SIAM REVIEW
Vol. 47, No. 1, pp. 50–66

On the Finite Element Solution of the Pure Neumann Problem*

Pavel Bochev†
R. B. Lehoucq†

Abstract. This paper considers the finite element approximation and algebraic solution of the pure Neumann problem. Our goal is to present a concise variational framework for the finite element solution of the Neumann problem that focuses on the interplay between the algebraic and variational problems. While many of the results that stem from our analysis are known by some experts, they are seldom derived in a rigorous fashion and remain part of numerical folklore. As a result, this knowledge is not accessible (or appreciated) by many practitioners—both novices and experts—in one source. Our paper contributes a simple, yet insightful link between the continuous and algebraic variational forms that will prove useful.

Key words. finite elements, Neumann problem, Rayleigh–Ritz minimization, regularization, quadratic programming

AMS subject classifications. 65N30, 65N22, 65F10, 65F35

DOI. 10.1137/S0036144503426074

1. Introduction.

This paper is concerned with the finite element solution of the pure Neumann problem

$$-\triangle u = f \text{ in } \Omega \quad \text{and} \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \Gamma, \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d, d = 1, 2, 3$, is a bounded open region with boundary Γ. Solutions of (1.1) are determined up to a constant[1] mode. The Fredholm alternative implies that the source f must be orthogonal to this mode, that is,

$$\int_\Omega f(x)dx = 0. \tag{1.2}$$

A direct Galerkin discretization of (1.1)–(1.2) leads to a linear system with similar properties: a stiffness matrix with a one-dimensional kernel and a source term that is orthogonal to this kernel. There are two basic approaches for computing a finite element solution from this system. One, favored by some practitioners, is to constrain

*Received by the editors April 15, 2003; accepted for publication (in revised form) April 7, 2004; published electronically February 1, 2005.
http://www.siam.org/journals/sirev/47-1/42607.html
†Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110 (pbboche@sandia.gov, rblehou@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC04-94AL85000.
[1]In the context of mechanical systems this mode is usually called a *rigid body motion*.

the candidate solution by specifying its value at a node. This eliminates the null-space and allows one to solve the linear system by a conventional (sparse) direct solver.

Alternatively, the solution can be computed from the consistent singular system either by a properly modified direct procedure that recognizes (machine) zero pivot, or by a minimization-based iterative solver such as conjugate gradient. This approach is less popular for three reasons: special purpose direct solvers are required; there is a general aversion towards solving singular systems; and many people are not aware that conjugate gradients work for positive semidefinite consistent linear systems.

In the extant literature, both solution techniques are formulated directly for the discrete problem without any connection to a variational problem. This situation is unsatisfactory because under closer scrutiny both approaches reveal some unsettling details. For instance, specifying solution datum at a node is inherently ambiguous, while roundoff error may render the singular system inconsistent and prevent convergence of the conjugate gradient algorithm. At the same time, many well-regarded finite element method textbooks [2, 13, 6, 19, 9, 22, 21, 20, 12] provide only scant information on these issues. As a rule, engineering texts limit their exposition to a brief, ad hoc discussion of the first approach; see the recent textbook by Gresho and Sani [12, p. 474], or the classic text [2]. Mathematically oriented finite element books, on the other hand, focus on variational problems posed in factor or zero mean spaces [6, 11, 3], but do not discuss the practical details of implementing conforming finite element methods in these settings. As for the second approach, the solution of singular systems by the conjugate gradient algorithm is rarely discussed outside the specialized literature on iterative solvers [1, 15] or sparse direct solvers [17, 10].

The contribution of our paper is threefold. First, we seek to develop a unifying variational framework for the finite element solution of the Neumann problem that embraces existing solution techniques and presents a lucid connection between the algebraic equations and well-posed variational problems. Second, with the aforementioned connection, we present several new results that have not appeared, to the best of our knowledge, in the literature. Third, we address the impact of our choices when using an iterative method of solution instead of the commonly studied impact on (sparse) direct methods for the solution of the linear system.

Since our analysis will recover widely practiced solution techniques, many of the results (and conclusions) in this paper will probably be known to an expert in the mathematical theory of finite elements or an experienced practitioner of the method. Nevertheless, we feel that there is a need to provide both novices and experts with a systematic presentation of the existing body of knowledge. Moreover, our treatment reveals the common variational origins of seemingly different solution techniques, allows for their rigorous mathematical analysis, and suggests new methods.

We mention that our approach can be applied with equal success to other problems where a finite element discretization leads to a matrix with a nontrivial kernel. We have intentionally chosen to focus on the Neumann problem so as to avoid unnecessary technical detail and instead focus on the germane idea.

Finite element solution of the Neumann problem and all ensuing approaches can be completely understood by realizing that there are two variational settings that give well-posed weak problems. Both are related to the energy functional of (1.1),

$$J(v, f) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega f v dx, \tag{1.3}$$

but differ in the type of optimization involved: constrained vs. unconstrained. Without constraints, minimizers of (1.3) are determined up to a constant. The first varia-

tional setting is to factor out the constants and minimize (1.3) on a *factor space*. This leads to finite element methods that require solution of a singular linear system.

If we impose a suitable constraint, then (1.3) will have a unique minimizer in a standard Sobolev space. This is the second variational setting and, depending on how the constraint is introduced and implemented, a number of different methods follow. The standard way to enforce a constraint is to use Lagrange multipliers. We show that the popular solution method of fixing the datum at a point is simply an instance of this technique. Ultimately, solutions of finite element problems obtained by Lagrange multipliers all reduce to variations of the null-space method [16] for equality-constrained quadratic programs.

A saddle-point Lagrangian formulation can also be regularized by relaxing the constraint. This leads to a class of finite element methods that have not been previously documented in the literature. Moreover, we show that these *regularized* finite element formulations have some attractive properties, especially in the context of iterative solution methods.

Throughout the paper we use the standard notation $H^s(\Omega)$ for a Sobolev space of order s with norm and inner product given by $\|\cdot\|_s$ and $(\cdot,\cdot)_s$, respectively. Seminorms on $H^s(\Omega)$ will be denoted by $|\cdot|_k$, $0 \le k \le s$. For example, $|u|_1 = \int_\Omega |\nabla u|^2 dx$. For $s = 0$ we write $L^2(\Omega)$ instead of $H^0(\Omega)$ and denote the resulting inner product by $(\cdot,\cdot)$. We also define the subspace $L^2_0(\Omega)$ of *zero-mean* functions.

Since our study also makes use of matrix theory, we introduce some useful notation. With $\{\mathbf{e}_i\}_{i=1}^N$ and $\mathbf{I}_N$ we denote the canonical basis on $\mathbb{R}^N$ and the identity matrix of order N. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ the standard Euclidean norm and inner product are denoted by $\mathbf{x}^T\mathbf{y}$ and $\|\cdot\|$, respectively. The ordering of the eigenvalues of an $N \times N$ matrix $\mathbf{A}$ is $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$.

We call attention to our specific use of bold font for matrices and vectors. Elements of matrices and vectors will be denoted by lowercase Greek letters.

2. Projections and Inequalities. Two projection operators will play a fundamental role throughout the paper. Let ω be a smooth function such that

$$(1,\omega) > 0 \tag{2.1}$$

and consider the subspace

$$H^1_\omega(\Omega) = \{u \in H^1(\Omega) \,|\, (u,\omega) = 0\} \tag{2.2}$$

of all functions in $H^1(\Omega)$ with zero ω-mean. For any $u \in H^1(\Omega)$ we define the operators

$$\mathcal{P}_\omega u = u - \frac{(u,\omega)}{(1,\omega)} = u - u_\omega, \tag{2.3}$$

where $u_\omega = (u,\omega)/(1,\omega)$ is the normalized ω-mean of u, and

$$\mathcal{P}^*_\omega u = u - \omega\frac{(1,u)}{(1,\omega)}, \tag{2.4}$$

respectively. A direct calculation shows that $\mathcal{P}_\omega u \in H^1_\omega(\Omega)$, $\mathcal{P}^*_\omega u \in H^1(\Omega) \cap L^2_0(\Omega)$, and $\mathcal{P}_\omega$ and $\mathcal{P}^*_\omega$ are projectors. Specifically, $\mathcal{P}_\omega$ is a projector $H^1(\Omega) \mapsto H^1_\omega(\Omega)$ parallel to span(1) and $\mathcal{P}^*_\omega$ is a projector $H^1(\Omega) \mapsto H^1(\Omega) \cap L^2_0(\Omega)$ parallel to span(ω).

LEMMA 2.1. *$\mathcal{P}_\omega$ and $\mathcal{P}_\omega^*$ are adjoint with respect to the L^2 inner product, that is,*

$$(\mathcal{P}_\omega u, v) = (u, \mathcal{P}_\omega^* v). \tag{2.5}$$

Proof.

$$(\mathcal{P}_\omega u, v) = (u - u_\omega, v) = (u, v) - \frac{(u,\omega)}{(1,\omega)}(1, v) = (u, v - \omega(1, v)/(1, \omega)) = (u, \mathcal{P}_\omega^* v). \qquad \square$$

We note that $\mathcal{P}_\omega^* f = 0$ for $\omega = f$ and that $\mathcal{P}_\omega$ is self-adjoint when $\omega = 1$. We end this section with a generalization of a well-known inequality that will prove useful in section 4. The Poincaré–Friedrichs inequality [3] bounds the L^2 norm by the H^1 seminorm for a function in H_0^1. The following generalized version extends this bound to H^1 functions using our notion of the normalized mean of a function.

LEMMA 2.2. *Assume that Ω is simply connected and that $H^1(\Omega) \subset L^2(\Omega)$ with compact imbedding. Then there exists a positive constant C such that*

$$\|\mathcal{P}_\omega u\|_0 \le C|u|_1 \quad \textit{and} \quad \|u\|_0 \le C(|u|_1 + |u_\omega|) \quad \textit{for every } u \in H^1(\Omega). \tag{2.6}$$

Proof. If the first inequality in (2.6) is not true, then there's a sequence $\{u_k\} \subset H^1(\Omega)$ such that $\|u_k\|_1 = 1$, $u_{\omega,k} = 0$, and $|u_k|_1 < 1/k$. This sequence has a subsequence, denoted again by $\{u_k\}$, that converges weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$ to some u. This and $|u_k|_1 < 1/k$ imply that $\nabla u = 0$ almost everywhere in Ω and so $u =$const almost everywhere in Ω. Likewise,

$$\int_\Omega u\omega dx = \lim_{k\mapsto\infty} \int_\Omega u_k \omega dx = 0.$$

By assumption $(1, \omega) > 0$ and so $u \equiv 0$. As a result, $u_k \mapsto 0$ in $H^1(\Omega)$, a contradiction. The second inequality follows by a similar compactness argument. $\square$

Compact imbedding is a standard assumption required by the compactness argument used to demonstrate the existence of the constant. This hypothesis places mild restrictions on the domain Ω. (See [3, p. 32] and [6, pp. 128–130] for further details on compact imbedding and compactness arguments.)

3. Unconstrained Optimization Setting. We consider the problem of minimizing (1.3) over the factor space $H^1(\Omega)/\mathbb{R}$:

$$\min_{\hat{u}\in H^1(\Omega)/\mathbb{R}} J(\hat{u}, f), \tag{3.1}$$

where $f \in L_0^2(\Omega)$ is given and

$$H^1(\Omega)/\mathbb{R} = \{\hat{u} \subset H^1(\Omega) \,|\, u, v \in \hat{u} \Leftrightarrow u - v = C\}. \tag{3.2}$$

$H^1(\Omega)/\mathbb{R}$ is a Hilbert space when equipped with the quotient norm

$$\|\hat{u}\|_{H^1(\Omega)/\mathbb{R}} = \inf_{u\in\hat{u}} \|u\|_1 \tag{3.3}$$

and the mapping $\hat{u} \mapsto |u|_1$, $u \in \hat{u}$, defines a norm equivalent to (3.3) [11, p. 13]. The Euler–Lagrange equation for (3.1) is to seek $\hat{u} \in H^1(\Omega)/\mathbb{R}$ such that

$$\hat{\mathcal{A}}(\hat{u}, \hat{v}) = \hat{F}(\hat{v}) \quad \forall \hat{v} \in H^1(\Omega)/\mathbb{R}, \tag{3.4}$$

where

$$\hat{\mathcal{A}}(\hat{u},\hat{v}) = \mathcal{A}(u,v) := \int_\Omega \nabla u \cdot \nabla v dx, \quad u \in \hat{u},\, v \in \hat{v}, \tag{3.5}$$

is a bilinear form $H^1(\Omega)/\mathbb{R} \times H^1(\Omega)/\mathbb{R} \mapsto \mathbb{R}$, and

$$\hat{F}(\hat{u}) = F(u) := (f,u), \quad u \in \hat{u}, \tag{3.6}$$

is a linear functional $H^1(\Omega)/\mathbb{R} \mapsto \mathbb{R}$. Both (3.5) and (3.6) are well-defined because $\hat{\mathcal{A}}(u_1-u_2,\cdot) = \hat{\mathcal{A}}(\cdot,v_1-v_2) = 0$ and $\hat{F}(u_1-u_2) = C\int_\Omega f dx = 0$ whenever $u_1,u_2 \in \hat{u}$ and $v_1,v_2 \in \hat{v}$. Because $|\cdot|_1$ is equivalent to the quotient norm (3.3), the bilinear form (3.5) is continuous and coercive on the quotient space. Hence (3.4) has a unique solution in $H^1(\Omega)/\mathbb{R}$.

4. Constrained Optimization Setting. To formulate a problem that has a unique minimizer in $H^1(\Omega)$ we will require all minimizers to have a vanishing ω-mean; that is, we consider the problem

$$\min_{u\in H^1(\Omega)} J(u,f) \quad \text{subject to} \quad u_\omega = 0. \tag{4.1}$$

The choice of ω and the handling of the constraint in (4.1) provide a template for all finite element methods for the Neumann problem.

4.1. A Saddle-Point Formulation. We introduce a Lagrange multiplier $\tau \in \mathbb{R}$ and consider the saddle-point optimization problem (see problem 4.21 in [3, p. 140])

$$\inf_{u\in H^1(\Omega)} \sup_{\tau\in\mathbb{R}} \left(J(u,f) + \tau u_\omega\right). \tag{4.2}$$

The saddle-point $(u,\tau) \in H^1(\Omega)\times\mathbb{R}$ of (4.2) solves the first-order optimality system

$$\begin{aligned} \mathcal{A}(u,v) \,+\, \tau v_\omega \;&=\; F(v) \qquad \forall v \in H^1(\Omega), \\ \sigma u_\omega \;&=\; 0 \qquad\quad \forall \sigma \in \mathbb{R}. \end{aligned} \tag{4.3}$$

THEOREM 4.1. *Problem* (4.3) *has a unique solution* (u,τ) *for any* $f \in L^2(\Omega)$.

Proof. We apply the abstract theory of [7] and so we must show that there exists a $\gamma > 0$ for every τ so that the form $b(\tau,u) = \tau u_\omega$ satisfies the inf-sup condition

$$\sup_{u\in H^1(\Omega)} \frac{b(\tau,u)}{\|u\|_1} \geq \gamma|\tau|.$$

We equivalently show that for a given $\tau \in \mathbb{R}$ there exists $u \in H^1(\Omega)$ such that $b(\tau,u) \geq \gamma\|u\|_1|\tau|$. Choosing $u = 1$ gives $\|u\|_1 = \sqrt{\text{meas}(\Omega)}$ and

$$b(\tau,u) = \tau(1,\omega)/(1,\omega) = \tau,$$

and so the inf-sup condition clearly holds with $\gamma = 1/\sqrt{\text{meas}(\Omega)}$. To show that $\mathcal{A}(\cdot,\cdot)$ is coercive on the kernel

$$Z = \{u \in H^1(\Omega) \,|\, b(\tau,u) = 0 \quad \forall \tau \in \mathbb{R}\},$$

we observe that $Z = H^1_\omega(\Omega)$. The generalized Friedrichs inequality (2.6) implies that $|u|_1$ is an equivalent norm on $H^1_\omega(\Omega)$, and because $\mathcal{A}(u,u) = |u|_1^2$, we conclude that

this form is coercive on Z. Existence and uniqueness of a saddle-point (u, τ) now follow from [7]. ∎

Restriction of (4.1) to Z gives the equivalent, unconstrained, *reduced* problem

$$\min_{u \in H^1_\omega(\Omega)} J(u, f). \tag{4.4}$$

The Euler–Lagrange equation of the reduced problem is

$$\text{seek } u \in H^1_\omega(\Omega) \text{ such that } \quad \mathcal{A}(u, v) = F(v) \quad \forall v \in H^1_\omega(\Omega). \tag{4.5}$$

Theorem 4.1 asserts that $\mathcal{A}(\cdot, \cdot)$ is a coercive bilinear form on $H^1_\omega(\Omega) \times H^1_\omega(\Omega)$. Therefore, the Lax–Milgram lemma implies that (4.5) is a well-posed problem for any $f \in L^2(\Omega)$.

In summary, we have the choice of either the saddle-point problem (4.3) or the coercive problem (4.5).

4.2. A Stabilized Saddle-Point Formulation. We can modify (4.2) by stabilizing the Lagrangian functional

$$\inf_{u \in H^1(\Omega)} \sup_{\tau \in \mathbb{R}} \left(J(u, f) + \tau u_\omega - \frac{1}{2\rho} \tau^2 \right), \tag{4.6}$$

where $\rho > 0$ is a stabilizing parameter. The optimality system for (4.6) is to seek $(u, \tau) \in H^1(\Omega) \times \mathbb{R}$ such that

$$\begin{aligned} \mathcal{A}(u, v) + \tau v_\omega &= F(v) \qquad \forall v \in H^1(\Omega), \\ \sigma u_\omega &= \rho^{-1} \sigma \tau \qquad \forall \sigma \in \mathbb{R}. \end{aligned} \tag{4.7}$$

The Lagrange multiplier can be eliminated from (4.7) to obtain the *regularized* problem

$$\mathcal{A}_\rho(u, v) = F(v) \quad \forall v \in H^1(\Omega), \tag{4.8}$$

where

$$\mathcal{A}_\rho(u, v) = \mathcal{A}(u, v) + \rho u_\omega v_\omega = \int_\Omega \nabla u \cdot \nabla v dx + \rho u_\omega v_\omega. \tag{4.9}$$

We remark that (4.8) is a first-order optimality system for the unconstrained minimization of the penalized energy functional:

$$\min_{u \in H^1(\Omega)} \left(J(u, f) + \frac{\rho}{2} u_\omega^2 \right) \equiv \min_{u \in H^1(\Omega)} J_\rho(u, f). \tag{4.10}$$

THEOREM 4.2. *For every $f \in L^2(\Omega)$ problem* (4.10) *has a unique minimizer $u \in H^1(\Omega)$.*

Proof. From (2.6) we see that

$$\mathcal{A}_\rho(u, u) = |u|_1^2 + \rho u_\omega^2 \geq C \|u\|_1^2;$$

that is, (4.9) is coercive on $H^1(\Omega) \times H^1(\Omega)$. Continuity of this form and $F(\cdot)$ are obvious, and so we can conclude that the regularized problem has a unique solution. ∎

Therefore, in the present setting we can choose between the regularized saddle-point problem (4.7) or the coercive problem (4.8).

4.3. Characterization of Solutions. We now consider the relationship between the solutions obtained from the constrained optimization setting and the original Neumann problem. Without stabilization we have the choice of (4.3) or (4.5); with stabilization the choice is between (4.7) or (4.8). However, within each pair the same weak solution u will be generated and so it suffices to consider the two coercive equations, that is, (4.5) and (4.8).

If $f \in L^2_0(\Omega)$, both (4.4) and (4.10) have solutions that belong to a minimizing class of (3.1). However, (3.1) is not well-posed unless $f \in L^2_0(\Omega)$, while the weak problems (4.5) and (4.8) are coercive and solvable for any $f \in L^2(\Omega)$. Our next result shows that when f does not satisfy the compatibility condition (1.2), the solution computed by (4.5) and (4.8) solves a Neumann problem with a modified source term.

THEOREM 4.3. *Let $\tilde{u}$ denote a solution of* (4.8) *(respectively,* (4.5)*). For any $f \in L^2(\Omega)$*

$$\tilde{u}_\omega = \alpha(f, 1), \tag{4.11}$$

where $\alpha = 1/\rho$ for (4.8) *and $\alpha = 0$ for* (4.5)*. If $\tilde{u} \in H^2(\Omega)$, then $\tilde{u}$ solves the Neumann problem*

$$-\triangle u = \mathcal{P}^*_\omega f \quad \textit{in } \Omega \quad \textit{and} \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \textit{on } \Gamma.$$

Proof. For $\tilde{u}$ computed by (4.5), formula (4.11) is trivially true since $\tilde{u} \in H^1_\omega(\Omega)$. To prove (4.11) for (4.8), insert $v = 1$ in (4.8) to obtain

$$(f, 1) = \frac{\rho}{(1,\omega)^2}(\tilde{u}, \omega)(1, \omega) = \rho \tilde{u}_\omega.$$

Let $\tilde{u} \in H^2(\Omega)$ solve (4.8). Integrating (4.8) by parts gives

$$\left(-\triangle \tilde{u} - f + \omega \frac{\rho}{(1,\omega)^2}(\tilde{u}, \omega), v\right) + \left\langle \frac{\partial \tilde{u}}{\partial \mathbf{n}}, v \right\rangle_\Gamma = 0 \quad \forall v \in H^1(\Omega).$$

From (4.11) $\rho(\tilde{u}, \omega)/(1, \omega)^2 = (f, 1)/(1, \omega)$, so $v \in H^1_0(\Omega)$ implies

$$-\triangle \tilde{u} - (f - \omega(f, 1)/(1, \omega)) = -\triangle \tilde{u} - \mathcal{P}^*_\omega f = 0.$$

Choosing $v \neq 0$ on Γ recovers the Neumann boundary condition.

Let $\tilde{u} \in H^2(\Omega)$ denote a solution of (4.5). Since $\mathcal{P}_\omega v \in H^1_\omega(\Omega)$,

$$\mathcal{A}(\tilde{u}, \mathcal{P}_\omega v) = F(\mathcal{P}_\omega v) \quad \forall v \in H^1(\Omega).$$

From the definition of $\mathcal{A}(\cdot, \cdot)$, (2.3), and Lemma 2.1,

$$\mathcal{A}(\tilde{u}, \mathcal{P}_\omega v) = \mathcal{A}(\tilde{u}, v) \quad \text{and} \quad (f, \mathcal{P}_\omega v) = (\mathcal{P}^*_\omega f, v),$$

and so

$$\mathcal{A}(\tilde{u}, v) = (\mathcal{P}^*_\omega f, v) \quad \forall v \in H^1(\Omega).$$

Integrating this identity by parts gives

$$(-\triangle \tilde{u} - \mathcal{P}^*_\omega f, v) + \left\langle \frac{\partial \tilde{u}}{\partial \mathbf{n}}, v \right\rangle_\Gamma = 0 \quad \forall v \in H^1(\Omega).$$

The theorem follows by first choosing $v \in H^1_0(\Omega)$ and then $v \in H^1(\Omega)$. ∎

COROLLARY 4.4. *If $f \in L_0^2(\Omega)$, solutions of the reduced and regularized problems coincide.*

Proof. Let u^R solve (4.8). From (4.11) it is clear that $u_\omega^R = 0$ whenever $f \in L_0^2(\Omega)$; that is, $u^R \in H_\omega^1(\Omega)$. Now it is easy to see that u^R also satisfies the weak problem (4.5). ∎

5. Finite Element Solution. Throughout this section $\mathcal{T}$ denotes a uniformly regular triangulation of Ω into finite elements. For brevity we restrict our attention to planar regions, triangular elements, and Lagrangian finite element spaces P^k; see [3] for details. The coefficient vector of $u_h \in P^k$ with respect to a nodal basis $\{\phi_i^h\}_{i=1}^N$ is denoted by $\mathbf{u}$.

Formulation of finite element methods is based on the link between the optimization and the Neumann problems established in sections 3–4. Thus, we identify the finite element solution of (1.1) with the computation of approximate minimizers or saddle-points out of some P^k. To state the algebraic problems that will arise in the solution process we shall need the symmetric positive semidefinite stiffness matrix $\mathbf{A}$ with element i, j:

$$\mathbf{A}_{i,j} = \mathcal{A}(\phi_j^h, \phi_i^h), \qquad i,j = 1,\dots,N. \tag{5.1}$$

We denote the jth column of $\mathbf{A}$ by $\mathbf{A}_j$; $\mathbf{f}_i = F(\phi_i^h)$ is the ith element of the discrete source term $\mathbf{f}$, and $\mathbf{w}_i = (\phi_i^h, \omega)$ is the weighted basis mean vector. When $\omega = 1$ we will use $\mathbf{z}$ instead of $\mathbf{w}$. For a nodal basis $\mathbf{A}$ has a kernel spanned by the constant vector $\mathbf{c} = (1,\dots,1)^T$. If $\mathbf{M}$ is the mass matrix with element $\mathbf{M}_{i,j} = (\phi_j^h, \phi_i^h)$, the relationships $\mathbf{z} = \mathbf{M}\mathbf{c}$ and $(u_h, v_h) = \mathbf{u}^T\mathbf{M}\mathbf{v}$ hold. The last expression is the discrete $L^2(\Omega)$ inner product of u_h and v_h.

5.1. Finite Elements in the Unconstrained Setting. In mathematical finite element texts the use of (3.4) as a well-posed weak form for the Neumann problem is standard. In contrast, this setting has found limited acceptance among practitioners because formally a finite element subspace $P^k/\mathbb{R}$ of $H^1(\Omega)/\mathbb{R}$ is required, formulation of the ensuing method is never clarified, and the matrix problem is singular. However, the ambiguities of a factor space setting can be easily avoided within the optimization framework. Since $P^k/\mathbb{R}$ is isomorphic to $\mathbb{R}^N/(\ker(\mathbf{A}) \equiv \mathbf{c})$ the discrete equivalent of (3.1) and its algebraic form are

$$\min_{\hat{u}^h \in P^k/\mathbb{R}} J(\hat{u}^h, f) \equiv \min_{\hat{\mathbf{u}} \in \mathbb{R}^N/\mathbf{c}} \frac{1}{2}\hat{\mathbf{u}}^T\mathbf{A}\hat{\mathbf{u}} - \hat{\mathbf{u}}^T\mathbf{f}. \tag{5.2}$$

Therefore, a finite element method in the factor space setting simply amounts to computation of an arbitrary member from the minimizing class $\hat{\mathbf{u}}^h$. Such a member can be determined by solving the linear system

$$\mathbf{A}\hat{\mathbf{u}} = \mathbf{f} \tag{5.3}$$

by a sparse direct method modified so that a zero pivot can be detected. However, floating point arithmetic complicates this decision because the solver needs to decide when a pivot is negligible. Instead we recommend that an iterative scheme be applied directly to (5.2). Indeed, as long as the $\mathbf{f}$ is in the range of $\mathbf{A}$ the quadratic functional in (5.2) has a finite lower bound. As a result, the conjugate gradient algorithm will generate a minimizing sequence that converges modulo $\ker(\mathbf{A})$; see Theorem 13.11 in [1, p. 583]. The rate of convergence of the conjugate gradient algorithm depends on the ratio $\kappa_\epsilon(\mathbf{A}) = \lambda_N(\mathbf{A})/\lambda_2(\mathbf{A})$ or the effective condition number.

Table 5.1 *Loss of consistency in* (5.3) *within CG.* $\mathbf{x}^{(j)}$ *denotes the CG solution at the* j*th iteration.*

P^1 elements, nonuniform				P^2 elements, nonuniform		
Quad.	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$
1	−2.522E-03	1000	0.38000	−1.857E-03	underintegrated $\mathbf{A}$	
3	3.069E-05	208	0.1242E-05	1.610E-04	1000	0.2128E+02
7	−2.744E-09	85	0.1329E-05	−2.023E-08	169	0.8327E-06
P_1 elements, uniform				P_2 elements, uniform		
3	4.628E-15	55	0.9786E-05	−3.123E-16	54	0.7439E-06

An important practical consideration for (5.3) is that the discrete source $\mathbf{f}$ must be discretely orthogonal to the constant vector $\mathbf{c}$ and $\mathbf{A}\mathbf{c} = \mathbf{0}$. Since $(1, f) = \mathbf{c}^T\mathbf{f}$ in exact arithmetic, the linear system will be consistent whenever the Neumann problem is solvable, that is, when f has zero mean. In practice the source $\mathbf{f}$ and the matrix $\mathbf{A}$ are computed in floating point arithmetic via quadrature. As a result, $\mathbf{c}^T\mathbf{f}$ equals $(1, f)$ only approximately and (5.3) may become inconsistent. To restore consistency we take a cue from Theorem 4.3 and introduce the discrete projector $(\mathbf{P}^T\mathbf{f})_i \equiv (\mathcal{P}_\omega^* f, \phi_i^h)$, $i = 1, \ldots, N$. A direct calculation shows that

$$\mathbf{P}^T = \mathbf{I} - \frac{\mathbf{w}\mathbf{c}^T}{\mathbf{w}^T\mathbf{c}}.$$

Application of the projector to the linear system results in

$$(\mathbf{P}^T\mathbf{A}\mathbf{P})\mathbf{u} = \mathbf{P}^T\mathbf{f}. \tag{5.4}$$

The matrix $\mathbf{P}$ is the discrete analogue of the projector $\mathcal{P}_\omega$ and so the finite element solution $\mathbf{P}\mathbf{u}$ has zero ω-mean; that is, $\mathbf{w}^T\mathbf{P}\mathbf{u} = 0$. We remark that the iterative solution of semidefinite systems and application of projectors is rarely discussed beyond specialized texts on iterative solvers and does not seem to be widely known among finite element practitioners. This is another reason for the limited use of (5.3).

Let us demonstrate that the use of a projector to maintain consistency of (5.3) is not unfounded, especially for unstructured meshes. To test the effects of numerical quadrature we consider the zero mean source f defined by evaluating (1.1) at $u(x, y) = \cos(\pi x^2)\cos(2\pi y)$ on the unit square. We solve (5.3) with discrete sources $\mathbf{f}$ computed using linear (one-point), quadratic (three-point), and quintic (seven-point) quadrature rules [8, p. 343].

Table 5.1 shows that for P^2 elements on nonuniform meshes, the three-point rule leads to a numerically inconsistent linear system and so the conjugate gradient algorithm diverges. For nonuniform P^1 elements the three-point rule does suffice but requires 2.5 times more conjugate gradient iterations than the seven-point rule.

On uniform grids all three quadrature rules led to a discrete source $\mathbf{f}$ with *exact* zero mean and a consistent (to machine precision) linear system. Table 5.1 shows that in this case conjugate gradients converged without a difficulty. This contrasting behavior clearly demonstrates the importance of maintaining consistency in (5.3).

5.2. Finite Elements in the Constrained Setting. The starting point now is the constrained problem (4.1). To define a finite element solution we restrict minimization of (4.1) to a subspace P^k of $H^1(\Omega)$ and note that $u_{h,\omega} = 0$ if and only if $\mathbf{u}^T\mathbf{w} = 0$.

As a result, the discrete equivalent of (4.1) and its algebraic form is

$$\min_{\substack{u_h \in P^k \\ u_{h,\omega}=0}} J(u_h, f) \equiv \min_{\substack{\mathbf{u}\in\mathbb{R}^N \\ \mathbf{w}^T\mathbf{u}=0}} \frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u} - \mathbf{u}^T\mathbf{f}. \tag{5.5}$$

In the optimization literature (5.5) is known as an equality-constrained quadratic program [16]. This quadratic program can be solved in a number of ways. In all cases, however, we are led to an algebraic equation that is related to one of the four variational problems (4.3), (4.5), (4.6), or (4.8). In what follows we consider the variational settings of sections 4.1 and 4.2 and demonstrate their relationship with (5.5).

5.2.1. The Saddle-Point Formulation. The algebraic equivalent of the saddle-point equation (4.3) is a symmetric, indefinite linear system:

$$\begin{pmatrix} \mathbf{A} & (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w} \\ (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \tau \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix}. \tag{5.6}$$

This system can be obtained directly from (5.5) by introducing a Lagrange multiplier for the algebraic constraint. The matrix in (5.6) is called the Karush–Kuhn–Tucker (KKT) matrix. One way to compute a finite element approximation is to solve (5.6) by either a sparse direct method or an iterative method. Another approach that exploits the structure in the KKT matrix is the *null-space* method. The alternative *range-space* method requires that $\mathbf{A}$ be nonsingular and not be applicable to (5.6).

The constraint $\mathbf{w}^T\mathbf{u} = 0$ implies that the minimizer belongs to the subspace $\text{span}(\mathbf{w})^\perp$ or, equivalently, the null-space of $\mathbf{w}^T$. Let $\mathbf{B} \in \mathbb{R}^{N\times(N-1)}$ denote a matrix whose columns form a basis for $\text{span}(\mathbf{w})^\perp$. Then $\mathbf{u} = \mathbf{B}\mathbf{v}$ and (5.5) is equivalent to an unconstrained problem

$$\min_{\mathbf{v}\in\mathbb{R}^{N-1}} \frac{1}{2}\mathbf{v}^T\mathbf{B}^T\mathbf{A}\mathbf{B}\mathbf{v} - \mathbf{v}^T\mathbf{B}^T\mathbf{f} \tag{5.7}$$

in terms of $\mathbf{v}$. The null-space method for (5.6) amounts to constructing the matrix $\mathbf{B}$ and solving the symmetric positive definite linear system

$$\mathbf{B}^T\mathbf{A}\mathbf{B}\mathbf{v} = \mathbf{B}^T\mathbf{f}. \tag{5.8}$$

The null-space method is the variational equivalent of "minimization on the kernel" that produced the reduced problem (4.4). Let us show that a conforming discretization of (4.4) is in turn equivalent to an explicit method for constructing the matrix $\mathbf{B}$. For this purpose we restrict (4.4) to a finite element subspace $P^k_\omega = P^k \cap H^1_\omega(\Omega)$ of $H^1_\omega(\Omega)$. Since P^k_ω is isomorphic with $\mathbb{R}^{N-1}$, the discrete minimization problem and its algebraic form are

$$\min_{u_h\in P^k_\omega} J(u_h, f) \equiv \min_{\mathbf{v}\in\mathbb{R}^{N-1}} \frac{1}{2}\mathbf{v}^T\mathbf{A}_\omega\mathbf{v} - \mathbf{v}^T\mathbf{f}_\omega, \tag{5.9}$$

where $\mathbf{A}_\omega$, $\mathbf{f}_\omega$, and $\mathbf{v}$ denote a stiffness matrix, right-hand side, and a coefficient vector relative to some basis $\{\psi_i\}_{i=1}^{N-1}$ of P^k_ω. Let $\mathbf{B}$ denote the transformation matrix between this basis and the standard nodal basis of P^k. If $\mathbf{u}$ contains the nodal coefficients of u_h relative to P^k and $\mathbf{v}$ are the coefficients of this function relative to the basis in P^k_ω, then

$$\mathbf{u} = \mathbf{B}\mathbf{v} \quad \text{and} \quad \mathbf{A}_\omega = \mathbf{B}^T\mathbf{A}\mathbf{B}.$$

Because $\mathcal{A}(\cdot,\cdot)$ is coercive on $H^1_\omega(\Omega)\times H^1_\omega(\Omega)$ and $P^k_\omega \subset H^1_\omega(\Omega)$, the matrix $\mathbf{A}_\omega$ is symmetric and positive definite.

In general, $\{\psi_i\}_{i=1}^{N-1}$ need not be a nodal basis. In this case the coefficients in $\mathbf{v}$ are linear combinations of the nodal values in $\mathbf{u}$. Because nodal bases are easier to work with we demonstrate how we can construct such a basis for a given weight ω. Suppose that[2] $(\phi^h_\ell,\omega)\neq 0$ for some ℓ between 1 and N. Solving $(u_h,\omega)\equiv\sum_{i=1}^N \alpha_i(\phi^h_i,\omega)=0$ for the ℓth term gives the set of functions

$$\psi^h_{i,\ell} = \phi^h_i - \phi^h_\ell \frac{(\phi^h_i,\omega)}{(\phi^h_\ell,\omega)}, \quad i=1,\dots,N, \quad i\neq\ell, \tag{5.10}$$

parameterized by ℓ, and a space $P^k_\omega = \operatorname{span}\{\psi^h_{i,\ell}\}_{i\neq\ell}\subset H^1_\omega(\Omega)$. Note that P^k_ω does not have a degree of freedom associated with the (arbitrarily chosen) triangulation node x_ℓ. By definition $(\psi^h_{i,\ell},\omega)=0$ and $\psi^h_{i,\ell}(x_j)=\delta_{ij}$, and so (5.10) is a nodal basis. A straightforward calculation shows that the transformation matrix for the basis (5.10) is

$$\mathbf{B}_{\ell,\omega} = \Big(\mathbf{I} - \frac{\mathbf{e}_\ell \mathbf{w}^T}{\mathbf{e}_\ell^T\mathbf{w}}\Big)\mathbf{I}^\ell_N.$$

Now consider the situation where $\mathbf{w}=\mathbf{e}_\ell$, so that the constraint in (5.5) is $\mathbf{e}_\ell^T\mathbf{u}=0$. In this case, the transformation matrix is

$$\mathbf{B} = \Big(\mathbf{I} - \mathbf{e}_\ell\mathbf{e}_\ell^T\Big)\mathbf{I}^\ell_N = \mathbf{I}^\ell_N.$$

Therefore, $\mathbf{A}_\omega$ is simply $\mathbf{A}$ with deleted ℓth row and column. Note that $\mathbf{e}_\ell^T\mathbf{u}=0$ is the same as $u_h(x_\ell)=0$ and so this is simply the standard method of specifying the solution value at a node. Therefore, this commonly used technique turns out to be a variant of the null-space method.

Our framework allows us to establish an interesting link between the linear system and the variational equation. Let ϕ^h_ℓ denote the basis function associated with node x_ℓ in some triangulation $\mathcal{T}_h$, and consider a weight function $\omega_{h,\ell}$ such that

$$(\phi^h_\ell,\omega_{h,\ell}) = 1 \quad \text{and} \quad (\phi^h_k,\omega_{h,\ell}) = 0 \quad \text{for } k\neq\ell. \tag{5.11}$$

Then $\mathbf{w}=\mathbf{e}_\ell$ and fixing the solution value can be viewed as a conforming discretization of the saddle-point (4.3) or the reduced (4.5) problem with a constraint given by

$$(\omega_{h,\ell},u)=0.$$

While the choice of $\omega_{h,\ell}$ is not unique, (5.11) formally implies that $\omega_{h,\ell}\mapsto\delta(x_\ell)$ as $\mathcal{T}_h$ is refined. Because the delta function is in the dual of $H^1(\Omega)$ only in one dimension, this constraint will become ill-posed in two and three dimensions as $h\to 0$. We conclude that specifying the solution at a node leads to an ill-posed variational problem in two and three dimensions and so impacts the resulting linear system.

The following discrete Poincaré inequality available in the domain decomposition literature is the key result for understanding the effect of using (5.11) upon the resulting stiffness matrix.

[2]This assumption is necessary because Lagrangian basis functions may have zero mean. One example is given by the P^2 basis functions associated with the nodes of a triangulation.

LEMMA 5.1. *Let $u_h \in P^k \subset H^1(\Omega)$, where Ω is the unit volume in $\mathbb{R}^d, d = 1, 2, 3$. If $u_h(p) = 0$ for some point $p \in \Omega$, then*

$$\frac{\|u_h\|_0^2}{|u_h|_1^2} \leq \begin{cases} C, & d = 1, \\ C(1 + |\ln h|), & d = 2, \\ Ch^{-1}, & d = 3 \end{cases} \tag{5.12}$$

for a constant C independent of h.

Proof. For $d = 1$, Sobolev imbedding implies that $\|u_h\|_\infty \leq C|u_h|$. For $d = 2$, Lemma 3.4 of [4] implies that $\|u_h\|_\infty \leq C(1 + |\ln h|)|u_h|$. For $d = 3$, Lemma 2.3 of [5] and Lemma 2.2 imply that

$$\|\mathcal{P}_\omega u_h\|_\infty^2 \leq C_1 h^{-1}(\|\mathcal{P}_\omega u_h\|_0^2 + |u_h|_1^2) \leq C_1(C_2 + 1)h^{-1}|u_h|_1$$

for constants C_1, C_2. The hypothesis gives that $u_h(p) = 0$ and so $|u_\omega| \leq \|\mathcal{P}_\omega u_h\|_\infty$ and the triangle inequality gives $\|u_h\|_\infty \leq 2C_1(C_2 + 1)h^{-1}|u_h|_1 \equiv Ch^{-1}|u_h|_1$.

Finally, $\|u_h\|_0 \leq \|1\|_0\|u_h\|_\infty = \|u_h\|_\infty$ for $d = 1, 2, 3$ and the lemma is proved. ∎

An elementary result established in finite element theory is that the condition number of the stiffness matrix is proportional to h^{-2} for a conforming discretization of the Laplacian with homogeneous Dirichlet boundary conditions (for instance, see [14, pp. 141–142]). Lemma 5.1 allows us to easily modify this elementary proof to determine the condition number of the stiffness matrix that arises by specifying a nodal value of the solution.

THEOREM 5.2. *Assume the same hypothesis of Lemma* 5.1 *and let $\mathbf{A}_\omega$ be the stiffness matrix relative to some basis $\{\psi_i\}_{i=1}^{N-1}$ of P_ω^k. If $\omega \equiv \omega_{h,\ell}$ is defined by* (5.11)*, then for a constant C independent of h*

$$\kappa(\mathbf{A}_\omega) \leq \begin{cases} Ch^{-2}, & d = 1, \\ C(1 + |\ln h|)h^{-2}, & d = 2, \\ Ch^{-3}, & d = 3, \end{cases} \tag{5.13}$$

where $\kappa(\mathbf{A}_\omega)$ is the condition number of $\mathbf{A}_\omega$.

The numerical experiments in section 5.2.2 show numerical support for this theorem.

5.2.2. The Stabilized Saddle-Point Formulation. The linear system

$$\begin{pmatrix} \mathbf{A} & (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w} \\ (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w}^T & -\rho^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \tau \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix} \tag{5.14}$$

is the algebraic equivalent of the stabilized saddle-point problem (4.6). As in section 4.2, the Lagrange multiplier can be eliminated to obtain a system only in terms of $\mathbf{u}$:

$$\mathbf{A}_\rho \mathbf{u} \equiv \left(\mathbf{A} + \frac{\rho}{(\mathbf{w}^T\mathbf{c})^2}\mathbf{w}\mathbf{w}^T\right)\mathbf{u} = \mathbf{f}. \tag{5.15}$$

This equation is the necessary condition for the quadratic program

$$\min_{u_h \in P^k} J_\rho(u_h, f) \equiv \min_{\mathbf{u} \in \mathbb{R}^N} \frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u} - \mathbf{u}^T\mathbf{f} + \rho\frac{(\mathbf{w}^T\mathbf{u})^2}{(\mathbf{w}^T\mathbf{c})^2}, \tag{5.16}$$

Table 5.2 *CG solution of* (5.15). $\mathbf{x}^{(j)}$ *denotes the CG solution at the* j*th iteration.*

P_1 elements, nonuniform				P_2 elements, nonuniform		
Quad.	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$
1	−2.522E-03	85	0.1358E-05	−1.857E-03	underintegrated $\mathbf{A}$	
3	3.069E-05	85	0.1242E-05	1.610E-04	169	0.8667E-06
7	−2.744E-09	85	0.1329E-05	−2.023E-08	169	0.8327E-06

which is a discrete counterpart of (4.10). From Theorem 4.2 it follows that $\mathbf{A}_\rho$ is symmetric and positive definite. The sparsity pattern of $\mathbf{A}_\rho$ depends on the choice of ω because $\mathbf{A}_\rho$ is a rank-1 correction of the singular matrix $\mathbf{A}$. If the support of ω overlaps with only a few elements in $\mathcal{T}_h$, the vector $\mathbf{w}$ will have only a few nonzero entries and $\mathbf{A}_\rho$ will have a sparsity pattern similar to that of $\mathbf{A}$. In this case a sparse direct solver can be used.

When the support of ω is larger, for instance, if $\omega = 1$, then $\mathbf{w}\mathbf{w}^T$ is dense, and formally $\mathbf{A}_\rho$ is also dense. While a direct elimination is not practical in this case, (5.15) can be solved iteratively for almost the same cost as (5.3). Typically, an iterative solver requires one matrix-vector product $\mathbf{A}_\rho\mathbf{u}$ per iteration. This product can be computed by

1. forming the vector $\mathbf{v} = \mathbf{A}\mathbf{u}$;
2. computing the scalar $\mu = \rho(\mathbf{w}^T\mathbf{u})$;
3. updating $\mathbf{v} + \mu\mathbf{w}$.

Step 1 is a standard part of any finite element solver, so the only additional work involved is the dot product in step 2 ($2N-1$ flops) and the update in step 3 ($2N$ flops). The row vector $\mathbf{w}^T$ can be precomputed and stored, rendering the computation of μ efficient.

Theorem 4.2 also implies that the regularized system (5.15) must be solvable for any discrete source $\mathbf{f}$. This means that iterative solver performance should not degrade as in Table 5.1 for low-order quadrature. Table 5.2 contains convergence history for preconditioned conjugate gradients with $\omega = 1$ and $\rho = h^2$ applied to (5.15) and the same exact solution as in Section 5.1. Regardless of the quadrature we see identical convergence of the solver.

The following theorem proves fundamental for understanding the structure of $\mathbf{A}_\rho$ and how the rank-1 update modifies the null-space of $\mathbf{A}$.

THEOREM 5.3. *Let* $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ *denote the eigendecomposition of the singular stiffness matrix* $\mathbf{A}$*, with* $\|\mathbf{c}\|\mathbf{Q}\mathbf{e}_1 = \mathbf{c}$ *and* $\mathbf{A}\mathbf{c} = \mathbf{0}$*. If* $\mathbf{w} = (\|\mathbf{w}\|/\|\mathbf{c}\|)\cos(\phi)\mathbf{c} + \mathbf{r}$*, where* $\mathbf{r}^T\mathbf{c} = 0$ *and* ϕ *measures the positive angle between* $\mathbf{c}$ *and* $\mathbf{w}$*, then*

$$\left\|\mathbf{A}_\rho - \mathbf{Q}\left(\mathbf{\Lambda} + \frac{\rho}{\|\mathbf{c}\|^2}\mathbf{e}_1\mathbf{e}_1^T\right)\mathbf{Q}^T\right\| \le \frac{\rho}{\|\mathbf{c}\|^2}\tan(\phi)(2+\tan(\phi)). \tag{5.17}$$

Proof. From the identity

$$\mathbf{A}_\rho = \mathbf{A} + \frac{\rho}{(\mathbf{w}^T\mathbf{c})^2}\mathbf{w}\mathbf{w}^T = \mathbf{Q}\left(\mathbf{\Lambda} + \frac{\rho}{(\mathbf{w}^T\mathbf{c})^2}(\mathbf{Q}^T\mathbf{w})(\mathbf{Q}^T\mathbf{w})^T\right)\mathbf{Q}^T \tag{5.18}$$

and the hypothesis on $\mathbf{Q}\mathbf{e}_1$, we have

$$\mathbf{Q}^T\mathbf{w} = \mathbf{Q}^T\Big((\|\mathbf{w}\|/\|\mathbf{c}\|)\cos(\phi)\mathbf{c} + \mathbf{r}\Big) = \|\mathbf{w}\|\cos(\phi)\mathbf{e}_1 + \mathbf{Q}^T\mathbf{r}$$

and hence

$$\begin{aligned}\mathbf{Q}^T\mathbf{w}(\mathbf{Q}^T\mathbf{w})^T &= \Big(\|\mathbf{w}\|\cos(\phi)\mathbf{e}_1 + \mathbf{Q}^T\mathbf{r}\Big)\Big(\|\mathbf{w}\|\cos(\phi)\mathbf{e}_1 + \mathbf{Q}^T\mathbf{r}\Big)^T \\ &= \|\mathbf{w}\|^2\cos^2(\phi)\mathbf{e}_1\mathbf{e}_1^T + \mathbf{Q}^T\mathbf{r}(\mathbf{Q}^T\mathbf{r})^T \\ &\quad + \|\mathbf{w}\|\cos(\phi)\Big(\mathbf{Q}^T\mathbf{r}\mathbf{e}_1^T + \mathbf{e}_1\mathbf{r}^T\mathbf{Q}\Big).\end{aligned}$$

Using (5.18), the previous expression and the easily established relationships

$$\|\mathbf{Q}^T\mathbf{r}(\mathbf{Q}^T\mathbf{r})^T\| = \|\mathbf{r}\|^2, \quad \|\mathbf{Q}^T\mathbf{r}\mathbf{e}_1^T\| = \|\mathbf{r}\| = \|\mathbf{w}\|\sin(\phi), \quad |\mathbf{w}^T\mathbf{c}| = \|\mathbf{c}\|\|\mathbf{w}\|\cos(\phi)$$

give

$$\mathbf{A}_\rho - \mathbf{Q}(\mathbf{\Lambda} + \tfrac{\rho}{(\mathbf{w}^T\mathbf{c})^2}\|\mathbf{w}\|^2\cos^2(\phi)\mathbf{e}_1\mathbf{e}_1^T)\mathbf{Q}^T = \mathbf{A}_\rho - \mathbf{Q}(\mathbf{\Lambda} + (\rho/\|\mathbf{c}\|^2)\mathbf{e}_1\mathbf{e}_1^T)\mathbf{Q}^T,$$

and so finally

$$\begin{aligned}\|\mathbf{A}_\rho - \mathbf{Q}(\mathbf{\Lambda} + (\rho/\|\mathbf{c}\|^2)\mathbf{e}_1\mathbf{e}_1^T)\mathbf{Q}^T\| &\le \tfrac{\rho\|\mathbf{w}\|^2}{(\mathbf{w}^T\mathbf{c})^2}(2|\cos(\phi)\sin(\phi)| + \sin^2(\phi)) \\ &= (\rho/\|\mathbf{c}\|^2)\tan(\phi)(2 + \tan(\phi))\end{aligned}$$

and the theorem is proved. ∎

This theorem shows that with a proper choice of ρ the rank-1 update modifies the zero eigenvalue of $\mathbf{A}$ to a positive one and only perturbs the eigenvectors. If ρ is at least as large as $\|\mathbf{c}\|^2\lambda_2$ and $\rho\tan(\phi) < \|\mathbf{c}\|^2$, then the condition number of $\mathbf{A}_\rho$ equals the effective condition number of $\mathbf{A}$. Note that as $\mathbf{w}$ improves as an approximation for the null-space vector $\mathbf{c}$, then ϕ is small and a ρ can be selected larger than λ_2. A final remark is that, given the previous remarks, the selection of the parameter ρ is robust because $\lambda_2 \approx h^d$, where $d = 1, 2, 3$ (see [18, pp. 195–196]).

Recall that Theorem 5.2 implies condition numbers higher than the effective condition number whenever a solution is being specified at a point. This can be confirmed by comparing the conjugate gradient convergence of (5.15) (with $\omega = 1$ and $\rho = h^2, h^3$, in two and three dimensions) and (5.8) when $\mathbf{B} = \mathbf{I}_N^\ell$. Figures 5.1–5.2 show the results when the pure Neumann problem is solved on the unit square and on the unit cube by bilinear and trilinear finite elements, respectively. The zero mean sources $\triangle u(x, y) = \triangle\cos(\pi x^2)\cos(2\pi y)$ and $\triangle u(x, y, z) = \triangle\cos(\pi x^2)\cos(2\pi y)\cos(z^3\pi)$ are used. The choices of $\mathbf{B} = \mathbf{I}_N^\ell$ correspond to specifying the center and the corner nodal coefficients. Figures 5.1–5.2 reveal a substantial and growing gap between the iteration counts for the regularized approach and that of specifying the solution at a point. Figure 5.3, on the other hand, shows that with respect to the mesh size this gap grows faster in three dimensions, and hence supports the conclusion of Theorem 5.2.

6. Conclusions. We have demonstrated that finite element methods for the Neumann problem originate from two optimization settings. The first requires minimization of a quadratic energy functional on a factor space and leads to singular linear systems. These systems can be solved iteratively provided consistency is maintained by a discrete projector to ensure that the source remains discretely orthogonal to the constant mode.

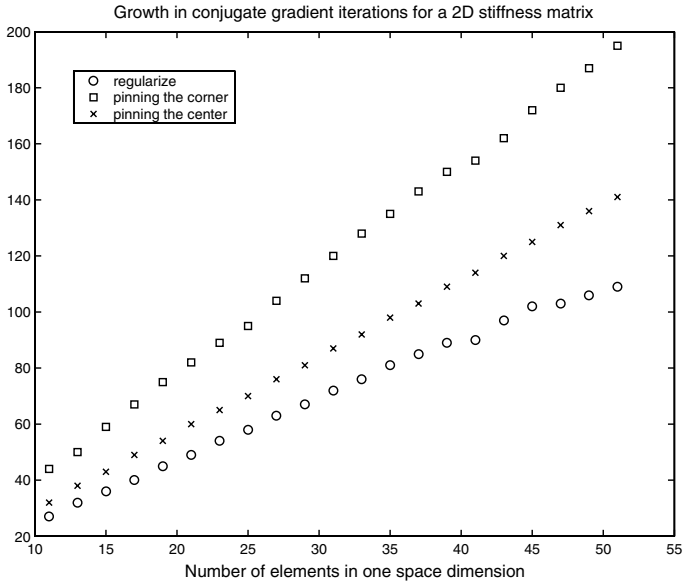

Fig. 5.1 *Growth in conjugate gradient iterations for the solution of* (5.15) *vs.* (5.8) *for a stiffness matrix from a bilinear quadrilateral approximation of a two-dimensional problem. Two choices of* $\mathbf{B} = \mathbf{I}_N^\ell$ *corresponding to specifying the center and corner nodal coefficient are used.*

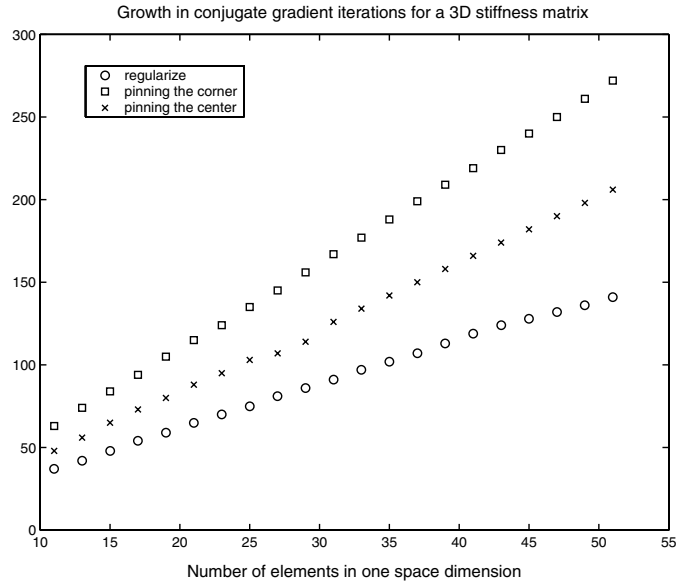

Fig. 5.2 *Growth in conjugate gradient iterations for the solution of* (5.15) *vs.* (5.8) *for a stiffness matrix from a trilinear quadrilateral approximation of a three-dimensional problem. Two choices of* $\mathbf{B} = \mathbf{I}_N^\ell$ *corresponding to specifying the center and corner nodal coefficient are used.*

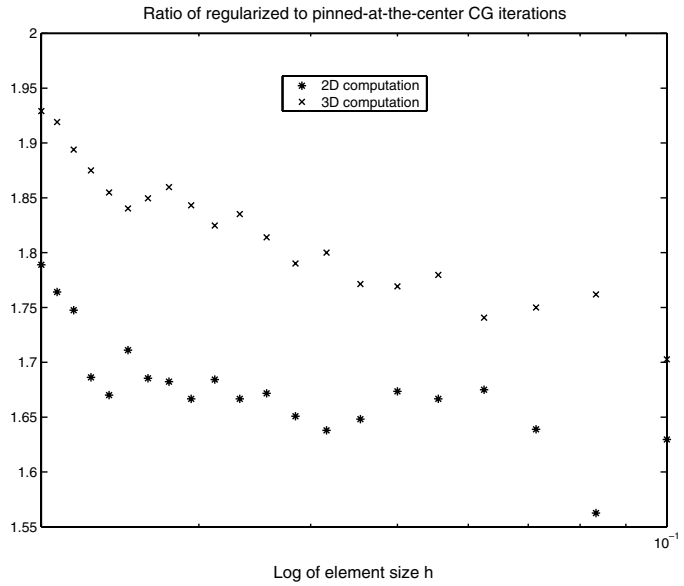

Fig. 5.3 *Ratio of the number of regularized to pinned-at-the-center conjugate gradient iterations.*

The second optimization setting involves constrained minimization of a quadratic functional and leads to an equality-constrained quadratic program. The manner in which the constraint is treated defines yet another two classes of finite element methods, while the choice of the constraint describes the different methods within each class.

The first class corresponds to the application of the null-space method for the solution of the quadratic program. The method of specifying a solution value at a node is an instance of this class. Moreover, we established that this method can be associated with a variational formulation involving a weight function approaching a delta function as $h \to 0$. As a result, condition numbers of the resulting matrices are larger than the effective condition number of the singular matrix.

The second class of finite element methods corresponds to a regularized formulation of the constrained minimization problem. Here we were led to a new class of methods for the Neumann problem that provide symmetric positive definite linear systems with effective condition numbers. Moreover, the sparsity pattern of the rank-1 update can be controlled so as to match the sparsity pattern of the singular matrix by taking a weight function with the appropriate support.

Acknowledgments. We would like to thank Doug Arnold, Martin Berggren, Quang Du, and Axel Klawonn for helpful discussions; David Silvester for a careful review of the nearly completed revision; and Ulrich Hetmaniuk for generating Matlab code for the matrices and source terms used in the experiments of the last section.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.
[2] E. Becker, G. Carey, and T. Oden, *Finite Elements. An Introduction*, Vol. 1, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[3] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.
[4] J. H. Bramle, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring.* I, Math. Comp., 47 (1986), pp. 103–134.
[5] J. H. Bramle and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–476.
[6] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.
[8] G. Carey and T. Oden, *Finite Elements. Computational Aspects*, Vol. 3, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[9] R. Cook, D. Malkus, and M. Plesha, *Concepts and Applications of Finite Element Analysis*, 3rd ed., John Wiley and Sons, New York, 1989.
[10] C. Farhat and M. Géradin, *On the general solution by a direct method of a large-scale singular system of linear equations: Application to the analysis of floating structures*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 675–696.
[11] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[12] P. Gresho and R. Sani, *Incompressible Flow and the Finite Element Method*, John Wiley and Sons, Chichester, UK, 1998.
[13] T. Hughes, *The Finite Element Method: Linear Static and Dynamic Analysis*, Dover, New York, 2000.
[14] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.
[15] S. F. McCormick, *Multilevel Adaptive Methods for Partial Differential Equations*, SIAM, Philadelphia, 1989.
[16] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Verlag, New York, 1999.
[17] M. Papadrakakis and Y. Fragakis, *An integrated geometric-algebraic method for solving semi-definite problems in structural mechanics*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6513–6532.
[18] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, 2nd ed., Springer-Verlag, Berlin, 1997.
[19] J. N. Reddy, *An Introduction to the Finite Element Method*, 2nd ed., McGraw-Hill, New York, 1993.
[20] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method: Fluid Dynamics*, Vol. 3, 5th ed., Butterworth-Heinemann, New York, 2000.
[21] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method: Solid Mechanics*, Vol. 2, 5th ed., Butterworth-Heinemann, New York, 2000.
[22] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method: The Basis*, Vol. 1, 5th ed., Butterworth-Heinemann, New York, 2000.