МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2003
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2003
Proceedings of the Thirty Second Spring Conference of
the Union of Bulgarian Mathematicians
Sunny Beach, April 5–8, 2003

SENSITIVITY OF THE STANDARD MATRIX ALGEBRAIC EQUATION*

M. M. Konstantinov

In this paper we study the sensitivity of the standard matrix algebraic equation $AX = B$. Asymptotic properties of perturbation bounds for this equation are analyzed.

1. Introduction. In this paper we present a perturbation analysis for the standard linear matrix algebraic equation. The estimates presented are valid both for real and complex equations.

We denote by $\mathbf{F}^{m\times n}$ the space of $m \times n$ matrices over the field of real ($\mathbf{F} = \mathbf{R}$) or complex ($\mathbf{F} = \mathbf{C}$) numbers, and $\mathbf{R}_+ = [0, \infty)$. The Frobenius and the spectral norms in $\mathbf{F}^{m\times n}$ are denoted as $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_2$, respectively. The matrix $|M| = [|m_{ij}|] \in \mathbf{R}_+^{m\times n}$ is the absolute value of $M = [m_{ij}] \in \mathbf{F}^{m\times n}$ and $M \otimes N = [m_{ij}N]$ is the Kronecker product of the matrices M, N. We use the notation $\mathcal{O}(m) \subset \mathbf{R}^{m\times m}$ and $\mathcal{U}(m) \in \mathbf{C}^{m\times m}$ for the multiplicative groups of real orthogonal and complex unitary $m \times m$ matrices. The component-wise partial order relation in $\mathbf{R}^{m\times n}$ is denoted by $\preceq$ while ':=' stands for 'equal by definition'.

2. Main results. Consider the standard linear matrix equation

$$AX = B, \tag{1}$$

where $A \in \mathbf{F}^{m\times m}$ is a non-singular matrix, while the coefficient B and the solution $X = A^{-1}B$ are $m \times n$ matrices over $\mathbf{F}$. This equation gives rise to some of the most popular and widely used perturbation bounds (norm-wise, component-wise, structured and backward) in numerical linear algebra [1, 2]. At the same time little is known about the tightness of these perturbation bounds. It is instructive to see how the concepts for various types of perturbation bounds are applied to this most 'unstructured' linear matrix equation.

We consider the non-trivial case $B \neq 0$ which implies $X \neq 0$. However, the results are valid also for the case $B = 0$ with the exception of those connected to relative perturbation bounds.

Let $E := (\delta B, \delta A)$ be a perturbation in the data (B, A) and $Y = X + \delta X$ be the solution of the perturbed equation $(A + \delta A)Y = B + \delta B$. For $\|\delta A\|_2\|A^{-1}\|_2 < 1$ the matrix $A + \delta A$ is non-singular and $\delta X = \delta X(E) = (A + \delta A)^{-1}(\delta B - \delta AX)$. Now the forward perturbation analysis problem is to estimate the norm $\|\delta X\|$ or the absolute

*MSC 2000: 15A06, 15A24

value $|\delta X|$ of the perturbation δX in the solution as functions of $\|\delta A\|$, $\|\delta B\|$ or $|\delta A|$, $|\delta B|$, respectively. In the following we shall use the Frobenius norm for the perturbations in the data and the solution.

Writing the perturbed equation as $\delta X = A^{-1}(\delta B - \delta AX) - A^{-1}\delta A\delta X$ (or using the explicit expression for δX) we get the following a posteriori bound, which is often used in practice

$$\delta_X = \delta_X(E) \le f(\delta) := \frac{\|A^{-1}\|_2(\delta_B + \|X\|_2\delta_A)}{1 - \|A^{-1}\|_2\delta_A}, \ \delta_A < \frac{1}{\|A^{-1}\|_2}, \tag{2}$$

where $\delta := [\delta_B, \delta_A]^\top \in \mathbf{R}_+^2$ and $\delta_Z := \|\delta Z\|_{\mathrm{F}}$. This bound is asymptotically sharp. But it is even asymptotically exact as shown below. We also prove that for $m > 1$ the bound (2) cannot in general be exact (for definitions of exactness see the paper "On Properties of Perturbation Bounds"by the author, P. Petkov, V. Mehrmann and D. Gu in these Proceedings).

We shall recall here some of these definitions. Let $\eta := [\eta_1, \eta_2]^\top \in \mathbf{R}_+^2$ and set $\omega(\eta) := \max\{\delta_X(E) : \delta \preceq \eta\}$.

The bound $\delta_X \le f(\delta)$, $\delta_A \in [0, a_0)$, $a_0 := 1/\|A^{-1}\|_2$, is:

– *asymptotically sharp* if there exist $\delta B \ne 0$, $\delta A \ne 0$ such that $\delta_X(\varepsilon E) = f(\varepsilon\delta) + o(\varepsilon)$ for $\varepsilon \to 0$;

– *asymptotically exact* if $\omega(\eta) = f(\eta) + o(\|\eta\|)$ for $\eta \to 0$;

– *exact* if $f = \omega$;

– *attainable* if there exists a one-dimensional manifold $\mathcal{M}$ such that $f(\eta) = \omega(\eta)$ for $\eta \in \mathcal{M}$ with $\eta_1, \eta_2 > 0$;

– *almost achievable* if for every positive $\tau < 1$ there exists E such that $\delta_X = \tau f(\delta)$.

Next the class of equations, for which the bound (2) is exact, is fully described. Note that here the exact domain for δ_A is the interval $[0, a_0)$.

Consider now the problem of estimating the linear combination $y = N_1x_1 + N_2x_2$, where y, x_i are vectors and N_i are matrices of corresponding size, satisfying $\|x_i\|_2 \le \eta_i$. The general case is considered in [3]. We have $\|y\|_2 \le \mathrm{est}(\eta; N)$, where $N = (N_1, N_2)$, $\mathrm{est}(\eta; N) := \min\{\mathrm{est}_2(\eta; N), \mathrm{est}_3(\eta; N)\}$ and $\mathrm{est}_2(\eta; N) := \|[N_1, N_2]\|_2\|\eta\|_2$, $\mathrm{est}_3(\eta; N) := \sqrt{\eta^\top N_0\eta}$. Here $N_0 = [n_{ij}] \in \mathbf{R}_+^{2\times 2}$ is a matrix with elements $n_{ij} = \|N_i^{\mathrm{H}}N_j\|_2$. Note that $\mathrm{est}_3(\eta; N) \le \mathrm{est}_1(\eta; N)$, where $\mathrm{est}_1(\eta; N) := \|N_1\|_2\eta_1 + \|N_2\|_2\eta_2$.

For equation (1) we have the bound

$$\delta_X \le \frac{\mathrm{est}(\delta_B, \delta_A; \Lambda, N_A)}{1 - \|\Lambda\|_2\delta_A}, \ \delta_A < 1/\|\Lambda\|_2 = a_0, \tag{3}$$

where $\Lambda := (I_n \otimes A)^{-1} = I_n \otimes A^{-1}$ and $N_A := -\Lambda(X^\top \otimes I_m) = -X^\top \otimes A^{-1}$.

In turn, the component-wise perturbation bound for equation (1) is obtained as follows. Suppose that $|\delta Z| \preceq \Delta_Z$, $Z = B, A$, where Δ_Z are given non-negative matrices of corresponding size. If the spectral radius of $|A^{-1}|\Delta_A$ is less than 1, we have

$$|\delta X| \preceq \left(I_m - |A^{-1}|\Delta_A\right)^{-1} |A^{-1}|(\Delta_B + \Delta_A|X|).$$

The only visible difference between the classical bound (2) and the bound (3) is in the numerator since the denominators in fact coincide in view of $\|\Lambda\|_2 = \|A^{-1}\|_2$. The numerator in (2) is $\|A^{-1}\|_2(\delta_B + \|X\|_2\delta_A) = \mathrm{est}_1(\delta_B, \delta_A; \Lambda, N_A)$. On the other hand we know that $\mathrm{est} \le \mathrm{est}_3 \le \mathrm{est}_1$ so that est is at least as good as est_1. In fact, both bounds coincide for this case. Indeed,

$$\begin{aligned} N_A &= -\Lambda(X^\top \otimes I_m) = -(I_n \otimes A^{-1})(X^\top \otimes I_m) = -X^\top \otimes A^{-1}, \\ N &= [\Lambda, N_A] = \left[I_n,\, -X^\top\right] \otimes A^{-1} \end{aligned}$$

and

$$\Lambda^\top N_A = -(I_n \otimes A^{-\top})(X^\top \otimes A^{-1}) = -X^\top \otimes (AA^\top)^{-1}.$$

Hence

$$\begin{aligned} \|N_A\|_2 &= \|A^{-1}\|_2 \|X\|_2,\ \left\|\Lambda^\top N_A\right\|_2 = \|A^{-1}\|_2^2 \|X\|_2, \\ \|[\Lambda, N_A]\|_2 &= \|A^{-1}\|_2 \left\|\left[I_n, -X^\top\right]\right\|_2 = \|A^{-1}\|_2 \sqrt{1 + \|X\|_2^2} \end{aligned}$$

and

$$\mathrm{est}_3(\delta_B, \delta_A; \Lambda, N_A) = \|A^{-1}\|_2(\delta_B + \|X\|_2 \delta_A) = \mathrm{est}_1(\delta_B, \delta_A; \Lambda, N_A).$$

We also have a bound

$$\begin{aligned} \psi(\gamma, \delta_B, \delta_A, \Lambda, N_A) &= \left\|\left[\Lambda, \frac{N_A}{\gamma}\right]\right\|_2 \sqrt{\delta_B^2 + \gamma^2 \delta_A^2} \\ &= \sqrt{\delta_B^2 + \|X\|_2^2 \delta_A^2 + \delta_A^2 \gamma^2 + \frac{\|X\|_2^2 \delta_B^2}{\gamma^2}}\,. \end{aligned}$$

The minimum of ψ in $\gamma > 0$ is achieved for $\gamma^0 = \|X\|_2 \delta_B / \delta_A$ and is equal to est_1 (we suppose that $\delta_A > 0$, since otherwise the results are trivial).

Thus all local bounds (with the exception of est_2) coincide with the bound est. The reason is that equation (1) has no specific structure.

We have shown that the bound $f(\delta)$ is asymptotically sharp. Next we shall show that it is also asymptotically exact. Finally we shall determine the class of equations of type (1) for which the bound is even exact.

Let

$$\begin{aligned} X &= Q\Sigma_X R^{\mathrm{H}} = Q\mathrm{diag}(\sigma_1(X), \ldots, \sigma_k(X), 0, \ldots, 0)R^{\mathrm{H}}, \\ A &= U\Sigma_A V^{\mathrm{H}} = U\mathrm{diag}(\sigma_1(A), \ldots, \sigma_m(A))V^{\mathrm{H}} \end{aligned}$$

be the singular value decompositions of X and A, respectively, where $k := \mathrm{rank}(X)$. Let q_j, r_i and u_j, v_j be the columns of the orthogonal matrices Q, R and U, V, respectively. Define the integers k_0 and ℓ_0 from

$$k_0 := \min\{i : \sigma_i(A) = \sigma_m(A)\},\ \ell_0 := \max\{i : \sigma_i(X) = \sigma_1(X)\}. \tag{4}$$

We have

$$\|N\mathrm{vec}(E)\|_2 = \left\|\mathrm{vec}^{-1}(m,n)(N\mathrm{vec}(E))\right\|_{\mathrm{F}} = \left\|A^{-1}(\delta B - \delta A X)\right\|_{\mathrm{F}},$$

where $\mathrm{vec}(E) := [\mathrm{vec}^\top(\delta B), \mathrm{vec}^\top(\delta A)]^\top$ and $\mathrm{vec}(B)$ is the column-wise vector representation of the matrix B (note that the inverse vec^{-1} of vec must contain information about the size of the matrix arguments of vec).

Let us fix the integers $i \in \{1, \ldots, \ell_0\}$ and $j \in \{k_0, \ldots, m\}$, and choose

$$\begin{aligned} \delta B &:= \delta_B \left(e_{ni}^\top \otimes u_j\right) R^{\mathrm{H}} = \delta_B u_j r_i^{\mathrm{H}}, \\ \delta A &:= -\delta_A \left(e_{mi}^\top \otimes u_j\right) Q^{\mathrm{H}} = -\delta_A u_j q_i^{\mathrm{H}}, \end{aligned}$$

where e_{ni} is the i-th column of I_n. Then

$$A^{-1}u_j = \|A^{-1}\|_2 v_j,\ q_i^{\mathrm{H}} Q\Sigma_X R^{\mathrm{H}} = \sigma_1(X) r_i.$$

Since $\|A^{-1}\|_2 = 1/\sigma_m(A)$ we get

$$\begin{aligned}
\|N\mathrm{vec}(E)\|_2 &= \left\|A^{-1}u_j\left(\delta_A r_i^{\mathrm{H}} + \delta_A q_i^{\mathrm{H}} Q\Sigma_X R^{\mathrm{H}}\right)\right\|_{\mathrm{F}}\\
&= (\delta_B + \|X\|_2\delta_A)\left\|A^{-1}u_j r_i^{\mathrm{H}}\right\|_{\mathrm{F}}\\
&= \|A^{-1}\|_2(\delta_B + \|X\|_2\delta_A)\left\|v_j r_i^{\mathrm{H}}\right\|_{\mathrm{F}}\\
&= \|A^{-1}\|_2(\delta_B + \|X\|_2\delta_A) = \mathrm{est}(\delta; N).
\end{aligned}$$

Hence $\mathrm{est}(\delta; N) \le \omega_1(\delta; N)$, where

$$\omega_1(\delta, N) := \max\{\|\Lambda z + N_A z_A\|_2 : \|z\|_2 \le \delta_B,\ \|z_A\|_2 \le \delta_A\}.$$

On the other hand $\mathrm{est}(\delta, N) \ge \omega_1(\delta, N)$ by construction. The last two inequalities yield $\mathrm{est}(\delta, N) = \omega_1(\delta, N)$. Thus we have proved the following result.

Proposition 1. *The bound* (2) *is asymptotically exact for all Sylvester equations of type* (1).

We are now going to find conditions for exactness of the bound (2). We consider mainly the case $n = 1$ when (1) is a vector equation, since it is equivalent to n vector equations for the columns of X.

Setting

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} := U^{\mathrm{H}} B,\ y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} := V^{\mathrm{H}} X,$$

where $b_i, y_i \in \mathbf{F}^{1\times n}$, we get $\Sigma_A y = b$, i.e.,

$$\sigma_i y_i = b_i,\ i = 1, \ldots, m, \tag{5}$$

where $\sigma_i := \sigma_i(A)$.

We look for extremal perturbations $b \to b + G_b$, $\Sigma_A \to \Sigma_A + G_{\Sigma_A}$, with $\|G_b\|_{\mathrm{F}} \le \delta_B$, $\|G_{\Sigma_A}\|_{\mathrm{F}} \le \delta_A < \sigma_m$ in the pair (Σ_A, b) for which the norm of the perturbation

$$\delta y = (\Sigma_A + G_{\Sigma_A})^{-1}(G_b - G_{\Sigma_A} y)$$

in the solution $y = \Sigma_A^{-1} b = V^{\mathrm{H}} X$ is maximum, i.e.,

$$\begin{aligned}
\omega(\delta) &= \max\left\{\|(\Sigma_A + \delta\Sigma)^{-1}(\delta b - \delta\Sigma y)\|_{\mathrm{F}} : \|\delta b\|_{\mathrm{F}} \le \delta_B,\ \|\delta\Sigma\|_{\mathrm{F}} \le \delta_A\right\}\\
&= \|(\Sigma_A + G_{\Sigma_A})^{-1}(G_b - G_{\Sigma_A} y)\|_2.
\end{aligned}$$

We also need the notion of an *acute perturbation* of a non-singular $m \times m$ matrix A.

Definition 1. *A perturbation δA of A is* acute *in the norm $\|\cdot\|$ if $\|\delta A\| < 1/\|A^{-1}\|$ and equality in*

$$\|(A + \delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|\delta A\|}$$

holds.

In many applications, however, we have to estimate $\|(A + \delta A)^{-1}\|_2$ as a function of $\|\delta A\|_{\mathrm{F}}$. Hence this definition must be slightly modified, since the F-norm is not an operator norm but satisfies the inequality $\|AB\|_{\mathrm{F}} \le \|A\|_2 \|B\|_{\mathrm{F}}$, which yields

$$\|(A + \delta A)^{-1}\|_2 \le \frac{\|A^{-1}\|_2}{1 - \|A^{-1}\|_2 \|\delta A\|_{\mathrm{F}}}.$$

Definition 2. *A perturbation δA of A with $\|\delta A\|_{\mathrm{F}} < \sigma_m(A)$ is said to be* F-acute *if*

$$\|(A + \delta A)^{-1}\|_2 = \frac{\|A^{-1}\|_2}{1 - \|A^{-1}\|_2 \|\delta A\|_{\mathrm{F}}} = \frac{1}{\sigma_m(A) - \|\delta A\|_{\mathrm{F}}}.$$

Given $0 < \alpha < 1/\|A^{-1}\|_2$ there are exactly $m - k_0 + 1$ different F-acute perturbations δA with $\|\delta A\|_{\mathrm{F}} = \alpha$, namely $\delta A = -\alpha\, u_j v_j^{\mathrm{H}}$, $j = k_0, \ldots, m$.

For the matrix Σ_A the F-acute perturbations are $\delta\Sigma_A = -\alpha\, E_{ii}(m,m)$ with $k_0 \le i \le m$, where the matrix $E_{ij}(m,n)\mathbf{R}^{m\times n}$ has a single non-zero element, equal to 0, in position (i,j). Generically $\sigma_{m-1} > \sigma_m$ and $k_0 = m$, i.e., there is only one F-acute perturbation $\delta A = -\alpha\, u_m v_m^{\mathrm{H}}$.

The properties of acute perturbations strongly depend on the underlying norm. Consider p-acute perturbations δA in the Hölder p-norm with $\|\delta A\|_p < \|A^{-1}\|_p^{-1}$, for which

$$\|(A + \delta A)^{-1}\|_p = \frac{\|A^{-1}\|_p}{1 - \|A^{-1}\|_p \|\delta A\|_p}.$$

For instance, if $m > 1$ there are infinitely many 2-acute perturbations.

It follows from the inequalities $\sigma_i > 0$ and the diagonal structure of system (5) that $G_{\Sigma_A} \preceq 0$ and that the i-th element of G_b must have the sign of the corresponding right-hand side b_i provided $n = 1$. Moreover, G_{Σ_A} must be diagonal, i.e.,

$$\begin{aligned} G_{\Sigma_A} &= -\mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_m),\ \varepsilon_i \ge 0, \\ G_b &= [\gamma_1 \,\mathrm{sign}(b_1), \ldots, \gamma_m \,\mathrm{sign}(b_m)]^\top,\ \gamma_i \ge 0. \end{aligned}$$

Hence

$$\delta y_i = \pm \frac{\gamma_i + |y_i|\varepsilon_i}{\sigma_i - \varepsilon_i}.$$

The extremal perturbation is now obtained as a solution of the maximization problem

$$\sum_{i=1}^{m} \left(\frac{\gamma_i + |y_i|\varepsilon_i}{\sigma_i - \varepsilon_i} \right)^2 \to \max \tag{6}$$

subject to the constraints

$$\sum_{i=1}^{m} \gamma_i^2 \le \delta_B^2,\ \sum_{i=1}^{m} \varepsilon_i^2 \le \delta_A^2, \tag{7}$$

where $\delta_A < \sigma_m$.

Using particular examples, it may be shown that in general the bound (2) is not exact when $m > 1$.

Example 1. Consider the system (5) with $m = 2$, $n = 1$ and $\delta_B = \delta_A = \eta$. The bound (2) here is

$$f(\eta, \eta) = \left(1 + \sqrt{y_1^2 + y_2^2}\right) \frac{\eta}{\sigma_2 - \eta}.$$

The maximization problem (6), (7) in γ_i, ε_i depends on five parameters σ_1, σ_2, $|y_1|$, $|y_2|$ and η, where $\sigma_1 > \sigma_2 > 0$, $0 \le \eta < \sigma_2$ and $|y_1| + |y_2| > 0$. Depending on the relations among these parameters we have the following two cases.

First, let $(\sigma_1 = \sigma_2)$ or $(\sigma_1 > \sigma_2$ and $|y_1| \leq |y_2|)$. Then

$$\omega(\eta, \eta) = (1 + \max\{|y_1|, |y_2|\}) \frac{\eta}{\sigma_2 - \eta}.$$

In this case the extremal perturbation G_{Σ_A} in Σ_A is F-acute. The bound $f(\eta, \eta)$ is exact if and only if $(\sigma_1 \geq \sigma_2$ and $b_1 = 0)$ or $(\sigma_1 = \sigma_2$ and $b_2 = 0)$.

Second, let $(\sigma_1 > \sigma_2)$ and $(|y_1| > |y_2|)$. Here the bound $f(\eta, \eta)$ is not exact. At the same time the extremal perturbation in Σ_A may not be F-acute. Indeed, the maximum norm of the perturbation δy in y for an F-acute perturbation G_{Σ_A} of Σ_A is

$$\nu_2 := (1 + |y_2|) \frac{\eta}{\sigma_2 - \eta}.$$

Suppose that $(1 + |y_1|)\sigma_2 > (1 + |y_2|)\sigma_1$ and

$$\eta < \frac{(1 + |y_1|)\sigma_2 - (1 + |y_2|)\sigma_1}{|y_1| - |y_2|}.$$

Then, taking the perturbations in b and Σ_A as

$$\delta b = \begin{bmatrix} \eta \\ 0 \end{bmatrix}, \ \delta \Sigma_A = \begin{bmatrix} -\eta & 0 \\ 0 & 0 \end{bmatrix}$$

we obtain that the norm of the perturbation in y now is

$$\nu_1 := (1 + |y_1|) \frac{\eta}{\sigma_2 - \eta} > \nu_2.$$

Hence the extremal perturbation, for which the norm of δy is at least ν_1, can not be F-acute.

The following proposition reveals the role of F-acute perturbations in exact bounds.

Proposition 2. *If the bound* (2) *is exact then every extremal perturbation G_A in A is F-acute* (*this is true in the general case $n \geq 1$*).

Proof. Suppose that the bound (2) is exact ($f(\delta) = \omega(\delta)$) but the extremal perturbation G_A in A is not acute. Then

$$\|(A + G_A)^{-1}\|_2 < \frac{1}{\sigma_m - \delta_A}$$

which yields

$$\begin{aligned} \omega(\delta) &= \left\|(A + G_A)^{-1}(G_B - G_A X)\right\|_{\mathrm{F}} \leq \left\|(A + G_A)^{-1}\right\|_2 \|G_B - G_A X\|_{\mathrm{F}} \\ &< \frac{\|G_B - G_A X\|_{\mathrm{F}}}{\sigma_m - \delta_A} \leq \frac{\delta_B + \|X\|_2 \delta_A}{\sigma_m - \delta_A} = f(\delta), \end{aligned}$$

i.e., the bound is not exact. This contradiction shows that G_A must be F-acute. $\square$

The converse statement to Proposition 2, namely that an extremal perturbation may be F-acute while the bound (2) is not exact, is not true as demonstrated in Example 1.

Hence it is important to determine the class of equations (1), for which the bound (2) is exact.

Proposition 3. *Let $n = 1$. Then the perturbation bound* (2) *is exact if and only if there exists an integer $j \in \{k_0, \ldots, m\}$, such that $b_i = u_i^{\mathrm{H}} B = 0$ for $i \neq j$ (or equivalently, such that $\left\|u_j^{\mathrm{H}} B\right\|_2 = \|B\|_2$), where $u_1, \ldots, u_m$ are the columns of the matrix U in the singular value decomposition $A = U \Sigma_A V^{\mathrm{H}}$ of A.*

Proof. *Necessity.* Suppose that the bound (2) is exact. Then according to Proposition 2 the extremal perturbation G_{Σ_A} in Σ_A is F-acute, i.e., there exists an integer $j \in \{k_0, \ldots, m\}$ such that

$$(8) \qquad \delta y_i = \begin{cases} \gamma_i/\sigma_i & \text{if} \quad i \neq j, \\ (\gamma_j + |y_j|\delta_A)/(\sigma_m - \delta_A) & \text{if} \quad i = j. \end{cases}$$

Since $\sigma_i \geq \sigma_j$ for all $i \in \{1, \ldots, m\}$, then the maximum of $\|\delta y\|_2$ in $\gamma_1, \ldots, \gamma_m$ is achieved for $\gamma_i = 0$ if $i \neq j$ and $\gamma_j = \delta_B$. Hence

$$\|\delta y\|_2 = |\delta y_j| = \frac{\delta_B + |y_j|\delta_A}{\sigma_m - \delta_A}.$$

Since the bound is exact it follows from the comparison with the right-hand side of (2) that $|y_j| = \|y\|_2$. Having in mind that $y_i = u_i^{\mathrm{H}} B/\sigma_i$ we see that y and hence B has all but one element (in the j-th position) equal to zero.

Sufficiency. Let $\left\|u_j^{\mathrm{H}} B\right\|_2 = \|B\|_2$. Then the only non-zero element of $U^{\mathrm{H}} B$ and hence of y is in the j-th position and (8) holds. Choosing $\gamma_i = 0$ if $i \neq j$ and $\gamma_j = \delta_B$ we get

$$\|\delta y\|_2 = |\delta y_j| = \frac{\delta_B + |y_j|\delta_A}{\sigma_m - \delta_A} = \frac{\delta_B + \|y\|_2\delta_A}{\sigma_m - \delta_A} = f(\delta),$$

i.e., the bound $f(\delta)$ is reached and is thus exact. $\square$

In the generic case $k_0 = m$ Proposition 3 tells us that the bound (2) is exact if and only if $B^{\mathrm{H}} U = [0, \ldots, 0, \pm\|B\|_2]^{\top}$.

Consider finally the case when the size in the perturbations is measured in 2-norm. We have

$$(9) \qquad \|\delta X\|_2 \leq \frac{\|A^{-1}\|_2(\|\delta B\|_2 + \|X\|_2\|\delta A\|_2)}{1 - \|A^{-1}\|_2\|\delta A\|_2}.$$

The bound (9) is asymptotically exact for all $n \geq 1$. Similarly to Proposition 3 we may prove the following result.

Proposition 4. *The bound* (9) *is exact for* $n = 1$ *if and only if*

$$\left\|B^{\mathrm{H}}[u_k, \ldots, u_m]\right\|_2 = \|B\|_2.$$

Proof. The proof is based on the use of the 2-acute perturbation $\delta\Sigma_A = \mathrm{diag}(0, -\delta_2 I_{m-k+1})$ in system (5). $\square$

It follows from $AX = B$ that $\|B\|_2 \leq \|A\|_2\|X\|_2$ and $1/\|X\|_2 \leq \|A\|_2/\|B\|_2$. Substituting the last inequality in (9) yields the well known a priori relative perturbation bound

$$(10) \qquad \varepsilon_X \leq \frac{\mathrm{cond}_2(A)\,(\varepsilon_B + \varepsilon_A)}{1 - \mathrm{cond}_2(A)\varepsilon_A},$$

where $\varepsilon_Z := \|\delta Z\|_2/\|Z\|_2$ and $\mathrm{cond}_2(A) := \|A\|_2\|A^{-1}\|_2$.

Unfortunately, in general **the bound (10) is not even asymptotically sharp** – this is the price of deleting the 'a posteriori' quantity $\|X\|_2$.

The asymptotically exact (and hence asymptoticaly sharp) relative perturbation bound here is

$$(11) \qquad \varepsilon_X \leq \frac{\mathrm{cond}_2(A)\,(\theta\varepsilon_B + \varepsilon_A)}{1 - \mathrm{cond}_2(A)\varepsilon_A},$$

where

$$\theta := \frac{\|B\|_2}{\|A\|_2\|X\|_2} = \frac{\|B\|_2}{\|A\|_2\|A^{-1}B\|_2}.$$

Since $\|A^{-1}B\|_2 \le \|A^{-1}\|_2\|B\|_2$ we have

$$1/\mathrm{cond}_2(A) \le \theta \le 1.$$

Thus, if: $\mathrm{cond}_2(A)$ is large, θ is close or equal to $1/\mathrm{cond}_2(A)$ and $\varepsilon_A/\varepsilon_B$ is small, then **the a priori bound (10) may be arbitrarily larger than the true a posteriori bound (11).**

Example 2. Let $A = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\delta A = \begin{bmatrix} 0 & 0 \\ 0 & -\varepsilon^2 \end{bmatrix}$, $\delta B = \begin{bmatrix} 0 \\ \varepsilon \end{bmatrix}$, where $\varepsilon > 0$ is a small parameter. The exact relative perturbation in X is $\varepsilon_X = 2\varepsilon/(1-\varepsilon)$. The a priori bound (10) here takes the form

$$\varepsilon_X \le \varphi_{\mathrm{ap}}(\varepsilon) := \frac{1+\varepsilon}{1-\varepsilon},$$

while the true a posteriori bound (11) is reduced to

$$\varepsilon_X \le \varphi_{\mathrm{tr}}(\varepsilon) := \frac{2\varepsilon}{1-\varepsilon}$$

(and is even exact for this particular case). We see that the ratio of the two bounds

$$\frac{\varphi_{\mathrm{ap}}(\varepsilon)}{\varphi_{\mathrm{tr}}(\varepsilon)} = \frac{1+\varepsilon}{2\varepsilon}$$

tends to infinity as ε tends to zero.

It follows from the above considerations that the bound (10) is asymptotically exact (for all $n \ge 1$) if and only if $\theta = 1$, which is equivalent to

$$\|B\|_2 = \|A\|_2\|X\|_2 = \|A\|_2\|A^{-1}B\|_2. \tag{12}$$

This condition may be reformulated as follows.

Proposition 5. *Set $m_0 := \max\{i : \sigma_i(A) = \sigma_1(A)\}$. The bound (10) is asymptotically exact for any $n \ge 1$ if and only if one of the following alternative conditions holds:*

1. *$A = \alpha Q$, where $0 \ne \alpha \in \mathbf{R}$ and $Q \in \mathcal{O}(m)$ when $m_0 = m$ in the real case, and $A = \alpha Q$, where $0 \ne \alpha \in \mathbf{C}$ and $Q \in \mathcal{U}(m)$ in the complex case;*
2. *$u_i^{\mathrm{H}} B = 0$ for $i > m_0$ when $m_0 < m$.*

Proof. 1. In the real case we have $m_0 = m$ if and only if $A = \alpha Q$, where $Q \in \mathcal{O}(m)$. In this case $X = Q^\top B/\alpha$ and $\|X\|_2 = \|B\|_2/|\alpha|$. The complex case is treated similarly. Since $\|A\|_2 = |\alpha|$ we have $\|B\|_2 = \|A\|_2\|X\|_2$.

2. Consider the transformed system (5). The condition (12) is equivalent to $\|b\|_2^2 = \|\Sigma_A\|_2^2\|y\|_2^2$ which in turn gives

$$\sum_{i=1}^{m_0} \|b_i\|_2^2 + \sigma_1^2 \sum_{i=m_0+1}^{m} \frac{\|b_i\|_2^2}{\sigma_i^2} = \sum_{i=1}^{m_0} \|b_i\|_2^2 + \sum_{i=m_0+1}^{m} \|b_i\|_2^2.$$

Since $\sigma_1 > \sigma_{m_0+1} \ge \cdots \ge \sigma_m$ it follows that $b_i = u_i^{\mathrm{H}} B = 0$ for $i > m_0$. $\square$

Combining Propositions 3 and 5 we also get the following necessary and sufficient condition for exactness of the bound (10).

Proposition 6. *The bound* (10) *is exact if and only if* $A = \alpha Q$, *where* $0 \neq \alpha \in \mathbf{R}$ *and* $Q \in \mathcal{O}(m)$ *in the real case, and* $A = \alpha Q$, *where* $0 \neq \alpha \in \mathbf{C}$ *and* $Q \in \mathcal{U}(m)$ *in the complex case.*

At the same time the relative bound (11) is exact together with the absolute bound (9) under the weaker condition of Proposition 4. When A is a scalar multiple of an orthogonal or unitary matrix as in the condition of Proposition 6 then $k_0 = 1$ and the condition of Proposition 4 holds.

3. Conclusions. We have analyzed perturbation bounds for the standard linear matrix equation from the viewpoint of their sharpness and exactness. The above results depend on the norm used. For Hölder p-norms with $p \neq 2$ the conditions for various types of exactness of the perturbation bounds will be different.

REFERENCES

[1] G. H. Golub, C. F. Van Loan. Matrix Computations. The Johns Hopkins University Press, Baltimore, third edition, 1996.

[2] N. J. Higham. Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia, PA, 1996.

[3] M. Konstantinov, M. Stanislavova, P. Petkov. Perturbation bounds and characterisation of the solution of the associated algebraic Riccati equation. *Linear Algebra Appl.*, **285** (1998), 7–31.

M. M. Konstantinov
UACEG, 1 Hr. Smirnenski Blvd.
1046 Sofia, Bulgaria
e-mail: mmk@uacg.bg

ЧУСТВИТЕЛНОСТ НА СТАНДАРТНОТО ЛИНЕЙНО МАТРИЧНО УРАВНЕНИЕ

М. М. Константинов

Изучена е чуствителността на стандартното матрично алгебрично уравнение $AX = B$. Анализирани са асимптотичните свойства на пертурбационните граници за това уравнение.