NORTH-HOLLAND

Condition Numbers and Their Condition Numbers

Desmond J. Higham

Department of Mathematics and Computer Science
University of Dundee
Dundee, DD1 4HN, Scotland

Submitted by Iain S. Duff

ABSTRACT

Various normwise relative condition numbers that measure the sensitivity of matrix inversion and the solution of linear systems are characterized. New results are derived for the cases where two common, noninduced matrix norms are used, and where different vector norms are used for the domain and range of the matrix. Condition numbers that respect the structure of symmetric problems are also analyzed. The sensitivity of the condition number itself is then investigated, and we obtain sharp examples of Demmel's general result that for certain problems in numerical analysis "the condition number of the condition number is the condition number." Finally, upper bounds are derived for the sensitivity of componentwise condition numbers.

1. INTRODUCTION

The classical normwise relative condition number measures the sensitivity of a matrix inverse. Given $A \in \mathbb{R}^{n\times n}$, which we will always assume to be nonsingular, and a matrix norm $\|\cdot\|$, this condition number may be defined as

$$\operatorname{cond}(A) := \lim_{\epsilon\to 0+} \sup_{\|\Delta A\|\le\epsilon\|A\|} \frac{\|(A+\Delta A)^{-1} - A^{-1}\|}{\epsilon\|A^{-1}\|} \tag{1.1}$$

LINEAR ALGEBRA AND ITS APPLICATIONS 214:193–213 (1995)

© Elsevier Science Inc., 1995
655 Avenue of the Americas, New York, NY 10010

0024-3795/95/$9.50
SSDI 0024-3795(93)00066-9

[5; 7; 8, p. 80; 15]. Note that in order to reduce the sensitivity measure to a single number, two simplifications have been introduced:

(1) We look at the *largest* relative change in A^{-1} compared with a relative change in A of size ϵ.

(2) We take the *limit* as $\epsilon \to 0_+$.

Hence a condition number records the *worst-case* sensitivity to *small* perturbations. When the matrix norm is induced by a vector norm, it is well known that $\mathrm{cond}(A)$ has the characterization

$$\mathrm{cond}(A) = \kappa(A) := \|A\|\|A^{-1}\|. \tag{1.2}$$

Since matrix inversion is rarely necessary in practice, it is of interest to define the corresponding condition number for the linear system $Ax = b$:

$$\mathrm{cond}(A, b) := \lim_{\epsilon \to 0_+} \sup_{\substack{\|\Delta A\| \le \epsilon \|A\| \\ \|\Delta b\| \le \epsilon \|b\|}} \frac{\|(A + \Delta A)^{-1}(b + \Delta b) - A^{-1}b\|}{\epsilon \|A^{-1}b\|}. \tag{1.3}$$

Here, we measure the sensitivity of the solution x to relative perturbations in A and b. For the case where $\|\cdot\|$ in (1.3) denotes any vector norm and the induced matrix norm, the characterization

$$\mathrm{cond}(A, b) = \kappa(A) + \frac{\|A^{-1}\|\,\|b\|}{\|A^{-1}b\|} \tag{1.4}$$

was derived by Bartels [2], and is quoted in [11]. Although terms like the right-hand side of (1.4) often appear in perturbation results (see, for example, [8, p. 79]), the author is not aware of any references before [2] that explicitly derive the condition number. Using the inequality $\|A\| \ge \|b\|/\|x\|$, it follows from (1.2) and (1.4) that

$$\mathrm{cond}(A) \le \mathrm{cond}(A, b) \le 2\,\mathrm{cond}(A), \tag{1.5}$$

with the right-hand inequality being attainable for some b. Hence, we see that "the condition number," $\mathrm{cond}(A)$, gives a reasonable order-of-magnitude sensitivity measure for any linear system $Ax = b$.

Condition numbers are useful for two distinct reasons. When the data $\{A, b\}$ contains errors, either experimental or numerical, the condition number bounds the level of uncertainty inherent in the solution before a numerical algorithm is applied. Also, when combined with a backward error estimate, the condition number provides an approximate upper bound on the error in a computed solution.

This work extends the standard condition-number theory in three ways. First we examine the definitions (1.1) and (1.3) in the case of noninduced matrix norms. Then we look at the corresponding definitions that arise when A and ΔA are

constrained to be symmetric. Finally, to determine the sensitivity of the problem of computing condition numbers, we look at the "condition number of the condition number." Our aim throughout is to obtain neat characterizations or bounds for the newly defined quantities, and to relate them to standard sensitivity measures.

Our notation for vector and matrix norms is as follows. The Hölder vector p-norms will be written $\|\cdot\|_p$; that is,

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad 1 \le p < \infty, \quad \text{and} \quad \|x\|_\infty := \max_{1\le i\le n} |x_i|. \tag{1.6}$$

Given arbitrary vector norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, we define the corresponding operator norm $\|\cdot\|_{\alpha,\beta}$ by [8, p. 57]

$$\|A\|_{\alpha,\beta} := \max_{\|x\|_\alpha = 1} \|Ax\|_\beta.$$

Note that, in general, the submultiplicative property $\|AB\|_{\alpha,\beta} \le \|A\|_{\alpha,\beta}\,\|B\|_{\alpha,\beta}$ does not hold, but we do have

$$\|AB\|_{\alpha,\beta} \le \|A\|_{\gamma,\beta}\|B\|_{\alpha,\gamma}, \tag{1.7}$$

for any third vector norm $\|\cdot\|_\gamma$. The choice $\|\cdot\|_\alpha = \|\cdot\|_1$ and $\|\cdot\|_\beta = \|\cdot\|_\infty$ produces the max norm,

$$\|A\|_{1,\infty} = \|A\|_{\max} := \max_{1\le i,\, j\le n} |a_{ij}|.$$

For simplicity, we write the induced norm $\|\cdot\|_{\alpha,\alpha}$ as $\|\cdot\|_\alpha$. The Frobenius norm is defined by $\|A\|_F := \sqrt{\sum_{i,\,j=1}^{n} a_{ij}^2}$, and we recall that if $A = U\Sigma V^T$ is a singular value decomposition (SVD) of A [8, p. 71], with the singular values ordered so that $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0$, then the Frobenius and spectral norms satisfy $\|A\|_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$ and $\|A\|_2 = \sigma_1$.

In the next section we give characterizations for the condition numbers $\mathrm{cond}(A)$ in (1.1) that arise when the Frobenius norm or the max norm is used, and also when a general $\|\cdot\|_{\alpha,\beta}$ norm is used to measure A with $\|\cdot\|_{\beta,\alpha}$ measuring A^{-1}. Section 3 relates the $\|\cdot\|_F$ condition number to the corresponding distance to the nearest singular matrix, showing that the two quantities are not reciprocal in general. For completeness, we also include the analogous result for the α, β case, which Kahan [14] attributes to Gastinel. In Section 4 we characterize $\mathrm{cond}(A, b)$ in (1.3) when $\|\cdot\|_2$ is used to measure x and b and $\|\cdot\|_F$ is used for A. We also characterize the case where $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are used to measure x and b respectively, and $\|\cdot\|_{\alpha,\beta}$ is used to measure A. We show that a natural generalization of (1.4) arises, Section 5 looks at the condition numbers that are obtained when A and ΔA are

symmetric. We find that $\mathrm{cond}(A)$ remains unchanged using $\|\cdot\|_2$, $\|\cdot\|_F$ or $\|\cdot\|_{\max}$. For linear systems, we show that even when A is not symmetric, forcing ΔA to be symmetric does not alter the $\|\cdot\|_2$ condition number and reduces the $\|\cdot\|_F$ condition number by no more than a factor of $1/\sqrt{2}$. In Section 6, we prove that the condition numbers are approximately as sensitive as the original problems that they describe. Finally, in Section 7 we review results on componentwise condition numbers and derive upper bounds on their sensitivity.

2. MATRIX INVERSION

We begin this section with a characterization of the Frobenius-norm version of $\mathrm{cond}(A)$ in (1.1). This result was derived by the author and Sven Bartels in 1991 and appeared in [2]. We include a proof here, since it will be referred to later when symmetric perturbations are considered.

THEOREM 2.1. *The condition number*

$$\mathrm{cond}_F(A) := \lim_{\epsilon \to 0+} \sup_{\|\Delta A\|_F \le \epsilon \|A\|_F} \frac{\|(A+\Delta A)^{-1} - A^{-1}\|_F}{\epsilon \|A^{-1}\|_F} \tag{2.1}$$

satisfies

$$\mathrm{cond}_F(A) = \frac{\|A\|_F \|A^{-1}\|_2^2}{\|A^{-1}\|_F}. \tag{2.2}$$

Proof. With $\|\Delta A\|_F \le \epsilon \|A\|_F$, neglecting $O(\epsilon^2)$ terms in a standard expansion (see, for example, [8, Lemma 2.3.3]) gives

$$(A+\Delta A)^{-1} - A^{-1} = -A^{-1}\,\Delta A\, A^{-1}. \tag{2.3}$$

Hence, the result is proved if we can show that

$$\sup_{\|\widehat{\Delta A}\|_F \le 1} \|A^{-1}\,\widehat{\Delta A}\,A^{-1}\|_F = \|A^{-1}\|_2^2. \tag{2.4}$$

The general inequalities $\|BC\|_F \le \|B\|_2 \|C\|_F$ and $\|BC\|_F \le \|B\|_F \|C\|_2$ (see [13, p. 313]) give "$\le$" in (2.4). Equality is found by taking $\widehat{\Delta A} = V e_1 e_1^T U^T$, where $A^{-1} = U\Sigma V^T$ is an SVD with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0$, and e_1 is the first column of the identity matrix. ∎

Writing $\kappa_F(A) := \|A\|_F \|A^{-1}\|_F$ for the analogue of (1.2), we see from Theorem 2.1 that $\mathrm{cond}_F(A) \ne \kappa_F(A)$ in general. Using the inequalities $\|A\|_2 \le$

$\|A\|_F \leq \sqrt{n}\|A\|_2$ [13, p. 314] it follows that

$$\frac{\kappa_F(A)}{n} \leq \operatorname{cond}_F(A) \leq \kappa_F(A).$$

Next we characterize the condition number that arises when $\|\cdot\|_{\max}$ is used in (1.1).

THEOREM 2.2. *The condition number*

$$\operatorname{cond}_{\max}(A) := \lim_{\epsilon \to 0_+} \sup_{\|\Delta A\|_{\max} \leq \epsilon \|A\|_{\max}} \frac{\|(A+\Delta A)^{-1} - A^{-1}\|_{\max}}{\epsilon \|A^{-1}\|_{\max}} \tag{2.5}$$

satisfies

$$\operatorname{cond}_{\max}(A) = \frac{\|A\|_{\max}\|A^{-1}\|_\infty \|A^{-1}\|_1}{\|A^{-1}\|_{\max}}. \tag{2.6}$$

Proof. By analogy with the proof of Theorem 2.1, we must prove that

$$\sup_{\|\widehat{\Delta A}\|_{\max} \leq 1} \|A^{-1}\widehat{\Delta A}\, A^{-1}\|_{\max} = \|A^{-1}\|_\infty \|A^{-1}\|_1. \tag{2.7}$$

Using (1.7) gives "$\leq$" in (2.7). To show that equality is attainable, let r_i^T and c_j denote the ith row and jth column of A^{-1}, respectively. Suppose that k and l are such that $\|r_k\|_1 = \|A^{-1}\|_\infty$ and $\|c_l\|_1 = \|A^{-1}\|_1$. Let E denote the matrix of ones, and let $D_1 = \operatorname{diag}(\operatorname{sign} r_k)$ and $D_2 = \operatorname{diag}(\operatorname{sign} c_l)$. Then choosing $\widehat{\Delta A} = D_1 E D_2$ gives $\|\widehat{\Delta A}\|_{\max} = 1$ and

$$\begin{aligned} \|A^{-1}\widehat{\Delta A}\, A^{-1}\|_{\max} &\geq (A^{-1}\widehat{\Delta A}\, A^{-1})_{kl} = |r_k|^T E |c_l| \\ &= \|r_k\|_1 \|c_l\|_1 = \|A^{-1}\|_\infty \|A^{-1}\|_1, \end{aligned}$$

as required. ∎

Finally, we consider the case where A is regarded as an operator with the norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ measuring vectors in the domain and range of A respectively. This suggests the use of $\|\cdot\|_{\alpha,\beta}$ to measure A and $\|\cdot\|_{\beta,\alpha}$ to measure A^{-1}. To analyze the corresponding condition number, we require the following lemma, which is essentially a simple application of the Hahn-Banach theorem.

LEMMA 2.1. *Given vector norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ and vectors $x, y \in \mathbb{R}^n$ such that $\|x\|_\alpha = \|y\|_\beta = 1$, there exists a matrix B with $\|B\|_{\alpha,\beta} = 1$ such that $Bx = y$.*

Proof. Recall that the dual norm of $\|\cdot\|_\alpha$ is defined by $\|z\|_\alpha^* = \max_{\|w\|_\alpha = 1} |z^T w|$. Now from a standard duality result [13, p. 288] there ex-

ists a vector $z \in \mathbb{R}^n$ such that $\|z\|_\alpha^* = 1$ and $z^T x = \|x\|_\alpha = 1$. Let $B = yz^T$. Then $Bx = y$ and

$$\|B\|_{\alpha,\beta} = \max_{\|w\|_\alpha=1} \|yz^T w\|_\beta = \|y\|_\beta \max_{\|w\|_\alpha=1} |z^T w| = \|y\|_\beta \|z\|_\alpha^* = 1,$$

as required. ∎

We remark that this result is the key to the characterization of many normwise condition number and backward error expressions [2, 5, 7, 14, 16]. The next theorem generalizes the standard condition-number characterization (1.2).

THEOREM 2.3. *The condition number*

$$\mathrm{cond}_{\alpha,\beta}(A) := \lim_{\epsilon \to 0_+} \sup_{\|\Delta A\|_{\alpha,\beta} \le \epsilon \|A\|_{\alpha,\beta}} \frac{\|(A+\Delta A)^{-1} - A^{-1}\|_{\beta,\alpha}}{\epsilon \|A^{-1}\|_{\beta,\alpha}} \tag{2.8}$$

satisfies

$$\mathrm{cond}_{\alpha,\beta}(A) = \|A\|_{\alpha,\beta} \|A^{-1}\|_{\beta,\alpha}. \tag{2.9}$$

Proof. Following the proofs of the previous two theorems, the required result is

$$\sup_{\|\widehat{\Delta A}\|_{\alpha,\beta} \le 1} \|A^{-1} \widehat{\Delta A}\, A^{-1}\|_{\beta,\alpha} = \|A^{-1}\|_{\beta,\alpha}^2, \tag{2.10}$$

and, once more, "$\le$" can be deduced using (1.7). To show the opposite inequality, we have

$$\|A^{-1} \widehat{\Delta A}\, A^{-1}\|_{\beta,\alpha} = \max_{\|\widehat{y}\|_\beta = 1} \|A^{-1} \widehat{\Delta A}\, A^{-1} \widehat{y}\|_\alpha \ge \|A^{-1} \widehat{\Delta A} \hat{x}\|_\alpha \|A^{-1}\|_{\beta,\alpha}, \tag{2.11}$$

where, for the lower bound, we have chosen $\widehat{y}$ so that $\|A^{-1}\widehat{y}\|_\alpha = \|A^{-1}\|_{\beta,\alpha}$, and hence $\widehat{x}$ is a vector such that $\|\widehat{x}\|_\alpha = 1$. Now, from Lemma 2.1, there exists a matrix $\widehat{\Delta A}$ with $\|\widehat{\Delta A}\|_{\alpha,\beta} = 1$ such that $\widehat{\Delta A}\widehat{x} = \widehat{y}$. In (2.11) this gives $\|A^{-1}\Delta A\, A^{-1}\|_{\beta,\alpha} \ge \|A^{-1}\|_{\beta,\alpha}^2$, as required. ∎

3. NEARNESS TO SINGULARITY

It is well known that when $\|\cdot\|$ denotes a vector norm and the induced matrix norm, the reciprocal of $\mathrm{cond}(A)$ in (1.1) gives the relative distance from A to the set of singular matrices. This idea extends readily to the general $\|\cdot\|_{\alpha,\beta}$ norm. The result is essentially proved in [14], where it is attributed to Gastinel. More precisely, Kahan assumes that $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are Hölder p-norms, and does not explicitly point out that $\|\cdot\|_{\beta,\alpha}$ should be used to measure A^{-1}.

THEOREM 3.1. *Defining*

$$\operatorname{dist}_{\alpha,\beta}(A) := \min\left\{\frac{\|\Delta A\|_{\alpha,\beta}}{\|A\|_{\alpha,\beta}} : A + \Delta A \text{ singular}\right\},$$

we have

$$\operatorname{dist}_{\alpha,\beta}(A) := (\|A\|_{\alpha,\beta}\|A^{-1}\|_{\beta,\alpha})^{-1} = \operatorname{cond}_{\alpha,\beta}(A)^{-1}.$$

Proof. See Kahan [14, pp. 775–776]. ∎

The next result concerns the Frobenius-norm distance to singularity.

THEOREM 3.2. *Defining*

$$\operatorname{dist}_F(A) := \min\left\{\frac{\|\Delta A\|_F}{\|A\|_F} : A + \Delta A \text{ singular}\right\},$$

we have

$$\operatorname{dist}_F(A) = (\|A\|_F\|A^{-1}\|_2)^{-1}.$$

Proof. Using Theorem 3.1 with $\|\cdot\|_2$, if $A + \Delta A$ is singular then $\|\Delta A\|_F \geq \|\Delta A\|_2 \geq 1/\|A^{-1}\|_2$. Now, to show that equality is possible, note that $(A + \Delta A)^{-1} = (I + A^{-1}\Delta A)^{-1}A^{-1}$, so that it is sufficient to find a suitable ΔA that makes $I + A^{-1}\Delta A$ singular. Let $A^{-1} = U\Sigma V^T$ be an SVD with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$, and let $\Delta A = VYU^T$. Then $I + A^{-1}\Delta A = U(I + \Sigma Y)U^T$. The choice $Y = -e_1e_1^T/\sigma_1$ makes $I + \Sigma Y$ singular, with $\|\Delta A\|_F = \|Y\|_F = 1/\sigma_1 = 1/\|A^{-1}\|_2$ as required. ∎

Comparing $\operatorname{cond}_{\max}(A)$ in (2.6) with $\operatorname{dist}_{1,\infty}(A)$ in Theorem 3.1, and $\operatorname{cond}_F(A)$ in (2.2) with $\operatorname{dist}_F(A)$ in Theorem 3.2, we see that the reciprocal relationship in Theorem 3.1 does not hold in general for noninduced matrix norms.

4. LINEAR SYSTEMS

In this section we look at condition numbers analogous to (1.3) that arise when noninduced matrix norms are used. We begin with a characterization for the case where the $\|\cdot\|_2$ vector norm is combined with the Frobenius norm.

THEOREM 4.1. *For the linear system $Ax = b$ the condition number*

$$\operatorname{cond}_F(A, b) := \lim_{\epsilon\to 0+} \sup_{\substack{\|\Delta A\|_F \leq \epsilon\|A\|_F \\ \|\Delta b\|_2 \leq \epsilon\|b\|_2}} \frac{\|(A + \Delta A)^{-1}(b + \Delta b) - A^{-1}b\|_2}{\epsilon\|A^{-1}b\|_2} \quad (4.1)$$

satisfies

$$\mathrm{cond}_F(A,b) = \|A\|_F\|A^{-1}\|_2 + \frac{\|A^{-1}\|_2\|b\|_2}{\|A^{-1}b\|_2}. \tag{4.2}$$

Proof. Suppose $\|\Delta A\|_F \le \epsilon\|A\|_F$ and $\|\Delta b\|_2 \le \epsilon\|b\|_2$. Let $(A+\Delta A)(x+\Delta x) = b + \Delta b$. Then

$$\Delta x = A^{-1}(\Delta b - \Delta A\, x) + O(\epsilon^2). \tag{4.3}$$

Now

$$\begin{aligned}\|A^{-1}(\Delta b - \Delta A\, x)\|_2 &\le \|A^{-1}\|_2\|\Delta b\|_2 + \|A^{-1}\|_2\|\Delta A\|_2\|x\|_2 \\ &\le \epsilon\|A^{-1}\|_2(\|b\|_2 + \|A\|_F\|x\|_2),\end{aligned}$$

giving "$\le$" in (4.2). Now suppose $\|\widehat{y}\|_2 = 1$ and $\|A^{-1}\widehat{y}\|_2 = \|A^{-1}\|_2$. Then let $\Delta b = \epsilon\widehat{y}\|b\|_2$ and $\Delta A = -\widehat{y}x^T\epsilon\|A\|_F/\|x\|_2$, so that $\|\Delta A\|_F = \epsilon\|A\|_F$. With these perturbations $\|A^{-1}(\Delta b - \Delta A\, x)\|_2 = \epsilon\|A^{-1}\|_2(\|b\|_2 + \|A\|_F\|x\|_2)$, giving equality in (4.2). ∎

The case where $\|\cdot\|_{\max}$ is used to measure A and $\|\cdot\|_\infty$ is used to measure x and b is covered by the componentwise analysis of [10]. For completeness, we quote the result below.

THEOREM 4.2. *For the linear system $Ax = b$ the condition number*

$$\mathrm{cond}_{\max}(A,b) := \lim_{\epsilon\to 0_+} \sup_{\substack{\|\Delta A\|_{\max}\le\epsilon\|A\|_{\max} \\ \|\Delta b\|_\infty\le\epsilon\|b\|_\infty}} \frac{\|(A+\Delta A)^{-1}(b+\Delta b) - A^{-1}b\|_\infty}{\epsilon\|A^{-1}b\|_\infty} \tag{4.4}$$

satisfies

$$\mathrm{cond}_{\max}(A,b) = \frac{\||A^{-1}|\widehat{E}|A^{-1}b| + |A^{-1}|\hat{e}\|_\infty}{\|A^{-1}b\|_\infty}, \tag{4.5}$$

where $\widehat{E} \in \mathbb{R}^{n\times n}$ has all components equal to $\|A\|_{\max}$ and $\widehat{e} \in \mathbb{R}^n$ has all components equal to $\|b\|_\infty$.

Proof. See the analysis leading up to Equation (3.5) of [10]. ∎

When two possibly different norms are used to measure vectors in the domain and range of A, we have the following generalization of (1.4).

THEOREM 4.3. *For the linear system $Ax = b$ the condition number*

$$\operatorname{cond}_{\alpha,\beta}(A, b) := \lim_{\epsilon \to 0+} \sup_{\substack{\|\Delta A\|_{\alpha,\beta} \le \epsilon \|A\|_{\alpha,\beta} \\ \|\Delta b\|_\beta \le \epsilon \|b\|_\beta}} \frac{\|(A + \Delta A)^{-1}(b + \Delta b) - A^{-1}b\|_\alpha}{\epsilon \|A^{-1}b\|_\alpha} \tag{4.6}$$

satisfies

$$\operatorname{cond}_{\alpha,\beta}(A, b) = \|A\|_{\alpha,\beta}\|A^{-1}\|_{\beta,\alpha} + \frac{\|A^{-1}\|_{\beta,\alpha}\|b\|_\beta}{\|A^{-1}b\|_\alpha}. \tag{4.7}$$

Proof. Suppose $\|\Delta A\|_{\alpha,\beta} \le \epsilon \|A\|_{\alpha,\beta}$ and $\|\Delta b\|_\beta \le \epsilon \|b\|_\beta$. As in the proof of Theorem 4.1, the key quantity is $A^{-1}(\Delta b - \Delta A\, x)$. We have, using (1.7),

$$\begin{aligned} \|A^{-1}(\Delta b - \Delta A\, x)\|_\alpha &\le \|A^{-1}\, \Delta b\|_\alpha + \|A^{-1}\, \Delta A\, x\|_\alpha \\ &\le \|A^{-1}\|_{\beta,\alpha}\|\Delta b\|_\beta + \|A^{-1}\, \Delta A\|_\alpha \|x\|_\alpha \\ &\le \|A^{-1}\|_{\beta,\alpha}\|\Delta b\|_\beta + \|A^{-1}\|_{\beta,\alpha}\|\Delta A\|_{\alpha,\beta}\|x\|_\alpha \\ &\le \epsilon \|A^{-1}\|_{\beta,\alpha}(\|b\|_\beta + \|A\|_{\alpha,\beta}\|x\|_\alpha), \end{aligned}$$

giving "$\le$" in (4.7). Now, suppose that $\|\widehat{y}\|_\beta = 1$ with $\|A^{-1}\widehat{y}\|_\alpha = \|A^{-1}\|_{\beta,\alpha}$, and choose $\Delta b = \epsilon \|b\|_\beta \widehat{y}$. From Lemma 2.1 there exists a matrix B with $\|B\|_{\alpha,\beta} = 1$ such that $Bx/\|x\|_\alpha = -\widehat{y}$. Letting $\Delta A = \epsilon \|A\|_{\alpha,\beta} B$, we have

$$\|A^{-1}(\Delta b - \Delta A\, x)\|_\alpha = \epsilon \|A^{-1}\|_{\beta,\alpha}(\|b\|_\beta + \|A\|_{\alpha,\beta}\|x\|_\alpha),$$

showing that equality is possible in (4.7). ■ ■

Since $\|A\|_{\alpha,\beta}\|x\|_\alpha \ge \|b\|_\beta$, with equality for some b, we see from (2.9) and (4.7) that with the $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ measures the inequalities (1.5) still hold. It is also clear from the proof of Theorem 4.3 that if we alter the definition of $\operatorname{cond}_{\alpha,\beta}(A, b)$ so that b cannot be perturbed, then $\operatorname{cond}_{\alpha,\beta}(A)$ and $\operatorname{cond}_{\alpha,\beta}(A, b)$ becomes equal.

5. SYMMETRY

If a matrix A is symmetric, then it is sometimes appropriate to define condition numbers that measure sensitivity to symmetric perturbations. For example, when a symmetric matrix is stored as a symmetric floating-point matrix, the inherent uncertainty in the solution is caused by symmetric, rather than general, perturbations. The effect of structure, including symmetry, on problem sensitivity was investigated in [10] for componentwise measures. In particular, a computable characterisation of the componentwise condition number for $Ax = b$ was derived.

The treatment here differs in that we consider normwise sensitivity, and we directly compare the structured and unstructured measures. We show that in many cases imposing symmetry on ΔA has little effect on the condition number.

THEOREM 5.1. *Given $A = A^T$ and a matrix norm $\|\cdot\|_M$, define a symmetry-respecting condition number with respect to inversion by*

$$\text{symmcond}(A) := \lim_{\epsilon \to 0_+} \sup_{\substack{\|\Delta A\|_M \le \epsilon \|A\|_M \\ \Delta A = \Delta A^T}} \frac{\|(A+\Delta A)^{-1} - A^{-1}\|_M}{\epsilon \|A^{-1}\|_M}. \tag{5.1}$$

Then when the matrix is $\|\cdot\|_2$, $\|\cdot\|_F$, or $\|\cdot\|_{\max}$, the condition number (5.1) is identical to the corresponding unstructured condition number; that is, the constraint $\Delta A = \Delta A^T$ has no effect.

Proof. It is clear that imposing symmetry on ΔA cannot make the condition number bigger. To show equality for the three norms, it is sufficient to show that the optimal ΔA derived in the proofs of Theorems 2.1 and 2.2 can be taken to be symmetric. For the Frobenius norm, the result follows from the fact that the symmetric matrix A^{-1} has an SVD of the form $A^{-1} = U\Sigma U^T$. The same ΔA can be used to give the spectral norm result. For the $\|\cdot\|_{\max}$ case, when A^{-1} is symmetric the largest row sum also gives the largest column sum. Hence we may take $k = l$ in the proof of Theorem 2.2, giving $D_1 = D_2$, as required. ∎

A similar result holds for the distance to singularity using $\|\cdot\|_2$ or $\|\cdot\|_F$.

THEOREM 5.2. *Given $A = A^T$ and a matrix norm $\|\cdot\|_M$, define a symmetry-respecting distance to singularity by*

$$\text{symmdist}(A) := \min \left\{ \frac{\|\Delta A\|_M}{\|A\|_M} : A + \Delta A \text{ singular}, \Delta A = \Delta A^T \right\}. \tag{5.2}$$

Then when the matrix norm is either $\|\cdot\|_2$ or $\|\cdot\|_F$, the distance (5.2) is identical to the corresponding unstructured distance; that is, the constraint $\Delta A = \Delta A^T$ has no effect.

Proof. Constraining ΔA to be symmetric cannot make the distance smaller. Equality follows from the fact that we may take $U = V$ in the proof of Theorem 3.2. ∎

The following theorem concerns the spectral-norm condition number for $Ax = b$, and does not require A to be symmetric.

THEOREM 5.3. *For any nonsingular $A \in \mathbb{R}^{n\times n}$ (not necessarily symmetric) the spectral-norm condition number with respect to symmetric perturbations for*

$Ax = b$,

$$\text{symmcond}_2(A, b) := \lim_{\epsilon \to 0_+} \sup_{\substack{\|\Delta A\|_2 \le \epsilon \|A\|_2,\ \Delta A = \Delta A^T \\ \|\Delta b\|_2 \le \epsilon \|b\|_2}} \frac{\|(A + \Delta A)^{-1}(b + \Delta b) - A^{-1}b\|_2}{\epsilon \|A^{-1}b\|_2}, \tag{5.3}$$

is identical to the corresponding unstructured condition number; that is, the constraint $\Delta A = \Delta A^T$ has no effect.

Proof. Imposing symmetry on ΔA cannot increase the condition number. Consider the proof of Theorem 4.3 in the case where $\|\cdot\|_\alpha = \|\cdot\|_\beta = \|\cdot\|_2$. Here $\|\widehat{y}\|_2 = 1$ and $\|A^{-1}\widehat{y}\|_2 = \|A^{-1}\|_2$. Our result is proved if we can find a symmetric matrix B such that $\|B\|_2 = 1$ and $Bx/\|x\|_2 = -\widehat{y}$. This can be done by taking B as a suitably chosen Householder transformation matrix (see, for example, [8, p. 195]). ∎

Since the Frobenius norm is always within a factor $\sqrt{n}$ of the spectral norm, Theorem 5.3 can be used to derive a lower bound on the corresponding symmetric condition number. However, the bound can be further sharpened by appealing to a result of Bunch, Demmel, and Van Loan [4]. The improvement is essentially due to the fact that there exists an optimal symmetric perturbation for the $\|\cdot\|_2$ case that has rank two, whereas the Householder matrix used in the proof of Theorem 5.3 has full rank n. This allows the $\sqrt{n}$ to be reduced to $\sqrt{2}$.

THEOREM 5.4. *For any nonsingular $A \in \mathbb{R}^{n\times n}$ (not necessarily symmetric) the Frobenius norm condition number with respect to symmetric perturbations for $Ax = b$,*

$$\text{symmcond}_F(A, b) := \lim_{\epsilon \to 0_+} \sup_{\substack{\|\Delta A\|_F \le \epsilon \|A\|_F,\ \Delta A = \Delta A^T \\ \|\Delta b\|_2 \le \epsilon \|b\|_2}} \frac{\|(A + \Delta A)^{-1}(b + \Delta b) - A^{-1}b\|_2}{\epsilon \|A^{-1}b\|_2}, \tag{5.4}$$

satisfies

$$\text{symmcond}_F(A, b) \ge \|A^{-1}\|_2 \frac{\|A\|_F}{\sqrt{2}} + \frac{\|A^{-1}\|_2 \|b\|_2}{\|A^{-1}b\|_2} \ge \frac{\text{cond}_F(A, b)}{\sqrt{2}}.$$

Proof. We are concerned with $\|A^{-1}(\Delta b - \Delta A\, x)\|_2$. Letting $\widehat{y}$ be such that $\|\widehat{y}\|_2 = 1$ and $\|A^{-1}\widehat{y}\|_2 = \|A^{-1}\|_2$, we may take $\Delta b = \epsilon \widehat{y} \|b\|_2$. We know from Lemma 2.1 that there exists a matrix B such that $\|B\|_2 = 1$ and $Bx/\|x\|_2 = -\widehat{y}$.

It follows from Theorem 3 of [4] that there exists a symmetric matrix C with $\|C\|_F \leq \sqrt{2}$ such that $Cx/\|x\|_2 = -\widehat{y}$. Letting $\Delta A = C\epsilon\|A\|_F/\sqrt{2}$, we have $\|\Delta A\|_F \leq \epsilon\|A\|_F$ and

$$\|A^{-1}(\Delta b - \Delta A\, x)\|_2 = \epsilon\|A^{-1}\|_2 \left\{\|b\|_2 + \frac{\|A\|_F}{\sqrt{2}}\|x\|_2\right\},$$

giving the required bound. ∎

Theorem 5.1 shows that $\mathrm{cond}_{\max}(A)$ in (2.5) is unaffected by symmetry. The related linear-system condition numbers $\mathrm{cond}_{\max}(A,\ b)$ in (4.4) and $\mathrm{cond}_{1,\infty}(A,\ b)$ in (4.6), however, do not share this property. A bound on the effect of symmetry can be obtained by applying norm inequalities to the $\|\cdot\|_2$ version in Theorem 5.3, but the result is unlikely to be sharp in general. It is possible, though, to characterize the symmetry-respecting version of $\mathrm{cond}_{\max}(A,\ b)$—the $\|\cdot\|_{\max}$, $\|\cdot\|_\infty$ measurements overlap with the componentwise measurements used in [10], and Equation (3.4) of [10] provides an explicit formula.

6. CONDITION-NUMBER SENSITIVITY

In general, condition numbers cannot be computed exactly, and hence it is of interest to know the sensitivity of the problem "compute the condition number," that is, the condition number of the condition number. This concept was investigated by Demmel [5], who showed that for certain fundamental problems in numerical analysis, including matrix inversion, and to within unspecified multiplicative constants, *the condition number of the condition number is the condition number.* Our results below are more specialized, since they apply only to matrix inversion and the solution of linear systems, and consequently they are sharper. (They involve fixed additive and multiplicative constants rather than unknown multiplicative constants).

To motivate the analysis, we consider a linear system $Ax = b$. Typically, an *a priori* rounding error analysis or an *a posteriori* residual check will allow us to conclude that a computed solution $\widehat{x}$ satisfies a nearby system $(A + \Delta A)\widehat{x} = b + \Delta b$, where $\|\Delta A\|$ and $\|\Delta b\|$ are small, say $\max\{\|\Delta A\|/\|A\|,\ \|\Delta b\|/\|b\|\} = c_1 u$, where c_1 is close to unity and u is the machine unit roundoff. Using appropriate norms for $\mathrm{cond}(A,\ b)$, it is clear that we have the approximate error bound

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \mathrm{cond}(A,\ b)\, c_1 u. \tag{6.1}$$

Now, even when $\mathrm{cond}(A,\ b)$ has a simple characterization, it cannot normally be computed exactly. Given that A and b may contain errors before an algorithm

to compute $\operatorname{cond}(A, b)$ is applied, perhaps the best that we can hope for is to compute $\operatorname{cond}(A + \widetilde{\Delta A}, b + \widetilde{\Delta b})$, where $\max\{\|\widetilde{\Delta A}\|/\|A\|, \|\widetilde{\Delta b}\|/\|b\|\} = c_2 u$, with c_2 close to unity. The error in the computed version of the bound (6.1) may be analyzed by considering the level-2 condition number

$$\operatorname{cond}^{[2]}(A, b) := \lim_{\epsilon \to 0_+} \sup_{\substack{\|\Delta A\| \le \epsilon \|A\| \\ \|\Delta b\| \le \epsilon \|b\|}} \frac{|\operatorname{cond}(A + \Delta A, b + \Delta b) - \operatorname{cond}(A, b)|}{\epsilon \operatorname{cond}(A, b)}. \tag{6.2}$$

We then have the approximate inequality

$$\begin{aligned} &\left|\operatorname{cond}(A + \widetilde{\Delta A}, b + \widetilde{\Delta b})\, c_1 u - \operatorname{cond}(A, b)\, c_1 u\right| \\ &\quad \le \operatorname{cond}(A, b)\, c_1 u \operatorname{cond}^{[2]}(A, b)\, c_2 u. \end{aligned}$$

We conclude that if $\operatorname{cond}^{[2]}(A, b) < u^{-1}$, then using $\operatorname{cond}(A + \widetilde{\Delta A}, b + \widetilde{\Delta b})$ instead of $\operatorname{cond}(A, b)$ in (6.1) will not affect the order of magnitude of the error bound.

The results below show that for inverting a matrix or solving a linear system the sensitivity of the condition number is approximately given by the condition number itself. For the $Ax = b$ case described above the result has the following implication. If $\operatorname{cond}(A, b) < u^{-1}$, so that the error bound (6.1) indicates that we have some relative accuracy in $\widehat{x}$, then $\operatorname{cond}(A, b)$ is sufficiently well conditioned for the computed error bound to be meaningful. For simplicity, we restrict attention to condition numbers based on the $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ and $\|\cdot\|_{\alpha,\beta}$ norms; similar results for other condition numbers studied earlier can be derived.

The first result concerns matrix inversion, and relies on the following lemma.

LEMMA 6.1. *As* $\epsilon \to 0_+$

$$\begin{aligned} &\max_{\|\Delta A\|_{\alpha,\beta} \le \epsilon \|A\|_{\alpha,\beta}} \left| \|(A + \Delta A)^{-1}\|_{\beta,\alpha} - \|A^{-1}\|_{\beta,\alpha} \right| \\ &\quad = \epsilon \|A^{-1}\|_{\beta,\alpha} \operatorname{cond}_{\alpha,\beta}(A) + O(\epsilon^2). \end{aligned} \tag{6.3}$$

Proof. Using (1.7), if $\|\Delta A\|_{\alpha,\beta} \le \epsilon \|A\|_{\alpha,\beta}$ then "$\le$" in (6.3) follows by taking norms in the expansion $(A + \Delta A)^{-1} = A^{-1} - A^{-1}\,\Delta A\, A^{-1} + O(\epsilon^2)$.

Now let $\widehat{y}$ be such that $\|\widehat{y}\|_\beta = 1$ and $\|A^{-1}\widehat{y}\|_\alpha = \|A^{-1}\|_{\beta,\alpha}$ and let $\widehat{x} = A^{-1}\widehat{y}/\|A^{-1}\widehat{y}\|_\alpha$. Then by Lemma 2.1 there exists a matrix B such that $\|B\|_{\alpha,\beta} = 1$ and $B\widehat{x} = -\widehat{y}$. Choosing $\Delta A = \epsilon B \|A\|_{\alpha,\beta}$ gives

$$\begin{aligned} \|A^{-1} - A^{-1}\Delta A A^{-1}\|_{\beta,\alpha} &\ge \|(A^{-1} - A^{-1}\Delta A A^{-1})\hat{y}\|_\alpha \\ &= \|A^{-1}\|_{\beta,\alpha}(1 + \epsilon \|A\|_{\alpha,\beta} \|A^{-1}\|_{\beta,\alpha}), \end{aligned} \tag{6.4}$$

showing that (6.3) is attainable. ∎

THEOREM 6.1. *The level-2 condition number*

$$\mathrm{cond}^{[2]}_{\alpha,\beta}(A) := \lim_{\epsilon\to 0_+} \sup_{\|\Delta A\|_{\alpha,\beta}\le\epsilon\|A\|_{\alpha,\beta}} \frac{|\mathrm{cond}_{\alpha,\beta}(A+\Delta A)-\mathrm{cond}_{\alpha,\beta}(A)|}{\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)} \tag{6.5}$$

satisfies

$$\mathrm{cond}_{\alpha,\beta}(A)-1 \le \mathrm{cond}^{[2]}_{\alpha,\beta}(A) \le \mathrm{cond}_{\alpha,\beta}(A)+1.$$

Proof. If $\|\Delta A\|_{\alpha,\beta}\le\epsilon\|A\|_{\alpha,\beta}$, then using $\|A+\Delta A\|_{\alpha,\beta}\le\|A\|_{\alpha,\beta}(1+\epsilon)$ and Lemma 6.1, it follows that

$$\begin{aligned}&\|A+\Delta A\|_{\alpha,\beta}\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\\ &\quad\le \mathrm{cond}_{\alpha,\beta}(A)(1+\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)+\epsilon)+O(\epsilon^2),\end{aligned} \tag{6.6}$$

so that

$$\frac{\mathrm{cond}_{\alpha,\beta}(A+\Delta A)-\mathrm{cond}_{\alpha,\beta}(A)}{\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)} \le \mathrm{cond}_{\alpha,\beta}(A)+1+O(\epsilon). \tag{6.7}$$

Similarly, using $\|A+\Delta A\|_{\alpha,\beta}\ge\|A\|_{\alpha,\beta}(1-\epsilon)$ and Lemma 6.1, we can derive a lower bound of $-\mathrm{cond}_{\alpha,\beta}(A)-1+O(\epsilon)$ for the right-hand side of (6.7), and hence, in (6.5),

$$\mathrm{cond}^{[2]}_{\alpha,\beta}(A) \le \mathrm{cond}_{\alpha,\beta}(A)+1.$$

To get a lower bound, we may choose ΔA as in (6.4), giving

$$\begin{aligned}&\|A+\Delta A\|_{\alpha,\beta}\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\\ &\quad\ge \|A\|_{\alpha,\beta}(1-\epsilon)\|A^{-1}\|_{\beta,\alpha}[1+\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)]+O(\epsilon^2),\end{aligned}$$

and hence

$$\mathrm{cond}_{\alpha,\beta}(A+\Delta A) \ge \mathrm{cond}_{\alpha,\beta}(A)\,[1-\epsilon+\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)]+O(\epsilon^2). \tag{6.8}$$

This rearranges to

$$\frac{\mathrm{cond}_{\alpha,\beta}(A+\Delta A)-\mathrm{cond}_{\alpha,\beta}(A)}{\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)} \ge \mathrm{cond}_{\alpha,\beta}(A)-1+O(\epsilon).$$

So, in (6.5),

$$\mathrm{cond}^{[2]}_{\alpha,\beta}(A) \ge \mathrm{cond}_{\alpha,\beta}(A)-1.$$

■

Next we consider linear systems.

THEOREM 6.2. *The level-2 condition number*

$$\operatorname{cond}^{[2]}_{\alpha,\beta}(A,b) := \lim_{\epsilon\to 0+} \sup_{\substack{\|\Delta A\|_{\alpha,\beta}\le\epsilon\|A\|_{\alpha,\beta}\\ \|\Delta b\|_\beta\le\epsilon\|b\|_\beta}} \frac{|\operatorname{cond}_{\alpha,\beta}(A+\Delta A, b+\Delta b) - \operatorname{cond}_{\alpha,\beta}(A,b)|}{\epsilon \operatorname{cond}_{\alpha,\beta}(A,b)}$$

satisfies

$$\frac{\operatorname{cond}_{\alpha,\beta}(A,\,b)}{4} - \frac{1}{2} \le \operatorname{cond}^{[2]}_{\alpha,\beta}(A,\,b) \le 3\operatorname{cond}_{\alpha,\beta}(A,\,b) + 2.$$

Proof. First we derive the upper bound. Suppose $\|\Delta A\|_{\alpha,\beta} \le \epsilon\|A\|_{\alpha,\beta}$ and $\|\Delta b\|_\beta \le \epsilon\|b\|_\beta$. From Lemma 6.1, we have

$$\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\|b+\Delta b\|_\beta \le \|A^{-1}\|_{\beta,\alpha}\|b\|_\beta[1+\epsilon\operatorname{cond}_{\alpha,\beta}(A)+\epsilon] + O(\epsilon^2). \tag{6.9}$$

Also, using the definition of $\operatorname{cond}_{\alpha,\beta}(A,\,b)$,

$$\begin{aligned}\frac{1}{\|x+\Delta x\|_\alpha} \le \frac{1}{\|x\|_\alpha - \|\Delta x\|_\alpha} &= \frac{1}{\|x\|_\alpha}\left(1 + \frac{\|\Delta x\|_\alpha}{\|x\|_\alpha}\right) + O(\epsilon^2)\\ &\le \frac{1}{\|x\|_\alpha}[1+\epsilon\operatorname{cond}_{\alpha,\beta}(A,\,b)] + O(\epsilon^2).\end{aligned} \tag{6.10}$$

Combining (6.9) and (6.10), we find

$$\begin{aligned}\frac{\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\|b+\Delta b\|_\beta}{\|x+\Delta x\|_\alpha} \le \frac{\|A^{-1}\|_{\beta,\alpha}\|b\|_\beta}{\|x\|_\alpha}&[1+\epsilon\operatorname{cond}_{\alpha,\beta}(A)\\ &+\epsilon\operatorname{cond}_{\alpha,\beta}(A,b)+\epsilon] + O(\epsilon^2),\end{aligned}$$

from which it follows that

$$\begin{aligned}&\frac{\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\|b+\Delta b\|_\beta/\|x+\Delta x\|_\alpha - \|A^{-1}\|_{\beta,\alpha}\|b\|_\beta/\|x\|_\alpha}{\epsilon(\operatorname{cond}_{\alpha,\beta}(A) + \|A^{-1}\|_{\beta,\alpha}\|b\|_\beta/\|x\|_\alpha)}\\ &\quad\le 1 + \operatorname{cond}_{\alpha,\beta}(A) + \operatorname{cond}_{\alpha,\beta}(A,\,b) + O(\epsilon).\end{aligned} \tag{6.11}$$

A similar analysis gives a lower bound of $-1-\operatorname{cond}_{\alpha,\beta}(A)-\operatorname{cond}_{\alpha,\beta}(A,b) + O(\epsilon)$ for the right-hand side of (6.11), and hence we have

$$\begin{aligned}&\frac{\left|\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\|b+\Delta b\|_\beta/\|x+\Delta x\|_\alpha - \|A^{-1}\|_{\beta,\alpha}\|b\|_\beta/\|x\|_\alpha\right|}{\epsilon(\operatorname{cond}_{\alpha,\beta}(A) + \|A^{-1}\|_{\beta,\alpha}\|b\|_\beta/\|x\|_\alpha)}\\ &\quad\le 1 + \operatorname{cond}_{\alpha,\beta}(A) + \operatorname{cond}_{\alpha,\beta}(A,b) + O(\epsilon).\end{aligned} \tag{6.12}$$

Now, from Theorem 6.1,

$$\begin{aligned}
&\frac{|\mathrm{cond}_{\alpha,\beta}(A+\Delta A)-\mathrm{cond}_{\alpha,\beta}(A)|}{\epsilon(\mathrm{cond}_{\alpha,\beta}(A)+\|A^{-1}\|_{\beta,\alpha}\|b\|_\beta/\|x\|_\alpha} \\
&\quad\le \frac{|\mathrm{cond}_{\alpha,\beta}(A+\Delta A)-\mathrm{cond}_{\alpha,\beta}(A)|}{\epsilon\,\mathrm{cond}_{\alpha,\beta}(A)} \\
&\quad\le \mathrm{cond}^{[2]}_{\alpha,\beta}(A)+O(\epsilon) \\
&\quad\le \mathrm{cond}_{\alpha,\beta}(A)+1+O(\epsilon). \qquad (6.13)
\end{aligned}$$

Using the characterization (4.7), it follows from (6.12) and (6.13) that

$$\mathrm{cond}^{[2]}_{\alpha,\beta}(A,b)\le 2+2\,\mathrm{cond}_{\alpha,\beta}(A)+\mathrm{cond}_{\alpha,\beta}(A,b)\le 2+3\,\mathrm{cond}_{\alpha,\beta}(A,b).$$

For a lower bound, we may choose ΔA to satisfy (6.8), which rearranges to

$$\begin{aligned}
&\mathrm{cond}_{\alpha,\beta}(A+\Delta A)-\mathrm{cond}_{\alpha,\beta}(A) \\
&\quad\ge \epsilon[\mathrm{cond}_{\alpha,\beta}(A)-1]\,\mathrm{cond}_{\alpha,\beta}(A)+O(\epsilon^2). \qquad (6.14)
\end{aligned}$$

Choosing $\Delta b=0$ gives

$$\|\Delta x\|_\alpha=\|A^{-1}\,\Delta A\,x\|_\alpha+O(\epsilon^2)\le \mathrm{cond}_{\alpha,\beta}(A)\,\|x\|_\alpha\epsilon+O(\epsilon^2),$$

so that the expression

$$\frac{\|(A+\Delta A)^{-1}\|_{\beta,\alpha}\|b+\Delta b\|_\beta}{\|x+\Delta x\|_\alpha}-\frac{\|A^{-1}\|_{\beta,\alpha}\|b\|_\beta}{\|x\|_\alpha}$$

is guaranteed to be nonnegative, ignoring $O(\epsilon^2)$ quantities. Combining this with (6.14), it follows that

$$\begin{aligned}
\mathrm{cond}^{[2]}_{\alpha,\beta}(A,\,b)&\ge \frac{[\mathrm{cond}_{\alpha,\beta}(A)-1]\,\mathrm{cond}_{\alpha,\beta}(A)}{\mathrm{cond}_{\alpha,\beta}(A)+\|A^{-1}\|_{\beta,\alpha}\|b\|_\beta/\|x\|_\alpha} \\
&\ge \frac{[\mathrm{cond}_{\alpha,\beta}(A)-1]\,\mathrm{cond}_{\alpha,\beta}(A)}{2\,\mathrm{cond}_{\alpha,\beta}(A)} \\
&\ge \frac{\mathrm{cond}_{\alpha,\beta}(A,b)}{4}-\frac{1}{2}.
\end{aligned}$$

■

As a final point we return to the question of the relevance of the level-2 condition numbers. In practice, condition numbers will usually be computed

via their characterisations; for example, $\mathrm{cond}_{\alpha,\beta}(A) = \|A\|_{\alpha,\beta}\|A^{-1}\|_{\beta,\alpha}$. In this case, it could be argued that the best that we can hope to compute is $\|A + \Delta A_1\|_{\alpha,\beta}\|(A + \Delta A_2)^{-1}\|_{\beta,\alpha}$, where ΔA_1 and ΔA_2 are *different* small perturbations. (Even asking for $\|(A + \Delta A_2)^{-1}\|_{\beta,\alpha}$ is unreasonable, in general. However, for the commonly used $\|\cdot\|_\infty$ and $\|\cdot\|_1$ matrix norms, condition-number estimators compute rows and columns of A^{-1} by solving linear systems [9, 12]. If the correct row or column index is found, then a stable solution of the linear system will provide the row or column of the inverse of a nearby matrix.) By examining the proofs of Theorems 6.1 and 6.2 it is clear that allowing different perturbations in this manner does not significantly affect the level-2 condition numbers—in fact, as we show below, for the case of matrix inversion the upper bound in Theorem 6.1 becomes an exact characterization.

THEOREM 6.3. *The alternative level-2 condition number*

$$\overline{\mathrm{cond}}^{[2]}_{\alpha,\beta}(A) := \lim_{\epsilon\to 0+} \sup_{\substack{\|\Delta A_1\|_{\alpha,\beta}\le\epsilon\|A\|_{\alpha,\beta} \\ \|\Delta A_2\|_{\alpha,\beta}\le\epsilon\|A\|_{\alpha,\beta}}} \left| \frac{\|A + \Delta A_1\|_{\alpha,\beta}\|(A + \Delta A_2)^{-1}\|_{\beta,\alpha} - \|A\|_{\alpha,\beta}\|A^{-1}\|_{\beta,\alpha}}{\epsilon\|A\|_{\alpha,\beta}\|A^{-1}\|_{\beta,\alpha}} \right|$$

satisfies

$$\overline{\mathrm{cond}}^{[2]}_{\alpha,\beta}(A) = \mathrm{cond}_{\alpha,\beta}(A) + 1.$$

Proof. By allowing different perturbations, we can achieve equality in (6.6). The result follows immediately. ∎

7. COMPONENTWISE MEASURES

As an alternative to the normwise measures considered in the previous sections, it is possible to treat perturbations in a componentwise manner. This style of analysis goes back at least as far as Bauer [3], and is particularly meaningful for problems with some structure [1, 10]. Further, appropriately chosen componentwise measures are insensitive to diagonal scaling, and often lead to sharper error bounds. We mention that the related matter of componentwise distance to singularity is addressed in [6].

Rohn [17] introduced the following componentwise relative condition number for the (i, j) element of the inverse:

$$c_{ij}(A) := \lim_{\epsilon\to 0+} \sup_{|\Delta A|\le\epsilon|A|} \frac{|(A + \Delta A)^{-1} - A^{-1}|_{ij}}{\epsilon|A^{-1}|_{ij}}. \tag{7.1}$$

Here $|A|$ means $(|a_{ij}|)$, and $A \leq B$ means $a_{ij} \leq b_{ij}$ for $1 \leq i, j \leq n$. We see that $c_{ij}(A)$ measures the worst-case relative change in $(A^{-1})_{ij}$ under small componentwise changes in A. The characterization

$$c_{ij}(A) = \frac{(|A^{-1}||A||A^{-1}|)_{ij}}{|A^{-1}|_{ij}} \tag{7.2}$$

was derived in [17]. Rohn also defined a componentwise condition number for the ith component of $Ax = b$ by

$$c_i(A, b) := \lim_{\epsilon \to 0+} \sup_{\substack{|\Delta A| \leq \epsilon |A| \\ |\Delta b| \leq \epsilon |b|}} \frac{|(A + \Delta A)^{-1}(b + \Delta b) - A^{-1}b|_i}{\epsilon |A^{-1}b|_i}, \tag{7.3}$$

and showed that

$$c_i(A, b) = \frac{(|A^{-1}||A||A^{-1}b| + |A^{-1}|\,|b|)_i}{|A^{-1}b|_i}. \tag{7.4}$$

The expressions (7.2) and (7.4) are valid when $(A^{-1})_{ij} \neq 0$ and $(A^{-1}b)_i \neq 0$, respectively, and we assume henceforth that these conditions hold.

Other componentwise condition numbers for a linear system have been put forward in the literature. Skeel [18] uses $\epsilon \|A^{-1}b\|_\infty$ rather than $\epsilon |A^{-1}b|_i$ in the right-hand side of (7.3), and in [10] componentwise perturbations relative to general tolerances, $|\Delta A| \leq \epsilon E$ and $|\Delta b| \leq \epsilon f$, are allowed. Altering the entries in E and f allows great flexibility, and in particular it is possible to mimic normwise measurements. For example, choosing $e_{ij} \equiv \|A\|_{\max}$ and $f_i \equiv \|b\|_\infty$ produces the normwise condition number $\mathrm{cond}_{\max}(A, b)$ in (4.4). The measure in (7.3), however, is the most appropriate for our analysis.

Overall componentwise relative condition numbers can be defined as

$$c_{\max}(A) := \max_{i,j}\{c_{ij}(A)\}, \qquad c_{\max}(A, b) := \max_i\{c_i(A, b)\}. \tag{7.5}$$

Note that because of the presence of the denominators $|A^{-1}|_{ij}$ and $|A^{-1}b|_i$ in (7.2) and (7.4), $c_{\max}(A)$ and $c_{\max}(A, b)$ can be arbitrarily larger than any given normwise condition number, and it is not possible to relate $c_{\max}(A)$ and $c_{\max}(A, b)$ via inequalities like (1.5).

The theorem below gives an upper bound on the level-2 condition numbers that correspond to $c_{ij}(A)$ and $c_i(A, b)$. Although it is clear from the proof that the bounds may be far from sharp, we do have the pleasing result that the level-2 condition numbers cannot be significantly larger than the level-1 condition numbers. As in the normwise case, allowing different perturbations ΔA to different factors in the characterizations (7.2) and (7.4) would not affect the results significantly.

THEOREM 7.1. *The level-2 condition numbers*

$$c_{ij}^{[2]}(A) := \lim_{\epsilon \to 0+} \sup_{|\Delta A| \le \epsilon |A|} \frac{|c_{ij}(A + \Delta A) - c_{ij}(A)|}{\epsilon c_{ij}(A)} \tag{7.6}$$

and

$$c_i^{[2]}(A, b) := \lim_{\epsilon \to 0+} \sup_{\substack{|\Delta A| \le \epsilon |A| \\ |\Delta b| \le \epsilon |b|}} \frac{|c_i(A + \Delta A, b + \Delta b) - c_i(A, b)|}{\epsilon c_i(A, b)} \tag{7.7}$$

satisfy

$$c_{ij}^{[2]}(A) \le 3c_{\max}(A) + 1$$

and

$$c_i^{[2]}(A, b) \le 3c_{\max}(A, b) + 2c_{\max}(A) + 2.$$

Proof. If $|\Delta A| \le \epsilon |A|$ then $(A + \Delta A)^{-1} = A^{-1} - A^{-1}\Delta A A^{-1} + O(\epsilon^2)$, and it follows that

$$\begin{aligned} |A^{-1}| - \epsilon |A^{-1}|\,|A|\,|A^{-1}| + O(\epsilon^2) &\le |(A + \Delta A)^{-1}| \\ &\le |A^{-1}| + \epsilon |A^{-1}|\,|A|\,|A^{-1}| + O(\epsilon^2). \end{aligned}$$

Hence, from (7.2),

$$\begin{aligned} |A^{-1}|_{ij}[1 - \epsilon c_{ij}(A)] + O(\epsilon^2) &\le |(A + \Delta A)^{-1}|_{ij} \\ &\le |A^{-1}|_{ij}[1 + \epsilon c_{ij}(A)] + O(\epsilon^2), \end{aligned}$$

which we weaken to

$$\begin{aligned} |A^{-1}|[1 - \epsilon c_{\max}(A)] + O(\epsilon^2) &\le |(A + \Delta A)^{-1}| \\ &\le |A^{-1}|[1 + \epsilon c_{\max}(A)] + O(\epsilon^2). \end{aligned} \tag{7.8}$$

Now from (7.8), and using $|A + \Delta A| \le |A|(1 + \epsilon)$,

$$\begin{aligned} &|(A + \Delta A)^{-1}|\,|A + \Delta A|\,|(A + \Delta A)^{-1}| \\ &\quad \le |A^{-1}|\,|A|\,|A^{-1}|[1 + \epsilon + 2\epsilon c_{\max}(A)] + O(\epsilon^2). \end{aligned}$$

Hence, again using (7.8),

$$\begin{aligned} &\frac{(|(A + \Delta A)^{-1}|\,|A + \Delta A|\,|(A + \Delta A)^{-1}|)_{ij}}{|(A + \Delta A)^{-1}|_{ij}} \\ &\quad \le \frac{(|A^{-1}|\,|A|\,|A^{-1}|)_{ij}}{|A^{-1}|_{ij}}[1 + \epsilon + 3\epsilon c_{\max}(A)] + O(\epsilon^2). \end{aligned} \tag{7.9}$$

Similarly, using (7.8) and $|A + \Delta A| \geq |A|(1 - \epsilon)$ it can be shown that

$$\frac{(|(A+\Delta A)^{-1}||A+\Delta A|(A+\Delta A)^{-1}|)_{ij}}{|(A+\Delta A)^{-1}|_{ij}} \geq \frac{(|A^{-1}||A||A^{-1}|)_{ij}}{|A^{-1}|_{ij}}(1 - \epsilon - 3\epsilon c_{\max}(A)) + O(\epsilon^2). \tag{7.10}$$

It follows from (7.9) and (7.10) that

$$c_{ij}^{[2]}(A) \leq 1 + 3c_{\max}(A). \tag{7.11}$$

Now, for the linear system, we have

$$|x|_i[1 - \epsilon c_{\max}(A, b)] + O(\epsilon^2) \leq |x + \Delta x|_i \leq |x|_i[1 + \epsilon c_{\max}(A, b)] + O(\epsilon^2).$$

Hence, using the result (7.8) for matrix inversion,

$$\frac{\left|\frac{\left(|(A+\Delta A)^{-1}||A+\Delta A||x+\Delta x|\right)_i}{|x+\Delta x|_i} - \frac{\left(|A^{-1}||A||x|\right)_i}{|x|_i}\right|}{\epsilon\frac{(|A^{-1}||A||x|)_i}{|x|_i}} \leq 1 + 2c_{\max}(A, b) + c_{\max}(A) + O(\epsilon), \tag{7.12}$$

and

$$\frac{\left|\frac{\left(|(A+\Delta A)^{-1}||b+\Delta b|\right)_i}{|x+\Delta x|_i} - \frac{\left(|A^{-1}||b|\right)_i}{|x|_i}\right|}{\epsilon\frac{(|A^{-1}||b|)_i}{|x|_i}} \leq 1 + c_{\max}(A, b) + c_{\max}(A) + O(\epsilon). \tag{7.13}$$

So, using (7.4) in (7.7), it follows from (7.12) and (7.13) that

$$c_i^{[2]}(A, b) \leq \lim_{\epsilon\to 0_+} \sup_{\substack{|\Delta A|\leq\epsilon|A| \\ |\Delta b|\leq\epsilon|b|}} \frac{\left|\frac{\left(|(A+\Delta A)^{-1}||A+\Delta A||x+\Delta x|\right)_i}{|x+\Delta x|_i} - \frac{\left(|A^{-1}||A||x|\right)_i}{|x|_i}\right|}{\epsilon\frac{(|A^{-1}||A||x|)_i}{|x|_i}}$$
$$+ \lim_{\epsilon\to 0_+} \sup_{\substack{|\Delta A|\leq\epsilon|A| \\ |\Delta b|\leq\epsilon|b|}} \frac{\left|\frac{\left(|(A+\Delta A)^{-1}||b+\Delta b|\right)_i}{|x+\Delta x|_i} - \frac{\left(|A^{-1}||b|\right)_i}{|x|_i}\right|}{\epsilon\frac{(|A^{-1}||b|)_i}{|x|_i}}$$
$$\leq 2 + 3c_{\max}(A, b) + 2c_{\max}(A). \qquad \blacksquare$$

I thank Nick Higham for suggesting perturbations to the original manuscript.

REFERENCES

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

2 S. G. Bartels, Two Topics in Matrix Analysis: Structured Sensitivity for Vandermonde-like Systems and a Subgradient Method for Matrix Norm Estimation, M.Sc. Thesis, Univ. of Dundee, 1991.

3 F. L. Bauer, Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme, *Z. Ang. New. Math. Mech.* 46:409–421 (1966).

4 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).

5 J. W. Demmel, On condition numbers and the distance to the nearest ill-posed problem, *Numer. Math.* 51:251–289 (1987).

6 J. W. Demmel, The componentwise distance to the nearest singular matrix, *SIAM J. Matrix Anal. Appl.* 13:10–19 (1992).

7 A. J. Geurts, A contribution to the theory of condition, *Numer. Math.* 39:85–96 (1982).

8 G. H. Golub and C. F. Van Loan, *Matrix Computations,* 2nd ed., Johns Hopkins U.P., Baltimore, 1989.

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

10 D. J. Higham and N. J. Higham, Backward error and condition of structured linear systems, *SIAM J. Matrix Anal. Appl.* 13:162–175 (1992).

11 D. J. Higham and N. J. Higham, Componentwise perturbation theory for linear systems with multiple right-hand sides, *Linear Algebra Appl.* 174:111–129 (1992).

12 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).

13 R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge U.P., 1985.

14 W. Kahan, Numerical linear algebra, *Canad. Math. Bull.* 9:757–801 (1966).

15 J. R. Rice, A theory of condition, *SIAM J. Numer. Anal.* 3:287–310 (1966).

16 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).

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

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

Received 10 June 1992; final manuscript accepted 22 March 1993