(1 - \sqrt{1 - 8\alpha_n \kappa C}\right), \tag{5.3}$$

satisfy $q_n \to 0$. Moreover, if the identifications $x_0 \mapsto x_n$ and $x_* \mapsto \mathbf{P}_n x_0$ are made, then the mapping, $\mathbf{A} = \mathbf{I} - \mathbf{T}_n$, satisfies the two conditions of Lemma 3.1, viz., (3.3) and (3.4), so that

$$\frac{\alpha_n}{1 + q_n} \leq \|x_n - \mathbf{P}_n x_0\| \leq \frac{\alpha_n}{1 - q_n} = \delta_n. \tag{5.4}$$

In particular, the asymptotic relation (5.2) holds.

Proof. Set $\epsilon_n = \frac{q_n}{\kappa}$. Then (3.3) holds, with the stated identifications. With the choice of $\delta_n = \frac{\epsilon_n}{2C}$, we have the equation,

$$\alpha_n = \frac{q_n(1 - q_n)}{2\kappa C} = \delta_n(1 - q_n), \tag{5.5}$$

so that (3.4) holds and hence (5.4) and (5.2). □

The final result of this paper deals with maintaining the essence of the upper bound in (5.4), while transferring the determination of x_n to the solution of an approximate linear problem, via one Newton iteration. We begin with a general theoretical result concerning the convergence of Newton's method for the approximation of x_n, defined via

$$u_n^{k+1} - u_n^k = -[\mathbf{A}_n'(u_n^k)]^{-1}\mathbf{A}_n(u_n^k). \tag{5.6}$$

The following lemma is based upon [8, Lemma 2.2], where the identifications, $\alpha = 3\delta_0/4$, and $\theta_1\tau_1 = 1/2$, are made.

Lemma 5.1. *Suppose that $\mathbf{T}_n'$ is Lipschitz continuous with constant $2C > 0$, independent of n, so that (3.11) holds with $\delta_\epsilon = \epsilon/(2C)$. Moreover, without loss of generality, select κ so that*

$$\|[\mathbf{A}_n'(y)]^{-1}\| \leq \kappa, \tag{5.7}$$

uniformly for $y \in \{x : \|x - x_0\| \leq \delta_0\} := \mathcal{B}_{\delta_0}$, where we have used the notation of Theorems 3.1 and 5.1. Suppose that the initial iterate satisfies $u_0 \in \mathcal{B}_{3\delta_0/4}$ and

$$\|\mathbf{A}_n(u^0)\| \leq \rho^{-1}. \tag{5.8}$$

Define

$$h = 2C\kappa^2\rho^{-1}, \tag{5.9}$$

and suppose that $h \leq \min\{\frac{1}{2}, \frac{2C\kappa\delta_0}{4}\}$. *For the Newton sequence defined by* (5.6), *the inequalities,*

$$\begin{aligned} \|u_n^k - u_n^{k-1}\| &\leq \kappa \|\mathbf{A}_n(u_n^{k-1})\|, \quad k \geq 1, \qquad (5.10) \\ \|\mathbf{A}_n(u_n^k)\| &\leq \frac{h\rho}{2} \|\mathbf{A}_n(u_n^{k-1})\|^2, \quad k \geq 1, \qquad (5.11) \end{aligned}$$

hold. In particular, the sequence is well-defined in $\mathcal{B}_{\delta_0}$, *and the convergence to* x_n *is described by the error estimate,*

$$\|x_n - u_n^k\| \leq \frac{\theta_k \kappa}{h\rho} \left(\prod_{j=0}^{k} \tau_j^{2^{k-j}} \right) \frac{(1 - \sqrt{1-2h})^{2^k}}{2^k}. \qquad (5.12)$$

Here, $\{\theta_k\}$ *and* $\{\tau_k\}$ *are decreasing sequences bounded by* 1. *For* $k = 1$, *the following (weaker) estimate is implied* :

$$\|x_n - u_n^1\| \leq \frac{h\rho^{-1}}{4} \leq \frac{C\kappa^2\rho^{-2}}{2}. \qquad (5.13)$$

Theorem 5.2. *Assume the Lipschitz continuity and uniform inverse boundedness properties described at the beginning of Lemma* 5.1. *Assume also the boundedness property,*

$$\|[\mathbf{A}_n'(y)]\| \leq \kappa' \qquad (5.14)$$

uniformly for $y \in \mathcal{B}_{\delta_0}$, *Select* n_* *such that, for* $n \geq n_* - 1$,

$$\sqrt{\kappa} \left(\kappa' + \frac{1}{2\kappa} \right) \max \left\{ \frac{1}{2} \left(1 - \sqrt{1 - \alpha_n 2\kappa C} \right), \frac{C\kappa}{\kappa'} \|\mathbf{A}_{n+1}(\mathbf{P}_n x_0)\| \right\} < \frac{1}{2}.$$

Then the numbers q_n, *defined for* $n_* - 1$ *by* (5.3), *and, for* $n \geq n_*$, *by*

$$q_n = \max \left\{ 4q_{n-1}^2 \kappa \left(\kappa' + \frac{1}{2\kappa} \right)^2, \frac{1}{2} \left(1 - \sqrt{1 - \alpha_n 2\kappa C} \right), \frac{C\kappa}{\kappa'} \|\mathbf{A}_{n+1}(\mathbf{P}_n x_0)\| \right\}, \qquad (5.15)$$

satisfy $q_n \searrow 0$. *Here,* n_* *is determined to satisfy the additional conditions* (cf. (5.20)) *that*

$$\begin{aligned} \mathbf{P}_n x_0 \in \mathcal{B}_{\delta_0/2},\ 2\delta_n &\leq \delta_0/4,\ n \geq n_* - 1, \qquad (5.16) \\ \|\tilde{x}_{n_*-1} - \mathbf{P}_{n_*-1} x_0\| &\leq 2\delta_{n_*-1} := \frac{q_{n_*-1}}{\kappa C}, \qquad (5.17) \\ q_n &\leq \frac{1}{8\kappa} \max\{1, (C\kappa\delta_0)^2/2\}, \quad n \geq n_*. \qquad (5.18) \end{aligned}$$

Moreover, a sequence $\{\tilde{x}_n\}$ *can be defined by linear solution procedures, as a replacement for* $\{x_n\}$, *such that, for* $n \geq n_*$,

$$\|\tilde{x}_n - \mathbf{P}_n x_0\| \leq 2\delta_n. \qquad (5.19)$$

Specifically, $\tilde{x}_n$ is the first Newton iterate, based upon the mapping $\mathbf{I}-\mathbf{T}_n$, with starting value, $\tilde{x}_{n-1}$:

$$\tilde{x}_n = \tilde{x}_{n-1} + [\mathbf{I} - \mathbf{T}_n'(\tilde{x}_{n-1})]^{-1}(\mathbf{I} - \mathbf{T}_n)(\tilde{x}_{n-1}). \tag{5.20}$$

Proof. It is clear that the numbers q_n, defined by (5.15), converge to 0. Set

$$\tilde{x}_n = u_n^1,\ n \geq n_*, \tag{5.21}$$

where the Newton iteration is based upon the mapping, $\mathbf{A}_n = \mathbf{I} - \mathbf{T}_n$, with starting value, $u_n^0 = \tilde{x}_{n-1}$. Note that $\tilde{x}_{n-1} \in \mathcal{B}_{3\delta_0/4}$ follows for $n = n_*$, via (5.16) and (5.17), and inductively, via (5.16) and (5.19), for general $n \geq n_*$. In order to employ the previous lemma, set

$$\rho^{-1} := \frac{\sqrt{q_n}}{C\kappa^{3/2}}. \tag{5.22}$$

With this definition, and the definition of h in (5.9), we see that h satisfies the conditions of the lemma, via the hypothesis, (5.18). If the small residual hypothesis, (5.8), can be verified, then the conclusion,

$$\|x_n - \tilde{x}_n\| \leq \delta_n, \tag{5.23}$$

is a consequence of the estimate (5.13), given the definition of ρ in (5.22). The principal conclusion of the theorem, stated in (5.19), follows from (5.23) and the estimate, which is the conclusion of Theorem 5.1,

$$\|\mathbf{P}_n x_0 - x_n\| \leq \delta_n. \tag{5.24}$$

We shall have occasion to use the residual in integral form during the course of the proof of the small residual condition:

$$\mathbf{A}_n(x) = \int_0^1 [\mathbf{A}_n'(y+t(x-y)) - \mathbf{A}_n'(y)](x-y)\ dt + \mathbf{A}_n(y) + \mathbf{A}_n'(y)(x-y). \tag{5.25}$$

Now make the identifications, $x = \tilde{x}_{n-1}$, $y = \mathbf{P}_{n-1}x_0$, to conclude that

$$\begin{aligned}
\|\mathbf{A}_n(\tilde{x}_{n-1})\| &\leq C\|\tilde{x}_{n-1} - \mathbf{P}_{n-1}x_0\|^2 + \|\mathbf{A}_n(\mathbf{P}_{n-1}x_0)\| + \kappa'\|\tilde{x}_{n-1} - \mathbf{P}_{n-1}x_0\| \\
&\leq C(2\delta_{n-1})^2 + \|\mathbf{A}_n(\mathbf{P}_{n-1}x_0)\| + \kappa'(2\delta_{n-1}) \\
&\leq C(2\delta_{n-1})^2 + 2\kappa'(2\delta_{n-1}) \\
&\leq \rho^{-1}.
\end{aligned}$$

Here we have used the induction hypothesis and the fact that the third inequality latent in (5.15) implies that $\|\mathbf{A}_n(\mathbf{P}_{n-1}x_0)\| \leq 2\kappa'\delta_{n-1}$, while the first inequality implies (5.8). This completes the proof. □

References

[1] J.-P. Aubin. Approximation des espaces de distributions et des opérateurs differéntials. *Bull. Soc. Math. France Suppl. Mém.*, 12, 1967.

[2] Ivo Babuška and A.K. Aziz. Survey lectures on the mathematical foundations of the finite element method. In A.K. Aziz, editor, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 5–359. Academic Press, 1972.

[3] J.H. Bramble and S.R. Hilbert. Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.*, 7:112–124, 1970.

[4] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York, 1994.

[5] F. Brezzi. On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers. *R.A.I.R.O. Anal. Numér.*, 12:129–151, 1974.

[6] T. Dupont and R. Scott. Polynomial approximations of functions in Sobolev spaces. *Math. Comp.*, 34:441–463, 1980.

[7] J.W. Jerome. *Approximation of Nonlinear Evolution Systems*. Academic Press, 1983.

[8] J.W. Jerome. Approximate Newton methods and homotopy for stationary operator equations. *Constructive Approximation*, 1:271–285, 1985.

[9] J.W. Jerome. The mathematical study and approximation of semiconductor models. In *Advances in Numerical Analysis: Large Scale Matrix Problems and the Numerical Solution of Partial Differential Equations* (J. Gilbert and D. Kershaw, eds.), pages 157–204. Oxford University Press, 1994.

[10] J.W. Jerome and T. Kerkhoven. A finite element approximation theory for the drift-diffusion semiconductor model. *SIAM J. Numer. Anal.*, 28:403–422, 1991.

[11] A.N. Kolmogorov. Über die beste Annähergung von Funktionen einer gegebenen Funktionklasse. *Ann. Math.*, 37(2):107–111, 1936.

[12] M.A. Krasnosel'skii, G.M. Vainikko, P.P. Zabreiko, Ya.B. Rititskii, and V.Ya. Stetsenko. *Approximate Solution of Operator Equations*. Wolters-Noordhoff, Groningen, 1972.

[13] J. Nitsche. Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numer. Math.*, 11:346–348, 1968.

[14] G.W. Strang and G. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood Cliffs, N.J., 1973.