SIAM J. MATRIX ANAL. APPL.
Vol. 13, No. 1, pp. 162–175, January 1992

013

BACKWARD ERROR AND CONDITION OF STRUCTURED LINEAR SYSTEMS*

DESMOND J. HIGHAM† AND NICHOLAS J. HIGHAM‡

Dedicated to Gene Golub on the occasion of his 60th birthday

Abstract. Existing definitions of backward error and condition number for linear systems do not cater to structure in the coefficient matrix, except possibly for sparsity. The definitions are extended so that when the coefficient matrix has structure the perturbed matrix has this structure too. It is shown that when the structure comprises linear dependence on a set of parameters, the structured componentwise backward error is given by the solution of minimal ∞-norm to an underdetermined linear system; an explicit expression for the condition number in this linear case is also obtained. Applications to symmetric matrices, Toeplitz matrices and the least squares problem are discussed and illustrated through numerical examples.

Key words. componentwise backward error, condition number, underdetermined system, symmetric matrix, Toeplitz matrix, least squares problem, augmented system

AMS(MOS) subject classification. 65F99

1. Introduction.

A backward error of an approximate solution y to a square linear system $Ax = b$ is a measure of the smallest perturbations ΔA and Δb such that

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

Backward error has two distinct uses. First, it can be compared with the size of any uncertainty in the data A and b to ascertain whether y solves a problem sufficiently close to the original one. Second, by invoking perturbation results a bound can be obtained on the forward error $y - x$ in terms of the backward error and an appropriate condition number.

Two classes of backward error definition are in current use, corresponding to different ways of measuring the size of the perturbations ΔA and Δb. The most familiar is the *normwise backward error*

$$\eta(y) = \min\{\epsilon : (A + \Delta A)y = b + \Delta b, \quad \|\Delta A\| \le \epsilon\|E\|, \quad \|\Delta b\| \le \epsilon\|f\|\}, \tag{1.1}$$

in which $\|\cdot\|$ denotes any vector norm and the corresponding subordinate matrix norm, and the matrix E and the vector f are arbitrary. Rigal and Gaches [19] derive the explicit expression

$$\eta(y) = \frac{\|r\|}{\|E\|\|y\| + \|f\|}, \tag{1.2}$$

where $r = b - Ay$; they also show that the minimum in (1.1) is achieved by the perturbations

$$\Delta A_{\min} = \frac{\|E\|\|y\|}{\|E\|\,\|y\| + \|f\|} r z^T, \qquad \Delta b_{\min} = -\frac{\|f\|}{\|E\|\,\|y\| + \|f\|} r, \tag{1.3}$$

* Received by the editors September 28, 1990; accepted for publication (in revised form) April 29, 1991.

† Department of Mathematics and Computer Science, University of Dundee, Dundee, DD1 4HN, Scotland (na.dhigham@na-net.ornl.gov). Part of this author's work was done while he was in the Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. The work of this author was supported by the Natural Sciences and Engineering Research Council of Canada and the Information Technology Research Centre of Ontario.

‡ Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (na.nhigham@na-net.ornl.gov). The work of this author was started during his visit to the Department of Computer Science, University of Toronto, Toronto, Ontario, Canada.

where z is a vector dual to y, that is,

$$z^T y = \|z\|_D \|y\| = 1, \quad \text{where } \|z\|_D = \max_{v \neq 0} \frac{|z^T v|}{\|v\|}.$$

For the particular choice $E = A$ and $f = b$, $\eta(y)$ is called the *normwise relative backward error*. The classic backward error analyses of Wilkinson [24], [25] for linear equation solvers provide bounds on the normwise relative backward error of a computed solution.

A more stringent measure of backward error results if the components of the perturbations ΔA and Δb are measured individually, rather than together in a norm. This way we obtain the *componentwise backward error*

$$\omega(y) = \min\{\epsilon : (A + \Delta A)y = b + \Delta b, \ |\Delta A| \leq \epsilon E, \ |\Delta b| \leq \epsilon f\}, \tag{1.4}$$

where $E \geq 0$ and $f \geq 0$ contain arbitrary tolerances, and inequalities between matrices hold componentwise. The current trend of using componentwise error analysis and perturbation theory began with the 1979 paper of Skeel [20]. However, componentwise backward error was introduced and studied much earlier in a 1964 paper of Oettli and Prager [18]. Oettli and Prager obtained the explicit formula

$$\omega(y) = \max_i \frac{|r_i|}{(E|y| + f)_i}, \tag{1.5}$$

in which $\xi/0$ is interpreted as zero if $\xi = 0$ and infinity otherwise. (A short proof of (1.5) is given in [15] and [20].) Perturbations that achieve the minimum in (1.4) are

$$\Delta A_{\min} = D_1 E D_2, \qquad \Delta b_{\min} = -D_1 f, \tag{1.6}$$

where $D_1 = \operatorname{diag}(r_i/(E|y| + f)_i)$ and $D_2 = \operatorname{diag}(\operatorname{sign}(y_i))$.

One reason for the current interest in componentwise backward error is that it provides a more meaningful measure of stability than the normwise version when the elements of A and b vary widely in magnitude. The most common choice of tolerances is $E = |A|$ and $f = |b|$, which yields the *componentwise relative backward error*. For this definition, zeros in A and b force zeros in the corresponding entries of ΔA and Δb in (1.4), and so if $\omega(y)$ is small, then y solves a problem that is relatively close to the original one and has the same sparsity pattern. Another attractive property of the componentwise relative backward error is that it is insensitive to the scaling of the system: if $Ax = b$ is scaled to $(S_1 A S_2)(S_2^{-1} x) = S_1 b$, where S_1 and S_2 are diagonal, and y is scaled to $S_2^{-1} y$, then ω remains unchanged. Recent work that makes use of componentwise backward error includes [1], [2], [13], [15], [16].

There are situations where even the componentwise backward error is not entirely appropriate, because it does not respect any structure (other than sparsity) in A or b. For example, if A is a Toeplitz matrix and $\eta(y)$ and $\omega(y)$ are small, it does not necessarily follow that y solves a nearby Toeplitz system, since ΔA in (1.1) or (1.4) is not required to be a Toeplitz matrix. Indeed, $\Delta A_{\min}$ in (1.3) or (1.6) is clearly not Toeplitz in general. Similar remarks can be made about condition numbers: the standard condition numbers are derived without requiring that perturbations preserve structure, hence they generally exceed the actual condition number for a linear system subject to structured perturbations. See Bunch [4] or Van Dooren [22] for a more detailed discussion of the desirability of preserving matrix structure in definitions of backward error; Van Dooren also discusses structured condition numbers and describes various structured linear algebra problems that arise in signal processing.

In this work, we extend the notions of componentwise backward error and condition number to allow for dependence of the data on a set of parameters. In §2 we define the *structured componentwise backward error* and show how to compute it when the dependence is linear. In §3 we define a *structured condition number* that measures the sensitivity of a linear system to structured perturbations measured componentwise. We derive an explicit expression for the condition number in the case where the parametrization of the data is linear.

In §4 we examine applications involving symmetric matrices, Toeplitz matrices, and the least squares (LS) problem. In particular, we explain why, when perturbations to a symmetric matrix are measured using the 2-norm, it makes little difference to the backward error or condition number whether the perturbations are required to preserve symmetry or not. For most algorithms for solving structured linear systems little is known about the size of the structured backward error of the computed solution. Some insight can be gained by computing the structured backward error in specific instances, as we illustrate with numerical examples in §5. We give some suggestions for further work in §6.

2. Structured componentwise backward error. Consider an approximate solution y to the linear system $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Suppose A belongs to a set $S \subseteq \mathbb{R}^{n\times n}$ whose members depend on t real parameters ($t \le n^2$); we write this dependence as $A = A[p]$ where $p \in \mathbb{R}^t$. We assume that b does not exhibit any such structure, although the analysis below could be modified to allow for structure in b. (An example of a problem where b has structure is the Yule–Walker Toeplitz system [9, p. 184] in which b depends on the same parameters as A.)

Given nonnegative vectors of tolerances $g \in \mathbb{R}^t$ and $f \in \mathbb{R}^n$, we define the *structured componentwise backward error*

$$\mu(y) = \min\{\epsilon : (A+\Delta A)y = b+\Delta b,\ \ A+\Delta A = A[p+\Delta p], \quad |\Delta p| \le \epsilon g,\quad |\Delta b| \le \epsilon f\}. \tag{2.1}$$

This definition differs from that of the componentwise relative backward error in two respects: we require $A + \Delta A \in S$, so that $A + \Delta A$ has the same structure as A, and we measure the size of the perturbation to A using Δp rather than ΔA. If $S = \mathbb{R}^{n\times n}$, then $\mu(y) \equiv \omega(y)$, assuming g and p comprise the elements of E in (1.4) and A, respectively.

The following transformation removes the absolute values from the constraints in the definition of $\mu(y)$ and replaces the inequalities by equalities. Let

$$\Delta p = D_1 v, \qquad \Delta b = D_2 w, \tag{2.2}$$

where $D_1 = \operatorname{diag}(g_i)$, $D_2 = \operatorname{diag}(f_i)$. Then the smallest ϵ satisfying $|\Delta p| \le \epsilon g$ and $|\Delta b| \le \epsilon f$ is $\epsilon = \max\{\|v\|_\infty, \|w\|_\infty\}$, and so

$$\mu(y) = \min\left\{ \left\| \begin{bmatrix} v \\ w \end{bmatrix} \right\|_\infty : (A+\Delta A)y = b+\Delta b,\ \ A+\Delta A = A[p+\Delta p], \right.$$
$$\left. \Delta p = D_1 v,\ \ \Delta b = D_2 w \right\}. \tag{2.3}$$

In general, this equality constrained nonlinear optimization problem has no closed form solution (and it will have no solution at all if the constraints cannot be satisfied). We therefore concentrate on the special case where S is a linear subspace

of $\mathbb{R}^{n\times n}$. Several classes of matrices of interest fall into this category, such as symmetric, Toeplitz, circulant, and Hankel matrices. Note that linearity implies $\Delta A = (A + \Delta A) - A \in S$. Furthermore, if each element of A is equal to a single element of p, that is, $a_{ij} = p_{k_{ij}}$, then we have the equivalence

$$|\Delta p| \le \epsilon g \iff |\Delta A| \le \epsilon E, \tag{2.4}$$

where $e_{ij} = g_{k_{ij}}$; we will use this equivalence below.

Defining $r = b - Ay$, the equation $(A+\Delta A)y = b+\Delta b$ may be written $\Delta A\, y - \Delta b = r$, or more usefully, $y^T \Delta A^T - \Delta b^T = r^T$, which places the variables ΔA to the right of the constant vector y. Applying the vec operator (which stacks the columns of a matrix into one long vector), we obtain

$$(I_n \otimes y^T)\mathrm{vec}(\Delta A^T) - \Delta b = r, \tag{2.5}$$

where $\otimes$ denotes the Kronecker product (see [17, Chap. 12] for properties of the vec operator and the Kronecker product).

By linearity we have

$$\mathrm{vec}(\Delta A^T) = B\Delta p \tag{2.6}$$

for some $B \in \mathbb{R}^{n^2\times t}$, which we assume to be of full rank.

Using (2.6) and (2.2) we can rewrite (2.5) as

$$(I_n \otimes y^T)BD_1 v - D_2 w = r,$$

or, with $Y = I_n \otimes y^T$,

$$[YBD_1,\ -D_2]\begin{bmatrix} v \\ w \end{bmatrix} = r. \tag{2.7}$$

This is an underdetermined system of the form $Cz = r$, with $C \in \mathbb{R}^{n\times(t+n)}$ and we seek the solution of minimal ∞-norm, the minimal value being $\mu(y)$.

Note that in the case where $t = n^2$ and $B = I$, the rows of C are "structurally independent," that is, there is at most one nonzero per column. Our minimization problem breaks into n independent problems of the form: minimize $\|x\|_\infty$ subject to $a^T x = \alpha$ (which has the solution $x = (\alpha/\|a\|_1)\mathrm{sign}(a)$). It is easy to see that we recover the Oettli–Prager formula (1.5).

If C is rank-deficient, then there may be no solution to $Cz = r$, in which case the structured componentwise backward error $\mu(y)$ may be regarded as being infinite. Assume, therefore, that C has full rank. If C^T has the QR factorization

$$C^T = Q\begin{bmatrix} R \\ 0 \end{bmatrix},$$

then $Cz = r$ may be written

$$r = \begin{bmatrix} R^T & 0^T \end{bmatrix} Q^T z \equiv \begin{bmatrix} R^T & 0^T \end{bmatrix}\begin{bmatrix} \overline{z}_1 \\ \overline{z}_2 \end{bmatrix} = R^T \overline{z}_1.$$

Thus $\overline{z}_1 = R^{-T} r$ is uniquely determined, and

$$z = Q\begin{bmatrix} \overline{z}_1 \\ \overline{z}_2 \end{bmatrix}.$$

Choosing $\overline{z}_2$ to minimize $\|z\|_\infty$ is equivalent to solving an overdetermined linear system in the ∞-norm sense, for which several methods are available [23, Chap. 2], [6].

We can obtain an approximation to the desired ∞-norm minimum by minimizing in the 2-norm, which amounts to setting $\overline{z}_2 = 0$ (and which yields $z = C^{+T} r$, where C^+ is the pseudo-inverse of C). In view of the fact that $n^{-1/2}\|x\|_2 \le \|x\|_\infty \le \|x\|_2$ for $x \in \mathbb{R}^n$, it follows that

$$\mu(y) \le \overline{\mu}(y) \equiv \left\| Q \begin{bmatrix} \overline{z}_1 \\ 0 \end{bmatrix} \right\|_\infty \le \sqrt{t+n}\,\mu(y). \tag{2.8}$$

How does the structured componentwise backward error $\mu(y)$ compare with the standard componentwise backward error $\omega(y)$? If we assume that (2.4) is valid and that E and f are the same for both backward errors then, clearly, $\mu(y) \ge \omega(y)$. More interestingly, if there are zeros in E and f, $\mu(y)$ can be infinite when $\omega(y)$ is finite. The reason is that there are more free parameters in the definition of $\omega(y)$ than in that of $\mu(y)$ and zeros in E or f reduce the number of free parameters in both definitions—potentially by enough for there to exist feasible perturbations ΔA and Δb for $\omega(y)$ but not for $\mu(y)$. Indeed, note that zeros in E and f introduce zero columns in C, making C more likely to be rank-deficient.

Two simple examples help to illustrate the points discussed above. Consider the system with

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad y = \begin{bmatrix} \epsilon \\ 1+\epsilon \end{bmatrix},$$

where $\epsilon > 0$, and let $E = |A|$, $f = 0$ in (1.1), (1.4), (2.1) and (2.4). It is easy to check that

$$\eta_\infty(y) = \frac{\epsilon}{1+\epsilon}, \qquad \omega(y) = 1,$$

and for symmetric structure, $\mu(y) = \infty$. If we alter A, b, and y to

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ \epsilon \end{bmatrix}, \quad y = \begin{bmatrix} \epsilon \\ 1 \end{bmatrix},$$

we find that

$$\eta_\infty(y) = \frac{\epsilon}{2\max(1,\epsilon)}, \quad \omega(y) = \frac{\epsilon}{1+\epsilon}, \quad \mu(y) = 1,$$

which shows that even when the structured backward error is finite it can be arbitrarily larger than the normwise and componentwise backward errors.

It is of interest to characterize when C has full rank, as this guarantees that $\mu(y)$ is finite. C is certainly of full rank if f has no zero elements, because then D_2 is nonsingular, but little more can be said about the rank of C in general.

Finally, we note that if C has full rank, then

$$\begin{aligned} \mu(y) \le \overline{\mu}(y) &= \|C^{+T} r\|_\infty \\ &\le \|C^{+T} r\|_2 \le \|C^+\|_2 \|r\|_2 \\ &= \sigma_{\min}(C)^{-1} \|r\|_2, \end{aligned}$$

where $\sigma_{\min}$ denotes the smallest singular value. This inequality will often be a reasonable approximation and so it would be useful to determine the behavior of $\sigma_{\min}(C)$ as B and y vary. Unfortunately, the rectangularity of B and Y makes it difficult to obtain any results in this direction.

3. Structured condition number. With the same notation as in §2, we define the *structured condition number*

$$\text{cond}_\infty(A,x) = \lim_{\epsilon\to 0}\sup\left\{ \frac{\|\Delta x\|_\infty}{\epsilon\|x\|_\infty} : (A+\Delta A)(x+\Delta x) = b+\Delta b, \right.$$

$$\left. A+\Delta A = A[p+\Delta p],\ \ |\Delta p| \le \epsilon g, |\Delta b| \le \epsilon f \right\}. \tag{3.1}$$

This definition employs the same class of perturbations as in the definition of the structured componentwise backward error $\mu(y)$, and so we have the perturbation result, for any y,

$$\frac{\|y-x\|_\infty}{\|x\|_\infty} \le \text{cond}_\infty(A,x)\mu(y) + O(\mu(y)^2).$$

(Strictly, this result requires that $\|\Delta p\|_\infty = O(\epsilon) \Rightarrow \|\Delta A\|_\infty = O(\epsilon)$; otherwise the order term has to be weakened to $o(\mu(y))$.)

An explicit expression for $\text{cond}_\infty(A,x)$ can be derived in the case where A depends linearly on its parameters. For a given ϵ and perturbed system in the definition of $\text{cond}_\infty(A,x)$, we have

$$(A+\Delta A)\Delta x = \Delta b - \Delta A x,$$

which yields

$$\begin{aligned} \Delta x &= A^{-1}(\Delta b - \Delta A x - \Delta A \Delta x) \\ &= A^{-1}\Delta b - A^{-1}\Delta A x + O(\epsilon^2). \end{aligned} \tag{3.2}$$

First, we analyze the term $A^{-1}\Delta A x$. We have

$$\begin{aligned} \Delta A x &= \text{vec}((\Delta A x)^T) \\ &= \text{vec}(x^T \Delta A^T) \\ &= (I_n \otimes x^T)\text{vec}(\Delta A^T) \\ &= XB\Delta p, \end{aligned}$$

where we have used (2.6) and defined $X = I_n \otimes x^T$. Since $|\Delta p| \le \epsilon g$, it follows that

$$|A^{-1}\Delta A x| = |A^{-1}XB\Delta p| \le \epsilon |A^{-1}XB| g.$$

Similarly,

$$|A^{-1}\Delta b| \le \epsilon |A^{-1}| f.$$

Taking norms we obtain

$$\| - A^{-1}\Delta A x + A^{-1}\Delta b\|_\infty \le \epsilon \| \, |A^{-1}XB| g + |A^{-1}| f \, \|_\infty. \tag{3.3}$$

It is easy to see that equality is attainable in (3.3) for suitable choice of Δp and Δb. It follows from (3.2) and (3.3) that

$$\frac{\|\Delta x\|_\infty}{\epsilon\|x\|_\infty} \le \frac{\| \, |A^{-1}XB| g + |A^{-1}| f \, \|_\infty}{\|x\|_\infty} + O(\epsilon)$$

is a sharp bound, and hence

$$\text{cond}_\infty(A,x) = \frac{\| \, |A^{-1}XB|g + |A^{-1}|f \, \|_\infty}{\|x\|_\infty}. \tag{3.4}$$

In the special case where no structure is imposed ($B = I_{n^2}$), this expression can be written in the form

$$\text{cond}'_\infty(A,x) = \frac{\| \, |A^{-1}|E|x| + |A^{-1}|f \, \|_\infty}{\|x\|_\infty}, \tag{3.5}$$

where E is defined in (2.4); this is a generalization of a condition number of Skeel [20], as described in [1], [15]. It is possible for $\text{cond}'_\infty(A,x)$ to exceed $\text{cond}_\infty(A,x)$ by an arbitrary factor. However, if $\| \, |A^{-1}|f \, \|_\infty \approx \| \, |A^{-1}|E|x| \, \|_\infty$ then the two condition numbers will be of similar magnitude.

More convenient to work with than $\text{cond}_\infty(A,x)$ is the quantity

$$\theta_\infty(A,x) = \frac{\|A^{-1}XBD_1\|_\infty + \|A^{-1}D_2\|_\infty}{\|x\|_\infty},$$

where $D_1 = \text{diag}(g_i)$ and $D_2 = \text{diag}(f_i)$. It is easy to show that

$$\tfrac{1}{2}\theta_\infty(A,x) \le \text{cond}_\infty(A,x) \le \theta_\infty(A,x).$$

The quantity $\theta_\infty(A,x)$ can be estimated without explicitly forming the matrices $A^{-1}XBD_1 \in \mathbb{R}^{n\times t}$ and $A^{-1}D_2 \in \mathbb{R}^{n\times n}$ (assuming a factorization of A is available) by using the method of Hager [10] and Higham [12], [14]; this method estimates $\|C\|_\infty$ at the cost of forming a few matrix-vector products Cx and C^Ty.

We also mention two interesting nonlinear structures, those of Vandermonde matrices $V = (\alpha_j^{i-1})$ and Cauchy matrices $H = ((\alpha_i + \beta_j)^{-1})$. In [11], explicit expressions are derived for $\text{cond}_\infty(V,x)$ and $\text{cond}_\infty(V^T,x)$ in the case where $f = 0$ and $g = (1,1,\cdots,1)^T$. In [8] a structured condition number with respect to the inversion of H is derived.

4. Applications. In this section, we look in detail at the structured componentwise backward error and structured condition number for three applications—those involving symmetric matrices, Toeplitz matrices, and the augmented system for a LS problem.

4.1. Symmetric matrices. For the property of symmetry, there are $t = \frac{1}{2}n(n+1)$ parameters in the vector p. It is natural to take these parameters to be the elements in the upper triangle of A, in which case every row of B contains a single nonzero entry equal to one. To illustrate the form of the underdetermined system (2.7), we consider the case $n = 3$. It is easy to derive the system without using B. Also, it is convenient to work with the independent elements of ΔA rather than Δp, and a symmetric matrix of tolerances E rather than g (see (2.4)).

The constraint $\Delta Ay - \Delta b = r$, that is,

$$\begin{bmatrix} \Delta a_{11} & \Delta a_{12} & \Delta a_{13} \\ \Delta a_{12} & \Delta a_{22} & \Delta a_{23} \\ \Delta a_{13} & \Delta a_{23} & \Delta a_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} - \begin{bmatrix} \Delta b_1 \\ \Delta b_2 \\ \Delta b_3 \end{bmatrix} = r,$$

is equivalent to the system

$$\begin{bmatrix} y_1 & y_2 & y_3 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & y_1 & 0 & y_2 & y_3 & 0 & 0 & -1 & 0 \\ 0 & 0 & y_1 & 0 & y_2 & y_3 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \Delta a_{11} \\ \Delta a_{12} \\ \Delta a_{13} \\ \Delta a_{22} \\ \Delta a_{23} \\ \Delta a_{33} \\ \Delta b_1 \\ \Delta b_2 \\ \Delta b_3 \end{bmatrix} = r. \tag{4.1}$$

On using the transformation (2.2) (where $\Delta p = \mathrm{vec}(\Delta A)$), we obtain the underdetermined system (2.7),

$$\begin{bmatrix} y_1 & y_2 & y_3 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & y_1 & 0 & y_2 & y_3 & 0 & 0 & -1 & 0 \\ 0 & 0 & y_1 & 0 & y_2 & y_3 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = r, \tag{4.2}$$

where $D_1 = \mathrm{diag}(e_{11}, e_{12}, e_{13}, e_{22}, e_{23}, e_{33})$ and $D_2 = \mathrm{diag}(f_1, f_2, f_3)$. Note that the $n \times (n^2/2 + 3n/2)$ coefficient matrix C is upper trapezoidal. Solutions to (4.2) are easily obtained by inspection, but in general, none of these solutions will be of near minimal norm.

The special structure of the matrix C enables computation of the QR factorization of C^T in $O(n^3)$ operations, by careful use of Givens rotations.

A structured *normwise* backward error for symmetric matrices has been considered by Bunch, Demmel, and Van Loan [5]. They consider $\eta(y)$ (see (1.1)) with $f \equiv 0$ and show that enforcing symmetry of ΔA when A is symmetric does not increase $\eta(y)$ for the 2-norm, and it increases it by at most a factor $\sqrt{2}$ for the Frobenius norm. No such result holds for componentwise backward errors because, as explained in §2, it is possible for $\mu(y)$ to be infinite when $\omega(y)$ is finite. However, we note that in the special case where E is diagonal, $\mu(y) = \omega(y)$, because the inequality $|\Delta A| \leq \epsilon E$ in (1.4) automatically forces ΔA to be diagonal and hence symmetric.

The result of [5] can be loosely verified using (2.7). If we set all the elements of f and g to 1 (thus $D_1 = I_t$ and $D_2 = I_n$), then $\mu(y)$ differs from $\eta(y)$ for the 2-norm by at most a factor $\sqrt{n}$ when $B = I_{n^2}$. We will assume that $y = e_1$; this entails no loss of generality because an orthogonal transformation

$$(A + \Delta A)y = b + \Delta b \quad \rightarrow \quad Q(A + \Delta A)Q^T \cdot Qy = Q(b + \Delta b)$$

does not change the class of admissible perturbations or the 2-norms of the perturbations, although it does require g to be multiplied by a factor $\sqrt{n}$. Comparing $Y = I \otimes y^T$ with YB, where B corresponds to the symmetry constraint, we find that they differ only in that Y has extra zero columns. Thus imposing symmetry does not affect the norm of the minimum ∞-norm solution to the system (2.7) when B, D_1,

and D_2 are identity matrices, and this confirms the result of [5], to within a factor n. (The factor n is a consequence of switching from using components to 2-norms.)

A very similar argument shows that when D_1 and D_2 are identity matrices the condition number $\mathrm{cond}_2(A,x)$ is the same when symmetry is imposed as when there is no structural constraint ($\mathrm{cond}_2(A,x)$ is defined as in (3.4) but with the 2-norm replacing the ∞-norm). We note that a condition number that respects symmetry has been derived in a different context by Fletcher [7]. Making statistical assumptions about the perturbations to a linear system, Fletcher shows that the expected condition number of a system is changed little by the imposition of symmetry.

To summarize, when perturbations to a symmetric matrix are measured using the 2-norm it makes little difference to the backward error or to the condition number whether symmetry is enforced or not.

4.2. Toeplitz matrices. Recall that $A \in \mathbb{R}^{n\times n}$ is a Toeplitz matrix if there exist scalars $\{a_k\}_{k=1-n}^{n-1}$ such that $a_{ij} = a_{j-i}$, that is,

$$A = \begin{bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ a_{-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_{1-n} & \cdots & a_{-1} & a_0 \end{bmatrix} = \mathrm{Toeplitz}(a_{1-n}, \cdots, a_0, \cdots, a_{n-1}).$$

In computing the "Toeplitz componentwise backward error," we have to distinguish between unsymmetric and symmetric Toeplitz matrices, for which the number of parameters in A is $t = 2n-1$ and $t = n$, respectively. As in the previous section, it is easy to derive the relevant underdetermined system (2.7).

For illustration we again consider the case $n = 3$. It is straightforward to obtain the following analogues of (4.1), where $\Delta A = \mathrm{Toeplitz}(\Delta a_{1-n}, \cdots, \Delta a_0, \cdots, \Delta a_{n-1})$:

$$\text{(4.3)}\quad \text{unsymmetric:} \quad \begin{bmatrix} y_3 & y_2 & y_1 & 0 & 0 & -1 & 0 & 0 \\ 0 & y_3 & y_2 & y_1 & 0 & 0 & -1 & 0 \\ 0 & 0 & y_3 & y_2 & y_1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \Delta a_2 \\ \Delta a_1 \\ \Delta a_0 \\ \Delta a_{-1} \\ \Delta a_{-2} \\ \Delta b_1 \\ \Delta b_2 \\ \Delta b_3 \end{bmatrix} = r,$$

$$\text{(4.4)}\quad \text{symmetric:} \quad \begin{bmatrix} y_3 & y_2 & y_1 & -1 & 0 & 0 \\ 0 & y_3 + y_1 & y_2 & 0 & -1 & 0 \\ y_1 & y_2 & y_3 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \Delta a_2 \\ \Delta a_1 \\ \Delta a_0 \\ \Delta b_1 \\ \Delta b_2 \\ \Delta b_3 \end{bmatrix} = r.$$

Note that the coefficient matrix in (4.3) loses its Toeplitz structure when we carry out the column scaling necessary to reach (2.7) (cf. (4.2)). Since the number of columns of C is $t+n = O(n)$ in both cases, the cost of computing the QR factorization of C is no more than $O(n^3)$ operations.

Note that if we set $f = 0$, then in the symmetric case the system $Cz = r$ reduces to a square system, corresponding to the fact that the number of parameters in ΔA and Δb is the same as the number of equations.

4.3. The augmented system for the least squares problem. Let $A \in \mathbb{R}^{m\times n}$ have full rank n, let $b \in \mathbb{R}^m$, and let y be an approximate solution to the LS problem $\min_x \|Ax-b\|_2$. Suppose we wish to determine the backward error of y, that is, $\eta_{LS}(y)$ or $\omega_{LS}(y)$, defined as $\eta(y)$ in (1.1) or $\omega(y)$ in (1.4), respectively, but with the condition $(A+\Delta A)y = b$ replaced by the requirement that $\|(A+\Delta A)y-(b+\Delta b)\|_2$ is minimal. Obtaining an explicit formula for $\eta_{LS}(y)$ or $\omega_{LS}(y)$, or an effective way of computing these quantities, is an open problem, as discussed in [13] and [21]; here we make some progress on the problem.

Observe that the LS minimizer x satisfies the augmented system

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \tag{4.5}$$

since this is simply a representation of the normal equations. Because this is a square system, the work in §2 can be exploited. The augmented system has a great deal of structure; to reflect this in the structured componentwise backward error $\mu(y)$, it is sufficient to impose symmetry and to take E (see (2.4)) and f of the form

$$E = \begin{bmatrix} 0 & E_A \\ E_A^T & 0 \end{bmatrix}, \qquad f = \begin{bmatrix} f_b \\ 0 \end{bmatrix}.$$

Let us denote this backward error by $\mu_{LS}(r,y)$. The main observation of this section is that $\mu_{LS}(r,y)$ respects the structure of the augmented system (unlike $\beta(r,y)$ below) and can be computed using standard methods (as described for $\mu(y)$ in §2).

A complicating factor is that r in (4.5) is effectively a vector of free parameters, so to obtain $\eta_{LS}(y)$ or $\omega_{LS}(y)$ we have to minimize $\mu_{LS}(r,y)$ over all r. Fortunately, in the applications of interest the naturally arising r is often a good approximation to the minimizer [16].

In [3] and [13] a "pseudo-componentwise backward error" $\beta(r,y)$ was defined for the augmented system in which different perturbations are allowed in the two occurrences of A. This quantity β is simply $\omega(y)$ of (1.4) applied to the augmented system with

$$E = \begin{bmatrix} 0 & |A| \\ |A^T| & 0 \end{bmatrix}, \qquad f = \begin{bmatrix} |b| \\ 0 \end{bmatrix},$$

and so an explicit formula is available for it from (1.5). In [16], β is proved to be small after one step of fixed precision iterative refinement, under suitable assumptions, when a QR factorization is used to solve the LS problem. Hence it is of interest to compare $\mu_{LS}(r,y)$ with $\beta(r,y)$ when $E_A = |A|$ and $f_b = |b|$. Clearly we have $\mu_{LS}(r,y) \geq \beta(r,y)$ because of the additional symmetry constraint in the definition of μ_{LS}. We report some numerical comparisons in the next section.

Finally, we note that it is possible to obtain a first-order approximation to the backward error $\omega_{LS}(y)$ by considering the perturbed normal equations

$$(A+\Delta A)^T(A+\Delta A)y = (A+\Delta A)^T(b+\Delta b).$$

Expanding and dropping the second-order terms $\Delta A^T\Delta A$ and $\Delta A^T\Delta b$, we have

$$A^T\Delta A y + \Delta A^T A y - A^T\Delta b - \Delta A^T b = A^T(b-Ay).$$

With manipulation similar to that in §2, this linearized problem can be reduced to the computation of a minimum ∞-norm solution to an underdetermined system.

5. Numerical experiments. In the three applications discussed in §3 it is more expensive to compute $\mu(y)$ or $\overline{\mu}(y)$ than to solve the original linear equation problem (but it is inexpensive to estimate $\mathrm{cond}_\infty(A,x)$). Thus, unlike the normwise and componentwise backward errors $\eta(y)$ and $\omega(y)$, $\mu(y)$ is not a quantity that we would compute routinely in the course of solving a problem. However, $\mu(y)$ is useful as a computational tool for studying the stability of a numerical algorithm for solving a structured linear equations problem. We describe some numerical experiments involving the applications of §3.

Our experiments were done using MATLAB, which has a unit roundoff $u \approx 2.2\times 10^{-16}$. Computing μ involves finding the minimal ∞-norm solution to the underdetermined system (2.7). To do this we used the QR factorization transformation to an overdetermined system described in §2, and solved this system in the ∞-norm sense using the method of [6]. The cost of the method of [6] when applied to $\min \|Ax-b\|_\infty$ is approximately the cost of solving k weighted LS problems $\min \|B_i^{-1}Ax_i - b_i\|_2$, where k depends mildly on the problem dimensions (typically $k \le 20$), and where B_i is diagonal except for one full column.

In the first test we solved the system $Ax = b$ using Gaussian elimination with partial pivoting (GEPP), where A is the 10×10 Hilbert matrix and $b = (1,1,\cdots,1)^T/3$. For several E and f, we evaluated the backward errors η, ω, and μ for the computed vector $\widehat{x}$, and the condition numbers $\mathrm{cond}_\infty(A,x)$ of (3.4) and $\mathrm{cond}'_\infty(A,x)$ of (3.5); we imposed the constraint of symmetry for μ (symmetry is denoted by S in the first column of Table 5.1). For this matrix GEPP interchanges rows, and so symmetry is lost in the solution process. We know from standard error analysis that $\eta(\widehat{x})$ will be of order u for $E = A$ (assuming there is no undue element growth in the elimination), irrespective of f, and the result of Bunch, Demmel, and Van Loan referred to in §4.1 shows that imposing symmetry of ΔA in (1.1) cannot significantly increase η. Comparing $\mu(\widehat{x})$ and $\omega(\widehat{x})$, and cond_∞ and cond'_∞, in Table 5.1, we see that requiring symmetry also has little or no effect on the componentwise backward error or the componentwise condition number in this example.

The reason why the numbers in the first two rows of Table 5.1 are the same is that $|A||x|$ is large compared with $|b|$ and hence it makes relatively little difference to the formulas (1.2), (1.5), (3.4), (3.5), and the matrix C, whether we take $f = 0$ or $f = |b|$.

In Table 5.2, A is the symmetric part of a 10×10 matrix with elements from the random normal (0,1) distribution and b is the same vector as in the first example. We see that for the computed solution $\widehat{x}$ from GEPP, $\mu(\widehat{x}) = \omega(\widehat{x})$ in each case; this behavior is not uncommon. Our limited experience indicates that for well-conditioned, full symmetric matrices, $\mu(\widehat{x})$ is usually of similar size to $\omega(\widehat{x})$ for the $\widehat{x}$ from GEPP.

The next example involves the symmetric positive definite 10×10 Toeplitz matrix $A = (\rho^{|i-j|})$, with $\rho = 1-3\times 10^{-5}$ and $b = (1,2,\cdots,10)^T/3$. We solved $Ax = b$ using GEPP and the $O(n^2)$ operations Levinson algorithm [9, p. 187]. [1] Tables 5.3 and 5.4 report the μ values obtained on imposing symmetry or Toeplitz structure alone (denoted by S or T in the first column), and both symmetry and Toeplitz structure. In the tables, $\|A\|_M$ denotes $\max_{i,j} |a_{ij}|$. Preserving the symmetric Toeplitz structure raises the backward errors three orders of magnitude. Note also that there is little difference in the backward errors between the two methods. This example shows that even when a method specific to Toeplitz systems is used, the computed solution is not guaranteed to be the solution to a nearby Toeplitz system.

[1] Note that the second plus should be a minus in the expression for α in [9, Algorithm 4.7.2].

TABLE 5.1
$A = Hilbert(10)$, $\kappa_2(A) = 1.60e13$, *GEPP*. ($E_1 = |\mathrm{diag}(A)|$).

	E	f	$\kappa_2(C)$	cond'_∞	cond_∞	$\eta_\infty(\widehat{x})$	$\omega(\widehat{x})$	$\mu(\widehat{x})$
S	$\lvert A\rvert$	$\lvert b\rvert$	2.40e0	3.05e12	3.05e12	1.99e-18	2.15e-17	2.18e-17
S	$\lvert A\rvert$	0	2.40e0	3.05e12	3.05e12	1.99e-18	2.15e-17	2.18e-17
S	0	$\lvert b\rvert$	1.00e0	1.72e6	1.72e6	4.08e-11	4.08e-11	4.08e-11
S	E_1	0	5.17e4	6.63e11	6.63e11	5.82e-18	3.70e-12	3.70e-12

TABLE 5.2
$A = \mathrm{symm}(\mathrm{rand}(10))$, $\kappa_2(A) = 6.24e1$, *GEPP*. ($E_1 = |\mathrm{diag}(A)|$).

	E	f	$\kappa_2(C)$	cond'_∞	cond_∞	$\eta_\infty(\widehat{x})$	$\omega(\widehat{x})$	$\mu(\widehat{x})$
S	$\lvert A\rvert$	$\lvert b\rvert$	3.63e0	5.06e1	4.79e1	3.81e-17	1.26e-16	1.26e-16
S	$\lvert A\rvert$	0	3.81e0	4.77e1	4.49e1	3.87e-17	1.42e-16	1.42e-16
S	0	$\lvert b\rvert$	1.00e0	2.97e0	2.97e0	2.33e-15	2.33e-15	2.33e-15
S	E_1	0	1.35e2	6.50e0	6.50e0	2.11e-16	4.77e-14	4.77e-14

It is perhaps surprising that the increase in the backward error $\mu(\widehat{x})$ between rows "S," "T," and "S,T" in Tables 5.3 and 5.4 is not matched by a decrease in $\mathrm{cond}_\infty(A, x)$. This means, for example, that a smaller forward error bound (equal to condition number times backward error) is obtained in this example if we do *not* utilize the full structure of the problem. Nevertheless, it is not difficult to find examples where $\mathrm{cond}'_\infty(A, x)/\mathrm{cond}_\infty(A, x)$ is large for symmetric Toeplitz structure if we set $f = 0$, which confines perturbations to the coefficient matrix.

We mention that in all the examples reported the approximation $\overline{\mu}$ in (2.8) satisfied $\overline{\mu} \leq 2\mu$.

In a further experiment we repeated some of the numerical tests from [16], which involve fixed precision iterative refinement of the LS problem using a QR factorization. We extended the testing of [16] by evaluating $\mu_{LS}(\widehat{r}, \widehat{y})$ in addition to $\beta(\widehat{r}, \widehat{y})$, where μ_{LS} and β are defined in §4.3 and $\widehat{r}$ and $\widehat{y}$ are the computed residual and LS solution (both after refinement), respectively. For [16, problem PR] (in which A is a 4×3 matrix with widely varying row norms) and problem set H (a parametrized set of problems involving a 6×5 submatrix of the inverse of the Hilbert matrix of order 6), we found that $\mu_{LS}(\widehat{r}, \widehat{y}) \approx \beta(\widehat{r}, \widehat{y}) \approx u$ in every case. Thus we can conclude that *in these examples* $\omega_{LS}(\widehat{y}) \equiv \min_r \mu_{LS}(r, \widehat{y}) \approx u$, that is, the computed solution obtained after iterative refinement is the exact solution to a small componentwise perturbation of the original LS problem.

6. Concluding remarks. The contribution of this work is to extend existing definitions of backward error and condition number in a way appropriate to structured linear systems and to show how these structure-respecting quantities can be computed in the important case of linear structure. Thus we have derived new theoretical and computational tools. Several questions merit further investigation:

(1) Are there any nonlinear structures for which $\mu(y)$ can be computed more efficiently than if it is treated as a general nonlinear optimization problem (for example, for Vandermonde matrices)?

(2) Is it possible to obtain further theoretical bounds on $\mu(y)$ that would help us to understand its behavior?

(3) Standard backward error analysis results for linear system solvers usually ignore structure. Are there problems and algorithms for which a structured backward

TABLE 5.3
A = *Toeplitz* (10), $\kappa_2(A) = 6.50e5$, *GEPP.* ($E_2 = \|A\|_M ee^T$, $f_2 = \|b\|_\infty e$.)

	E	f	$\kappa_2(C)$	cond'_∞	cond_∞	$\eta_\infty(\widehat{x})$	$\omega(\widehat{x})$	$\mu(\widehat{x})$
S	$\|A\|$	$\|b\|$	1.73e0	1.33e5	1.33e5	2.13e-17	1.07e-16	2.13e-16
S	$\|A\|$	0	1.73e0	1.33e5	1.33e5	2.13e-17	1.07e-16	2.13e-16
T	$\|A\|$	$\|b\|$	1.73e0	1.33e5	1.33e5	2.13e-17	1.07e-16	2.13e-16
T	$\|A\|$	0	1.73e0	1.33e5	1.33e5	2.13e-17	1.07e-16	2.13e-16
S,T	$\|A\|$	$\|b\|$	4.28e3	1.33e5	1.33e5	2.13e-17	1.07e-16	3.23e-13
S,T	$\|A\|$	0	6.06e3	1.33e5	1.33e5	2.13e-17	1.07e-16	6.46e-13
S,T	E_2	f_2	2.92e3	1.33e5	1.33e5	2.13e-17	1.07e-16	2.29e-13

TABLE 5.4
A = *Toeplitz* (10), *Levinson algorithm. Condition numbers as in Table* 5.3.

	E	f	$\eta_\infty(\widehat{x})$	$\omega(\widehat{x})$	$\mu(\widehat{x})$
S	$\|A\|$	$\|b\|$	4.07e-17	2.04e-16	2.32e-16
S	$\|A\|$	0	4.07e-17	2.04e-16	2.32e-16
T	$\|A\|$	$\|b\|$	4.07e-17	2.04e-16	2.32e-16
T	$\|A\|$	0	4.07e-17	2.04e-16	2.33e-16
S,T	$\|A\|$	$\|b\|$	4.07e-17	2.04e-16	4.22e-13
S,T	$\|A\|$	0	4.07e-17	2.04e-16	8.45e-13
S,T	$\|A\|_M ee^T$	$\|b\|_\infty e$	4.07e-17	2.03e-16	3.00e-13

error result can be developed? See [22] for further examples of structured problems.

(4) What can be said about the ratio $\text{cond}'_\infty(A,x)/\text{cond}_\infty(A,x)$ for particular structures and choices of tolerances, that is, how much can the imposition of structure change the condition number? We have answered this question in a particular case involving the property of symmetry.

Acknowledgments. We thank Yuying Li for providing us with MATLAB M-files that implement the method of [6]. The second author thanks the Numerical Analysis Group at the University of Toronto, for their hospitality.

REFERENCES

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

[2] M. ARIOLI, I. S. DUFF AND P. P. M. DE RIJK, *On the augmented system approach to sparse least-squares problems*, Numer. Math., 55 (1989), pp. 667–684.

[3] A. BJÖRCK, *Component-wise perturbation analysis and error bounds for linear least squares solutions*, BIT, 31 (1991), pp. 238–244.

[4] J. R. BUNCH, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

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

[6] T. F. COLEMAN AND Y. LI, *A global and quadratically convergent method for linear L_∞ problems*, Tech. Report 90-1121, Department of Computer Science, Cornell University, Ithaca, NY, 1990.

[7] R. FLETCHER, *Expected conditioning*, IMA J. Numer. Anal., 5 (1985), pp. 247–273.

[8] I. GOHBERG AND I. KOLTRACHT, *On the inversion of Cauchy matrices*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proceedings of the International Symposium MTNS-89, Volume III, M.A. Kaashoek, J.H. van Schuppen and A.C.M. Ran, eds., 1990, pp. 381–392.

[9] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Second Edition, Johns Hopkins University Press, Baltimore, MD, 1989.

[10] W. W. HAGER, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.
[11] N. J. HIGHAM, *Error analysis of the Björck–Pereyra algorithms for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.
[12] ———, FORTRAN *codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm* 674), ACM Trans. Math. Software, 14 (1988), pp. 381–396.
[13] ———, *Computing error bounds for regression problems*, in Statistical Analysis of Measurement Error Models and Applications, P. J. Brown and W. A. Fuller, eds., Contemporary Mathematics 112, American Mathematical Society, Providence, RI, 1990, pp. 195–208.
[14] ———, *Experience with a matrix norm estimator*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 804–809.
[15] ———, *How accurate is Gaussian elimination?*, in Numerical Analysis 1989, Proceedings of the 13th Dundee Conference, D. F. Griffiths and G. A. Watson, eds., Pitman Research Notes in Mathematics 228, Longman Scientific and Technical, Harlow, Essex, 1990, pp. 137–154.
[16] ———, *Iterative refinement enhances the stability of QR factorization methods for solving linear equations*, Numerical Analysis Report No. 182, University of Manchester, England, 1990; BIT, to appear.
[17] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Second Edition, Academic Press, New York, 1985.
[18] 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.
[19] J. L. RIGAL AND J. GACHES, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.
[20] R. D. SKEEL, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.
[21] G. W. STEWART, *Research, development, and* LINPACK, in Mathematical Software III, J.R. Rice, ed., Academic Press, New York, 1977, pp. 1–14.
[22] P. M. VAN DOOREN, *Structured linear algebra problems in digital signal processing*, Proceedings of North Atlantic Treaty Organization ASI, Leuven 1988, Series F, Springer-Verlag, Berlin, 1990.
[23] G. A. WATSON, *Approximation Theory and Numerical Methods*, John Wiley, Chichester, 1980.
[24] J. H. WILKINSON, *Rounding errors in algebraic processes*, Notes on Applied Science, No. 32, Her Majesty's Stationery Office, London, 1963.
[25] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.