Electronic Transactions on Numerical Analysis.
Volume 58, pp. 394–401, 2023.
Copyright © 2023, Kent State University.
ISSN 1068–9613.
DOI: 10.1553/etna_vol58s394

ETNA
Kent State University and
Johann Radon Institute (RICAM)

CONDITIONING OF LINEAR SYSTEMS ARISING FROM PENALTY METHODS[*]

WILLIAM LAYTON[†] AND SHUXIAN XU[†]

We dedicate this paper to Professor Owe Axelsson.

Abstract. Penalizing incompressibility in the Stokes problem leads, under mild assumptions, to matrices with condition numbers $\kappa = \mathcal{O}(\varepsilon^{-1}h^{-2})$, with ε = penalty parameter $\ll 1$ and h = meshwidth < 1. Although $\kappa = \mathcal{O}(\varepsilon^{-1}h^{-2})$ is large, practical tests seldom report difficulty in solving these systems. In the SPD case, using the conjugate gradient method, this is usually explained by spectral gaps occurring in the penalized coefficient matrix. Herein we point out a second contributing factor. Since the solution is approximately incompressible, solution components in the eigenspaces associated with the penalty terms can be small. As a result, the *effective* condition number can be much smaller than the standard condition number.

Key words. penalty method, effective condition number

AMS subject classifications. 65F35, 15A12

1. Introduction.

The difficulty in solving a linear system approximately in finite precision is quantified by the coefficient matrix condition number. For a symmetric, positive definite matrix A and measured in the Euclidean norm, this is $\kappa = \lambda_{\max}/\lambda_{\min}$. For some methods and systems, the effective condition number (Definition 1.1) provides a sharper estimate by taking into account the size of solution components in the relevant eigenspaces. This report shows that one important system where the effective condition number satisfies $\kappa(\text{solution}) \ll \kappa$ arises from penalty methods (Theorem 2.1, Section 2).

Penalty methods have advantages and disadvantages. Disadvantages include the need to pick a specific value of the penalty parameter ε and the ill-conditioning of the associated linear system. We show herein that, measured by the system's effective condition number, this ill-conditioning is not severe; cf. Theorem 2.1, Section 2. This observation is developed herein for the standard penalty approximation of the Stokes problem

$$-\triangle u + \nabla p = f(x) \qquad \text{and} \qquad \nabla \cdot u = 0$$

in a polygonal domain Ω with boundary condition $u = 0$ on $\partial\Omega$ and normalization $\int_\Omega p dx = 0$. Let $(\cdot,\cdot)$, $||\cdot||$ denote the $L^2(\Omega)$-inner product and -norm and $|\cdot|_2$ the usual Euclidean norm of a vector and matrix. X^h denotes a standard, conforming, velocity finite element space of continuous, piecewise polynomials that vanish on $\partial\Omega$. The penalty approximation results from replacing $\nabla \cdot u = 0$ by $\nabla \cdot u^\varepsilon = -\varepsilon p^\varepsilon$ and eliminating p^ε. It is: find $u^\varepsilon \in X^h$ satisfying

$$(\nabla u^\varepsilon, \nabla v) + \varepsilon^{-1}(\nabla \cdot u^\varepsilon, \nabla \cdot v) = (f, v) \qquad \text{for all } v \in X^h. \tag{1.1}$$

Picking a basis $\{\phi_1, \ldots, \phi_N\}$ for X^h leads to a linear system with coefficient matrix

$$A_{ij} = (\nabla\phi_i, \nabla\phi_j) + \varepsilon^{-1}(\nabla\cdot\phi_i, \nabla\cdot\phi_j), \qquad i, j = 1, \ldots, N.$$

Standard condition number estimates for this system yield bounds like $\kappa \leq C\varepsilon^{-1}h^{-2}$.

Recall that the standard condition number [2, 23], introduced in [21, 22] and still of interest, cf., e.g., [20], measuring the correlation between relative error and relative residual, the distance to the nearest singular matrix, and the difficulty, cost, and worse-case accuracy

[*] Received June 13, 2022. Accepted March 8, 2023. Published online on April 25, 2023. Recommended by Maya Neycheva. Partially supported by NSF Grant DMS 2110379.

[†] Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260, USA ({wjl,shx34}@pitt.edu).

in solving a linear system with A, is $\kappa := |A|_2|A^{-1}|_2$. However, estimates for the above with κ are often (but not always, cf. [5]) too rough because they do not consider the size of solution components in the matrix eigenspaces. Thus, generalizations exist such as the Kaporin condition number [16] and extensions based on generalized inverses [12] that can be used to obtain better predictions. One important, easy to compute and interpret generalization is the *condition number at the solution*, also called the *effective condition number*.

DEFINITION 1.1. *Let $Ac = b$, and select the Euclidean norm. Then $\kappa(c)$, the condition number at the solution c, is*

$$\kappa(c) := |A^{-1}|_2 \frac{|Ac|_2}{|c|_2}.$$

Clearly $\kappa(c) \le \kappa$, and $\kappa(c)$ takes into account both the spectrum and the magnitude of solution components across eigenspaces. It is also known, e.g., [2], that the relative error of approximations is bounded by $\kappa(c)$ times the relative residual. To our knowledge this extension is due to Chan and Foulser [9]. It was soon thereafter developed by Axelsson [1, 2] and has been further developed in important works in [3, 4, 7, 10, 18, 19].

Section 2 gives the proof that $\kappa(u^\varepsilon) \ll \kappa$. Section 3 presents consistent numerical tests. Throughout, "C" denotes an $\mathcal{O}(1)$-constant, independent of ε and h.

2. Analysis of $\kappa(u^\varepsilon)$.

We assume X^h satisfies the following two assumptions that are typical, e.g. [6, 11], for finite element spaces on quasi-uniform meshes.

A1: (Inverse estimate) For all $v \in X^h$, $||\nabla v|| \le Ch^{-1}||v||$.

A2: (Norm equivalence) Let $v = \sum_{i=1}^N a_i\phi_i(x), N = Ch^{-d}, d = \dim(\Omega) = 2$ or 3. Then $||v||$ and $\sqrt{\frac{1}{N}\sum_{i=1}^N a_i^2}$ are uniform-in-h equivalent norms.

THEOREM 2.1. *Let A1 and A2 hold. Select $|\cdot|_2$, the Euclidean norm. Let f^h be the projection of f onto the finite element space and u^ε the solution of* (1.1). *Then,*

$$\max_{f^h} \kappa(u^\varepsilon) = \kappa \le C(h^{-2} + \varepsilon^{-1}h^{-2})$$
$$\kappa(u^\varepsilon) \le C\frac{||f^h||}{||u^\varepsilon||}$$
$$\kappa(u^\varepsilon) \le Ch^{-2}\left(1 + \frac{h}{\varepsilon}\frac{||\nabla\cdot u^\varepsilon||}{||u^\varepsilon||}\right).$$

Proof. We first estimate $|A^{-1}|_2^2 = \max_b |A^{-1}b|_2^2/|b|_2^2$. Given an arbitrary right-hand side $\vec{b}$, let the linear system for the undetermined coefficients $(c_1, c_2, \dots, c_N)^T = \vec{c}$ be denoted by $A\vec{c} = \vec{b}$. We convert in a standard way this linear system to an equivalent formulation similar to (1.1). Recall that the mass matrix associated with this basis is

$$M_{ij} = (\phi_i, \phi_j), \qquad i, j = 1, \dots, N.$$

The matrix M is symmetric and positive definite. Under A1 and A2, its eigenvalues satisfy $0 < C_1N^{-1} \le \lambda(M) \le C_2N^{-1}$. Given the vector $\vec{b}$, let $\vec{a} = M^{-1}\vec{b}$. By construction

$$g(x) = \sum_{i=1}^N a_i\phi_i(x) \quad \text{satisfies} \quad (g(x), \phi_j) = b_j, \quad j = 1, \dots, N.$$

By A2, $||g||$ and $(N^{-1}\sum a_i^2)^{1/2}$ are uniformly equivalent norms. The bounds for $\lambda(M)$ (also from A2) imply that $(N^{-1}\sum a_i^2)^{1/2}$ and $(N^{-1}\sum b_i^2)^{1/2}$ are also uniformly equivalent norms.

Next define

$$w = \sum_{i=1}^{N} c_i \phi_i(x).$$

By A2 again, $||w||$, $(N^{-1}\sum c_i^2)^{1/2}$ are equivalent norms. Thus,

$$|A^{-1}|_2^2 = \max_b \frac{|A^{-1}b|_2^2}{|b|_2^2} \leq C \max_{g \in X^h} \frac{||w||^2}{||g||^2}.$$

By construction $w, g \in X^h$ satisfy

$$(\nabla w, \nabla v) + \varepsilon^{-1}(\nabla \cdot w, \nabla \cdot v) = (g, v) \qquad \text{for all } v \in X^h.$$

Setting $v = w$ and using simple inequalities gives $||w||^2 \leq C||g||^2$. Indeed,

$$C||w||^2 \leq ||\nabla w||^2 + \varepsilon^{-1}||\nabla \cdot w||^2 = (g, w) \ \leq \frac{C}{2}||w||^2 + C||g||^2.$$

Thus $||w||^2 \leq C||\nabla w||^2 \leq C||g||^2$ and $||w||^2/||g||^2 \leq C$. This implies $|A^{-1}|_2 \leq C$, uniformly in h, ε.

Next we estimate $|Ac|_2/|c|_2$ where $u^\varepsilon = \sum_{i=1}^N c_i\phi_i(x)$ is the solution of (1.1). By norm equivalence, as above, and (1.1), this is equivalent to $||f^h||/||u^\varepsilon||$ where f^h is the L^2-projection of $f(x)$ onto the finite element space. This gives the first estimate $\kappa(u^\varepsilon) \leq C||f^h||/||u^\varepsilon||$. For the second estimate, we have

$$\begin{aligned}
||f^h|| &= \max_{v \in X^h} \frac{(f^h, v)}{||v||} = \max_{v \in X^h} \frac{(\nabla u^\varepsilon, \nabla v) + \varepsilon^{-1}(\nabla \cdot u^\varepsilon, \nabla \cdot v)}{||v||} \\
&\leq \max_{v \in X^h} \frac{||\nabla u^\varepsilon||||\nabla v|| + \varepsilon^{-1}||\nabla \cdot u^\varepsilon||||\nabla \cdot v||}{||v||} \\
&\overset{\text{(using A1)}}{\leq} \max_{v \in X^h} \frac{Ch^{-2}||u^\varepsilon||||v|| + \varepsilon^{-1}Ch^{-1}||\nabla \cdot u^\varepsilon||||v||}{||v||} \\
&\leq Ch^{-2}||u^\varepsilon|| + \varepsilon^{-1}Ch^{-1}||\nabla \cdot u^\varepsilon||,
\end{aligned}$$

which implies

$$\frac{||f^h||}{||u^\varepsilon||} \leq Ch^{-2} + C\varepsilon^{-1}h^{-1}\frac{||\nabla \cdot u^\varepsilon||}{||u^\varepsilon||}.$$

This yields

$$\kappa(u^\varepsilon) \leq C\frac{||f^h||}{||u^\varepsilon||} \leq C\left(h^{-2} + \varepsilon^{-1}h^{-1}\frac{||\nabla \cdot u^\varepsilon||}{||u^\varepsilon||}\right).$$

Using the inequalities $||\nabla \cdot u^\varepsilon|| \leq C||\nabla u|| \leq Ch^{-1}||u^\varepsilon||$ and A1 yields the standard estimate $\kappa \leq C(h^{-2} + \varepsilon^{-1}h^{-2})$. □

3. An illustration.

The result in Section 2 gives three estimates for the conditioning. We explore if the three estimates

$$\kappa \leq C(h^{-2} + \varepsilon^{-1}h^{-2}), \quad \kappa(u^\varepsilon) \leq C\frac{||f^h||}{||u^\varepsilon||}, \quad \kappa(u^\varepsilon) \leq C\left(h^{-2} + \varepsilon^{-1}h^{-1}\frac{||\nabla \cdot u^\varepsilon||}{||u^\varepsilon||}\right)$$

TABLE 3.1
Values of Est. 2 $\simeq \frac{||f^h||}{||u^\varepsilon||}$, *P1-elements.*

$\varepsilon \Downarrow \backslash h \Rightarrow$	0.125	6.2 e-2	3.1 e-2	1.6 e-2	7.8 e-3	3.9 e-3
1	39	38	37	37	37	37
6.3 e-2	1.9 e2	1.7 e2	1.7 e2	1.7 e2	1.7 e2	1.7 e2
3.9 e-3	1.8 e3	7.8 e2	4.3 e2	3.3 e2	3.0 e2	2.9 e2
2.4 e-4	2.5 e4	7.5 e3	2.2 e3	8.0 e2	4.3 e2	3.3 e2
1.5 e-5	4.0 e5	1.1 e5	2.9 e4	7.5 e3	2.1 e3	8.0 e2
9.5 e-7	6.4 e6	1.8 e6	4.5 e5	1.1 e5	2.9 e4	7.4 e3
6.0 e-8	1.0 e8	2.9 e7	7.2 e6	1.8 e6	4.5 e5	1.1 e5

yield significantly different predictions. We investigate this question for two test problems and two finite element spaces. The first test problem has a smooth solution and the second has a right-hand side $f(x, y)$ that is oscillating as fast as possible on the given mesh. The domain is the unit square, and the mesh is a uniform mesh of squares each divided into two right triangles. The first finite element space is P1, i.e., conforming linear elements. The second is P2, i.e., conforming quadratics.

In penalty methods, the substitution $p^h = -\frac{1}{\varepsilon}\nabla \cdot u^h$ means that the analysis of pressure stability and errors introduces the implicitly defined pressure space $\nabla \cdot X^h$. It is known in the theory of penalty methods that the recovered pressure accuracy is higher when $(X^h, \nabla \cdot X^h)$ satisfies a discrete inf-sup condition. For these two spaces $(X^h, \nabla \cdot X^h)$ does not satisfy the discrete inf-sup condition. The results herein indicate that for small ε the conditioning depends instead on the existence of a non-trivial divergence-free subspace rather than an inf-sup condition. The space of conforming linear elements does not contain a divergence-free subspace. Thus, the coefficient matrix A is expected to show higher ill-conditioning as $\varepsilon \to 0$ than with conforming quadratics.

As "C" in Theorem 2.1 is an $\mathcal{O}(1)$-constant, independent of ε and h, we compute and compare the right-hand sides

$$\begin{aligned} &\textit{Est. } 1 \simeq h^{-2} + \varepsilon^{-1}h^{-2}, \\ &\textit{Est. } 2 \simeq \frac{||f^h||}{||u^\varepsilon||}, \qquad\qquad \text{and} \\ &\textit{Est. } 3 \simeq \varepsilon^{-1}h^{-1}\frac{||\nabla \cdot u^\varepsilon||}{||u^\varepsilon||}. \end{aligned}$$

We computed these values starting with $\varepsilon = 1, h = 0.125$. We successively cut ε by 16 and h by 2 until $\varepsilon = 5.96\text{e-}8$ and $h = 0.00195312$. In the data, the number $x\text{e}y$ denotes $x \cdot 10^y$.

Test problem 1. We solve this test problem with P1, conforming, linear elements and then with P2, conforming, quadratic elements. Choose $f(x, y) = (\sin(x + y), \cos(x + y))^T$. For P1-elements we present the results for *Est.* 2 in Table 3.1.

Down the columns (fixing h and decreasing ε), the data for *Est.* 2 show that this quantity grows as $\varepsilon \downarrow 0$, roughly like ε^{-1} for fixed h. Across the rows, for fixed ε and $h \downarrow 0$, *Est.* 2 decreases. The diagonals (necessary if we choose $\varepsilon \sim h$) show mild growth. Comparing the last row of the table with the row of the *Est.* 1-values shows that *Est.* 2 consistently yields a lower estimate of the ill-conditioning than *Est.* 1. Concerning the growth of *Est.* 2 as $\varepsilon \downarrow 0$ for fixed h, since $h, ||f||$ do not change as ε varies, this growth is due to $||u^\varepsilon|| \to 0$ as $\varepsilon \downarrow 0$. The P1 finite element space does not have a non-trivial divergence-free subspace [15]. Thus,

TABLE 3.2
Values of Est. 3 $\simeq \varepsilon^{-1}h^{-1}\frac{||\nabla\cdot u^{\varepsilon}||}{||u^{\varepsilon}||}$, *P1-elements.*

$\varepsilon \Downarrow \backslash h \Rightarrow$	0.125	6.2 e-2	3.1 e-2	1.6 e-2	7.8 e-3	3.9 e-3
1	25	50	1.0 e2	2.0 e2	4.0 e2	8.0 e2
6.3 e-2	3.5 e2	6.6 e2	1.3 e3	2.5 e3	5.0 e3	1.0 e4
3.9 e-3	4.2 e3	4.1 e3	4.5 e3	6.3 e3	1.1 e4	2.0 e4
2.4 e-4	6.2 e4	4.9 e4	4.2 e4	4.5 e4	3.5 e2	5.1 e4
1.5 e-5	9.8 e5	7.6 e5	6.3 e5	4.2 e4	6.0 e5	6.4 e5
9.5 e-7	1.5 e7	1.2 e7	1.0 e7	9.5 e6	9.4 e6	9.4 e6
6.0 e-8	2.5 e8	1.9 e8	1.6 e8	1.5 e8	1.5 e8	1.5 e8

TABLE 3.3
Values of Est. 3 $\simeq \varepsilon^{-1}h^{-1}\frac{||\nabla\cdot u^{\varepsilon}||}{||u^{\varepsilon}||}$, *P2-elements.*

$\varepsilon \Downarrow \backslash h \Rightarrow$	0.125	6.2 e-2	3.1 e-2	1.6 e-2	7.8 e-3	3.9 e-3
1	25	50	1.0 e2	2.0 e2	4.0 e2	8.0 e2
6.3 e-2	3.1 e2	6.2 e2	1.2 e3	2.5 e3	5.0 e3	1.0 e4
3.9 e-3	6.7 e2	1.3 e3	2.5 e3	5.0 e3	1.0 e4	2.0 e4
2.4 e-4	7.1 e2	1.3 e3	2.6 e3	5.1 e3	1.0 e4	2.0 e4
1.5 e-5	7.1 e2	1.3 e3	2.6 e3	5.2 e3	1.0 e4	2.0 e4
9.5 e-7	7.1 e2	1.3 e3	2.7 e3	5.3 e3	1.1 e4	2.1 e4
6.0 e-8	7.1 e2	1.3 e3	2.7 e3	5.3 e3	1.1 e4	2.1 e4

$\varepsilon \downarrow 0$ forces $||\nabla \cdot u|| \to 0$, which forces $||u|| \to 0$. This indicates that the ill-conditioning of penalty methods *with P1-elements* as $\varepsilon \downarrow 0$ is an essential feature caused by the *lack of a divergence-free subspace*. For comparison with the last row in Table 3.1, the values of *Est.* $1 \simeq h^{-2} + \varepsilon^{-1}h^{-2}$ for $\varepsilon = 6.0\text{e-}8$ are

$$1.0\text{e}9 \quad 4.3\text{e}9 \quad 1.7\text{e}10 \quad 6.9\text{e}10 \quad 2.7\text{e}11 \quad 1.1\text{e}12.$$

Clearly *Est.* 2 provides a smaller estimate than *Est.* 1. Another interpretation is that a poor choice of the finite element space (made to accentuate ill-conditioning) makes the problem of selecting ε acute.

Table 3.2 presents the analogous results for *Est.* 3 and P1-elements. For *Est.* 3, the data show a similar behavior to *Est.* 2 except for fixed ε and $h \downarrow 0$. In this case, the ill-conditioning for small ε and $h \downarrow 0$ is over-estimated compared with Table 3.1. This is expected because the data come from P1-elements and $||\nabla \cdot u^{\varepsilon}||$ occurs in *Est.* 3. Again, *Est.* 3 still provides a smaller estimate of the ill-conditioning than *Est.* 1.

We have attributed the ill-conditioning observed above as $\varepsilon \to 0$ to the use of P1-elements. To test if this explanation is plausible we repeated this test with P2-elements. Since this finite element space contains a non-trivial divergence-free subspace [15], our intuition is that ill-conditioning would be reduced. Table 3.3 presents the values of *Est.* $3 \simeq \varepsilon^{-1}h^{-1}\frac{||\nabla\cdot u^{\varepsilon}||}{||u^{\varepsilon}||}$. The values of *Est.* $2 \simeq ||f^h||/||u^{\varepsilon}||$, not presented, were significantly smaller than that for *Est.* 3 and converged to 1.13 as $\varepsilon, h \to 0$.

The above data indicate that with P2-elements and a problem with a smooth solution, the contribution of the penalty term to the ill-conditioning is small.

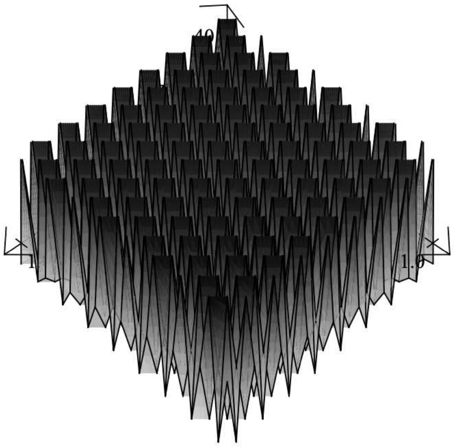

FIG. 3.1. *Plot of* $10 + 50\sin(\frac{2\pi x}{0.125} + \frac{2\pi y}{0.125})$.

TABLE 3.4
Values of Est. 2 $\simeq \frac{||f^h||}{||u^\varepsilon||}$, *P1-elements.*

$\varepsilon \Downarrow \backslash h \Rightarrow$	0.125	6.2 e-2	3.1 e-2	1.6 e-2	7.8 e-3	3.9 e-3
1	1.1 e2	1.0 e2	1.0 e2	1.0 e2	1.0 e2	1.0 e2
6.3 e-2	5.6 e2	5.6 e2	5.6 e2	5.6 e2	5.6 e2	5.6 e2
3.9 e-3	7.4 e3	7.5 e3	7.6 e3	7.6 e3	7.7 e3	7.7 e3
2.4 e-4	1.1 e5	1.1 e5	1.2 e5	1.2 e5	1.2 e5	1.2 e5
1.5 e-5	1.8 e6	1.8 e6	1.8 e6	1.8 e6	1.8 e6	1.9 e6
9.5 e-7	2.9 e7	2.9 e7	2.9 e7	2.9 e7	2.9 e7	2.9 e7
6.0 e-8	4.7 e8	4.6 e8	4.6 e8	4.6 e8	4.6 e8	4.6 e8

Test problem 2. We select the forcing function

$$f(x,y) = \left(10 + 50\sin\left(\frac{2\pi x}{h} + \frac{2\pi y}{h}\right), 10 + 50\sin\left(\frac{2\pi x}{h} + \frac{2\pi y}{h}\right)\right)^T.$$

Each component oscillates as rapidly as the given mesh allows; see Figure 3.1 for $h = 0.125$. The forcing function and thus the solution change with the mesh size. For comparison with the last row in Table 3.4, the values of *Est.* 1 $\simeq h^{-2} + \varepsilon^{-1}h^{-2}$ for $\varepsilon = 6.0$e-8 are

1.1 e9 4.3 e9 1.7 e10 6.9 e10 2.7 e11 1.1 e12.

In Table 3.4, as $\varepsilon \downarrow 0$, *Est.* 2 grows roughly like ε^{-1}. The stable behavior of *Est.* 2 as $h \downarrow 0$ is unexpected. The behavior of *Est.* 3 $\simeq \varepsilon^{-1}h^{-1}\frac{||\nabla\cdot u^\varepsilon||}{||u^\varepsilon||}$ for P1-elements was similar.

We now present tests of P2-elements for this problem. In this test, *Est.* 3 was larger than *Est.* 2. We thus present the *Est.* 3-data in Table 3.5. For comparison with the last row in Table 3.5, the *Est.* 1-values for $\varepsilon = 6.0$e-8 are

1.1e9 4.3e9 1.7e10 6.9e10 2.7e11 1.1e12.

TABLE 3.5
Values of Est. 3 $\simeq \varepsilon^{-1}h^{-1}\frac{||\nabla\cdot u^{\varepsilon}||}{||u^{\varepsilon}||}$, *P2-elements.*

$\varepsilon \Downarrow \backslash h \Rightarrow$	0.125	6.2e-2	3.1 e-2	1.6 e-2	7.8 e-3	3.9 e-3
1	25	49	1.0 e2	2.0 e2	4.0 e2	7.9 e2
6.3 e-2	3.8 e2	7.5 e2	1.5 e3	3.0 e3	6.0 e3	1.2 e4
3.9 e-3	6.0 e3	1.2 e4	2.4 e4	4.8 e4	9.6 e4	1.9 e5
2.4 e-4	9.6 e4	1.9 e5	3.8 e5	7.7 e5	1.5 e6	3.1 e6
1.5 e-5	1.5 e6	3.1 e6	6.2 e6	1.2 e7	2.5 e7	4.9 e7
9.5 e-7	2.5 e7	4.9 e7	9.8 e7	2.0 e8	3.9 e8	7.9 e8
6.0 e-8	3.9 e8	7.9 e8	1.6 e9	3.2 e9	6.3 e9	1.3 e10

For this problem, down the columns (fixed h, $\varepsilon \downarrow 0$), the computed estimate of $\kappa(u^{\varepsilon})$ grows roughly like ε^{-1}. Reading across the columns, the values in each row grow like h^{-1} (not h^{-2}). In all cases, *Est.* 3 was smaller than *Est.* 1.

4. Conclusions. The classical view is that there are two significant difficulties with penalty methods. The first is the ill-conditioning of the resulting system matrix. We have shown that the effective condition number $\kappa(u^{\varepsilon})$ is much smaller than the usual condition number due to the magnitude of the components in the penalized eigenspaces being small in a precise sense. Motivated by this theoretical result, we then compared the derived estimates of the ill-conditioning for two test problems and for two elements. With P1-elements, the ill-conditioning was not as severe as $\mathcal{O}(\varepsilon^{-1}h^{-2})$ but followed an ε^{-1}-growth as $\varepsilon \to 0$. For P2-elements and approximating a smooth solution, $\kappa(u^{\varepsilon})$ was smaller than the expected rate $\kappa \sim h^{-2}$ for the discrete Laplacian. With both P1- and P2-elements and for an academic test problem with data oscillating as fast as the mesh allows, the ill-conditioning was not as severe as the expected $\mathcal{O}(\varepsilon^{-1}h^{-2})$ but also followed the ε^{-1}-pattern as $\varepsilon \to 0$.

The second difficulty is the selection of an effective value of ε. While the most commonly recommended choices, e.g., [8, 14], are $\varepsilon =$ *time step, mesh width*, and (*machine epsilon)*$^{1/2}$, none have proven reliably effective. Recent work of [17, 24] may resolve this impediment by an algorithmic, self-adaptive selection of ε based on some indicator of violation of incompressibility. The estimates in Theorem 2.1 give insight into the resulting conditioning when this is done. Let TOL denote a specified tolerance. If ε is adapted so that the penalized solution u^{ε} satisfies $\frac{||\nabla\cdot u^{\varepsilon}||}{||u^{\varepsilon}||} \leq TOL$, then Theorem 2.1 immediately implies that $\kappa(u^{\varepsilon})$ satisfies

$$\kappa(u^{\varepsilon}) \leq Ch^{-2}\left(1 + h\frac{TOL}{\varepsilon}\right).$$

If ε is adapted so that the penalized solution u^{ε} satisfies $\frac{||\nabla\cdot u^{\varepsilon}||}{||\nabla u^{\varepsilon}||} \leq TOL$, then, similarly, Theorem 2.1 implies that $\kappa(u^{\varepsilon})$ satisfies

$$\kappa(u^{\varepsilon}) \leq Ch^{-2}\left(1 + \frac{TOL}{\varepsilon}\right). \tag{4.1}$$

For (4.1), rewrite $||\nabla\cdot u^{\varepsilon}||/||u^{\varepsilon}||$ as $(||\nabla\cdot u^{\varepsilon}||/||\nabla u^{\varepsilon}||)\,(||\nabla u^{\varepsilon}||/||u^{\varepsilon}||)$. The inverse estimate A1 implies $||\nabla u^{\varepsilon}||/||u^{\varepsilon}|| \leq Ch^{-1}$, yielding (4.1).

The numerical tests suggest that two factors not considered in Theorem 2.1 are significant. The first is whether the finite element space has a nontrivial, divergence-free subspace. The second is the influence of smoothness of the sought solution or its problem data on the effective conditioning. In addition, highly refined meshes are used in practical flow simulations, and the linear system often has large skew-symmetric parts. The extension of the analysis herein to include these effects is an open problem.

REFERENCES

[1] O. AXELSSON, *Condition numbers for the study of the rate of convergence of the conjugate gradient method*, in Proc. 2nd IMACS Internat. Symposium on Iterative Methods in Linear Algebra, S. Margenov and P. S. Vassilevski, eds., IMACS, New Jersey, 1996, pp. 3–33.

[2] ———, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1994.

[3] O. AXELSSON AND I. KAPORIN, *On the sublinear and superlinear rate of convergence of conjugate gradient methods*, Numer. Algorithms, 25 (2000), pp. 1–22.

[4] ———, *Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations*, Numer. Linear Algebra Appl., 8 (2001), pp. 265–286.

[5] J. M. BANOCZI, N.-C. CHIU, G. E. CHO, AND I. C. F. IPSEN, *The lack of influence of the right-hand side on the accuracy of linear system solution*, SIAM J. Sci. Comput., 20 (1998), pp. 203–227.

[6] S. BRENNER AND L. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 2008.

[7] C. BREZINSKI, *Error estimates for the solution of linear systems*, SIAM J. Sci. Comput., 21 (1999), pp. 764–781.

[8] G. F. CAREY AND R. KRISHNAN, *Penalty approximation of Stokes flow*, Comput. Methods Appl. Mech. Engrg., 35 (1982), pp. 169–206.

[9] T. F. CHAN AND D. E. FOULSER, *Effectively well-conditioned linear systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 963–969.

[10] S. CHRISTIANSEN AND P. C. HANSEN, *The effective condition number applied to error analysis of certain boundary collocation methods*, J. Comput. Appl. Math., 54 (1994), pp. 15–36.

[11] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, SIAM, Philadelphia, 2002.

[12] S. DEMKO, *Condition numbers of rectangular systems and bounds for generalized inverses*, Linear Algebra Appl., 78 (1986), pp. 199–206.

[13] Q. DU, D. WANG, AND L. ZHU, *On mesh geometry and stiffness matrix conditioning for general finite element spaces*, SIAM J. Numer. Anal., 47, 2009, pp. 1421–1444.

[14] J. C. HEINRICH AND C. A. VIONNET, *The penalty method for the Navier-Stokes equations*, Arch. Comput. Methods Engrg., 2 (1995), pp. 51–65.

[15] V. JOHN, A. LINKE, C. MERDON, M. NEILAN, AND L. G. REBHOLZ, *On the divergence constraint in mixed finite element methods for incompressible flows*, SIAM Rev., 59 (2017), pp. 492–544.

[16] I. E. KAPORIN, *New convergence results and preconditioning strategies for the conjugate gradient method*, Numer. Linear Algebra Appl., 1 (1994), pp. 179–210.

[17] K. KEAN, X. XIE, AND S. XU, *A doubly adaptive penalty method for the Navier Stokes Equation*, Preprint on arXiv, 2022. https://arxiv.org/abs/2201.03978. To appear in Int. J. Numer. Anal. Model. Ser. B, 2023.

[18] Z.-C. LI AND H.-T. HUANG, *Effective condition number for numerical partial differential equations*, Numer. Linear Algebra Appl., 15 (2008), pp. 575–594.

[19] Z.-C. LI, H.-T. HUANG, AND J. HUANG, *Superconvergence and stability for boundary penalty techniques of finite difference methods*, Numer. Methods Partial Differential Equations, 24 (2008), pp. 972–990.

[20] T. TAO AND V. VU, *Smooth analysis of the condition number and the least singular value*, Math. Comp., 79 (2010), pp. 2333–2352.

[21] A. M. TURING, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 287–308.

[22] J. VON NEUMANN AND H. H. GOLDSTINE, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc., 53 (1947), pp. 1021–1099.

[23] J. WILKINSON, *Rounding Errors in Algebraic Processes*, National Physical Laboratory Notes on Applied Science Vol. 32, HMSO, London, 1963.

[24] X. XIE, *On adaptive grad-div parameter selection*, J. Sci. Comput., 92 (2022), Paper No. 108, 23 pages.