On the symmetric componen twise relative backward error for linear systems of equations

Stanley Eisenstadt, Serge Gratton, David Titley-Peloquin

ON THE SYMMETRIC COMPONEN TWISE RELATIVE BACKWARD ERROR FOR LINEAR SYSTEMS OF EQUATIONS*

STANLEY C. EISENSTAT†, SERGE GRATTON‡, AND DAVID TITLEY-PELOQUIN§

Abstract. We derive an upper bound on the symmetric componentwise relative backward error for symmetric linear systems of equations. Since the bound can be computed efficiently and, except for some artificial examples, seems to be of the same order of magnitude as the true symmetric componentwise backward error, we believe that it is suitable for practical use. Our results also provide new insight into the relationship between the symmetric and unsymmetric backward errors.

Key words. componentwise backward error, symmetric backward error

DOI. 10.1137/140986566

1. Introduction. During an "open problems" session at the Householder Symposium XIX on Numerical Linear Algebra, held in Spa, Belgium, James Demmel presented the following challenge: Solve the *symmetric componentwise relative backward error* problem for linear systems of equations. In this note we present some progress that we have made toward its solution.

Suppose that we are given an approximate solution $x \in \mathbb{R}^n$ to the linear system $Au = b$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular, $b \in \mathbb{R}^n$, and $u \in \mathbb{R}^n$. The backward error problem is to find minimal (in some sense) perturbations $\Delta A \in \mathbb{R}^{n\times n}$ and $\Delta b \in \mathbb{R}^n$ such that x is the exact solution of the perturbed problem

$$(A + \Delta A)x = b + \Delta b.$$

The size of the smallest perturbation is called the *backward error* and can be measured normwise or componentwise. Here we consider a *componentwise relative* measure that is often used in conjunction with sparse direct solvers (see, e.g., [1]).

When the matrix A has some structure, it is natural to require that the perturbation ΔA also have this structure. Here we consider the case where A is symmetric, and we require that ΔA be symmetric as well.

The unstructured componentwise and normwise backward error problems have been solved for nearly fifty years [8, 9]. The symmetric normwise backward error problem has also long been settled [2]. The symmetric componentwise variant is more difficult: Although some results exist [13, 4, 12], it remains open.

In this work we derive an upper bound on the symmetric componentwise relative backward error, whose precise definition is given in (6) below. The bound can be computed efficiently and, except for some artificial examples, seems to be of the same

Funding: The work of the authors was conducted with the support of the "Fondation Sciences et Technologies pour l'Aéronautique et l'Espace (STAE)," within the "Réseau Thématique de Recherche Avancée (RTRA)," Toulouse, France.

†Department of Computer Science, Yale University, New Haven, CT 06520 (stanley.eisenstat@yale.edu).

‡INPT-IRIT-ENSEEIHT, B. P. 7122, 31071 Toulouse, France (serge.gratton@enseeiht.fr).

§Department of Bioresource Engineering, McGill University, Ste-Anne-de-Bellevue, Qc, H9X 3V9, Canada (david.titley-peloquin@mcgill.ca).

order of magnitude as the true symmetric backward error. Therefore, we believe that it is suitable for practical use. Our results also provide new insight into the relationship between the symmetric and unsymmetric backward errors.

Error bounds for the approximate solution x can be obtained by combining the structured backward error and a structured perturbation analysis. For an overview of structured perturbation results for linear systems of equations we refer to [4, 5, 10, 11] or [6, Chapter 7] and the references therein.

1.1. Notation. We use subscripts to denote elements of vectors or subvectors of partitioned vectors, and similarly for matrices, while superscripts denote terms in a sequence. For example, $A_{ij}^{(k)}$ denotes the (i,j)th element of the kth term in the sequence $\{A^{(k)}\}$. For $u \in \mathbb{R}^n$ and $v \in \mathbb{R}$ we define

$$\operatorname{diag}(u) = \begin{bmatrix} u_1 & & \\ & \ddots & \\ & & u_n \end{bmatrix}, \qquad \operatorname{sign}(v) = \begin{cases} +1 & \text{if } v > 0, \\ -1 & \text{if } v < 0, \\ 0 & \text{if } v = 0. \end{cases}$$

Absolute values and inequalities are meant componentwise; that is, $|A|$ is the matrix whose entries are $|A_{ij}|$, and $|A| \le |B|$ means that $|A_{ij}| \le |B_{ij}|$ for all pairs (i,j).

1.2. The componentwise backward error. Given $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, and $x \in \mathbb{R}^n$, the componentwise relative backward error for the linear system $Au = b$ is defined as

$$\epsilon^* = \min_{\epsilon, \Delta A, \Delta b} \left\{ \epsilon : (A + \Delta A)x = b + \Delta b,\ |\Delta A| \le \epsilon |A|,\ |\Delta b| \le \epsilon |b| \right\}. \tag{1}$$

Oettli and Prager [8] showed that

$$\epsilon^* = \max_i \frac{(|b - Ax|)_i}{(|A|\,|x| + |b|)_i}$$

with the convention that $0/0 = 0$. Since $(|b - Ax|)_i \le (|b| + |A||x|)_i$, it must be the case that $0 \le \epsilon^* \le 1$ with $\epsilon^* = 0$ if and only if $Ax = b$.

Let $r = b - Ax$, let d be the vector whose elements are

$$d_i = \begin{cases} (|A||x| + |b|)_i & \text{if } (|A||x| + |b|)_i \neq 0, \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

and let

$$D = \operatorname{diag}(d), \quad z = D^{-1}r, \quad Z = \operatorname{diag}(z), \quad S = \operatorname{diag}(\operatorname{sign}(x)). \tag{3}$$

Then

$$\epsilon^* = \|z\|_\infty, \tag{4}$$

and the minimum in (1) is attained by the perturbations

$$\Delta A^* = Z|A|S, \qquad \Delta b^* = -Z|b|. \tag{5}$$

See [6, section 7.2] for a short proof.

Analogously to (1), for symmetric A the *symmetric* componentwise relative backward error for $Ax = b$ is defined as

$$\epsilon^*_{\rm sym} = \min_{\epsilon,\Delta A,\Delta b} \big\{\, \epsilon : (A+\Delta A)x = b + \Delta b, \quad |\Delta A| \le \epsilon |A|,\ |\Delta b| \le \epsilon |b|,\ \Delta A = (\Delta A)^T \,\big\}. \tag{6}$$

Since the perturbations $\Delta A = -A$ and $\Delta b = -b$ satisfy the constraints with $\epsilon = 1$, we have $0 \le \epsilon^*_{\rm sym} \le 1$ with $\epsilon^*_{\rm sym} = 0$ if and only if $Ax = b$.

Because (6) is identical to (1) except for the additional symmetry constraint, it must be the case that $\epsilon^*_{\rm sym} \ge \epsilon^*$. Rump [12] has recently given an example (see (20)) where $\epsilon^*_{\rm sym}$ can be arbitrarily larger than ϵ^*. However, this result cannot be used to compute or bound $\epsilon^*_{\rm sym}$ given a specific triplet (A, b, x).

Variants of (6) have appeared in the literature. For example, Smoktunowicz [13] (see also [6, Problem 7.12]) considers perturbations only in the matrix A,

$$\min_{\epsilon,\Delta A} \big\{\, \epsilon \,:\, (A+\Delta A)x = b,\ |\Delta A| \le \epsilon |A| \big\}, \tag{7a}$$

$$\min_{\epsilon,\Delta A} \big\{\, \epsilon \,:\, (A+\Delta A)x = b,\ |\Delta A| \le \epsilon |A|,\ \Delta A = (\Delta A)^T \,\big\}, \tag{7b}$$

and shows that if A is symmetric and diagonally dominant, the minimum in (7b) is at most a factor of three larger than that in (7a). This is also true for symmetric positive definite matrices [13, Theorem 3.1]. On the other hand, Higham and Higham [4] give an example of a general symmetric A (see (19)) where the minimum in (7b) can be arbitrarily larger than that in (7a). Since perturbations to b are readily accounted for in structured perturbation analyses, we include them in (6).

More generally, one can also replace $|A|$ and $|b|$ in (1) and (6) with arbitrary nonnegative tolerances E and f. For example, in the symmetric case Higham and Higham [4] consider

$$\min_{\epsilon,\Delta A,\Delta b} \big\{\, \epsilon \,:\, (A+\Delta A)x = b + \Delta b,\ |\Delta A| \le \epsilon E,\ |\Delta b| \le \epsilon f,\ \Delta A = (\Delta A)^T \,\big\}. \tag{8}$$

However, $E = |A|$ and $f = |b|$ is a natural and common choice, and for simplicity we have chosen to work with the less general (but still useful) formulation (6).

Each of these backward error problems is a linear program (LP) and can in principle be solved using any general LP solver. Alternatively, Higham and Higham [4] show how the solution can be computed as the minimum ∞-norm solution of an underdetermined linear system of equations. However, when the problem dimensions are large, these approaches are too expensive to be of practical use. It would be useful to have an expression for (or a tight upper bound on) the symmetric backward error $\epsilon^*_{\rm sym}$ that can be computed cheaply, analogously to (4) in the unsymmetric case. This was the goal of the present work.

1.3. Our contributions. We make some progress toward solving (6) by deriving an upper bound $\tilde{\epsilon}_{\rm sym}$ on $\epsilon^*_{\rm sym}$. Combined with a structured perturbation analysis, it leads to upper bounds on the error in x. It can also be used to compute a bound on the ratio $\epsilon^*_{\rm sym}/\epsilon^*$ for a given triplet (A, b, x).

In the following section we derive our upper bound. In section 3 we investigate whether it is tight. In section 4 we show how it can be computed efficiently, even when the problem dimensions are very large. In section 5 we give some illustrative examples and a summary of our numerical experiments.

2. A bound on the symmetric componentwise backward error. We start by presenting our main result.

THEOREM 2.1. *Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and vectors $b \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$, define $r = b - Ax$, the vector d as in (2), and the vector z and the matrices D and S as in (3). Then the matrix*

$$N = D^{-1}\big(\operatorname{diag}\big(\tfrac{1}{2}|A||x| + |b|\big) + \tfrac{1}{2}S|A|S\operatorname{diag}(|x|)\big) \tag{9}$$

is diagonally dominant with nonnegative diagonal elements, and the linear system $N\tilde{z} = z$ is consistent. Moreover, for any solution $\tilde{z}$, the perturbations

$$\widetilde{\Delta A} = \tfrac{1}{2}\big(\widetilde{Z}|A|S + S|A|\widetilde{Z}\big), \qquad \widetilde{\Delta b} = -\widetilde{Z}|b|, \tag{10}$$

where $\widetilde{Z} = \operatorname{diag}(\tilde{z})$, satisfy

$$(A + \widetilde{\Delta A})x = b + \widetilde{\Delta b}, \qquad \widetilde{\Delta A} = (\widetilde{\Delta A})^T; \tag{11}$$

and the symmetric backward error $\epsilon^*_{\mathrm{sym}}$ in (6) satisfies

$$\epsilon^*_{\mathrm{sym}} \le \max\big\{\,\tilde{\epsilon}_A,\, \tilde{\epsilon}_b\,\big\} \le \|\tilde{z}\|_\infty \equiv \tilde{\epsilon}_{\mathrm{sym}}, \tag{12}$$

where

$$\tilde{\epsilon}_A = \max_{i,j} \frac{|\widetilde{\Delta A}_{ij}|}{|A_{ij}|}, \qquad \tilde{\epsilon}_b = \max_j \frac{|\widetilde{\Delta b}_j|}{|b_j|} \tag{13}$$

with the convention that $0/0 = 0$.

Note the similarity to the Oettli and Prager [8] result for the unsymmetric backward error. The final bound $\|\tilde{z}\|_\infty$ in (12) has the same form as (4) and the perturbations $\widetilde{\Delta A}$ and $\widetilde{\Delta b}$ in (10) are the same as those in (5), except that z and Z have been replaced by $\tilde{z}$ and $\widetilde{Z}$, respectively, and $\widetilde{\Delta A}$ has been symmetrized.

Proof. Consider the perturbations in (10). Clearly $\widetilde{\Delta A}$ is symmetric. In order to satisfy the other equality constraint in (11) we must have

$$\begin{aligned}
0 &= (A + \widetilde{\Delta A})x - (b + \widetilde{\Delta b}) \\
&= \tfrac{1}{2}\big(\widetilde{Z}|A|S + S|A|\widetilde{Z}\big)x + \widetilde{Z}|b| - (b - Ax) \\
&= \widetilde{Z}\big(\tfrac{1}{2}|A|Sx + |b|\big) + \tfrac{1}{2}S|A|\widetilde{Z}x - r \\
&= \operatorname{diag}\big(\tfrac{1}{2}|A||x| + |b|\big)\,\tilde{z} + \tfrac{1}{2}S|A|S\operatorname{diag}(|x|)\,\tilde{z} - Dz \\
&= D\big(N\tilde{z} - z\big).
\end{aligned}$$

Thus, if $N\tilde{z} = z$, then the perturbations $\widetilde{\Delta A}$ and $\widetilde{\Delta b}$ satisfy both equality constraints in (11), and by (6) we have

$$\epsilon^*_{\mathrm{sym}} \le \max\big\{\,\tilde{\epsilon}_A,\, \tilde{\epsilon}_b\,\big\}.$$

From (10) it follows that $\widetilde{\Delta A}_{ij} = 0$ if $A_{ij} = 0$, and

$$\frac{|\widetilde{\Delta A}_{ij}|}{|A_{ij}|} \le \tfrac{1}{2}|\tilde{z}_i| + \tfrac{1}{2}|\tilde{z}_j| \le \|\tilde{z}\|_\infty$$

otherwise, while $\widetilde{\Delta b}_j = 0$ if $b_j = 0$, and

$$\frac{|\widetilde{\Delta b}_j|}{|b_j|} = |\tilde{z}_j| \leq \|\tilde{z}\|_\infty$$

otherwise. Therefore,

$$\epsilon^*_{\text{sym}} \leq \|\tilde{z}\|_\infty = \tilde{\epsilon}_{\text{sym}}.$$

We now show that N is diagonally dominant with nonnegative diagonal entries. From (9) it follows that

$$N_{ii} = \frac{\left(\frac{1}{2}|A||x| + |b|\right)_i}{d_i} + \frac{|A_{ii}||x_i|}{2d_i} = \frac{|A_{ii}||x_i| + |b_i|}{d_i} + \sum_{j \neq i} \frac{|A_{ij}||x_j|}{2d_i} \geq 0,$$

while, for $j \neq i$,

$$|N_{ij}| = \begin{cases} 0 & \text{if } x_i = 0, \\ \frac{1}{2d_i}|A_{ij}||x_j| & \text{otherwise.} \end{cases}$$

Therefore

$$|N_{ii}| - \sum_{j \neq i} |N_{ij}| \geq \frac{|A_{ii}||x_i| + |b_i|}{d_i} \geq 0. \tag{14}$$

Finally we show by contradiction that the linear system $N\tilde{z} = z$ is consistent. Suppose it is not. Then the matrix $N = N(A, b, x)$ must be singular.

If any entries in x are zero, we can symmetrically permute the rows and columns so that

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix},$$

where $|x_1| > 0$, and partition the other variables conformally. Then $N\tilde{z} = z$ is equivalent to

$$\begin{bmatrix} N(A_{11}, b_1, x_1) & 0 \\ 0 & D_2^{-1} \operatorname{diag}\left(\frac{1}{2}|A_{21}||x_1| + |b_2|\right)) \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Because

$$|z_2| = \left|D_2^{-1} r_2\right| = D_2^{-1}|b_2 - A_{21}x_1| \leq 2D_2^{-1}\left(\tfrac{1}{2}|A_{21}||x_1| + |b_2|\right),$$

if a diagonal entry in the $(2,2)$ block of the coefficient matrix above is zero, then the corresponding entry on the right-hand side must also be zero and that equation is consistent. If the diagonal entry is nonzero, that equation has a unique solution. That is, the second block of equations is consistent, so the system $N(A_{11}, b_1, x_1)\tilde{z}_1 = z_1$ must be inconsistent. Thus without loss of generality we may assume that $|x| > 0$.

If A is reducible we can symmetrically permute the rows and columns so that

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

and partition the other variables conformally. Then $N\tilde{z} = z$ is equivalent to

$$\begin{bmatrix} N(A_{11}, b_1, x_1) & 0 \\ 0 & N(A_{22}, b_2, x_2) \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

and at least one of the systems $N(A_{ii}, b_i, x_i)\tilde{z}_i = z_i$ must be inconsistent. By applying the same reasoning recursively, we see that without loss of generality we may assume that A is irreducible.

Since $|x| > 0$, the matrix N has the same off-diagonal nonzero structure as A and must also be irreducible; and by (14) it is diagonally dominant. If any entry in b were nonzero, then the corresponding row of N would be strictly diagonally dominant, so that N would be irreducibly diagonally dominant and nonsingular. Thus without loss of generality we may assume that $b = 0$.

Because N is singular, there exists a vector $u \neq 0$ such that

$$0 = Nu = D^{-1} \operatorname{diag}\left(\tfrac{1}{2}|A||x|\right) u + \tfrac{1}{2} D^{-1} S|A|S \operatorname{diag}(|x|)\, u.$$

Let $u_i \neq 0$ be an element of u of maximum magnitude. Then

$$0 = (Nu)_i = \frac{u_i}{2d_i} \sum_{j=1}^{n} |A_{ij}||x_j| \left(1 + \operatorname{sign}(x_i) \operatorname{sign}(x_j) \frac{u_j}{u_i}\right).$$

Since $|u_j| \leq |u_i|$, all terms in the above sum are nonnegative. For the sum to be zero we must have $A_{ii} = 0$ and $u_j = -u_i \operatorname{sign}(x_i x_j)$ for every j for which $|A_{ij}| > 0$; that is, u_j must also be an element of maximum magnitude. Applying the same argument repeatedly, we must have $|A_{kk}| = 0$ and $u_k = (-1)^\ell u_i \operatorname{sign}(x_i x_k)$ whenever there is a path of length ℓ from node i to node k in the graph of A. Since A is irreducible, every node k is reachable and all entries in u are so defined. Thus u is unique up to a scaling factor, and the null-space of N is one-dimensional.

Since the sign of u_k depends only on whether the lengths of all paths from node i to node k in the graph of A are all even or all odd, the matrix A must have Young's property $\mathcal{A}$: there exist two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathcal{W} = \{1, \ldots, n\}$ such that $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{W}$; and if $A_{ij} \neq 0$, then $i \in \mathcal{S}_1$ and $j \in \mathcal{S}_2$ or $i \in \mathcal{S}_2$ and $j \in \mathcal{S}_1$. It follows (see, e.g., [15, section 2.6]) that after symmetrically permuting the rows and columns of A we can write

$$A = \begin{bmatrix} 0 & A_{21}^T \\ A_{21} & 0 \end{bmatrix}.$$

Partitioning all other variables conformally, we have

$$N = \tfrac{1}{2} D^{-1} \begin{bmatrix} \operatorname{diag}\left(|A_{21}^T||x_2|\right) & S_1|A_{21}^T|S_2 \operatorname{diag}(|x_2|) \\ S_2|A_{21}|S_1 \operatorname{diag}(|x_1|) & \operatorname{diag}(|A_{21}|x_1|) \end{bmatrix}.$$

Let $v = D \left[\begin{smallmatrix} x_1 \\ -x_2 \end{smallmatrix}\right]$. Then

$$N^T v = \tfrac{1}{2} \begin{bmatrix} \operatorname{diag}\left(|A_{21}^T||x_2|\right) x_1 - \operatorname{diag}(|x_1|)\, S_1|A_{21}^T|S_2 x_2 \\ \operatorname{diag}(|x_2|)\, S_2|A_{21}|S_1 x_1 - \operatorname{diag}(|A_{21}|x_1|)\, x_2 \end{bmatrix} = 0;$$

that is, v spans the null space of N^T. On the other hand, since $r = -Ax$ and $z = D^{-1} r$, we have

$$v^T z = -\begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}^T \begin{bmatrix} A_{21}^T x_2 \\ A_{21} x_1 \end{bmatrix} = x_1^T A_{21}^T x_2 - x_2^T A_{21} x_1 = 0;$$

that is, z is orthogonal to the null space of N^T. Thus z lies in the range of N and $N\tilde{z} = z$ must have a solution, contrary to our original assumption. ∎

3. Improvements to the bound. While Theorem 2.1 gives a simple expression for an upper bound, it is possible to improve on this result. We mention some of these improvements in this section.

First, we show how to eliminate zero elements on the diagonal of N, since such elements would complicate the algorithms for computing $\tilde{\epsilon}_{\text{sym}}$ that are presented in section 4. Suppose that $N_{ii} = 0$ for some i. Since N is diagonally dominant, we must have $N_{ij} = 0$ for all j, so that $(Ny)_i = 0$ for any $y \in \mathbb{R}^n$. Since the nonzero structure of N is symmetric, we must have $N_{ji} = 0$ for all j; and since $N\tilde{z} = z$ has a solution, we also have $z_i = (N\tilde{z})_i = 0$. Thus the linear system $N\tilde{z} = z$ does not place any constraints on $\tilde{z}_i$. Changing N_{ii} from 0 to 1 leaves N diagonally dominant with nonnegative diagonal entries but imposes the added constraint $\tilde{z}_i = 0$. However, setting $\widetilde{z}_i = 0$ in any solution to the original system yields a solution to both the original and the new systems that does not increase $\|\tilde{z}\|_\infty$. After eliminating all zero diagonal entries in this manner, we get the following result.

COROLLARY 3.1. *In the notation of Theorem* 2.1, *let the matrix* $\bar{N}$ *be defined by*

$$\bar{N}_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } N_{ii} = 0, \\ N_{ij} & \text{otherwise.} \end{cases}$$

Then $\bar{N}$ *is diagonally dominant with positive diagonal entries,* $\bar{N}\tilde{z} = z$ *is consistent, and* $\epsilon^*_{\text{sym}} \le \|\tilde{z}\|_\infty$ *for any solution* $\tilde{z}$.

It is also possible to improve our bound by taking advantage of zeros in x, b, and the diagonal of A.

If any entries in x are zero, then without loss of generality we can write A, b, and x as

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix},$$

where $|x_1| > 0$, and any symmetric perturbation ΔA and Δb as

$$\Delta A = \begin{bmatrix} \Delta A_{11} & \Delta A_{21}^T \\ \Delta A_{21} & \Delta A_{22} \end{bmatrix}, \qquad \Delta b = \begin{bmatrix} \Delta b_1 \\ \Delta b_2 \end{bmatrix}.$$

To satisfy $(A + \Delta A)x = b + \Delta b$ this perturbation must satisfy

$$(A_{11} + \Delta A_{11})x_1 = b_1 + \Delta b_1, \qquad (A_{21} + \Delta A_{21})x_1 = b_2 + \Delta b_2.$$

But replacing ΔA_{21} and Δb_2 by the Oettli–Prager perturbation for $A_{21}x_1 = b_2$ and replacing ΔA_{22} by the zero matrix gives another perturbation that satisfies these requirements without increasing ϵ. Indeed, the new minimum ϵ can be nearly a factor of 2 smaller (see (21)).

If corresponding entries on the diagonal of A and in b are 0, then the second term $\max\{\tilde{\epsilon}_A, \tilde{\epsilon}_b\}$ in (12) can be as much as $2n - 1$ times smaller than $\tilde{\epsilon}_{\text{sym}}$ (see (22)).

If a diagonal entry of A is 0 and the corresponding entry of b is not, then $\tilde{\epsilon}_A$ can be smaller than $\tilde{\epsilon}_b$. More generally, whenever $\tilde{\epsilon}_A < \tilde{\epsilon}_b$ there is a gap between ϵ^*_{sym} and $\max\{\tilde{\epsilon}_A, \tilde{\epsilon}_b\}$ that can be reduced by the following procedure. Suppose that we only allow perturbations in A as in problem (7b). Analogously to Theorem 2.1, let

$$\widehat{Z} = \operatorname{diag}(\hat{z}), \qquad \widehat{\Delta A} = \tfrac{1}{2}\left(\widehat{Z}|A|S + S|A|\widehat{Z}\right), \qquad \hat{\epsilon}_A = \max_{i,j} \frac{|\widehat{\Delta A}_{ij}|}{|A_{ij}|},$$

where $\hat{z}$ is any vector such that

$$\tfrac{1}{2}D^{-1}\big(\operatorname{diag}(|A||x|) + S|A|S\operatorname{diag}(|x|)\big)\hat{z} = z.$$

(Such a $\hat{z}$ will exist if $|b_i| = 0$ whenever $(|A||x|)_i = 0$; the proof is similar to that of Theorem 2.1.) Then $(A+\widehat{\Delta A})x = b$ and $\widehat{\Delta A}$ is symmetric, so that $\hat{\epsilon}_A$ is also an upper bound on $\epsilon^*_{\mathrm{sym}}$. Now consider the blended perturbations

$$\Delta A^{(\beta)} = \beta\widetilde{\Delta A} + (1-\beta)\widehat{\Delta A}, \qquad \Delta b^{(\beta)} = \beta\widetilde{\Delta b},$$

for $0 \le \beta < 1$. We still have

$$(A + \Delta A^{(\beta)})x = b + \Delta b^{(\beta)}, \qquad \Delta A^{(\beta)} = (\Delta A^{(\beta)})^T,$$

while

$$\max_{i,j}\frac{|\Delta A^{(\beta)}_{ij}|}{|A_{ij}|} \le \beta\tilde{\epsilon}_A + (1-\beta)\hat{\epsilon}_A, \qquad \max_j \frac{|\Delta b^{(\beta)}_j|}{|b_j|} \le \beta\tilde{\epsilon}_b.$$

Since $\tilde{\epsilon}_A < \tilde{\epsilon}_b$,

$$\epsilon^*_{\mathrm{sym}} \le \min_{0\le\beta<1}\max\left\{\beta\tilde{\epsilon}_A + (1-\beta)\hat{\epsilon}_A,\, \beta\tilde{\epsilon}_b\right\} = \begin{cases}\hat{\epsilon}_A & \text{if } \hat{\epsilon}_A < \tilde{\epsilon}_A,\\ \beta_{\min}\tilde{\epsilon}_b & \text{otherwise,}\end{cases} \tag{15}$$

where $\beta_{\min} = \hat{\epsilon}_A/(\hat{\epsilon}_A + (\tilde{\epsilon}_b - \tilde{\epsilon}_A)) < 1$. This bound is smaller than $\tilde{\epsilon}_b = \max\{\tilde{\epsilon}_A, \tilde{\epsilon}_b\}$ and $\tilde{\epsilon}_{\mathrm{sym}}$, the bounds from Theorem 2.1 (see (23)).

In the tests performed in section 5.3, the bound $\tilde{\epsilon}_{\mathrm{sym}}$ almost always gives a good order of magnitude estimate of $\epsilon^*_{\mathrm{sym}}$, so that these improvements are usually unneeded. However, the blending procedure may provide a much tighter bound in some particular examples (see (23)).

4. Efficient computation of the bound. The perturbations $\widetilde{\Delta A}$ and $\widetilde{\Delta b}$ in (10) and the upper bound $\tilde{\epsilon}_{\mathrm{sym}}$ in (12) for the symmetric componentwise backward error are based on a solution $\tilde{z}$ of the linear system $N\tilde{z} = z$ with N given in (9) and z in (3). In principle one can explicitly compute a $\tilde{z}$ by solving the linear system directly, taking advantage of the fact that no pivoting is required for stability since N is diagonally dominant. However, this is not always practical. Furthermore, since N is based on $|A|$ (as opposed to A), even if exact or inexact triangular factors of A are available, it is not clear how to reuse them to obtain a cheap estimate of $\tilde{z}$.

For large problems it is typically much more efficient to compute $\tilde{z}$ using an iterative procedure. In practice one does not usually require the backward error and corresponding perturbations with a great deal of accuracy; often only an order-of-magnitude estimate is sufficient. This makes the use of iterative strategies particularly attractive. Below we outline two such strategies that we have found to be quite efficient in our numerical experiments. To simplify the presentation we assume that $N_{ii} > 0$ for all i. If this is not the case, one can simply replace N by $\bar{N}$ (see Corollary 3.1).

4.1. Gauss–Seidel. Write $N = E + L + U$, where E is diagonal with positive diagonal entries and L and U are strictly lower and upper triangular, respectively. To solve $N\tilde{z} = z$ the Gauss–Seidel method with initial guess 0 computes the sequence of iterates

$$\tilde{z}^{(k)} = (E+L)^{-1}\left(z - U\tilde{z}^{(k-1)}\right), \quad k > 0; \qquad \tilde{z}^{(0)} = 0. \tag{16}$$

Let $\tilde{\epsilon}_{\rm sym}^{(k)} = \|\tilde{z}^{(k)}\|_\infty$. Since

$$\tilde{z}^{(k)} - \tilde{z} = -(E+L)^{-1}U\left(\tilde{z}^{(k-1)} - \tilde{z}\right), \quad k > 0; \qquad \tilde{z}^{(0)} - \tilde{z} = -\tilde{z},$$

we have

$$\left|\tilde{\epsilon}_{\rm sym}^{(k)} - \tilde{\epsilon}_{\rm sym}\right| = \left|\left\|\tilde{z}^{(k)}\right\|_\infty - \|\tilde{z}\|_\infty\right| \le \left\|\tilde{z}^{(k)} - \tilde{z}\right\|_\infty \le \left\|\left((E+L)^{-1}U\right)^k\right\|_\infty \tilde{\epsilon}_{\rm sym}.$$

Since E is diagonal with positive diagonal entries,

$$\left\|\left((E+L)^{-1}U\right)^k\right\|_\infty \le \left\|\left((E-|L|)^{-1}|U|\right)^k\right\|_\infty = \left\|\left((E-|L|)^{-1}|U|\right)^k e\right\|_\infty \equiv \alpha^{(k)},$$

where $e \in \mathbb{R}^n$ is the vector whose entries are all 1. Thus

$$\left(1+\alpha^{(k)}\right)^{-1} \tilde{\epsilon}_{\rm sym}^{(k)} \le \tilde{\epsilon}_{\rm sym} \le \left(1-\alpha^{(k)}\right)^{-1} \tilde{\epsilon}_{\rm sym}^{(k)}. \tag{17}$$

The iteration can be stopped when the upper and lower bounds are within a small factor, say, 2, of each other, corresponding to $\alpha^{(k)} \le 1/3$. Note that $\alpha^{(k)}$ can be computed as $\alpha^{(k)} = \|\tilde{q}^{(k)}\|_\infty$, where

$$\tilde{q}^{(k)} = (E-|L|)^{-1}|U|\tilde{q}^{(k-1)}, \quad k > 0; \qquad \tilde{q}^{(0)} = e$$

are the Gauss–Seidel iterates for the linear system $(E-|L|-|U|)\tilde{q} = 0$ with initial guess e. This doubles the cost per iteration.

4.2. Convergence of Gauss–Seidel. To show that Gauss–Seidel converges even when N is singular, we need a special case of Keller [7, Corollary 2.1].

THEOREM 4.1. *Let T be a symmetric nonnegative definite matrix with positive diagonal entries. Then the Gauss–Seidel iterates $u^{(k)}$ for the linear system $Tu = v$ converge for any initial guess $u^{(0)}$. Moreover, if u satisfies $Tu = v$ and the initial error $u^{(0)} - u$ is orthogonal to the null space of T, then $u^{(k)}$ converges to u.*

Since N need not be symmetric, we cannot apply this result directly. However, let $P = \operatorname{diag}(p)$, where

$$p_i = \begin{cases} \sqrt{|x_i|d_i} & \text{if } x_i \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Then PNP^{-1} is symmetric and has the same positive diagonal entries as N. Moreover, it has the same eigenvalues as N (by similarity), and these eigenvalues are real (by symmetry) and lie in the right half plane (by Gershgorin's theorem applied to N). That is, PNP^{-1} is symmetric nonnegative definite with positive diagonal entries.

Let $T = P^2N$. Since $T = P(PNP^{-1})P$ and P is diagonal, T is symmetric with positive diagonal entries, and by Sylvester's theorem it is nonnegative definite. Since Gauss–Seidel is invariant under row scaling, the iterates for $T\tilde{z} = P^2z$ with initial guess 0 are the same as those for $N\tilde{z} = z$. Finally, since P is nonsingular, the null spaces of T and N are the same.

By Theorem 2.1 the linear system $N\tilde{z} = z$ is consistent. Let $\bar{z}$ denote the projection of some solution $\tilde{z}$ onto the orthogonal complement of the null space of N. Then $N\bar{z} = z$ so that $T\bar{z} = P^2z$. Since the initial error $0 - \bar{z}$ is orthogonal to the null space of T, by Theorem 4.1 the Gauss–Seidel iterates must converge to $\bar{z}$.

A similar argument applied to $(E - |L| - |U|)\tilde{q} = 0$ shows that the Gauss–Seidel iterates $\tilde{q}^{(k)}$ with the initial guess e must converge. If $E - |L| - |U|$ is nonsingular, its null space is $\{0\}$, and the initial error $e - 0$ corresponding to the solution $\tilde{q} = 0$ is orthogonal to it. Thus the iterates converge to 0, as does $\alpha^{(k)}$. However, if $E-|L|-|U|$ is singular, then it can be shown that neither sequence does, so that $\alpha^{(k)}$ does not provide a useful bound.

4.3. GMRES. While the Gauss–Seidel iterates always converge to a solution, they may converge slowly and $\alpha^{(k)}$ need not converge to 0. An alternative is to use GMRES with initial guess 0 to solve the preconditioned system

$$\left(G_L^{-1} N G_R^{-1}\right) y = G_L^{-1} z,$$

where G_L and G_R are nonsingular, and compute $\tilde{z} = G_R^{-1} y$.

Let $\tilde{y}^{(k)}$ denote the kth iterate, let $\tilde{z}^{(k)} = G_R^{-1}\tilde{y}^{(k)}$, and let $\tilde{\epsilon}_{\text{sym}}^{(k)} = \|\tilde{z}^{(k)}\|_\infty$. If N is nonsingular, the error can be bounded as follows:

$$\begin{aligned}
\left|\tilde{\epsilon}_{\text{sym}}^{(k)} - \tilde{\epsilon}_{\text{sym}}\right| &\le \left\|\tilde{z}^{(k)} - \tilde{z}\right\|_\infty \le \left\|G_R^{-1} G_R N^{-1} G_L G_L^{-1} \left(N\tilde{z}^{(k)} - z\right)\right\|_2 \\
&\le \left\|G_R^{-1}\right\|_2 \left\|\left(G_L^{-1} N G_R^{-1}\right)^{-1}\right\|_2 \left\|G_L^{-1}(N\tilde{z}^{(k)} - z)\right\|_2 \\
&\approx \left\|G_R^{-1}\right\|_2 \sigma_{\min}(H_k)^{-1} \left\|G_L^{-1}\left(N\tilde{z}^{(k)} - z\right)\right\|_2,
\end{aligned}$$

where H_k is the $(k+1) \times k$ upper Hessenberg matrix generated by GMRES and the last factor is the norm of the residual of the preconditioned system. Letting

$$\alpha^{(k)} = \left\|G_R^{-1}\right\|_2 \sigma_{\min}(H_k)^{-1} \left\|G_L^{-1}\left(N\tilde{z}^{(k)} - z\right)\right\|_2 \Big/ \left\|\tilde{z}^{(k)}\right\|_\infty,$$

we get

$$\left(1 - \alpha^{(k)}\right) \tilde{\epsilon}_{\text{sym}}^{(k)} \lesssim \tilde{\epsilon}_{\text{sym}} \lesssim \left(1 + \alpha^{(k)}\right) \tilde{\epsilon}_{\text{sym}}^{(k)}. \tag{18}$$

Again the upper bound and lower bound are within a factor of 2 of each other when $\alpha^{(k)} \le 1/3$. Since $\sigma_{\min}(H_k)^{-1}$ is a lower bound on $\|(G_L^{-1} N G_R^{-1})^{-1}\|_2$, this bound is only approximate, unlike that for the Gauss–Seidel iteration. However, the cost of computing it is negligible.

In our tests the best preconditioner was Gauss–Seidel ($G_L = E + L$ and $G_R = I$, while taking advantage of the fact that

$$G_L^{-1} N G_R^{-1} = I + (E + L)^{-1} U$$

to reduce the cost of multiplying by the preconditioned matrix to that of one Gauss–Seidel iteration). Since Gauss–Seidel with initial guess $\tilde{z}^{(k)} = 0$ always converges and GMRES with the Gauss–Seidel preconditioner generates the same Krylov subspaces but chooses iterates that minimize the norm of the residual, the latter combination must also converge to a solution even when N is singular.

As noted in section 4.2, the matrix PNP^{-1} is symmetric nonnegative definite. Thus GMRES with the preconditioner $G_L = P^{-1}$ and $G_R = P$ reduces to the conjugate residual method, which unlike GMRES requires a constant amount of work per iteration. However, in our experiments this advantage was not enough to overcome a slower rate of convergence.

5. Numerical experiments.

5.1. Some illustrative examples. We start by giving a few small examples that illustrate some important points.

Higham and Higham [4] show that allowing perturbations in the right-hand side vector b can reduce the symmetric componentwise backward error. Let

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ \delta \end{bmatrix}, \quad x = \begin{bmatrix} \delta \\ 1 \end{bmatrix}, \tag{19}$$

where $\delta > 0$. If only perturbations in A are allowed, then that minimum is $\delta/(1+\delta)$ in the unsymmetric case (7a) and 1 in the symmetric case (7b). On the other hand, if perturbations in b are allowed, we have $\epsilon^* = \epsilon^*_{\mathrm{sym}} = \delta/(2+\delta)$. Furthermore, $\tilde{\epsilon}_{\mathrm{sym}} = 3\delta/(4+3\delta)$, which is of the same order of magnitude.

In contrast, Rump [12] shows that the symmetric backward error can be arbitrarily larger than the unsymmetric one, even when perturbations to b are allowed. Let

$$A = \begin{bmatrix} \delta & 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & -1 & 1 & 0 & 1 \\ -1 & 0 & 1 & 1 & 0 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ 4 \\ 0 \\ 0 \end{bmatrix}. \tag{20}$$

Then $\epsilon^* = \delta/(2+\delta)$, while $\epsilon^*_{\mathrm{sym}} = 1$. Furthermore, $\tilde{\epsilon}_{\mathrm{sym}} = 1$.

As shown in section 3, we can take advantage of zeros in x. Let

$$A = \begin{bmatrix} 2 & -2 & 1 \\ -2 & 2 & -2 \\ 1 & -2 & 2 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ \delta \end{bmatrix}. \tag{21}$$

Then $\tilde{\epsilon}_{\mathrm{sym}} = 2(1+\delta)/(3+2\delta)$ corresponding to the perturbation

$$\widetilde{\Delta A} = \begin{bmatrix} \frac{2}{5} & \frac{2}{5} & \frac{1+\delta}{3+2\delta} \\ \frac{2}{5} & \frac{2}{5} & \frac{2(1+\delta)}{3+2\delta} \\ \frac{1+\delta}{3+2\delta} & \frac{2(1+\delta)}{3+2\delta} & 0 \end{bmatrix}, \quad \widetilde{\Delta b} = -\begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{\delta(1+\delta)}{3+2\delta} \end{bmatrix}.$$

However, if we replace the third row/column of $\widetilde{\Delta A}$ and the third row of $\widetilde{\Delta b}$ by the Oettli–Prager perturbation for

$$\begin{bmatrix} 1 & -2 & 2 \end{bmatrix} x = \begin{bmatrix} \delta \end{bmatrix},$$

then we get the perturbation

$$\Delta A = \begin{bmatrix} \frac{2}{5} & \frac{2}{5} & \frac{1+\delta}{3+\delta} \\ \frac{2}{5} & \frac{2}{5} & \frac{2(1+\delta)}{3+\delta} \\ \frac{1+\delta}{3+\delta} & \frac{2(1+\delta)}{3+\delta} & 0 \end{bmatrix}, \quad \Delta b = -\begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{\delta(1+\delta)}{3+\delta} \end{bmatrix},$$

which corresponds to $\epsilon^*_{\mathrm{sym}} = (1+\delta)/(3+\delta)$. The latter is smaller than $\tilde{\epsilon}_{\mathrm{sym}}$ by a factor of 2 asymptotically as $\delta \to 0$.

While $\|\tilde{z}\|_\infty$ is easier to compute, $\max\{\tilde{\epsilon}_A, \tilde{\epsilon}_b\}$ can be as much as $2n-1$ times smaller. Let

$$A = \begin{bmatrix} 1 & -1 & & & \\ -1 & 0 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & 0 & (-1)^{n-1} \\ & & & (-1)^{n-1} & 0 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{22}$$

Then $\epsilon^* = \epsilon^*_{\text{sym}} = \max\{\tilde{\epsilon}_A, \tilde{\epsilon}_b\} = 1$, but $\tilde{\epsilon}_{\text{sym}} = 2n-1$.

As shown in section 3, blending can lead to tighter bounds. Let

$$A = \begin{bmatrix} 1 & -1 & & & \\ -1 & 0 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & 0 & (-1)^{n-1} \\ & & & (-1)^{n-1} & (-1)^{n} \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad b = \delta \begin{bmatrix} -1 \\ 1 \\ \vdots \\ (-1)^n \end{bmatrix}. \tag{23}$$

Then $\epsilon^* = \delta/(2+\delta)$, while $\epsilon^*_{\text{sym}} = n\delta/(2+n\delta)$. Thus the ratio $\epsilon^*_{\text{sym}}/\epsilon^*$ tends to n as $\delta \to 0$. Furthermore, for sufficiently small δ we have $\tilde{\epsilon}_{\text{sym}} \approx n^2\delta/4$ if n is even and $\tilde{\epsilon}_{\text{sym}} \approx (n^2+1)\delta/4$ if n is odd. Thus the ratio $\tilde{\epsilon}_{\text{sym}}/\epsilon^*_{\text{sym}}$ tends to roughly $n/2$ as $\delta \to 0$. On the other hand, by using the blending procedure and taking $\beta = 0$ in (15), we have $\hat{\epsilon}_A = n\delta/2$. In other words, in this example,

$$\epsilon^* \ll \epsilon^*_{\text{sym}} \approx \hat{\epsilon}_A \ll \tilde{\epsilon}_{\text{sym}}.$$

5.2. Comparison with known bounds. Smoktunowicz [13, Theorem 2.2] shows that, in our notation

$$\epsilon^*_{\text{sym}} \le \max_i \frac{|r_i|}{(|A||x|)_i}\big(2(n-1)\gamma + 1\big), \qquad \text{where } \gamma = \max_{i,j} \frac{|A_{ij}|}{|A_{ii}|}, \tag{24}$$

and has also suggested that [14]

$$\epsilon^*_{\text{sym}} \le 2\beta\epsilon^*, \qquad \text{where } \beta = \max_i \frac{(|A||x| + |b|)_i}{|b_i|} \tag{25}$$

with the convention that $\alpha/0 = \infty$ if $\alpha \neq 0$ and $0/0 = 0$.

The bound (24) is ∞ when any diagonal entry of A is 0 (unless the entire row/column and right-hand side are also 0). As the tests in the next subsection show, it is not very tight.

The bound (25) is at least as large as $\|\tilde{z}\|_\infty$ for any solution $\tilde{z}$ of the linear system $\bar{N}\tilde{z} = z$ defined in Corollary 3.1. If $(|A||x| + |b|)_i = 0$, then $d_i = 1$ and

$$N_{ii} = \big(\tfrac{1}{2}|A||x| + |b|\big)_i + \tfrac{1}{2}|A_{ii}||x_i| = 0$$

so that $\bar{N}_{ii} = 1$. Since N is diagonally dominant, $\bar{N}_{ij} = N_{ij} = 0$ for all $j \neq i$. Thus we have $(|I - \bar{N}|e)_i = 0$.

If $(|A||x| + |b|)_i > 0$, then $d_i = (|A||x| + |b|)_i$,

$$(I - \bar{N})_{ii} = \frac{1}{d_i}\left(d_i - \big(\tfrac{1}{2}|A||x| + |b|\big)_i - \tfrac{1}{2}|A_{ii}||x_i|\right) = \frac{1}{2d_i}\left((|A||x|)_i - |A_{ii}||x_i|\right),$$

and

$$\sum_{j\neq i}|(I-\bar{N})_{ij}| = \sum_{j\neq i}|N_{ij}| \leq \frac{1}{2d_i}\sum_{j\neq i}|A_{ij}||x_j| = \frac{1}{2d_i}\left((|A||x|)_i - |A_{ii}||x_i|\right).$$

Thus we have

$$(|I-\bar{N}|e)_i \leq \frac{1}{d_i}\left((|A||x|)_i - |A_{ii}||x_i|\right) = 1 - \frac{|A_{ii}||x_i|+|b_i|}{(|A||x|+|b|)_i}.$$

Putting these together,

$$\|I-\bar{N}\|_\infty = \max_i(|I-\bar{N}|e)_i \leq \max_{i\,:\,(|A||x|+|b|)_i>0}\left(1 - \frac{|A_{ii}||x_i|+|b_i|}{(|A||x|+|b|)_i}\right)$$

with the convention that the maximum is 0 if $|A||x|+|b|=0$. If $\|I-\bar{N}\|_\infty < 1$, then $\bar{N}^{-1}$ exists,

$$\|\bar{N}^{-1}\|_\infty \leq \frac{1}{1-\|I-\bar{N}\|_\infty} \leq \max_{i\,:\,(|A||x|+|b|)_i>0}\frac{(|A||x|+|b|)_i}{|A_{ii}||x_i|+|b_i|} \leq \beta,$$

and we have

$$\epsilon^*_{\text{sym}} \leq \tilde{\epsilon}_{\text{sym}} = \|\tilde{z}\|_\infty \leq \|\bar{N}^{-1}\|_\infty\|z\|_\infty \leq \beta\epsilon^*.$$

If $\|I-\bar{N}\|_\infty = 1$, then $|A_{ii}||x_i|+|b_i| = 0$ for some i with $(|A||x|+|b|)_i > 0$, so that the last expression is ∞ and the result still holds.

5.3. Numerical experiments. It is difficult to compare ϵ^*_{sym} and $\tilde{\epsilon}_{\text{sym}}$ on large problems directly because the symmetric backward error ϵ^*_{sym} is too expensive to compute via linear programming. However, recall that

$$\epsilon^* \leq \epsilon^*_{\text{sym}} \leq \tilde{\epsilon}_{\text{sym}},$$

where ϵ^* is the unsymmetric backward error in (1). Both ϵ^* and $\tilde{\epsilon}_{\text{sym}}$ can be computed cheaply. Therefore, we can investigate the ratio $\tilde{\epsilon}_{\text{sym}}/\epsilon^*$, which is an upper bound on both $\tilde{\epsilon}_{\text{sym}}/\epsilon^*_{\text{sym}}$ and $\epsilon^*_{\text{sym}}/\epsilon^*$, to gain insight into the relationship between $\tilde{\epsilon}_{\text{sym}}$ and ϵ^*_{sym}, as well as the relationship between ϵ^*_{sym} and ϵ^*.

We ran tests on the 589 matrices from the University of Florida Sparse Matrix Collection [3] that satisfied the following criteria:

- A is real and symmetric.
- A is not binary (i.e., some A_{ij} is neither 0 nor 1).
- A is not structurally singular (i.e., some permuted diagonal is nonzero).
- $100 \leq n \leq 100000$ and $4 \leq \text{nnz}(A)/n \leq 200$.
- At most 10^{10} flops are needed to factor a symmetric positive definite matrix with the off-diagonal nonzero structure of A.

After choosing x as described below, we created b as follows (in MATLAB notation):

```
[i,j] = find(A);
E     = 10^(-p) * randn(nnz(A),1) ;
E     = sparse(i,j,E,n,n) .* A ;
E     = E + E';
b     = (A+E)*x;
f     = 10^(-p) * randn(n,1) .* b ;
b     = b-f;
```

TABLE 1
Test results for $p = 4$ and $x_i = \sin(i)$.

Matrix	$\epsilon^*/10^{-p}$	$\tilde{\epsilon}_{\rm sym}/\epsilon^*$	$(24)/\epsilon^*_{\rm sym}$	C_{GS}	D_{GS}	C_{GM}	D_{GM}
Mean (median)	2.97	1.47	1×10^{15}	1	3	1	2
Median (median)	3.08	1.43	3×10^{4}	1	3	1	2
Maximum (median)	4.80	2.06	3×10^{17}	2	17	2	5
Median (maximum)	4.21	2.02	4×10^{4}	2	3	2	3
Maximum (maximum)	6.09	2.93	4×10^{17}	9	17	7	7

Thus, up to rounding error, x satisfies $(A + E)x = b + f$, where E and f are random perturbations of A and b that preserve their sparsity structure and E is symmetric. Since the componentwise relative errors $|E_{ij}|/|A_{ij}|$ and $|f_j|/|b_j|$ are of the order of magnitude 10^{-p}, we expect $\epsilon^*_{\rm sym}$ to be of the same order.

The results for $p = 4$ and $x_i = \sin(i)$ are summarized in Table 1, but the results were similar for other values of p. Because b is generated randomly, all values reported are the mean, median, or maximum over the problems of the medians or maxima over 101 runs for each of the problems. The value of $\tilde{\epsilon}_{\rm sym}$ was computed by solving $\bar{N}\tilde{z} = z$ directly. The values listed in column $(24)/\epsilon^*_{\rm sym}$ include only the 415 problems out of 589 in which the bound (24) is finite.

We also computed approximations to $\tilde{\epsilon}_{\rm sym}$ by solving $\bar{N}\tilde{z} = z$ using Gauss–Seidel (GS) and GMRES with the Gauss–Seidel preconditioner (GM) (see section 4). Columns C_{GS} and C_{GM} give the first steps k at which

$$\left|\tilde{\epsilon}_{\rm sym} - \tilde{\epsilon}^{(k)}_{\rm sym}\right| \le \tfrac{1}{10}\,\tilde{\epsilon}_{\rm sym}.$$

Columns D_{GS} and D_{GM} give the number of steps until convergence is *detected* by the practical stopping criterion $\alpha^{(k)} \le 1/3$ (see (17) and (18)). The reported means are rounded to the nearest integer.

The results are remarkable. The computed ratios $\tilde{\epsilon}_{\rm sym}/\epsilon^*$ are close to 1, indicating that in these tests $\epsilon^*_{\rm sym} \approx \tilde{\epsilon}_{\rm sym}$ and $\epsilon^* \approx \epsilon^*_{\rm sym}$. Convergence of Gauss–Seidel and Gauss–Seidel-preconditioned GMRES to an order-of-magnitude estimate of the symmetric backward error is achieved within a few iterations, and our practical stopping criteria are also triggered very quickly, especially for Gauss–Seidel-preconditioned GMRES.

In the above experiments, the choice $x_i = \sin(i)$ ensures that $|b| > 0$. In an attempt to create problems in which $\tilde{\epsilon}_{\rm sym}/\epsilon^* \gg 1$, we also ran tests with sparse vectors b. In MATLAB notation,

```
w = sum(abs(A),2) ./ sum(A~=0,2);
b = w .* randn(n,1) .* (randperm(n)'<=q*n);
```

for different values of the density q. Thus b contains qn randomly positioned and generated entries. For each such b we picked x to be the computed solution of $Ax = b$, computed using an LDL^T factorization of A with scaling and pivoting. (In order to efficiently compute the factorization and reliably compute such x, of the previous 589 matrices we restricted ourselves to testing only those 343 numerically nonsingular matrices of dimension at most 20000.) Finally, we applied sparsity- and symmetry-preserving perturbations to A and b as in the previous set of tests, so that once again up to rounding errors $(A + E)x = b + f$.

The results for this choice of x for $p = 6$ and density $q = 0.1$ are summarized in Table 2. The values listed in column $(24)/\epsilon^*_{\rm sym}$ include only the 221 problems out of 343 in which the bound (24) is finite. In these tests, the ratio $\tilde{\epsilon}_{\rm sym}/\epsilon^*$ is as large

TABLE 2
Test results for $p = 6$ and b sparse.

Matrix	$\epsilon^*/10^{-p}$	$\tilde{\epsilon}_{\rm sym}/\epsilon^*$	$(24)/\epsilon^*_{\rm sym}$	C_{GS}	D_{GS}	C_{GM}	D_{GM}
Mean (median)	3.42	1.85	7×10^{14}	7	32	3	9
Median (median)	3.42	1.75	2×10^{4}	1	4	2	3
Maximum (median)	4.82	8.18	2×10^{17}	99	99	99	99
Median (maximum)	4.75	2.42	2×10^{4}	2	4	2	3
Maximum (maximum)	7.05	18.91	3×10^{17}	99	99	99	99

as 18.91, and more iterations are needed to achieve and detect convergence. (The number of iterations was capped at 98. A value of 99 indicates that the iterations failed to converge to the desired tolerance within 98 iterations.) While not as good as those in Table 1, these results nevertheless support the idea that $\epsilon^*_{\rm sym} \ll \tilde{\epsilon}_{\rm sym}$ and $\epsilon^* \ll \epsilon^*_{\rm sym}$ are very uncommon occurrences that arise from some special structure in A, b, and/or x.

Although we have not been able to fully quantify this observation, we offer the following partial explanation. Suppose that N is strictly diagonally dominant. Then

$$\epsilon^*_{\rm sym} = \|\tilde{z}\|_\infty = \left\|N^{-1}z\right\|_\infty \le \left\|N^{-1}\right\|_\infty \|z\|_\infty \le \left\|N^{-1}\right\|_\infty \epsilon^*.$$

For $\epsilon^*_{\rm sym}$ to be much larger than ϵ^*, $\|N^{-1}\|_\infty$ must be large. But from (14)

$$\left\|N^{-1}\right\|_\infty \le \frac{1}{\min_i |N_{ii}| - \sum_{j\neq i} |N_{ij}|} \le \max_i \frac{(|A||x| + |b|)_i}{|A_{ii}||x_i| + |b_i|} = 1 + \max_i \frac{\sum_{j\neq i} |A_{ij}x_j|}{|A_{ii}||x_i| + |b_i|}.$$

Thus a necessary condition for $\epsilon^* \ll \epsilon^*_{\rm sym}$ is

$$\sum_{j\neq i} |A_{ij}x_j| \gg |A_{ii}||x_i| + |b_i|$$

for some i. Although such examples can be created artificially, they seem to be extremely rare in practice.

6. Concluding remarks. We have given an upper bound on the symmetric componentwise relative backward error that can be computed efficiently and in our numerical experiments is usually of the same order of magnitude as the true symmetric componentwise backward error. Therefore, we believe that our bound is suitable for use in practice. It also provides new insight into the relationship between the symmetric backward error and the unsymmetric one, showing that both are usually of the same order of magnitude. Whether the techniques presented in this note can be extended to solve backward error problems involving structural properties other than symmetry remains to be seen.

Acknowledgments. The authors would like to thank Xiao-Wen Chang, Alicja Smoktunowicz, as well as two anonymous referees for their very helpful comments.

REFERENCES

[1] M. ARIOLI, J. DEMMEL, AND I. DUFF, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] J. BUNCH, W. DEMMEL, AND C. VAN LOAN, *The strong stability of algorithms for solving symmetric linear systems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 494–499.

[3] T. A. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM Trans. Math. Software, 38 (2011).
[4] D. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175, 1992.
[5] N. J. Higham, *A survey of componentwise perturbation theory in numerical linear algebra*, in Mathematics of Computation 1943–1993: A Half Century of Computational Mathematics, W. Gautschi, ed., Proc. Sympos. Appl. Math. 48, AMS, Providence, RI, 1994, pp. 49–77.
[6] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
[7] H. B. Keller, *On the solution of singular and semidefinite linear systems by iteration*, SIAM J. Numer. Anal., 2 (1965), pp. 281–290.
[8] W. Oettli and W. Prager, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409.
[9] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. ACM, 14 (1967), pp. 543–548.
[10] S. M. Rump, *Structured perturbations Part* I*: Normwise distances*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 1–30.
[11] S. M. Rump, *Structured perturbations Part* II*: Componentwise distances*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 31–56.
[12] S. M. Rump, *The componentwise structure and unstructured backward errors can be arbitrarily far apart*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 385–392.
[13] A. Smoktunowicz, *A note on the strong componentwise stability of algorithms for solving symmetric linear systems*, Demonstr. Math., 28 (1995), pp. 443–448.
[14] A. Smoktunowicz, *private communication*, 2015.
[15] D. M. Young, *Iterative Solution of Large Linear Systems*, Dover, Mineola, NY, 2003.