APPROXIMATE ERROR BOUNDS ON SOLUTIONS OF NONLINEARLY PRECONDITIONED PDES*

LULU LIU[†] AND DAVID E. KEYES[‡]

Abstract. In many multiphysics applications, an ultimate quantity of interest can be written as a linear functional of the solution to the discretized governing nonlinear partial differential equations and finding a sufficiently accurate pointwise solution may be regarded as a step toward that end. In this paper, we derive a posteriori approximate error bounds for linear functionals corresponding to quantities of interest using two kinds of nonlinear preconditioning techniques. Nonlinear preconditioning, such as the inexact Newton with backtracking and nonlinear elimination algorithm and the multiplicative Schwarz preconditioned inexact Newton algorithm, may be effective in improving global convergence for Newton's method. It may prevent stagnation of the nonlinear residual norm and reduce the number of solutions of large ill-conditioned linear systems involving a global Jacobian required at each nonlinear iteration. We illustrate the effectiveness of the new bounds using canonical nonlinear PDE models: a flame sheet model and a nonlinear coupled lid-driven cavity problem.

Key words. nonlinear equations, nonlinear preconditioning, error bounds, inexact Newton methods

AMS subject classifications. 49M15, 65H10, 65H20, 65N15, 65N22

DOI. 10.1137/19M1285044

1. Introduction. Multiphysics simulation rides the wave of increasing computational capability to model interacting phenomena that, if modeled separately, would require unrealistically restrictive assumptions about interface conditions and/or coupling terms. Examples include systems coupled in bulk, such as electromagnetism with hydrodynamics in magnetohydrodynamics [7], chemical kinetics with hydrodynamics in combustion [29], and electrophysiology with tissue deformation cardiodynamics [27], as well as systems coupled at interfaces, such as fluids with structures in aeroelasticity [13] and oceanic with atmospheric dynamics in climate [15].

To obtain numerical solutions of large systems arising from the discretization of weakly coupled nonlinear multiphysics problems, we may employ a field-by-field approach through operator-splitting in analogy with Gauss–Seidel iterations. However, for strongly coupled multiphysics systems, fully implicit Newton-like solvers are favored, including variants of Jacobian-free Newton–Krylov methods [17, 19]. Without a good initial iterate, however, Newton-like methods may suffer from lack of robustness in the form of stagnation or slow convergence before the approximate solution is brought into the domain of superlinear or quadratic convergence. Nonlinear preconditioning may be effective in desensitizing Newton to the initial guess to enable faster global convergence.

*Submitted to the journal's Methods and Algorithms for Scientific Computing section September 3, 2019; accepted for publication (in revised form) April 29, 2021; published electronically July 15, 2021.

<https://doi.org/10.1137/19M1285044>

Funding: The work of the first author was supported by the National Natural Science Foundation of China through grant 11901296 and by the Natural Science Foundation of Jiangsu through grant BK20180450. The work of the second author was supported by the Extreme Computing Research Center at KAUST.

[†]Department of Mathematics, School of Science, Nanjing University of Science and Technology, Nanjing, 210094, China (lulu.liu@njust.edu.cn).

[‡]Extreme Computing Research Center, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia (david.keyes@kaust.edu.sa).

Nonlinear preconditioning techniques deal with strong nonlinearities by replacing the nonlinear functions or changing the unknowns. In analogy with linear preconditioning, these can be classified into “left” and “right” forms, respectively. Left nonlinear preconditioning for Newton was introduced in the form of the additive Schwarz preconditioned inexact Newton method (ASPIN) [5], originally conceived of as a nonlinear domain decomposition method. For multiphysics problems, a Gauss–Seidel-like variant of ASPIN based on field splitting rather than domain-splitting was proposed in the form of the multiplicative Schwarz preconditioned inexact Newton (MSPIN) algorithm [21]. Right nonlinear preconditioners are derived from nonlinear elimination (NE) [20]. Based on different partitioning strategies, they lead to algorithms such as NKS-RAS [6] in a geometry-based domain decomposition form and nonlinear elimination preconditioned inexact Newton algorithm (INB-NE) [16, 31] or RS-NE [32] in a physics-based field-split form.

Some theoretical support has been provided for nonlinear preconditioning, including a convergence analysis of ASPIN in the context of semilinear elliptic PDEs [24] and local convergence rates for ASPIN and MSPIN [1, 22]. Here, we contribute theory relevant to bounds on quantities of interest written as linear functionals of the solution, such as pointwise or regionally integrated values or fluxes. In a typical design problem, the full PDE solution is often merely a means to evaluate such pointwise or integrated quantities. For example, in cooling electronic components engineers focus on the average temperature over small boundary patches of a fluid-filled domain [25]. Corresponding error bounds are obtained using linear or nonlinear block Gauss–Seidel iterations in [30]. We derive a posteriori approximate error bounds for linear functionals for nonlinear preconditioners—INB-NE and MSPIN—employed as solvers for multiphysics problems.

The paper is organized as follows. In section 2, we introduce notation and the definition of row block diagonal dominance. In section 3, we show some properties of a matrix that plays an important role in bounding errors. Our main results are in section 4, namely a posteriori approximate error bounds on solutions of nonlinearly preconditioned inexact Newton algorithms INB-NE and MSPIN. In section 5, the effectiveness of bounds is illustrated on two canonical nonlinear model problems—flame sheet combustion and lid-driven cavity flow. Conclusions are drawn in section 6. Appendix A provides a two-component analytical example intended to cement the concept of nonlinear preconditioning.

2. Notation. For the matrix $A \in R^{n \times n}$, let $\|\cdot\|$ denote the 2-norm for vectors and matrices, and let $\sigma(A)$ and $\rho(A)$ be the eigenvalue spectrum and the spectral radius of A , respectively. For a symmetric positive definite matrix A , let $A^{\frac{1}{2}}$ be a square root of A such that $A^{\frac{1}{2}}$ is symmetric and $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$. The spectral condition number of A is denoted by $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$, where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the maximum and minimum eigenvalues of A , respectively.

Following [12], we give a generalized definition of row diagonal dominance.

DEFINITION 2.1. *The matrix $X \in R^{n \times n}$ is partitioned into the $p \times p$ block form*

$$(2.1) \quad X = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{p1} & \cdots & X_{pp} \end{bmatrix},$$

where $X_{ij} \in R^{n_i \times n_j}$, $i, j = 1, \dots, p$, $n_1 + n_2 + \dots + n_p = n$ and the diagonal blocks $X_{ii} \in R^{n_i \times n_i}$ are nonsingular, $i = 1, \dots, p$. The matrix X is said to be row block diagonally dominant if

$$(2.2) \quad \sum_{\substack{j=1 \\ j \neq i}}^p \|X_{ii}^{-1} X_{ij}\| \leq 1, \quad i = 1, 2, \dots, p.$$

If inequalities in (2.2) are strict, the matrix X is called row block strictly diagonally dominant.

3. Some properties. In this section, we state some properties of a 2×2 nonsingular partitioned matrix $\mathcal{A}$, which are required to derive approximate error bounds on solutions of nonlinearly preconditioned inexact Newton algorithms such as INB-NE and MSPIN.

Assumption 3.1. Let the nonsingular matrix be denoted by

$$(3.1) \quad \mathcal{A} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix},$$

and let the diagonal blocks $\mathcal{A}_{11} \in R^{n_1 \times n_1}$ and $\mathcal{A}_{22} \in R^{n_2 \times n_2}$ be invertible.

For a matrix $\mathcal{A}$ in the form of (3.1), the matrix

$$(3.2) \quad \mathcal{C} = \mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12}$$

plays an important role in bounding errors in the next section, so we give some results pertaining to it.

LEMMA 3.2. Under Assumption 3.1, let $\mathcal{I}_1 \in R^{n_1 \times n_1}$ and $\mathcal{I}_2 \in R^{n_2 \times n_2}$ be identity matrices. Then it holds that $\mathcal{I}_2 - \mathcal{C}$ is invertible.

Proof. Since matrices $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$ are invertible, we can write $\mathcal{A}$ in the factored form

$$(3.3) \quad \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{11} & \mathbf{0} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{I}_1 & \mathcal{A}_{11}^{-1} \mathcal{A}_{12} \\ \mathbf{0} & \mathcal{I}_2 - \mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12} \end{bmatrix}.$$

Taking the determinant of the above equation, we have

$$(3.4) \quad \det(\mathcal{A}) = \det(\mathcal{A}_{11}) \det(\mathcal{A}_{22}) \det(\mathcal{I}_2 - \mathcal{C}), \quad \mathcal{C} = \mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12}.$$

According to assumptions, determinants of $\mathcal{A}$, $\mathcal{A}_{11}$, and $\mathcal{A}_{22}$ are nonzero, so it is immediate from (3.4) that $\mathcal{I}_2 - \mathcal{C}$ is invertible. $\square$

The next lemma shows a relation between the matrix $\mathcal{C}$ and the block Gauss–Seidel iterative method, which tells us how the convergence of the block Gauss–Seidel method depends on matrix partitioning.

LEMMA 3.3. Under Assumption 3.1, we consider the block Gauss–Seidel splitting

$$(3.5) \quad \mathcal{A} = \mathcal{M} - \mathcal{N}, \quad \mathcal{M} = \begin{bmatrix} \mathcal{A}_{11} & \mathbf{0} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix}, \quad \mathcal{N} = \begin{bmatrix} \mathbf{0} & -\mathcal{A}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with the companion matrix being $\mathcal{M}^{-1} \mathcal{N}$. Then

$$(3.6) \quad \rho(\mathcal{C}) = \rho(\mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12}) = \rho(\mathcal{M}^{-1} \mathcal{N}).$$

Proof. We may write the companion matrix as

$$(3.7) \quad \mathcal{M}^{-1}\mathcal{N} = \begin{bmatrix} \mathcal{A}_{11}^{-1} & \mathbf{0} \\ -\mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1} & \mathcal{A}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathcal{A}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathcal{A}_{11}^{-1}\mathcal{A}_{12} \\ \mathbf{0} & \mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12} \end{bmatrix},$$

which shows that (3.6) holds. $\square$

In section 4, we bound $\|(I - \mathcal{C})^{-1}\|$ in order to derive error bounds using the result: $\|(I - \mathcal{C})^{-1}\| \leq \frac{1}{1 - \|\mathcal{C}\|}$ if $\|\mathcal{C}\| < 1$. The following lemma demonstrates that $\|\mathcal{C}\| < 1$ can be guaranteed by row block strictly diagonal dominance of the matrix $\mathcal{A}$.

LEMMA 3.4. *Under Assumption 3.1, let $\mathcal{A}$ be row block strictly diagonally dominant. Then*

$$(3.8) \quad \|\mathcal{C}\| = \|\mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\| < 1.$$

Proof. Following Definition 2.1, the 2×2 block matrix $\mathcal{A}$ is row block strictly diagonally dominant. Then it holds that

$$(3.9) \quad \|\mathcal{A}_{22}^{-1}\mathcal{A}_{21}\| < 1, \quad \|\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\| < 1,$$

and it follows that

$$(3.10) \quad \|\mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\| \leq \|\mathcal{A}_{22}^{-1}\mathcal{A}_{21}\| \|\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\| < 1,$$

which concludes the proof of the lemma. $\square$

In the next section, the evaluation of error bounds depends on the action of $(I - \mathcal{C})^{-1}$ or $(I - \mathcal{C})^{-T}$ on a given vector, which is equivalent to solving a linear system. Usually, we obtain the solution of the linear system using Krylov subspace methods, and the condition number of the coefficient matrix characterizes the convergence rate associated with iterative methods. The next lemma provides an estimate of $\kappa(\mathcal{I} - \mathcal{C})$ and shows $\rho(\mathcal{C}) < 1$ under reasonable conditions. Actually, when $\mathcal{C}$ is normal, we further have $\rho(\mathcal{C}) = \|\mathcal{C}\| < 1$.

LEMMA 3.5. *Under Assumption 3.1, we assume that $\mathcal{A}$ in (3.1) is a symmetric positive definite matrix and define $\gamma_{\mathcal{A}}$ as*

$$(3.11) \quad \gamma_{\mathcal{A}} = \sup_{\substack{0 \neq w_1 \in W_1 \\ 0 \neq w_2 \in W_2}} \frac{|w_1^T \mathcal{A} w_2|}{(w_1^T \mathcal{A} w_1)^{\frac{1}{2}} (w_2^T \mathcal{A} w_2)^{\frac{1}{2}}},$$

where the subspaces W_1 and W_2 are defined by

$$W_1 = \left\{ v = \begin{bmatrix} x_1 \\ 0 \end{bmatrix} : x_1 \in R^{n_1} \right\}, \quad W_2 = \left\{ v = \begin{bmatrix} 0 \\ x_2 \end{bmatrix} : x_2 \in R^{n_2} \right\}.$$

Then the following hold:

- Let $\mathcal{D} = \begin{bmatrix} \mathcal{A}_{11} & \\ & \mathcal{A}_{22} \end{bmatrix}$ be the block diagonal matrix of $\mathcal{A}$ and $\mathcal{I} \in R^{n_2 \times n_2}$ be an identity matrix. Then $\kappa(\mathcal{I} - \mathcal{C}) \leq \kappa(\mathcal{D}^{-1}\mathcal{A})$.
- $\gamma_{\mathcal{A}}^2 = \rho(\mathcal{C}) < 1$.

Proof. Let $\mathcal{S} = \mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}$ be the Schur complement of $\mathcal{A}$ with respect to $\mathcal{A}_{11}$. According to [2, 26], it holds that

$$(3.12) \quad \gamma_A < 1, \quad \kappa(\mathcal{D}^{-1}\mathcal{A}) = \frac{1 + \gamma_A}{1 - \gamma_A}, \quad \kappa(\mathcal{A}_{22}^{-1}\mathcal{S}) \leq \frac{1}{1 - \gamma_A^2},$$

and then we obtain the inequality

$$(3.13) \quad \kappa(\mathcal{I} - \mathcal{C}) = \kappa(\mathcal{A}_{22}^{-1}\mathcal{S}) \leq \frac{1}{1 - \gamma_A^2} = \frac{1}{(1 + \gamma_A)^2} \frac{1 + \gamma_A}{1 - \gamma_A} \leq \frac{1 + \gamma_A}{1 - \gamma_A} = \kappa(\mathcal{D}^{-1}\mathcal{A}).$$

Since $\mathcal{A}$ is a symmetric and positive definite matrix and $\mathcal{A}_{22}^{-\frac{1}{2}}$ is also symmetric, it is easy to verify that $\mathcal{A}_{22}^{-\frac{1}{2}}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-\frac{1}{2}}$ is symmetric and positive semidefinite, and then all of its eigenvalues are nonnegative. We may write the similarity transformation

$$(3.14) \quad \mathcal{C} = \mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12} = \mathcal{A}_{22}^{-\frac{1}{2}} \left(\mathcal{A}_{22}^{-\frac{1}{2}}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-\frac{1}{2}} \right) \mathcal{A}_{22}^{\frac{1}{2}},$$

which implies that $\sigma(\mathcal{C}) = \sigma(\mathcal{A}_{22}^{-\frac{1}{2}}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-\frac{1}{2}})$, and it follows that

$$(3.15) \quad 0 \leq \lambda \leq \rho(\mathcal{C}) = \rho(\mathcal{A}_{22}^{-\frac{1}{2}}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-\frac{1}{2}}) \quad \forall \lambda \in \sigma(\mathcal{C}).$$

Since $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$ are invertible, we have

$$\begin{aligned} (3.16) \quad \gamma_A^2 &= \sup_{\substack{0 \neq v_1 \in R^{n_1} \\ 0 \neq v_2 \in R^{n_2}}} \frac{(v_1^T \mathcal{A}_{12} v_2)^2}{(v_1^T \mathcal{A}_{11} v_1)(v_2^T \mathcal{A}_{22} v_2)} \\ &= \sup_{\substack{0 \neq \hat{v}_1 \in R^{n_1} \\ 0 \neq \hat{v}_2 \in R^{n_2}}} \frac{(\hat{v}_1^T \mathcal{A}_{11}^{-\frac{1}{2}} \mathcal{A}_{12} \mathcal{A}_{22}^{-\frac{1}{2}} \hat{v}_2)^2}{(\hat{v}_1^T \hat{v}_1)(\hat{v}_2^T \hat{v}_2)} \\ &= \|\mathcal{A}_{11}^{-\frac{1}{2}} \mathcal{A}_{12} \mathcal{A}_{22}^{-\frac{1}{2}}\|^2 \\ &= \rho(\mathcal{A}_{22}^{-\frac{1}{2}} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12} \mathcal{A}_{22}^{-\frac{1}{2}}). \end{aligned}$$

By (3.12) and (3.15), we complete the proof of (b). $\square$

Remark 3.6. Let σ_1 be the largest singular value of $A \in R^{m \times n}$; then

$$(3.17) \quad \|A\| = \sigma_1 = \max\{|v_1^T A v_2| \mid v_1 \in R^m, \|v_1\| = 1, v_2 \in R^n, \|v_2\| = 1\},$$

which is used in (3.16), and the proof can be found in [9].

Remark 3.7. Under Assumption 3.1, we assume that $\mathcal{A}$ in (3.1) is a symmetric positive definite matrix. Since $\mathcal{A}_{22}(\mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12})\mathcal{A}_{22}^{-1} = \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-1}$, i.e., $\sigma(\mathcal{C}) = \sigma(\mathcal{C}^T)$, it follows that

$$(3.18) \quad \kappa(\mathcal{I} - \mathcal{C}) = \kappa((\mathcal{I} - \mathcal{C})^T).$$

4. Derivation of approximate error bounds for nonlinear problems.

Given a nonlinear system of equations $F : \hat{D} \subset R^n \rightarrow R^n$, we want to find a vector $x^* \in R^n$ such that

$$(4.1) \quad F(x^*) = 0.$$

Here $F(x) = [F_1, F_2, \dots, F_n]^T$, $F_i = F_i(x) = F_i(x_1, \dots, x_n)$, and $x = [x_1, x_2, \dots, x_n]^T$. The nonlinear system $F(x)$ is split as

$$(4.2) \quad F(x) = F(u, v) = \begin{bmatrix} G(u, v) \\ H(u, v) \end{bmatrix} = 0, \quad x = [u, v]^T,$$

where G is a vector of length n_1 , H is a vector of length n_2 , the discretization of u has n_1 unknowns, the discretization of v has n_2 unknowns, and $n_1 + n_2 = n$.

Here we are interested in a linear functional of the coupled solution as follows:

$$(4.3) \quad J(u, v) = \langle \psi_1, u \rangle + \langle \psi_2, v \rangle$$

for prescribed $\psi_1 \in R^{n_1}$ and $\psi_2 \in R^{n_2}$, where $\langle y, z \rangle = y^T z$ is the standard vector inner product for vectors y and z , and $\|y\| = \sqrt{\langle y, y \rangle}$. In this paper, we will bound the error in approximating $J(u, v)$ by an approximate solution for u, v .

For the nonlinear problem (4.2) we make the following assumptions for $F(x)$.

Assumption 4.1. Let $U = B(x^*, r) = \{y \mid \|y - x^*\| < r\}$ represent an open ball centered at x^* with its radius r .

(A1) The function $F(x)$ is well-defined in $U \subset \hat{D}$.

(A2) The Jacobian

$$(4.4) \quad F'(x) = \begin{bmatrix} \frac{\partial G}{\partial u} & \frac{\partial G}{\partial v} \\ \frac{\partial H}{\partial u} & \frac{\partial H}{\partial v} \end{bmatrix}$$

is continuous in U and $F'(x^*)$ is nonsingular.

(A3) The submatrices $\frac{\partial G}{\partial u}$ and $\frac{\partial H}{\partial v}$ are invertible at the exact solution x^* .

Suppose $x^{(k)}$ is the current approximate solution. We describe a global inexact Newton method with backtracking (INB) [10], which is also the basis of nonlinear preconditioners such as INB-NE and MSPIN. The major computational cost of INB is spent on the (approximate) solution of the linear system

$$(4.5) \quad F'(x^{(k)})d^{(k)} = F(x^{(k)})$$

for the inexact Newton direction $d^{(k)}$ by enforcing that

$$(4.6) \quad \|F(x^{(k)}) - F'(x^{(k)})d^{(k)}\| \leq \eta_k \|F(x^{(k)})\|$$

is satisfied for a given η_k , and then the updated approximate solution is computed through

$$(4.7) \quad x^{(k+1)} = x^{(k)} + \lambda^{(k)} d^{(k)}.$$

The parameter η_k , called a “forcing term,” determines how accurately the Jacobian linear system needs to be solved using iterative, e.g., Krylov subspace, methods. Another parameter $\lambda^{(k)}$ determines the step size along the inexact Newton direction $d^{(k)}$. It is picked such that

$$(4.8) \quad \hat{f}(x^{(k)} - \lambda^{(k)} d^{(k)}) \leq \hat{f}(x^{(k)}) - \alpha \lambda^{(k)} \nabla \hat{f}(x^{(k)})^T d^{(k)},$$

where the merit function $\hat{f}(x) = \frac{1}{2} \|F(x)\|^2$. The parameter α guarantees the reduction of $\|F(x^{(k)})\|$ and is set small in practice (e.g., $\alpha = 10^{-4}$).

4.1. Approximate error bounds for INB-NE. In the INB-NE algorithm, we consider the variable u as the “bad” component to be eliminated from the nonlinear problem (4.2). The solvability of $G(u, v) = 0$ for v in an appropriate set results in an implicit function $u = f(v)$ such that

$$(4.9) \quad G(u, v) = G(f(v), v) = 0,$$

$$(4.10) \quad H(u, v) = H(f(v), v) = 0.$$

To simplify the coming discussion, we write the Jacobian matrix of $F(x)$ as

$$(4.11) \quad F'(x) = F'(u, v) = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix}.$$

Differentiating (4.9) as a function of v leads to

$$(4.12) \quad \mathcal{A}_{11}f'(v) + \mathcal{A}_{12} = 0,$$

and we assume that $\mathcal{A}_{11}$ is nonsingular. Then

$$(4.13) \quad f'(v) = -\mathcal{A}_{11}^{-1}\mathcal{A}_{12}.$$

Let $\mathcal{H}(v) := H(f(v), v) = 0$. Then

$$(4.14) \quad \mathcal{H}'(v) = \mathcal{A}_{21}f'(v) + \mathcal{A}_{22} = \mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}.$$

The Newton direction equation associated with $\mathcal{H}(v)$ at $\bar{v}$,

$$(4.15) \quad \mathcal{H}'(\bar{v})\delta v = -\mathcal{H}(\bar{v}),$$

can be embedded into

$$(4.16) \quad F'(\bar{u}, \bar{v})\delta x = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix} \begin{bmatrix} \delta u \\ \delta v \end{bmatrix} = \begin{bmatrix} 0 \\ -\mathcal{H}(\bar{v}) \end{bmatrix} = -F(\bar{u}, \bar{v}), \quad \bar{u} = f(\bar{v}),$$

which implies that we can obtain the Newton direction δv in (4.15) via solving the global Jacobian system at the point $[\bar{u}, \bar{v}]^T$ instead of assembling the Schur complement $\mathcal{H}'(\bar{v})$ in (4.14).

In the framework of INB-NE, the k th approximate solution is denoted by $x^{(k)} = [u^{(k)}, \hat{v}^{(k)}]^T$. A new starting point $\hat{x}^{(k)} = [\hat{u}^{(k)}, \hat{v}^{(k)}]^T$ is provided for the outer Newton iteration in the subspace correction phase, which satisfies

$$(4.17) \quad G(\hat{u}^{(k)}, \hat{v}^{(k)}) = G(f(\hat{v}^{(k)}), \hat{v}^{(k)}) = 0, \quad \hat{u}^{(k)} = f(\hat{v}^{(k)}),$$

and residuals associated with the solutions $\hat{u}^{(k)}$ and $\hat{v}^{(k)}$ are defined as

$$(4.18) \quad R^{(k)} = H(f(\hat{v}^{(k)}), \hat{v}^{(k)}).$$

The Jacobian matrix at the approximate solution $\hat{x}^{(k)}$ is computed by

$$(4.19) \quad F'(\hat{x}^{(k)}) = F'(\hat{u}^{(k)}, \hat{v}^{(k)}) = \begin{bmatrix} \mathcal{A}_{11}^{(k)} & \mathcal{A}_{12}^{(k)} \\ \mathcal{A}_{21}^{(k)} & \mathcal{A}_{22}^{(k)} \end{bmatrix},$$

Integrating (4.11), we linearize (4.2) with respect to $\hat{u}^{(k)}$ and $\hat{v}^{(k)}$ and obtain

$$(4.20) \quad G(u, v) = G(\hat{u}^{(k)}, \hat{v}^{(k)}) + \mathcal{A}_{11}^{(k)}(u - \hat{u}^{(k)}) + \mathcal{A}_{12}^{(k)}(v - \hat{v}^{(k)}) + R_G^{(k)},$$

$$(4.21) \quad H(u, v) = H(\hat{u}^{(k)}, \hat{v}^{(k)}) + \mathcal{A}_{21}^{(k)}(u - \hat{u}^{(k)}) + \mathcal{A}_{22}^{(k)}(v - \hat{v}^{(k)}) + R_H^{(k)},$$

or

$$(4.22) \quad F(x) = F(\hat{x}^{(k)}) + F'(\hat{x}^{(k)})(x - \hat{x}^{(k)}) + \begin{bmatrix} R_G^{(k)} \\ R_H^{(k)} \end{bmatrix},$$

and the final term tells us how nonlinear the function $F(x)$ is in the neighborhood of $\hat{x}^{(k)}$. According to Assumption 4.1, we know that U is convex. Therefore, it holds that $\hat{x}^{(k)} + t(x - \hat{x}^{(k)}) \in U$ for $t \in [0, 1]$ and $x \in U$. Applying the integral form of the mean value theorem, we have

$$(4.23) \quad F(x) = F(\hat{x}^{(k)}) + \int_0^1 F'(\hat{x}^{(k)} + t(x - \hat{x}^{(k)}))(x - \hat{x}^{(k)}) dt \quad \forall x, \hat{x}^{(k)} \in U,$$

and then it follows that

$$(4.24) \quad F(x) = F(\hat{x}^{(k)}) + F'(\hat{x}^{(k)})(x - \hat{x}^{(k)}) + \int_0^1 \left(F'(\hat{x}^{(k)} + t(x - \hat{x}^{(k)})) - F'(\hat{x}^{(k)}) \right) (x - \hat{x}^{(k)}) dt.$$

Comparing (4.22) with (4.24), it is easy to obtain

$$(4.25) \quad \begin{bmatrix} R_G^{(k)} \\ R_H^{(k)} \end{bmatrix} = \int_0^1 \left(F'(\hat{x}^{(k)} + t(x - \hat{x}^{(k)})) - F'(\hat{x}^{(k)}) \right) (x - \hat{x}^{(k)}) dt.$$

In order to give an estimate of $R_G^{(k)}$ and $R_H^{(k)}$, we introduce the following assumption under the framework of affine invariance [11, 16].

Assumption 4.2. The affine contravariant Lipschitz condition holds: there exists a positive real constant ω_F such that

$$(4.26) \quad \|(F'(y) - F'(z))(y - z)\| \leq \omega_F \|F'(z)(y - z)\|^2 \quad \forall y, z \in U.$$

Here we define ω_F as a global affine contravariant Lipschitz constant.

Using the above assumption, we obtain the estimate of the remainder terms in (4.22) and (4.24):

$$(4.27) \quad \left\| \begin{bmatrix} R_G^{(k)} \\ R_H^{(k)} \end{bmatrix} \right\| = \left\| \int_0^1 \left(F'(\hat{x}^{(k)} + t(x - \hat{x}^{(k)})) - F'(\hat{x}^{(k)}) \right) (x - \hat{x}^{(k)}) dt \right\| \\ \leq \int_0^1 t \omega_F \|F'(\hat{x}^{(k)})(x - \hat{x}^{(k)})\|^2 dt \\ = \frac{1}{2} \omega_F \|F'(\hat{x}^{(k)})(x - \hat{x}^{(k)})\|^2, \quad x = [u, v]^T.$$

Thus, it makes sense to have the linear approximation

$$(4.28) \quad F(x) \approx F(\hat{x}^{(k)}) + F'(\hat{x}^{(k)})(x - \hat{x}^{(k)}) \quad \text{as } x \rightarrow \hat{x}^{(k)}.$$

As a matter of fact, we claim that linearizations in (4.20) and (4.21) are reasonable, even if $F(x) = 0$ is a nonlinearly stiff problem. In the subspace correction phase of INB-NE, nonlinear “imbalance” is handled by removing appropriate variables deemed to cause trouble, and it will provide a new starting point $\hat{x}^{(k)} = [\hat{u}^{(k)}, \hat{v}^{(k)}]^T$ for the outer Newton iteration. The function $F(x)$ has more balanced nonlinearities in the neighborhood of $\hat{x}^{(k)}$, namely, the difference between $F(x)$ and the linear model $F(\hat{x}^{(k)}) + F'(\hat{x}^{(k)})(x - \hat{x}^{(k)})$ is not very large. When $\hat{x}^{(k)}$ is approaching the exact solution x^* , therefore, it is plausible to ignore the remainder terms $R_G^{(k)}$ and $R_H^{(k)}$.

Using (4.9), (4.10), and (4.17), substituting (4.18) into (4.21), and discarding high order terms yields

$$(4.29) \quad \mathcal{A}_{11}^{(k)}(u - \hat{u}^{(k)}) + \mathcal{A}_{12}^{(k)}(v - \hat{v}^{(k)}) \approx 0,$$

$$(4.30) \quad R^{(k)} + \mathcal{A}_{22}^{(k)}(I - \mathcal{C}_1^{(k)})(v - \hat{v}^{(k)}) \approx 0,$$

where $\mathcal{C}_1^{(k)} = \mathcal{A}_{22}^{(k)-1} \mathcal{A}_{21}^{(k)} \mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)}$. Under Assumption 4.1, it is easy to know that $F'(\hat{x}^{(k)})$, $\mathcal{A}_{11}^{(k)}$, and $\mathcal{A}_{22}^{(k)}$ are invertible in a suitable neighborhood of the exact solution x^* . According to Lemma 3.2, $(I - \mathcal{C}_1^{(k)})^{-1}$ exists and it follows from (4.29) and (4.30) that

$$(4.31) \quad u - \hat{u}^{(k)} \approx \mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)} (I - \mathcal{C}_1^{(k)})^{-1} \mathcal{A}_{22}^{(k)-1} R^{(k)},$$

$$(4.32) \quad v - \hat{v}^{(k)} \approx -(I - \mathcal{C}_1^{(k)})^{-1} \mathcal{A}_{22}^{(k)-1} R^{(k)}.$$

Next, we compute the error induced in the linear functional (4.3),

$$(4.33) \quad \begin{aligned} J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)}) &= \langle \psi_1, u - \hat{u}^{(k)} \rangle + \langle \psi_2, v - \hat{v}^{(k)} \rangle \\ &\approx \langle \psi_1, \mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)} (I - \mathcal{C}_1^{(k)})^{-1} \mathcal{A}_{22}^{(k)-1} R^{(k)} \rangle \\ &\quad - \langle \psi_2, (I - \mathcal{C}_1^{(k)})^{-1} \mathcal{A}_{22}^{(k)-1} R^{(k)} \rangle, \end{aligned}$$

from which we may deduce the following theorem describing error bounds.

THEOREM 4.3. *At the k th iteration in the INB-NE algorithm, the approximate solution $\hat{x}^{(k)} = [\hat{u}^{(k)}, \hat{v}^{(k)}]^T$ as a new starting point for the outer Newton iteration is provided by the subspace correction phase. Let $\mathcal{C}_1^{(k)} = \mathcal{A}_{22}^{(k)-1} \mathcal{A}_{21}^{(k)} \mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)}$. We define the error induced in the linear functional (4.3) as*

$$(4.34) \quad \Delta J_k = |J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)})|.$$

Then the approximate error bound is given by

$$(4.35) \quad \Delta J_k \lesssim \|\mathcal{A}_{22}^{(k)-1} R^{(k)}\| \left(\|(I - \mathcal{C}_1^{(k)})^{-T} \mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1\| + \|(I - \mathcal{C}_1^{(k)})^{-T} \psi_2\| \right).$$

If $\|\mathcal{C}_1^{(k)}\| < 1$, we further derive

$$(4.36) \quad \Delta J_k \lesssim \|\mathcal{A}_{22}^{(k)-1} R^{(k)}\| (\|\mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1\| + \|\psi_2\|) \frac{1}{1 - \|\mathcal{C}_1^{(k)}\|}.$$

Proof. Following (4.33), it is easy to obtain

$$\begin{aligned} & |J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)})| \\ & \approx |\langle (I - \mathcal{C}_1^{(k)})^{-T} \mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1, \mathcal{A}_{22}^{(k)-1} R^{(k)} \rangle - \langle (I - \mathcal{C}_1^{(k)})^{-T} \psi_2, \mathcal{A}_{22}^{(k)-1} R^{(k)} \rangle| \\ & \leq \|\mathcal{A}_{22}^{(k)-1} R^{(k)}\| (\|(I - \mathcal{C}_1^{(k)})^{-T} \mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1\| + \|(I - \mathcal{C}_1^{(k)})^{-T} \psi_2\|). \end{aligned}$$

Now, we prove the last inequality (4.36). If $\|\mathcal{C}_1^{(k)}\| < 1$, it follows from (4.35) that

$$\begin{aligned} & |J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)})| \\ & \lesssim \|\mathcal{A}_{22}^{(k)-1} R^{(k)}\| (\|\mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1\| + \|\psi_2\|) \|(I - \mathcal{C}_1^{(k)})^{-T}\| \\ & \leq \|\mathcal{A}_{22}^{(k)-1} R^{(k)}\| (\|\mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1\| + \|\psi_2\|) \frac{1}{1 - \|\mathcal{C}_1^{(k)}\|}, \end{aligned}$$

which completes the proof. $\square$

It is noted that approximate error bounds given by (4.36) are independent of the choice of ψ_2 , i.e., we always get the same error bound for different ψ_2 that satisfy $\|\psi_2\| = 1$. In contrast, the bound given by (4.35) depends on the action of $(I - \mathcal{C}_1^{(k)})^{-T}$ on ψ_1 and ψ_2 , and therefore it will give rise to different predictions for convergence performance corresponding to linear functionals defined in (4.3) with different ψ_1 and ψ_2 . In practical computation in (4.35), the multiplication of a given vector with $(I - \mathcal{C}_1^{(k)})^{-T}$ is equivalent to solving a linear system. However, the matrix $(I - \mathcal{C}_1^{(k)})^T$ has never been formed explicitly; we simply provide the matrix-vector multiplication for Krylov subspace methods such as GMRES [28]. Similar to Jacobian matrices in left nonlinear preconditioners, further preconditioning using matrix-based techniques is often difficult due to lack of the explicit format of $(I - \mathcal{C}_1^{(k)})^T$. When the Jacobian matrix $F'(\hat{x}^{(k)})$ is symmetric and positive definite, following Lemma 3.5, $(I - \mathcal{C}_1^{(k)})^T$ is not ill-conditioned if the condition number of the Jacobian matrix with the block Jacobi preconditioner is not very large.

Remark 4.4. The condition $\|\mathcal{C}_1^{(k)}\| < 1$ is not unreasonable strict. According to Lemma 3.4, if $\mathcal{A}^{(k)}$ is a row block strictly diagonally dominant, then it satisfies

$$(4.37) \quad \|\mathcal{C}_1^{(k)}\| = \|\mathcal{A}_{22}^{(k)-1} \mathcal{A}_{21}^{(k)} \mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)}\| < 1.$$

In addition, under the assumption of sufficiently weak dependency of the first model with respect to the component v , it is reasonable to assume that $\|\mathcal{A}_{12}^{(k)}\|$ is sufficiently small that

$$(4.38) \quad \|\mathcal{C}_1^{(k)}\| \leq \|\mathcal{A}_{22}^{(k)-1}\| \|\mathcal{A}_{21}^{(k)}\| \|\mathcal{A}_{11}^{(k)-1}\| \|\mathcal{A}_{12}^{(k)}\| < 1,$$

which also allows us to employ (4.36) to get error bounds.

Remark 4.5. The error bounds given by (4.35) and (4.36) are not exact. For example, we may more precisely write (4.35) as

$$\Delta J_k \leq \|\mathcal{A}_{22}^{(k)-1} R^{(k)}\| \left(\|(I - \mathcal{C}_1^{(k)})^{-T} \mathcal{A}_{12}^{(k)T} \mathcal{A}_{11}^{(k)-T} \psi_1\| + \|(I - \mathcal{C}_1^{(k)})^{-T} \psi_2\| \right) + |r_F^{(k)}|,$$

where

$$(4.39) \quad r_F^{(k)} = -\langle \psi_1, \mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)} (I - \mathcal{C}_1^{(k)})^{-1} \mathcal{A}_{22}^{(k)-1} (\mathcal{A}_{21}^{(k)} \mathcal{A}_{11}^{(k)-1} R_G^{(k)} - R_H^{(k)}) \rangle \\ - \langle \psi_1, \mathcal{A}_{11}^{(k)-1} R_G^{(k)} \rangle + \langle \psi_2, (I - \mathcal{C}_1^{(k)})^{-1} \mathcal{A}_{22}^{(k)-1} (\mathcal{A}_{21}^{(k)} \mathcal{A}_{11}^{(k)-1} R_G^{(k)} - R_H^{(k)}) \rangle.$$

We give a rough estimate of $r_F^{(k)}$:

$$|r_F^{(k)}| \leq \|\psi_1\| \|\mathcal{A}_{11}^{(k)-1} R_G^{(k)}\| \\ + \|\psi_1\| \|\mathcal{A}_{11}^{(k)-1} \mathcal{A}_{12}^{(k)}\| \|(I - \mathcal{C}_1^{(k)})^{-1}\| \\ \times \left(\|\mathcal{A}_{22}^{(k)-1} \mathcal{A}_{21}^{(k)}\| \|\mathcal{A}_{11}^{(k)-1} R_G^{(k)}\| + \|\mathcal{A}_{22}^{(k)-1} R_H^{(k)}\| \right) \\ + \|\psi_2\| \|(I - \mathcal{C}_1^{(k)})^{-1}\| \left(\|\mathcal{A}_{22}^{(k)-1} \mathcal{A}_{21}^{(k)}\| \|\mathcal{A}_{11}^{(k)-1} R_G^{(k)}\| + \|\mathcal{A}_{22}^{(k)-1} R_H^{(k)}\| \right).$$

Assuming that $\mathcal{A}^{(k)}$ is a row block strictly diagonally dominant and using (4.27), the above inequality can be further simplified as follows:

$$(4.40) \quad |r_F^{(k)}| \leq \|\psi_1\| \|\mathcal{A}_{11}^{(k)-1} R_G^{(k)}\| + \frac{\|\psi_1\| + \|\psi_2\|}{1 - \|\mathcal{C}_1^{(k)}\|} \left(\|\mathcal{A}_{11}^{(k)-1} R_G^{(k)}\| + \|\mathcal{A}_{22}^{(k)-1} R_H^{(k)}\| \right) \\ \leq \frac{1}{2} \omega_F \|\psi_1\| \|\mathcal{A}_{11}^{(k)-1}\| \|F'(\hat{x}^{(k)})(x^* - \hat{x}^{(k)})\|^2 \\ + \frac{\omega_F (\|\psi_1\| + \|\psi_2\|) \left(\|\mathcal{A}_{11}^{(k)-1}\| + \|\mathcal{A}_{22}^{(k)-1}\| \right)}{2(1 - \|\mathcal{C}_1^{(k)}\|)} \|F'(\hat{x}^{(k)})(x^* - \hat{x}^{(k)})\|^2.$$

Remark 4.6. The INB-NE algorithm can be regarded as the generalization of Gaussian elimination in the nonlinear case. If we have the linear system

$$(4.41) \quad \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

In the INB-NE algorithm, we equivalently solve the linear system

$$(4.42) \quad \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathbf{0} & \mathcal{A}_{22} - \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 - \mathcal{A}_{21} \mathcal{A}_{11}^{-1} b_1 \end{bmatrix}$$

instead of the original linear system (4.41), which is exactly the same as Gaussian elimination.

4.2. Approximate error bounds for MSPIN. We consider (4.2) as the multiphysics problem with two components. In the MSPIN algorithm, the submodels are solved sequentially for the physical variable corrections, which implicitly forms the preconditioned system. For any given $x = [u, v]^T \in R^n$, the preconditioned nonlinear system

$$(4.43) \quad \mathcal{F}(x) = \begin{bmatrix} g(u, v) \\ h(u, v) \end{bmatrix} = 0$$

is obtained by solving

$$(4.44) \quad G(u - g, v) = 0$$

for g . With values of u, v, g , the system

$$(4.45) \quad H(u - g, v - h) = 0$$

is solved for h .

Let $p = u - g(u, v)$ and $q = v - h(u, v)$. We define

$$(4.46) \quad \mathcal{B} = \begin{bmatrix} \frac{\partial G}{\partial p} & \frac{\partial G}{\partial v} \\ \frac{\partial H}{\partial p} & \frac{\partial H}{\partial q} \end{bmatrix} = \begin{bmatrix} \mathcal{B}_{11} & \mathcal{B}_{12} \\ \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix}.$$

According to Assumption 4.1, we know that $\mathcal{B}$, $\mathcal{B}_{11}$, and $\mathcal{B}_{22}$ are invertible in the suitable neighborhood of the exact solution x^* . As shown in [21], the derivatives of g and h with respect to u and v are written as

$$(4.47) \quad \frac{\partial g}{\partial u} = I_u, \quad \frac{\partial g}{\partial v} = \mathcal{B}_{11}^{-1} \mathcal{B}_{12},$$

$$(4.48) \quad \frac{\partial h}{\partial u} = 0, \quad \frac{\partial h}{\partial v} = I_v - \mathcal{B}_{22}^{-1} \mathcal{B}_{21} \mathcal{B}_{11}^{-1} \mathcal{B}_{12},$$

where I_u and I_v are the identity matrices that have the same dimension as the u and v blocks, respectively.

In the framework of MSPIN, the k th approximate solution components $\hat{u}^{(k)}$ and $\hat{v}^{(k)}$ satisfy

$$\begin{aligned} G(\hat{u}^{(k)} - g(\hat{u}^{(k)}, \hat{v}^{(k)}), \hat{v}^{(k)}) &= 0, \\ H(\hat{u}^{(k)} - g(\hat{u}^{(k)}, \hat{v}^{(k)}), \hat{v}^{(k)} - h(\hat{u}^{(k)}, \hat{v}^{(k)})) &= 0, \end{aligned}$$

and residuals associated with the solutions $\hat{u}$ and $\hat{v}$ are defined as

$$(4.49) \quad R_1^{(k)} = g(\hat{u}^{(k)}, \hat{v}^{(k)}),$$

$$(4.50) \quad R_2^{(k)} = h(\hat{u}^{(k)}, \hat{v}^{(k)}).$$

Integrating (4.47) and (4.48), we linearize (4.43) with respect to $\hat{u}$ and $\hat{v}$ and obtain

$$(4.51) \quad g(u, v) = g(\hat{u}^{(k)}, \hat{v}^{(k)}) + (u - \hat{u}^{(k)}) + \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)}(v - \hat{v}^{(k)}) + R_g^{(k)},$$

$$(4.52) \quad h(u, v) = h(\hat{u}^{(k)}, \hat{v}^{(k)}) + (I - \mathcal{B}_{22}^{(k)-1} \mathcal{B}_{21}^{(k)} \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)})(v - \hat{v}^{(k)}) + R_h^{(k)}.$$

Applying the integral form of the mean value theorem and comparing with (4.51) and (4.52), we have

$$(4.53) \quad \begin{bmatrix} R_g^{(k)} \\ R_h^{(k)} \end{bmatrix} = \int_0^1 \left(\mathcal{F}'(\hat{x}^{(k)} + t(x - \hat{x}^{(k)})) - \mathcal{F}'(\hat{x}^{(k)}) \right) (x - \hat{x}^{(k)}) dt.$$

In order to give an estimate of $R_g^{(k)}$ and $R_h^{(k)}$, we introduce the following assumption.

Assumption 4.7. $F(x)$ is twice continuously differentiable in U .

Under Assumption 4.7 and Lemma 4.2 in [22], there is a neighborhood $U_1 \subset B(x^*, \frac{\tau}{2})$ of the exact solution x^* such that the Lipschitz condition holds:

$$(4.54) \quad \|\mathcal{F}'(y) - \mathcal{F}'(z)\| \leq L_{\mathcal{F}} \|y - z\| \quad \forall y, z \in U_1.$$

Here $L_{\mathcal{F}}$ is the Lipschitz constant. Then, it follows that the remainder terms in (4.51) and (4.52) can be estimated by

$$(4.55) \quad \left\| \begin{bmatrix} R_g^{(k)} \\ R_h^{(k)} \end{bmatrix} \right\| \leq \frac{1}{2} L_{\mathcal{F}} \|x - \hat{x}^{(k)}\|^2 \quad \forall x, \hat{x}^{(k)} \in U_1.$$

In fact, when the original system $F(x)$ has strong nonlinearities, high order terms tell us how nonlinear the function is, so it is not appropriate to approximate $F(x)$ using the linear model $F(\hat{x}^{(k)}) + F'(\hat{x}^{(k)})(x - \hat{x}^{(k)})$ in the neighborhood of $\hat{x}^{(k)}$; in contrast, the linearizations above of $\mathcal{F}(x)$ in (4.51) and (4.52) are reasonable. The MSPIN algorithm creates a modified global residual function $\mathcal{F}(x)$ with better conditioning and better balanced nonlinearities than the original system from the union of subspace corrections of inner Newton iterations. In the neighborhood of $\hat{x}^{(k)} = [\hat{u}^{(k)}, \hat{v}^{(k)}]^T$, therefore, it makes sense to discard the remainder terms $R_g^{(k)}$ and $R_h^{(k)}$.

Using (4.43), (4.49), and (4.50), we may write (4.51) and (4.52) as

$$\begin{aligned} R_1^{(k)} + (u - \hat{u}^{(k)}) + \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)}(v - \hat{v}^{(k)}) &\approx 0, \\ R_2^{(k)} + (I - \mathcal{C}_2^{(k)-1} \mathcal{B}_{21}^{(k)} \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)})(v - \hat{v}^{(k)}) &\approx 0, \end{aligned}$$

which allow us to deduce that

$$(4.56) \quad v - \hat{v}^{(k)} \approx -(I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)},$$

$$(4.57) \quad u - \hat{u}^{(k)} \approx -R_1^{(k)} + \mathcal{B}_{11}^{-1} \mathcal{B}_{12} (I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)},$$

where $\mathcal{C}_2^{(k)} = \mathcal{B}_{22}^{(k)-1} \mathcal{B}_{21}^{(k)} \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)}$.

Next, we are interested in bounding the error induced in the linear functional (4.3) as follows:

$$\begin{aligned} J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)}) &= \langle \psi_1, u - \hat{u}^{(k)} \rangle + \langle \psi_2, v - \hat{v}^{(k)} \rangle \\ (4.58) \quad &\approx -\langle \psi_1, R_1^{(k)} \rangle + \langle \psi_1, \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)} (I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)} \rangle \\ &\quad - \langle \psi_2, (I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)} \rangle; \end{aligned}$$

we deduce the following result similar to Theorem 4.3 for the MSPIN algorithm.

THEOREM 4.8. *At the k th iteration in the MSPIN algorithm, the approximate solution is denoted by $\hat{x}^{(k)} = [\hat{u}^{(k)}, \hat{v}^{(k)}]^T$ and we let $\mathcal{C}_2^{(k)} = \mathcal{B}_{22}^{(k)-1} \mathcal{B}_{21}^{(k)} \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)}$. We define the error induced in the linear functional (4.3) as*

$$(4.59) \quad \Delta J_k = |J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)})|.$$

Then the approximate error bound is given by

$$(4.60) \quad \begin{aligned} \Delta J_k &\lesssim |\langle \psi_1, R_1^{(k)} \rangle| + \|(I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)}\| \|\mathcal{B}_{12}^{(k)T} \mathcal{B}_{11}^{(k)-T} \psi_1\| \\ &\quad + \|R_2\| \|(I - \mathcal{C}_2^{(k)})^{-T} \psi_2\|. \end{aligned}$$

If $\|\mathcal{C}_2^{(k)}\| < 1$, we further derive

$$(4.61) \quad \Delta J_k \lesssim |\langle \psi_1, R_1^{(k)} \rangle| + \|R_2^{(k)}\| (\|\mathcal{B}_{12}^{(k)T} \mathcal{B}_{11}^{(k)-T} \psi_1\| + \|\psi_2\|) \frac{1}{1 - \|\mathcal{C}_2^{(k)}\|}.$$

Proof. We prove the last inequality (4.61) only; the rest can be proved in the same way. When $\|\mathcal{C}_2^{(k)}\| < 1$, using the Cauchy–Schwarz inequality, it follows from (4.58) that

$$\begin{aligned}
& |J(u, v) - J(\hat{u}^{(k)}, \hat{v}^{(k)})| \\
& \lesssim |\langle \psi_1, R_1^{(k)} \rangle| + |\langle \mathcal{B}_{12}^{(k)T} \mathcal{B}_{11}^{(k)-T} \psi_1, (I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)} \rangle| + |\langle \psi_2, (I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)} \rangle| \\
& \leq |\langle \psi_1, R_1^{(k)} \rangle| + (\|\mathcal{B}_{12}^{(k)T} \mathcal{B}_{11}^{(k)-T} \psi_1\| + \|\psi_2\|) \|(I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)}\| \\
& \leq |\langle \psi_1, R_1^{(k)} \rangle| + \|R_2^{(k)}\| (\|\mathcal{B}_{12}^{(k)T} \mathcal{B}_{11}^{(k)-T} \psi_1\| + \|\psi_2\|) \frac{1}{1 - \|\mathcal{C}_2^{(k)}\|}.
\end{aligned}$$

Remark 4.9. It is not convenient to directly compute B in (4.46); in practice, we instead use the approximation $F'(z)|_{z=[p,v]}^T$.

Remark 4.10. The error bounds given by (4.60) and (4.61) are not exact. For instance, we may more precisely write (4.60) as

$$\begin{aligned}
(4.62) \quad \Delta J_k & \leq |\langle \psi_1, R_1^{(k)} \rangle| + \|(I - \mathcal{C}_2^{(k)})^{-1} R_2^{(k)}\| \|\mathcal{B}_{12}^{(k)T} \mathcal{B}_{11}^{(k)-T} \psi_1\| \\
& \quad + \|R_2\| \|(I - \mathcal{C}_2^{(k)})^{-T} \psi_2\| + |r_{\mathcal{F}}^{(k)}|,
\end{aligned}$$

where

$$\begin{aligned}
(4.63) \quad r_{\mathcal{F}}^{(k)} & = \langle \psi_1, \mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)} (I - \mathcal{C}_2^{(k)})^{-1} R_h^{(k)} \rangle - \langle \psi_1, R_g^{(k)} \rangle \\
& \quad - \langle \psi_2, (I - \mathcal{C}_2^{(k)})^{-1} R_h^{(k)} \rangle.
\end{aligned}$$

Using (4.14), we give a rough estimate of $r_{\mathcal{F}}^{(k)}$:

$$\begin{aligned}
|r_{\mathcal{F}}^{(k)}| & \leq \|\psi_1\| \|\mathcal{B}_{11}^{(k)-1} \mathcal{B}_{12}^{(k)}\| \|(I - \mathcal{C}_2^{(k)})^{-1}\| \|R_h^{(k)}\| + \|\psi_1\| \|R_g^{(k)}\| \\
& \quad + \|\psi_2\| \|(I - \mathcal{C}_2^{(k)})^{-1}\| \|R_h^{(k)}\|.
\end{aligned}$$

Assuming that $\mathcal{B}^{(k)}$ is a row block strictly diagonally dominant and using (4.55), the above inequality can be further simplified:

$$\begin{aligned}
(4.64) \quad |r_{\mathcal{F}}^{(k)}| & \leq \|\psi_1\| \|R_g^{(k)}\| + \frac{\|\psi_1\| + \|\psi_2\|}{1 - \|\mathcal{C}_2^{(k)}\|} \|R_h^{(k)}\| \\
& \leq \frac{1}{2} L_{\mathcal{F}} (\|\psi_1\| + \frac{\|\psi_1\| + \|\psi_2\|}{1 - \|\mathcal{C}_2^{(k)}\|}) \|x^* - \hat{x}^{(k)}\|^2.
\end{aligned}$$

Remark 4.11. In the linear case with two components, the MSPIN algorithm is the same as Gaussian elimination. If we have the linear system

$$(4.65) \quad F(x) = b - \mathcal{A}x = 0,$$

where

$$(4.66) \quad \mathcal{A} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \mathcal{M} = \begin{bmatrix} \mathcal{A}_{11} & 0 \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix},$$

in the MSPIN algorithm, we will solve the preconditioned system

$$(4.67) \quad \mathcal{F}_{MSPIN}(x) = \mathcal{M}^{-1}(b - \mathcal{A}x)$$

instead of the original linear system (4.65). Through the action of the linear preconditioner $\mathcal{M}^{-1}$, the coefficient matrix $\mathcal{A}$ is transformed into an upper triangular matrix,

$$(4.68) \quad \mathcal{M}^{-1}\mathcal{A} = \begin{bmatrix} I & \mathcal{A}_{11}^{-1}\mathcal{A}_{12} \\ 0 & I - \mathcal{A}_{22}^{-1}\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12} \end{bmatrix},$$

which is equivalent to the row echelon form in Gaussian elimination. The difference between $\mathcal{M}^{-1}\mathcal{A}$ and the coefficient matrix in (4.42) corresponding to INB-NE is given by a scaling matrix

$$(4.69) \quad \mathcal{D} = \begin{bmatrix} \mathcal{A}_{11} & \\ & \mathcal{A}_{22} \end{bmatrix}.$$

5. Numerical examples. In this section, we apply approximate error bounds to linear functionals of solutions for two prototype multiphysics problems solved using nonlinear preconditioning techniques. In the first case, we consider a flame sheet model in an axisymmetric opposed jet configuration and employ field-split INB-NE as the solver. This couples the conservation of mass and momentum with a Shvab–Zeldovich formulation for the energy and species fraction. In the second, we implement INB-NE and MSPIN for a combination of lid-driven flow and buoyancy-driven flow by coupling a Navier–Stokes formula in velocity-vorticity form and the energy equation. All tests are done in the Portable Extensible Toolkit for Scientific Computation (PETSc) [3].

5.1. One-dimensional flame sheet. Following [18, 23], we consider a flame sheet model for a diluted methane-air counterflow diffusion flame in a cylindrical double-jet configuration. The governing equations are

$$(5.1) \quad \frac{dV}{dz} + 2a\rho f' = 0,$$

$$(5.2) \quad \frac{d}{dz} \left(\mu \frac{df'}{dz} \right) - V \frac{df'}{dz} + a(\rho_\infty - \rho(f')^2) = 0,$$

$$(5.3) \quad \frac{1}{\text{Pr}_{\text{ref}}} \frac{d}{dz} \left(\mu \frac{dS}{dz} \right) - V \frac{dS}{dz} = 0,$$

where f' is the similarity function, V is the mass flux in the transverse direction, S is the conserved scalar, the Prandtl number is $\text{Pr}_{\text{ref}} = 0.75$, and the strain rate is $a = 40\text{s}^{-1}$. The jets are located at $z = -0.6\text{ cm}$ and $z = 0.8876\text{ cm}$. We impose the boundary conditions $V = 2.8 \times 10^{-2}\text{ gm/cm}^2 \cdot \text{s}$, $f' = 1.216$, $S = 1$ at the fuel jet and $f' = 1$, $S = 0$ at the oxidizer jet. In order to close the nonlinear two-point boundary value problems, we specify a power relationship between the temperature and viscosity given by

$$(5.4) \quad \mu = \mu_0 \left(\frac{T}{T_0} \right)^{0.7}$$

for prescribed constants μ_0 and T_0 . We can track the location of the flame front by solving $S(z_f) = S_f$, where the constant S_f is derived from the stoichiometry of the reaction. Then computational domain is divided into the fuel-rich side ($S(z) > S_f$) and the oxidizer-rich side ($S(z) < S_f$). By assuming an infinite reaction rate and complete consumption at the flame front, the fuel and oxidizer cannot coexist on either side of the flame front. This assumption that the fuel concentration is zero on one side and the oxidizer on the other allows recovery of the temperature and the mass fraction of species from the conserved scalar S with different formulae in each region. Full details are found in [18].

We discretize the fully coupled two-point boundary value problem (5.1)–(5.3) using the finite difference method. It is difficult to solve directly the nonlinear system

using Newton-like methods without a good initial guess. For the sake of fast convergence as well as low computational cost, a pseudotransient continuation technique is often used to provide the initial guess. By appending the terms $\frac{\partial f'}{\partial t}$ and $\frac{\partial S}{\partial t}$ to the left-hand side of (5.2) and (5.3), respectively, we solve the pseudotransient system

$$(5.5) \quad D \frac{X_k - X_{k-1}}{\Delta t_k} + F(X_k) = 0, \quad k = 0, 1, \dots,$$

which brings the approximate solution into the convergence domain of the steady-state solutions as time marching, and then implement full Newton iterations (when $\Delta t_k \gg 1$). No pseudotime-stepping is used for the algebraic constraint equation and boundary conditions [8], so the diagonal matrix D is introduced with entry 1 for the pseudotime-stepped variables and 0 for the rest.

In the interests of space, we report only results for the steady-state problem after preparatory pseudotransient continuation. The numerical solution is obtained on a uniform grid consisting of 501 grid points. For this example, a field-split INB-NE algorithm is carried out by eliminating the conserved scalar S before applying a global Newton iteration. The nonlinear solver requires an initial iterate, and we start with the constant mass flux for V and the linear interpolation using boundary values for f' and S .

Here we consider the following pointwise linear functionals:

$$(5.6) \quad J_1 = V(z), \quad J_2 = f'(z), \quad J_3 = S(z), \quad z = z_1, z_2,$$

where J_1 , J_2 , J_3 refer to the mass flux in the transverse direction, the similarity function, and the conserved scalar, respectively, evaluated at $z_1 = -0.123986$ and $z_2 = 0.29256$. Figure 1 shows convergence history in terms of three unknowns, and we note that the convergence rates at different points z_1 and z_2 are quite different even for the same component of the solution. The solid and dotted lines represent the absolute errors and approximate error bounds, respectively, of J_1 , J_2 , and J_3 evaluated at z_1 and z_2 . The black solid lines represent the norms of the true residual, i.e., $\|F(X_k)\|$. It is evident that the bounds effectively predict the convergence profiles corresponding to the different linear functionals defined in (5.6).

We consider the following three convergence criteria:

- the absolute residual norm,

$$\|F(X_k)\| \leq \epsilon_1;$$

- the relative residual norm,

$$\frac{\|F(X_k)\|}{\|F(X_0)\|} \leq \epsilon_2;$$

- the solution update norm,

$$\|X_{k+1} - X_k\| = \|\Delta X_k\| \leq \epsilon_3.$$

Here ϵ_1 , ϵ_2 , ϵ_3 are the given tolerance parameters. Figure 2 shows convergence history corresponding to the different criteria. The solid line and the dotted line with red stars refer to the absolute error and the approximate error bound using (4.35) of J_3 evaluated at z_2 , respectively; the solid line, the dotted line, and the dashed line with black dots represent the norms of the residual, the relative norms of the residual, and

FIG. 1. In the left-hand column, figures show the convergence history of V , f' , and S . In the right-hand column, the solid lines with blue circles and red stars refer to the absolute errors, respectively, of J_1 , J_2 , and J_3 computed at z_1 and z_2 ; the dotted lines with blue circles and red stars refer to approximate error bounds at z_1 and z_2 computed using (4.35); the black lines represent the norms of the true residual.

the norms of the solution update, respectively. Observe, for example, that the absolute residual norm, the relative residual norm, and the solution update norm are several orders of magnitude larger than both the absolute error and the bound corresponding to the linear functional $J_3 = S(z_2)$. As the default indicators of convergence, they would lead to more Newton iterations than required to resolve this quantity of interest.

5.2. Driven cavity flow problem. Following [4, 8], we consider ex19.c in the scalable nonlinear equations solvers subdirectory of PETSc [3], which describes the two-dimensional flow driven partly by a moving lid and partly by buoyancy (in the Boussinesq approximation) in an upright square cavity $\Omega = (0, 1) \times (0, 1)$ with four unknowns: the velocity fields (u, v) in the (x, y) directions, the vorticity ω , and the

FIG. 2. The solid line and the dotted line with red stars refer to the absolute error and the approximate error bound using (4.35) of J_3 evaluated at z_2 , respectively. The solid line, the dotted line, and the dashed line with black dots represent the norms of the residual, the relative norms of the residual, and the norms of the solution update, respectively.

temperature T . The governing system consists of the nondimensional steady-state Navier–Stokes equations in vorticity-velocity form and energy equation

$$(5.7) \quad \begin{cases} -\Delta u - \frac{\partial \omega}{\partial y} = 0, \\ -\Delta v + \frac{\partial \omega}{\partial x} = 0, \\ -\Delta \omega + u \frac{\partial \omega}{\partial x} + v \frac{\partial \omega}{\partial y} - Gr \frac{\partial T}{\partial x} = 0, \\ -\Delta T + Pr(u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y}) = 0, \end{cases}$$

where Gr and Pr denote the Grashof number and the Prandtl number, respectively. We impose $u = V_{lid}$, $v = 0$ on the top wall (a lid moving with a steady velocity gives rise to a principal vortex and some corner vortices by viscous forces) and no-slip boundary conditions on other rigid walls. Dirichlet boundary conditions are used for the vorticity, derived from its definition,

$$(5.8) \quad \omega(x, y) = -\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}.$$

The insulating homogeneous Neumann conditions and Dirichlet conditions are imposed on the horizontal walls and the vertical walls, respectively. The temperature T is set as 0 on the left wall and 1 on the right wall, and the temperature difference leads to a buoyant vortex flow in the cavity.

The computational domain Ω is divided into a uniform Cartesian grid, and we obtain a numerical solution of the fully coupled boundary value problem (5.7) using the finite difference method. A standard five-point stencil, first-order upwinding, and central differencing are used to approximate the Laplace operators for each component, the convective terms, and the other first-order gradients, respectively. For the initial guess, we set up a motionless, vorticity free flow field and $T_0(x, y) = x$ as the interpolatory temperature field. The stopping condition for the global nonlinear solvers is

$$(5.9) \quad \|F(\hat{x}^{(k)})\| \leq 10^{-10}.$$

For the purpose of applying the field-based nonlinear preconditioners, both the unknowns and the nonlinear system are partitioned into two groups. Taking as one partition the velocity degrees of freedom and the other the vorticity and temperature degrees of freedom, we split the system (5.7) into two submodels:

$$(5.10) \quad F_{u,v} : \begin{cases} -\Delta u - \frac{\partial \omega}{\partial y} = 0, \\ -\Delta v + \frac{\partial \omega}{\partial x} = 0, \end{cases}$$

and

$$(5.11) \quad F_{\omega,T} : \begin{cases} -\Delta \omega + u \frac{\partial \omega}{\partial x} + v \frac{\partial \omega}{\partial y} - Gr \frac{\partial T}{\partial x} = 0, \\ -\Delta T + Pr(u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y}) = 0. \end{cases}$$

As a result, the submodels $F_{u,v}$ and $F_{\omega,T}$ are both linear among their own unknowns, which may be regarded as benefits of physics-based field-split partitioning.

For this example, we are interested in the following four linear functionals of u , v , ω , and T :

$$(5.12) \quad J_1 = \int_0^1 u(0.5, y) dy,$$

$$(5.13) \quad J_2 = v_x(0.5, 0.5),$$

$$(5.14) \quad J_3 = \omega(0.5, 0.5),$$

$$(5.15) \quad J_4 = \int_0^1 \int_0^1 T dx dy.$$

The physical interpretations are given for these functionals as follows: J_1 is the mass flux across the vertical line through the geometric center of the cavity; J_2 is the partial derivative of v along horizontal lines through the geometric center; J_3 is the vorticity at the central point of the computational domain Ω ; J_4 is the average temperature in the cavity.

We use dimensionless values of $Gr = 2 \times 10^4$, $Pr = 1$ and the lid velocity $V_{lid} = 100$ for numerical tests, and Figure 3 shows the converged solutions of the fully coupled

FIG. 3. The velocity and temperature profiles for the solution of the driven cavity flow problem ($Gr = 2 \times 10^4$, $Pr = 1$, and the lid velocity $V_{lid} = 100$).

problem (5.7) for u , v , and T on the 257×257 mesh. We solve the driven cavity flow problem (5.7) using standard INB and nonlinearly preconditioned INB-NE and MSPIN, respectively. For the INB-NE algorithm, we carry out a subspace correction phase by eliminating velocity components u and v . For the MSPIN algorithm, we first solve (5.11) for vorticity and temperature fields, and then solve (5.10) for velocity components. In contrast to 15 iterations using INB, nonlinear preconditioning techniques such as INB-NE and MSPIN are effective in reducing the number of Newton iterations and converge within seven and eight iterations, respectively. Figures 4 and 5 show errors and approximate error bounds for the linear functionals defined by (5.12)–(5.15) corresponding to INB-NE and MSPIN. In all plots, the solid line represents the true value of the absolute errors, and the dotted line refers to the bounds computed using (4.35) and (4.60). It is seen that the approximate error bounds predict different rates of convergence for different linear functionals and become sharper as the approximate solution $\hat{x}^{(k)}$ approaches the exact solution.

Next, we keep the parameters such as Gr , Pr , and V_{lid} fixed but run the tests using INB-NE on the uniform mesh with different resolutions $h = 1/64$, $h = 1/128$, $h = 1/256$, $h = 1/512$, and $h = 1/1024$. Figure 6 shows the true values of the absolute errors and the approximate error bounds in J_1 , J_2 , J_3 , and J_4 on problems of different resolutions. For all tests, it is observed that the approximate error bounds always successfully predict different rates of convergence for different linear functionals. Moreover, the INB-NE method benefits from the application of right nonlinear preconditioners such that the number of nonlinear iterations does not change much when the mesh size varies from 65×65 up to 1025×1025 . At each Newton iteration, the order of magnitude of the error bounds grows as the mesh size is increased for fixed Gr , Pr , and V_{lid} . In Figure 7, we vary the mesh size, and the scaling behaviors

FIG. 4. The absolute errors in J_1 , J_2 , J_3 , and J_4 with $Gr = 2 \times 10^4$, $Pr = 1$, and the lid velocity $V_{lid} = 100$ on the 257×257 mesh. In all plots, the solid lines and the dotted lines with circles refer to the true values and the values given by (4.35) using INB-NE, respectively. The black lines represent the norms of the true residual.

FIG. 5. The absolute errors in J_1 , J_2 , J_3 , and J_4 with $Gr = 2 \times 10^4$, $Pr = 1$, and the lid velocity $V_{lid} = 100$ on the 257×257 mesh. In all plots, the solid lines and the dotted lines with squares refer to the true values and the values given by (4.60) using MSPIN, respectively. The black lines represent the norms of the true residual.

FIG. 6. The absolute errors in J_1 , J_2 , J_3 , and J_4 with $Gr = 2 \times 10^4$, $Pr = 1$, and the lid velocity $V_{lid} = 100$ on different mesh sizes. In all plots, the solid lines refer to the true values and the dotted lines to the values given by (4.35) using INB-NE.

FIG. 7. The true values of absolute errors (the left column) and the approximate error bounds (the right column) given by (4.35) in J_1 , J_2 , J_3 , and J_4 with $Gr = 2 \times 10^4$, $Pr = 1$, and the lid velocity $V_{lid} = 100$ at different global Newton iterations.

are shown for both the true values of the absolute errors and the approximate error bounds given by (4.35) corresponding to successive global Newton iterations. At a fixed iteration error bounds scale well for up to the 1025×1025 mesh. The absolute errors grow faster as the approximate solution $\hat{x}^{(k)}$ approaches the exact solution, while error bounds reflect the same trend and also grow accordingly.

FIG. 8. The absolute errors in J_1 , J_2 , J_3 , and J_4 with $Gr = 10, 10^3, 10^4$, $Pr = 1$, and the lid velocity $V_{lid} = 0.1$ on the 257×257 mesh. In all plots, the solid lines refer to the true values and the dotted lines to the values given by (4.60) using MSPIN.

Finally, we set $Pr = 1$ and the lid velocity $V_{lid} = 0.1$ on the 257×257 mesh for numerical tests, but change the values of Gr . In this case, buoyancy as a result of temperature difference plays a more important role than the moving lid to drive the circulation in the cavity. Figure 8 gives absolute errors and approximate error bounds at different Gr for the linear functionals corresponding to MSPIN. The growth of Gr intensifies the temperature change of the fluid, and therefore it is harder to obtain a convergent solution to the corresponding nonlinear system with stronger nonlinearities. It is seen that the values of absolute errors increase as Gr varies from 10 to 10^4 . By comparison, error bounds increase accordingly and no violation of bounds happens for any of the tests.

6. Conclusions. For multiphysics simulation, we often target some linear functionals of the solution that possess a physical interpretation. Since we wish to employ the INB-NE and MSPIN algorithms to tame strong nonlinearities and improve global convergence properties for fully coupled multiphysics problems, we derive a posteriori approximate error bounds on the linear functionals of interest using readily available by-products of the INB-NE and MSPIN iterations and some auxiliary computations. Numerical experiments show that our approximate error bounds can predict different convergence rates corresponding to different linear functionals.

Additional computation is required for the error bounds proposed in this paper. Namely, we have to calculate the action of $(I - C_i^{(k)})^{-1}$ or $(I - C_i^{(k)})^{-T}$ on a vector at each global Newton iteration. This extra work contributes to the confidence in the accuracy of the approximate linear functional. This optional investment in confidence

can be regarded as philosophically similar to the role of extra computation in adaptive mesh refinement based on a posteriori error analysis for the solution using finite element methods [14].

Appendix A. A simple illustrative example. Consider the system $F(x)$ of two nonlinear equations in two unknowns,

$$(A.1) \quad F(x) = F(u, v) = \begin{bmatrix} F_1(u, v) \\ F_2(u, v) \end{bmatrix} = \begin{bmatrix} (u - \alpha v^3 + \alpha)^3 - v^3 \\ u + 2v - 3 \end{bmatrix} = 0.$$

The first equation becomes less dependent on $x_2 \equiv v$, which appears to the ninth power, and thus couples less strongly with the other equation, as $\alpha \rightarrow 0$. It is easy to verify that $x^* = [u^*, v^*]^T = [1, 1]^T$ is a root of this system.

For the INB-NE method, we choose u as the bad component to be eliminated, i.e., the new starting point (obtained by subspace correction) for updating the solution is defined as

$$(A.2) \quad \tilde{x} = \begin{bmatrix} v + \alpha v^3 - \alpha \\ v \end{bmatrix},$$

provided that $x = [u, v]^T$ is given. The Jacobian matrix at $\tilde{x}$ is computed by

$$(A.3) \quad F'(\tilde{x}) = \begin{bmatrix} 3v^2 & -3v^2(1 + 3\alpha v^2) \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix},$$

and then we have

$$(A.4) \quad |\mathcal{C}^{\text{INB-NE}}| = |\mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} \mathcal{A}_{12}| = \frac{1 + 3\alpha v^2}{2} < 1, \quad \text{if } v < \frac{1}{\sqrt{3\alpha}},$$

and

$$(A.5) \quad R_G = -3\alpha v^2(1 + 2v)(1 - v)^2, \quad R_H = 0.$$

For the MSPIN method, we can derive a nonlinearly preconditioned system $\mathcal{F}(x)$,

$$(A.6) \quad \mathcal{F}(x) = \mathcal{F}(u, v) = \begin{bmatrix} g(u, v) \\ h(u, v) \end{bmatrix} = \begin{bmatrix} u - \alpha v^3 + \alpha - v \\ \frac{\alpha}{2}v^3 + \frac{3}{2}v - \frac{\alpha}{2} - \frac{3}{2} \end{bmatrix} = 0,$$

by solving

$$(A.7) \quad F_1(u - g, v) = (u - g - \alpha v^3 + \alpha)^3 - v^3 = 0,$$

$$(A.8) \quad F_2(u - g^*, v - h) = (u - g^*) + 2(v - h) - 3 = 0,$$

where g^* is the solution of (A.7). Following (A.6), it is easy to obtain

$$(A.9) \quad \mathcal{F}'(x) = \begin{bmatrix} \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \\ \frac{\partial h}{\partial u} & \frac{\partial h}{\partial v} \end{bmatrix} = \begin{bmatrix} 1 & -3\alpha v^2 - 1 \\ 0 & 1 - \frac{1}{2}(-3\alpha v^2 - 1) \end{bmatrix} = \begin{bmatrix} 1 & \mathcal{B}_{11}^{-1} \mathcal{B}_{12} \\ 0 & 1 - \mathcal{B}_{22}^{-1} \mathcal{B}_{21} \mathcal{B}_{11}^{-1} \mathcal{B}_{12} \end{bmatrix},$$

and then we have

$$(A.10) \quad |\mathcal{C}^{\text{MSPIN}}| = |\mathcal{B}_{22}^{-1} \mathcal{B}_{21} \mathcal{B}_{11}^{-1} \mathcal{B}_{12}| = \frac{1 + 3\alpha v^2}{2} < 1, \quad \text{if } v < \frac{1}{\sqrt{3\alpha}},$$

and

$$(A.11) \quad R_g = -\alpha(1+2v)(1-v)^2, \quad R_h = \frac{\alpha(1+2v)}{2}(1-v)^2.$$

We run tests starting from the four different initial estimates with the same distance from the exact solution: $x_0 = [0, 0]^T$, $x_0 = [0, 2]^T$, $x_0 = [2, 0]^T$, and $x_0 = [2, 2]^T$. We set $\alpha = 0.01$ for all of numerical tests. Here we consider the linear functionals as follows:

$$(A.12) \quad J_1(u, v) = u = \langle \psi_{1,1}, u \rangle + \langle \psi_{1,2}, v \rangle, \quad \text{where } \psi_{1,1} = 1 \text{ and } \psi_{1,2} = 0,$$

$$(A.13) \quad J_2(u, v) = u + 3v = \langle \psi_{2,1}, u \rangle + \langle \psi_{2,2}, v \rangle, \quad \text{where } \psi_{2,1} = 1 \text{ and } \psi_{2,2} = 3.$$

Using (4.39) and (A.5), we compute

$$\begin{aligned} (A.14) \quad r_{F,1}^{\text{INB-NE}} &= -\langle \psi_{1,1}, \mathcal{A}_{11}^{-1} \mathcal{A}_{12} (I - \mathcal{C}^{\text{INB-NE}})^{-1} \mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} R_G \rangle - \langle \psi_{1,1}, \mathcal{A}_{11}^{-1} R_G \rangle \\ &= \frac{2\alpha(1+2v)}{3(1+\alpha v^2)}(1-v)^2, \\ r_{F,2}^{\text{INB-NE}} &= -\langle \psi_{2,1}, \mathcal{A}_{11}^{-1} \mathcal{A}_{12} (I - \mathcal{C}^{\text{INB-NE}})^{-1} \mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} R_G \rangle \\ &\quad - \langle \psi_{2,1}, \mathcal{A}_{11}^{-1} R_G \rangle + \langle \psi_{2,2}, (I - \mathcal{C}^{\text{INB-NE}})^{-1} \mathcal{A}_{22}^{-1} \mathcal{A}_{21} \mathcal{A}_{11}^{-1} R_G \rangle \\ &= -\frac{\alpha(1+2v)}{3(1+\alpha v^2)}(1-v)^2. \end{aligned}$$

Using (4.63) and (A.11), we compute

$$\begin{aligned} (A.15) \quad r_{\mathcal{F},1}^{\text{MSPIN}} &= \langle \psi_{1,1}, \mathcal{B}_{11}^{-1} \mathcal{B}_{12} (I - \mathcal{C}^{\text{MSPIN}})^{-1} R_h \rangle - \langle \psi_{1,1}, R_g \rangle \\ &= \frac{2\alpha(1+2v)}{3(1+\alpha v^2)}(1-v)^2, \\ r_{\mathcal{F},2}^{\text{MSPIN}} &= \langle \psi_{2,1}, \mathcal{B}_{11}^{-1} \mathcal{B}_{12} (I - \mathcal{C}^{\text{MSPIN}})^{-1} R_h \rangle - \langle \psi_{2,1}, R_g \rangle \\ &\quad - \langle \psi_{2,2}, (I - \mathcal{C}^{\text{MSPIN}})^{-1} R_h \rangle. \\ &= -\frac{\alpha(1+2v)}{3(1+\alpha v^2)}(1-v)^2. \end{aligned}$$

Figures 9 and 10 show the absolute error values and approximate error bounds corresponding to INB-NE and MSPIN, respectively. For INB-NE, the error bound given by (4.35) looks sharper than the error bound given by (4.36). The former is actually not strict in some cases because of linearization at each Newton iteration, but no significant violation happens in numerical tests, and $r_{F,1}^{\text{INB-NE}}$ and $r_{F,2}^{\text{INB-NE}}$ in (A.14) quantify how approximate the error bounds are. For MSPIN, the error bound given by (4.60) always performs better than the bound predicted by (4.61).

Remark A.1. In large-scale problems, it is rarely possible to explicitly write down the corrected starting point in (A.2) or the modified systems in (A.6), which are feasible here due to the algebraic simplicity of the system $F(x)$.

FIG. 9. For INB-NE, the green circles refer to the absolute errors in J_1 and J_2 , the blue squares refer to the error bound computed using (4.35), and the red x-marks refer to the error bound computed using (4.36).

FIG. 10. For MSPIN, the green circles refer to the absolute errors in J_1 and J_2 , the blue squares refer to the error bound computed using (4.60), and the red x-marks refer to the error bound computed using (4.61).

Acknowledgment. The authors are grateful for the use of a Cray XC40 (“Shaheen”) operated by the Supercomputing Laboratory of the King Abdullah University of Science and Technology.

REFERENCES

- [1] H.-B. AN, *On convergence of the additive Schwarz preconditioned inexact Newton method*, SIAM J. Numer. Anal., 43 (2005), pp. 1850–1871.
- [2] O. AXELSSON, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1996.
- [3] S. BALAY, S. ABHYANKAR, M. F. ADAMS, J. BROWN, P. BRUNE, K. BUSCHELMAN, L. DALCIN, A. DENER, V. ELJKHOUT, W. D. GROPP, D. KARPEYEV, D. KAUSHIK, M. G. KNEPLEY, D. A. MAY, L. C. MCINNES, R. T. MILLS, T. MUNSON, K. RUPP, P. SANAN, B. F. SMITH, S. ZAMPINI, H. ZHANG, AND H. ZHANG, *PETSc Web Page*, <https://www.mcs.anl.gov/petsc>, 2019.
- [4] P. BRUNE, M. G. KNEPLEY, B. F. SMITH, AND X. TU, *Composing scalable nonlinear algebraic solvers*, SIAM Rev., 57 (2015), pp. 535–565.
- [5] X.-C. CAI AND D. E. KEYES, *Nonlinearly preconditioned inexact Newton algorithms*, SIAM J. Sci. Comput., 24 (2002), pp. 183–200.
- [6] X.-C. CAI AND X.-F. LI, *Inexact Newton methods with restricted additive Schwarz based nonlinear elimination for problems with high local nonlinearity*, SIAM J. Sci. Comput., 33 (2011), pp. 746–762.
- [7] L. CHACÓN AND D. A. KNOLL, *A 2D high- β Hall MHD implicit nonlinear solver*, J. Comput. Phys., 188 (2003), pp. 573–592.
- [8] T. S. COFFEY, C. T. KELLEY, AND D. E. KEYES, *Pseudotransient continuation and differential-algebraic equations*, SIAM J. Sci. Comput., 25 (2003), pp. 553–569.
- [9] A. DAX, *On extremum properties of orthogonal quotients matrices*, Linear Algebra Appl., 432 (2010), pp. 1234–1257.
- [10] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.
- [11] P. DEUFLHARD, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer Ser. Comput. Math. 35, Springer, New York, 2011.
- [12] C. ECHEVERRÍA, J. LIESEN, AND R. NABBEN, *Block diagonal dominance of matrices revisited: Bounds for the norms of inverses and eigenvalue inclusion sets*, Linear Algebra Appl., 553 (2018), pp. 365–383.
- [13] M. W. GEE, U. KÜTTLER, AND W. A. WALL, *Truly monolithic algebraic multigrid for fluid–structure interaction*, Internat. J. Numer. Methods Engrg., 85 (2011), pp. 987–1016.
- [14] M. B. GILES AND E. SÜLI, *Adjoint methods for PDEs: A posteriori error analysis and post-processing by duality*, Acta Numer., 11 (2002), pp. 145–236.
- [15] A. E. GILL, *Atmosphere–Ocean Dynamics*, Elsevier, Amsterdam, 2016.
- [16] S.-H. GONG AND X.-C. CAI, *A nonlinear elimination preconditioned inexact Newton method for heterogeneous hyperelasticity*, SIAM J. Sci. Comput., 41 (2019), pp. S390–S408.
- [17] D. E. KEYES, L. C. MCINNES, C. WOODWARD, W. GROPP, E. MYRA, M. PERNICE, J. BELL, J. BROWN, A. CLO, J. CONNORS, ET AL., *Multiphysics simulations: Challenges and opportunities*, Int. J. High Perform. Comput. Appl., 27 (2013), pp. 4–83.
- [18] D. E. KEYES AND M. D. SMOOKE, *Flame sheet starting estimates for counterflow diffusion flame problems*, J. Comput. Phys., 73 (1987), pp. 267–288.
- [19] D. A. KNOLL AND D. E. KEYES, *Jacobian-free Newton–Krylov methods: A survey of approaches and applications*, J. Comput. Phys., 193 (2004), pp. 357–397.
- [20] P. J. LANZKRON, D. J. ROSE, AND J. T. WILKES, *An analysis of approximate nonlinear elimination*, SIAM J. Sci. Comput., 17 (1996), pp. 538–559.
- [21] L. LIU AND D. E. KEYES, *Field-split preconditioned inexact Newton algorithms*, SIAM J. Sci. Comput., 37 (2015), pp. 1388–1409.
- [22] L. LIU AND D. E. KEYES, *Convergence analysis for the multiplicative Schwarz preconditioned inexact Newton algorithm*, SIAM J. Numer. Anal., 54 (2016), pp. 3145–3166.
- [23] L. LIU, D. E. KEYES, AND R. KRAUSE, *A note on adaptive nonlinear preconditioning techniques*, SIAM J. Sci. Comput., 40 (2018), pp. A1171–A1186.
- [24] S. H. LUI, *Nonlinearly preconditioned Newton’s method*, in Proceedings of the 14th International Conference on Domain Decomposition Methods, Cocoyoc, Mexico, 2002, pp. 95–105.
- [25] L. MACHIELS, J. PERAIRE, AND A. PATERA, *Output bound approximations for partial differential equations; Application to the incompressible Navier-Stokes equations*, in Industrial and

- Environmental Applications of Direct and Large-Eddy Simulation, Springer, New York, 1999, pp. 93–108.
- [26] J. MANDEL, *On block diagonal and Schur complement preconditioning*, Numer. Math., 58 (1990), pp. 79–93.
 - [27] P. PATHMANATHAN AND J. P. WHITELEY, *A numerical method for cardiac mechanoelectric simulations*, Ann. Biomed. Eng., 37 (2009), pp. 860–873.
 - [28] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
 - [29] M. D. SMOOKE, *The computation of laminar flames*, Proc. Combustion Institute, 34 (2013), pp. 65–98.
 - [30] J. P. WHITELEY, K. GILLOW, S. J. TAVENER, AND A. C. WALTER, *Error bounds on block Gauss–Seidel solutions of coupled multiphysics problems*, Internat. J. Numer. Methods Engrg., 88 (2011), pp. 1219–1237.
 - [31] H.-J. YANG, F.-N. HWANG, AND X.-C. CAI, *Nonlinear preconditioning techniques for full-space Lagrange–Newton solution of PDE-constrained optimization problems*, SIAM J. Sci. Comput., 38 (2016), pp. A2756–A2778.
 - [32] H.-J. YANG, C. YANG, AND S.-Y. SUN, *Active-set reduced-space methods with nonlinear elimination for two-phase flow problems in porous media*, SIAM J. Sci. Comput., 38 (2016), pp. B593–B618.