BIT 33 (1993), 124–136.

PERTURBATION THEORY AND BACKWARD ERROR FOR $AX - XB = C$

NICHOLAS J. HIGHAM*

Department of Mathematics, University of Manchester, Manchester, M 13 9PL, England.
email: na.nhigham@na-net.ornl.gov.

Abstract.

Because of the special structure of the equations $AX - XB = C$ the usual relation for linear equations "backward error = relative residual" does not hold, and application of the standard perturbation result for $Ax = b$ yields a perturbation bound involving $\mathrm{sep}(A, B)^{-1}$ that is not always attainable. An expression is derived for the backward error of an approximate solution Y; it shows that the backward error can exceed the relative residual by an arbitrary factor. A sharp perturbation bound is derived and it is shown that the condition number it defines can be arbitrarily smaller than the $\mathrm{sep}(A, B)^{-1}$-based quantity that is usually used to measure sensitivity. For practical error estimation using the residual of a computed solution an "LAPACK-style" bound is shown to be efficiently computable and potentially much smaller than a sep-based bound. A Fortran 77 code has been written that solves the Sylvester equation and computes this bound, making use of LAPACK routines.

AMS (MOS) subject classifications: 65F05, 65G05.

Key words. Sylvester equation, Lyapunov equation, backward error, perturbation bound, condition number, error estimate, LAPACK.

1. Introduction.

The matrix equation

$$AX - XB = C, \tag{1.1}$$

where $A \in C^{m \times m}$, $B \in C^{n \times n}$, and $C \in C^{m \times n}$, arises in various mathematical settings. Linear equations arising from finite difference discretization of a separable elliptic boundary value problem on a rectangular domain can be written in this form, where A and B represent application of a difference operator in the "y" and "x" directions, respectively [26]. The discretized equations are more commonly written in the form

$$(I_n \otimes A - B^T \otimes I_m)\mathrm{vec}(X) = \mathrm{vec}(C), \tag{1.2}$$

---

* Nuffield Science Research Fellow. This work was carried out while the author was a visitor at the Institute for Mathematics and its Applications, University of Minnesota.

Received April 1992. Revised July 1992.

which is equivalent to (1.1). Here, $A \otimes B \equiv (a_{ij}B)$ is a Kronecker product and the vec operator stacks the columns of a matrix into one long vector. (See [21, Ch. 4] for properties of the Kronecker product and the vec operator.) This "big", standard linear system has a coefficient matrix of order mn with very special structure.

The equation (1.1) plays an important role in the eigenproblem. In particular, the equation often has to be solved in algorithms that manipulate a real Schur decomposition. Examples of such algorithms include an algorithm for block diagonalizing a matrix described in [10, sec. 7.6.3], the algorithm used in LAPACK for re-ordering the eigenvalues in the quasi-triangular form [3], and an algorithm for computing real square roots of a real matrix [17]. In the latter two applications $m, n \in \{1, 2\}$, so the system (1.2) has order 1, 2 or 4. Related to (1.1) is the separation of A and B,

$$\text{sep}(A, B) = \min_{X \neq 0} \frac{\|AX - XB\|_F}{\|X\|_F}, \tag{1.3}$$

which is an important tool in measuring invariant subspace sensitivity [10, sec. 7.2.5], [27, 28]. Here, we are using the Frobenius norm, $\|A\|_F = (\sum_{i,j} |a_{ij}|^2)^{1/2}$. It is easy to see that $\text{sep}(A, B) \neq 0$ if and only if (1.1) has a unique solution for each C, or that, equivalently, A and B do not have a common eigenvalue.

Equation (1.1) is known as the Sylvester equation (see [4] for a historical reference that justifies this terminology). The special case with $B = -A^*$ is the Lyapunov equation $AX + XA^* = C$, which has many applications in control theory [14, 20].

The main purposes of this work are to evaluate the backward error of an approximate solution Y to (1.1) and to determine the sensitivity of (1.1) to perturbations in the data. In doing so we necessarily take full account of the structure of the Sylvester equation. Expressions for the backward error and condition number can be obtained from the work in [16], which applies to linear systems $Ax = b$ in which A depends linearly on a set of parameters. However, in the particular case of the Sylvester equation it is easy to derive even simpler expressions directly, and the main contribution of this work is to analyse these expressions and explain their implications.

Backward error measures how much the data A, B and C must be perturbed in order for an approximate solution Y to (1.1) to be the exact solution of the perturbed system. An important point explained in section 3 is that a small value for the residual $R = C - (AY - YB)$ does not imply a small backward error, unlike for a standard linear system $Ax = b$. Although this point may not be widely appreciated, it is not surprising, because in the particular case where $m = n$, $B = 0$ and $C = I$, we have $AX = I$, and it is well-known that an approximate matrix inverse does not necessarily have a small backward error, even if it has a small residual (see [8, 15], for example). In section 2 we derive an explicit expression for the normwise relative backward error of an approximate solution Y, and determine under what conditions it can greatly exceed the relative residual. This analysis answers the open question raised in [5] of whether the Bartels-Stewart method for solving the

Sylvester equation is backwards stable (indeed it answers the same question for any method for solving the Sylvester equation, including the method of Golub, Nash and Van Loan [9]).

In section 4 we give a perturbation result for the Sylvester equation; this yields a condition number that reflects the structure of the problem. We show that this condition number can be arbitrarily smaller than the quantity involving $\mathrm{sep}(A, B)^{-1}$ that has previously been employed in perturbation bounds in the literature. Of particular practical interest is how to obtain, in terms of the residual, a forward error bound for a computed solution $\hat{X}$ to (1.1). We explain in section 5 how to compute efficiently an "LAPACK-style" bound that is potentially much smaller than the usual sep-based bound.

We have written a Fortran 77 subroutine dggsvx that solves the Sylvester equation and, optionally, estimates our suggested forward error bound and $\mathrm{sep}(A, B)$. The subroutine dggsvx makes use of LAPACK routines [1] and is in the style of an LAPACK driver (release 1.0 of LAPACK does not include a driver for the Sylvester equation).

2. Solving $AX - XB = C$.

In this section we briefly review methods for solving the Sylvester equation and examine what can be said about the residual of the computed solution $\hat{X}$. Knowledge of the residual is useful in the following sections.

Bartels and Stewart [5] showed how to solve (1.1) with the aid of Schur decompositions of A and B. Suppose A and B are real and have real Schur decompositions $A = URU^T$, $B = VSV^T$, where U and V are orthogonal and R and S are upper quasi-triangular, that is, block triangular with 1×1 or 2×2 diagonal blocks, and with any 2×2 diagonal blocks having complex conjugate eigenvalues. (If A and B are complex, the triangular Schur form is used and the following discussion is simplified.) Then the equation transforms to $U^TAU \cdot U^TXV - U^TXV \cdot V^TBV = U^TCV$, that is, $RZ - ZS = D$, or equivalently $Pz = d$, where $P = I_n \otimes R - S^T \otimes I_m$, $z = \mathrm{vec}(Z)$ and $d = \mathrm{vec}(D)$.

If R and S are both triangular then so is P, up to row and column permutations. Therefore z can be obtained by back substitution, and standard backward error analysis [10, sec. 3.1] shows that[1]

$$(P + \Delta P)\hat{z} = d, \qquad |\Delta P| \leq c_{m,n} u |P|, \tag{2.1}$$

where $c_{m,n}$ is a modest constant that depends on the dimensions m and n, and u is the

[1] In fact, this result holds only for the usual "with guard digit" model of floating point arithmetic, namely $fl(x \ op \ y) = (x \ op \ y)(1 + \delta)$, $|\delta| \leq u$, $op = *, /, +, -$. If the model is weakened to $fl(x \pm y) = x(1 + \alpha) \pm y(1 + \beta)$, $|\alpha|, |\beta| \leq u$, as is necessary for machines that lack a guard digit, then (2.1) is vitiated by the rounding errors in forming P, but (2.2) is still valid.

unit roundoff. Here, inequalities and absolute values are interpreted componentwise. Thus $|d - P\hat{z}| \leq c_{m,n}u|P|\,|\hat{z}|$, which implies the weaker inequality

$$|D - (R\hat{Z} - \hat{Z}S)| \leq c_{m,n}u(|R|\,|\hat{Z}| + |\hat{Z}|\,|S|). \tag{2.2}$$

If R or S is quasi-triangular then the computation of $\hat{Z}$ involves the solution of systems of dimension 2 or 4 by Gaussian elimination with pivoting. If iterative refinement is used for each of these systems "$\bar{P}\bar{z} = \bar{d}$", and if for each system $\bar{P}$ is not too ill-conditioned and the vector $|\bar{P}|\,|\bar{z}|$ is not too badly scaled, then (2.1) and (2.2) remain valid [25]. Otherwise, we have only a normwise bound.

$$\|D - (R\hat{Z} - \hat{Z}S)\|_F \leq c'_{m,n}u(\|R\|_F + \|S\|_F)\,\|\hat{Z}\|_F.$$

Because the transformation of a matrix to Schur form is a stable process, it is true overall that

$$\|C - (A\hat{X} - \hat{X}B)\|_F \leq c''_{m,n}u(\|A\|_F + \|B\|_F)\,\|\hat{X}\|_F. \tag{2.3}$$

Thus the relative residual is guaranteed to be bounded by a modest multiple of the unit roundoff u, as was noted in [5].

Golub, Nash and Van Loan [9] suggested a modification of the Bartels-Stewart algorithm in which A is reduced only to upper Hessenberg form: $A = UHU^T$. The reduced system $HZ - ZS = D$ can be solved by solving n upper Hessenberg systems. As shown in [9], the Hessenberg-Schur algorithm can be more efficient than the Bartels-Stewart algorithm, depending on the problem dimensions, and the computed solution $\hat{X}$ again satisfies (2.3).

The use of iterative methods to solve (1.1) has attracted attention recently for applications where A and B are large and sparse [22, 26, 29]. The iterations are usually terminated when an inequality of the form (2.3) holds, so here the size of the relative residual is known a priori (assuming the method converges).

3. Backward error.

The normwise backward error of an approximate solution Y to (1.1) is defined by

$$\eta(Y) = \min\{\varepsilon: (A + E)Y - Y(B + F) = C + G,\ \|E\|_F \leq \varepsilon\alpha,\ \|F\|_F \leq \varepsilon\beta,\ \|G\|_F \leq \varepsilon\gamma\}. \tag{3.1}$$

The tolerances α, β and γ provide some freedom in how we measure the perturbations. Of most interest is the choice $\alpha = \|A\|_F$, $\beta = \|B\|_F$, $\gamma = \|C\|_F$, which yields the *normwise relative backward error*. The equation $(A + E)Y - Y(B + F) = C + G$ may be written

$$EY - YF - G = R, \tag{3.2}$$

where the residual $R = C - (AY - YB)$. For a standard linear system $Ax = b$

a small relative residual is equivalent to a small backward error. Specifically, it can be shown [24] that

$$\min\{\varepsilon: (A+E)y = b+f,\ \|E\|_2 \le \varepsilon\alpha,\ \|f\|_2 \le \varepsilon\beta\} = \frac{\|r\|_2}{\alpha\|y\|_2 + \beta}, \tag{3.3}$$

where $\|\cdot\|_2$ denotes the vector 2-norm, $\|x\|_2 = (x^Tx)^{1/2}$, and the corresponding subordinate matrix norm. For the Sylvester equation a small backward error implies a small relative residual since, using the optimal perturbations from (3.1) in (3.2), we have

$$\|R\|_F = \|EY - YF - G\|_F \le ((\alpha+\beta)\|Y\|_F + \gamma)\eta(Y). \tag{3.4}$$

However, the reverse implication does not always hold. To see this we write (3.2) in the form

$$(Y^T \otimes I_m)\mathrm{vec}(E) - (I_n \otimes Y)\mathrm{vec}(F) - \mathrm{vec}(G) = \mathrm{vec}(R),$$

that is,

$$[\alpha(Y^T \otimes I_m),\ -\beta(I_n \otimes Y),\ -\gamma I_{mn}]\begin{bmatrix}\mathrm{vec}(E)/\alpha\\ \mathrm{vec}(F)/\beta\\ \mathrm{vec}(G)/\gamma\end{bmatrix} = \mathrm{vec}(R). \tag{3.5}$$

This is an underdetermined system of the form $Hz = r$, where H is $mn \times (m^2 + n^2 + mn)$, and H is certainly of full rank if $\gamma \neq 0$. There are many solutions to this system, but there is a unique one of minimum 2-norm, given by $z = H^+r$, where H^+ is the pseudo-inverse of H. It follows that

$$(1/\sqrt{3})\|H^+r\|_2 \le \eta(Y) \le \|H^+r\|_2. \tag{3.6}$$

Since $\|H^+r\|_2 \le \|H^+\|_2\|r\|_2$, with equality for suitable r, we see that the maximum size of the backward error relative to the residual is dependent on $\|H^+\|_2$. We now derive an expression for $\|H^+\|_2$. In view of the general formula $\|A^+\|_2 = \sigma_{\min}(A)^{-1}$ for full rank A, where $\sigma_{\min}$ denotes the smallest singular value, our task is to determine the smallest singular value of H.

If Y has the singular value decomposition $Y = U\Sigma V^*$, then H is unitarily equivalent to the matrix

$$\begin{aligned}\tilde{H} &= (V^T \otimes U^*)\cdot H \cdot \mathrm{diag}(\bar{U} \otimes U, \bar{V} \otimes V, \bar{V} \otimes U)\\ &= [\alpha(\Sigma^T \otimes I_m),\ -\beta(I_n \otimes \Sigma),\ -\gamma I_{mn}].\end{aligned} \tag{3.7}$$

Therefore H has the same singular values as $\tilde{H}$, and these are the square roots of the eigenvalues of the diagonal matrix

$$\tilde{H}\tilde{H}^* = \alpha^2(\Sigma^T\Sigma \otimes I_m) + \beta^2(I_n \otimes \Sigma\Sigma^T) + \gamma^2 I_{mn}.$$

It follows that the singular values of H are given by

$$\sigma_{ij} = (\alpha^2\sigma_j^2 + \beta^2\sigma_i^2 + \gamma^2)^{1/2}, \quad 1 \le i \le m, \quad 1 \le j \le n,$$

where $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_{\min(m,n)} \ge 0$ are the singular values of Y and we define

$\sigma_{\min(m,n)+1} = \ldots = \sigma_{\max(m,n)} = 0$. Hence, assuming that H has full rank,

$$\|H^+\|_2 = (\alpha^2\sigma_n^2 + \beta^2\sigma_m^2 + \gamma^2)^{-1/2}.$$

Combining this result with (3.6) we obtain

$$\eta(Y) \le \mu \frac{\|R\|_F}{(\alpha+\beta)\|Y\|_F + \gamma}, \tag{3.8}$$

where

$$\mu = \frac{(\alpha+\beta)\|Y\|_F + \gamma}{(\alpha^2\sigma_n^2 + \beta^2\sigma_m^2 + \gamma^2)^{1/2}}. \tag{3.9}$$

The scalar $\mu \ge 1$ is an amplification factor that measures by how much, at worst, the backward error can exceed the relative residual. We now examine μ more closely, concentrating on the normwise relative backward error, for which $\alpha = \|A\|_F$, $\beta = \|B\|_F$ and $\gamma = \|C\|_F$.

First, note that if $n = 1$ and $B = 0$, so that the Sylvester equation reduces to a linear system $Ay = c$, then $\sigma_1 = \|y\|_2$ and $\sigma_k = 0$ for $k > 1$, so $\mu = (\|A\|_F\|y\|_2 + \|c\|_2)/(\|A\|_F^2\|y\|_2^2 + \|c\|_2^2)^{1/2}$. Since $1 \le \mu \le \sqrt{2}$, we recover the result (3.3) from (3.4) and (3.8), to within a factor $\sqrt{2}$.

If $m = n$ then

$$\mu = \frac{(\|A\|_F + \|B\|_F)\|Y\|_F + \|C\|_F}{((\|A\|_F^2 + \|B\|_F^2)\sigma_{\min}(Y)^2 + \|C\|_F^2)^{1/2}}. \tag{3.10}$$

We see that μ is large only when

$$\|Y\|_F \gg \sigma_{\min}(Y) \text{ and } \|Y\|_F \gg \frac{\|C\|_F}{\|A\|_F + \|B\|_F}, \tag{3.11}$$

that is, when Y is ill-conditioned and Y is a large-normed solution to the Sylvester equation. In the general case, with $m \ne n$, one of σ_m^2 and σ_n^2 is always zero and hence μ can be large for a third reason: A (if $m < n$) or B (if $m > n$) greatly exceeds the rest of the data in norm; in these cases the Sylvester equation is badly scaled. However, if we set $\alpha = \beta = \|A\|_F + \|B\|_F$, which corresponds to regarding A and B as comprising a single set of data, then bad scaling does not affect μ.

If we allow only A and B to be perturbed in (3.1) (as may be desirable if the right-hand side C is known exactly), then $\gamma = 0$ and (3.10) and (3.11) remain valid with $\|C\|_F$ replaced by zero. In this case $\mu \ge \|Y\|_F\|Y^+\|_2 \approx \kappa_2(Y)$ (for any m and n), so μ is large whenever Y is ill-conditioned (and included in this case is matrix inversion). Conditions involving controllability which guarantee that the solution to the Sylvester equation with $m = n$ is nonsingular are given in [12], while in [7] a determinantal condition for nonsingularity is given. It appears to be an open problem to derive conditions for the Sylvester equation to have a well-conditioned solution.

The following numerical example illustrates the above analysis. This particular

example was carefully chosen so that the entries of A and B are of a simple form, but equally effective examples are easily generated using random, ill-conditioned A and B of dimension $m, n \geq 2$. Let

$$A = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad B = A - \alpha \begin{bmatrix} 1+\alpha & 0 \\ 0 & 1 \end{bmatrix}.$$

Define C by the property that $\operatorname{vec}(C)$ is the singular vector corresponding to the smallest singular value of $I_n \otimes A - B^T \otimes I_m$. With $\alpha = 10^{-6}$, we solved the Sylvester equation in Matlab by the Bartels-Stewart algorithm and found that the computed $\hat{X}$ satisfies

$$\frac{\|R\|_F}{(\|A\|_F + \|B\|_F)\|\hat{X}\|_F + \|C\|_F} = 2.82 \times 10^{-17}, \quad \sigma(\hat{X}) = \{2 \times 10^{18}, 5 \times 10^{5}\},$$

$$\eta(\hat{X}) \approx \|H^+ r\|_2 = 2.21 \times 10^{-8}, \quad \mu = 5.66 \times 10^{12}.$$

Matlab has unit roundoff $u \approx 1.1 \times 10^{-16}$, so although $\hat{X}$ has a very acceptable residual (as it must in view of (2.3)), its backward error is eight orders of magnitude larger than is necessary to achieve backward stability. We solved the same Sylvester equation using Gaussian elimination with partial pivoting on the system (1.2). The relative residual was again less than u, but the backward error was appreciably larger: $\eta(\hat{X}) \approx 1.53 \times 10^{-5}$.

The analysis above makes no assumption on the structure of the matrices A and B. If A and B are (quasi-) triangular then one may wish to restrict the perturbations E and F in (3.1) to have the same structure. This requirement can be met by removing those elements of $\operatorname{vec}(E)$ and $\operatorname{vec}(F)$ in (3.5) that correspond to the "zero triangles" of A and B, and deleting the corresponding columns of the matrix H. If H_i denotes H with column i removed then $\sigma_{\min}(H_i) \leq \sigma_{\min}(H)$, so one would expect forcing preservation of triangularity to make the backward error no smaller and potentially much bigger.

For the Lyapunov equation, in which $B = -A^*$, we need to modify the definition (3.1) of backward error so that $F = -E^*$, in order to make a single perturbation to the matrix A. Clearly, the modified backward error is no smaller than (3.1). The analogue of (3.2) is $EY + YE^* - G = R$. Assuming that the data are real this can be written as

$$[\alpha((Y^T \otimes I_n) + (I_n \otimes Y)\Pi^T), \ -\gamma I_{n^2}] \begin{bmatrix} \operatorname{vec}(E)/\alpha \\ \operatorname{vec}(G)/\gamma \end{bmatrix} = \operatorname{vec}(R),$$

where $\operatorname{vec}(E^T) = \Pi^T \operatorname{vec}(E)$, and where Π is a permutation matrix known as the vec-permutation matrix [13]. Unlike for the general Sylvester equation, no explicit formula is available for the norm of the pseudo-inverse of the coefficient matrix. Thus the added structure of the Lyapunov equation makes the backward error much less analytically tractable.

To summarise, the backward error of an approximate solution to the Sylvester

equation can be arbitrarily larger than its relative residual. The key quantity is the amplification factor μ in (3.9), which bounds the ratio of relative residual to backward error.

In [5], Bartels and Stewart state that they were unable to show that the computed solution $\hat{X}$ from their algorithm has a small backward error, although they could show that it has a small normwise relative residual, as in (2.3). Our analysis, and the numerical example, make it clear that $\hat{X}$ will not always have a small backward error – for $\|H^+r\|_2 \approx \|H^+\|_2\|r\|_2$ holds for some rounding errors (for example, if there is just a single rounding error, so that $r = \theta e_k$, where the kth column of H^+ has maximal norm), and then (3.8) is an approximate equality, with μ possibly large.

4. Perturbation result.

To derive a perturbation result we consider the perturbed Sylvester equation

$$(A + \Delta A)(X + \Delta X) - (X + \Delta X)(B + \Delta B) = C + \Delta C,$$

which, on dropping second order terms, becomes

$$A\Delta X - \Delta X B = \Delta C - \Delta A X + X \Delta B.$$

This system may be written in the form

$$P\,\mathrm{vec}(\Delta X) = -[X^T \otimes I_m, \quad -I_n \otimes X, \quad -I_{mn}] \begin{bmatrix} \mathrm{vec}(\Delta A) \\ \mathrm{vec}(\Delta B) \\ \mathrm{vec}(\Delta C) \end{bmatrix}, \tag{4.1}$$

where $P = I_n \otimes A - B^T \otimes I_m$. If we measure the perturbations normwise by

$$\varepsilon = \max\{\|\Delta A\|_F/\alpha, \|\Delta B\|_F/\beta, \|\Delta C\|_F/\gamma\},$$

where α, β and γ are tolerances as in (3.1), then

$$\|\Delta X\|_F/\|X\|_F \le 3^{1/2}\Psi_\varepsilon \tag{4.2}$$

is a sharp bound (to first order in ε), where

$$\Psi = \|P^{-1}[\alpha(X^T \otimes I_m), \quad -\beta(I_n \otimes X), \quad -\gamma I_{mn}]\|_2/\|X\|_F \tag{4.3}$$

is the corresponding condition number for the Sylvester equation. The bound (4.2) can be weakened to

$$\frac{\|\Delta X\|_F}{\|X\|_F} \le 3^{1/2}\Phi\varepsilon, \tag{4.4}$$

where

$$\Phi = \|P^{-1}\|_2 \frac{(\alpha + \beta)\|X\|_F + \gamma}{\|X\|_F}.$$

If $\|P^{-1}\|_2(\alpha+\beta)\varepsilon < 1/2$ then twice the upper bound in (4.4) can be shown to be a strict bound for the error. The perturbation bound (4.4) with $\alpha = \|A\|_F$, $\beta = \|B\|_F$ and $\gamma = \|C\|_F$ is the one that is usually quoted for the Sylvester equation (see [9, 14], for example); it can also be obtained by applying standard perturbation theory for $Ax = b$ to (1.2). Note that the term $\|P^{-1}\|_2$ is equal to the reciprocal of $\operatorname{sep}(A, B)$ in (1.3).

For the real Lyapunov equation, a similar derivation to the one above shows that the condition number is

$$\|(I_n \otimes A + A \otimes I_n)^{-1}[\alpha((X^T \otimes I_n) + (I_n \otimes X)\Pi^T),\ -\gamma I_{n^2}]\,\|_2/\|X\|_F,$$

where Π is the vec-permutation matrix.

How much can the bounds (4.2) and (4.4) differ? The answer is: by an arbitrary factor. To show this we consider the case where B is normal (or equivalently, A is normal if we transpose the Sylvester equation). We can assume B is in Schur form, thus $B = \operatorname{diag}(\mu_j)$ (with the μ_j possibly complex). Then $P = \operatorname{diag}(A - \mu_{jj}I_m)^{-1}$, and it is straightforward to show that if $X = [x_1, \ldots, x_n]$, and if we approximate the 2-norms in the definitions of Ψ and Φ by Frobenius norms, then

$$\Psi^2 \approx \Bigg(\alpha^2 \sum_{j=1}^{n} \|x_j\|_2^2 \|(A - \mu_{jj}I_m)^{-1}\|_F^2 + \beta^2 \sum_{j=1}^{n} \|(A - \mu_{jj}I_m)^{-1}X\|_F^2 + \gamma^2 \sum_{j=1}^{n} \|(A - \mu_{jj}I_m)^{-1}\|_F^2\Bigg)/\|X\|_F^2,$$

while

$$\Phi^2 \approx \sum_{j=1}^{n} \|(A - \mu_{jj}I_m)^{-1}\|_F^2((\alpha+\beta) + \gamma/\|X\|_F)^2.$$

These formulas show that in general Ψ and Φ will be of similar magnitude, and we know that $\Psi \le \Phi$ from the definitions. However, Ψ can be much smaller than Φ. For example, suppose that $\gamma = 0$ and

$$\|(A - \mu_{nn}I_m)^{-1}\|_F \gg \max_{j \ne n} \|(A - \mu_{jj}I_m)^{-1}\|_F.$$

Then if

$$\|x_n\|_2/\|X\|_F \ll 1 \text{ and } \|(A - \mu_{nn}I_m)^{-1}X\|_F/\|X\|_F \ll \|(A - \mu_{nn}I_m)^{-1}\|_F,$$

we have $\Psi \ll \Phi$. Such examples are easily constructed. To illustrate, let $A = \operatorname{diag}(2, 2, \ldots, 2, 1)$ and $B = \operatorname{diag}(1/2, 1/2, \ldots, 1/2, 1-\varepsilon)$, with $\varepsilon > 0$, so that $A - \mu_{nn}I_m = \operatorname{diag}(1+\varepsilon,\ 1+\varepsilon, \ldots, 1+\varepsilon, \varepsilon)$, and let $X = (A - \mu_{nn}I_m)Y$, where $Y = [y, y, \ldots, y, 0]$ with $\|(A - \mu_{nn}I_m)y\|_2 = \|A - \mu_{nn}I_m\|_2$ and $\|y\|_2 = 1$. Then, if $\gamma = O(\varepsilon)$,

$$\Psi = O(\alpha^2 + \beta^2), \quad \Phi \approx \varepsilon^{-1}(\alpha^2 + \beta^2).$$

To summarise, the "traditional" perturbation bound (4.4) for the Sylvester equation can severely overestimate the effect of a perturbation on the data when only

A and B are perturbed, because it does not take account of the special structure of the problem. In contrast, the perturbation bound (4.2) does respect the Kronecker structure, and consequently is attainable for any given A, B and C.

To obtain an a posteriori error bound for a computed solution $\hat{X} \equiv X + \Delta X$ we can set $\Delta A = 0$, $\Delta B = 0$ and $\Delta C = A\hat{X} - \hat{X}B - C = R$ in (4.1), which leads to

$$\|X - \hat{X}\|_F / \|X\|_F \leq \|P^{-1}\|_2 \, \|R\|_F / \|X\|_F. \tag{4.5}$$

A similar but potentially much smaller bound is described in the next section.

5. Practical error bounds.

For an approximate solution $\hat{x}$ to a linear system $Ax = b$ of order n, we have for $r = b - A\hat{x}$,

$$\|x - \hat{x}\|_\infty = \|A^{-1}r\|_\infty \leq \| \, |A^{-1}| \, |r| \, \|_\infty,$$

and this bound is optimal if we are prepared to ignore signs in the elements of A^{-1} and r. To obtain a strict computed bound it is necessary to add a term that takes account of any rounding errors in forming r. The overall bound is

$$\frac{\|x - \hat{x}\|_\infty}{\|\hat{x}\|_\infty} \leq \frac{\| \, |A^{-1}|(|r| + (n+1)u(|A| \, |x| + |b|)) \, \|_\infty}{\|\hat{x}\|_\infty}. \tag{5.1}$$

The numerator in the bound is of the form $\| \, |A^{-1}|d| \, \|_\infty$, and as in [2] we have

$$\begin{aligned} \| \, |A^{-1}|d| \, \|_\infty &= \| \, |A^{-1}|De| \, \|_\infty = \| \, |A^{-1}D|e| \, \|_\infty \\ &= \| \, |A^{-1}D| \, \|_\infty = \|A^{-1}D\|_\infty, \end{aligned}$$

where $D = \operatorname{diag}(d)$ and $e = (1, 1, \ldots 1)^T$. Hence $\| \, |A^{-1}|d| \, \|_\infty$ can be estimated using the norm estimator of [11, 18, 19], which estimates $\|B\|_1$ at the cost of forming a few matrix-vector products involving B and B^T. With $B = (A^{-1}D)^T$ we need to solve a few linear systems involving A and A^T. The bound (5.1) is the one returned by the linear equation solvers in the Fortran linear algebra library LAPACK [1]; it is estimated in the way described above.

For the Sylvester equation we can use the same approach if we identify $Ax = b$ with (1.2). For the computed residual we have

$$\begin{aligned} \hat{R} &= fl(C - (A\hat{X} - \hat{X}B)) = R + \Delta R, \\ |\Delta R| &\leq u(3|C| + (m+3)|A| \, |\hat{X}| + (n+3)|\hat{X}| \, |B|) \equiv R_u. \end{aligned}$$

Therefore the bound is

$$\frac{\|X - \hat{X}\|_M}{\|\hat{X}\|_M} \leq \frac{\| \, |P^{-1}|(|\operatorname{vec}(\hat{R})| + \operatorname{vec}(R_u)) \, \|_M}{\|\hat{X}\|_M}, \tag{5.2}$$

where $\|X\|_M = \max_{i,j}|x_{ij}|$. Using the technique described above, this bound can be estimated at the cost of solving a few linear systems with coefficient matrices $I_n \otimes A - B^T \otimes I_m$ and its transpose – in other words, solving a few Sylvester equations $AX - XB = C$ and $A^TX - XB^T = D$. If the Bartels-Stewart algorithm is used, these solutions can be computed with the aid of the previously computed Schur decompositions of A and B. The condition number Ψ in (4.3) and $\mathrm{sep}(A, B) = \|P^{-1}\|_2^{-1}$ can both be estimated in much the same way. Alternative algorithms for efficiently estimating $\mathrm{sep}(A, B)$ given Schur decompositions of A and B are given in [6, 23].

The attraction of (5.2) is that large elements in the jth column of P^{-1} may be countered by a small jth element of $|\mathrm{vec}(\hat{R})| + \mathrm{vec}(R_u)$, making the bound much smaller than (4.5). In this sense (5.2) has better scaling properties than (4.5), although (5.2) is not actually invariant under diagonal scalings of the Sylvester equation.

We give a numerical example to illustrate the advantage of (5.2) over (4.5). Let

$$A = J_3(0), \quad B = J_3(10^{-3}), \quad c_{ij} \equiv 1,$$

where $J_n(\lambda)$ denotes a Jordan block of size n with eigenvalue λ. Solving the Sylvester equation by the Bartels-Stewart algorithm we found that the bounds are

$$(4.5): 8.00 \times 10^{-3}, \quad (5.2): 6.36 \times 10^{-15}$$

(where in evaluating (4.5) we replaced R by $|\hat{R}| + R_u$, as in (5.2)). Here, $\mathrm{sep}(A, B) = 1.67 \times 10^{-16}$, and the bound (5.2) is small because relatively large elements of $|\mathrm{vec}(\hat{R})| + \mathrm{vec}(R_u)$ are nullified by relatively small columns of P^{-1}. For this example, with $\alpha = \|A\|_F$, $\beta = \|B\|_F$, $\gamma = \|C\|_F$, we have

$$\Psi = 7.00 \times 10^9, \quad \Phi = 1.70 \times 10^{16},$$

confirming that the usual perturbation bound (4.4) for the Sylvester equation can be very pessimistic. Furthermore,

$$\frac{\|R\|_F}{(\|A\|_F + \|B\|_F)\|\hat{X}\|_F + \|C\|_F} = 7.02 \times 10^{-24},$$

$$(\hat{X}) = \{6 \times 10^{15}, 5 \times 10^4, 3 \times 10^2\},$$

$$\eta(\hat{X}) \approx \|H^+ r\|_2 = 1.00 \times 10^{-19}, \quad \mu = 2.26 \times 10^{13},$$

so we have an example where the backward error is small despite a large-normed H^+, since $\|H^+ r\|_2 \ll \|H^+\|_2 \|r\|_2$.

Finally, we mention that the backward error of a computed solution $\hat{X}$ can be bounded by estimating $\sigma_{\min}(\hat{X})$ and then evaluating the bound in (3.8). If a QR factorization $\hat{X} = QR$ is computed, then any available condition estimator can be used to estimate $\sigma_{\min}(R) = \sigma_{\min}(\hat{X})$. Note that the backward error can be computed "exactly" as $\|H^+ r\|_2$ (see (3.6)) using only the SVD of $\hat{X}$, since the SVD of H is given in terms of that of $\hat{X}$ as described in (3.7).

6. Software.

The computations discussed above can all be done using the LAPACK software [1]. The Bartels-Stewart algorithm can be implemented by calling xGEES[2] to compute the Schur decomposition, using the level 3 BLAS routine xGEMM to transform the right-hand side C, calling xTRSYL to solve the (quasi-) triangular Sylvester equation, and using xGEMM to transform back to the solution X. The error bound (5.2) can be estimated using xLACON (which implements the estimator of [11, 18, 19]) in conjunction with the above routines. We have written a Fortran 77 code dggsvx that follows the above outline. It is in the style of an LAPACK driver and follows the LAPACK naming conventions.

Acknowledgements.

I thank Zhaojun Bai for bringing the question of backward error for the Sylvester equation to my attention, and Bai and Jim Demmel for fruitful discussions on this work and for their comments on the manuscript.

REFERENCES

1. E. Anderson, Z. Bai, C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, S. Ostrouchov, and D. C. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.
2. 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.
3. Z. Bai and J. W. Demmel, *On a direct algorithm for computing invariant subspaces with specified eigenvalues*, Technical Report CS-91-139, Department of Computer Science, University of Tennessee, Nov. 1991. (LAPACK Working Note #38).
4. J. B. Barlow, M. M. Monahemi, and D. P. O'Leary, *Constrained matrix Sylvester equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1–9.
5. R. H. Bartels and G. W. Stewart, *Algorithm 432: Solution of the matrix equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.
6. R. Byers, *A LINPACK-style condition estimator for the equation $AX - XB^T = C$*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 926–928.
7. K. Datta, *The matrix equation $XA - BX = R$ and its applications*, Linear Algebra and Appl., 109 (1988), pp. 91–105.
8. J. J. Du Croz and N. J. Higham, *Stability of methods for matrix inversion*, IMA Journal of Numerical Analysis, 12 (1992), pp. 1–19.
9. G. H. Golub, S. Nash, and C. F. Van Loan, *A Hessenberg-Schur method for the problem $AX + XB = C$*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.
10. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, second ed., 1989.
11. W. W. Hager, *Condition estimates*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 311–316.

[2] The leading "x" stands for S, C, D, or Z, which indicates the data type: single precision, complex, double precision or complex double precision.

12. J. Z. Hearon, *Nonsingular solutions of* $TA - BT = C$, Linear Algebra and Appl., 16 (1977), pp. 57–63.
13. H. V. Henderson and S. R. Searle, *The* vec-permutation matrix, the vec operator and Kronecker products: A review, Linear and Multilinear Algebra, 9 (1981), pp. 271–288.
14. G. Hewer and C. Kenney, *The sensitivity of the stable Lyapunov equation*, SIAM J. Control and Optimization, 26 (1988), pp. 321–344.
15. D. J. Higham and N. J. Higham, *Componentwise perturbation theory for linear systems with multiple right-hand sides*, Numerical Analysis Report No. 200, University of Manchester, England, July 1991. To appear in Linear Algebra and Appl.
16. ———, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.
17. N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra and Appl., 88/89 (1987), pp. 405–430.
18. ———, *FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm 674)*, ACM Trans. Math. Soft., 14 (1988), pp. 381–396.
19. ———, *Experience with a matrix norm estimator*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 804–809.
20. A. S. Hodel, *Recent applications of the Lyapunov equation in control theory*, Manuscript, Dept. of Electrical Engineering, Auburn University, 1991. To appear in Proceedings of the IMACS International Symposium on Iterative Methods in Linear Algebra.
21. R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991.
22. D. Y. Hu and L. Reichel, *Krylov subspace methods for the Sylvester equation*, Linear Algebra and Appl., 172 (1992), pp. 283–314.
23. B. Kågström and P. Poromaa, *Distributed and shared memory block algorithms for the triangular Sylvester equation with* sep^{-1} estimators, SIAM J. Matrix Anal., 13 (1992), pp. 90–101.
24. 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.
25. R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.
26. G. Starke and W. Niethammer, *SOR for* $AX - XB = C$, Linear Algebra and Appl., 154–156 (1991), pp. 355–375.
27. G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Review, 15 (1973), pp. 727–764.
28. J. M. Varah, *On the* separation of two matrices, SIAM J. Numer. Anal., 16 (1979), pp. 216–222.
29. E. L. Wachspress, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Lett. 1 (1988), pp. 87–90.