# On Numerical Behavior of Matrix Splitting Iteration Methods *

Zhong-Zhi Bai

*State Key Laboratory of Scientific/Engineering Computing*
*Institute of Computational Mathematics and Scientific/Engineering Computing*
*Academy of Mathematics and Systems Science*
*Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, P.R. China*
*Email: bzz@lsec.cc.ac.cn*

Miroslav Rozložník

*Institute of Computer Science*
*Academy of Sciences of the Czech Republic*
*Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic*
*Email: miro@cs.cas.cz*

September 20, 2014

**Abstract**

We study numerical behavior of stationary single- or two-step matrix splitting iteration methods for solving large sparse systems of linear equations. We show that inexact solutions of inner linear systems associated with the matrix splittings may considerably influence the convergence and the accuracy of the approximate solutions computed in finite precision arithmetic. For a general stationary matrix splitting iteration method, we analyze two mathematically equivalent implementations and find the corresponding componentwise or normwise forward or backward stable implementation.

**Keywords:** matrix splitting, stationary iteration method, convergence rate, rounding error analysis, backward error.

**AMS(MOS) Subject Classifications:** 65F10, 65F35, 65G30, 65G50; CR: G1.3.

## 1 Introduction

We consider an iterative solution of the large sparse system of linear equations

$$Ax = b, \quad A \in \mathbb{C}^{n,n} \quad \text{and} \quad b \in \mathbb{C}^n, \tag{1.1}$$

where $A$ is a nonsingular and, in general, a non-Hermitian matrix, and $b$ is the corresponding right-hand side vector. Many iteration methods for the linear system (1.1) are based on efficient splittings of the coefficient matrix $A$ in

*The research of Z.-Z. Bai is supported by The National Basic Research Program (No. 2011CB309703), The National Natural Science Foundation (No. 91118001) and The National Natural Science Foundation for Creative Research Groups (No. 11321061), P.R. China. The research of M. Rozložník is supported by the Grant Agency of the Czech Republic under the project 13-06684S and by the international collaboration support of the Academy of Sciences of the Czech Republic.

the form $A = M - N$, where $M$ is a nonsingular matrix such that a linear system with the coefficient matrix $M$ is easily solvable. The classical examples are the *Jacobi*, the *Gauss-Seidel* and the *successive overrelaxation* (**SOR**) iteration methods [31, 19, 18, 20], in which the matrix $A$ is split into its diagonal, off-diagonal and triangular parts, giving rise to the diagonal and the lower/upper triangular matrices $M$, respectively; see [32, 33] and the references therein. The modern examples are the *Hermitian and skew-Hermitian splitting* (**HSS**) iteration method [9] and its variants such as PMHSS (*preconditioned and modified Hermitian and skew-Hermitian splitting*) [7], in which the matrix $A$ is split into its Hermitian and skew-Hermitian parts, giving rise to the shifted Hermitian and the shifted skew-Hermitian matrices $M$; see also [11, 5, 6] and the references therein. In general, the HSS iteration method belongs to the framework of two-step matrix splitting iteration methods [14, 3, 4], which, for given two splittings $A = M_1 - N_1$ and $A = M_2 - N_2$ with $M_1$ and $M_2$ being nonsingular, iterates alternately between these two splittings in an analogous fashion to the classical *alternating direction implicit* (**ADI**) iteration method for solving partial differential equations [27, 16]; see also [8, 10] and the references therein.

In some cases, computing the exact solution of a linear system with the coefficient matrix $M$ (or $M_1$ or $M_2$) can be expensive and impractical in actual implementations. To further improve the computing efficiency, we usually solve this linear system, called the inner linear system, by another iteration scheme to some prescribed accuracy, resulting in an inexact or an inner/outer iteration method; see [12, 9, 11, 6]. For example, in the category of two-stage matrix splitting iteration methods, a linear system with the coefficient matrix $M$ is solved iteratively by an inner iteration scheme based on another splitting $M = F - G$, with $F$ being a nonsingular matrix; see [26, 25, 13]. This two-stage matrix splitting iteration method has been studied intensively by many authors in the literature, see, e.g., [17, 12, 2, 15] and the references therein. The inexact solution of the inner linear system may cause two important effects on the numerical behavior of the overall matrix splitting iteration process, i.e., a certain convergence delay of the iteration sequence and a possible accuracy limit on the computed approximate solution. By the componentwise or the normwise backward error analysis [20], in this paper we will prescribe the tolerance $\tau$ (or the tolerances $\tau_1$ and $\tau_2$) for the inner iteration method, with respect to the splitting matrix $M$ (or the splitting matrices $M_1$ and $M_2$), in a single (or a two-step) iteration process, which equivalently determines the number of the inner iteration steps. In other words, we interpret each computed approximate solution of an inner linear system as an exact solution of a perturbed linear system, where the relative perturbation of the coefficient matrix of the inner linear system, measured either by the size of its components or by its norm, is bounded by the parameter $\tau$ (or the parameters $\tau_1$ and $\tau_2$), being of the ideal order $\tau = \mathcal{O}(u)$ (or $\tau_1, \tau_2 = \mathcal{O}(u)$) for a backward stable method, but being much larger than the roundoff unit $u$ in practical implementations.

In this paper, we concentrate on the question what is the best accuracy we can obtain from such inexact schemes when implemented in finite precision arithmetic. The fact that the inner solution tolerance strongly influences the accuracy of the computed iterates is known and was studied in several contexts [9, 29, 30, 11, 23, 24]. Stationary iterative methods with the inner linear systems solved to working accuracy have been analyzed in [21, 15]. However, significantly less is known for iteration methods that use the inexact nontrivial splittings. We will also analyze the maximum attainable convergence delay of inexact two-step splitting iteration methods in terms of these parameters and in terms of spectral properties of corresponding splitting matrices. In this sense we extend the work achieved in [21] and give similar results to [23, 24]. In our work, we will analyze two mathematically equivalent implementations and point out that the one that is componentwise or normwise forward or backward stable. Given a computed approximate solution $\hat{x}$ to the linear system (1.1), an iteration method is called componentwise forward stable if the error $\hat{x} - x$ satisfies the bound $|\hat{x} - x| \leq \mathcal{O}(u)\,|A^{-1}||A||x|$, and is called normwise forward stable if the Euclidean norm of the error satisfies the bound $\|\hat{x} - x\| \leq \mathcal{O}(u)\,\|A^{-1}\|\|A\|\|x\|$. Similarly, an iteration method is called componentwise backward stable if the residual $b - A\hat{x}$ satisfies $|b - A\hat{x}| \leq \mathcal{O}(u)\,(|A||\hat{x}| + |b|)$, and is called normwise backward stable if the Euclidean norm of the residual satisfies $\|b - A\hat{x}\| \leq \mathcal{O}(u)\,(\|A\|\|\hat{x}\| + \|b\|)$.

The organization of the paper is as follows. In Section 2 we derive the main results on the convergence delay and maximum attainable accuracy for stationary (single-step) matrix splitting iteration methods. Section 3 is devoted to the analysis of the stationary two-step matrix splitting iteration methods. In Section 4, we review the HSS and the PMHSS iteration methods [9, 7], describe two experimental examples where the tested linear systems arise, and state the computing settings that are followed in the implementations. The numerical results are given in Section 5. Finally, in Section 6, we end the paper by a few concluding remarks.

Throughout the paper, we adopt the following notations and concepts. The term $I$ denotes the identity matrix

of suitable dimension and the symbol $\|\cdot\|$ indicates the Euclidean norm of either a vector or a matrix. For a given vector $x$ and matrix $X$, $|x|$ and $|X|$ stand for their absolute values, and $\|x\|$ and $\|X\|$ stand for their Euclidean norms, respectively. When $X$ is a square and nonsingular matrix, we use the quantity $\kappa(X) = \|X\|\|X^{-1}\|$ to represent its Euclidean-norm condition number. Note that $\kappa(X) = \kappa(X^{-1})$. For a square matrix $X$, we denote by $\rho(X)$ its spectral radius. For distinction with their exact arithmetic counterparts, we denote quantities computed in finite precision arithmetic by using an extra upper-hat. In addition, we assume the standard model for floating-point computations and denote by $u$ the unit roundoff. The term $\mathcal{O}(u)$ is a low-degree polynomial in the problem dimension $n$ multiplied by the unit roundoff $u$. It is independent of the system parameters but is dependent on details of the computer arithmetic. For simplicity, we do not evaluate the terms proportional to higher powers of $u$ and also occasionally skip the technical details that would negatively affect the presentation of our results.

## 2 Stationary Matrix Splitting Iteration Methods

Assume that $A = M - N$ is a splitting of the coefficient matrix $A$ of the linear system (1.1), with $M$ being nonsingular. Starting from an arbitrary initial vector $x_0$, a stationary (single-step) matrix splitting iteration method for solving the linear system (1.1) produces a sequence of approximate solutions $x_{k+1}$, $k = 0, 1, 2, \ldots$, with

$$x_{k+1} = M^{-1}(Nx_k + b) \tag{2.1}$$

or

$$x_{k+1} = x_k + M^{-1}(b - Ax_k). \tag{2.2}$$

Note that the iteration schemes (2.1) and (2.2) are mathematically equivalent, but as we will see later they are numerically different in actual implementations. From (2.1) and (2.2) we see that the error of the approximate solution $x_{k+1} - x$ and the associated residual $b - Ax_{k+1}$ satisfy, respectively, the recurrences

$$x_{k+1} - x = (I - M^{-1}A)(x_k - x) = G(x_k - x) \tag{2.3}$$

and

$$b - Ax_{k+1} = (I - AM^{-1})(b - Ax_k) = F(b - Ax_k), \tag{2.4}$$

with

$$G = I - M^{-1}A \quad \text{and} \quad F = I - AM^{-1}.$$

Note that the matrices $G$ and $F$ have the equivalent expressions $G = M^{-1}N$ and $F = NM^{-1}$.

In practical situations, the inner linear systems, induced by the iteration schemes (2.1) and (2.2), with respect to the coefficient matrix $M$, cannot be solved exactly. Instead, we will assume that every computed solution of a linear system with the coefficient matrix $M$ will be given by an approximate solution that can be interpreted as an exact solution of a linear system with the same right-hand side vector, but with a perturbed coefficient matrix $M + \Delta M$. Note that under reasonable assumption on the size of the increment $\Delta M$, the inverse of the matrix $M + \Delta M$ can be written in the form

$$(M + \Delta M)^{-1} = (I + \Delta H)M^{-1} = M^{-1}(I + \Delta E),$$

with

$$\Delta H = -(M + \Delta M)^{-1}\Delta M \quad \text{and} \quad \Delta E = -\Delta M(M + \Delta M)^{-1}.$$

If $M^{-1}$ in the iteration matrices $G$ and $F$ appearing in (2.3) and (2.4) are straightforwardly replaced by $(M+\Delta M)^{-1}$, then we could obtain the recurrences with the iteration matrices $G + \Delta G$ and $F + \Delta F$, where

$$\Delta G = \Delta H(G - I) \quad \text{and} \quad \Delta F = (F - I)\Delta E.$$

Hence, inexact solutions of the inner linear systems with respect to the coefficient matrix $M$ affect the convergence rate of the corresponding overall iteration scheme. Roughly speaking, a potential delay in the convergence is determined by the sizes of the increments $\Delta H$ and $\Delta E$. For stationary iteration methods, this phenomenon has been analyzed by several authors; see, e.g., [28, 21, 20, 15].

The accuracies of the approximate solutions computed by two equivalent iteration schemes (2.1) and (2.2) can be estimated by the standard tools of rounding error analysis [20]. The iteration scheme (2.1) has been analyzed by Higham and Knight in [21], where they discussed the recurrence for the computed approximate solutions $\hat{x}_{k+1}$, $k = 0,1,2,\ldots$, in the form

$$(M+\Delta M_k)\hat{x}_{k+1} = N\hat{x}_k + b + \Delta s_k, \tag{2.5}$$

with

$$|\Delta M_k| \leq \mathscr{O}(u)\,|M| \quad \text{and} \quad |\Delta s_k| \leq \mathscr{O}(u)\,(|N||\hat{x}_k| + |b|); \tag{2.6}$$

see also [20, Chapter 17]). The bound on $|\Delta M_k|$ is valid if the matrix $M$ is triangular, which is the case for the stationary relaxation iteration methods such as Jacobi, Gauss-Seidel and SOR [18, 31]. These classical matrix splitting iteration methods can be shown to be forward stable in a componentwise sense and backward stable in a normwise sense. The inner linear systems with respect to the coefficient matrix $M$ are, in general, not easily solvable, so they are solved iteratively in practical implementations. As a result, we cannot expect that all these inner linear systems can be solved in a backward stable way. Instead, we assume that the relative componentwise backward error associated with $\hat{x}_{k+1}$ is bounded by the parameter $\tau$ ($\tau \leq 1$), i.e., we use the stopping criterion based on the backward error and terminate the inner iteration process once $|\Delta M_k| \leq \tau|M|$ is satisfied. As a matter of fact, assuming $\tau \cdot \mathrm{cond}(M) < 1$ seems reasonable and some accuracy could be achieved in computing the approximate solutions for all inner linear systems.

In the following, we will analyze the maximum attainable accuracy of the computed approximate solutions caused by the inexact solutions of the inner linear systems with the coefficient matrix $M$. More specifically, we are going to show how the level of inexactness given by the tolerance $\tau$ affects the maximum attainable accuracy of the computed approximate solution $\hat{x}_{k+1}$ defined by (2.5), together with

$$|\Delta M_k| \leq \tau|M|, \quad \text{and} \quad |\Delta s_k| \leq \mathscr{O}(u)\,(|N||\hat{x}_k| + |b|),$$

while as will be shown later for the scheme (2.2) the maximum attainable accuracy will be proportional to the roundoff unit $u$.

Given an initial guess $\hat{x}_0$, the computed approximate solution $\hat{x}_{k+1}$, for $k = 0,1,2,\ldots$, is thus the exact solution of (2.5), which can be reformulated as

$$\hat{x}_{k+1} = G\hat{x}_k + M^{-1}(b+\Delta y_k) = G^{k+1}\hat{x}_0 + \sum_{i=0}^{k} G^i M^{-1}(b+\Delta y_{k-i}), \tag{2.7}$$

where

$$\Delta y_{k-i} = \Delta s_{k-i} - \Delta M_{k-i}\hat{x}_{k-i+1}, \quad i = 0,1,\ldots,k. \tag{2.8}$$

For the residual vectors corresponding to the solution $\hat{x}_{k+1}$, by making use of the identities

$$AG = AM^{-1}N = NM^{-1}A = FA \quad \text{and} \quad I - AM^{-1} = NM^{-1} = F$$

we can derive the recurrence in the form

$$b - A\hat{x}_{k+1} = F(b - A\hat{x}_k) - (I-F)\Delta y_k = F^{k+1}(b - A\hat{x}_0) + \sum_{i=0}^{k} F^i (I-F)\,\Delta y_{k-i}. \tag{2.9}$$

Using the identities

$$x = Gx + M^{-1}b = G^{k+1}x + \sum_{i=0}^{k} G^i M^{-1} b,$$

together with (2.7), we then obtain the formula for the error $\hat{x}_{k+1} - x$ of the $(k+1)$-th approximate solution $\hat{x}_{k+1}$ computed by the scheme (2.1) as follows:

$$\hat{x}_{k+1} - x = G^{k+1}(\hat{x}_0 - x) + \sum_{i=0}^{k} G^i M^{-1} \Delta y_{k-i}.$$

Therefore, the componentwise bound for the error $\hat{x}_{k+1} - x$ is given by

$$|\hat{x}_{k+1} - x| \leq |G^{k+1}(\hat{x}_0 - x)| + \sum_{i=0}^{k} \left|G^i\right| |M^{-1}| \max_{0 \leq i \leq k} |\Delta y_i|. \tag{2.10}$$

Analogously, using (2.9) we can obtain the componentwise bound for the corresponding residual $b - A\hat{x}_{k+1}$ as follows:

$$|b - A\hat{x}_{k+1}| \leq |F^{k+1}(b - A\hat{x}_0)| + \sum_{i=0}^{k} |F^i||I - F| \max_{0 \leq i \leq k} |\Delta y_i|. \tag{2.11}$$

If the spectral radius of the iteration matrix $G$ is less than 1, i.e., $\rho(G) < 1$, then the term $|G^{k+1}(\hat{x}_0 - x)|$ converges to the zero vector and, hence, for a large $k$ the bound for the maximum attainable accuracy of the computed approximate solution (measured in terms of its error) is given by the supremum of the second term in (2.10). Equivalently, if $\rho(G) < 1$, then $\rho(F) < 1$ and the term $|F^{k+1}(b - A\hat{x}_0)|$ converges to the zero vector, too. As a result, for a large $k$ the bound for the maximum attainable accuracy of the computed approximate solution (measured in terms of the residual) is given by the supremum of the second term in (2.11). Indeed, then the series $\sum_{i=0}^{\infty} G^i$ and $\sum_{i=0}^{\infty} F^i$ converge and, with

$$|\Delta M_i| \leq \tau |M|$$

and

$$|\Delta y_i| \leq |\Delta M_i||\hat{x}_{i+1}| + |\Delta s_i| \leq \tau |M||\hat{x}_{i+1}| + \mathcal{O}(u)\,(|N||\hat{x}_i| + |b|),$$

corresponding to the recurrence (2.5) we obtain the bounds

$$|\hat{x}_{k+1} - x| \lessapprox \left(\sum_{i=0}^{\infty} \left|G^i\right|\right) \left( \left[\tau |M^{-1}||M| + \mathcal{O}(u)\,|M^{-1}||N|\right] \max_{0 \leq i \leq k+1} |\hat{x}_i| + \mathcal{O}(u)\,|M^{-1}||b| \right) \tag{2.12}$$

and

$$|b - A\hat{x}_{k+1}| \lessapprox \left(\sum_{i=0}^{\infty} \left|F^i\right|\right) |I - F| \left( \left[\tau |M| + \mathcal{O}(u)\,|N|\right] \max_{0 \leq i \leq k+1} |\hat{x}_i| + \mathcal{O}(u)\,|b| \right). \tag{2.13}$$

Using $\tau \gg \mathcal{O}(u)$ and

$$|b| = |Ax| \leq |A||x| \leq (|M| + |N|)|x|,$$

we can rewrite (2.12) and (2.13) into

$$|\hat{x}_{k+1} - x| \lessapprox \tau \left(\sum_{i=0}^{\infty} |G^i|\right) |M^{-1}|(|M| + |N|) \left( \max_{0 \leq i \leq k+1} |\hat{x}_i| + |x| \right) \tag{2.14}$$

and

$$|b - A\hat{x}_{k+1}| \lessapprox \tau \left(\sum_{i=0}^{\infty} |F^i|\right) |I - F| \left( (|M| + |N|) \max_{0 \leq i \leq k+1} |\hat{x}_i| + |b| \right). \tag{2.15}$$

Provided that the entries of $\sum_{i=0}^{\infty}|G^i|$ or $\sum_{i=0}^{\infty}|F^i|$ are not too large, in the case of backward stable solutions of all inner linear systems with $\tau=\mathcal{O}(u)$, the estimates in (2.14) and (2.15) guarantee small forward and backward errors in the componentwise sense, respectively. These bounds contain the factor $\max_{0\le i\le k+1}|\hat{x}_i|$ that can be also significant depending on the convergence behavior of our stationary iteration method. Provided that this factor is not too large, i.e., $\max_{0\le i\le k+1}|\hat{x}_i|\approx|x|$, the componentwise forward or backward stability are then ensured if $|M^{-1}|\approx|A^{-1}|$ and $|M|+|N|\approx|A|$. However, in practice we have $\tau\gg\mathcal{O}(u)$ and, therefore, the maximum attainable accuracy in general does depend on the parameter $\tau$.

The normwise approach is similar. The componentwise bounds in (2.6) can be replaced by the normwise ones

$$\|\Delta M_k\|\le\tau\|M\| \quad\text{and}\quad \|\Delta s_k\|\le\mathcal{O}(u)\,(\|N\|\|\hat{x}_k\|+\|b\|).$$

So from (2.8) we can correspondingly obtain the estimate

$$\|\Delta y_i\|\le\|\Delta M_i\|\|\hat{x}_{i+1}\|+\|\Delta s_i\|\le\tau\|M\|\|\hat{x}_{i+1}\|+\mathcal{O}(u)\,(\|N\|\|\hat{x}_i\|+\|b\|). \tag{2.16}$$

Now, analogously to (2.10) and (2.11) we have the normwise bounds

$$\|\hat{x}_{k+1}-x\|\le\|G^{k+1}(\hat{x}_0-x)\|+\sum_{i=0}^{k}\|G^i\|\|M^{-1}\|\max_{0\le i\le k}\|\Delta y_i\| \tag{2.17}$$

and

$$\|b-A\hat{x}_{k+1}\|\le\|F^{k+1}(b-A\hat{x}_0)\|+\sum_{i=0}^{k}\|F^i\|\|I-F\|\max_{0\le i\le k}\|\Delta y_i\|. \tag{2.18}$$

Provided that $\|G\|<1$ and $\|F\|<1$, it holds that

$$\left\|\sum_{i=0}^{k}G^i\right\|\le\sum_{i=0}^{k}\|G^i\|\le\sum_{i=0}^{k}\|G\|^i\le\frac{1}{1-\|G\|}$$

and

$$\left\|\sum_{i=0}^{k}F^i\right\|\le\sum_{i=0}^{k}\|F^i\|\le\sum_{i=0}^{k}\|F\|^i\le\frac{1}{1-\|F\|}.$$

Similarly to [20] we define the normwise growth factor

$$\theta_{k+1}=\sup_{0\le i\le k+1}\left\{\frac{\|\hat{x}_i\|}{\|x\|}\right\},$$

so that

$$\|\hat{x}_i\|\le\theta_{k+1}\|x\|,\qquad i=0,1,\ldots,k+1.$$

By making use of (2.16) and

$$\|b\|\le(\|M\|+\|N\|)\|x\|, \tag{2.19}$$

we have for $i=0,1,\ldots,k$ that

$$\|\Delta y_i\|\le\theta_{k+1}(\tau\|M\|+\mathcal{O}(u)\,\|N\|)\|x\|+\mathcal{O}(u)\|b\|$$

and

$$\begin{aligned}\|\Delta y_i\| &\leq \theta_{k+1}(\tau\|M\|+\mathscr{O}(u)\|N\|)\|x\|+\mathscr{O}(u)(\|M\|+\|N\|)\|x\| \\ &\leq (1+\theta_{k+1})(\tau\|M\|+\mathscr{O}(u)\|N\|)\|x\|,\end{aligned}$$

where we have used the fact $\tau \gg \mathscr{O}(u)$. From (2.17) and (2.18) we then have

$$\|\hat{x}_{k+1}-x\| \leq \|G^{k+1}(\hat{x}_0-x)\| + \frac{1+\theta_{k+1}}{1-\|G\|}\left[\tau\,\kappa(M)+\mathscr{O}(u)\,\|M^{-1}\|\|N\|\right]\|x\| \tag{2.20}$$

and

$$\|b-A\hat{x}_{k+1}\| \leq \|F^{k+1}(b-A\hat{x}_0)\| + \frac{\|I-F\|}{1-\|F\|}\left[\theta_{k+1}(\tau\|M\|+\mathscr{O}(u)\|N\|)\|x\|+\mathscr{O}(u)\,\|b\|\right]. \tag{2.21}$$

Here in the derivation of (2.20) we have also applied the estimate (2.19).

In practical situations, when $\tau \gg \mathscr{O}(u)$, the relative error of the computed approximate solution will be proportional to the parameter $\tau$. Provided that $\|G\|$ and $\|F\|$ are not too close to 1, and $\theta_{k+1}$ is not too large, neglecting the terms with $\mathscr{O}(u)$ in (2.20) and (2.21) we see that the normwise relative error and the normwise residual will approximately satisfy

$$\frac{\|\hat{x}_{k+1}-x\|}{\|x\|} \lessapprox \tau\,\frac{1+\theta_{k+1}}{1-\|G\|}\,\kappa(M) \quad \text{and} \quad \|b-A\hat{x}_{k+1}\| \lessapprox \tau\,\frac{\theta_{k+1}\|I-F\|}{1-\|F\|}\|M\|\|x\|,$$

respectively. In the case of backward stable solutions of all inner linear systems with $\tau = \mathscr{O}(u)$, the bounds (2.20) and (2.21) reduce to the error bound (17.11) and the residual bound (17.19) in [20]. This guarantees a small normwise forward error if $\kappa(M) \approx \kappa(A)$ and a small normwise backward error if $\|M\| \approx \|A\|$ under the above-mentioned conditions.

As also noted in [22], if greater computing accuracy is required, we are better to work with the recurrence (2.2). This iteration scheme is similar to the iterative refinement, which is a popular technique for improving the computing accuracy of linear solvers; see [18]. We will show that under mild conditions this iteration scheme will deliver approximate solution with the accuracy being proportional to the roundoff unit $u$, but independent of the parameter $\tau$. This indicates a significant difference from the iteration scheme (2.1).

Given an initial guess $\hat{x}_0$, at the $(k+1)$-th step of the iteration scheme (2.2), we first compute the residual of the previously computed approximate solution $\hat{x}_k$ as follows:

$$\hat{r}_k = b-A\hat{x}_k+\Delta r_k, \quad \text{with} \quad |\Delta r_k| \leq \mathscr{O}(u)\,(|b|+|A||\hat{x}_k|). \tag{2.22}$$

Then we solve approximately the correction equation with the matrix $M$ so that the computed correction vector $\hat{z}_k$ satisfies

$$(M+\Delta M_k)\hat{z}_k = \hat{r}_k, \quad \text{with} \quad |\Delta M_k| \leq \tau|M|, \tag{2.23}$$

where the stopping criterion in the inner iteration is again assumed to be based on the backward error smaller than the parameter $\tau$. We finally obtain the approximate solution $\hat{x}_{k+1}$ that satisfies

$$\hat{x}_{k+1} = \hat{x}_k+\hat{z}_k+\Delta x_k, \quad \text{with} \quad |\Delta x_k| \leq u(|\hat{x}_k|+|\hat{z}_k|). \tag{2.24}$$

This computing procedure is well defined if the matrix $M+\Delta M_k$ is nonsingular, which is guaranteed under relatively mild conditions on the accuracy in the inner iterations (measured by the parameter $\tau$), e.g., $\sigma_{\min}(M) > \|\Delta M_k\|$, $k = 0,1,\ldots$, where $\sigma_{\min}(M)$ represents the smallest singular value of the matrix $M$. By using (2.24) we can derive the following recurrences for the error $\hat{x}_{k+1}-x$ and the residual $b-A\hat{x}_{k+1}$ corresponding to the computed approximate solution $\hat{x}_{k+1}$:

$$\hat{x}_{k+1}-x = \left[I-(M+\Delta M_k)^{-1}A\right](\hat{x}_k-x)+(M+\Delta M_k)^{-1}\Delta r_k+\Delta x_k, \tag{2.25}$$

$$b-A\hat{x}_{k+1} = \left[I-A(M+\Delta M_k)^{-1}\right](b-A\hat{x}_k)-A(M+\Delta M_k)^{-1}\Delta r_k-A\Delta x_k. \tag{2.26}$$

We now derive componentwise bounds for the error $\hat{x}_{k+1}-x$ and the residual $b-A\hat{x}_{k+1}$ based on the identities (2.25) and (2.26). To this end, from the definition of the update $\hat{z}_k$ we have

$$\begin{aligned}\hat{z}_k &= (M+\Delta M_k)^{-1}\left[(b-A\hat{x}_k)+\Delta r_k\right]\\ &= (M+\Delta M_k)^{-1}\left[A(x-\hat{x}_k)+\Delta r_k\right].\end{aligned} \tag{2.27}$$

Therefore,

$$\begin{aligned}|\hat{z}_k| &\leq |(M+\Delta M_k)^{-1}|\left[|b-A\hat{x}_k|+|\Delta r_k|\right]\\ &\leq |(M+\Delta M_k)^{-1}|\left[|b-A\hat{x}_k|+\mathcal{O}(u)\left(|b|+|A||\hat{x}_k|\right)\right]\end{aligned} \tag{2.28}$$

and

$$\begin{aligned}|\hat{z}_k| &\leq |(M+\Delta M_k)^{-1}|\left[|A||x-\hat{x}_k|+|\Delta r_k|\right]\\ &\leq (1+\mathcal{O}(u))\,|(M+\Delta M_k)^{-1}||A|(|x|+|\hat{x}_k|).\end{aligned} \tag{2.29}$$

It follows straightforwardly from these estimates, the bounds (2.22) and (2.24), as well as the identities (2.25) and (2.26) that

$$|\hat{x}_{k+1}-x| \leq |I-(M+\Delta M_k)^{-1}A||\hat{x}_k-x|+\mathcal{O}(u)\,|(M+\Delta M_k)^{-1}||A|(|x|+|\hat{x}_k|)+u|\hat{x}_k| \tag{2.30}$$

and

$$\begin{aligned}|b-A\hat{x}_{k+1}| \leq\ & \left[|I-A(M+\Delta M_k)^{-1}|+u|(M+\Delta M_k)^{-1}||A|\right]|b-A\hat{x}_k|\\ &+\mathcal{O}(u)\,|A(M+\Delta M_k)^{-1}|(|b|+|A||\hat{x}_k|)+u|A||\hat{x}_k|.\end{aligned} \tag{2.31}$$

If $\rho(\tau|M^{-1}||M|)<1$, then from $|\Delta M_k|\leq\tau|M|$ we have

$$|(M+\Delta M_k)^{-1}| \leq \sum_{i=0}^{\infty}\left(\tau|M^{-1}||M|\right)^i|M^{-1}| = \left(I-\tau|M^{-1}||M|\right)^{-1}|M^{-1}|$$

and

$$|A(M+\Delta M_k)^{-1}| \leq |I-F|\sum_{i=0}^{\infty}\left(\tau|M^{-1}||M|\right)^i = |I-F|\left(I-\tau|M^{-1}||M|\right)^{-1}.$$

Moreover, we claim that there exist matrices $\Delta G$ and $\Delta F$ such that

$$|I-(M+\Delta M_k)^{-1}A| \leq |G+\Delta G|$$

and

$$|I-A(M+\Delta M_k)^{-1}|+u|(M+\Delta M_k)^{-1}||A| \leq |F+\Delta F|.$$

Indeed, such matrices $\Delta G$ and $\Delta F$ do exist and they can be bounded as

$$\begin{aligned}|\Delta G| &\leq \tau\sum_{i=0}^{\infty}(\tau|M^{-1}||M|)^i|M^{-1}||M||M^{-1}A|\\ &= \tau|M^{-1}||M||M^{-1}A|\left(I-\tau|M^{-1}||M|\right)^{-1}\end{aligned}$$

and

$$\begin{aligned}|\Delta F| &\leq \left(\tau|AM^{-1}||M|+u|A|\right)|M^{-1}|\sum_{i=0}^{\infty}(\tau|M^{-1}||M|)^i\\ &= \left(\tau|AM^{-1}||M|+u|A|\right)|M^{-1}|\left(I-\tau|M^{-1}||M|\right)^{-1}.\end{aligned}$$

As a result, we can obtain the following bounds for $|\hat{x}_{k+1}-x|$ and $|b-A\hat{x}_{k+1}|$:

$$\begin{aligned}
|\hat{x}_{k+1}-x| &\leq |G+\Delta G||\hat{x}_k-x| + \mathscr{O}(u)\left(I-\tau|M^{-1}||M|\right)^{-1}|M^{-1}||A|(|x|+|\hat{x}_k|)+u|\hat{x}_k| \\
&\leq |G+\Delta G|^{k+1}|\hat{x}_0-x| + \sum_{i=0}^{k}|G+\Delta G|^i \\
&\quad \cdot\left[\mathscr{O}(u)\left(I-\tau|M^{-1}||M|\right)^{-1}|M^{-1}||A|\left(|x|+\max_{0\leq i\leq k}|\hat{x}_i|\right)+u\max_{0\leq i\leq k}|\hat{x}_i|\right] \qquad (2.32)
\end{aligned}$$

and

$$\begin{aligned}
|b-A\hat{x}_{k+1}| &\leq |F+\Delta F||b-A\hat{x}_k| + \mathscr{O}(u)\,|I-F|\left(I-\tau|M^{-1}||M|\right)^{-1}(|b|+|A||\hat{x}_k|)+u|A||\hat{x}_k| \\
&\leq |F+\Delta F|^{k+1}|b-A\hat{x}_0| + \sum_{i=0}^{k}|F+\Delta F|^i \\
&\quad \cdot\left[\mathscr{O}(u)\,|I-F|\left(I-\tau|M^{-1}||M|\right)^{-1}\left(|b|+|A|\max_{0\leq i\leq k}|\hat{x}_i|\right)+u|A|\max_{0\leq i\leq k}|\hat{x}_i|\right]. \qquad (2.33)
\end{aligned}$$

Provided that the spectral radii $\rho(|G+\Delta G|)$ and $\rho(|F+\Delta F|)$ are less than 1, the first terms in (2.32) and (2.33) will be small after sufficiently large number of iteration steps. Then the error $\hat{x}_{k+1}-x$ and the residual $b-A\hat{x}_{k+1}$ will be proportional to the roundoff unit $u$ as

$$|\hat{x}_{k+1}-x| \lessapprox \sum_{i=0}^{k}|G+\Delta G|^i\left[\mathscr{O}(u)\left(I-\tau|M^{-1}||M|\right)^{-1}|M^{-1}||A|\left(|x|+\max_{0\leq i\leq k}|\hat{x}_i|\right)+u\max_{0\leq i\leq k}|\hat{x}_i|\right]$$

and

$$|b-A\hat{x}_{k+1}| \lessapprox \sum_{i=0}^{k}|F+\Delta F|^i\left[\mathscr{O}(u)\,|I-F|\left(I-\tau|M^{-1}||M|\right)^{-1}\left(|b|+|A|\max_{0\leq i\leq k}|\hat{x}_i|\right)+u|A|\max_{0\leq i\leq k}|\hat{x}_i|\right].$$

These bounds are significantly better than the bounds we have obtained for the recurrence (2.1). Although in practical situations it is $\tau \gg \mathscr{O}(u)$ that is used in the iteration scheme (2.2), we will obtain very accurate approximate solutions after sufficiently many iterations.

For the normwise approach, now the componentwise bounds in (2.22), (2.23) and (2.24) are, respectively, replaced by the normwise ones

$$\hat{r}_k = b-A\hat{x}_k+\Delta r_k, \quad \text{with} \quad \|\Delta r_k\| \leq \mathscr{O}(u)\,(\|b\|+\|A\|\|\hat{x}_k\|),$$

$$(M+\Delta M_k)\hat{z}_k = \hat{r}_k, \quad \text{with} \quad \|\Delta M_k\| \leq \tau\|M\|$$

and

$$\hat{x}_{k+1} = \hat{x}_k+\hat{z}_k+\Delta x_k, \quad \text{with} \quad \|\Delta x_k\| \leq u(\|\hat{x}_k\|+\|\hat{z}_k\|).$$

Based on the identities (2.25) and (2.26), using an analogous approach we can derive the normwise bounds

$$\frac{\|\hat{x}_{k+1}-x\|}{\|x\|} \lessapprox \mathscr{O}(u)\frac{1+\theta_k}{1-\|G\|-\tau\kappa(M)\,(1-\|G\|+\|I-G\|)}\left(\|M^{-1}\|\|A\|+1-\tau\kappa(M)\right)$$

and

$$\begin{aligned}
\|b-A\hat{x}_{k+1}\| &\lessapprox \mathscr{O}(u)\frac{\|I-F\|}{1-\|F\|-\tau\kappa(M)\,(1-\|F\|+\|I-F\|)-u\|M^{-1}\|\|A\|} \\
&\quad \cdot\left[\|b\|+\theta_k\|A\|\|x\|\left(1+\frac{1-\tau\kappa(M)}{\|I-F\|}\right)\right]
\end{aligned}$$

for the error $\hat{x}_{k+1} - x$ and the residual $b - A\hat{x}_{k+1}$, respectively, under the assumptions

$$\|G + \Delta G\| \leq \|G\| + \tau\,\kappa(M)\,(1 - \tau\,\kappa(M))^{-1}\|I - G\| < 1 \tag{2.34}$$

and

$$\|F + \Delta F\| \leq \|F\| + (1 - \tau\,\kappa(M))^{-1}\left(\tau\,\kappa(M)\,\|I - F\| + u\|M^{-1}\|\|A\|\right) < 1. \tag{2.35}$$

Recall that $\theta_k$ is the growth factor depending on all preceding computed iterates $\{\hat{x}_i\}_{i=0}^{k}$. Again, these bounds guarantee small normwise forward and backward errors, respectively, under mild conditions as stated in (2.34) and (2.35).

In summary, if the iteration schemes (2.1) and (2.2) are either componentwise or normwise forward or backward stable, and if the splitting matrix $N$ is as sparse and structured as the coefficient matrix $A$, then, at the $k$-th iteration step of these two schemes, computing the vector $Nx_k + b$ should be as costly as computing the residual $b - Ax_k$. So the iteration scheme (2.1) costs about the same workloads as the iteration scheme (2.2) at each iteration step. Roughly speaking, provided that the inner linear systems having the same coefficient matrix $M$ are solved inexactly in accuracies controlled by the same tolerance $\tau$, the iteration scheme (2.2) can always achieve higher computational efficiency than the iteration scheme (2.1).

# 3 Stationary Two-Step Matrix Splitting Iteration Methods

In this section, we study the numerical behavior of the stationary two-step matrix splitting iteration methods [27, 16, 3, 4, 12, 14] and give results similar to the stationary matrix splitting iteration methods in Section 2. The stationary two-step matrix splitting iteration framework has been studied extensively by several authors from several perspectives, see, e.g., [9, 8, 10, 5, 6] and the references therein. We consider two splittings of the matrix $A$ in the form $A = M_1 - N_1$ and $A = M_2 - N_2$. Given an initial vector $x_0$, we define the stationary two-step matrix splitting iteration method by the following two successive recurrences

$$\begin{aligned} M_1 x_{k+1/2} &= N_1 x_k + b, \\ M_2 x_{k+1} &= N_2 x_{k+1/2} + b. \end{aligned}$$

Alternatively, we can use these recurrences in the most straightforward way as

$$x_{k+1/2} = M_1^{-1}(N_1 x_k + b), \tag{3.1}$$

$$x_{k+1} = M_2^{-1}(N_2 x_{k+1/2} + b). \tag{3.2}$$

Denote by $G_1 = M_1^{-1}N_1 = I - H_1 A$ and $G_2 = M_2^{-1}N_2 = I - H_2 A$, with $H_1 = M_1^{-1}$ and $H_2 = M_2^{-1}$. Then (3.1) and (3.2) can be rewritten as

$$x_{k+1/2} = G_1 x_k + H_1 b$$

and

$$x_{k+1} = G_2 x_{k+1/2} + H_2 b.$$

These give rise to the alternative recurrences

$$x_{k+1/2} = x_k + H_1(b - Ax_k), \tag{3.3}$$

$$x_{k+1} = x_{k+1/2} + H_2(b - Ax_{k+1/2}). \tag{3.4}$$

At each iteration step, the recurrences (3.3) and (3.4) involve the computations of two residuals $b - Ax_k$ and $b - Ax_{k+1/2}$, which require two matrix-vector multiplications with respect to the matrix $A$. According to Lemma 2.1

in [9], this can be avoided, however, by the substitution of $x_{k+1/2}$ in (3.3) into $x_{k+1}$ in (3.4), leading to

$$\begin{aligned} x_{k+1} &= x_k + H_1(b - Ax_k) + H_2\left[b - A\left(x_k + H_1(b - Ax_k)\right)\right] \\ &= (I - H_2A)(I - H_1A)x_k + [(I - H_2A)H_1 + H_2]b \\ &= G_2G_1x_k + (G_2H_1 + H_2)b \\ &= Gx_k + Hb, \end{aligned}$$

where

$$G = G_2G_1 \quad \text{and} \quad H = G_2H_1 + H_2. \tag{3.5}$$

We remark that the matrix $H$ admits the following equivalent expressions

$$H = H_1 + H_2G_1 = H_1 + H_2 - H_2AH_1 = H_2(M_1 + M_2 - A)H_1, \tag{3.6}$$

and the matrices $G$ and $H$ satisfy the identity $G = I - HA$. Thus, instead of (3.3) and (3.4) we can use only one single recurrence

$$x_{k+1} = x_k + H(b - Ax_k). \tag{3.7}$$

The detailed convergence analysis about the alternating splitting iteration method can be found in [14, 3, 4] and the references therein.

In practical situations, the inner linear systems, induced by the iteration schemes (3.1) and (3.2) with respect to the coefficient matrices $M_1$ and $M_2$, cannot be solved exactly, and they are often solved inexactly by some other iteration schemes; see [9, 11] and the references therein. It follows that inexact solutions of the inner linear systems with respect to the coefficient matrices $M_1$ and $M_2$ affect the convergence rate of the corresponding overall iteration scheme.

In the following, we estimate the maximum attainable accuracy for approximate solution, computed with (3.1) and (3.2), to the linear system (1.1). Using the same approach as for the stationary matrix splitting iteration method defined by (2.1) in Section 2, we can write

$$\hat{x}_{k+1/2} = M_1^{-1}(N_1\hat{x}_k + b + \Delta s_{k+1/2}), \tag{3.8}$$

$$\hat{x}_{k+1} = M_2^{-1}(N_2\hat{x}_{k+1/2} + b + \Delta s_{k+1}), \tag{3.9}$$

where

$$\begin{aligned} |\Delta s_{k+1/2}| &\leq \tau_1|M_1||\hat{x}_{k+1/2}| + \mathscr{O}(u)\left(|N_1||\hat{x}_k| + |b|\right), \\ |\Delta s_{k+1}| &\leq \tau_2|M_2||\hat{x}_{k+1}| + \mathscr{O}(u)\left(|N_2||\hat{x}_{k+1/2}| + |b|\right). \end{aligned}$$

Again, $\tau_1$ and $\tau_2$ are the tolerances employed to describe the accuracies in solving the inner linear systems with respect to the matrices $M_1$ and $M_2$, respectively. Substituting $\hat{x}_{k+1/2}$ into the formula of $\hat{x}_{k+1}$, we obtain the expression

$$\hat{x}_{k+1} = G\hat{x}_k + Hb + \Delta y_k, \tag{3.10}$$

with

$$\Delta y_k = G_2M_1^{-1}\Delta s_{k+1/2} + M_2^{-1}\Delta s_{k+1}.$$

Denote by $F_1 = N_1M_1^{-1}$ and $F_2 = N_2M_2^{-1}$. Then it follows from direct manipulation that

$$A\Delta y_k = F_2(I - F_1)\Delta s_{k+1/2} + (I - F_2)\Delta s_{k+1},$$

where we have used the commutative property of the matrices $A$ and $M_2^{-1}N_2$, i.e., $AM_2^{-1}N_2 = M_2^{-1}N_2A$.

From the bounding conditions on $\Delta s_{k+1/2}$ and $\Delta s_{k+1}$ we have the estimates

$$\begin{aligned} |\Delta s_{k+1/2}| &\leq \tau_1|M_1||\hat{x}_{k+1/2}| + \mathcal{O}(u)\,(|N_1||\hat{x}_k| + |b|) \\ &\leq (\tau_1|M_1| + \mathcal{O}(u)\,|N_1|) \max_{i=k,k+1/2} |\hat{x}_i| + \mathcal{O}(u)|b| \end{aligned}$$

and

$$\begin{aligned} |\Delta s_{k+1/2}| &\leq (\tau_1|M_1| + \mathcal{O}(u)\,|N_1|) \max_{i=k,k+1/2} |\hat{x}_i| + \mathcal{O}(u)(|M_1| + |N_1|)|x| \\ &\leq (\tau_1|M_1| + \mathcal{O}(u)\,|N_1|) \left( \max_{i=k,k+1/2} |\hat{x}_i| + |x| \right), \end{aligned}$$

as well as

$$\begin{aligned} |\Delta s_{k+1}| &\leq \tau_2|M_2||\hat{x}_{k+1}| + \mathcal{O}(u)\,(|N_2||\hat{x}_{k+1/2}| + |b|) \\ &\leq (\tau_2|M_2| + \mathcal{O}(u)\,|N_2|) \max_{i=k+1/2,k+1} |\hat{x}_i| + \mathcal{O}(u)\,|b| \end{aligned}$$

and

$$\begin{aligned} |\Delta s_{k+1}| &\leq (\tau_2|M_2| + \mathcal{O}(u)\,|N_2|) \max_{i=k+1/2,k+1} |\hat{x}_i| + \mathcal{O}(u)\,(|M_2| + |N_2|)|x| \\ &\leq (\tau_2|M_2| + \mathcal{O}(u)\,|N_2|) \left( \max_{i=k+1/2,k+1} |\hat{x}_i| + |x| \right). \end{aligned}$$

Here in the estimates of $|\Delta s_{k+1/2}|$ and $|\Delta s_{k+1}|$ we have used the facts $\tau_1 \gg \mathcal{O}(u)$, $\tau_2 \gg \mathcal{O}(u)$, and applied the bounds

$$\begin{aligned} |b| &= |(M_1 - N_1)x| \leq (|M_1| + |N_1|)|x|, \\ |b| &= |(M_2 - N_2)x| \leq (|M_2| + |N_2|)|x|. \end{aligned}$$

Therefore, according to the formulas of $\Delta y_k$ and $A\Delta y_k$, it holds that

$$\begin{aligned} |\Delta y_k| &\leq |G_2 M_1^{-1}||\Delta s_{k+1/2}| + |M_2^{-1}||\Delta s_{k+1}| \\ &\leq |G_2 M_1^{-1}|(\tau_1|M_1| + \mathcal{O}(u)\,|N_1|) \left( \max_{i=k,k+1/2} |\hat{x}_i| + |x| \right) \\ &\quad + |M_2^{-1}|(\tau_2|M_2| + \mathcal{O}(u)\,|N_2|) \left( \max_{i=k+1/2,k+1} |\hat{x}_i| + |x| \right) \\ &\leq \left[ |G_2 M_1^{-1}|(\tau_1|M_1| + \mathcal{O}(u)\,|N_1|) + |M_2^{-1}|(\tau_2|M_2| + \mathcal{O}(u)\,|N_2|) \right] \\ &\quad \cdot \left( \max_{i=k,k+1/2,k+1} |\hat{x}_i| + |x| \right) \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} |A\Delta y_k| &\leq |F_2(I - F_1)||\Delta s_{k+1/2}| + |I - F_2||\Delta s_{k+1}| \\ &\leq |F_2(I - F_1)| \left[ (\tau_1|M_1| + \mathcal{O}(u)\,|N_1|) \max_{i=k,k+1/2} |\hat{x}_i| + |b| \right] \\ &\quad + |I - F_2| \left[ (\tau_2|M_2| + \mathcal{O}(u)\,|N_2|) \max_{i=k+1/2,k+1} |\hat{x}_i| + |b| \right]. \end{aligned} \tag{3.12}$$

Because (3.10) immediately implies

$$\hat{x}_{k+1} = G^{k+1}\hat{x}_0 + \sum_{i=0}^{k} G^i(Hb + \Delta y_{k-i}),$$

by making use of the identity

$$x = G^{k+1}x + \sum_{i=0}^{k} G^i H b$$

and the relationship $AG = FA$, with $F = I - AH = F_2F_1$, we can obtain the recurrences

$$\hat{x}_{k+1} - x = G^{k+1}(\hat{x}_0 - x) + \sum_{i=0}^{k} G^i \Delta y_{k-i} \tag{3.13}$$

and

$$b - A\hat{x}_{k+1} = F^{k+1}(b - A\hat{x}_0) + \sum_{i=0}^{k} F^i A \Delta y_{k-i} \tag{3.14}$$

for the error $\hat{x}_{k+1} - x$ and the residual $b - A\hat{x}_{k+1}$. It then follows straightforwardly that

$$|\hat{x}_{k+1} - x| \leq |G^{k+1}||\hat{x}_0 - x| + \sum_{i=0}^{k} |G^i||\Delta y_{k-i}|, \tag{3.15}$$

$$|b - A\hat{x}_{k+1}| \leq |F^{k+1}||b - A\hat{x}_0| + \sum_{i=0}^{k} |F^i||A\Delta y_{k-i}|. \tag{3.16}$$

If $\rho(G) < 1$, then $\rho(F) < 1$. Hence, the terms $|G^{k+1}||\hat{x}_0 - x|$ and $|F^{k+1}||b - A\hat{x}_0|$ converge to the zero vector. As a result, for a large $k$ the bounds for the maximum attainable accuracies of the computed approximate solutions (measured in terms of both error and residual) are given by the supremums of the second terms in (3.15) and (3.16), respectively. Indeed, corresponding to the recurrence (3.8), after substitutions of (3.11) and (3.12) into (3.15) and (3.16), respectively, we obtain the bounds

$$\begin{aligned} |\hat{x}_{k+1} - x| \lessapprox & \sum_{i=0}^{k} |G^i| \left[ |G_2 M_1^{-1}| (\tau_1 |M_1| + \mathscr{O}(u)\,|N_1|) + |M_2^{-1}| (\tau_2 |M_2| + \mathscr{O}(u)\,|N_2|) \right] \\ & \cdot \left( \max_{1 \le i \le 2k+3} |\hat{x}_{(i-1)/2}| + |x| \right) \end{aligned}$$

and

$$\begin{aligned} |b - A\hat{x}_{k+1}| \lessapprox & \sum_{i=0}^{k} |F^i| \left[ |F_2(I - F_1)| \left( (\tau_1 |M_1| + \mathscr{O}(u)\,|N_1|) \max_{1 \le i \le 2k+3} |\hat{x}_{(i-1)/2}| + |b| \right) \right. \\ & + \left. |I - F_2| \left( (\tau_2 |M_2| + \mathscr{O}(u)\,|N_2|) \max_{1 \le i \le 2k+3} |\hat{x}_{(i-1)/2}| + |b| \right) \right]. \end{aligned}$$

Provided that the entries of the vector $\max\limits_{1 \le i \le 2k+3} |\hat{x}_{(i-1)/2}|$ and the entries in the matrix $\sum\limits_{i=0}^{k} |G^i|$ or $\sum\limits_{i=0}^{k} |F^i|$ are not too large, in the case of backward stable solutions of all inner linear systems with $\tau_1 = \mathscr{O}(u)$ and $\tau_2 = \mathscr{O}(u)$, these estimates then guarantee small forward and backward errors in the componentwise sense, respectively, if $|M_1^{-1}| \approx |M_2^{-1}| \approx |A^{-1}|$ and $|M_1| + |N_1| \approx |M_2| + |N_2| \approx |A|$. However, in practice we have $\tau_1 \gg \mathscr{O}(u)$ and $\tau_2 \gg \mathscr{O}(u)$. Therefore, the maximum attainable accuracies in general do depend on the parameters $\tau_1$ and $\tau_2$.

The normwise approach can be conducted in a similar fashion. In fact, by introducing the normwise growth factor

$$\theta_{k+1} = \sup_{1 \le i \le 2k+3} \frac{\|\hat{x}_{(i-1)/2}\|}{\|x\|}$$

so that

$$\|\hat{x}_{(i-1)/2}\| \le \theta_{k+1}\|x\|, \quad i = 0,1,\ldots,2k+3,$$

if $\|F\| < 1$ and $\|G\| < 1$ we can correspondingly obtain the normwise bounds for the error $\hat{x}_{k+1} - x$ and the residual $b - A\hat{x}_{k+1}$ as follows:

$$\begin{aligned}
\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} &\lessapprox \frac{1+\theta_{k+1}}{1-\|G\|}\left[(\tau_1\|G_2\|\,\kappa(M_1) + \tau_2\,\kappa(M_2))\right. \\
&\quad \left. + \mathscr{O}(u)\ (\|G_2\|\|M_1^{-1}\|\|N_1\| + \|M_2^{-1}\|\|N_2\|)\right], \\
\|b - A\hat{x}_{k+1}\| &\lessapprox \frac{\theta_{k+1}}{1-\|F\|}\left[(\tau_1\|F_2(I-F_1)\|\|M_1\| + \tau_2\|I-F_2\|\|M_2\|)\right. \\
&\quad \left. + \mathscr{O}(u)\ (\|F_2(I-F_1)\|\|N_1\| + \|I-F_2\|\|N_2\|)\right]\|x\| \\
&\quad + \frac{\|b\|}{1-\|F\|}(\|F_2(I-F_1)\| + \|I-F_2\|).
\end{aligned}$$

Provided that $\tau_1 \gg \mathscr{O}(u)$ and $\tau_2 \gg \mathscr{O}(u)$, these bounds can be approximately reduced to

$$\begin{aligned}
\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} &\lessapprox \frac{1+\theta_{k+1}}{1-\|G\|}(\tau_1\|G_2\|\,\kappa(M_1) + \tau_2\,\kappa(M_2)), \\
\|b - A\hat{x}_{k+1}\| &\lessapprox \frac{\theta_{k+1}}{1-\|F\|}(\tau_1\|F_2(I-F_1)\|\|M_1\| + \tau_2\|I-F_2\|\|M_2\|)\|x\| \\
&\quad + \frac{\|b\|}{1-\|F\|}(\|F_2(I-F_1)\| + \|I-F_2\|).
\end{aligned}$$

Roughly speaking, the limiting accuracy level measured in terms of the error is given by the quantity $\tau_1\|G_2\|\,\kappa(M_1) + \tau_2\,\kappa(M_2)$, so the $\tau_1$-term is damped by the quantity $\|G_2\|$. In actual implementations, we should balance the choices of the tolerances $\tau_1$ and $\tau_2$ in such a way that a desired overall accuracy of the error is achieved. For example, we may set $\tau_1 = \frac{\tau}{\|G_2\|}$ and $\tau_2 = \tau$, where $\tau$ is a prescribed tolerance. Consequently, it holds that

$$\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \lessapprox \frac{1+\theta_{k+1}}{1-\|G\|}\,\tau(\kappa(M_1) + \kappa(M_2)).$$

The quantity $\tau_1\|F_2(I-F_1)\|\|M_1\| + \tau_2\|I-F_2\|\|M_2\|$ plays an analogous role in the result for the norm of the residual, and the tolerances $\tau_1$ and $\tau_2$ should be chosen in a similar fashion to the above, e.g., through a prescribed common tolerance $\tau$. Definitely, the maximum attainable accuracies depend on the levels of inexactness (measured in terms of $\tau$) in solving the inner linear systems either with the matrix $M_1$ or with the matrix $M_2$.

In some applications one needs the maximum attainable accuracy to be proportional to the machine precision $u$. Hence, it makes sense that we discuss the recurrence (3.7) by straightforwardly applying the theory for the recurrence (2.2) established in Section 2. In this manner, we can derive the componentwise and the normwise bounds for the error $\hat{x}_{k+1} - x$ and the residual $b - A\hat{x}_{k+1}$ of the computed solution $\hat{x}_{k+1}$ of the linear system (1.1).

To this end, we recall that the matrix $H$, defined in (3.5) and reformulated in (3.6), adopts the equivalent expression

$$H = M_2^{-1}(M_1 + M_2 - A)M_1^{-1}.$$

Similar to the computing model described in (2.22), (2.23) and (2.24), at the $(k+1)$-th step of the iteration scheme (3.7) we assume that the computed solution $\hat{x}_{k+1}$ is obtained by the procedure

$$\hat{r}_k = b - A\hat{x}_k + \Delta r_k, \quad \hat{z}_k = (H + \Delta H^{(k)})\hat{r}_k \quad \text{and} \quad \hat{x}_{k+1} = \hat{x}_k + \hat{z}_k + \Delta x_k,$$

with $\hat{x}_0$ a given initial guess, where

$$|\Delta r_k| \le \mathscr{O}(u)\,(|b| + |A||\hat{x}_k|), \quad |\Delta x_k| \le u(|\hat{x}_k| + |\hat{z}_k|), \tag{3.17}$$

and $\Delta H^{(k)}$ is a perturbation to the matrix $H$, which is defined implicitly by

$$H+\Delta H^{(k)} = (M_2+\Delta M_2^{(k)})^{-1}(M_1+M_2-A+\Delta M^{(k)})(M_1+\Delta M_1^{(k)})^{-1},$$

with $\Delta M_1^{(k)}$, $\Delta M_2^{(k)}$ and $\Delta M^{(k)}$ being imposed to satisfy

$$|\Delta M_1^{(k)}| \le \tau_1|M_1|, \quad |\Delta M_2^{(k)}| \le \tau_2|M_2| \quad \text{and} \quad |\Delta M^{(k)}| \le \mathscr{O}(u)\,(|M_1|+|M_2|+|A|). \tag{3.18}$$

Again, $\tau_1$ and $\tau_2$ are two prescribed tolerances used to measure the accuracies in solving the inner linear systems with respect to the matrices $M_1$ and $M_2$, respectively.

Again, we omit the superscripts of the perturbation matrices such as $\Delta H^{(k)}$, $\Delta M^{(k)}$ and $\Delta M_1^{(k)}$, $\Delta M_2^{(k)}$, which are used to label the iterate indices. Moreover, there exist matrices $\Delta G$ and $\Delta F$, independent of the iterate index $k$, such that

$$|I-(H+\Delta H)A| \le |G+\Delta G| \quad \text{and} \quad |I-A(H+\Delta H)|+u|H+\Delta H||A| \le |F+\Delta F|.$$

Provided that the spectral radii $\rho(|G+\Delta G|)$ and $\rho(|F+\Delta F|)$ are strictly less than 1, the maximum attainable accuracies will be proportional to the roundoff unit $u$ and also independent of the parameters $\tau_1$ and $\tau_2$ as

$$|\hat{x}_{k+1}-x| \lessapprox \sum_{i=0}^{k} |G+\Delta G|^i \left[\mathscr{O}(u)\,|H+\Delta H||A|\left(|x|+\max_{0\le i\le k}|\hat{x}_k|\right)+u\max_{0\le i\le k}|\hat{x}_k|\right]$$

and

$$|b-A\hat{x}_{k+1}| \lessapprox \sum_{i=0}^{k} |F+\Delta F|^i \left[\mathscr{O}(u)\,|A(H+\Delta H)|\left(|b|+|A|\max_{0\le i\le k}|\hat{x}_k|\right)+u|A|\max_{0\le i\le k}|\hat{x}_k|\right].$$

As

$$|H+\Delta H| \le |I-G-\Delta G||A^{-1}| \quad \text{and} \quad |A(H+\Delta H)| \le |I-F-\Delta F|,$$

we can further obtain the bounds

$$|\hat{x}_{k+1}-x| \lessapprox \sum_{i=0}^{k} |G+\Delta G|^i \left[\mathscr{O}(u)\,|I-G-\Delta G||A^{-1}||A|\left(|x|+\max_{0\le i\le k}|\hat{x}_k|\right)+u\max_{0\le i\le k}|\hat{x}_k|\right]$$

and

$$|b-A\hat{x}_{k+1}| \lessapprox \sum_{i=0}^{k} |F+\Delta F|^i \left[\mathscr{O}(u)\,|I-F-\Delta F|\left(|b|+|A|\max_{0\le i\le k}|\hat{x}_k|\right)+u|A|\max_{0\le i\le k}|\hat{x}_k|\right].$$

These bounds are significantly better than the bounds we have obtained for the recurrence defined in (3.1) and (3.2). Although in practical situations it is $\tau_1 \gg \mathscr{O}(u)$ and $\tau_2 \gg \mathscr{O}(u)$ that are used in the iteration scheme (3.7), we will obtain very accurate approximate solutions after sufficiently many iterations.

For the normwise approach, replacing the componentwise bounds in (3.17) and (3.18) by the normwise ones

$$\|\Delta r_k\| \le \mathscr{O}(u)\,(\|b\|+\|A\|\|\hat{x}_k\|), \quad \|\Delta x_k\| \le u(\|\hat{x}_k\|+\|\hat{z}_k\|)$$

and

$$\|\Delta M_1\| \le \tau_1\|M_1\|, \quad \|\Delta M_2\| \le \tau_2\|M_2\|, \quad \|\Delta M\| \le \mathscr{O}(u)\,(\|M_1\|+\|M_2\|+\|A\|),$$

respectively, we can analogously obtain the normwise bounds for the error $\hat{x}_{k+1}-x$ and the residual $b-A\hat{x}_{k+1}$ as follows:

$$\|\hat{x}_{k+1}-x\| \lessapprox \sum_{i=0}^{k} \|G+\Delta G\|^i \left[\mathscr{O}(u)\,\|I-(G+\Delta G)\|\,\kappa(A)\left(\|x\|+\max_{0\le i\le k}\|\hat{x}_k\|\right)+u\max_{0\le i\le k}\|\hat{x}_k\|\right]$$

and

$$\|b-A\hat{x}_{k+1}\| \lessapprox \sum_{i=0}^{k} \|F+\Delta F\|^i \left[\mathscr{O}(u)\,\|I-(F+\Delta F)\|\left(\|b\|+\|A\|\max_{0\le i\le k}\|\hat{x}_k\|\right)+u\|A\|\max_{0\le i\le k}\|\hat{x}_k\|\right].$$

Provided that $\|G+\Delta G\|<1$ and $\|F+\Delta F\|<1$, and assuming that the normwise growth factor

$$\theta_k = \sup_{0\le i\le k}\left\{\frac{\|\hat{x}_i\|}{\|x\|}\right\}$$

is not too large, the above normwise bounds can be further simplified to

$$\frac{\|\hat{x}_{k+1}-x\|}{\|x\|} \lessapprox \mathscr{O}(u)\,\frac{1+\theta_k}{1-\|G+\Delta G\|}\left(\|I-(G+\Delta G)\|\,\kappa(A)+\frac{\theta_k}{1+\theta_k}\right)$$

and

$$\|b-A\hat{x}_{k+1}\| \lessapprox \mathscr{O}(u)\,\frac{\|I-(F+\Delta F)\|}{1-\|F+\Delta F\|}\left[\|b\|+\left(1+\frac{1}{\|I-(F+\Delta F)\|}\right)\theta_k\|A\|\|x\|\right].$$

Consequently, these bounds guarantee small normwise forward and backward errors, respectively, under mild conditions on the coefficient matrix $A$ as well as the splitting matrices $M_1$, $N_1$ and $M_2$, $N_2$.

In summary, if the iteration schemes (3.1)-(3.2) and (3.3)-(3.4) are either componentwise or normwise forward or backward stable, and if the splitting matrices $N_1$ and $N_2$ are as sparse and structured as the coefficient matrix $A$, then at the $k$-th iteration step of these two schemes computing the vectors $N_1x_k+b$ and $N_2x_{k+1/2}+b$ should be as costly as computing the residuals $b-Ax_k$ and $b-Ax_{k+1/2}$, respectively. So the iteration scheme (3.1)-(3.2) costs about the same workloads as the iteration scheme (3.3)-(3.4) at each iteration step. Roughly speaking, provided that the inner linear systems having the same coefficient matrices $M_1$ and $M_2$ are solved inexactly in accuracies controlled by the same tolerances $\tau_1$ and $\tau_2$, respectively, the iteration scheme (3.3)-(3.4) can always achieve higher computational efficiency than the iteration scheme (3.1)-(3.2).

# 4 Description of Implementations

In this section, we review the PMHSS and the HSS iteration methods [9, 7] and clarify their implementation settings. We remark that PMHSS and HSS are, respectively, the typical examples of the stationary single- and two-step matrix splitting iteration methods for solving the large sparse linear system (1.1); see also [12, 11, 6] and the references therein. Besides, we describe two experimental examples where a complex symmetric and a non-symmetric positive-definite linear systems arise.

The PMHSS iteration method is used to solve the linear system (1.1), with its coefficient matrix $A\in\mathbb{C}^{n,n}$ being complex symmetric and given by $A=W+\mathrm{i}T$, where $W,T\in\mathbb{R}^{n,n}$ are real, symmetric, and positive semidefinite matrices with, at least, one of them being positive definite. Here and in the sequel, we use $\mathrm{i}=\sqrt{-1}$ to denote the imaginary unit. A specific form of this iteration method is given by setting the iteration parameter $\alpha$ to be 1 and choosing the preconditioning matrix to be $W$, which has the following algorithmic description.

**Method 4.1. (The PMHSS Iteration Method [7])**
Let $x_0\in\mathbb{C}^n$ be an arbitrary initial guess. For $k=0,1,2,\ldots$ until the sequence of iterates $\{x_k\}_{k=0}^{\infty}\subset\mathbb{C}^n$ converges, compute the next iterate $x_{k+1}$ according to the following procedure:

$$(W+T)x_{k+1}=\frac{1+\mathrm{i}}{2}(W-\mathrm{i}T)x_k+\frac{1-\mathrm{i}}{2}b.$$

The PMHSS iteration scheme is induced by the matrix splitting $A = M - N$, with

$$M = (1+\mathrm{i})(W+T) \quad \text{and} \quad N = \mathrm{i}(W - \mathrm{i}T).$$

It alternatively admits the following equivalent form in terms of the residual.

**Method 4.2. (The PMHSS Iteration Method [7])**
Let $x_0 \in \mathbb{C}^n$ be an arbitrary initial guess. For $k = 0,1,2,\ldots$ until the sequence of iterates $\{x_k\}_{k=0}^{\infty} \subset \mathbb{C}^n$ converges, compute the next iterate $x_{k+1}$ according to the following procedure:

$$x_{k+1} = x_k + \frac{1-\mathrm{i}}{2}(W+T)^{-1}(b - Ax_k).$$

In fact, the PMHSS iteration method is a stationary single-step matrix splitting iteration method. It converges unconditionally to the unique solution of the complex symmetric linear system (1.1) for any initial guess if $\mathrm{null}(W) \cap \mathrm{null}(T) = \{0\}$. For distinction, we call Methods 4.1 and 4.2, respectively, the PMHSS iteration schemes I and II, or shortly, PMHSS-I and PMHSS-II, in the subsequent discussion.

In actual computations, we solve the linear sub-systems with respect to the coefficient matrix $W + T$ iteratively by the *preconditioned conjugate gradient* (**PCG**) method, with the incomplete Cholesky factorization [18] preconditioner (MATLAB code `ichol(sparse(·))`).

The HSS iteration method is used to solve the linear system (1.1) with its coefficient matrix $A \in \mathbb{C}^{n,n}$ being non-Hermitian and positive definite, i.e., its Hermitian part $\mathscr{H}(A) = \frac{1}{2}(A + A^*)$ is positive definite; see [9]. Denote by $\mathscr{S}(A) = \frac{1}{2}(A - A^*)$ the skew-Hermitian part of the matrix $A$. Then it holds that $A = \mathscr{H}(A) + \mathscr{S}(A)$, and the HSS iteration method can be algorithmically described as follows.

**Method 4.3. (The HSS Iteration Method [9])**
Let $x_0 \in \mathbb{C}^n$ be an arbitrary initial guess. For $k = 0,1,2,\ldots$ until the sequence of iterates $\{x_k\}_{k=0}^{\infty} \subset \mathbb{C}^n$ converges, compute the next iterate $x_{k+1}$ according to the following procedure:

$$\begin{cases} (\alpha I + \mathscr{H}(A))x_{k+1/2} &= (\alpha I - \mathscr{S}(A))x_k + b, \\ (\alpha I + \mathscr{S}(A))x_{k+1} &= (\alpha I - \mathscr{H}(A))x_{k+1/2} + b, \end{cases}$$

where $\alpha$ is a given positive constant.

The HSS iteration scheme is induced by the matrix splitting $A = M(\alpha) - N(\alpha)$, with

$$M(\alpha) = \frac{1}{2\alpha}(\alpha I + \mathscr{H}(A))(\alpha I + \mathscr{S}(A)) \quad \text{and} \quad N(\alpha) = \frac{1}{2\alpha}(\alpha I - \mathscr{H}(A))(\alpha I - \mathscr{S}(A)).$$

It alternatively admits the following equivalent form in terms of the residual.

**Method 4.4. (The HSS Iteration Method [9])**
Let $x_0 \in \mathbb{C}^n$ be an arbitrary initial guess. For $k = 0,1,2,\ldots$ until the sequence of iterates $\{x_k\}_{k=0}^{\infty} \subset \mathbb{C}^n$ converges, compute the next iterate $x_{k+1}$ according to the following procedure:

$$\begin{cases} x_{k+1/2} &= x_k + (\alpha I + \mathscr{H}(A))^{-1}(b - Ax_k), \\ x_{k+1} &= x_{k+1/2} + (\alpha I + \mathscr{S}(A))^{-1}(b - Ax_{k+1/2}), \end{cases}$$

where $\alpha$ is a given positive constant.

In fact, the HSS iteration method is a stationary two-step matrix splitting iteration method induced by the matrix splittings

$$\begin{array}{rclcrcl} M_1 &=& \alpha I + \mathscr{H}(A), & & N_1 &=& \alpha I - \mathscr{S}(A), \\ M_2 &=& \alpha I + \mathscr{S}(A), & & N_2 &=& \alpha I - \mathscr{H}(A). \end{array}$$

It converges unconditionally to the unique solution of the non-Hermitian positive definite linear system (1.1) for any initial guess. For distinction, we call Methods 4.3 and 4.4, respectively, the HSS iteration schemes I and II, or shortly, HSS-I and HSS-II, in the subsequent discussion.

In actual computations, the iteration parameter $\alpha$ is chosen to be the experimentally optimal one that minimizes the number of iteration steps of the HSS iteration method. We solve the linear sub-systems with respect to the coefficient matrices $\alpha I + \mathscr{H}(A)$ and $\alpha I + \mathscr{S}(A)$ iteratively by the PCG or the PCGNE (*preconditioned conjugate gradient for normal equation*) methods, with the incomplete Cholesky (MATLAB code `ichol(sparse(·))`) or the incomplete LU (MATLAB code `ilu(sparse(·))`) factorization preconditioners [18].

We describe in detail two experimental examples in the following.

**Example 4.1.** *(See [1, 6]) The linear system (1.1) is of the form*

$$\left[\left(K+\frac{3-\sqrt{3}}{\eta}I\right)+\mathrm{i}\left(K+\frac{3+\sqrt{3}}{\eta}I\right)\right]x=b, \tag{4.1}$$

*where $\eta$ is the time step-size and $K$ is the five-point centered difference matrix approximating the negative Laplacian operator $L=-(u_{xx}+u_{yy}+u_{zz})$ with homogeneous Dirichlet boundary conditions, on a uniform mesh in the unit cube $\Omega=[0,1]\times[0,1]\times[0,1]$ with the mesh-size $h=\frac{1}{m+1}$. The matrix $K\in\mathbb{R}^{n,n}$ possesses the tensor-product form $K=B_m\otimes I\otimes I+I\otimes B_m\otimes I+I\otimes I\otimes B_m$, with $B_m=h^{-2}\cdot\mathrm{tridiag}(-1,2,-1)\in\mathbb{R}^{m,m}$. Hence, $K$ is an $n\times n$ block-pentadiagonal matrix, with $n=m^3$. We take*

$$W=K+\frac{3-\sqrt{3}}{\eta}I \quad \textit{and} \quad T=K+\frac{3+\sqrt{3}}{\eta}I,$$

*and the right-hand side vector $b$ with its $j$th entry $[b]_j$ being given by*

$$[b]_j=\frac{(1-\mathrm{i})j}{\eta(j+1)^2}, \qquad j=1,2,\ldots,n.$$

*Furthermore, we normalize coefficient matrix and right-hand side by multiplying both by $h^2$.*

In our tests we take $\eta=h$. For more details about the practical backgrounds of this class of problems, we refer to [1, 6] and the references therein.

**Example 4.2.** *(See [9, 8]) Consider the linear system (1.1), for which $A\in\mathbb{R}^{n,n}$ is the upwind difference matrix of the three-dimensional convection-diffusion equation*

$$-(u_{xx}+u_{yy}+u_{zz})+\frac{q\cdot\exp(x+y+z)}{x+y+z}(xu_x+yu_y+zu_z)=f(x,y,z)$$

*on the unit cube $\Omega=[0,1]\times[0,1]\times[0,1]$ with the homogeneous Dirichlet boundary conditions. The step-sizes along all $x$, $y$ and $z$ directions are the same, i.e., $h=\frac{1}{m+1}$, and the right-hand side vector $b$ is taken to be $b=Ae$, with $e\in\mathbb{R}^n$ being the vector of all entries equal to 1. We denote by $Re=qh$ the mesh Reynolds number.*

All iteration processes are started from zero and terminated once the Euclidean norms of the current residuals are reduced by a factor of $10^8$ from those of the initial residuals. In addition, all codes are run in MATLAB (version R2013a) in double precision and all experiments are performed on a personal computer with 2.66GHz central processing unit (Intel(R) Core(TM)2 Duo CPU E6750), 2.00G memory and Windows operating system.

# 5 Numerical Results

By implementing the two equivalent schemes of the PMHSS iteration method used to solve Example 4.1 and those of the HSS iteration method used to solve Example 4.2, we show that the residual-update schemes, i.e., PMHSS-II

Table 1: Numerical Results of PMHSS Iteration Schemes for Example 4.1 at IT = 50

| $m$ | Method | Index | $\tau$ | | | | |
|---|---|---|---|---|---|---|---|
| | | | $10^{-4}$ | $10^{-6}$ | $10^{-8}$ | $10^{-10}$ | $10^{-12}$ |
| 32 | PMHSS-I | CPU | 2.69 | 1.73 | 3.02 | 4.69 | 6.94 |
| | | BERR | 1.06E-04 | 1.72E-06 | 1.34E-08 | 1.17E-10 | 1.49E-12 |
| | PMHSS-II | CPU | 10.35 | 10.88 | 11.57 | 12.76 | 15.92 |
| | | BERR | 5.47E-16 | 5.45E-16 | 5.48E-16 | 5.45E-16 | 5.47E-16 |
| 64 | PMHSS-I | CPU | 11.01 | 24.38 | 48.67 | 72.25 | 106.50 |
| | | BERR | 1.08E-04 | 1.83E-06 | 1.41E-08 | 1.32E-10 | 1.08E-12 |
| | PMHSS-II | CPU | 205.12 | 214.01 | 236.21 | 256.48 | 286.18 |
| | | BERR | 5.77E-16 | 5.64E-16 | 5.62E-16 | 5.63E-16 | 5.61E-16 |

and HSS-II, are always significantly more stable than the direct-splitting schemes, i.e., PMHSS-I and HSS-I, for large spectrums of the stopping tolerance(s) $\tau$ (or $\tau_1$ and $\tau_2$) of the inner iteration method(s). To this end, we report numerical results with respect to the number of iteration steps (denoted as "**IT**"), the computing time in seconds (denoted as "**CPU**"), and the norm of the backward error (denoted as "**BERR**") for these iteration schemes. Here BERR is defined as

$$\text{BERR} = \frac{\|b - Ax_k\|}{\|b\| + \|A\|\|x_k\|},$$

with $k$ being the iterate index.

At IT = 50, in Table 1 we list CPU and BERR for PMHSS-I and PMHSS-II when they are used to solve Example 4.1 with respect to different problem sizes and variant stopping tolerances. We observe that for each fixed $m$, the CPU for each scheme increases significantly when the tolerance $\tau$ becomes smaller; and for fixed $m$ and $\tau$, PMHSS-I always costs much less CPU than PMHSS-II. In Figures 1 and 2 we depict the curves of BERR versus IT when $m = 32$ and 64, with respect to variant stopping tolerances for PMHSS-I and PMHSS-II when they are used to solve Example 4.1. From Table 1 and Figures 1-2 we observe that for fixed $m$ the norm of backward error of PMHSS-I is of the same order of magnitude as $\tau$, but that of PMHSS-II is always of the order $\mathcal{O}(u)$ of the machine precision $u$ by no matter whether $\tau$ is large or small. Hence, in actual computations PMHSS-II is always backward stable independent of the tolerance $\tau$, but PMHSS-I may be backward stable only for those $\tau$ of about the order $\mathcal{O}(u)$ of magnitude. As a result, the two equivalent implementations of the exact PMHSS iteration method have about the same stability property and convergence behavior.

At IT = 250, in Table 2 we list CPU and BERR for HSS-I and HSS-II when they are used to solve Example 4.2 with respect to $m = 64$, Re = 10, and variant stopping tolerances. We observe that the CPU for each scheme increases significantly when either of the tolerances $\tau_1$ and $\tau_2$ becomes smaller; and for fixed $\tau_1$ and $\tau_2$, HSS-I always costs much less CPU than HSS-II. Moreover, the norm of backward error of HSS-I is of an order of magnitude like $\mathcal{O}(\max\{\tau_1, \tau_2\})$, but that of HSS-II is always of the order $\mathcal{O}(u)$ of the machine precision $u$ by no matter whether $\tau_1$ or $\tau_2$ is large or small; see Figure 3 in which $\tau \equiv \tau_1 = \tau_2$. Hence, in actual computations HSS-II is always backward stable independent of the tolerances $\tau_1$ and $\tau_2$, but HSS-I may be backward stable only for those $\tau_1$ and $\tau_2$ of about the order $\mathcal{O}(u)$ of magnitude. As a result, the two equivalent implementations of the exact HSS iteration method have about the same stability property and convergence behavior.

With regard to Tables 1 and 2, one reason for the CPUs of PMHSS-I and HSS-I being much less than the CPUs of PMHSS-II and HSS-II is that the stopping criterions of the inner iteration methods adopted in PMHSS-I and HSS-I are much more easily achievable than those adopted in PMHSS-II and HSS-II, respectively, especially when the iterates are approaching to the exact solution of the system of linear equations (1.1). Admittedly, as the inexactly computed solutions have very different accuracy, the CPUs here do not reflect the computing efficiency of both iteration schemes, and they only show the overall (or the average) computational costs of the inner iterations, or in other words, the average numbers of inner iteration steps.

Table 2: Numerical Results of HSS for Example 4.2 with $m = 64$ and Re $= 10$ at IT $= 250$

| $\tau_1$ | Method | Index | $\tau_2$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $10^{-4}$ | $10^{-6}$ | $10^{-8}$ | $10^{-10}$ | $10^{-12}$ | $10^{-14}$ |
| $10^{-4}$ | HSS-I | CPU | 250.17 | 339.56 | 423.30 | 510.85 | 580.82 | 583.76 |
| | | BERR | 3.01E-04 | 8.66E-05 | 1.23E-04 | 1.21E-04 | 1.21E-04 | 1.21E-04 |
| | HSS-II | CPU | 389.48 | 431.18 | 495.28 | 539.88 | 621.17 | 703.32 |
| | | BERR | 7.33E-17 | 7.46E-17 | 7.34E-17 | 7.34E-17 | 7.48E-17 | 7.41E-17 |
| $10^{-6}$ | HSS-I | CPU | 253.96 | 345.61 | 426.30 | 519.16 | 584.64 | 585.40 |
| | | BERR | 2.52E-05 | 1.56E-06 | 2.51E-06 | 1.09E-06 | 1.09E-06 | 1.09E-06 |
| | HSS-II | CPU | 402.32 | 442.74 | 502.64 | 559.40 | 626.10 | 714.77 |
| | | BERR | 7.44E-17 | 7.42E-17 | 7.46E-17 | 7.44E-17 | 7.44E-17 | 7.39E-17 |
| $10^{-8}$ | HSS-I | CPU | 272.11 | 352.52 | 449.93 | 524.44 | 597.51 | 622.80 |
| | | BERR | 2.50E-05 | 1.54E-07 | 9.74E-09 | 7.98E-09 | 1.30E-08 | 1.30E-08 |
| | HSS-II | CPU | 455.02 | 480.71 | 513.60 | 574.09 | 641.13 | 731.26 |
| | | BERR | 7.47E-17 | 7.47E-17 | 7.34E-17 | 7.41E-17 | 7.46E-17 | 7.35E-17 |
| $10^{-10}$ | HSS-I | CPU | 304.95 | 391.21 | 466.61 | 578.35 | 697.57 | 941.22 |
| | | BERR | 2.50E-05 | 1.48E-07 | 1.94E-09 | 3.64E-10 | 2.20E-10 | 2.20E-10 |
| | HSS-II | CPU | 437.73 | 482.88 | 535.62 | 585.11 | 655.26 | 739.58 |
| | | BERR | 7.46E-17 | 7.52E-17 | 7.53E-17 | 7.48E-17 | 7.34E-17 | 7.45E-17 |
| $10^{-12}$ | HSS-I | CPU | 508.51 | 614.75 | 735.35 | 824.93 | 903.22 | 985.34 |
| | | BERR | 2.50E-05 | 1.48E-07 | 1.90E-09 | 2.48E-11 | 3.78E-12 | 3.78E-12 |
| | HSS-II | CPU | 459.49 | 532.42 | 579.38 | 609.75 | 683.28 | 768.80 |
| | | BERR | 7.61E-17 | 7.44E-17 | 7.50E-17 | 7.47E-17 | 7.49E-17 | 7.19E-17 |
| $10^{-14}$ | HSS-I | CPU | 520.55 | 668.16 | 764.35 | 756.78 | 624.93 | 628.16 |
| | | BERR | 2.50E-05 | 1.48E-07 | 1.90E-09 | 2.48E-11 | 3.78E-12 | 3.78E-12 |
| | HSS-II | CPU | 481.66 | 521.08 | 589.94 | 642.01 | 713.43 | 806.06 |
| | | BERR | 7.49E-17 | 7.51E-17 | 7.40E-17 | 7.46E-17 | 7.42E-17 | 7.45E-17 |

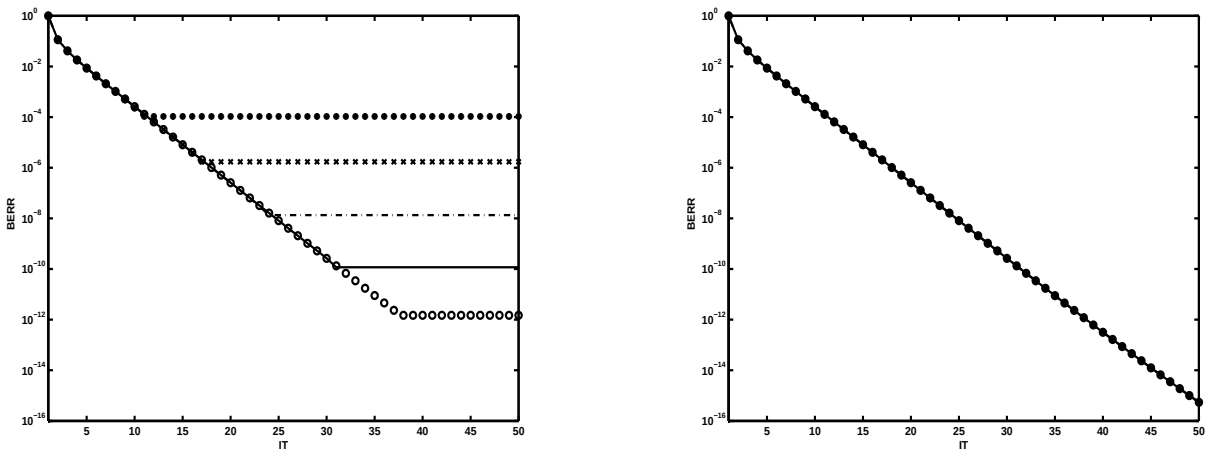


Figure 1: Pictures of BERR versus IT for PMHSS when $m = 32$, with PMHSS-I (left) and PMHSS-II (right). $\tau = 10^{-4}$: the star line "$***$", $\tau = 10^{-6}$: the cross line "$\times\times\times$", $\tau = 10^{-8}$: the dash-dotted line "$-\cdot-\cdot-\cdot$", $\tau = 10^{-10}$: the solid line "$---$", and $\tau = 10^{-12}$: the circle line "$\circ\circ\circ$".

# 6 Concluding Remarks

Stationary matrix splitting iteration methods for solving large sparse systems of linear equations have two typical equivalent reformulations: the residual-update scheme and the direct-splitting scheme. Both theoretical analyses and numerical experiments have shown that the former is always significantly more stable than the later for a large spectrum of the stopping tolerance of the inner iteration method. Moreover, for both reformulations, inexact solutions of inner linear systems associated with the matrix splittings may considerably influence the convergence and the accuracy of the approximate solutions computed in finite precision arithmetic, a finer tolerance often costs more computing time, and their exact implementations have about the same stability property and convergence behavior. These conclusions hold equally true for both single- and two-step matrix splitting iteration methods.

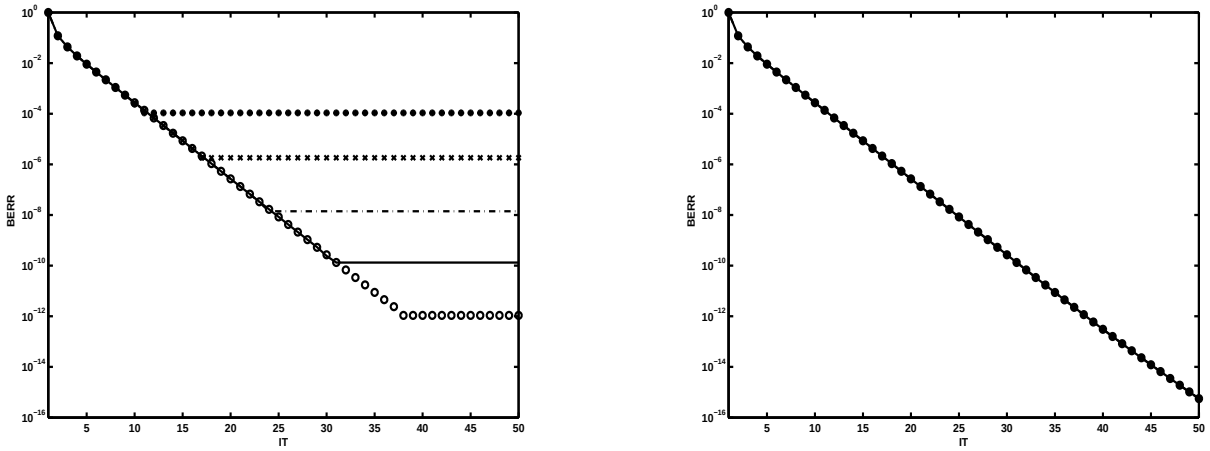


Figure 2: Pictures of BERR versus IT for PMHSS when $m = 64$, with PMHSS-I (left) and PMHSS-II (right). $\tau = 10^{-4}$: the star line "$***$", $\tau = 10^{-6}$: the cross line "$\times\times\times$", $\tau = 10^{-8}$: the dash-dotted line "$-\cdot-\cdot-\cdot$", $\tau = 10^{-10}$: the solid line "$---$", and $\tau = 10^{-12}$: the circle line "$\circ\circ\circ$".

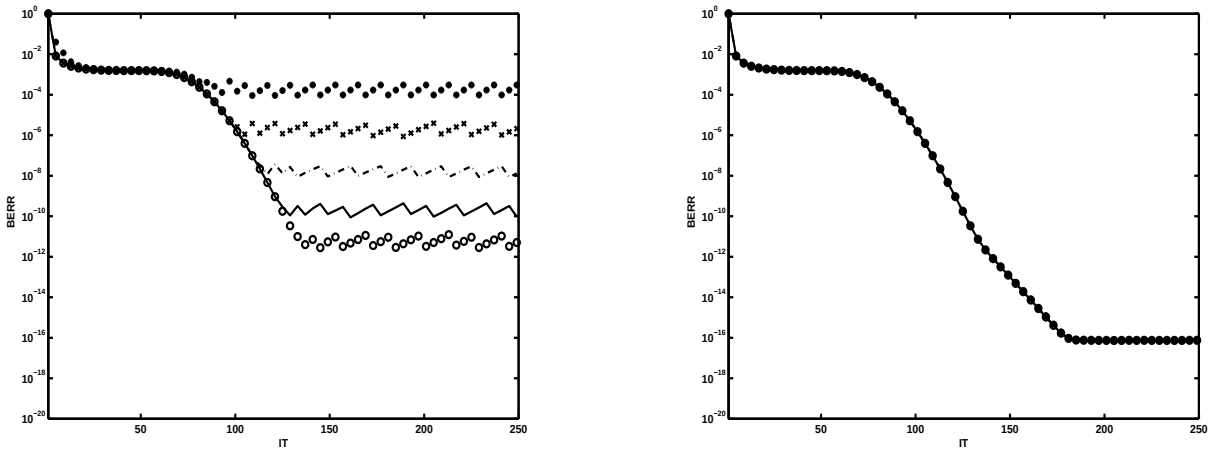


Figure 3: Pictures of BERR versus IT for HSS when $m = 64$ and $\tau_1 = \tau_2 \equiv \tau$, with HSS-I (left) and HSS-II (right). $\tau = 10^{-4}$: the star line "$***$", $\tau = 10^{-6}$: the cross line "$\times\times\times$", $\tau = 10^{-8}$: the dash-dotted line "$-\cdot-\cdot-\cdot$", $\tau = 10^{-10}$: the solid line "$---$", and $\tau = 10^{-12}$: the circle line "$\circ\circ\circ$".

# References


[1] O. Axelsson and A. Kucherov, Real valued iterative methods for solving complex symmetric linear systems, *Numer. Linear Algebra Appl.*, 7(2000), 197-218.

[2] Z.-Z. Bai, A class of two-stage iterative methods for systems of weakly nonlinear equations, *Numer. Algorithms*, 14(1997), 295-319.

[3] Z.-Z. Bai, On the convergence of additive and multiplicative splitting iterations for systems of linear equations, *J. Comput. Appl. Math.*, 154(2003), 195-214.

[4] Z.-Z. Bai, An algebraic convergence theorem for the multiplicative Schwarz iteration, *Numer. Math.-JCU (English Ser.)*, 12(2003), 179-182.

[5] Z.-Z. Bai, Splitting iteration methods for non-Hermitian positive definite systems of linear equations, *Hokkaido Math. J.*, 36(2007), 801-814.

[6] Z.-Z. Bai, M. Benzi and F. Chen, Modified HSS iteration methods for a class of complex symmetric linear systems, *Computing*, 87(2010), 93-111.

[7] Z.-Z. Bai, M. Benzi and F. Chen, On preconditioned MHSS iteration methods for complex symmetric linear systems, *Numer. Algorithms*, 56(2011), 297-317.

[8] Z.-Z. Bai, G.H. Golub, L.-Z. Lu and J.-F. Yin, Block triangular and skew-Hermitian splitting methods for positive-definite linear systems, *SIAM J. Sci. Comput.*, 26(2005), 844-863.

[9] Z.-Z. Bai, G.H. Golub and M.K. Ng, Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems, *SIAM J. Matrix Anal. Appl.*, 24(2003), 603-626.

[10] Z.-Z. Bai, G.H. Golub and M.K. Ng, On successive overrelaxation acceleration of the Hermitian and skew-Hermitian splitting iterations, *Numer. Linear Algebra Appl.*, 14(2007), 319-335.

[11] Z.-Z. Bai, G.H. Golub and M.K. Ng, On inexact Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems, *Linear Algebra Appl.*, 428(2008), 413-440.

[12] Z.-Z. Bai, J.-C. Sun and D.-R. Wang, A unified framework for the construction of various matrix multi-splitting iterative methods for large sparse system of linear equations, *Computers Math. Appl.*, 32(1996), 51-76.

[13] Z.-Z. Bai and D.-R. Wang, The monotone convergence of the two-stage iterative method for solving large sparse systems of linear equations, *Appl. Math. Lett.*, 10(1997), 113-117.

[14] M. Benzi and D.B. Szyld, Existence and uniqueness of splittings for stationary iterative methods with applications to alternating methods, *Numer. Math.*, 76(1997), 309-321.

[15] Z.-H. Cao, Rounding error analysis of two-stage iterative methods for large linear systems, *Appl. Math. Comput.*, 139(2003), 371-381.

[16] J. Douglas and H.H. Rachford, On the numerical solution of heat conduction problems in two and three space variables, *Trans. Amer. Math. Soc.*, 82(1956), 421-439.

[17] A. Frommer and D.B. Szyld, $H$-splittings and two-stage iterative methods, *Numer. Math.*, 63(1992), 345-356.

[18] G.H. Golub and C.F. Van Loan, Matrix Computations, Third Edition, *The Johns Hopkins University Press*, Baltimore and London, 1996.

[19] L.A. Hageman and D.M. Young, Applied Iterative Methods, *Academic Press*, New York, 1981.

[20] N.J. Higham, Accuracy and Stability of Numerical Algorithms, Second Edition, *SIAM*, Philadelphia, PA, 2002.

[21] N.J. Higham and P.A. Knight, Componentwise error analysis for stationary iterative methods, In "Linear Algebra, Markov Chains, and Queueing Models" (C.D. Meyer and R.J. Plemmons eds.), *IMA Volumes in Mathematics and its Applications*, 48(1993), 29-46.

[22] E. Isaacson and H.B. Keller, Analysis of Numerical Methods, *Dover Publications Inc.*, New York, 1966.

[23] P. Jiránek and M. Rozložník, Maximum attainable accuracy of inexact saddle point solvers, *SIAM J. Matrix Anal. Appl.*, 29(2008), 1297-1321.

[24] P. Jiránek and M. Rozložník, Limiting accuracy of segregated solution methods for nonsymmetric saddle point problems, *J. Comput. Appl. Math.*, 215(2008), 28-37.

[25] R.J. Lanzkron, D.J. Rose and D.B. Szyld, Convergence of nested classical iterative methods for linear systems, *Numer. Math.*, 58(1991), 685-702.

[26] N.K. Nichols, On the convergence of two-stage iterative processes for solving linear equations, *SIAM J. Numer. Anal.*, 10(1973), 460-469.

[27] D.W. Peaceman and H.H. Rachford, The numerical solution of parabolic and elliptic differential equations, *J. Soc. Indust. Appl. Math.*, 3(1955), 28-41.

[28] H.H. Rachford, Rounding errors in alternating direction methods for parabolic problems, *SIAM J. Numer. Anal.*, 5(1968), 407-421.

[29] V. Simoncini and D.B. Szyld, Theory of inexact Krylov subspace methods and applications to scientific computing, *SIAM J. Sci. Comput.*, 25(2003), 454-477.

[30] J. Van den Eshof and G.L.G. Sleijpen, Inexact Krylov subspace methods for linear systems, *SIAM J. Matrix Anal. Appl.*, 26(2004), 125-153.

[31] R.S. Varga, Matrix Iterative Analysis, *Springer-Verlag*, Berlin and Heidelberg, 2000.

[32] Q. Xiang and S.-P. Wu, A modified alternating direction method for positive definite systems, In "Information Engineering and Applications" (R.-B. Zhu and Y. Ma eds.), *Lecture Notes in Electrical Engineering*, 154(2012), 437-444.

[33] Q. Xiang, S.-P. Wu and Y. Xu, Alternating lower-upper splitting iterative method and its application in CFD, *J. Beijing Univ. Aeron. Astron.*, 38(2012), 953-956. (In Chinese)