Componentwise Perturbation Theory for Linear Systems With Multiple Right-Hand Sides

Desmond J. Higham*
Department of Mathematics and Computer Science
University of Dundee
Dundee DD1 4HN, Scotland

and

Nicholas J. Higham†
Department of Mathematics
University of Manchester
Manchester M13 9PL, England

Submitted by Richard A. Brualdi

ABSTRACT

Existing definitions of componentwise backward error and componentwise condition number for linear systems are extended to systems with multiple right-hand sides and to a general class of componentwise measure of perturbations involving Hölder p-norms. It is shown that for a system of order n with r right-hand sides, the componentwise backward error can be computed by finding the minimum p-norm solutions to n underdetermined linear systems, and an explicit expression is obtained in the case $r = 1$. A perturbation bound is derived, and from this the componentwise condition number is obtained to within a multiplicative constant. Applications of the results are discussed to invariant subspace computations, quasi-Newton methods based on multiple secant equations, and an inverse ODE problem.

*Supported by the University of Dundee Research Initiatives Fund. E-mail: na.dhigham@na-net.ornl.gov.

†E-mail: na.nhigham@na-net.ornl.gov.

655 Avenue of the Americas, New York, NY 10010 0024-3795/92/$5.00

1. INTRODUCTION

The concepts of backward error and condition number are widely used in the study of linear systems $Ax = b$, where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. For a given approximate solution y, a backward error measures how much the data A and b have to be perturbed in order for the perturbed system to have y as a solution. A condition number quantifies the worst-case sensitivity of the solution x to small perturbations in the data, for a particular A and b. To make these concepts precise we must specify what kinds of perturbations are allowed and how they are to be measured. A general definition of backward error is

$$\mathrm{be}(y) = \min\{\phi(\Delta A, \Delta b) : (A + \Delta A)y = b + \Delta b\},$$

and the corresponding condition number is

$$\mathrm{cond}(A, x) = \lim_{\epsilon \to 0} \sup \left\{ \frac{\psi(\Delta x)}{\epsilon} : (A + \Delta A)(x + \Delta x) \right.$$
$$\left. = b + \Delta b,\ \phi(\Delta A, \Delta b) \leqslant \epsilon \right\}.$$

Here, ϕ and ψ are normlike functions on $\mathbb{R}^{n \times (n+1)}$ and $\mathbb{R}^n$ respectively; they may involve arbitrary parameters, and they may be infinite even when their arguments have finite entries. The latter property allows ϕ to impose a particular sparsity structure on the perturbations, as we will see below.

Two special cases of these definitions are well known. Let $\|\cdot\|$ denote an arbitrary vector norm and the corresponding subordinate matrix norm, and let the matrix E and vector f be arbitrary. If we take

$$\phi(\Delta A, \Delta b) = \max\left\{ \frac{\|\Delta A\|}{\|E\|}, \frac{\|\Delta b\|}{\|f\|} \right\}, \qquad \psi(\Delta x) = \frac{\|\Delta x\|}{\|x\|},$$

we obtain the *normwise backward error* and *normwise condition number*. Rigal and Gaches [23] and Kovarik [21] show that the normwise backward error is given by the explicit formula

$$\mathrm{be}(y) = \frac{\|b - Ay\|}{\|E\|\,\|y\| + \|f\|}.$$

The normwise condition number can be shown to be

$$\operatorname{cond}(A, x) = \frac{\|A^{-1}\| \, \|f\|}{\|x\|} + \|A^{-1}\| \, \|E\|.$$

If $E = A$ and $f = b$ then $\kappa(A) \leqslant \operatorname{cond}(A, x) \leqslant 2\kappa(A)$, where $\kappa(A) = \|A\| \, \|A^{-1}\|$ is the standard matrix condition number.

The other common choice of ψ and ϕ is $\psi(\Delta x) = \|\Delta x\|_\infty / \|x\|_\infty$ and

$$\phi(\Delta A, \Delta b) = \min\{\epsilon : |\Delta A| \leqslant \epsilon E, |\Delta b| \leqslant \epsilon f\},$$

where E and f are now assumed to have nonnegative entries and the absolute values and inequalities are interpreted componentwise. This yields the *componentwise backward error* and *componentwise condition number*. Oettli and Prager [22] derive the expression

$$\operatorname{be}(y) = \max_i \frac{(|b - Ay|)_i}{(E|y| + f)_i}. \tag{1.1}$$

Here, and throughout, $\xi/0$ is interpreted as zero if $\xi = 0$ and infinity otherwise. The componentwise condition number is given by

$$\operatorname{cond}(A, x) = \frac{\big\| |A^{-1}|(E|x| + f) \big\|_\infty}{\|x\|_\infty}, \tag{1.2}$$

as shown by Skeel [26] for $E = |A|$, $f = |b|$, and in [2, 19] for general E and f.

The purpose of this work is to extend the above normwise and componentwise definitions in two useful ways and to show how the resulting backward errors and condition numbers can be computed. The new aspects are that we treat systems with multiple right-hand sides and we use a general class of componentwise measures of ΔA, Δx, and Δb. These extensions are motivated by some practical applications that we describe in Section 4.

We consider a multiple right-hand-side linear system $AX = B$, where $A \in \mathbb{R}^{n \times n}$ and X, $B \in \mathbb{R}^{n \times r}$. We define the componentwise backward error and condition number by

$$\operatorname{be}_p(Y) = \min\{\phi_p(\Delta A, \Delta B) : (A + \Delta A)Y = B + \Delta B\}, \tag{1.3}$$

$$\operatorname{cond}_p(A,X) = \lim_{\epsilon\to 0}\sup\left\{\frac{\psi_p(\Delta X)}{\epsilon} : (A+\Delta A)(X+\Delta X) = B+\Delta B,\right.$$
$$\left.\phi_p(\Delta A,\Delta B) \leqslant \epsilon\right\}. \quad (1.4)$$

As the subscript p indicates, we restrict our attention to a particular class of functions ϕ_p and ψ_p, namely

$$\phi_p(\Delta A,\Delta B) = \nu_p\Big(\big[\,(\Delta a_{ij}/e_{ij}) \quad (\Delta b_{ij}/f_{ij})\,\big]\Big), \qquad (1.5a)$$

$$\psi_p(\Delta X) = \nu_p\big((\Delta x_{ij}/g_{ij})\big), \qquad (1.5b)$$

where E, F and G have nonnegative elements and ν_p is the Hölder p-norm

$$\nu_p(A) = \Big(\sum_i\sum_j |a_{ij}|^p\Big)^{1/p}, \qquad 1 \leqslant p \leqslant \infty.$$

We use the notation $\nu_p(\cdot)$ to avoid confusion with $\|\cdot\|_p$, which, as usual, denotes the matrix norm subordinate to the vector p-norm.

When $p = \infty$, the p-norm is the "max norm" $\nu_\infty(A) = \max_{i,j}|a_{ij}|$, and when also $r = 1$, the backward error $\mathrm{be}_\infty(Y)$ and the condition number $\operatorname{cond}_\infty(A, X)$ reduce to the usual componentwise backward error and componentwise condition number discussed above. Other instances of the p-norm of practical interest are the Frobenius norm ($p = 2$) and the "sum norm" ($p = 1$).

For most purposes it is sufficient to choose $g_{ij} \equiv \nu_p(X)$, as in the single right-hand side cases described above, but we consider arbitrary weights g_{ij}, since they are easily accommodated in the analysis. We mention that Gohberg and Koltracht [14] define and derive a componentwise condition number for a general map; for the $AX = B$ problem this corresponds to (1.4) with $p = \infty$, $E = |A|$, $F = |B|$, and $G = |X|$. Also, Rohn [24] determines the condition number in (1.4) in the case $r = 1$, $p = \infty$, $E = |A|$, $F = |b|$, and $G = |x|$.

The $n^2 + nr$ parameters in E and F allow a great deal of control in the backward error be_p. For the choice $E = |A|$ and $F = |B|$, be_p has two particularly attractive properties. First, it is invariant under row and column scalings of the form $AX = B \to D_1AD_2\cdot D_2^{-1}X = D_1B$ (D_i diagonal), since if we scale ΔA and ΔB accordingly, then $\phi_p(\Delta A, \Delta B)$ is unchanged. (The componentwise condition number is likewise invariant under row scalings, and also under column scalings if $G = |X|$.) Second, a zero element in A or

B forces a corresponding zero entry in ΔA or ΔB, in order to keep $\phi_p(\Delta A, \Delta B)$ finite, and so perturbations are forced to preserve the sparsity structure of the data.

In the next section we show that $\mathrm{be}_p(Y)$ can be computed by finding the minimum p-norm solutions to n underdetermined systems, and we derive an explicit expression for $\mathrm{be}_p(Y)$ when $r = 1$. A perturbation bound, and explicit bounds for the componentwise condition number $\mathrm{cond}_p(A, X)$, are derived in Section 3. Applications are described in Section 4, and an extension to structured systems is discussed in Section 5.

2. COMPONENTWISE BACKWARD ERROR

In this section we show how to compute the componentwise backward error $\mathrm{be}_p(Y)$.

The constraints in (1.3) can be written $\Delta AY - \Delta B = B - AY \equiv R$, or

$$[\Delta A \quad \Delta B]\begin{bmatrix} Y \\ -I_r \end{bmatrix} = R.$$

Defining $C = [\Delta A \;\; \Delta B]^T \in \mathbb{R}^{(n+r)\times n}$ and $Z = [Y^T \;\; -I_r] \in \mathbb{R}^{r\times(n+r)}$, we have the multiple right-hand-side system $ZC = R^T$, where C is to be determined. We will assume that Z has full rank; if Z is rank-deficient, there may be no solution to $ZC = R^T$ and hence no feasible perturbations ΔA and ΔB in the definition of $\mathrm{be}_p(Y)$, and in this case we regard the backward error as infinite. Let $H = [E \;\; F]^T$ be the matrix of tolerances corresponding to C.

First, we identify a special case where an explicit formula can be obtained for the solution. If $p = 2$, $e_{ij} \equiv \alpha$, and $f_{ij} \equiv \beta$ ($\alpha = \|A\|_F$ and $\beta = \|B\|_F$ being natural choices) then the problem is to minimize $\|\overline{C}\|_F$ subject to $\overline{Z}\overline{C} = R^T$, where

$$\overline{Z} = \begin{bmatrix} \alpha Y^T & -\beta I_r \end{bmatrix}, \qquad \overline{C} = \begin{bmatrix} \frac{1}{\alpha}\Delta A & \frac{1}{\beta}\Delta B \end{bmatrix}^T.$$

The solution is $\overline{C} = \overline{Z}^+ R^T$, where $\overline{Z}^+$ is the pseudoinverse of $\overline{Z}$.

In general, some further manipulation is required. One possibility is to convert the system to the matrix-vector system $(I_n \otimes Z)\mathrm{vec}(C) = \mathrm{vec}(R^T)$, where $\otimes$ denotes the Kronecker product and the vec operator stacks the columns of a matrix into a vector [20, Chapter 4]. If D is the diagonal matrix $\mathrm{diag}(\mathrm{vec}(H))$ and we write $\mathrm{vec}(C) = Dx$, then the problem is to find the minimum p-norm solution to the underdetermined system $(I_n \otimes Z)Dx = \mathrm{vec}(R^T)$. This can be done using standard methods for $p = 1, 2, \infty$ (see below). This approach is expensive, since the coefficient matrix has dimension $nr \times n(n+r)$.

A more efficient alternative is to exploit the property that $\nu_p(A)$ is an increasing function of the norms $\|a_j\|_p$ of the columns of A, so that $\nu_p(A)$ can be minimized by minimizing the norm of each column independently.

Equating jth columns in the system of constraints, we have $Zc_j = r_j$, where c_j and r_j are the jth columns of C and R^T respectively. Defining $D_j = \mathrm{diag}(h_{1j}, \ldots, h_{n+r,j})$ and writing $c_j = D_j x_j$, we have the underdetermined system $(ZD_j)x_j = r_j$, for which we seek the solution of minimum p-norm. We have the following result.

THEOREM 2.1. *In the notation above,*

$$\mathrm{be}_p(Y) = \nu_p([x_1, \ldots, x_n]) = \left\| [\|x_1\|_p, \ldots, \|x_n\|_p]^T \right\|_p,$$

where x_j is the minimum p-norm solution to $(ZD_j)x_j = r_j$, $j = 1, \ldots, n$.

Now we consider how to compute the required minimum p-norm solutions. Consider an underdetermined system $Ax = b$, and assume that A has full rank, which guarantees that the system is consistent. The minimum 2-norm solution is $\bar{x} = A^+b$. This can be computed using a QR factorization

$$A^T = Q\begin{bmatrix} R \\ 0 \end{bmatrix}.$$

We have

$$b = Ax = \begin{bmatrix} R^T & 0 \end{bmatrix} Q^T x \equiv \begin{bmatrix} R^T & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = R^T y_1. \tag{2.1}$$

Thus $y_1 = R^{-T}b$ is uniquely determined, and the minimum 2-norm solution is

$$\bar{x} = Q\begin{bmatrix} y_1 \\ 0 \end{bmatrix}.$$

In general, we require the solution $\bar{x}$ to $Ax = b$ of minimum p-norm. Using (2.1), this solution can be expressed as

$$\bar{x} = Q\begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

where $y_1 = R^{-T}b$ and y_2 minimizes

$$\left\| Q\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right\|_p \equiv \|Q_1 y_1 + Q_2 y_2\|_p.$$

Thus y_2 minimizes the p-norm of the residual of an overdetermined system, and it can therefore be computed by standard methods when $p = 1$ or ∞ [9; 28, Chapters 2, 6]. Alternatively, the minimum ∞-norm solution $\bar{x}$ to the underdetermined system can be computed directly, using methods in [7, 8].

Note that if either E or F has a zero element in the jth row, then $h_{ij} = 0$ for some i, and so $(D_j)_{ii} = 0$. For each such i the column dimension of the system $(ZD_j)x_j = r_j$ can be reduced by one by deleting the ith column of the coefficient matrix and the ith unknown; this reduction is not strictly necessary in theory or practice, but it has the benefit of reducing the computational cost. In any case, since the optimal perturbations ΔA and ΔB are made up from the vectors $\bar{c}_j = D_j \bar{x}_j$, it is easy to see that $e_{ij} = 0 \Rightarrow \Delta a_{ij} = 0$ an d $f_{ij} = 0 \Rightarrow \Delta b_{ij} = 0$, as must be the case in order to achieve a finite backward error.

When $r = 1$, $\text{be}_p(y)$ and the optimal perturbations can be obtained explicitly, as we now show. As in the general case above, for $j = 1, \dots, n$ we wish to minimize $\|x_j\|_p$ subject to $(ZD_j)x_j = r_j$, which we write as $w_j^T x_j = r_j$, where

$$w_j = (ZD_j)^T = D_j\begin{bmatrix} y \\ -1 \end{bmatrix} \in \mathbb{R}^{n+1}.$$

Now the Hölder inequality states that $|y^T x| \leqslant \|y\|_q \|x\|_p$ where $1/p + 1/q = 1$, with equality for $p, q > 1$ when the vectors $(|y_i|^q)$ and $(x_i|^p)$ are linearly dependent and when $\text{sign}(y_i x_i)$ is constant for all i. (If $p = 1$ or $q = 1$, equality is also attainable, as is easily seen.) It follows that

$$\min\left\{\|x_j\|_p : w_j^T x_j = r_j\right\} = \frac{|r_j|}{\|w_j\|_q}.$$

Hence, from Theorem 2.1 we have the following result.

COROLLARY 2.1. *If $r = 1$, then in the notation above,*

$$\mathrm{be}_p(y) = \left\| \left(\frac{r_j}{\left\| D_j \begin{bmatrix} y \\ -1 \end{bmatrix} \right\|_q} \right) \right\|_p . \tag{2.2}$$

Moreover, the minimum in the definition of $\mathrm{be}_p(y)$ *is attained when the jth row of* $[\Delta A, \Delta B]$ *is given by*

$$\frac{r_j}{\left\| D_j \begin{bmatrix} y \\ -1 \end{bmatrix} \right\|_q} \mathrm{dual} \left(D_j \begin{bmatrix} y \\ -1 \end{bmatrix} \right)^T D_j, \qquad j = 1, \dots, n,$$

where $v =$ dual u *denotes that v is any vector of unit p-norm such that* $v^T u = \|v\|_p \|u\|_q$.

When $p = \infty$ we have $q = 1$ and

$$\left\| D_j \begin{bmatrix} y \\ -1 \end{bmatrix} \right\|_1 = \sum_{k=1}^{n} e_{jk} |y_k| + f_j = (E|y| + f)_j, \tag{2.3}$$

and so (2.2) reduces to the Oettli-Prager formula (1.1).

3. COMPONENTWISE CONDITION NUMBER

Consider the perturbed system

$$(A + \Delta A)(X + \Delta X) = B + \Delta B. \tag{3.1}$$

In this section we obtain an almost sharp bound for $\psi_p(\Delta X)$ in terms of $\phi_p(\Delta A, \Delta B)$, where both these quantities are defined in (1.5). From this bound we are able to deduce the condition number $\mathrm{cond}_p(A, X)$ in (1.4), to within a constant factor depending on n.

To motivate the analysis we note that if $V \in \mathbb{R}^{n \times r}$ then

$$\max_j \|v_j\|_p \leqslant \nu_p(V) \leqslant r^{1/p} \max_j \|v_j\|_p, \tag{3.2}$$

where v_j is the jth column of V, there being equality on both sides for $p = \infty$. It follows that if we can obtain a bound for $\max_j \|v_j\|_p$, then we have a corresponding bound for $\nu_p(V)$, and if the former bound is attainable, then the latter bound is attainable to within a factor $r^{1/p}$. Therefore our approach will be to bound $\|v_j\|_p$, where $V = (\Delta x_{ij}/g_{ij})$. This involves analyzing single right-hand side systems only, and is the natural approach in that the sensitivity of $AX = B$ is approximately the same as the worst-case sensitivity of the individual systems $Ax_i = b_i$.

We analyze the perturbed system

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

where x and b represent the jth column of X and B respectively. Since $Ax = b$, we have

$$\begin{aligned} \Delta x &= A^{-1}(\Delta b - \Delta A(x + \Delta x)) \\ &= -A^{-1}[\Delta A \quad \Delta b]\begin{bmatrix} x + \Delta x \\ -1 \end{bmatrix}. \end{aligned} \tag{3.3}$$

Defining

$$c = [\Delta A \quad \Delta b]\begin{bmatrix} x + \Delta x \\ -1 \end{bmatrix}$$

we have

$$\begin{aligned} c_i &= [\Delta a_{i1}, \dots, \Delta a_{in}, \Delta b_i]\begin{bmatrix} x + \Delta x \\ -1 \end{bmatrix} \\ &= u_i^T z_i, \end{aligned}$$

where

$$u_i^T = [\Delta a_{i1}, \dots, \Delta a_{in}, \Delta b_i]D_i^{-1}, \qquad z_i = D_i\begin{bmatrix} x + \Delta x \\ -1 \end{bmatrix},$$

$$D_i = \operatorname{diag}(e_{i1}, \dots, e_{in}, f_{ij}). \tag{3.4}$$

[Although the definitions involve D_i^{-1}, the analysis remains valid when a zero tolerance makes D_i singular: the final perturbation bound contains a factor $\phi_p(\Delta A, \Delta B)$, which is finite if and only if $\Delta a_{ij} = 0$ whenever $e_{ij} = 0$ and similarly for Δb and f.]

An application of the Hölder inequality yields

$$|c_i| \leqslant \|u_i\|_p \|z_i\|_q, \tag{3.5}$$

where $1/p + 1/q = 1$, and so for the whole vector c,

$$|c| \leqslant \operatorname{diag}(\|z_i\|_q)\, w,$$

where $w_i = \|u_i\|_p$. Premultiplying (3.3) by G_j^{-1}, where

$$G_j = \operatorname{diag}(g_{1j}, \dots, g_{nj}),$$

and using the above bound for $|c|$, we obtain

$$|G_j^{-1}\,\Delta x| \leqslant |G_j^{-1}A^{-1}|\,|c| \leqslant |G_j^{-1}A^{-1}| \operatorname{diag}(\|z_i\|_q)\, w. \tag{3.6}$$

Taking norms, using the subordinate matrix norm $\|\cdot\|_p$, we have

$$\|G_j^{-1}\,\Delta x\|_p \leqslant \left\| G_j^{-1}|A^{-1}| \operatorname{diag}(\|z_i\|_q) \right\|_p \|w\|_p. \tag{3.7}$$

At this point it is desirable to remove the dependence on Δx (through z_i) from the right-hand side. To do so we note that

$$\begin{aligned} \|z_i\|_q &= \left\| D_i \begin{bmatrix} x \\ -1 \end{bmatrix} + D_i \begin{bmatrix} \Delta x \\ 0 \end{bmatrix} \right\|_q \\ &\leqslant \theta_i + \|D_i\|_q \|G_j\|_q \|G_j^{-1}\,\Delta x\|_q \\ &\leqslant \theta_i + \|D_i\|_q \|G_j\|_q c_n \|G_j^{-1}\,\Delta x\|_p, \end{aligned}$$

where $c_n = \max\{1, n^{1/q - 1/p}\}$ and

$$\theta_i = \left\| D_i \begin{bmatrix} x \\ -1 \end{bmatrix} \right\|_q, \tag{3.8}$$

and where we have used an inequality between p-norms [13, p. 28]. Hence we have

$$\|G_j^{-1}\,\Delta x\|_p \leqslant \frac{\big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\theta_i)\big\|_p\|w\|_p}{1 - c_n\|G_j\|_q\big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|D_i\|_q)\big\|_p\|w\|_p}. \tag{3.9}$$

Now

$$\|w\|_p^p = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\frac{\Delta a_{ij}}{e_{ij}}\right|^p + \sum_{i=1}^{n}\left|\frac{\Delta b_i}{f_{ij}}\right|^p,$$

and identifying b with the jth column of B, it follows from (3.2) that

$$\frac{1}{r^{1/p}}\phi_p(\Delta A, \Delta B) \leqslant \max_j\|w\|_p \leqslant \phi_p(\Delta A, \Delta B). \tag{3.10}$$

Hence, using (3.9) and (3.10), setting $\epsilon = \phi_p(\Delta A, \Delta B)$, and again using (3.2), we have the desired bound

$$\psi_p(\Delta X) \leqslant r^{1/p}\epsilon\max_j\frac{\big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\theta_i^{(j)})\big\|_p}{\left(1 - c_n\|G_j\|_q\big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|D_i\|_q)\big\|_p\epsilon\right)}, \tag{3.11}$$

where the superscript in $\theta_i^{(j)}$ reminds us that θ_i in (3.8) depends, via D_i, on the jth column of B [and of course the denominator in (3.11) is assumed to be positive].

Now we consider the sharpness of this perturbation bound. In view of our two invocations of (3.2), equality will be attainable in (3.11) to within a factor $r^{2/p}$ (to first order in ϵ) if (3.7) is attainable, so we consider the latter inequality in detail.

First, we consider the extreme p-norms, 1 and ∞. For $p = \infty$, (3.7) is

$$\|G_j^{-1}\,\Delta x\|_\infty \leqslant \big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|z_i\|_1)\big\|_\infty\|w\|_\infty.$$

This bound is sharp. If we choose $u_i = s_i\|u_i\|_\infty \operatorname{sign}(z_i)$, where $s_i = \pm 1$, then there is equality in (3.5), and if we choose the signs s_i appropriately and set $\|u_i\|_\infty \equiv \|w\|_\infty$, then

$$\begin{aligned}\|G_j^{-1}\,\Delta x\|_\infty = \|G_j^{-1}A^{-1}c\|_\infty &= \Big\|G_j^{-1}A^{-1}\operatorname{diag}(\|z_i\|_1)\,(s_i\|u_i\|_\infty)\Big\|_\infty\\ &= \Big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|z_i\|_1)\,\|w\|_\infty e\Big\|_\infty\\ &= \Big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|z_i\|_1)\Big\|_\infty\|w\|_\infty,\end{aligned}$$

where e is the vector of 1's.

For $p = 1$, (3.7) is

$$\|G_j^{-1}\,\Delta x\|_1 \leqslant \Big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|z_i\|_\infty)\Big\|_1\|w\|_1,$$

which, again, is sharp. There is equality in (3.5) when $u_i = \|u_i\|_1 e_k$, where e_k is the kth column of the identity matrix and where the kth element of z_i is one of maximal modulus. With the u_i so chosen, and on further setting $u_i = 0$ for every i except a single value corresponding to a column of $G_j^{-1}A^{-1}\operatorname{diag}(\|z_i\|_\infty)$ of maximal 1-norm, we have

$$\begin{aligned}\|G_j^{-1}\,\Delta x\|_1 = \|G_j^{-1}A^{-1}c\|_1 &= \Big\|G_j^{-1}A^{-1}\operatorname{diag}(\|z_i\|_\infty)\,(\|u_i\|_1)\Big\|_\infty\\ &= \Big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\|z_i\|_\infty)\Big\|_1\|w\|_1.\end{aligned}$$

For $1 < p < \infty$, equality is not attainable in (3.7) in general. However, since the inequality is sharp for $p = 1, \infty$, it follows that equality is attainable for $1 < p < \infty$ to within a factor n^2.

Thus, to summarize, (3.11) is attainable, to first order, to within a constant factor depending on n and r, and so we have the following result.

THEOREM 3.1. *In the notation of this section*

$$\begin{aligned}r^{-1/p}n^{-2}\max_j \Big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\theta_i^{(j)})\Big\|_p &\leqslant \operatorname{cond}_p(A, X)\\ &\leqslant r^{1/p}\max_j\Big\|G_j^{-1}|A^{-1}|\operatorname{diag}(\theta_i^{(j)})\Big\|_p,\end{aligned} \tag{3.12}$$

and for $p = 1, \infty$ the factor n^{-2} in the lower bound can be removed.

Note that, in accord with our convention about division by zero, the condition number is infinite if, for some j, $(G_j)_{ii} = 0$ while $|A^{-1}|\operatorname{diag}(\theta_i^{(j)})$ has a nonzero in the ith row. An infinite condition number means that for some i and j, arbitrarily small feasible perturbations can yield a nonzero $(\Delta X)_{ij}$ when $g_{ij} = 0$.

It is instructive to examine the special case where $r = 1$ and $p = \infty$. Since $r = 1$, there is no dependence on j. We have [cf. (2.3)]

$$\theta_i = (E|x| + f)_i,$$

and so

$$\begin{aligned}\left\| G_j^{-1}|A^{-1}|\operatorname{diag}(\theta_i)\right\|_\infty &= \left\| G_j^{-1}|A^{-1}|\operatorname{diag}(\theta_i)\, e\right\|_\infty \\ &= \left\| G_j^{-1}|A^{-1}|(\theta_i)\right\|_\infty \\ &= \left\| G_j^{-1}|A^{-1}|(E|x| + f)\right\|_\infty .\end{aligned}$$

If $G = |x|$ then

$$\operatorname{cond}_\infty(A, x) = \left\| \operatorname{diag}(|x_i|)^{-1}|A^{-1}|(E|x| + f)\right\|_\infty,$$

which reduces to the condition number determined by Rohn [24] when $E = |A|$ and $f = |b|$. If $g_{ij} \equiv \|x\|_\infty$, then $G_j \equiv \|x\|_\infty^{-1} I$ and we recover the expression for $\operatorname{cond}_\infty(A, x)$ in (1.2).

Finally, we note that when $p = 1$ or ∞, the bounds in (3.12) can be estimated in $O(n^2 r)$ operations without forming A^{-1} if a QR or LU factorization of A is available; this can be done using the method of Hager [15] and Higham [17, 18], which estimates $\|B\|_1$ or $\|B\|_\infty$ by evaluating several matrix-vector products involving B and B^T. The use of this method to estimate a componentwise condition number was first suggested in [2], in connection with the condition number (1.2), and the latter condition number is estimated this way in LAPACK [5].

4. APPLICATIONS

In this section we describe some applications where it is fruitful to examine the backward error and condition number of a multiple right-hand side linear system.

4.1. *Eigensystem Residual Bounds*

If the columns of $X \in \mathbb{R}^{n\times r}$ form a basis for an invariant subspace of $A \in \mathbb{R}^{n\times n}$, then $AX - XM = 0$ for a unique matrix M. If the columns of Y span only an approximate invariant subspace of A, then $R = AY - YM \neq 0$ for any M, but $\|R\|_F$ is minimized when $M = Y^+AY$. One measure of the quality of the approximation Y is its backward error $\mathrm{be}_p(Y)$, where, in the notation of (1.3), $B = YM$. If $\mathrm{be}_p(Y)$ is known, then perturbation results can be used to assess how well the eigenvalues of M approximate those of A. Stewart and Sun [27, p. 176] show that if Y has orthonormal columns and $M = Y^TAY$, then for any unitarily invariant norm, $\|(A + \Delta A)Y - YM\|$ is minimized when $\Delta A = RY^T$. This result intersects with our analysis in Section 2 in the case of the Frobenius norm ($p = 2$) when $E = ee^T$ and $F = 0$ (that is, no perturbations are allowed in the right-hand side). To our knowledge, componentwise perturbation theory for invariant subspaces has not yet been developed, except in the special case when A is symmetric, $r = n$, and the columns of Y are approximate eigenvectors [3, 10, 11]. Development of such a theory would be an interesting topic for future research.

4.2. *Multiple Secant Equations for Nonlinear Systems*

In the standard quasi-Newton methods for solving a nonlinear system $F(x) = 0$, where $F : \mathbb{R}^n \to \mathbb{R}^n$, approximations $A_{k+1} \approx (\partial f_i(x_{k+1})/\partial x_j)$ are computed that satisfy conditions $A_{k+1}s_k = y_k$, where s_k and y_k are known vectors and the subscript denotes the iteration number. More general quasi-Newton methods have been proposed in which A_{k+1} satisfies r secant equations

$$A_{k+1}s_j = y_j, \qquad j = k - r + 1, \dots, k; \tag{4.1}$$

see [4; 25; 12, pp. 190, 192]. The quasi-Newton method philosophy dictates that the freedom in the choice of the A_{k+1} be used by choosing A_{k+1} to minimize $\|A_{k+1} - A_k\|_F$. Hence $A_{k+1} = A_k + \Delta A_k$, where $\|\Delta A_k\|_F$ is minimized subject to the constraints $(A_k + \Delta A_k)S_k = Y_k$, where

$$S_k = [s_{k-r+1}, s_{k-r}, \dots, s_k], \qquad Y_k = [y_{k-r+1}, y_{k-r}, \dots, y_k].$$

This is precisely the problem of determining $\mathrm{be}_p(S_k)$, with $p = 2$, $E = ee^T$, and $F = 0$. The optimal perturbation is $\Delta A_k = (Y_k - A_kS_k)S_k^+$, and it has rank 1, since the first $r - 1$ columns of $Y_k - A_kS_k$ are zero, in view of the conditions (4.1) for A_k.

By choosing the tolerance matrix E suitably we can impose further restrictions on the quasi-Newton updates. For example, consider the sparse update problem [12, Section 11.2]

$$\min \|\Delta A\|_F \quad \text{subject to} \quad (A + \Delta A)s = y, \quad A + \Delta A \in \mathscr{Z},$$

where the last condition states that $A + \Delta A$ has a given sparsity pattern, i.e., it has zeros in specified entries. To obtain the solution via the backward-error results of Section 2 we set $p = 2$, $r = 1$, $F = 0$ and choose $e_{ij} \in \{0, 1\}$ to match the required sparsity pattern. From Corollary 2.1 a solution is the matrix ΔA whose jth row is given by

$$(y - As)_j \frac{s^T D_j^2}{\|D_j s\|_2^2},$$

where $D_j = \operatorname{diag}(e_{j1}, \dots, e_{jn})$. Since $e_{ij} \in \{0, 1\}$ we have $D_j^2 = D_j$, and we can write

$$\Delta A = \operatorname{sparse}\Big(\operatorname{diag}\big(s^T D_j s\big)^{-1} (y - As) s^T\Big),$$

where the operator sparse$(\cdot)$ imposes the required sparsity pattern on its argument by zeroing entries as necessary. This expression for ΔA is the same as the one given in [12, p. 244].

4.3. *Inverse ODE Problem*

Let $X \in \mathbb{R}^{n \times n}$, and let the function $u(t) : \mathbb{R} \to \mathbb{R}^n$ be a solution of the linear, autonomous, constant-coefficient system of ordinary differential equations

$$\frac{d}{dt} u(t) = X u(t), \qquad t \in [0, 1]. \tag{4.2}$$

Suppose we are given values of $u(t)$ at discrete points t_j, $j = 1, 2, \dots, s$, and we wish to recover the matrix X. Allen and Pruess [1] mention various areas of science where this type of inverse ODE problem arises, and they suggest the following algorithm for computing $\hat{X} \approx X$:

1. Using the discrete data $\{u(t_j)\}$, construct a function $\hat{u}(t) \approx u(t)$ (for example a cubic spline approximation).

2. Choose equally spaced points, $s_i = i/n$, $1 \leqslant i \leqslant n$, and form

$$\hat{A} = \left[\int_0^{s_1} \hat{u}(s)\,ds, \dots, \int_0^{s_n} \hat{u}(s)\,ds\right],$$

$$\hat{B} = \left[\hat{u}(s_1) - \hat{u}(0), \dots, \hat{u}(s_n) - \hat{u}(0)\right].$$

3. Solve $\hat{A}^T\hat{X}^T = \hat{B}^T$.

Note that if

$$A = \left[\int_0^{s_1} u(s)\,ds, \dots, \int_0^{s_n} u(s)\,ds\right],$$

$$B = \left[u(s_1) - u(0), \dots, u(s_n) - u(0)\right],$$

then by integrating (4.2) it follows that $A^TX^T = B^T$. Hence, in exact arithmetic, $\hat{X}^T$ solves a "nearby problem" whose perturbations $\hat{A}^T - A^T$ and $\hat{B}^T - B^T$ are introduced in step 2. The perturbations satisfy

$$|\hat{a}_{ij} - a_{ij}| = \left|\int_0^{s_j} [\hat{u}_i(s) - u_i(s)]\,ds\right| \leqslant s_j \max_{[0,\,s_j]} |\hat{u}_i(t) - u_i(t)|,$$

$$|\hat{b}_{ij} - b_{ij}| = \left|[\hat{u}_i(s_j) - u_i(s_j)] - [\hat{u}_i(0) - u_i(0)]\right| \leqslant 2 \max_{[0,\,s_j]} |\hat{u}_i(t) - u_i(t)|.$$

Allen and Pruess majorize these perturbation bounds into bounds on $\|\hat{A} - A\|_F$ and $\|\hat{B} - B\|_F$ involving the expression $\sum_{i=1}^n (\max_{[0,1]} |\hat{u}_i(t) - u_i(t)|)^2$ and use traditional normwise perturbation analysis to give an asymptotic bound for $\|\hat{X} - X\|_F$ as $\max(t_{i+1} - t_i) \to 0$ in the case of cubic splines. Our analysis in Section 3 is relevant to the case where a numerical estimate of $\|\hat{X} - X\|_p$ is required, and individual estimates of $\max_{[0,\,s_j]} |\hat{u}_i(t) - u_i(t)|$ can be computed for each i and j. Here, since we have a different perturbation bound for each element of A^T and B^T, we can choose the tolerances E and F in (1.5) accordingly and invoke the componentwise perturbation bound (3.11).

5. STRUCTURED SYSTEMS

Our results on componentwise backward error and condition for $AX = B$, $X \in \mathbb{R}^{n \times r}$, can be generalized by allowing for structure in A and B other than sparsity—for example symmetry, Hamiltonian structure, or Toeplitz structure. For the case $p = \infty$ and $r = 1$, a structured componentwise backward error and condition number are defined in [16]; for structure comprising linear dependence on a set of parameters it is shown in [16] how to compute the structured backward error, and an explicit expression is derived for the corresponding condition number. It is straightforward, using the approach described here, to extend the results of [16] to multiple right-hand side systems. However, it is important to realize that for multiple right-hand side systems it is not always possible to achieve structured perturbations. To see this, let $A = A^T \in \mathbb{R}^{n\times n}$ and $Y \in \mathbb{R}^{n\times r}$, and consider the symmetric normwise backward error

$$\eta(Y) = \min\left\{\|\Delta A\|_F : (A + \Delta A)Y = B, \Delta A = (\Delta A)^T\right\}, \qquad (5.1)$$

in which only A is perturbed. If $R = B - AY$ and

$$Q^T[Y \quad R] = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \qquad T_{11} \in \mathbb{R}^{r\times r},$$

is a QR factorization, then the constraint $\Delta A Y = R$ transforms to $Q^T \Delta A Q \cdot Q^T Y = Q^T R$, that is,

$$M \begin{bmatrix} T_{11} \\ 0 \end{bmatrix} = \begin{bmatrix} T_{12} \\ T_{22} \end{bmatrix}, \qquad M = Q^T \Delta A Q,$$

and provided that Y has full rank, this implies that

$$M = \begin{bmatrix} T_{12}T_{11}^{-1} & M_{12} \\ T_{22}T_{11}^{-1} & M_{22} \end{bmatrix},$$

where M_{12} and M_{22} are arbitrary. A symmetric M exists if and only if $T_{12}T_{11}^{-1}$ is symmetric, and this condition is equivalent to Y^TR being symmetric. If $r > 1$, this condition will usually not hold, and so there is usually no feasible perturbation ΔA. However, if in this example we allow both A and B to be perturbed, then it is easy to show that feasible perturbations do exist and hence $\eta(Y)$ is finite. When $r = 1$, $\eta(Y)$ in (5.1) is always finite, and moreover, the interesting result holds that $\eta(Y)$ is no more than twice as big as it would be if the symmetry constraint were not present [6, 16, 21].

The problem of obtaining a solution ΔA to (5.1) arises when multiple secant equations are imposed in quasi-Newton methods for optimization, where the symmetry constraint models the symmetry of the Hessian. An excellent discussion of the existence and computation of ΔA is given in [25], together with a technique for perturbing Y to ensure that $Y^T R$ is symmetric.

REFERENCES

1 R. C. Allen and S. A. Pruess, An analysis of an inverse problem in ordinary differential equations, *SIAM J. Sci. Statist. Comput.* 2:176–185 (1981).

2 M. Arioli, J. W. Demmel, and I. S. Duff, Solving sparse linear systems with sparse backward error, *SIAM J. Matrix Anal. Appl.* 10:165–190 (1989).

3 J. L. Barlow and J. W. Demmel, Computing accurate eigensystems of scaled diagonally dominant matrices, *SIAM J. Numer. Anal.* 27:762–791 (1990).

4 J. G. P. Barnes, An algorithm for solving non-linear equations based on the secant method, *Comput. J.* 8:66–72 (1965).

5 C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, and D. C. Sorensen, Provisional Contents, LAPACK Working Note 5, Report ANL-88-38, Mathematics and Computer Science Div., Argonne National Lab., Argonne, Ill., 1988.

6 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.* 10:494–499 (1989).

7 J. A. Cadzow, A finite algorithm for the minimum l_∞ solution to a system of consistent linear equations, *SIAM J. Numer. Anal.* 10:607–617 (1973).

8 J. A. Cadzow, An efficient algorithmic procedure for obtaining a minimum l_∞-norm solution to a system of consistent linear equations, *SIAM J. Numer. Anal.* 11:1151–1165 (1974).

9 T. F. Coleman and Y. Li, A Global and Quadratically-Convergent Method for Linear L_∞ Problems, Technical Report 90-1121, Dept. of Computer Science, Cornell Univ., Ithaca, N.Y., 1990; *SIAM J. Optimization*, to appear.

10 P. Deift, J. W. Demmel, L. C. Li, and C. Tomei, The Bidiagonal Singular Value Decomposition and Hamiltonian Mechanics, LAPACK Working Note 11, Mathematics and Computer Science Div., Argonne National Lab., Argonne, Ill., 1989; *SIAM J. Numer. Anal.*, to appear.

11 J. W. Demmel and K. Veselić, Jacobi's Method Is More Accurate than *QR*, LAPACK Working Note 15, Dept. of Computer Science, Univ. of Tennessee, Knoxville, 1989; *SIAM J. Matrix Anal. Appl.*, to appear.

12 J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1983.

13 N. Gastinel, *Linear Numerical Analysis*, Academic, London, 1970.

14 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, Vol. III (M. A. Kaashoek, J. H. van Schuppen and A. C. M. Ran, Eds.), 1990, pp. 381–392.

15 W. W. Hager, Condition estimates, *SIAM J. Sci. Statist. Comput.* 5:311–316 (1984).

16 D. J. Higham and N. J. Higham, Backward Error and Condition of Structured Linear Systems, Numerical Analysis Report 192, Univ. of Manchester, 1990; *SIAM J. Matrix Anal. Appl.*, to appear.

17 N. J. Higham, FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm 674), *ACM Trans. Math. Software* 14:381–396 (1988).

18 N. J. Higham, Experience with a matrix norm estimator, *SIAM J. Sci. Statist. Comput.* 11:804–809 (1990).

19 N. J. Higham, How accurate is Gaussian elimination?, in *Numerical Analysis* 1989, *Proceedings of the* 13*th Dundee Conference*, Pitman Research Notes in Mathematics 228 (D. F. Griffiths and G. A. Watson, Eds.), Longman Scientific and Technical, 1990, pp. 137–154.

20 R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge U.P., 1991.

21 Z. V. Kovarik, Compatibility of approximate solutions of inaccurate linear equations, *Linear Algebra Appl.* 15:217–225 (1976).

22 W. Oettli and W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, *Numer. Math.* 6:405–409 (1964).

23 J. L. Rigal and J. Gaches, On the compatibility of a given solution with the data of a linear system, *J. Assoc. Comput. Mach.* 14:543–548 (1967).

24 J. Rohn, New condition numbers for matrices and linear systems, *Computing* 41:167–169 (1989).

25 R. B. Schnabel, Quasi-Newton Methods Using Multiple Secant Updates, Report CU-CS-247-83, Dept. of Computer Science, Univ. of Colorado, Boulder, 1983.

26 R. D. Skeel, Scaling for numerical stability in Gaussian elimination, *J. Assoc. Comput. Mach.* 26:494–526 (1979).

27 G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic, London, 1990.

28 G. A. Watson, *Approximation Theory and Numerical Methods*, Wiley, Chichester, 1980.

Received August 1991; final manuscript accepted 31 August 1991