

Chapter 4: Matrix Norms

The analysis of matrix-based algorithms often requires use of matrix norms. These algorithms need a way to quantify the "size" of a matrix or the "distance" between two matrices. For example, suppose an algorithm only works well with full-rank, $n \times n$ matrices, and it produces inaccurate results when supplied with a nearly rank deficit matrix. Obviously, the concept of *e-rank* (also known as *numerical rank*), defined by

$$\text{rank}(A, \epsilon) = \min_{\|A-B\| \leq \epsilon} \text{rank}(B) \quad (4-1)$$

is of interest here. All matrices B that are "within" ϵ of A are examined when computing the *e-rank* of A .

We define a matrix norm in terms of a given vector norm; in our work, we use only the p -vector norm, denoted as $\|\vec{X}\|_p$. Let A be an $m \times n$ matrix, and define

$$\|A\|_p = \sup_{\vec{X} \neq 0} \frac{\|A\vec{X}\|_p}{\|\vec{X}\|_p}, \quad (4-2)$$

where "sup" stands for supremum, also known as least upper bound. Note that we use the same $\|\cdot\|_p$ notation for both vector and matrix norms. However, the meaning should be clear from context.

Since the matrix norm is defined in terms of the vector norm, we say that the matrix norm is *subordinate* to the vector norm. Also, we say that the matrix norm is *induced* by the vector norm.

Now, since $\|\vec{X}\|_p$ is a scalar, we have

$$\|A\|_p = \sup_{\vec{X} \neq 0} \frac{\|A\vec{X}\|_p}{\|\vec{X}\|_p} = \sup_{\vec{X} \neq 0} \left\| A\vec{X} / \|\vec{X}\|_p \right\|_p. \quad (4-3)$$

In (4-3), note that $\vec{X} / \|\vec{X}\|_p$ has unit length; for this reason, we can express the norm of A in terms of a supremum over all vectors of unit length,

$$\|A\|_p = \sup_{\|\vec{X}\|_p=1} \|A\vec{X}\|_p. \quad (4-4)$$

That is, $\|A\|_p$ is the supremum of $\|A\vec{X}\|_p$ on the *unit ball* $\|\vec{X}\|_p = 1$.

Careful consideration of (4-2) reveals that

$$\|A\vec{X}\|_p \leq \|A\|_p \|\vec{X}\|_p \quad (4-5)$$

for *all* \vec{X} . However, $\|A\vec{X}\|_p$ is a continuous function of \vec{X} , and the unit ball $\|\vec{X}\|_p = 1$ is closed and bounded (real analysis books, it is said to be *compact*). Now, on a closed and bounded set, a continuous function always achieves its maximum and minimum values. Hence, in the definition of the matrix norm, we can replace the "sup" with "max" and write

$$\|A\|_p = \max_{\vec{X} \neq 0} \frac{\|A\vec{X}\|_p}{\|\vec{X}\|_p} = \max_{\|\vec{X}\|_p=1} \|A\vec{X}\|_p. \quad (4-6)$$

When computing the norm of A, the definition is used as a starting point. The process has two steps.

- 1) Find a "candidate" for the norm, call it **K** for now, that satisfies $\|A\vec{X}\|_p \leq \mathbf{K} \|\vec{X}\|_p$ for all \vec{X} .
- 2) Find at least one nonzero \vec{X}_0 for which $\|A\vec{X}_0\|_p = \mathbf{K} \|\vec{X}_0\|_p$

Then, you have your norm: set $\|A\|_p = \mathbf{K}$.

MatLab's Matrix Norm Functions

From an application standpoint, the 1-norm, 2-norm and the ∞ -norm are among the most

important; MatLab computes these matrix norms. In MatLab, the 1-norm, 2-norm and ∞ -norm are invoked by the statements $\text{norm}(A,1)$, $\text{norm}(A,2)$, and $\text{norm}(A,\text{inf})$, respectively. The 2-norm is the default in MatLab. The statement $\text{norm}(A)$ is interpreted as $\text{norm}(A,2)$ by MatLab. Since the 2-norm used in the majority of applications, we will adopt it as our default. **In what follows, an "un-designated" norm $\|A\|$ is to be interpreted as the 2-norm $\|A\|_2$.**

The Matrix 1-Norm

Recall that the vector 1-norm is given by

$$\|\vec{X}\|_1 = \sum_{i=1}^n |x_i|. \quad (4-7)$$

Subordinate to the vector 1-norm is the matrix 1-norm

$$\|A\|_1 = \max_j \left(\sum_i |a_{ij}| \right). \quad (4-8)$$

That is, the *matrix 1-norm is the maximum of the column sums*. To see this, let $m \times n$ matrix A be represented in the column format

$$A = [\vec{A}_1 \mid \vec{A}_2 \mid \cdots \mid \vec{A}_n]. \quad (4-9)$$

Then we can write

$$A\vec{X} = [\vec{A}_1 \mid \vec{A}_2 \mid \cdots \mid \vec{A}_n] \vec{X} = \sum_{k=1}^n \vec{A}_k x_k, \quad (4-10)$$

where x_k , $1 \leq k \leq n$, are the components of arbitrary vector \vec{X} . The triangle inequality and standard analysis applied to the norm of (4-10) yields


$$\begin{aligned}
 \|A\vec{X}\|_1 &= \left\| \sum_{k=1}^n \bar{A}_k x_k \right\|_1 \leq \sum_{k=1}^n |x_k| \|\bar{A}_k\|_1 \\
 &\leq \max_j \|\bar{A}_j\|_1 \left(\sum_{k=1}^n |x_k| \right) \\
 &= \max_j \|\bar{A}_j\|_1 \|\vec{X}\|_1.
 \end{aligned} \tag{4-11}$$

With the development of (4-11), we have completed step #1 for computing the matrix 1-norm. That is, we have found a constant

$$\mathbf{K} = \max_j \|\bar{A}_j\|_1$$

such that $\|A\vec{X}\|_1 \leq \mathbf{K} \|\vec{X}\|_1$ for all \vec{X} . Step 2 requires us to find at least one vector for which we have equality in (4-11). But this is easy; to maximize $\|A\vec{X}\|_1$, it is natural to select the \vec{X}_0 that puts "all of the allowable weight" on the component that will "pull-out" the maximum column sum. That is, the "optimum" vector is

$$\vec{X}_0 = [0 \quad 0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0]$$


 $j^{\text{th}}\text{-position}$

where j is the index that satisfies

$$\|\bar{A}_j\|_1 = \max_k \|\bar{A}_k\|_1$$

(the sum of the magnitudes in the j^{th} column is equal to, or larger than, the sum of the magnitudes in any column). When \vec{X}_0 is used, we have equality in (4-11), and we have completed step #2, so (4-8) is the matrix 1-norm.

The Matrix ∞ -Norm

Recall that the vector ∞ -norm is given by

$$\|\vec{X}\|_{\infty} = \max_k |x_k|, \quad (4-12)$$

the vector's largest component. Subordinate to the vector ∞ -norm is the matrix ∞ -norm

$$\|A\|_{\infty} = \max_i \left(\sum_j |a_{ij}| \right). \quad (4-13)$$

That is, the ***matrix ∞ -norm is the maximum of the row sums***. To see this, let A be an arbitrary $m \times n$ matrix and compute

$$\begin{aligned} \|\vec{A}\vec{X}\|_{\infty} &= \left\| \begin{array}{c} \sum_k a_{1k} x_k \\ \sum_k a_{2k} x_k \\ \vdots \\ \sum_k a_{mk} x_k \end{array} \right\|_{\infty} = \max_i \left| \sum_k a_{ik} x_k \right| \leq \max_i \left(\sum_k |a_{ik} x_k| \right) \\ &\leq \max_i \left(\sum_k |a_{ik}| \right) \max_k |x_k| \end{aligned} \quad (4-14)$$

$$= \max_i \left(\sum_k |a_{ik}| \right) \|\vec{X}\|_\infty$$

We have completed the first step. We have found a constant $\mathbf{K} = \max_i \left(\sum_k |a_{ik}| \right)$ for which

$$\|A\vec{X}\|_\infty \leq \mathbf{K} \|\vec{X}\|_\infty \quad (4-15)$$

for all \vec{X} .

Step #2 requires that we find a non-zero \vec{X}_0 for which equality holds in (4-14) and (4-15).

A close examination of these formulas leads to the conclusion that equality prevails if \vec{X}_0 is defined to have the components

$$\begin{aligned} x_k &= \frac{\bar{a}_{ik}}{|a_{ik}|}, \quad a_{ik} \neq 0, \quad 1 \leq k \leq n, \\ &= 1, \quad a_{ik} = 0, \end{aligned} \quad (4-16)$$

(the overbar denotes complex conjugate) where i is the index for the maximum row sum. That is, in (4-16), use index i for which

$$\sum_k |a_{ik}| \geq \sum_k |a_{jk}|, \quad i \neq j. \quad (4-17)$$

Hence, (4-13) is the matrix ∞ -norm as claimed.

The Matrix 2-Norm

Recall that the vector 2-norm is given by

$$\|\vec{X}\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2} = \sqrt{\langle \vec{X}, \vec{X} \rangle}. \quad (4-18)$$

Subordinate to the vector 2-norm is the matrix 2-norm

$$\|A\|_2 = \sqrt{\text{largest eigenvalue of } A^*A}. \quad (4-19)$$

Due to this connection with eigenvalues, the matrix 2-norm is called the *spectral norm*.

To see (4-19) for an arbitrary $m \times n$ matrix A , note that A^*A is $n \times n$ and Hermitian. By Theorem 4.2.1 (see Appendix 4.1), the eigenvalues of A^*A are real-valued. Also, A^*A is at least positive semi-definite since $\vec{X}^*(A^*A)\vec{X} = (A\vec{X})^*(A\vec{X}) \geq 0$ for all \vec{X} . Hence, the eigenvalues of A^*A are both real-valued and non-negative; denote them as

$$\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq \dots \geq \sigma_n^2 \geq 0. \quad (4-20)$$

Note that these eigenvalues are arranged according to size with σ_1^2 being the largest. These eigenvalues are known as the *singular values* of matrix A .

Corresponding to these eigenvalues are n orthonormal (hence, they are independent) eigenvectors $\vec{U}_1, \vec{U}_2, \dots, \vec{U}_n$ with

$$(A^*A)\vec{U}_k = (\sigma_k^2)\vec{U}_k, \quad 1 \leq k \leq n. \quad (4-21)$$

The n eigenvectors form the columns of a unitary $n \times n$ matrix U that diagonalizes matrix A^*A under similarity (matrix $U^*(A^*A)U$ is diagonal with eigenvalues (4-20) on the diagonal).

Since the n eigenvectors $\vec{U}_1, \vec{U}_2, \dots, \vec{U}_n$ are independent, they can be used as a basis, and vector \vec{X} can be expressed as

$$\vec{X} = \sum_{k=1}^n c_k \vec{U}_k, \quad (4-22)$$

where the c_k are \vec{X} -dependent constants. Multiply \vec{X} , in the form of (4-22), by A^*A to obtain

$$A^*A\vec{X} = A^*A \sum_{k=1}^n c_k \vec{U}_k = \sum_{k=1}^n c_k \sigma_k^2 \vec{U}_k, \quad (4-23)$$

which leads to

$$\begin{aligned} \|\vec{X}\|_2^2 &= (A\vec{X})^* A\vec{X} = \vec{X}^* (A^*A\vec{X}) = \left(\sum_{k=1}^n c_k^* \vec{U}_k^* \right) \left(\sum_{j=1}^n c_j \sigma_j^2 \vec{U}_j \right) = \sum_{k=1}^n |c_k|^2 \sigma_k^2 \\ &\leq \sigma_1^2 \left(\sum_{k=1}^n |c_k|^2 \right) \\ &= \sigma_1^2 \|\vec{X}\|_2^2 \end{aligned} \quad (4-24)$$

for arbitrary \vec{X} . Hence, we have completed step#1: we found a constant $\mathbf{K} = \sqrt{\sigma_1^2}$ such that $\|A\vec{X}\|_2 \leq \mathbf{K} \|\vec{X}\|_2$ for all \vec{X} . Step#2 requires us to find at least one vector \vec{X}_0 for which equality holds; that is, we must find an \vec{X}_0 with the property that $\|A\vec{X}_0\|_2 = \mathbf{K} \|\vec{X}_0\|_2$. But, it is obvious that $\vec{X}_0 = \vec{U}_1$, the unit-length eigenvector associated with eigenvalue σ_1^2 , will work. Hence, the matrix 2-norm is given by $\|A\|_2 = \sqrt{\sigma_1^2}$, the square root of the largest eigenvalue of A^*A .

The 2-norm is the default in MatLab. Also, it is the default here. **From now on, unless specified otherwise, the 2-norm is assumed: $\|A\|$ means $\|A\|_2$.**

Example

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}$$

Find $\|A\|_2$, and find the unit-length vector \vec{X}_0 that maximizes $\|A\vec{X}\|_2$. First, compute the product

$$A^*A = A^T A = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}.$$

The eigenvalues of this matrix are $\sigma_1^2 = 6.8541$ and $\sigma_2^2 = .1459$; note that $A^T A$ is positive definite symmetric. The 2-norm of A is $\|A\|_2 = \sqrt{\sigma_1^2} = 2.618$, the square root of the largest eigenvalue of $A^T A$.

The corresponding eigenvectors are $\vec{U}_1 = [-.8507 \ -.5257]$ and $\vec{U}_2 = [.5257 \ -.8507]$. They are columns in an orthogonal matrix $U = [\vec{U}_1 \ \vec{U}_2]$; note that $U^T U = I$ or $U^T = U^{-1}$. Furthermore, matrix U diagonalizes $A^T A$

$$U^T (A^T A) U = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 6.8541 & 0 \\ 0 & .1459 \end{bmatrix}$$

The unit-length vector that maximizes $\|A\vec{X}\|_2$ is $\vec{X}_0 = \vec{U}_1$, and

$$\max_{\|\vec{X}\|=1} \|A\vec{X}\|_2 = \sqrt{\sigma_1^2} = \sqrt{6.8541}$$

p-Norm of Matrix Product

For $m \times n$ matrix A we know that

$$\|A\vec{X}\|_p \leq \|A\|_p \|\vec{X}\|_p \tag{4-25}$$

for all \vec{X} . Now, consider $m \times n$ matrix A and $n \times q$ matrix B . The product AB is $m \times q$. For q -dimensional \vec{X} , we have

$$\|AB\vec{X}\|_p = \|A(B\vec{X})\|_p \leq \|A\|_p \|B\vec{X}\|_p \leq \|A\|_p \|B\|_p \|\vec{X}\|_p. \quad (4-26)$$

by applying (4-25) twice in a row. Hence, we have

$$\frac{\|AB\vec{X}\|_p}{\|\vec{X}\|_p} \leq \|A\|_p \|B\|_p \quad (4-27)$$

for all non-zero \vec{X} . As a result, it follows that

$$\|AB\|_p \leq \|A\|_p \|B\|_p, \quad (4-28)$$

a useful, simple result.

2-Norm Bound

Let A be an $m \times n$ matrix with elements a_{ij} , $1 \leq i \leq m$, $1 \leq j \leq n$. Let

$$\max_{i,j} |a_{ij}|$$

denote the magnitude of the largest element in A . Let (i_0, j_0) be the indices of the largest element so that

$$|a_{i_0 j_0}| \geq |a_{ij}| \quad (4-29)$$

for all i, j .

Theorem 4.1 (2-norm bound):

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq \sqrt{mn} \max_{i,j} |a_{ij}| \quad (4-30)$$

Proof: Note that

$$\|A\vec{X}\|_2 = \sqrt{\left|\sum_k a_{1k}x_k\right|^2 + \left|\sum_k a_{2k}x_k\right|^2 + \cdots + \left|\sum_k a_{mk}x_k\right|^2} \quad (4-31)$$

$$\|A\|_2 = \max_{\|\vec{X}\|_2=1} \|A\vec{X}\|_2.$$

Define \vec{X}_0 as the vector with 1 in the j_0 position and zero elsewhere (see (4-29) for definition of i_0 and j_0). Since

$$\|A\vec{X}_0\|_2 \leq \max_{\|\vec{X}\|_2=1} \|A\vec{X}\|_2 = \|A\|_2 \quad (4-32)$$

we have

$$|a_{i_0 j_0}| \leq \sqrt{\sum_k |a_{k j_0}|^2} \leq \max_{\|\vec{X}\|_2=1} \|A\vec{X}\|_2 = \|A\|_2 \quad (4-33)$$

Hence, we have

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \quad (4-34)$$

as claimed. Now, we show the remainder of the 2-norm bound. Observe that

$$\begin{aligned}
\|A\|_2 &= \max_{\|\vec{X}\|_2=1} \sqrt{\left| \sum_k a_{1k} x_k \right|^2 + \left| \sum_k a_{2k} x_k \right|^2 + \cdots + \left| \sum_k a_{mk} x_k \right|^2} \\
&\leq \max_{\|\vec{X}\|_2=1} \sqrt{\sum_k |a_{1k} x_k|^2 + \sum_k |a_{2k} x_k|^2 + \cdots + \sum_k |a_{mk} x_k|^2} \\
&\leq \sqrt{\sum_k |a_{1k}|^2 + \sum_k |a_{2k}|^2 + \cdots + \sum_k |a_{mk}|^2} \\
&\leq \sqrt{mn} \max_{i,j} |a_{ij}|.
\end{aligned} \tag{4-35}$$

When combined with (4-34), we have $\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq \sqrt{mn} \max_{i,j} |a_{ij}|$.

Triangle Inequality for the p-Norm

Recall the Triangle inequality for real numbers: $|\alpha + \beta| \leq |\alpha| + |\beta|$. A similar result is valid for the matrix p-norm.

Theorem 4.2 (Triangle Inequality for the matrix 2-norm)

Let A and B be $m \times n$ matrices. Then

$$\|A + B\|_p \leq \|A\|_p + \|B\|_p \tag{4-36}$$

Application of Matrix Norm: Inverse of Matrices “Close” to a Non-singular Matrix.

Let A be an $n \times n$ non-singular matrix. Can we state conditions on the “size” (i.e., norm) of $n \times n$ matrix E which guarantees that $A+E$ is nonsingular? Before developing these conditions, we derive a closely related result, the geometric series for matrices. Recall the geometric series

$$\frac{1}{1-x} = \sum_{k=1}^{\infty} x^k, \quad |x| < 1 \tag{4-37}$$

for real variables. We are interested in the “matrix version” of this result.

Theorem 4.3 (Geometric Series for Matrices)

Let F be any $n \times n$ matrix with $\|F\|_2 < 1$. Let I denote the $n \times n$ identity matrix. Then, the difference $I - F$ is nonsingular and

$$(I - F)^{-1} = \sum_{k=0}^{\infty} F^k \quad (4-38)$$

with

$$\|(I - F)^{-1}\|_2 \leq \frac{1}{1 - \|F\|_2}. \quad (4-39)$$

Proof: First, we show that $I - F$ is nonsingular. Suppose that this is not true; suppose that $I - F$ is singular. Then there exists at least one non-zero \vec{X} such that $(I - F)\vec{X} = 0$ so that $\|\vec{X}\|_2 = \|F\vec{X}\|_2$. But since $\|F\vec{X}\|_2 \leq \|F\|_2 \|\vec{X}\|_2$, we must have $\|\vec{X}\|_2 = \|F\vec{X}\|_2 \leq \|F\|_2 \|\vec{X}\|_2$ which requires that $\|F\|_2 \geq 1$, a contradiction. Hence, the $n \times n$ matrix $I - F$ must be nonsingular. To obtain a series for $(I - F)^{-1}$, consider the obvious identity

$$\left(\sum_{k=0}^N F^k \right) (I - F) = I - F^{N+1} \quad (4-40)$$

However, $\|F^k\|_2 \leq \|F\|_2^k$, and since $\|F\|_2 < 1$, we have $F^k \rightarrow 0$ as $k \rightarrow \infty$. As a result of this,

$$\lim_{N \rightarrow \infty} \left(\sum_{k=0}^N F^k \right) (I - F) = I \quad (4-41)$$

so that $(I - F)^{-1} = \sum_{k=0}^{\infty} F^k$ as claimed. Finally

$$\begin{aligned}
\|(I-F)^{-1}\|_2 &= \left\| \sum_{k=0}^{\infty} F^k \right\|_2 \leq \sum_{k=0}^{\infty} \|F^k\|_2 \\
&\leq \sum_{k=0}^{\infty} \|F\|_2^k \quad (\text{this is an "ordinary" scalar geometric series}). \\
&= \frac{1}{1 - \|F\|_2}
\end{aligned} \tag{4-42}$$

as claimed. ♥

In many applications, Theorem 4.3 is used in the form

$$(I - \varepsilon F)^{-1} = \sum_{k=0}^{\infty} F^k \varepsilon^k \quad \text{for } |\varepsilon| < 1 / \|F\|_2, \tag{4-43}$$

where ε is considered to be a "small" parameter. Next, we generalize Theorem 4.3 and obtain a result that will be used in our study of how small errors in A and \vec{b} influence the solution of the linear algebraic problem $A\vec{X} = \vec{b}$.

Theorem 4-4

Assume that A is $n \times n$ and nonsingular. Let E be an arbitrary $n \times n$ matrix. If

$$\rho = \|A^{-1}E\|_2 < 1, \tag{4-44}$$

then $A + E$ is nonsingular and

$$\|(A + E)^{-1} - A^{-1}\|_2 \leq \frac{\|E\|_2 \|A^{-1}\|_2^2}{1 - \rho} \tag{4-45}$$

Proof: Since A is nonsingular, we have

$$A + E = A(I - F), \text{ where } F \equiv -A^{-1}E. \quad (4-46)$$

Since $\rho = \|F\|_2 = \|A^{-1}E\|_2 < 1$, it follows from Theorem 4.3 that $I - F$ is nonsingular and

$$\|(I - F)^{-1}\|_2 \leq \frac{1}{1 - \|F\|_2} = \frac{1}{1 - \rho}. \quad (4-47)$$

From (4-46) we have $(A + E)^{-1} = (I - F)^{-1}A^{-1}$; with the aid of (4-47), this can be used to write

$$\|(A + E)^{-1}\|_2 \leq \frac{\|A^{-1}\|_2}{1 - \rho}. \quad (4-48)$$

Now, multiply both sides of

$$(A + E)^{-1} - A^{-1} = -A^{-1}E(A + E)^{-1} \quad (4-49)$$

by $A + E$ to see this matrix identity. Finally, take the norm of (4-49) to obtain

$$\begin{aligned} \|(A + E)^{-1} - A^{-1}\|_2 &= \|A^{-1}E(A + E)^{-1}\|_2 \\ &\leq \|A^{-1}\|_2 \|E\|_2 \|(A + E)^{-1}\|_2 \end{aligned} \quad (4-50)$$

Now use (4-48) in this last result to obtain

$$\|(A + E)^{-1} - A^{-1}\|_2 \leq \|A^{-1}\|_2 \|E\|_2 \frac{\|A^{-1}\|_2}{1 - \rho} = \frac{\|A^{-1}\|_2^2 \|E\|_2}{1 - \rho} \quad (4-51)$$

as claimed.♥

We will use Theorem 4.4 when studying the sensitivity of the linear equation $A\vec{X} = \vec{b}$. That is, we want to relate changes in solution \vec{X} to small changes (errors) in both A and \vec{b} .