Chapter 2

Matrix Analysis

The analysis and derivation of algorithms in the matrix computation area requires a facility with certain aspects of linear algebra. Some of the basics are reviewed in §2.1. Norms and their manipulation are covered in §2.2 and §2.3. In §2.4 we develop a model of finite precision arithmetic and then use it in a typical roundoff analysis.

The next two sections deal with orthogonality, which has a prominent role to play in matrix computations. The singular value decomposition and the CS decomposition are a pair of orthogonal reductions that provide critical insight into the important notions of rank and distance between subspaces. In §2.7 we examine how the solution of a linear system $Ax = b$ changes if A and b are perturbed. The important concept of matrix condition is introduced.

Before You Begin

References that complement this chapter include Forsythe and Moler (1967), Stewart (1973), Stewart and Sun (1990), and Higham (1996).

2.1 Basic Ideas from Linear Algebra

This section is a quick review of linear algebra. Readers who wish a more detailed coverage should consult the references at the end of the section.

2.1.1 Independence, Subspace, Basis, and Dimension

A set of vectors $\{a_1, \ldots, a_n\}$ in $\mathbb{R}^m$ is *linearly independent* if $\sum_{j=1}^{n} \alpha_j a_j = 0$ implies $\alpha(1{:}n) = 0$. Otherwise, a nontrivial combination of the a_i is zero and $\{a_1, \ldots, a_n\}$ is said to be *linearly dependent*.

A *subspace* of $\mathbb{R}^m$ is a subset that is also a vector space. Given a collection of vectors $a_1, \ldots, a_n \in \mathbb{R}^m$, the set of all linear combinations of these vectors is a subspace referred to as the *span* of $\{a_1, \ldots, a_n\}$:

$$\text{span}\{a_1, \ldots, a_n\} = \left\{ \sum_{j=1}^{n} \beta_j a_j : \beta_j \in \mathbb{R} \right\}.$$

If $\{a_1, \ldots, a_n\}$ is independent and $b \in \text{span}\{a_1, \ldots, a_n\}$, then b is a unique linear combination of the a_j.

If $S_1, \ldots, S_k$ are subspaces of $\mathbb{R}^m$, then their sum is the subspace defined by $S = \{ a_1 + a_2 + \cdots + a_k : a_i \in S_i,\ i = 1{:}k \}$. S is said to be a *direct sum* if each $v \in S$ has a unique representation $v = a_1 + \cdots + a_k$ with $a_i \in S_i$. In this case we write $S = S_1 \oplus \cdots \oplus S_k$. The intersection of the S_i is also a subspace, $S = S_1 \cap S_2 \cap \cdots \cap S_k$.

The subset $\{a_{i_1}, \ldots, a_{i_k}\}$ is a *maximal linearly independent subset* of $\{a_1, \ldots, a_n\}$ if it is linearly independent and is not properly contained in any linearly independent subset of $\{a_1, \ldots, a_n\}$. If $\{a_{i_1}, \ldots, a_{i_k}\}$ is maximal, then $\text{span}\{a_1, \ldots, a_n\} = \text{span}\{a_{i_1}, \ldots, a_{i_k}\}$ and $\{a_{i_1}, \ldots, a_{i_k}\}$ is a *basis* for $\text{span}\{a_1, \ldots, a_n\}$. If $S \subseteq \mathbb{R}^m$ is a subspace, then it is possible to find independent basic vectors $a_1, \ldots, a_k \in S$ such that $S = \text{span}\{a_1, \ldots, a_k\}$. All bases for a subspace S have the same number of elements. This number is the *dimension* and is denoted by $\dim(S)$.

2.1.2 Range, Null Space, and Rank

There are two important subspaces associated with an m-by-n matrix A. The *range* of A is defined by

$$\text{ran}(A) = \{y \in \mathbb{R}^m : y = Ax \text{ for some } x \in \mathbb{R}^n\},$$

and the *null space* of A is defined by

$$\text{null}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

If $A = [\, a_1, \ldots, a_n \,]$ is a column partitioning, then

$$\text{ran}(A) = \text{span}\{a_1, \ldots, a_n\}.$$

The *rank* of a matrix A is defined by

$$\text{rank}(A) = \dim(\text{ran}(A)).$$

It can be shown that $\text{rank}(A) = \text{rank}(A^T)$. We say that $A \in \mathbb{R}^{m \times n}$ is *rank deficient* if $\text{rank}(A) < \min\{m, n\}$. If $A \in \mathbb{R}^{m \times n}$, then

$$\dim(\text{null}(A)) + \text{rank}(A) = n.$$

2.1.3 Matrix Inverse

The n-by-n *identity matrix* I_n is defined by the column partitioning

$$I_n = [\, e_1, \ldots, e_n \,]$$

where e_k is the kth "canonical" vector:

$$e_k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k})^T.$$

The canonical vectors arise frequently in matrix analysis and if their dimension is ever ambiguous, we use superscripts, i.e., $e_k^{(n)} \in \mathbb{R}^n$.

If A and X are in $\mathbb{R}^{n\times n}$ and satisfy $AX = I$, then X is the *inverse* of A and is denoted by A^{-1}. If A^{-1} exists, then A is said to be *nonsingular*. Otherwise, we say A is *singular*.

Several matrix inverse properties have an important role to play in matrix computations. The inverse of a product is the reverse product of the inverses:

$$(AB)^{-1} = B^{-1}A^{-1}. \tag{2.1.1}$$

The transpose of the inverse is the inverse of the transpose:

$$(A^{-1})^T = (A^T)^{-1} \equiv A^{-T}. \tag{2.1.2}$$

The identity

$$B^{-1} = A^{-1} - B^{-1}(B - A)A^{-1} \tag{2.1.3}$$

shows how the inverse changes if the matrix changes.

The *Sherman-Morrison-Woodbury formula* gives a convenient expression for the inverse of $(A+UV^T)$ where $A \in \mathbb{R}^{n\times n}$ and U and V are n-by-k:

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^TA^{-1}U)^{-1}V^TA^{-1}. \tag{2.1.4}$$

A rank k correction to a matrix results in a rank k correction of the inverse. In (2.1.4) we assume that both A and $(I + V^TA^{-1}U)$ are nonsingular.

Any of these facts can be verified by just showing that the "proposed" inverse does the job. For example, here is how to confirm (2.1.3):

$$B\left(A^{-1} - B^{-1}(B - A)A^{-1}\right) = BA^{-1} - (B - A)A^{-1} = I.$$

2.1.4 The Determinant

If $A = (a) \in \mathbb{R}^{1\times 1}$, then its *determinant* is given by $\det(A) = a$. The determinant of $A \in \mathbb{R}^{n\times n}$ is defined in terms of order $n-1$ determinants:

$$\det(A) = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} \det(A_{1j}).$$

Here, A_{1j} is an $(n-1)$-by-$(n-1)$ matrix obtained by deleting the first row and jth column of A. Useful properties of the determinant include

$$\begin{array}{lcll} \det(AB) & = & \det(A)\det(B) & A, B \in \mathbb{R}^{n\times n} \\ \det(A^T) & = & \det(A) & A \in \mathbb{R}^{n\times n} \\ \det(cA) & = & c^n\det(A) & c \in \mathbb{R}, A \in \mathbb{R}^{n\times n} \\ \det(A) \neq 0 & \Leftrightarrow & A \text{ is nonsingular} & A \in \mathbb{R}^{n\times n} \end{array}$$

2.1.5 Differentiation

Suppose α is a scalar and that $A(\alpha)$ is an m-by-n matrix with entries $a_{ij}(\alpha)$. If $a_{ij}(\alpha)$ is a differentiable function of α for all i and j, then by $\dot{A}(\alpha)$ we mean the matrix

$$\dot{A}(\alpha) = \frac{d}{d\alpha}A(\alpha) = \left(\frac{d}{d\alpha}a_{ij}(\alpha)\right) = (\dot{a}_{ij}(\alpha)) .$$

The differentiation of a parameterized matrix turns out to be a handy way to examine the sensitivity of various matrix problems.

Problems

P2.1.1 Show that if $A \in \mathbb{R}^{m\times n}$ has rank p, then there exists an $X \in \mathbb{R}^{m\times p}$ and a $Y \in \mathbb{R}^{n\times p}$ such that $A = XY^T$, where $\text{rank}(X) = \text{rank}(Y) = p$.

P2.1.2 Suppose $A(\alpha) \in \mathbb{R}^{m\times r}$ and $B(\alpha) \in \mathbb{R}^{r\times n}$ are matrices whose entries are differentiable functions of the scalar α. Show

$$\frac{d}{d\alpha}[A(\alpha)B(\alpha)] = \left[\frac{d}{d\alpha}A(\alpha)\right]B(\alpha) + A(\alpha)\left[\frac{d}{d\alpha}B(\alpha)\right] .$$

P2.1.3 Suppose $A(\alpha) \in \mathbb{R}^{n\times n}$ has entries that are differentiable functions of the scalar α. Assuming $A(\alpha)$ is always nonsingular, show

$$\frac{d}{d\alpha}\left[A(\alpha)^{-1}\right] = -A(\alpha)^{-1}\left[\frac{d}{d\alpha}A(\alpha)\right]A(\alpha)^{-1}.$$

P2.1.4 Suppose $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and that $\phi(x) = \frac{1}{2}x^TAx - x^Tb$. Show that the gradient of ϕ is given by $\nabla\phi(x) = \frac{1}{2}(A^T + A)x - b$.

P2.1.5 Assume that both A and $A+uv^T$ are nonsingular where $A \in \mathbb{R}^{n\times n}$ and $u, v \in \mathbb{R}^n$. Show that if x solves $(A + uv^T)x = b$, then it also solves a perturbed right hand side problem of the form $Ax = b + \alpha u$. Give an expression for α in terms of A, u, and v.

Notes and References for Sec. 2.1

There are many introductory linear algebra texts. Among them, the following are particularly useful:

P.R. Halmos (1958). *Finite Dimensional Vector Spaces*, 2nd ed., Van Nostrand-Reinhold, Princeton.

S.J. Leon (1980). *Linear Algebra with Applications*. Macmillan, New York.
G. Strang (1993). *Introduction to Linear Algebra*, Wellesley-Cambridge Press, Wellesley MA.
D. Lay (1994). *Linear Algebra and Its Applications*, Addison-Wesley, Reading, MA.
C. Meyer (1997). *A Course in Applied Linear Algebra*, SIAM Publications, Philadelphia, PA.

More advanced treatments include Gantmacher (1959), Horn and Johnson (1985, 1991), and

A.S. Householder (1964). *The Theory of Matrices in Numerical Analysis*, Ginn (Blaisdell), Boston.
M. Marcus and H. Minc (1964). *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston.
J.N. Franklin (1968). *Matrix Theory* Prentice Hall, Englewood Cliffs, NJ.
R. Bellman (1970). *Introduction to Matrix Analysis, Second Edition*, McGraw-Hill, New York.
P. Lancaster and M. Tismenetsky (1985). *The Theory of Matrices, Second Edition*, Academic Press, New York.
J.M. Ortega (1987). *Matrix Theory: A Second Course*, Plenum Press, New York.

2.2 Vector Norms

Norms serve the same purpose on vector spaces that absolute value does on the real line: they furnish a measure of distance. More precisely, $\mathbb{R}^n$ together with a norm on $\mathbb{R}^n$ defines a metric space. Therefore, we have the familiar notions of neighborhood, open sets, convergence, and continuity when working with vectors and vector-valued functions.

2.2.1 Definitions

A *vector norm* on $\mathbb{R}^n$ is a function $f:\mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties:

$$\begin{array}{lll} f(x) \geq 0 & x \in \mathbb{R}^n, & (f(x) = 0 \text{ iff } x = 0) \\ f(x+y) \leq f(x) + f(y) & x, y \in \mathbb{R}^n & \\ f(\alpha x) = |\alpha| f(x) & \alpha \in \mathbb{R}, x \in \mathbb{R}^n & \end{array}$$

We denote such a function with a double bar notation: $f(x) = \| x \|$. Subscripts on the double bar are used to distinguish between various norms.

A useful class of vector norms are the *p-norms* defined by

$$\| x \|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}} \qquad p \geq 1 . \tag{2.2.1}$$

Of these the 1, 2, and ∞ norms are the most important:

$$\begin{aligned} \| x \|_1 &= |x_1| + \cdots + |x_n| \\ \| x \|_2 &= \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}} = \left(x^T x\right)^{\frac{1}{2}} \\ \| x \|_\infty &= \max_{1 \leq i \leq n} |x_i| \end{aligned}$$

A *unit vector* with respect to the norm $\|\cdot\|$ is a vector x that satisfies $\|x\| = 1$.

2.2.2 Some Vector Norm Properties

A classic result concerning p-norms is the *Holder inequality*:

$$|x^T y| \leq \|x\|_p \|y\|_q \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{2.2.2}$$

A very important special case of this is the *Cauchy-Schwartz inequality*:

$$|x^T y| \leq \|x\|_2 \|y\|_2. \tag{2.2.3}$$

All norms on $\mathbb{R}^n$ are *equivalent*, i.e., if $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are norms on $\mathbb{R}^n$, then there exist positive constants, c_1 and c_2 such that

$$c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha \tag{2.2.4}$$

for all $x \in \mathbb{R}^n$. For example, if $x \in \mathbb{R}^n$, then

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2 \tag{2.2.5}$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty \tag{2.2.6}$$

$$\|x\|_\infty \leq \|x\|_1 \leq n\,\|x\|_\infty. \tag{2.2.7}$$

2.2.3 Absolute and Relative Error

Suppose $\hat{x} \in \mathbb{R}^n$ is an approximation to $x \in \mathbb{R}^n$. For a given vector norm $\|\cdot\|$ we say that

$$\epsilon_{abs} = \|\hat{x} - x\|$$

is the *absolute error* in $\hat{x}$. If $x \neq 0$, then

$$\epsilon_{rel} = \frac{\|\hat{x} - x\|}{\|x\|}$$

prescribes the *relative error* in $\hat{x}$. Relative error in the ∞-norm can be translated into a statement about the number of correct significant digits in $\hat{x}$. In particular, if

$$\frac{\|\hat{x} - x\|_\infty}{\|x\|_\infty} \approx 10^{-p},$$

then the largest component of $\hat{x}$ has approximately p correct significant digits.

Example 2.2.1 If $x = (1.234 \;\; .05674)^T$ and $\hat{x} = (1.235 \;\; .05128)^T$, then $\|\hat{x} - x\|_\infty / \|x\|_\infty \approx .0043 \approx 10^{-3}$. Note than $\hat{x}_1$ has about three significant digits that are correct while only one significant digit in $\hat{x}_2$ is correct.

2.2.4 Convergence

We say that a sequence $\{x^{(k)}\}$ of n-vectors *converges* to x if

$$\lim_{k \rightarrow \infty} \| x^{(k)} - x \| = 0 .$$

Note that because of (2.2.4), convergence in the α-norm implies convergence in the β-norm and vice versa.

Problems

P2.2.1 Show that if $x \in \mathbb{R}^n$, then $\lim_{p \rightarrow \infty} \| x \|_p = \| x \|_\infty$.

P2.2.2 Prove the Cauchy-Schwartz inequality (2.2.3) by considering the inequality $0 \leq (ax + by)^T (ax + by)$ for suitable scalars a and b.

P2.2.3 Verify that $\| \cdot \|_1$, $\| \cdot \|_2$, and $\| \cdot \|_\infty$ are vector norms.

P2.2.4 Verify (2.2.5)-(2.2.7). When is equality achieved in each result?

P2.2.5 Show that in $\mathbb{R}^n$, $x^{(i)} \rightarrow x$ if and only if $x_k^{(i)} \rightarrow x_k$ for $k = 1{:}n$.

P2.2.6 Show that any vector norm on $\mathbb{R}^n$ is uniformly continuous by verifying the inequality $| \| x \| - \| y \| | \leq \| x - y \|$.

P2.2.7 Let $\| \cdot \|$ be a vector norm on $\mathbb{R}^m$ and assume $A \in \mathbb{R}^{m \times n}$. Show that if $\text{rank}(A) = n$, then $\| x \|_A = \| Ax \|$ is a vector norm on $\mathbb{R}^n$.

P2.2.8 Let x and y be in $\mathbb{R}^n$ and define $\psi: \mathbb{R} \rightarrow \mathbb{R}$ by $\psi(\alpha) = \| x - \alpha y \|_2$. Show that ψ is minimized when $\alpha = x^T y / y^T y$.

P2.2.9 (a) Verify that $\| x \|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$ is a vector norm on $\mathbb{C}^n$. (b) Show that if $x \in \mathbb{C}^n$ then $\| x \|_p \leq c \left(\| \text{Re}(x) \|_p + \| \text{Im}(x) \|_p \right)$. (c) Find a constant c_n such that $c_n (\| \text{Re}(x) \|_2 + \| \text{Im}(x) \|_2) \leq \| x \|_2$ for all $x \in \mathbb{C}^n$.

P2.2.10 Prove or disprove:

$$v \in \mathbb{R}^n \Rightarrow \| v \|_1 \| v \|_\infty \leq \frac{1 + \sqrt{n}}{2} \| v \|_2 .$$

Notes and References for Sec. 2.2

Although a vector norm is "just" a generalization of the absolute value concept, there are some noteworthy subtleties:

J.D. Pryce (1984). "A New Measure of Relative Error for Vectors," *SIAM J. Num. Anal. 21*, 202-21.

2.3 Matrix Norms

The analysis of matrix algorithms frequently requires use of matrix norms. For example, the quality of a linear system solver may be poor if the matrix of coefficients is "nearly singular." To quantify the notion of near-singularity we need a measure of distance on the space of matrices. Matrix norms provide that measure.

2.3.1 Definitions

Since $\mathbb{R}^{m\times n}$ is isomorphic to $\mathbb{R}^{mn}$, the definition of a matrix norm should be equivalent to the definition of a vector norm. In particular, $f:\mathbb{R}^{m\times n} \rightarrow \mathbb{R}$ is a matrix norm if the following three properties hold:

$$\begin{array}{lll} f(A) \geq 0 & A \in \mathbb{R}^{m\times n}, & (f(A) = 0 \text{ iff } A = 0) \\ f(A+B) \leq f(A) + f(B) & A, B \in \mathbb{R}^{m\times n}, & \\ f(\alpha A) = |\alpha| f(A) & \alpha \in \mathbb{R}, A \in \mathbb{R}^{m\times n}. & \end{array}$$

As with vector norms, we use a double bar notation with subscripts to designate matrix norms, i.e., $\| A \| = f(A)$.

The most frequently used matrix norms in numerical linear algebra are the Frobenius norm,

$$\| A \|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} \tag{2.3.1}$$

and the p-norms

$$\| A \|_p = \sup_{x\neq 0} \frac{\| Ax \|_p}{\| x \|_p}. \tag{2.3.2}$$

Note that the matrix p-norms are defined in terms of the vector p-norms that we discussed in the previous section. The verification that (2.3.1) and (2.3.2) are matrix norms is left as an exercise. It is clear that $\| A \|_p$ is the p-norm of the largest vector obtained by applying A to a unit p-norm vector:

$$\| A \|_p = \sup_{x\neq 0} \left\| A\left(\frac{x}{\| x \|_p}\right) \right\|_p = \max_{\|x\|_p = 1} \| Ax \|_p .$$

It is important to understand that (2.3.1) and (2.3.2) define families of norms—the 2-norm on $\mathbb{R}^{3\times 2}$ is a different function from the 2-norm on $\mathbb{R}^{5\times 6}$. Thus, the easily verified inequality

$$\| AB \|_p \leq \| A \|_p \| B \|_p \qquad A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times q} \tag{2.3.3}$$

is really an observation about the relationship between three different norms. Formally, we say that norms f_1, f_2, and f_3 on $\mathbb{R}^{m\times q}$, $\mathbb{R}^{m\times n}$, and $\mathbb{R}^{n\times q}$ are *mutually consistent* if for all $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times q}$ we have $f_1(AB) \leq f_2(A) f_3(B)$.

Not all matrix norms satisfy the submultiplicative property

$$\| AB \| \leq \| A \| \| B \|. \tag{2.3.4}$$

For example, if $\| A \|_\Delta = \max |a_{ij}|$ and

$$A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

then $\| AB \|_\Delta > \| A \|_\Delta \| B \|_\Delta$. For the most part we work with norms that satisfy (2.3.4).

The p-norms have the important property that for every $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$ we have $\| Ax \|_p \leq \| A \|_p \| x \|_p$. More generally, for any vector norm $\| \cdot \|_\alpha$ on $\mathbb{R}^n$ and $\| \cdot \|_\beta$ on $\mathbb{R}^m$ we have $\| Ax \|_\beta \leq \| A \|_{\alpha,\beta} \| x \|_\alpha$ where $\| A \|_{\alpha,\beta}$ is a matrix norm defined by

$$\| A \|_{\alpha,\beta} = \sup_{x \neq 0} \frac{\| Ax \|_\beta}{\| x \|_\alpha}. \tag{2.3.5}$$

We say that $\| \cdot \|_{\alpha,\beta}$ is *subordinate* to the vector norms $\| \cdot \|_\alpha$ and $\| \cdot \|_\beta$. Since the set $\{x \in \mathbb{R}^n : \| x \|_\alpha = 1\}$ is compact and $\| \cdot \|_\beta$ is continuous, it follows that

$$\| A \|_{\alpha,\beta} = \max_{\|x\|_\alpha = 1} \| Ax \|_\beta = \| Ax^* \|_\beta \tag{2.3.6}$$

for some $x^* \in \mathbb{R}^n$ having unit α-norm.

2.3.2 Some Matrix Norm Properties

The Frobenius and p-norms (especially $p = 1, 2, \infty$) satisfy certain inequalities that are frequently used in the analysis of matrix computations. For $A \in \mathbb{R}^{m \times n}$ we have

$$\| A \|_2 \leq \| A \|_F \leq \sqrt{n} \| A \|_2 \tag{2.3.7}$$

$$\max_{i,j} |a_{ij}| \leq \| A \|_2 \leq \sqrt{mn} \max_{i,j} |a_{ij}| \tag{2.3.8}$$

$$\| A \|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}| \tag{2.3.9}$$

$$\| A \|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}| \tag{2.3.10}$$

$$\frac{1}{\sqrt{n}} \| A \|_\infty \leq \| A \|_2 \leq \sqrt{m} \| A \|_\infty \tag{2.3.11}$$

$$\frac{1}{\sqrt{m}} \| A \|_1 \leq \| A \|_2 \leq \sqrt{n} \| A \|_1 \tag{2.3.12}$$

If $A \in \mathbb{R}^{m\times n}$, $1 \le i_1 \le i_2 \le m$, and $1 \le j_1 \le j_2 \le n$, then

$$\| A(i_1:i_2, j_1:j_2) \|_p \le \| A \|_p \tag{2.3.13}$$

The proofs of these relations are not hard and are left as exercises.

A sequence $\{A^{(k)}\} \in \mathbb{R}^{m\times n}$ *converges* if $\lim_{k\to\infty} \| A^{(k)} - A \| = 0$. Choice of norm is irrelevant since all norms on $\mathbb{R}^{m\times n}$ are equivalent.

2.3.3 The Matrix 2-Norm

A nice feature of the matrix 1-norm and the matrix ∞-norm is that they are easily computed from (2.3.9) and (2.3.10). A characterization of the 2-norm is considerably more complicated.

Theorem 2.3.1 *If $A \in \mathbb{R}^{m\times n}$, then there exists a unit 2-norm n-vector z such that $A^TAz = \mu^2 z$ where $\mu = \| A \|_2$.*

Proof. Suppose $z \in \mathbb{R}^n$ is a unit vector such that $\| Az \|_2 = \| A \|_2$. Since z maximizes the function

$$g(x) = \frac{1}{2}\frac{\| Ax \|_2^2}{\| x \|_2^2} = \frac{1}{2}\frac{x^TA^TAx}{x^Tx}$$

it follows that it satisfies $\nabla g(z) = 0$ where ∇g is the gradient of g. But a tedious differentiation shows that for $i = 1{:}n$

$$\frac{\partial g(z)}{\partial z_i} = \left[(z^Tz)\sum_{j=1}^{n}(A^TA)_{ij}z_j - (z^TA^TAz)z_i \right] \Big/ (z^Tz)^2 .$$

In vector notation this says $A^TAz = (z^TA^TAz)z$. The theorem follows by setting $\mu = \| Az \|_2$. $\square$

The theorem implies that $\| A \|_2^2$ is a zero of the polynomial $p(\lambda) = \det(A^TA - \lambda I)$. In particular, the 2-norm of A is the square root of the largest eigenvalue of A^TA. We have much more to say about eigenvalues in Chapters 7 and 8. For now, we merely observe that 2-norm computation is iterative and decidedly more complicated than the computation of the matrix 1-norm or ∞-norm. Fortunately, if the object is to obtain an order-of-magnitude estimate of $\| A \|_2$, then (2.3.7), (2.3.11), or (2.3.12) can be used.

As another example of "norm analysis," here is a handy result for 2-norm estimation.

Corollary 2.3.2 *If $A \in \mathbb{R}^{m\times n}$, then $\| A \|_2 \le \sqrt{\| A \|_1 \| A \|_\infty}$.*

Proof. If $z \neq 0$ is such that $A^TAz = \mu^2 z$ with $\mu = \| A \|_2$, then $\mu^2 \| z \|_1 = \| A^TAz \|_1 \le \| A^T \|_1 \| A \|_1 \| z \|_1 = \| A \|_\infty \| A \|_1 \| z \|_1$. $\square$

2.3.4 Perturbations and the Inverse

We frequently use norms to quantify the effect of perturbations or to prove that a sequence of matrices converges to a specified limit. As an illustration of these norm applications, let us quantify the change in A^{-1} as a function of change in A.

Lemma 2.3.3 *If $F \in \mathbb{R}^{n\times n}$ and $\| F \|_p < 1$, then $I - F$ is nonsingular and*

$$(I - F)^{-1} = \sum_{k=0}^{\infty} F^k$$

with

$$\| (I - F)^{-1} \|_p \leq \frac{1}{1 - \| F \|_p}.$$

Proof. Suppose $I - F$ is singular. It follows that $(I - F)x = 0$ for some nonzero x. But then $\| x \|_p = \| Fx \|_p$ implies $\| F \|_p \geq 1$, a contradiction. Thus, $I - F$ is nonsingular. To obtain an expression for its inverse consider the identity

$$\left(\sum_{k=0}^{N} F^k \right) (I - F) = I - F^{N+1}.$$

Since $\| F \|_p < 1$ it follows that $\lim_{k\to\infty} F^k = 0$ because $\| F^k \|_p \leq \| F \|_p^k$. Thus,

$$\left(\lim_{N\to\infty} \sum_{k=0}^{N} F^k \right) (I - F) = I.$$

It follows that $(I - F)^{-1} = \lim_{N\to\infty} \sum_{k=0}^{N} F^k$. From this it is easy to show that

$$\| (I - F)^{-1} \|_p \leq \sum_{k=0}^{\infty} \| F \|_p^k = \frac{1}{1 - \| F \|_p}. \quad \square$$

Note that $\| (I - F)^{-1} - I \|_p \leq \| F \|_p/(1 - \| F \|_p)$ as a consequence of the lemma. Thus, if $\epsilon \ll 1$, then $O(\epsilon)$ perturbations in I induce $O(\epsilon)$ perturbations in the inverse. We next extend this result to general matrices.

Theorem 2.3.4 *If A is nonsingular and $r \equiv \| A^{-1}E \|_p < 1$, then $A + E$ is nonsingular and $\| (A + E)^{-1} - A^{-1} \|_p \leq \| E \|_p \| A^{-1} \|_p^2/(1 - r)$.*

Proof. Since A is nonsingular $A + E = A(I - F)$ where $F = -A^{-1}E$. Since $\| F \|_p = r < 1$ it follows from Lemma 2.3.3 that $I - F$ is nonsingular and $\| (I - F)^{-1} \|_p < 1/(1 - r)$. Now $(A + E)^{-1} = (I - F)^{-1}A^{-1}$ and so

$$\| (A + E)^{-1} \|_p \leq \frac{\| A^{-1} \|_p}{1 - r}.$$

Equation (2.1.3) says that $(A+E)^{-1} - A^{-1} = -A^{-1}E(A+E)^{-1}$ and so by taking norms we find

$$\begin{aligned} \| (A+E)^{-1} - A^{-1} \|_p &\leq \| A^{-1} \|_p \| E \|_p \| (A+E)^{-1} \|_p \\ &\leq \frac{\| A^{-1} \|_p^2 \| E \|_p}{1-r}. \quad \square \end{aligned}$$

Problems

P2.3.1 Show $\| AB \|_p \leq \| A \|_p \| B \|_p$ where $1 \leq p \leq \infty$.

P2.3.2 Let B be any submatrix of A. Show that $\| B \|_p \leq \| A \|_p$.

P2.3.3 Show that if $D = \text{diag}(\mu_1, \ldots, \mu_k) \in \mathbb{R}^{m \times n}$ with $k = \min\{m, n\}$, then $\| D \|_p = \max |\mu_i|$.

P2.3.4 Verify (2.3.7) and (2.3.8).

P2.3.5 Verify (2.3.9) and (2.3.10).

P2.3.6 Verify (2.3.11) and (2.3.12).

P2.3.7 Verify (2.3.13).

P2.3.8 Show that if $0 \neq s \in \mathbb{R}^n$ and $E \in \mathbb{R}^{m \times n}$, then

$$\left\| E\left(I - \frac{ss^T}{s^Ts}\right) \right\|_F^2 = \| E \|_F^2 - \frac{\| Es \|_2^2}{s^Ts}.$$

P2.3.9 Suppose $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$. Show that if $E = uv^T$ then $\| E \|_F = \| E \|_2 = \| u \|_2 \| v \|_2$ and that $\| E \|_\infty \leq \| u \|_\infty \| v \|_1$.

P2.3.10 Suppose $A \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, and $0 \neq s \in \mathbb{R}^n$. Show that $E = (y - As)s^T/s^Ts$ has the smallest 2-norm of all m-by-n matrices E that satisfy $(A+E)s = y$.

Notes and References for Sec. 2.3

For deeper issues concerning matrix/vector norms, see

F.L. Bauer and C.T. Fike (1960). "Norms and Exclusion Theorems," *Numer. Math. 2*, 137-44.

L. Mirsky (1960). "Symmetric Gauge Functions and Unitarily Invariant Norms," *Quart. J. Math. 11*, 50-59.

A.S. Householder (1964). *The Theory of Matrices in Numerical Analysis*, Dover Publications, New York.

N.J. Higham (1992). "Estimating the Matrix p-Norm," *Numer. Math. 62*, 539–556.

2.4 Finite Precision Matrix Computations

In part, rounding errors are what makes the matrix computation area so nontrivial and interesting. In this section we set up a model of floating point arithmetic and then use it to develop error bounds for floating point dot products, saxpy's, matrix-vector products and matrix-matrix products. For

a more comprehensive treatment than what we offer, see Higham (1996) or Wilkinson (1965). The coverage in Forsythe and Moler (1967) and Stewart (1973) is also excellent.

2.4.1 The Floating Point Numbers

When calculations are performed on a computer, each arithmetic operation is generally affected by *roundoff error*. This error arises because the machine hardware can only represent a subset of the real numbers. We denote this subset by $\mathbf{F}$ and refer to its elements as *floating point numbers*. Following conventions set forth in Forsythe, Malcolm, and Moler (1977, pp. 10-29), the floating point number system on a particular computer is characterized by four integers: the *base* β, the *precision* t, and the *exponent range* $[L, U]$. In particular, $\mathbf{F}$ consists of all numbers f of the form

$$f = \pm .d_1 d_2 \ldots d_t \times \beta^e \qquad 0 \leq d_i < \beta, \quad d_1 \neq 0, \quad L \leq e \leq U$$

together with zero. Notice that for a nonzero $f \in \mathbf{F}$ we have $m \leq |f| \leq M$ where

$$m = \beta^{L-1} \quad \text{and} \quad M = \beta^U(1 - \beta^{-t}). \tag{2.4.1}$$

As an example, if $\beta = 2$, $t = 3$, $L = 0$, and $U = 2$, then the non-negative elements of $\mathbf{F}$ are represented by hash marks on the axis displayed in FIG. 2.4.1. Notice that the floating point numbers are not equally spaced. A

FIGURE 2.4.1 *Sample Floating Point Number System*

typical value for (β, t, L, U) might be (2, 56, -64, 64).

2.4.2 A Model of Floating Point Arithmetic

To make general pronouncements about the effect of rounding errors on a given algorithm, it is necessary to have a model of computer arithmetic on $\mathbf{F}$. To this end define the set G by

$$G = \{\, x \in \mathbb{R} : m \leq |x| \leq M \,\} \cup \{0\} \tag{2.4.2}$$

and the operator fl: $G \rightarrow \mathbf{F}$ by

$$fl(x) = \begin{Bmatrix} \text{nearest } c \in \mathbf{F} \text{ to } x \text{ with ties handled} \\ \text{by rounding away from zero.} \end{Bmatrix}$$

The fl operator can be shown to satisfy

$$fl(x) = x(1+\epsilon) \qquad |\epsilon| \leq \mathbf{u} \tag{2.4.3}$$

where $\mathbf{u}$ is the *unit roundoff* defined by

$$\mathbf{u} = \frac{1}{2}\beta^{1-t}. \tag{2.4.4}$$

Let a and b be any two floating point numbers and let "op" denote any of the four arithmetic operations $+, -, \times, \div$. *If a op $b \in G$, then in our model of floating point arithmetic we assume that the computed version of $(a$ op $b)$ is given by $fl(a$ op $b)$.* It follows that $fl(a \text{ op } b) = (a \text{ op } b)(1+\epsilon)$ with $|\epsilon| \leq \mathbf{u}$. Thus,

$$\frac{|fl(a \text{ op } b) - (a \text{ op } b)|}{|a \text{ op } b|} \leq \mathbf{u} \qquad a \text{ op } b \neq 0 \tag{2.4.5}$$

showing that there is small relative error associated with individual arithmetic operations[1]. It is important to realize, however, that this is not necessarily the case when a sequence of operations is involved.

Example 2.4.1 If $\beta = 10$, $t = 3$ floating point arithmetic is used, then it can be shown that $fl[fl(10^{-4} + 1) - 1] = 0$ implying a relative error of 1. On the other hand the exact answer is given by $fl[fl(10^{-4} + fl(1-1)] = 10^{-4}$. Floating point arithmetic is not always associative.

If a op $b \notin G$, then an *arithmetic exception* occurs. *Overflow* and *underflow* results whenever $|a \text{ op } b| > M$ or $0 < |a \text{ op } b| < m$ respectively. The handling of these and other exceptions is hardware/system dependent.

2.4.3 Cancellation

Another important aspect of finite precision arithmetic is the phenomenon of *catastrophic cancellation*. Roughly speaking, this term refers to the extreme loss of correct significant digits when small numbers are additively computed from large numbers. A well-known example taken from Forsythe, Malcolm and Moler (1977, pp. 14-16) is the computation of e^{-a} via Taylor series with $a > 0$. The roundoff error associated with this method is

[1] There are important examples of machines whose additive floating point operations satisfy $fl(a \pm b) = (1+\epsilon_1)a \pm (1+\epsilon_2)b$ where $|\epsilon_1|, |\epsilon_2| \leq \mathbf{u}$. In such an environment, the inequality $|fl(a \pm b) - (a \pm b)| \leq \mathbf{u}|a \pm b|$ need not hold.

approximately u times the largest partial sum. For large a, this error can actually be larger than the exact exponential and there will be no correct digits in the answer no matter how many terms in the series are summed. On the other hand, if enough terms in the Taylor series for e^a are added and the result reciprocated, then an estimate of e^{-a} to full precision is attained.

2.4.4 The Absolute Value Notation

Before we proceed with the roundoff analysis of some basic matrix calculations, we acquire some useful notation. Suppose $A \in \mathbb{R}^{m\times n}$ and that we wish to quantify the errors associated with its floating point representation. Denoting the stored version of A by $fl(A)$, we see that

$$[fl(A)]_{ij} = fl(a_{ij}) = a_{ij}(1+\epsilon_{ij}) \qquad |\epsilon_{ij}| \leq \mathbf{u} \tag{2.4.6}$$

for all i and j. A better way to say the same thing results if we adopt two conventions. If A and B are in $\mathbb{R}^{m\times n}$, then

$$B = |A| \quad \Rightarrow \quad b_{ij} = |a_{ij}|,\ i = 1{:}m,\ j = 1{:}n$$

$$B \leq A \quad \Rightarrow \quad b_{ij} \leq a_{ij},\ i = 1{:}m,\ j = 1{:}n\,.$$

With this notation we see that (2.4.6) has the form

$$|fl(A) - A| \leq \mathbf{u}|A|\,.$$

A relation such as this can be easily turned into a norm inequality, e.g., $\| fl(A) - A \|_1 \leq \mathbf{u}\| A \|_1$. However, when quantifying the rounding errors in a matrix manipulation, the absolute value notation can be a lot more informative because it provides a comment on each (i,j) entry.

2.4.5 Roundoff in Dot Products

We begin our study of finite precision matrix computations by considering the rounding errors that result in the standard dot product algorithm:

$$\begin{array}{l} s = 0 \\ \textbf{for } k = 1{:}n \\ \qquad s = s + x_k y_k \\ \textbf{end} \end{array} \tag{2.4.7}$$

Here, x and y are n-by-1 floating point vectors.

In trying to quantify the rounding errors in this algorithm, we are immediately confronted with a notational problem: the distinction between computed and exact quantities. When the underlying computations are clear, we shall use the $fl(\cdot)$ operator to signify computed quantities.

Thus, $fl(x^Ty)$ denotes the computed output of (2.4.7). Let us bound $|fl(x^Ty) - x^Ty|$. If

$$s_p = fl\left(\sum_{k=1}^{p} x_k y_k\right),$$

then $s_1 = x_1y_1(1+\delta_1)$ with $|\delta_1| \leq \mathbf{u}$ and for $p = 2{:}n$

$$\begin{aligned} s_p &= fl(s_{p-1} + fl(x_py_p)) \\ &= (s_{p-1} + x_py_p(1+\delta_p))(1+\epsilon_p) \qquad |\delta_p|, |\epsilon_p| \leq \mathbf{u}\,. \end{aligned} \tag{2.4.8}$$

A little algebra shows that

$$fl(x^Ty) = s_n = \sum_{k=1}^{n} x_k y_k (1+\gamma_k)$$

where

$$(1+\gamma_k) = (1+\delta_k)\prod_{j=k}^{n}(1+\epsilon_j)$$

with the convention that $\epsilon_1 = 0$. Thus,

$$|fl(x^Ty) - x^Ty| \leq \sum_{k=1}^{n} |x_k y_k||\gamma_k|. \tag{2.4.9}$$

To proceed further, we must bound the quantities $|\gamma_k|$ in terms of $\mathbf{u}$. The following result is useful for this purpose.

Lemma 2.4.1 *If* $(1+\alpha) = \prod_{k=1}^{n}(1+\alpha_k)$ *where* $|\alpha_k| \leq \mathbf{u}$ *and* $n\mathbf{u} \leq .01$, *then* $|\alpha| \leq 1.01n\mathbf{u}$.

Proof. See Higham (1996, p. 75). □

Applying this result to (2.4.9) under the "reasonable" assumption $n\mathbf{u} \leq .01$ gives

$$|fl(x^Ty) - x^Ty| \leq 1.01n\mathbf{u}|x|^T|y|. \tag{2.4.10}$$

Notice that if $|x^Ty| \ll |x|^T|y|$, then the relative error in $fl(x^Ty)$ may not be small.

2.4.6 Alternative Ways to Quantify Roundoff Error

An easier but less rigorous way of bounding α in Lemma 2.4.1 is to say $|\alpha| \leq n\mathbf{u} + O(\mathbf{u}^2)$. With this convention we have

$$|fl(x^Ty) - x^Ty| \leq n\mathbf{u}|x|^T|y| + O(\mathbf{u}^2). \tag{2.4.11}$$

Other ways of expressing the same result include

$$|fl(x^T y) - x^T y| \leq \phi(n)\mathrm{u}|x|^T|y| \tag{2.4.12}$$

and

$$|fl(x^T y) - x^T y| \leq cn\mathrm{u}|x|^T|y|, \tag{2.4.13}$$

where in (2.4.12) $\phi(n)$ is a "modest" function of n and in (2.4.13) c is a constant of order unity.

We shall not express a preference for any of the error bounding styles shown in (2.4.10)-(2.4.13). This spares us the necessity of translating the roundoff results that appear in the literature into a fixed format. Moreover, paying overly close attention to the details of an error bound is inconsistent with the "philosophy" of roundoff analysis. As Wilkinson (1971, p. 567) says,

> There is still a tendency to attach too much importance to the precise error bounds obtained by an à priori error analysis. In my opinion, the bound itself is usually the least important part of it. The main object of such an analysis is to expose the potential instabilities, if any, of an algorithm so that hopefully from the insight thus obtained one might be led to improved algorithms. Usually the bound itself is weaker than it might have been because of the necessity of restricting the mass of detail to a reasonable level and because of the limitations imposed by expressing the errors in terms of matrix norms. À priori bounds are not, in general, quantities that should be used in practice. Practical error bounds should usually be determined by some form of à posteriori error analysis, since this takes full advantage of the statistical distribution of rounding errors and of any special features, such as sparseness, in the matrix.

It is important to keep these perspectives in mind.

2.4.7 Dot Product Accumulation

Some computers have provision for accumulating dot products in *double precision*. This means that if x and y are floating point vectors with length t mantissas, then the running sum s in (2.4.7) is built up in a register with a $2t$ digit mantissa. Since the multiplication of two t-digit floating point numbers can be stored exactly in a double precision variable, it is only when s is written to single precision memory that any roundoff occurs. In this situation one can usually assert that a computed dot product has good *relative* error, i.e., $fl(x^T y) = x^T y(1+\delta)$ where $|\delta| \approx \mathrm{u}$. Thus, the ability to accumulate dot products is very appealing.

2.4.8 Roundoff in Other Basic Matrix Computations

It is easy to show that if A and B are floating point matrices and α is a floating point number, then

$$fl(\alpha A) = \alpha A + E \qquad |E| \leq \mathbf{u}|\alpha A| \tag{2.4.14}$$

and

$$fl(A + B) = (A + B) + E \qquad |E| \leq \mathbf{u}|A + B|. \tag{2.4.15}$$

As a consequence of these two results, it is easy to verify that computed saxpy's and outer product updates satisfy

$$fl(\alpha x + y) = \alpha x + y + z \qquad |z| \leq \mathbf{u}\,(2|\alpha x| + |y|) + O(\mathbf{u}^2) \tag{2.4.16}$$

$$fl(C + uv^T) = C + uv^T + E \qquad |E| \leq \mathbf{u}\,(|C| + 2|uv^T|) + O(\mathbf{u}^2). \tag{2.4.17}$$

Using (2.4.10) it is easy to show that a dot product based multiplication of two floating point matrices A and B satisfies

$$fl(AB) = AB + E \qquad |E| \leq n\mathbf{u}|A||B| + O(\mathbf{u}^2). \tag{2.4.18}$$

The same result applies if a gaxpy or outer product based procedure is used. Notice that matrix multiplication does not necessarily give small relative error since $|AB|$ may be much smaller than $|A||B|$, e.g.,

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -.99 & 0 \end{bmatrix} = \begin{bmatrix} .01 & 0 \\ 0 & 0 \end{bmatrix}.$$

It is easy to obtain norm bounds from the roundoff results developed thus far. If we look at the 1-norm error in floating point matrix multiplication, then it is easy to show from (2.4.18) that

$$\| fl(AB) - AB \|_1 \leq n\mathbf{u} \| A \|_1 \| B \|_1 + O(\mathbf{u}^2). \tag{2.4.19}$$

2.4.9 Forward and Backward Error Analyses

Each roundoff bound given above is the consequence of a *forward error analysis*. An alternative style of characterizing the roundoff errors in an algorithm is accomplished through a technique known as *backward error analysis*. Here, the rounding errors are related to the data of the problem rather than to its solution. By way of illustration, consider the $n = 2$ version of triangular matrix multiplication. It can be shown that:

$$fl(AB) = \begin{bmatrix} a_{11}b_{11}(1+\epsilon_1) & (a_{11}b_{12}(1+\epsilon_2) + a_{12}b_{22}(1+\epsilon_3))(1+\epsilon_4) \\ 0 & a_{22}b_{22}(1+\epsilon_5) \end{bmatrix}$$

where $|\epsilon_i| \leq u$, for $i = 1{:}5$. However, if we define

$$\hat{A} = \begin{bmatrix} a_{11} & a_{12}(1+\epsilon_3)(1+\epsilon_4) \\ 0 & a_{22}(1+\epsilon_5) \end{bmatrix}$$

and

$$\hat{B} = \begin{bmatrix} b_{11}(1+\epsilon_1) & b_{12}(1+\epsilon_2)(1+\epsilon_4) \\ 0 & b_{22} \end{bmatrix},$$

then it is easily verified that $fl(AB) = \hat{A}\hat{B}$. Moreover,

$$\hat{A} = A + E \qquad |E| \leq 2\mathbf{u}|A| + O(\mathbf{u}^2)$$

$$\hat{B} = B + F \qquad |F| \leq 2\mathbf{u}|B| + O(\mathbf{u}^2).$$

In other words, the computed product is the exact product of slightly perturbed A and B.

2.4.10 Error in Strassen Multiplication

In §1.3.8 we outlined an unconventional matrix multiplication procedure due to Strassen (1969). It is instructive to compare the effect of roundoff in this method with the effect of roundoff in any of the conventional matrix multiplication methods of §1.1.

It can be shown that the Strassen approach (Algorithm 1.3.1) produces a $\hat{C} = fl(AB)$ that satisfies an inequality of the form (2.4.19). This is perfectly satisfactory in many applications. However, the $\hat{C}$ that Strassen's method produces does *not* always satisfy an inequality of the form (2.4.18). To see this, suppose

$$A = B = \begin{bmatrix} .99 & .0010 \\ .0010 & .99 \end{bmatrix}$$

and that we execute Algorithm 1.3.1 using 2-digit floating point arithmetic. Among other things, the following quantities are computed:

$$\begin{aligned} \hat{P}_3 &= fl(.99(.001 - .99)) = -.98 \\ \hat{P}_5 &= fl((.99 + .001).99) = .98 \\ \hat{c}_{12} &= fl(\hat{P}_3 + \hat{P}_5) = 0.0 \end{aligned}$$

Now in exact arithmetic $c_{12} = 2(.001)(.99) = .00198$ and thus Algorithm 1.3.1 produces a $\hat{c}_{12}$ with no correct significant digits. The Strassen approach gets into trouble in this example because small off-diagonal entries are combined with large diagonal entries. Note that in conventional matrix multiplication neither b_{12} and b_{22} nor a_{11} and a_{12} are summed. Thus the contribution of

the small off-diagonal elements is not lost. Indeed, for the above A and B a conventional matrix multiply gives $\hat{c}_{12} = .0020$.

Failure to produce a componentwise accurate $\hat{C}$ can be a serious shortcoming in some applications. For example, in Markov processes the a_{ij}, b_{ij}, and c_{ij} are transition probabilities and are therefore nonnegative. It may be critical to compute c_{ij} accurately if it reflects a particularly important probability in the modeled phenomena. Note that if $A \geq 0$ and $B \geq 0$, then conventional matrix multiplication produces a product $\hat{C}$ that has small componentwise relative error:

$$|\hat{C} - C| \leq nu|A|\,|B| + O(u^2) = nu|C| + O(u^2).$$

This follows from (2.4.18). Because we cannot say the same for the Strassen approach, we conclude that Algorithm 1.3.1 is not attractive for *certain* nonnegative matrix multiplication problems *if* relatively accurate $\hat{c}_{ij}$ are required.

Extrapolating from this discussion we reach two fairly obvious but important conclusions:

- Different methods for computing the same quantity can produce substantially different results.
- Whether or not an algorithm produces satisfactory results depends upon the type of problem solved and the goals of the user.

These observations are clarified in subsequent chapters and are intimately related to the concepts of algorithm stability and problem condition.

Problems

P2.4.1 Show that if (2.4.7) is applied with $y = x$, then $fl(x^Tx) = x^Tx(1+\alpha)$ where $|\alpha| \leq nu + O(u^2)$.

P2.4.2 Prove (2.4.3).

P2.4.3 Show that if $E \in \mathbb{R}^{m\times n}$ with $m \geq n$, then $\| \,|E|\,\|_2 \leq \sqrt{n}\,\|E\|_2$. This result is useful when deriving norm bounds from absolute value bounds.

P2.4.4 Assume the existence of a square root function satisfying $fl(\sqrt{x}) = \sqrt{x}(1+\epsilon)$ with $|\epsilon| \leq u$. Give an algorithm for computing $\|x\|_2$ and bound the rounding errors.

P2.4.5 Suppose A and B are n-by-n upper triangular floating point matrices. If $\hat{C} = fl(AB)$ is computed using one of the conventional §1.1 algorithms, does it follow that $\hat{C} = \hat{A}\hat{B}$ where $\hat{A}$ and $\hat{B}$ are close to A and B?

P2.4.6 Suppose A and B are n-by-n floating point matrices and that A is nonsingular with $\| \,|A^{-1}|\,|A|\,\|_\infty = r$. Show that if $\hat{C} = fl(AB)$ is obtained using any of the algorithms in §1.1, then there exists a $\hat{B}$ so $\hat{C} = A\hat{B}$ and $\|\hat{B} - B\|_\infty \leq nur\|B\|_\infty + O(u^2)$.

P2.4.7 Prove (2.4.18).

Notes and References for Sec. 2.4

For a general introduction to the effects of roundoff error, we recommend

J.H. Wilkinson (1963). *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, NJ.
J.H. Wilkinson (1971). "Modern Error Analysis," *SIAM Review 13*, 548–68.
D. Kahaner, C.B. Moler, and S. Nash (1988). *Numerical Methods and Software*, Prentice-Hall, Englewood Cliffs, NJ.
F. Chaitin-Chatelin and V. Frayssé (1996). *Lectures on Finite Precision Computations*, SIAM Publications, Philadelphia.

More recent developments in error analysis involve interval analysis, the building of statistical models of roundoff error, and the automating of the analysis itself:

T.E. Hull and J.R. Swensen (1966). "Tests of Probabilistic Models for Propagation of Roundoff Errors," *Comm. ACM. 9*, 108–13.
J. Larson and A. Sameh (1978). "Efficient Calculation of the Effects of Roundoff Errors," *ACM Trans. Math. Soft. 4*, 228–36.
W. Miller and D. Spooner (1978). "Software for Roundoff Analysis, II," *ACM Trans. Math. Soft. 4*, 369–90.
J.M. Yohe (1979). "Software for Interval Arithmetic: A Reasonable Portable Package," *ACM Trans. Math. Soft. 5*, 50–63.

Anyone engaged in serious software development needs a thorough understanding of floating point arithmetic. A good way to begin acquiring knowledge in this direction is to read about the IEEE floating point standard in

D. Goldberg (1991). "What Every Computer Scientist Should Know About Floating Point Arithmetic," *ACM Surveys 23*, 5–48.

See also

R.P. Brent (1978). "A Fortran Multiple Precision Arithmetic Package," *ACM Trans. Math. Soft. 4*, 57–70.
R.P. Brent (1978). "Algorithm 524 MP, a Fortran Multiple Precision Arithmetic Package," *ACM Trans. Math. Soft. 4*, 71–81.
J.W. Demmel (1984). "Underflow and the Reliability of Numerical Software," *SIAM J. Sci. and Stat. Comp. 5*, 887–919.
U.W. Kulisch and W.L. Miranker (1986). "The Arithmetic of the Digital Computer," *SIAM Review 28*, 1–40.
W.J. Cody (1988). "ALGORITHM 665 MACHAR: A Subroutine to Dynamically Determine Machine Parameters," *ACM Trans. Math. Soft. 14*, 303–311.
D.H. Bailey, H.D. Simon, J. T. Barton, M.J. Fouts (1989). "Floating Point Arithmetic in Future Supercomputers," *Int'l J. Supercomputing Appl. 3*, 86–90.
D.H. Bailey (1993). "Algorithm 719: Multiprecision Translation and Execution of FORTRAN Programs," *ACM Trans. Math. Soft. 19*, 288–319.

The subtleties associated with the development of high-quality software, even for "simple" problems, are immense. A good example is the design of a subroutine to compute 2-norms

J.M. Blue (1978). "A Portable FORTRAN Program to Find the Euclidean Norm of a Vector," *ACM Trans. Math. Soft. 4*, 15–23.

For an analysis of the Strassen algorithm and other "fast" linear algebra procedures see

R.P. Brent (1970). "Error Analysis of Algorithms for Matrix Multiplication and Triangular Decomposition Using Winograd's Identity," *Numer. Math. 16*, 145–156.
W. Miller (1975). "Computational Complexity and Numerical Stability," *SIAM J. Computing 4*, 97–107.
N.J. Higham (1992). "Stability of a Method for Multiplying Complex Matrices with Three Real Matrix Multiplications," *SIAM J. Matrix Anal. Appl. 13*, 681–687.
J.W. Demmel and N.J. Higham (1992). "Stability of Block Algorithms with Fast Level-3 BLAS," *ACM Trans. Math. Soft. 18*, 274–291.

2.5 Orthogonality and the SVD

Orthogonality has a very prominent role to play in matrix computations. After establishing a few definitions we prove the extremely useful singular value decomposition (SVD). Among other things, the SVD enables us to intelligently handle the matrix rank problem. The concept of rank, though perfectly clear in the exact arithmetic context, is tricky in the presence of roundoff error and fuzzy data. With the SVD we can introduce the practical notion of numerical rank.

2.5.1 Orthogonality

A set of vectors $\{x_1, \ldots, x_p\}$ in $\mathbf{R}^m$ is *orthogonal* if $x_i^T x_j = 0$ whenever $i \neq j$ and *orthonormal* if $x_i^T x_j = \delta_{ij}$. Intuitively, orthogonal vectors are maximally independent for they point in totally different directions.

A collection of subspaces $S_1, \ldots, S_p$ in $\mathbf{R}^m$ is *mutually orthogonal* if $x^T y = 0$ whenever $x \in S_i$ and $y \in S_j$ for $i \neq j$. The *orthogonal complement* of a subspace $S \subseteq \mathbf{R}^m$ is defined by

$$S^{\perp} = \{y \in \mathbf{R}^m : y^T x = 0 \text{ for all } x \in S\}$$

and it is not hard to show that $\text{ran}(A)^{\perp} = \text{null}(A^T)$. The vectors $v_1, \ldots, v_k$ form an *orthonormal* basis for a subspace $S \subseteq \mathbf{R}^m$ if they are orthonormal and span S.

A matrix $Q \in \mathbf{R}^{m \times m}$ is said to be *orthogonal* if $Q^T Q = I$. If $Q = [\, q_1, \ldots, q_m \,]$ is orthogonal, then the q_i form an orthonormal basis for $\mathbf{R}^m$. It is always possible to extend such a basis to a full orthonormal basis $\{v_1, \ldots, v_m\}$ for $\mathbf{R}^m$:

Theorem 2.5.1 *If $V_1 \in \mathbf{R}^{m \times r}$ has orthonormal columns, then there exists $V_2 \in \mathbf{R}^{m \times (m-r)}$ such that*

$$V = [\, V_1 \; V_2 \,]$$

is orthogonal. Note that $\text{ran}(V_1)^{\perp} = \text{ran}(V_2)$.

Proof. This is a standard result from introductory linear algebra. It is also a corollary of the QR factorization that we present in §5.2. □

2.5.2 Norms and Orthogonal Transformations

The 2-norm is invariant under orthogonal transformation, for if $Q^TQ = I$, then $\| Qx \|_2^2 = x^TQ^TQx = x^Tx = \| x \|_2^2$. The matrix 2-norm and the Frobenius norm are also invariant with respect to orthogonal transformations. In particular, it is easy to show that for all orthogonal Q and Z of appropriate dimensions we have

$$\| QAZ \|_F = \| A \|_F \tag{2.5.1}$$

and

$$\| QAZ \|_2 = \| A \|_2 . \tag{2.5.2}$$

2.5.3 The Singular Value Decomposition

The theory of norms developed in the previous two sections can be used to prove the extremely useful singular value decomposition.

Theorem 2.5.2 (Singular Value Decomposition (SVD)) *If A is a real m-by-n matrix, then there exist orthogonal matrices*

$$U = [\, u_1, \ldots, u_m \,] \in \mathbb{R}^{m\times m} \quad \textit{and} \quad V = [\, v_1, \ldots, v_n \,] \in \mathbb{R}^{n\times n}$$

such that

$$U^TAV = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n} \qquad p = \min\{m, n\}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

Proof. Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be unit 2-norm vectors that satisfy $Ax = \sigma y$ with $\sigma = \| A \|_2$. From Theorem 2.5.1 there exist $V_2 \in \mathbb{R}^{n\times(n-1)}$ and $U_2 \in \mathbb{R}^{m\times(m-1)}$ so $V = [\, x \;\; V_2 \,] \in \mathbb{R}^{n\times n}$ and $U = [\, y \;\; U_2 \,] \in \mathbb{R}^{m\times m}$ are orthogonal. It is not hard to show that U^TAV has the following structure:

$$U^TAV = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix} \equiv A_1 .$$

Since

$$\left\| A_1 \left(\begin{bmatrix} \sigma \\ w \end{bmatrix} \right) \right\|_2^2 \geq (\sigma^2 + w^Tw)^2$$

we have $\| A_1 \|_2^2 \geq (\sigma^2 + w^Tw)$. But $\sigma^2 = \| A \|_2^2 = \| A_1 \|_2^2$, and so we must have $w = 0$. An obvious induction argument completes the proof of the theorem. □

The σ_i are the *singular values* of A and the vectors u_i and v_i are the ith *left singular vector* and the ith *right singular vector* respectively. It

is easy to verify by comparing columns in the equations $AV = U\Sigma$ and $A^TU = V\Sigma^T$ that

$$\left.\begin{array}{rcl} Av_i &=& \sigma_i u_i \\ A^T u_i &=& \sigma_i v_i \end{array}\right\} \; i = 1{:}\min\{m,n\}$$

It is convenient to have the following notation for designating singular values:

$$\begin{array}{rcl} \sigma_i(A) &=& \text{the } i\text{th largest singular value of } A, \\ \sigma_{max}(A) &=& \text{the largest singular value of } A, \\ \sigma_{min}(A) &=& \text{the smallest singular value of } A. \end{array}$$

The singular values of a matrix A are precisely the lengths of the semi-axes of the hyperellipsoid E defined by $E = \{\, Ax : \| x \|_2 = 1 \,\}$.

Example 2.5.1

$$A = \begin{bmatrix} .96 & 1.72 \\ 2.28 & .96 \end{bmatrix} = U\Sigma V^T = \begin{bmatrix} .6 & -.8 \\ .8 & .6 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} .8 & .6 \\ .6 & -.8 \end{bmatrix}^T .$$

The SVD reveals a great deal about the structure of a matrix. If the SVD of A is given by Theorem 2.5.2, and we define r by

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0,$$

then

$$\begin{array}{rcll} \text{rank}(A) &=& r & (2.5.3) \\ \text{null}(A) &=& \text{span}\{v_{r+1}, \ldots, v_n\} & (2.5.4) \\ \text{ran}(A) &=& \text{span}\{u_1, \ldots, u_r\}\,, & (2.5.5) \end{array}$$

and we have the *SVD expansion*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \,. \tag{2.5.6}$$

Various 2-norm and Frobenius norm properties have connections to the SVD. If $A \in \mathbb{R}^{m\times n}$, then

$$\begin{array}{rcll} \| A \|_F^2 &=& \sigma_1^2 + \cdots + \sigma_p^2 \qquad p = \min\{m,n\} & (2.5.7) \\ \| A \|_2 &=& \sigma_1 & (2.5.8) \end{array}$$

$$\min_{x\neq 0} \frac{\| Ax \|_2}{\| x \|_2} = \sigma_n \qquad (m \geq n)\,. \tag{2.5.9}$$

2.5.4 The Thin SVD

If $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$ is the SVD of A and $m \geq n$, then

$$A = U_1\Sigma_1 V^T$$

where

$$U_1 = U(:,1{:}n) = [\, u_1, \ldots, u_n \,] \in \mathbb{R}^{m\times n}$$

and

$$\Sigma_1 = \Sigma(1{:}n, 1{:}n) = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}.$$

We refer to this much-used, trimmed down version of the SVD as the *thin SVD*.

2.5.5 Rank Deficiency and the SVD

One of the most valuable aspects of the SVD is that it enables us to deal sensibly with the concept of matrix rank. Numerous theorems in linear algebra have the form "if such-and-such a matrix has full rank, then such-and-such a property holds." While neat and aesthetic, results of this flavor do not help us address the numerical difficulties frequently encountered in situations where near rank deficiency prevails. Rounding errors and fuzzy data make rank determination a nontrivial exercise. Indeed, for some small ϵ we may be interested in the ϵ-rank of a matrix which we define by

$$\mathrm{rank}(A,\epsilon) = \min_{\|A-B\|_2 \leq \epsilon} \mathrm{rank}(B).$$

Thus, if A is obtained in a laboratory with each a_{ij} correct to within $\pm .001$, then it might make sense to look at $\mathrm{rank}(A, .001)$. Along the same lines, if A is an m-by-n floating point matrix then it is reasonable to regard A as *numerically rank deficient* if $\mathrm{rank}(A,\epsilon) < \min\{m,n\}$ with $\epsilon = \mathbf{u}\| A \|_2$.

Numerical rank deficiency and ϵ-rank are nicely characterized in terms of the SVD because the singular values indicate how near a given matrix is to a matrix of lower rank.

Theorem 2.5.3 *Let the SVD of $A \in \mathbb{R}^{m\times n}$ be given by Theorem 2.5.2. If $k < r = \mathrm{rank}(A)$ and*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \tag{2.5.10}$$

then

$$\min_{\mathrm{rank}(B)=k} \| A - B \|_2 = \| A - A_k \|_2 = \sigma_{k+1}. \tag{2.5.11}$$

Proof. Since $U^T A_k V = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ it follows that $\text{rank}(A_k) = k$ and that $U^T(A - A_k)V = \text{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$ and so $\| A - A_k \|_2 = \sigma_{k+1}$.

Now suppose $\text{rank}(B) = k$ for some $B \in \mathbb{R}^{m \times n}$. It follows that we can find orthonormal vectors $x_1, \ldots, x_{n-k}$ so $\text{null}(B) = \text{span}\{x_1, \ldots, x_{n-k}\}$. A dimension argument shows that

$$\text{span}\{x_1, \ldots, x_{n-k}\} \cap \text{span}\{v_1, \ldots, v_{k+1}\} \neq \{0\}.$$

Let z be a unit 2-norm vector in this intersection. Since $Bz = 0$ and

$$Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$$

we have

$$\| A - B \|_2^2 \geq \| (A - B)z \|_2^2 = \| Az \|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2$$

completing the proof of the theorem. □

Theorem 2.5.3 says that the smallest singular value of A is the 2-norm distance of A to the set of all rank-deficient matrices. It also follows that the set of full rank matrices in $\mathbb{R}^{m \times n}$ is both open and dense.

Finally, if $r_\epsilon = \text{rank}(A, \epsilon)$, then

$$\sigma_1 \geq \cdots \geq \sigma_{r_\epsilon} > \epsilon \geq \sigma_{r_\epsilon + 1} \geq \cdots \geq \sigma_p \qquad p = \min\{m, n\}.$$

We have more to say about the numerical rank issue in §5.5 and §12.2.

2.5.6 Unitary Matrices

Over the complex field the unitary matrices correspond to the orthogonal matrices. In particular, $Q \in \mathbb{C}^{n \times n}$ is *unitary* if $Q^H Q = QQ^H = I_n$. Unitary matrices preserve 2-norm. The SVD of a complex matrix involves unitary matrices. If $A \in \mathbb{C}^{m \times n}$, then there exist unitary matrices $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ such that

$$U^H A V = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n} \qquad p = \min\{m, n\}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

Problems

P2.5.1 Show that if S is real and $S^T = -S$, then $I - S$ is nonsingular and the matrix $(I - S)^{-1}(I + S)$ is orthogonal. This is known as the *Cayley transform* of S.

P2.5.2 Show that a triangular orthogonal matrix is diagonal.

P2.5.3 Show that if $Q = Q_1 + iQ_2$ is unitary with $Q_1, Q_2 \in \mathbb{R}^{n\times n}$, then the $2n$-by-$2n$ real matrix

$$Z = \begin{bmatrix} Q_1 & -Q_2 \\ Q_2 & Q_1 \end{bmatrix}$$

is orthogonal.

P2.5.4 Establish properties (2.5.3)-(2.5.9).

P2.5.5 Prove that

$$\sigma_{max}(A) = \max_{y \in \mathbb{R}^m, x \in \mathbb{R}^n} \frac{y^T A x}{\| x \|_2 \| y \|_2}$$

P2.5.6 For the 2-by-2 matrix $A = \begin{bmatrix} w & x \\ y & z \end{bmatrix}$, derive expressions for $\sigma_{max}(A)$ and $\sigma_{min}(A)$ that are functions of w, x, y, and z.

P2.5.7 Show that any matrix in $\mathbb{R}^{m\times n}$ is the limit of a sequence of full rank matrices.

P2.5.8 Show that if $A \in \mathbb{R}^{m\times n}$ has rank n, then $\| A(A^TA)^{-1}A^T \|_2 = 1$.

P2.5.9 What is the nearest rank-one matrix to $A = \begin{bmatrix} 1 & M \\ 0 & 1 \end{bmatrix}$ in the Frobenius norm?

P2.5.10 Show that if $A \in \mathbb{R}^{m\times n}$ then $\| A \|_F \leq \sqrt{\text{rank}(A)}\, \| A \|_2$, thereby sharpening (2.3.7).

Notes and References for Sec. 2.5

Forsythe and Moler (1967) offer a good account of the SVD's role in the analysis of the $Ax = b$ problem. Their proof of the decomposition is more traditional than ours in that it makes use of the eigenvalue theory for symmetric matrices. Historical SVD references include

E. Beltrami (1873). "Sulle Funzioni Bilineari," *Gionale di Mathematiche 11*, 98–106.

C. Eckart and G. Young (1939). "A Principal Axis Transformation for Non-Hermitian Matrices," *Bull. Amer. Math. Soc. 45*, 118–21.

G.W. Stewart (1993). "On the Early History of the Singular Value Decomposition," *SIAM Review 35*, 551–566.

One of the most significant developments in scientific computation has been the increased use of the SVD in application areas that require the intelligent handling of matrix rank. The range of applications is impressive. One of the most interesting is

C.B. Moler and D. Morrison (1983). "Singular Value Analysis of Cryptograms," *Amer. Math. Monthly 90*, 78–87.

For generalizations of the SVD to infinite dimensional Hilbert space, see

I.C. Gohberg and M.G. Krein (1969). *Introduction to the Theory of Linear Non-Self Adjoint Operators* , Amer. Math. Soc., Providence, R.I.

F. Smithies (1970). *Integral Equations*, Cambridge University Press, Cambridge.

Reducing the rank of a matrix as in Theorem 2.5.3 when the perturbing matrix is constrained is discussed in

J.W. Demmel (1987). "The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems, *SIAM J. Numer. Anal. 24*, 199–206.

G.H. Golub, A. Hoffman, and G.W. Stewart (1988). "A Generalization of the Eckart-Young-Mirsky Approximation Theorem." *Lin. Alg. and Its Applic. 88/89*, 317–328.
G.A. Watson (1988). "The Smallest Perturbation of a Submatrix which Lowers the Rank of the Matrix," *IMA J. Numer. Anal. 8*, 295–304.

2.6 Projections and the CS Decomposition

If the object of a computation is to compute a matrix or a vector, then norms are useful for assessing the accuracy of the answer or for measuring progress during an iteration. If the object of a computation is to compute a subspace, then to make similar comments we need to be able to quantify the distance between two subspaces. Orthogonal projections are critical in this regard. After the elementary concepts are established we discuss the CS decomposition. This is an SVD-like decomposition that is handy when having to compare a pair of subspaces. We begin with the notion of an orthogonal projection.

2.6.1 Orthogonal Projections

Let $S \subseteq \mathbb{R}^n$ be a subspace. $P \in \mathbb{R}^{n\times n}$ is the *orthogonal projection* onto S if $\text{ran}(P) = S$, $P^2 = P$, and $P^T = P$. From this definition it is easy to show that if $x \in \mathbb{R}^n$, then $Px \in S$ and $(I - P)x \in S^{\perp}$.

If P_1 and P_2 are each orthogonal projections, then for any $z \in \mathbb{R}^n$ we have

$$\| (P_1 - P_2)z \|_2^2 = (P_1 z)^T (I - P_2)z + (P_2 z)^T (I - P_1)z.$$

If $\text{ran}(P_1) = \text{ran}(P_2) = S$, then the right-hand side of this expression is zero showing that the orthogonal projection for a subspace is unique. If the columns of $V = [\, v_1, \ldots, v_k \,]$ are an orthonormal basis for a subspace S, then it is easy to show that $P = VV^T$ is the unique orthogonal projection onto S. Note that if $v \in \mathbb{R}^n$, then $P = vv^T/v^Tv$ is the orthogonal projection onto $S = \text{span}\{v\}$.

2.6.2 SVD-Related Projections

There are several important orthogonal projections associated with the singular value decomposition. Suppose $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$ is the SVD of A and that $r = \text{rank}(A)$. If we have the U and V partitionings

$$U = \underset{r \quad m-r}{[\, U_r \quad \tilde{U}_r \,]} \qquad V = \underset{r \quad n-r}{[\, V_r \quad \tilde{V}_r \,]}$$

then

$$\begin{aligned}
V_r V_r^T &= \text{projection on to null}(A)^{\perp} = \text{ran}(A^T) \\
\tilde{V}_r \tilde{V}_r^T &= \text{projection on to null}(A) \\
U_r U_r^T &= \text{projection on to ran}(A) \\
\tilde{U}_r \tilde{U}_r^T &= \text{projection on to ran}(A)^{\perp} = \text{null}(A^T)
\end{aligned}$$

2.6.3 Distance Between Subspaces

The one-to-one correspondence between subspaces and orthogonal projections enables us to devise a notion of distance between subspaces. Suppose S_1 and S_2 are subspaces of $\mathbf{R}^n$ and that $\dim(S_1) = \dim(S_2)$. We define the *distance* between these two spaces by

$$\text{dist}(S_1, S_2) = \| P_1 - P_2 \|_2 \tag{2.6.1}$$

where P_i is the orthogonal projection onto S_i. The distance between a pair of subspaces can be characterized in terms of the blocks of a certain orthogonal matrix.

Theorem 2.6.1 *Suppose*

$$W = \underset{k}{\left[\begin{array}{cc} W_1 & W_2 \end{array}\right.}\ \underset{n-k}{\left.\begin{array}{c} \\ \\ \end{array}\right]} \qquad Z = \underset{k}{\left[\begin{array}{cc} Z_1 & Z_2 \end{array}\right.}\ \underset{n-k}{\left.\begin{array}{c} \\ \\ \end{array}\right]}$$

are n-by-n orthogonal matrices. If $S_1 = \text{ran}(W_1)$ and $S_2 = \text{ran}(Z_1)$, then

$$\text{dist}(S_1, S_2) = \| W_1^T Z_2 \|_2 = \| Z_1^T W_2 \|_2.$$

Proof.

$$\begin{aligned} \text{dist}(S_1, S_2) &= \| W_1 W_1^T - Z_1 Z_1^T \|_2 = \| W^T (W_1 W_1^T - Z_1 Z_1^T) Z \|_2 \\ &= \left\| \begin{bmatrix} 0 & W_1^T Z_2 \\ -W_2^T Z_1 & 0 \end{bmatrix} \right\|_2 . \end{aligned}$$

Note that the matrices $W_2^T Z_1$ and $W_1^T Z_2$ are submatrices of the orthogonal matrix

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \equiv \begin{bmatrix} W_1^T Z_1 & W_1^T Z_2 \\ W_2^T Z_1 & W_2^T Z_2 \end{bmatrix} = W^T Z.$$

Our goal is to show that $\| Q_{21} \|_2 = \| Q_{12} \|_2$. Since Q is orthogonal it follows from

$$Q \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} Q_{11} x \\ Q_{21} x \end{bmatrix}$$

that

$$1 = \| Q_{11} x \|_2^2 + \| Q_{21} x \|_2^2$$

for all unit 2-norm $x \in \mathbf{R}^k$. Thus,

$$\begin{aligned} \| Q_{21} \|_2^2 &= \max_{\| x \|_2 = 1} \| Q_{21} x \|_2^2 = 1 - \min_{\| x \|_2 = 1} \| Q_{11} x \|_2^2 \\ &= 1 - \sigma_{\min}(Q_{11})^2. \end{aligned}$$

Analogously, by working with Q^T (which is also orthogonal) it is possible to show that

$$\| Q_{12}^T \|_2^2 = 1 - \sigma_{min}(Q_{11}^T)^2.$$

and therefore

$$\| Q_{12} \|_2^2 = 1 - \sigma_{min}(Q_{11})^2.$$

Thus, $\| Q_{21} \|_2 = \| Q_{12} \|_2$. □

Note that if S_1 and S_2 are subspaces in $\mathbb{R}^n$ with the same dimension, then

$$0 \leq \text{dist}(S_1, S_2) \leq 1.$$

The distance is zero if $S_1 = S_2$ and one if $S_1 \bigcap S_2^{\perp} \neq \{0\}$.

A more refined analysis of the blocks of the Q matrix above sheds more light on the difference between a pair of subspaces. This requires a special SVD-like decomposition for orthogonal matrices.

2.6.4 The CS Decomposition

The blocks of an orthogonal matrix partitioned into 2-by-2 form have highly related SVDs. This is the gist of the *CS decomposition*. We prove a very useful special case first.

Theorem 2.6.2 (The CS Decomposition (Thin Version)) *Consider the matrix*

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \qquad Q_1 \in \mathbb{R}^{m_1 \times n},\ Q_2 \in \mathbb{R}^{m_2 \times n}$$

where $m_1 \geq n$ and $m_2 \geq n$. If the columns of Q are orthonormal, then there exist orthogonal matrices $U_1 \in \mathbb{R}^{m_1 \times m_1}$, $U_2 \in \mathbb{R}^{m_2 \times m_2}$, and $V_1 \in \mathbb{R}^{n \times n}$ such that

$$\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}^T \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} V_1 = \begin{bmatrix} C \\ S \end{bmatrix}$$

where

$$\begin{aligned} C &= \text{diag}(\cos(\theta_1), \ldots, \cos(\theta_n)), \\ S &= \text{diag}(\sin(\theta_1), \ldots, \sin(\theta_n)), \end{aligned}$$

and

$$0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n \leq \frac{\Pi}{2}.$$

Proof. Since $\| Q_{11} \|_2 \leq \| Q \|_2 = 1$, the singular values of Q_{11} are all in the interval $[0, 1]$. Let

$$U_1^T Q_1 V_1 = C = \text{diag}(c_1, \ldots, c_n) = \begin{array}{c} \\ t \\ m_1 - t \end{array} \begin{array}{c} \begin{array}{cc} t & n-t \end{array} \\ \begin{bmatrix} I_t & 0 \\ 0 & \Sigma \end{bmatrix} \end{array}$$

be the SVD of Q_1 where we assume

$$1 = c_1 = \cdots = c_t > c_{t+1} \geq \cdots \geq c_n \geq 0.$$

To complete the proof of the theorem we must construct the orthogonal matrix U_2. If

$$Q_2 V_1 = \begin{array}{c} \\ t \end{array} \begin{array}{c} \begin{bmatrix} W_1 & W_2 \end{bmatrix} \\ \begin{array}{cc} t & n-t \end{array} \end{array},$$

then

$$\begin{bmatrix} U_1 & 0 \\ 0 & I_{m_2} \end{bmatrix}^T \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} V_1 = \begin{bmatrix} I_t & 0 \\ 0 & \Sigma \\ W_1 & W_2 \end{bmatrix}.$$

Since the columns of this matrix have unit 2-norm, $W_1 = 0$. The columns of W_2 are nonzero and mutually orthogonal because

$$W_2^T W_2 = I_{n-t} - \Sigma^T \Sigma \equiv \text{diag}(1 - c_{t+1}^2, \ldots, 1 - c_n^2)$$

is nonsingular. If $s_k = \sqrt{1 - c_k^2}$ for $k = 1{:}n$, then the columns of

$$Z = W_2 \, \text{diag}(1/s_{t+1}, \ldots, 1/s_n)$$

are orthonormal. By Theorem 2.5.1 there exists an orthogonal matrix $U_2 \in \mathbb{R}^{m_2 \times m_2}$ with $U_2(:, t+1{:}n) = Z$. It is easy to verify that

$$U_2^T Q_2 V_1 = \text{diag}(s_1, \ldots, s_n) \equiv S.$$

Since $c_k^2 + s_k^2 = 1$ for $k = 1{:}n$, it follows that these quantities are the required cosines and sines. □

Using the same sort of techniques it is possible to prove the following more general version of the decomposition:

Theorem 2.6.3 (CS Decomposition (General Version)) *If*

$$Q = \left[\begin{array}{c|c} Q_{11} & Q_{12} \\ \hline Q_{21} & Q_{22} \end{array}\right]$$

is a 2-by-2 (arbitrary) partitioning of an n-by-n orthogonal matrix, then there exist orthogonal

$$U = \left[\begin{array}{c|c} U_1 & 0 \\ \hline 0 & U_2 \end{array}\right] \quad \text{and} \quad V = \left[\begin{array}{c|c} V_1 & 0 \\ \hline 0 & V_2 \end{array}\right]$$

such that

$$U^T Q V = \left[\begin{array}{ccc|ccc} I & 0 & 0 & 0 & 0 & 0 \\ 0 & C & 0 & 0 & S & 0 \\ 0 & 0 & 0 & 0 & 0 & I \\ \hline 0 & 0 & 0 & I & 0 & 0 \\ 0 & S & 0 & 0 & -C & 0 \\ 0 & 0 & I & 0 & 0 & 0 \end{array}\right]$$

where $C = \text{diag}(c_1, \ldots, c_p)$ *and* $S = \text{diag}(s_1, \ldots, s_p)$ *are square diagonal matrices with* $0 < c_i, s_i < 1$.

Proof. See Paige and Saunders (1981) for details. We have suppressed the dimensions of the zero submatrices, some of which may be empty. □

The essential message of the decomposition is that the SVDs of the Q_{ij} are highly related.

Example 2.6.1 The matrix

$$Q = \left[\begin{array}{rr|rrr} -0.7576 & 0.3697 & 0.3838 & 0.2126 & -0.3112 \\ -0.4077 & -0.1552 & -0.1129 & 0.2676 & 0.8517 \\ -0.0488 & 0.7240 & -0.6730 & -0.1301 & 0.0602 \\ \hline -0.2287 & 0.0088 & 0.2235 & -0.9235 & 0.2120 \\ 0.4530 & 0.5612 & 0.5806 & 0.1162 & 0.3595 \end{array}\right]$$

is orthogonal and with the indicated partitioning can be reduced to

$$U^T Q V = \left[\begin{array}{rr|rrr} 0.9837 & 0.0000 & 0.1800 & 0.0000 & 0.0000 \\ 0.0000 & 0.6781 & 0.0000 & 0.7349 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ \hline 0.1800 & 0.0000 & -0.9837 & 0.0000 & 0.0000 \\ 0.0000 & 0.7349 & 0.0000 & -0.6781 & 0.0000 \end{array}\right]$$

The angles associated with the cosines and sines turn out to be very important in a number of applications. See §12.4.

Problems

P2.6.1 Show that if P is an orthogonal projection, then $Q = I - 2P$ is orthogonal.

P2.6.2 What are the singular values of an orthogonal projection?

P2.6.3 Suppose $S_1 = \text{span}\{x\}$ and $S_2 = \text{span}\{y\}$, where x and y are unit 2-norm vectors in $\mathbf{R}^2$. Working only with the definition of $\text{dist}(\cdot,\cdot)$, show that $\text{dist}(S_1, S_2) = \sqrt{1 - (x^T y)^2}$ verifying that the distance between S_1 and S_2 equals the sine of the angle between x and y.

Notes and References for Sec. 2.6

The following papers discuss various aspects of the CS decomposition:

C. Davis and W. Kahan (1970). "The Rotation of Eigenvectors by a Perturbation III," *SIAM J. Num. Anal. 7*, 1–46.

G.W. Stewart (1977). "On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems," *SIAM Review 19*, 634–662.

C.C. Paige and M. Saunders (1981). "Toward a Generalized Singular Value Decomposition," *SIAM J. Num. Anal. 18*, 398–405.

C.C. Paige and M. Wei (1994). "History and Generality of the CS Decomposition," *Lin. Alg. and Its Applic. 208/209*, 303–326.

See §8.7 for some computational details.

For a deeper geometrical understanding of the CS decomposition and the notion of distance between subspaces, see

T.A. Arias, A. Edelman, and S. Smith (1996). "Conjugate Gradient and Newton's Method on the Grassman and Stiefel Manifolds," to appear in *SIAM J. Matrix Anal. Appl.*

2.7 The Sensitivity of Square Systems

We now use some of the tools developed in previous sections to analyze the linear system problem $Ax = b$ where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $b \in \mathbb{R}^n$. Our aim is to examine how perturbations in A and b affect the solution x. A much more detailed treatment may be found in Higham (1996).

2.7.1 An SVD Analysis

If

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T = U\Sigma V^T$$

is the SVD of A, then

$$x = A^{-1}b = (U\Sigma V^T)^{-1}b = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i. \tag{2.7.1}$$

This expansion shows that small changes in A or b can induce relatively large changes in x if σ_n is small.

It should come as no surprise that the magnitude of σ_n should have a bearing on the sensitivity of the $Ax = b$ problem when we recall from Theorem 2.5.3 that σ_n is the distance from A to the set of singular matrices. As the matrix of coefficients approaches this set, it is intuitively clear that the solution x should be increasingly sensitive to perturbations.

2.7.2 Condition

A precise measure of linear system sensitivity can be obtained by considering the parameterized system

$$(A + \epsilon F)x(\epsilon) = b + \epsilon f \qquad x(0) = x$$

where $F \in \mathbb{R}^{n\times n}$ and $f \in \mathbb{R}^n$. If A is nonsingular, then it is clear that $x(\epsilon)$ is differentiable in a neighborhood of zero. Moreover, $\dot{x}(0) = A^{-1}(f - Fx)$ and thus, the Taylor series expansion for $x(\epsilon)$ has the form

$$x(\epsilon) = x + \epsilon\dot{x}(0) + O(\epsilon^2).$$

Using any vector norm and consistent matrix norm we obtain

$$\frac{\| x(\epsilon) - x \|}{\| x \|} \leq |\epsilon| \| A^{-1} \| \left\{ \frac{\| f \|}{\| x \|} + \| F \| \right\} + O(\epsilon^2). \tag{2.7.2}$$

For square matrices A define the *condition number* $\kappa(A)$ by

$$\kappa(A) = \| A \| \| A^{-1} \| \tag{2.7.3}$$

with the convention that $\kappa(A) = \infty$ for singular A. Using the inequality $\| b \| \leq \| A \| \| x \|$ it follows from (2.7.2) that

$$\frac{\| x(\epsilon) - x \|}{\| x \|} \leq \kappa(A)(\rho_A + \rho_b) + O(\epsilon^2) \tag{2.7.4}$$

where

$$\rho_A = |\epsilon| \frac{\| F \|}{\| A \|} \quad \text{and} \quad \rho_b = |\epsilon| \frac{\| f \|}{\| b \|}$$

represent the relative errors in A and b, respectively. Thus, the relative error in x can be $\kappa(A)$ times the relative error in A and b. In this sense, the condition number $\kappa(A)$ quantifies the sensitivity of the $Ax = b$ problem.

Note that $\kappa(\cdot)$ depends on the underlying norm and subscripts are used accordingly, e.g.,

$$\kappa_2(A) = \| A \|_2 \| A^{-1} \|_2 = \frac{\sigma_1(A)}{\sigma_n(A)}. \tag{2.7.5}$$

Thus, the 2-norm condition of a matrix A measures the elongation of the hyperellipsoid $\{Ax : \| x \|_2 = 1\}$.

We mention two other characterizations of the condition number. For p-norm condition numbers, we have

$$\frac{1}{\kappa_p(A)} = \min_{A + \Delta A \text{ singular}} \frac{\| \Delta A \|_p}{\| A \|_p}. \tag{2.7.6}$$

This result may be found in Kahan (1966) and shows that $\kappa_p(A)$ measures the relative p-norm distance from A to the set of singular matrices.

For any norm, we also have

$$\kappa(A) = \lim_{\epsilon \to 0} \sup_{\| \Delta A \| \leq \epsilon \| A \|} \frac{\| (A + \Delta A)^{-1} - A^{-1} \|}{\epsilon} \frac{1}{\| A^{-1} \|}. \tag{2.7.7}$$

This imposing result merely says that the condition number is a normalized Frechet derivative of the map $A \to A^{-1}$. Further details may be found in Rice (1966b). Recall that we were initially led to $\kappa(A)$ through differentiation.

If $\kappa(A)$ is large, then A is said to be an *ill-conditioned* matrix. Note that this is a norm-dependent property[2]. However, any two condition numbers $\kappa_\alpha(\cdot)$ and $\kappa_\beta(\cdot)$ on $\mathbb{R}^{n\times n}$ are equivalent in that constants c_1 and c_2 can be found for which

$$c_1\kappa_\alpha(A) \leq \kappa_\beta(A) \leq c_2\kappa_\alpha(A) \qquad A \in \mathbb{R}^{n\times n}.$$

For example, on $\mathbb{R}^{n\times n}$ we have

$$\begin{aligned} \frac{1}{n}\kappa_2(A) &\leq \kappa_1(A) \leq n\kappa_2(A) \\ \frac{1}{n}\kappa_\infty(A) &\leq \kappa_2(A) \leq n\kappa_\infty(A) \\ \frac{1}{n^2}\kappa_1(A) &\leq \kappa_\infty(A) \leq n^2\kappa_1(A). \end{aligned} \tag{2.7.8}$$

Thus, if a matrix is ill-conditioned in the α-norm, it is ill-conditioned in the β-norm modulo the constants c_1 and c_2 above.

For any of the p-norms, we have $\kappa_p(A) \geq 1$. Matrices with small condition numbers are said to be *well-conditioned*. In the 2-norm, orthogonal matrices are perfectly conditioned in that $\kappa_2(Q) = 1$ if Q is orthogonal.

2.7.3 Determinants and Nearness to Singularity

It is natural to consider how well determinant size measures ill-conditioning. If $\det(A) = 0$ is equivalent to singularity, is $\det(A) \approx 0$ equivalent to near singularity? Unfortunately, there is little correlation between $\det(A)$ and the condition of $Ax = b$. For example, the matrix B_n defined by

$$B_n = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 0 & 1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{2.7.9}$$

has determinant 1, but $\kappa_\infty(B_n) = n2^{n-1}$. On the other hand, a very well conditioned matrix can have a very small determinant. For example,

$$D_n = \mathrm{diag}(10^{-1}, \ldots, 10^{-1}) \in \mathbb{R}^{n\times n}$$

satisfies $\kappa_p(D_n) = 1$ although $\det(D_n) = 10^{-n}$.

2.7.4 A Rigorous Norm Bound

Recall that the derivation of (2.7.4) was valuable because it highlighted the connection between $\kappa(A)$ and the rate of change of $x(\epsilon)$ at $\epsilon = 0$. However,

[2] It also depends upon the definition of "large." The matter is pursued in §3.5

it is a little unsatisfying because it is contingent on ϵ being "small enough" and because it sheds no light on the size of the $O(\epsilon^2)$ term. In this and the next subsection we develop some additional $Ax = b$ perturbation theorems that are completely rigorous.

We first establish a useful lemma that indicates in terms of $\kappa(A)$ when we can expect a perturbed system to be nonsingular.

Lemma 2.7.1 *Suppose*

$$Ax = b \qquad A \in \mathbb{R}^{n\times n},\ 0 \neq b \in \mathbb{R}^n$$

$$(A+\Delta A)y = b+\Delta b \qquad \Delta A \in \mathbb{R}^{n\times n},\ \Delta b \in \mathbb{R}^n$$

with $\| \Delta A \| \leq \epsilon \| A \|$ *and* $\| \Delta b \| \leq \epsilon \| b \|$. *If* $\epsilon \kappa(A) = r < 1$, *then* $A + \Delta A$ *is nonsingular and*

$$\frac{\| y \|}{\| x \|} \leq \frac{1+r}{1-r}.$$

Proof. Since $\| A^{-1}\Delta A \| \leq \epsilon \| A^{-1} \| \| A \| = r < 1$ it follows from Theorem 2.3.4 that $(A + \Delta A)$ is nonsingular. Using Lemma 2.3.3 and the equality $(I + A^{-1}\Delta A)y = x + A^{-1}\Delta b$ we find

$$\begin{aligned} \| y \| &\leq \| (I + A^{-1}\Delta A)^{-1} \| \left(\| x \| + \epsilon \| A^{-1} \| \| b \| \right) \\ &\leq \frac{1}{1-r} \left(\| x \| + \epsilon \| A^{-1} \| \| b \| \right) = \frac{1}{1-r} \left(\| x \| + r \frac{\| b \|}{\| A \|} \right). \end{aligned}$$

Since $\| b \| = \| Ax \| \leq \| A \| \| x \|$ it follows that

$$\| y \| \leq \frac{1}{1-r} \left(\| x \| + r \| x \| \right). \quad \square$$

We are now set to establish a rigorous $Ax = b$ perturbation bound.

Theorem 2.7.2 *If the conditions of Lemma 2.7.1 hold, then*

$$\frac{\| y - x \|}{\| x \|} \leq \frac{2\epsilon}{1-r} \kappa(A) \tag{2.7.10}$$

Proof. Since

$$y - x = A^{-1}\Delta b - A^{-1}\Delta A y \tag{2.7.11}$$

we have $\| y - x \| \leq \epsilon \| A^{-1} \| \| b \| + \epsilon \| A^{-1} \| \| A \| \| y \|$ and so

$$\begin{aligned} \frac{\| y - x \|}{\| x \|} &\leq \epsilon \kappa(A) \frac{\| b \|}{\| A \| \| x \|} + \epsilon \kappa(A) \frac{\| y \|}{\| x \|} \\ &\leq \epsilon \kappa(A) \left(1 + \frac{1+r}{1-r} \right) = \frac{2\epsilon}{1-r} \kappa(A). \quad \square \end{aligned}$$

Example 2.7.1 The $Ax = b$ problem

$$\begin{bmatrix} 1 & 0 \\ 0 & 10^{-6} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 10^{-6} \end{bmatrix}$$

has solution $x = (1, 1)^T$ and condition $\kappa_\infty(A) = 10^6$. If $\Delta b = (10^{-6}, 0)^T$, $\Delta A = 0$, and $(A + \Delta A)y = b + \Delta b$, then $y = (1 + 10^{-6}, 1)^T$ and the inequality (2.7.10) says

$$10^{-6} = \frac{\| x - y \|_\infty}{\| x \|_\infty} \ll \frac{\| \Delta b \|_\infty}{\| b \|_\infty} \kappa_\infty(A) = 10^{-6} 10^6 = 1 .$$

Thus, the upper bound in (2.7.10) can be a gross overestimate of the error induced by the perturbation. On the other hand, if $\Delta b = (0, 10^{-6})^T$, $\Delta A = 0$, and $(A + \Delta A)y = b + \Delta b$, then this inequality says

$$\frac{10^0}{10^0} \leq 2 \times 10^{-6} 10^6 .$$

Thus, there are perturbations for which the bound in (2.7.10) is essentially attained.

2.7.5 Some Rigorous Componentwise Bounds

We conclude this section by showing that a more refined perturbation theory is possible if componentwise perturbation bounds are in effect and if we make use of the absolute value notation.

Theorem 2.7.3 *Suppose*

$$Ax = b \qquad A \in \mathbb{R}^{n \times n},\ 0 \neq b \in \mathbb{R}^n$$

$$(A + \Delta A)y = b + \Delta b \quad \Delta A \in \mathbb{R}^{n \times n},\ \Delta b \in \mathbb{R}^n$$

and that $|\Delta A| \leq \epsilon |A|$ *and* $|\Delta b| \leq \epsilon |b|$. *If* $\delta \kappa_\infty(A) = r < 1$, *then* $(A + \Delta A)$ *is nonsingular and*

$$\frac{\| y - x \|_\infty}{\| x \|_\infty} \leq \frac{2\delta}{1 - r} \| \, |A^{-1}| \, |A| \, \|_\infty .$$

Proof. Since $\| \Delta A \|_\infty \leq \epsilon \| A \|_\infty$ and $\| \Delta b \|_\infty \leq \epsilon \| b \|_\infty$ the conditions of Lemma 2.7.1 are satisfied in the infinity norm. This implies that $A + \Delta A$ is nonsingular and

$$\frac{\| y \|_\infty}{\| x \|_\infty} \leq \frac{1 + r}{1 - r} .$$

Now using (2.7.11) we find

$$\begin{aligned} |y - x| &\leq |A^{-1}| |\Delta b| + |A^{-1}| |\Delta A| |y| \\ &\leq \epsilon |A^{-1}| |b| + \epsilon |A^{-1}| |A| |y| \leq \epsilon |A^{-1}| |A| (|x| + |y|) . \end{aligned}$$

If we take norms, then

$$\| y - x \|_\infty \leq \epsilon \| \, |A^{-1}| \, |A| \, \|_\infty \left(\| x \|_\infty + \frac{1 + r}{1 - r} \| x \|_\infty \right) .$$

The theorem follows upon division by $\| x \|_\infty$. $\square$

We refer to the quantity $\| \, |A^{-1}| \, |A| \, \|_\infty$ as the *Skeel condition number*. It has been effectively used in the analysis of several important linear system computations. See §3.5.

Lastly, we report on the results of Oettli and Prager (1964) that indicate when an approximate solution $\hat{x} \in \mathbb{R}^n$ to the n-by-n system $Ax = b$ satisfies a perturbed system with prescribed structure. In particular, suppose $E \in \mathbb{R}^{n \times n}$ and $f \in \mathbb{R}^n$ are given and have nonnegative entries. We seek $\Delta A \in \mathbb{R}^{n \times n}$, $\Delta b \in \mathbb{R}^n$, and $\omega \geq 0$ such that

$$(A + \Delta A)\hat{x} = b + \Delta b \qquad |\Delta A| \leq \omega E \, , \, |\Delta b| \leq \omega f \, . \tag{2.7.12}$$

Note that by properly choosing E and f the perturbed system can take on certain qualities. For example, if $E = |A|$ and $f = |b|$ and ω is small, then $\hat{x}$ satisfies a nearby system in the componentwise sense. Oettli and Prager (1964) show that for a given A, b, $\hat{x}$, E, and f the smallest ω possible in (2.7.12) is given by

$$\omega_{min} = \max_{1 \leq i \leq n} \frac{|A\hat{x} - b|_i}{(E|\hat{x}| + f)_i} \, .$$

If $A\hat{x} = b$ then $\omega_{min} = 0$. On the other hand, if $\omega_{min} = \infty$, then $\hat{x}$ does not satisfy any system of the prescribed perturbation structure.

Problems

P2.7.1 Show that if $\| I \| \geq 1$, then $\kappa(A) \geq 1$.

P2.7.2 Show that for a given norm, $\kappa(AB) \leq \kappa(A)\kappa(B)$ and that $\kappa(\alpha A) = \kappa(A)$ for all nonzero α.

P2.7.3 Relate the 2-norm condition of $X \in \mathbb{R}^{m \times n}$ $(m \geq n)$ to the 2-norm condition of the matrices

$$B = \begin{bmatrix} I_m & X \\ 0 & I_n \end{bmatrix}$$

and

$$C = \begin{bmatrix} X \\ I_n \end{bmatrix}.$$

Notes and References for Sec. 2.7

The condition concept is thoroughly investigated in

J. Rice (1966). "A Theory of Condition," *SIAM J. Num. Anal. 3*, 287–310.
W. Kahan (1966). "Numerical Linear Algebra," *Canadian Math. Bull. 9*, 757–801.

References for componentwise perturbation theory include

W. Oettli and W. Prager (1964). "Compatibility of Approximate Solutions of Linear Equations with Given Error Bounds for Coefficients and Right Hand Sides," *Numer. Math. 6*, 405–409.

J.E. Cope and B.W. Rust (1979). "Bounds on solutions of systems with accurate data," *SIAM J. Num. Anal. 16*, 950–63.

R.D. Skeel (1979). "Scaling for numerical stability in Gaussian Elimination," *J. ACM 26*, 494–526.

J.W. Demmel (1992). "The Componentwise Distance to the Nearest Singular Matrix," *SIAM J. Matrix Anal. Appl. 13*, 10–19.

D.J. Higham and N.J. Higham (1992). "Componentwise Perturbation Theory for Linear Systems with Multiple Right-Hand Sides," *Lin. Alg. and Its Applic. 174*, 111–129.

N.J. Higham (1994). "A Survey of Componentwise Perturbation Theory in Numerical Linear Algebra," in *Mathematics of Computation 1943–1993: A Half Century of Computational Mathematics*, W. Gautschi (ed.), Volume 48 of *Proceedings of Symposia in Applied Mathematics*, American Mathematical Society, Providence, Rhode Island.

S. Chandrasekaran and I.C.F. Ipsen (1995). "On the Sensitivity of Solution Components in Linear Systems of Equations," *SIAM J. Matrix Anal. Appl. 16*, 93–112.

The reciprocal of the condition number measures how near a given $Ax = b$ problem is to singularity. The importance of knowing how near a given problem is to a difficult or insoluble problem has come to be appreciated in many computational settings. See

A. Laub(1985). "Numerical Linear Algebra Aspects of Control Design Computations," *IEEE Trans. Auto. Cont. AC-30*, 97–108.

J. L. Barlow (1986). "On the Smallest Positive Singular Value of an M-Matrix with Applications to Ergodic Markov Chains," *SIAM J. Alg. and Disc. Struct. 7*, 414–424.

J.W. Demmel (1987). "On the Distance to the Nearest Ill-Posed Problem," *Numer. Math. 51*, 251–289.

J.W. Demmel (1988). "The Probability that a Numerical Analysis Problem is Difficult," *Math. Comp. 50*, 449–480.

N.J. Higham (1989). "Matrix Nearness Problems and Applications," in *Applications of Matrix Theory*, M.J.C. Gover and S. Barnett (eds), Oxford University Press, Oxford UK, 1–27.