

EE731 Lecture Notes: Matrix Computations for Signal Processing

James P. Reilly©
Department of Electrical and Computer Engineering
McMaster University

November 2, 2006

Lecture 4

In this lecture, we first discuss the *quadratic form* associated with a matrix. We look at three relevant examples where the quadratic form is used in signal processing: the idea of *positive definiteness*, the *Gaussian multivariate probability density function*, and the *Rayleigh quotient*.

We then briefly discuss floating point number systems in computers, and we investigate the effect of these errors in algebraic systems. Specifically, we look at the important idea of the *condition number* of a matrix. Then, we look at some methods of implementing matrix operations using highly parallel computational architectures, called *systolic arrays*. These have the advantage of very fast execution times and relatively simple implementations.

5 The Quadratic Form

We introduce the quadratic form by considering the idea of *positive definiteness* of a matrix \mathbf{A} . A square matrix $\mathbf{A} \in \mathfrak{R}^{n \times n}$ is *positive definite* if and

only if, for any $0 \neq \mathbf{x} \in \mathfrak{R}^n$,

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0. \quad (1)$$

The matrix \mathbf{A} is *positive semi-definite* if and only if, for any $\mathbf{x} \neq \mathbf{0}$ we have

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0, \quad (2)$$

which, as we see later, includes the possibility that \mathbf{A} is rank deficient. The quantity on the left in (1) is referred to as a *quadratic form* of \mathbf{A} . It may be verified by direct multiplication that the quadratic form can also be expressed in the form

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j. \quad (3)$$

It is only the symmetric part of \mathbf{A} which is relevant in a quadratic form expression. This fact may be verified as follows. We can define the symmetric part \mathbf{T} of \mathbf{A} as $\mathbf{T} \triangleq \frac{1}{2}[\mathbf{A} + \mathbf{A}^T]$, and the asymmetric part \mathbf{S} of \mathbf{A} as $\mathbf{S} \triangleq \frac{1}{2}[\mathbf{A} - \mathbf{A}^T]$. Then we have the desired properties that $\mathbf{T}^T = \mathbf{T}$, $\mathbf{S} = -\mathbf{S}^T$, and $\mathbf{A} = \mathbf{T} + \mathbf{S}$.

We can express (3) as

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n t_{ij} x_i x_j + \sum_{i=1}^n \sum_{j=1}^n s_{ij} x_i x_j. \quad (4)$$

We now consider only the second term on the right in (4):

$$\sum_{i=1}^n \sum_{j=1}^n s_{ij} x_i x_j. \quad (5)$$

Since $\mathbf{S} = -\mathbf{S}^T$, the quantity $s_{ij} = -s_{ji}$, $j \neq i$, and $s_{ij} = 0$, $i = j$. Therefore, the sum in (5) is zero. Thus, when considering quadratic forms, it suffices to consider only the symmetric part \mathbf{T} of the matrix; i.e., we have the result $\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{x}^T \mathbf{T} \mathbf{x}$.

This result generalizes to the case where \mathbf{A} is complex. It is left as an exercise to show that i) $\mathbf{x}^H \mathbf{A} \mathbf{x} = \mathbf{x}^H \mathbf{T} \mathbf{x}$, where $\mathbf{T} \triangleq \frac{1}{2}[\mathbf{A} + \mathbf{A}^H]$, and ii) that the quantity $\mathbf{x}^H \mathbf{A} \mathbf{x}$ is pure real.

Quadratic forms on positive definite matrices are used very frequently in least-squares and adaptive filtering applications. Also as we see later, quadratic forms play a fundamental role in defining the multivariate Gaussian probability density function.

Theorem 1 *A matrix \mathbf{A} is positive definite if and only if all eigenvalues of the symmetric part of \mathbf{A} are positive.*

Proof: Let the eigendecomposition on the symmetric part \mathbf{T} of \mathbf{A} be represented as $\mathbf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. Since only the symmetric part of \mathbf{A} is relevant, the quadratic form on \mathbf{A} may be expressed as $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{T}\mathbf{x} = \mathbf{x}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{x}$. Let us define the variable \mathbf{z} as $\mathbf{z} \triangleq \mathbf{V}^T\mathbf{x}$. As we have seen previously in Chapters 1 and 2, \mathbf{z} is a rotation of \mathbf{x} due to the fact \mathbf{V} is orthonormal. Thus we have

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{z}^T\mathbf{\Lambda}\mathbf{z} \\ &= \sum_{i=1}^n z_i^2 \lambda_i.\end{aligned}\tag{6}$$

Thus (6) is greater than zero for arbitrary \mathbf{x} if and only if $\lambda_i > 0, i = 1, \dots, n$.

□.

We also see from (6) that if the equality in the quadratic form is satisfied, ($\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ for some \mathbf{x} and corresponding \mathbf{z}) then at least one eigenvalue of \mathbf{T} must be zero. Hence, if \mathbf{A} is symmetric, then \mathbf{A} being positive *semidefinite* implies that at least one eigenvalue of \mathbf{A} is zero, which means that \mathbf{A} is rank deficient.

5.1 The Locus of Points $\{z|z^T \Lambda z = 1\}$

Let us assume that \mathbf{A} is positive definite. Then quantity $z^T \Lambda z$ can be written as

$$\begin{aligned} z^T \Lambda z &= \sum_{i=1}^n z_i^2 \lambda_i \\ &= \sum_{i=1}^n \frac{z_i^2}{\frac{1}{\lambda_i}}. \end{aligned} \tag{7}$$

Eq. (7) is the canonical form of an ellipse in the variables z_i , with principal axis lengths $\sqrt{\frac{1}{\lambda_i}}$. The principal axes are aligned along the corresponding elementary basis directions $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$.

Since $\mathbf{z} = \mathbf{V}^T \mathbf{x}$ where \mathbf{V} is orthonormal, the locus of points $\{\mathbf{x}|\mathbf{x}^T \mathbf{A} \mathbf{x} = 1\}$ is a rotated version of the ellipse in (7). This ellipse has the same principal axes lengths as before, but the i th principal axis now lines up along the i th eigenvector \mathbf{v}_i of \mathbf{A} .

The locus of points $\{\mathbf{x}|\mathbf{x}^T \mathbf{A} \mathbf{x} = k, k > 0\}$, defines a scaled version of the ellipse above. In this case, the i th principal axis length is given by the quantity $\sqrt{\frac{k}{\lambda_i}}$.

Example: We now discuss an example to illustrate the above discussion. A three-dimensional plot of $y = \mathbf{x}^T \mathbf{A} \mathbf{x}$ is shown plotted in Fig. 1 for \mathbf{A} given by

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{8}$$

The corresponding contour plot is plotted in Fig. 2. Note that this curve is elliptical in cross-section in a plane $y = k$ as discussed above. A calculation verifies the eigenvalues of \mathbf{A} are 3, 1 with corresponding eigenvectors $[1, 1]^T$ and $[1, -1]^T$. For $y = k = 1$, the lengths of the principal axes of the ellipse are then $1/\sqrt{3}$ and 1. It can be verified from the figure these principal axis lengths are indeed the lengths indicated, and are lined up along the directions of the eigenvectors as required.

Positive definiteness of \mathbf{A} in the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ is the matrix analog

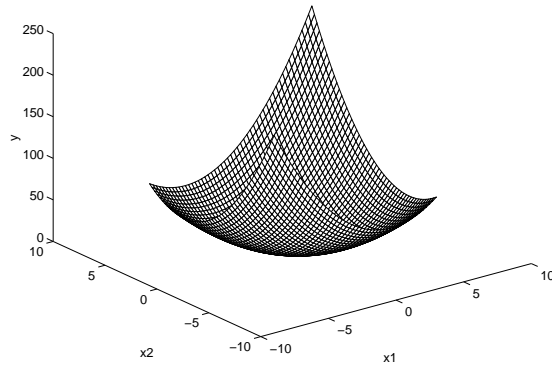


Figure 1: Three-dimensional plot of quadratic form.

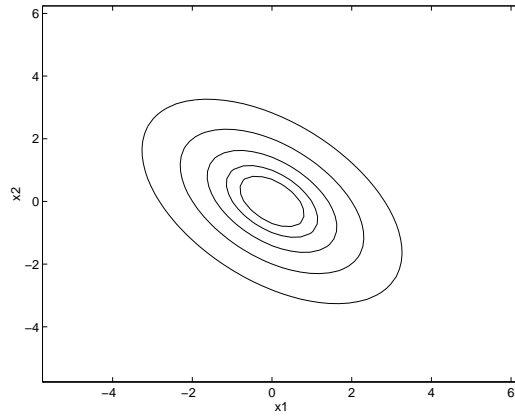


Figure 2: Plots of $\mathbf{x}^T \mathbf{A} \mathbf{x} = k$, for $k = 1, 2, 4, 8$ and 16 .

to the scalar a being positive in the scalar expression ax^2 . The scalar equation $y = ax^2$ is a parabola which faces upwards if a is positive. Likewise, the equation $y = \mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n z_i^2 \lambda_i$, where $\mathbf{z} = \mathbf{V} \mathbf{x}$ as before, is a multi-dimensional parabola. The parabola faces upwards in all directions if \mathbf{A} is positive definite. If \mathbf{A} is not positive (semi) definite, then some eigenvalues are negative and the curve faces down in the orientations corresponding to the negative eigenvalues, and up in those corresponding to the positive eigenvalues.

Theorem 2 *A (square) symmetric matrix \mathbf{A} can be decomposed into the form $\mathbf{A} = \mathbf{B}\mathbf{B}^T$ if and only if \mathbf{A} is positive definite or positive semi-definite.*

Proof: (Necessary condition; i.e., if $\mathbf{A} = \mathbf{B}\mathbf{B}^T$, then \mathbf{A} is positive definite.) Let us define \mathbf{z} as $\mathbf{B}^T \mathbf{x}$. Then

$$\begin{aligned} \mathbf{x}^T \mathbf{A} \mathbf{x} &= \mathbf{x}^T \mathbf{B}\mathbf{B}^T \mathbf{x} \\ &= \mathbf{z}^T \mathbf{z} \\ &\geq 0. \end{aligned} \tag{9}$$

Conversely (sufficient condition): Since \mathbf{A} is symmetric, we can write \mathbf{A} as $\mathbf{A} = \mathbf{V}^T \mathbf{\Lambda} \mathbf{V}$. Since \mathbf{A} is positive definite by hypothesis, we can write $\mathbf{A} = (\mathbf{V} \mathbf{\Lambda}^{1/2})(\mathbf{V} \mathbf{\Lambda}^{1/2})^T$. Let us define $\mathbf{B} \triangleq \mathbf{V} \mathbf{\Lambda}^{1/2} \mathbf{Q}^T$ where \mathbf{Q} is a matrix of appropriate size whose columns are orthonormal, such that $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$. Then $\mathbf{A} = \mathbf{V} \mathbf{\Lambda}^{1/2} \mathbf{Q}^T \mathbf{Q} \mathbf{\Lambda}^{1/2} \mathbf{V}^T = \mathbf{B}\mathbf{B}^T$.

□

Recall that \mathbf{A} is $n \times n$; thus, \mathbf{Q} can be of size $m \times n$, where $m \geq n$. It thus is clear that \mathbf{Q} is not unique, and therefore it follows this factorization of \mathbf{A} is unique only up to an orthogonal ambiguity.

The fact that \mathbf{A} can be decomposed into two symmetric factors in this way is the fundamental idea behind the Cholesky factorization, which is a major topic of the following chapter.

5.2 Differentiation of the Quadratic Form

We see that the quadratic form is a scalar. To differentiate a scalar with respect to the vector \mathbf{x} , we differentiate with respect to each element of \mathbf{x} in turn, and then assemble all the results back into a vector. We proceed as follows:

We write the quadratic form as

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n x_i x_j a_{ij}. \quad (10)$$

When differentiating the above with respect to a particular element x_k , we need only consider the terms when either index i or j equals k . Therefore:

$$\frac{d}{dx_k} \mathbf{x}^T \mathbf{A} \mathbf{x} = \frac{d}{dx_k} \left[\sum_{\substack{j=1 \\ j \neq k}}^n x_k x_j a_{kj} + \sum_{\substack{i=1 \\ i \neq k}}^n x_i x_k a_{ik} + x_k^2 a_{kk} \right] \quad (11)$$

where the first term of (11) corresponds to holding i constant at the value k , and the second corresponds to holding j constant at k . Care must be taken to include the term $x_k^2 a_{kk}$ corresponding to $i = j = k$ only once; therefore, it is excluded in the first two terms and added in separately. Eq. (11) evaluates to

$$\begin{aligned} \frac{d}{dx_k} \mathbf{x}^T \mathbf{A} \mathbf{x} &= \sum_{j \neq k} x_j a_{kj} + \sum_{i \neq k} x_i a_{ik} + 2x_k a_{kk} \\ &= \sum_j x_j a_{kj} + \sum_i x_i a_{ik} \end{aligned}$$

We note the first term is the inner product of \mathbf{x} with the k th row of \mathbf{A} , whereas the second term is the inner product of \mathbf{x} with the k th column of \mathbf{A} . It is straightforward to show that the asymmetric part of \mathbf{A} cancels out in the above expression for the first derivative, just as it did for the quadratic form itself. We may therefore take \mathbf{A} as effectively symmetric, and the two terms above are then equal. This gives

$$\frac{d}{dx_k} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2(\mathbf{A} \mathbf{x})_k \quad (12)$$

where $(\cdot)_k$ denotes k^{th} element of the argument. By assembling these individual terms corresponding to $k = 1, \dots, n$ back into a vector, we have the result that

$$\frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A} \mathbf{x}. \quad (13)$$

It is interesting to find the stationary points of the quadratic form subject to a norm constraint; i.e., we seek the solution to

$$\max_{\|\mathbf{x}\|_2^2=1} \mathbf{x}^T \mathbf{A} \mathbf{x}. \quad (14)$$

To solve this, we form the Lagrangian

$$\mathbf{x}^T \mathbf{A} \mathbf{x} + \lambda(1 - \mathbf{x}^T \mathbf{x}). \quad (15)$$

Differentiating, and setting the result to zero (realizing that $d/d\mathbf{x} (\mathbf{x}^T \mathbf{x}) = 2\mathbf{x}$) gives

$$\mathbf{A} \mathbf{x} = \lambda \mathbf{x}. \quad (16)$$

Thus, the eigenvectors are stationary points of the quadratic form, and the \mathbf{x} which gives the maximum (or minimum), subject to a norm constraint, is the maximum (minimum) eigenvector of \mathbf{A} .

5.3 The Gaussian Multi-Variate Probability Density Function

Here, we very briefly introduce this topic so we can use this material for an example of the application of the Cholesky decomposition later in this course, and also in least-squares analysis to follow shortly. This topic is a good application of quadratic forms. More detail is provided in several books.¹

First we consider the uni-variate case of the Gaussian probability distribution function (*pdf*). The *pdf* $p(x)$ of a Gaussian-distributed random variable

¹e.g. H. Van Trees, "Detection, Estimation and Modulation Theory", Part 1. L.L. Scharf, Statistical Signal Processing: Detection, Estimation, and Time Series Analysis, pg. 55.

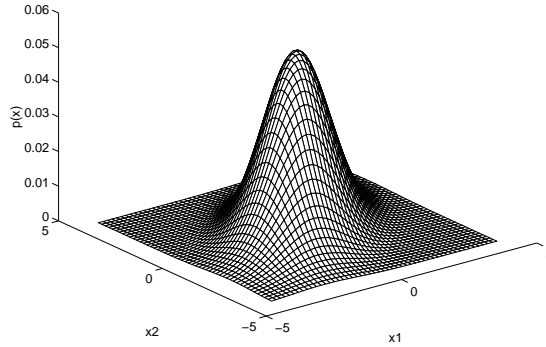


Figure 3: A Gaussian probability density function.

x with mean μ and variance σ^2 is given as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right]. \quad (17)$$

This is the familiar bell-shaped curve. It is completely specified by two parameters– the mean μ which determines the position of the peak, and the variance σ^2 which determines the width or spread of the curve.

We now consider the more interesting multi-dimensional case. Consider a Gaussian-distributed random vector $\mathbf{x} \in \Re^n$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The multivariate *pdf* describing \mathbf{x} is

$$p(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]. \quad (18)$$

We can see that the multi-variate case collapses to the uni-variate case when the number of variables becomes one. A plot of $p(\mathbf{x})$ vs. \mathbf{x} is shown in Fig. 3, for a mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ defined as

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \quad (19)$$

Because the exponent in (18) is a quadratic form, the set of points satisfied by the equation $[\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})] = k$ where k is a constant, is an ellipse. Therefore this ellipse defines a contour of equal probability density. The interior of this ellipse defines a region into which an observation will

fall with a specified probability α which is dependent on k . This probability level α is given as

$$\alpha = \int_{\mathcal{R}} (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right] d\mathbf{x}, \quad (20)$$

where \mathcal{R} is the interior of the ellipse. Stated another way, an *ellipse* is the region in which any observation governed by the probability distribution (18) will fall with a specified probability level α . As k increases, the ellipse gets larger, and α increases. These ellipses are referred to as *joint confidence regions* (JCRs) at probability level α .

The covariance matrix $\mathbf{\Sigma}$ controls the shape of the ellipse. Because the quadratic form in this case involves $\mathbf{\Sigma}^{-1}$, the length of the i th principal axis is $\sqrt{2k\lambda_i}$ instead of $\sqrt{2k/\lambda_i}$ as it would be if the quadratic form were in $\mathbf{\Sigma}$. Therefore as the eigenvalues of $\mathbf{\Sigma}$ increase, the size of the JCRs increase (i.e., the variances of the distribution increase) for a given value of k .

We now investigate the relationship of the covariances between the variables (i.e., off-diagonal terms of the covariance matrix) and the shape of the Gaussian pdf. We have seen previously in Lecture 2 that covariance is a measure of dependence between individual random variables. We have also seen that as the off-diagonal covariance terms become larger, there is a larger disparity between the largest and smallest eigenvalues of the covariance matrix. Thus, as the covariances increase, the eigenvalues, and thus the lengths of the semi-axes of the JCRs become more disparate; i.e., the JCRs of the Gaussian pdf become elongated. This behaviour is illustrated in Fig. 4, which shows a multi-variate Gaussian pdf for a mean $\boldsymbol{\mu} = \mathbf{0}$ and for a covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ given as

$$\mathbf{\Sigma}_2 = \begin{bmatrix} 2 & 1.9 \\ 1.9 & 2 \end{bmatrix}. \quad (21)$$

Note that in this case, the covariance elements of $\mathbf{\Sigma}_2$ have increased substantially relative to those of $\mathbf{\Sigma}_1$ in Fig. 3, although the variances themselves (the main diagonal elements) have remained unchanged. By examining the pdf of Figure 4, we see that the joint confidence ellipsoid has become elongated, as expected. (For $\mathbf{\Sigma}_1$ of Fig. 3 the eigenvalues are (3,1), and for $\mathbf{\Sigma}_2$ of Fig. 4, the eigenvalues are (3.9,0.1)). This elongation results in the conditional probability $p(x_1|x_2)$ for Fig. 4 having a much smaller variance (spread) than that for Fig. 3; i.e., when the covariances are larger, knowledge

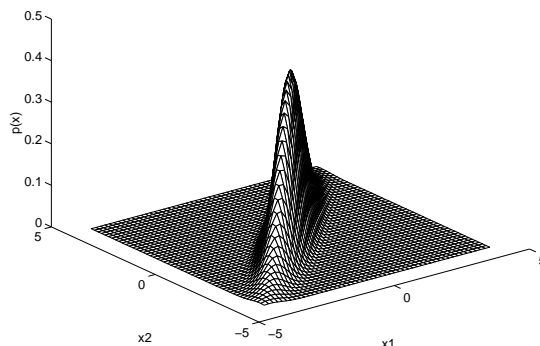


Figure 4: A Gaussian *pdf* with larger covariance elements.

of one variable tells us more about the other. This is how the probability density function incorporates the information contained in the covariances between the variables. With regard to Gaussian probability density functions, the following concepts: 1) larger correlations between the variables, 2) larger disparity between the eigenvalues, 3) elongated joint confidence regions, and 4) lower variances of the conditional probabilities, are all closely inter-related.

5.4 The Rayleigh Quotient

The *Rayleigh quotient* is a simple mathematical structure that has a great deal of interesting uses. The Rayleigh quotient $r(\mathbf{x})$ is defined as

$$r(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \quad (22)$$

It is easily verified that if \mathbf{x} is the i th eigenvector \mathbf{v}_i of \mathbf{A} , (not necessarily normalized to unit norm), then $r(\mathbf{x}) = \lambda_i$:

$$\begin{aligned} \frac{\mathbf{v}_i^T \mathbf{A} \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{v}_i} &= \frac{\lambda_i \mathbf{v}_i^T \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{v}_i} \\ &= \lambda_i. \end{aligned} \quad (23)$$

In fact, it can be shown by differentiating $r(\mathbf{x})$ with respect to \mathbf{x} , that $\mathbf{x} = \mathbf{v}_i$ is a stationary point of $r(\mathbf{x})$.

Further along this line of reasoning, let us define a subspace S_k as $S_k = [\mathbf{v}_1, \dots, \mathbf{v}_k]$, $k = 1, \dots, n$, where \mathbf{v}_i is the i th eigenvector of $\mathbf{A} \in \mathfrak{R}^{n \times n}$, where \mathbf{A} is symmetric. Then, the Courant Fischer minimax theorem ² says that

$$\lambda_k = \min_{0 \neq \mathbf{x} \in S_k} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \quad (24)$$

The Rayleigh quotient leads naturally to an iterative method for computing an eigenvalue/eigenvector of a symmetric matrix \mathbf{A} . If \mathbf{x} is an approximate eigenvector, then $r(\mathbf{x})$ gives us a reasonable approximation to the corresponding eigenvalue. Further, the inverse perturbation theory of Golub and Van Loan says that if u is an eigenvalue, then the solution to $(\mathbf{A} - u\mathbf{I})\mathbf{z} = \mathbf{b}$, where \mathbf{b} is an approximate eigenvector, gives us a better estimate of the eigenvector. These two ideas lead to the following *Rayleigh Quotient* technique for calculating an eigenvector/eigenvalue pair:

initialize \mathbf{x}_0 to an appropriate value; set $\|\mathbf{x}_0\|_2 = 1$.
 for $k = 0, 1, \dots$,
 $\mu_k = r(\mathbf{x}_k)$
 Solve $(\mathbf{A} - \mu_k \mathbf{I})\mathbf{z}_{k+1} = \mathbf{x}_k$ for \mathbf{z}_{k+1}
 $\mathbf{x}_{k+1} = \mathbf{z}_{k+1} / \|\mathbf{z}_{k+1}\|_2$

This procedure exhibits cubic convergence to the eigenvector. At convergence, μ is an eigenvalue, and \mathbf{z} is the corresponding eigenvector. Therefore the matrix $(\mathbf{A} - \mu\mathbf{I})$ is singular and \mathbf{z} is in its nullspace. The solution \mathbf{z} becomes extremely large and the system of equations $(\mathbf{A} - \mu\mathbf{I})\mathbf{z} = \mathbf{x}$ is satisfied only because of numerical error. Nevertheless, accurate values of the eigenvalue and eigenvector are obtained.

6 Floating Point Arithmetic Systems

A real number x can be represented in floating point form (denoted $fl(x)$) as

$$fl(x) = s \cdot f \cdot b^k \quad (25)$$

²See Wilkinson, "The Algebraic Eigenvalue Problem", pp. 100 – 101.

where

$$\begin{aligned}
 s &= \text{sign bit} = \pm 1 \\
 f &= \text{fractional part of } x \text{ of length } t \text{ bits} \\
 b &= \text{machine base} = 2 \text{ for binary systems} \\
 k &= \text{exponent}
 \end{aligned}$$

Note that the operation $\text{fl}(x)$ (i.e., conversion from a real number x to its floating point representation) maps a real number x into a set of *discrete* points on the real number line. These points are determined by (25). This mapping has the property that the separation between points is small for $|x|$ small, and large for $|x|$ large. Because the operation $\text{fl}(x)$ maps a continuous range of numbers into a discrete set, there is error associated with the representation $\text{fl}(x)$.

In the conversion process, the exponent is adjusted so that the most significant bit (msb) of the fractional part is 1, and so that the binary point is immediately to the right of the msb. For example, the binary number

$$x = .0000100111101011011 \quad (26)$$

could be represented as a floating point number with $t = 9$ bits as:

$$1.00111101 \times 2^{-5}.$$

Since it is known that the msb of the fractional part is a one, it does not need to be present in the actual floating-point number. This way, we get an extra bit, “for free”. This means the number x in (26) may be represented as

$$\underbrace{00111101}_f \times 2^{-5}.$$

↑ leading 1 assumed present

This above form only takes 8 bits instead of 9 to represent $\text{fl}(x)$ with the same precision.

The range of possible real numbers which can be mapped into the representation $|\text{fl}(x)|$ is:

$$1.00 \dots 00 \times 2^L \leq |\text{fl}(x)| \leq \left. \begin{array}{c} | \leftarrow t \text{ bits} \rightarrow | \\ 1.111111 \dots 1 \end{array} \right\} \times 2^U$$

where L and U are the minimum and maximum values of the exponent, respectively. Note that any arithmetic operation which produces a result outside of these bounds results in a floating point overflow or underflow error.

Note that because the leading one in the msb position is absent, it is now impossible to represent the number zero. Thus, a special convention is needed. This is usually done by reserving a special value of the exponent field.

6.1 Machine Epsilon u

Since the operation $\text{fl}(x)$ maps the set of real numbers into a discrete set, the quantity $\text{fl}(x)$ involves error. The quantity *machine epsilon*, represented by the symbol u is the maximum relative error possible in $\text{fl}(x)$.

The relative error ϵ_r in the quantity $\text{fl}(x)$ is given by

$$\epsilon_r = \frac{|\text{fl}(x) - x|}{|x|} \quad (27)$$

But u is the maximum relative error. Therefore

$$\begin{aligned} u = \max \epsilon_r &= \frac{\max |\text{fl}(x) - x|}{\min |x|} \\ &= \frac{0.00\dots0 \overset{|\leftarrow t\text{bits}\rightarrow|}{1111111111\dots}}{1} \\ &\simeq 2^{1-t} \end{aligned}$$

if the machine chops. By “chopping”, we mean the machine constructs the fractional part of $\text{fl}(x)$ by retaining only the most significant t bits, and truncating the rest. If the machine rounds, then the relative error is one half that due to chopping; hence

$$u = 2^{-t}$$

if the machine rounds.

Thus, the number $\text{fl}(x)$ may be represented as $\text{fl}(x) = x(1 + \epsilon)$ where $|\epsilon| \leq u$.

In an actual computer implementation, s is a single bit (usually 0 to indicate a positive number, and 1 to represent a negative number). In single precision, the total length of a floating point number is typically 32 bits. Of these, 8 are used for the exponent k , one for s , leaving 23 for the fractional part f ($t = 24$ bits effective precision). This means for single precision arithmetic with chopping, $u = 2^{-23} = 1.19 \times 10^{-7}$.

6.2 Catastrophic Cancellation

Significant reduction in precision may result when subtracting two nearly equal floating-point numbers. If the fractional part of two numbers A and B are identical in their first r digits ($r \leq t$), then $\text{fl}(A - B)$ has only $t - r$ bits significance; i.e., we have lost r bits of significance in representing the difference. As r approaches t , the difference has very few significant bits in its fractional part. This reduction in precision is referred to as *catastrophic cancellation*, and can be the cause of serious numerical problems in floating point computational systems.

We can demonstrate this phenomenon by example as follows: Let A and B be two numbers whose fractional parts are identical in their first $r = 7$ digits. Then for the case $t = 10$ and $b = 2$ (binary arithmetic)

$$\begin{aligned} \text{frac}(A) &= \begin{array}{c} \leftarrow r \text{ bits} \rightarrow \\ 1011011 \quad 101 \end{array} \\ \text{frac}(B) &= \begin{array}{c} \leftarrow r \text{ bits} \rightarrow \\ 1011011 \quad 001 \end{array} \end{aligned}$$

where $\text{frac}(\cdot)$ is the fractional part of the number. Because the numbers are nearly equal, it may be assumed that their exponents have the same value. Then, we see that the difference $\text{frac}(A - B)$ is $(100)_2$, which has only $t - r = 3$ bits significance. We have lost 7 bits of significance in representing the difference, which results in a drastic increase in u . Thus the difference can be in significant error.

Another example of catastrophic cancellation is as follows: Find roots of the quadratic equation

$$x^2 + 1958.63x + 0.00253 = 0$$

Solution:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (28)$$

computed roots: $x_1 = -1958.62998$, $x_2 = -0.00000150$

true roots: $x_1 = -1958.6299$, $x_2 = -0.0000012917$

There are obviously serious problems with the accuracy of x_2 , which corresponds to the “+” sign in (28) above. In this case, since $b^2 \gg 4ac$, $\sqrt{b^2 - 4ac} \simeq b$. Hence, we are subtracting two nearly equal numbers when calculating x_2 , which results in catastrophic cancellation.

Another example of catastrophic cancellation is evaluating the inner product of two nearly orthogonal vectors. This is because we are adding a group of not necessarily small numbers whose sum is small. This operation implicitly involves subtracting two nearly equal numbers.

It is shown in *Golub and Van Loan*, that

$$|\text{fl}(\mathbf{x}^T \mathbf{y}) - \mathbf{x}^T \mathbf{y}| \leq nu |\mathbf{x}|^T |\mathbf{y}| + O(u^2) \quad (29)$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $O(u^2)$, read “order u squared”, indicates the presence of terms in u^2 and higher, which can be ignored due to the fact they may be considered small in comparison to the first-order term in u . Hence (29) tells us that if $|\mathbf{x}^T \mathbf{y}| \ll |\mathbf{x}|^T |\mathbf{y}|$, which happens when \mathbf{x} is nearly orthogonal to \mathbf{y} , then the relative error in $\text{fl}(\mathbf{x}^T \mathbf{y})$ may not be small.

Fix: If the partial products are accumulated in a *double precision* register (length of fractional part = $2t$), little error results. This is because multiplication of two t -digit numbers can be stored exactly in a $2t$ digit mantissa. Hence, roundoff only occurs when converting to single precision, and the result is significant to approximately t bits significance in single precision.

6.3 Absolute Value Notation

It turns out that in order to perform error analysis on floating-point matrix computations, we need the absolute value notation:

If \mathbf{A} and \mathbf{B} are in $\mathfrak{R}^{m \times n}$ then

$$\begin{aligned} \mathbf{B} = |\mathbf{A}| &\Rightarrow b_{ij} = |a_{ij}|, \quad i = 1 : m, \quad j = 1 : n \\ \text{also } \mathbf{B} \leq \mathbf{A} &\Rightarrow b_{ij} \leq a_{ij}, \quad i = 1 : m, \quad j = 1 : n \end{aligned}$$

This notation is used often in the sequel of the course.

From discussion on floating point numbers, we then have

$$|\text{fl}(\mathbf{A}) - \mathbf{A}| \leq u|\mathbf{A}|. \quad (30)$$

6.4 The Sensitivity of Linear Systems

In this section, we discuss how errors in the floating point representation of numbers affects the error in the solution of the solution of a system of equations. In this respect, the idea of the matrix *condition number* is developed.

Consider the system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (31)$$

where $\mathbf{A} \in \mathfrak{R}^{n \times n}$ is nonsingular, and $\mathbf{b} \in \mathfrak{R}^n$. How do perturbations in \mathbf{A} or \mathbf{b} affect the solution \mathbf{x} ?

To gain insight, we consider several situations where perturbations can induce large errors in \mathbf{x} . For the first example, we perform the singular value decomposition on \mathbf{A} :

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \quad (32)$$

Therefore, because $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, we have

$$\mathbf{x} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{b},$$

or, using the outer product representation for matrix multiplication we have

$$\mathbf{x} = \sum_{i=1}^n \mathbf{v}_i \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i}. \quad (33)$$

Let us now consider a perturbed version $\tilde{\mathbf{A}}$, where $\tilde{\mathbf{A}} = \mathbf{A} + \epsilon \mathbf{F}$, where \mathbf{F} is an error matrix and ϵ , which can be taken to be small, controls the magnitude

of error. Now let \mathbf{F} be taken as the outer product $\mathbf{F} = \mathbf{u}_n \mathbf{v}_n^T$. Then, the singular value decomposition of $\tilde{\mathbf{A}}$ is identical to that for \mathbf{A} , except the new σ_n , denoted $\tilde{\sigma}_n$, is replaced with $\sigma_n + \epsilon$.

We see that $\tilde{\sigma}_n$ contains large relative error for σ_n suitably small. Further, since σ_n is small, the term for $i = n$ in (33) contributes strongly to \mathbf{x} . Thus, a small change in \mathbf{A} of the specific structure in the form of $\epsilon \mathbf{F}$ can result in large changes in the solution \mathbf{x} .

For a second example, we consider the “Interesting Theorem” of Sect. 3.7. Here, we see that the smallest singular value σ_n is the 2-norm distance of \mathbf{A} from the set of singular matrices. Consider the matrix \mathbf{A}_{n-1} defined in the theorem. Then \mathbf{A}_{n-1} is the closest singular matrix in the 2-norm sense to \mathbf{A} , and this 2-norm distance is σ_n . Thus, if σ_n is small, then \mathbf{A} is close to singularity and the computed \mathbf{x} becomes more sensitive to changes in either \mathbf{A} or \mathbf{b} . Note that if \mathbf{A} is singular, then the solution is “infinitely sensitive” to any perturbations, and hence produce meaningless results.³

These examples indicate that a small σ_n can cause large errors in \mathbf{x} . But we don’t have a precise idea of what “small” means in this context. “Small” relative to what? The following section addresses this question.

6.4.1 Derivation of condition number

We now develop the idea of the *condition number*, which gives us a precise definition of the sensitivity of \mathbf{x} to changes in \mathbf{A} or \mathbf{b} in eq. (31). Now consider the perturbed system

$$(\mathbf{A} + \epsilon \mathbf{F})\mathbf{x}(\epsilon) = \mathbf{b} + \epsilon \mathbf{f} \tag{34}$$

where

ϵ is a small scalar

$\mathbf{F} \in \mathfrak{R}^{n \times n}$ and $\mathbf{f} \in \mathfrak{R}^n$ are errors

$\mathbf{x}(\epsilon)$ is the perturbed solution, such that $\mathbf{x}(0) = \mathbf{x}$.

³Later in the course, we discuss methods of producing quite meaningful solutions to singular systems of equations.

We wish to place a lower bound on the relative error in \mathbf{x} due to the perturbations. Since \mathbf{A} is nonsingular, we can differentiate (34) implicitly wrt ϵ :

$$(\mathbf{A} + \epsilon\mathbf{F})\dot{\mathbf{x}}(\epsilon) + \mathbf{F}\mathbf{x}(\epsilon) = \mathbf{f} \quad (35)$$

For $\epsilon = 0$ we get

$$\dot{\mathbf{x}}(0) = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{F}\mathbf{x}). \quad (36)$$

The Taylor series expansion for $\mathbf{x}(\epsilon)$ about $\epsilon = 0$ has the form:

$$\mathbf{x}(\epsilon) = \mathbf{x} + \epsilon\dot{\mathbf{x}}(0) + O(\epsilon^2). \quad (37)$$

Substituting (36) into (37), we get

$$\mathbf{x}(\epsilon) - \mathbf{x} = \epsilon\mathbf{A}^{-1}(\mathbf{f} - \mathbf{F}\mathbf{x}) + O(\epsilon^2) \quad (38)$$

Hence by taking norms, we have

$$\begin{aligned} \|\mathbf{x}(\epsilon) - \mathbf{x}\| &= \|\epsilon\mathbf{A}^{-1}(\mathbf{f} - \mathbf{F}\mathbf{x}) + O(\epsilon^2)\| \\ &\leq \epsilon\|\mathbf{A}^{-1}(\mathbf{f} - \mathbf{F}\mathbf{x})\| + O(\epsilon^2) \end{aligned}$$

where the triangle inequality has been used; i.e., $\|\mathbf{A} + \mathbf{b}\| \leq \|\mathbf{A}\| + \|\mathbf{b}\|$. Using the property of p-norms, $\|\mathbf{A}\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{b}\|$, we have

$$\begin{aligned} \|\mathbf{x}(\epsilon) - \mathbf{x}\| &\leq \epsilon\|\mathbf{A}^{-1}\|\|\mathbf{f} - \mathbf{F}\mathbf{x}\| + O(\epsilon^2) \\ &\leq \epsilon\|\mathbf{A}^{-1}\|\{\|\mathbf{f}\| + \|\mathbf{F}\mathbf{x}\|\} + O(\epsilon^2) \\ &\leq \epsilon\|\mathbf{A}^{-1}\|\{\|\mathbf{f}\| + \|\mathbf{F}\|\|\mathbf{x}\|\} + O(\epsilon^2). \end{aligned}$$

Therefore the relative error in $\mathbf{x}(\epsilon)$ can be expressed as

$$\begin{aligned} \frac{\|\mathbf{x}(\epsilon) - \mathbf{x}\|}{\|\mathbf{x}\|} &\leq \epsilon\|\mathbf{A}^{-1}\|\left\{\frac{\|\mathbf{f}\|}{\|\mathbf{x}\|} + \|\mathbf{F}\|\right\} + O(\epsilon^2) \\ &= \epsilon\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\left\{\frac{\|\mathbf{f}\|}{\|\mathbf{A}\|\|\mathbf{x}\|} + \frac{\|\mathbf{F}\|}{\|\mathbf{A}\|}\right\} + O(\epsilon^2). \end{aligned}$$

But since $\mathbf{A}\mathbf{x} = \mathbf{b}$, then $\|\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$ and we have

$$\frac{\|\mathbf{x}(\epsilon) - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \epsilon\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\left\{\frac{\|\mathbf{f}\|}{\|\mathbf{b}\|} + \frac{\|\mathbf{F}\|}{\|\mathbf{A}\|}\right\} + O(\epsilon^2). \quad (39)$$

There are many interesting things about (39):

1. The left-hand side = $\frac{\|\mathbf{x}(\epsilon) - \mathbf{x}\|}{\|\mathbf{x}\|}$ is the *relative* error in \mathbf{x} due to the perturbation.

2. $\epsilon \frac{\|\mathbf{f}\|}{\|\mathbf{b}\|}$ is the relative error in $\mathbf{b} \triangleq \rho_b$
3. $\epsilon \frac{\|\mathbf{F}\|}{\|\mathbf{A}\|}$ is the relative error in $\mathbf{A} \triangleq \rho_A$
4. $\|\mathbf{A}^{-1}\| \|\mathbf{A}\|$ is defined as the *condition number* $\kappa(\mathbf{A})$ of \mathbf{A} .

From (39) we write

$$\frac{\|\mathbf{x}(\epsilon) - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A})(\rho_A + \rho_B) + O(\epsilon^2) \quad (40)$$

Thus we have the important result: Eq.(40) says that, to a first-order approximation, the relative error in the computed solution \mathbf{x} is bounded by the expression $\kappa(\mathbf{A}) \times$ (relative error in \mathbf{A} + relative error in \mathbf{b}). This is a rather intuitively satisfying result. Thus the condition number $\kappa(\mathbf{A})$ is the maximum amount the relative error in $\mathbf{A} + \mathbf{b}$ is magnified to give the relative error in the solution \mathbf{x} .

The condition number $\kappa(\mathbf{A})$ is norm-dependent. The most common norm is the 2-norm. In this case, $\|\mathbf{A}\|_2 = \sigma_1$. Further, since the singular values of \mathbf{A}^{-1} are the reciprocals of those of \mathbf{A} , it is easy to verify that $\|\mathbf{A}^{-1}\|_2 = \sigma_n^{-1}$. Therefore, from the definition of condition number, we have

$$\kappa_2(\mathbf{A}) = \frac{\sigma_1}{\sigma_n} \quad (41)$$

6.4.2 Alternative derivation of condition number:

We now develop the condition number again, but in a different way. It is hoped that with these two different derivations, you will get a better intuitive understanding of the concept. Consider the perturbed system where there are errors in both \mathbf{A} and \mathbf{b} . Here, the notation is simpler if we denote the errors as $\Delta\mathbf{A}$ and $\Delta\mathbf{b}_1$ respectively. The perturbed system becomes

$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}_1 \quad (42)$$

Assuming that $\Delta\mathbf{A}\Delta\mathbf{x}$ is small in comparison with $\Delta\mathbf{A}\mathbf{x}$ we can write the above as

$$\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}_1 - \Delta\mathbf{b}_2 \quad (43)$$

where $\Delta \mathbf{b}_2 = \Delta \mathbf{A}(\mathbf{x} + \Delta \mathbf{x}) \approx \Delta \mathbf{A} \mathbf{x}$. The error $\Delta \mathbf{b}_2$ is the error in \mathbf{A} transformed to appear as an error in \mathbf{b} . Lumping these errors together, we have

$$\mathbf{A}(\mathbf{x} + \Delta \mathbf{x}) = \mathbf{b} + \Delta \mathbf{b}, \quad (44)$$

where

$$\Delta \mathbf{b} \triangleq \Delta \mathbf{b}_1 - \Delta \mathbf{b}_2. \quad (45)$$

From the relation $\mathbf{A} \mathbf{x} = \mathbf{b}$ we have

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b}. \quad (46)$$

Subtracting $\mathbf{A} \mathbf{x} = \mathbf{b}$ from (44) we have

$$\mathbf{A} \Delta \mathbf{x} = \Delta \mathbf{b} \quad (47)$$

from which

$$\Delta \mathbf{x} = \mathbf{A}^{-1} \Delta \mathbf{b}. \quad (48)$$

We now consider what is the worst possible relative error $\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|}$ in \mathbf{x} in the 2-norm sense. This occurs when $\Delta \mathbf{b}$ from (48) is such that the corresponding $\|\Delta \mathbf{x}\|_2$ is maximum, and simultaneously, when \mathbf{b} from (46) is such that the corresponding $\|\mathbf{x}\|_2$ is minimum.

To find the respective \mathbf{b} and $\Delta \mathbf{b}$, we resort to the ellipsoidal interpretation of the svd of \mathbf{A}^{-1} shown in Fig. 1 Sect. 3.5. If $\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$, then

$$\mathbf{A}^{-1} = \mathbf{V} \mathbf{\Sigma}^{-1} \mathbf{U}^T. \quad (49)$$

Using (49) in (48) we have

$$\Delta \mathbf{x} = \mathbf{A}^{-1} \Delta \mathbf{b} = \mathbf{V} \mathbf{\Sigma}^{-1} \mathbf{U}^T \Delta \mathbf{b}; \quad (50)$$

a similar relation holds between \mathbf{x} and \mathbf{b} .

From (50), we can see that $\|\Delta \mathbf{x}\|_2$ is maximum with respect to $\Delta \mathbf{b}$ (for $\|\Delta \mathbf{b}\|_2$ fixed at a constant value, say k_1), when $\Delta \mathbf{b}$ lines up with the component of \mathbf{U} corresponding to the largest singular value of $\mathbf{\Sigma}^{-1}$, which is $\frac{1}{\sigma_n}$. Thus, over all possible values of $\Delta \mathbf{b}$ of constant magnitude, the largest $\Delta \mathbf{x}$ occurs when $\Delta \mathbf{b} \in \text{span}[\mathbf{u}_n]$, in which case the growth factor has the largest possible value σ_n^{-1} . Thus

$$\max_{\|\Delta \mathbf{b}\|_2 = k_1} \|\Delta \mathbf{x}\|_2 = \frac{1}{\sigma_n} \|\Delta \mathbf{b}\|_2 \quad (51)$$

Likewise, we wish to find \mathbf{b} such that $\|\mathbf{x}\|_2$ is minimum. This occurs for $\|\mathbf{b}\|_2$ fixed at a constant value, say k_2 , when \mathbf{b} lines up with the component of \mathbf{U} corresponding to the smallest singular value of $\mathbf{\Sigma}^{-1}$, which is $\frac{1}{\sigma_1}$. Therefore, over all possible values of \mathbf{b} of constant magnitude, the smallest \mathbf{x} occurs when $\mathbf{b} \in \text{span}[\mathbf{u}_1]$, in which case the growth factor is the minimum possible value σ_1^{-1} . Thus

$$\min_{\|\mathbf{b}\|_2=k_2} \|\mathbf{x}\|_2 = \frac{1}{\sigma_1} \|\mathbf{b}\|_2. \quad (52)$$

Note that we must fix the magnitudes of $\Delta\mathbf{b}$ and \mathbf{b} , because we are interested in the worst relative error in \mathbf{x} for a relative error in \mathbf{b} of fixed norm. Substituting (51) and (52) into the expression $\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ for maximum relative error in \mathbf{x} , we have

$$\begin{aligned} \max \frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} &= \frac{\sigma_1}{\sigma_n} \frac{\|\Delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2} \\ &= \kappa_2(\mathbf{A}) \frac{\|\Delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2} \\ &\leq \kappa_2(\mathbf{A}) \left[\frac{\|\Delta\mathbf{b}_1\|_2 + \|\Delta\mathbf{b}_2\|_2}{\|\mathbf{b}\|_2} \right] \end{aligned} \quad (53)$$

where in the last line we have used (45) and the triangle inequality. Eq. (53) which represents the maximum relative error in \mathbf{x} is the product of $\kappa(\mathbf{A})$ and the sum of relative errors. The first relative error term is due to errors in \mathbf{b} itself; the second is an error in \mathbf{b} transformed from an error in \mathbf{A} . Thus, (53) is roughly equivalent to (40).

The analysis for this section gives an interpretation of the meaning of the condition number $\kappa_2(\mathbf{A})$. It also indicates in what directions \mathbf{b} and $\Delta\mathbf{b}$ must point to result in the maximum relative error in \mathbf{x} . We see for worst error performance, $\Delta\mathbf{b}$ points along the direction of \mathbf{u}_n , and \mathbf{b} points along \mathbf{u}_1 . If the “svd ellipsoid” is elongated, then there is a large disparity in the relative growth factors in $\Delta\mathbf{x}$ and \mathbf{x} , and large relative error in \mathbf{x} can result.

Questions:

What is the condition number of an orthonormal matrix?

What is the condition number of a singular matrix?

What happens if $\Delta\mathbf{b} \in \text{span}(\mathbf{u}_1)$ and $\mathbf{b} \in \text{span}(\mathbf{u}_n)$?

6.5 More About the Condition Number

The condition number has these properties:

1. $\kappa(\mathbf{A}) \geq 1$.
2. If $\kappa(\mathbf{A}) \sim 1$, we say the system is *well-conditioned*, and the error in the solution is of the same magnitude as that of \mathbf{A} and \mathbf{b} .
3. If $\kappa(\mathbf{A})$ is large, then the system is poorly conditioned, and small errors in \mathbf{A} or \mathbf{b} could result in large errors in \mathbf{x} . In the practical case, the errors can be treated as random variables and hence are likely to have components along *all* the vectors \mathbf{u}_i , including \mathbf{u}_n . Thus in a practical situation with poor conditioning, error growth in the solution is almost certain to occur.

We still must consider how bad the condition number can be before it starts to seriously affect the accuracy of the solution for a given floating-point precision. In ordinary numerical systems, the errors in \mathbf{A} or \mathbf{b} result from the floating point representation of the numbers. The maximum relative error in the floating point number is u . The condition number $\kappa(\mathbf{A})$ is the worst-case factor by which this floating-point error is magnified in the solution. Thus, the relative error in the solution \mathbf{x} is bounded from above by the quantity $O(u\kappa(\mathbf{A}))$.⁴ Therefore, if $\kappa(\mathbf{A}) \sim \frac{1}{u}$, then the relative error in the solution can approach unity, which means the result is meaningless. If $\kappa(\mathbf{A}) \sim \frac{10^{-r}}{u}$, then the relative error in the solution can be taken as 10^{-r} , and the solution is approximately correct to r decimal places.

Property 3, Section 1 gives some interesting insight into how the effects of large condition number can be partially mitigated. Consider the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where \mathbf{A} is poorly conditioned. Now consider the modified system $(\mathbf{A} + s\mathbf{I})\mathbf{x} = \mathbf{b}$, where s is a small scalar, chosen so that it is small relative to the main diagonal of \mathbf{A} , yet significant with respect to the smallest singular value σ_n of \mathbf{A} . (The fact that \mathbf{A} is poorly conditioned allows such a number to exist). Now, the condition number of the modified system using Property 3 Section 1 is $\frac{\sigma_1 + s}{\sigma_n + s}$, which can be significantly smaller than the condition

⁴This bound only applies to the best or most stable algorithms for solving systems of equations. Poor algorithms will do a lot worse than this bound.

number of \mathbf{A} . Because s is small, the error in the solution due to the modification of the system can usually be tolerated. However, in cases where $\kappa(\mathbf{A})$ approaches $1/u$, the improvement in relative error in the solution due to error magnification can be enormous, and so great gains in numerical stability can be made using this technique.

Along these lines, it is interesting to note that the presence of noise, especially white noise, can significantly improve the conditioning of the system. For example, in least squares problems, we solve systems of the form $\mathbf{R}\mathbf{x} = \mathbf{b}$, where \mathbf{R} is a covariance matrix of some process which produces vector samples \mathbf{x} . Suppose \mathbf{x} is noise-free. The covariance matrix is given as $\mathbf{R} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$ and we assume for these purposes that \mathbf{R} is poorly conditioned. Now suppose we can observe only a noise-contaminated version $\tilde{\mathbf{x}}$ of \mathbf{x} given by $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{w}$, where \mathbf{w} is a vector of white noise samples with power σ^2 , which we take as relatively small compared to the power in \mathbf{x} . Because we can assume the noise is uncorrelated with the signal component \mathbf{x} , the covariance matrix $\tilde{\mathbf{R}}$ of $\tilde{\mathbf{x}}$ is given as $\mathbf{R} + \sigma^2\mathbf{I}$. Thus, the addition of noise adds the quantity σ^2 to the main diagonal of \mathbf{R} , thus improving the conditioning as discussed above. Therefore, in this special sense, we see that noise can actually help us in improving the accuracy of our solution, rather than working against us as it usually does.

We now consider an interesting theorem which relates the condition number of a covariance matrix of a process to its power spectral density.

Theorem 3 *The condition number of a covariance matrix representing a random process is bounded from above by the ratio of the maximum to minimum value of the corresponding power spectrum of the process.*

Proof:⁵ Let $\mathbf{R} \in \mathfrak{R}^{n \times n}$ be the covariance matrix of a stationary or wide-sense stationary random process \mathbf{x} , with corresponding eigenvectors \mathbf{v}_i and eigenvalues λ_i . In this treatment, the eigenvectors do not necessarily have unit 2-norm. Consider the *Rayleigh quotient* discussed in Sect. 5.4

$$\lambda_i = \frac{\mathbf{v}_i^T \mathbf{R} \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{v}_i}. \quad (54)$$

⁵This proof is taken from Haykin, "Adaptive Filter Theory", 2nd. ed., ch.2.

The quadratic form in the numerator may be expressed in an expanded form as

$$\mathbf{v}_i^T \mathbf{R} \mathbf{v}_i = \sum_{k=1}^n \sum_{m=1}^n v_{ik} r(k-m) v_{im} \quad (55)$$

where v_{ik} denotes the k th element of the i th eigenvector \mathbf{v}_i matrix \mathbf{V} , and $r(k-m)$ is the (k, m) th element of \mathbf{R} . Using the Wiener–Khinchine relation⁶ we may write

$$r(k-m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) e^{j\omega(k-m)} d\omega. \quad (56)$$

where $S(\omega)$ is the power spectral density of the process. Substituting (56) into (55) we have

$$\begin{aligned} \mathbf{v}_i^T \mathbf{R} \mathbf{v}_i &= \frac{1}{2\pi} \sum_{k=1}^n \sum_{m=1}^n v_{ik} v_{im} \int_{-\pi}^{\pi} S(\omega) e^{j\omega(k-m)} d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) d\omega \sum_{k=1}^n v_{ik} e^{j\omega k} \sum_{m=1}^n v_{im} e^{-j\omega m}. \end{aligned} \quad (57)$$

At this point, we interpret the eigenvector \mathbf{v}_i as a waveform in time. Let its corresponding Fourier transform $V_i(e^{j\omega})$ be given as

$$V_i(e^{j\omega}) = \sum_{k=1}^n v_{ik} e^{-j\omega k}. \quad (58)$$

We may therefore express (57) as

$$\mathbf{v}_i^T \mathbf{R} \mathbf{v}_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 S(\omega) d\omega. \quad (59)$$

It may also be shown that

$$\mathbf{v}_i^T \mathbf{v}_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 d\omega. \quad (60)$$

Substituting (59) and (60) into (54) we have

$$\lambda_i = \frac{\int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 S(\omega) d\omega}{\int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 d\omega}. \quad (61)$$

⁶This relation states that the autocorrelation sequence $r(\cdot)$ and the power spectral density $S(\omega)$ are a Fourier transform pair. See Haykin, “Adaptive Filter Theory”, ch. 2.

As an aside, (61) has an interesting interpretation in itself. The numerator may be regarded as the integral of the output power spectral density of a filter with coefficients \mathbf{v}_i , driven by the input process \mathbf{x} . The i th eigenvalue is this quantity normalized by the squared norm of \mathbf{v}_i .

Let S_{\min} and S_{\max} be the absolute minimum and maximum values of $S(\omega)$ respectively. Then it follows that

$$\int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 S(\omega) d\omega \geq S_{\min} \int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 d\omega \quad (62)$$

and

$$\int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 S(\omega) d\omega \leq S_{\max} \int_{-\pi}^{\pi} |V_i(e^{j\omega})|^2 d\omega \quad (63)$$

Hence, from (61) we can say that the eigenvalues λ_i are bounded by the maximum and minimum values of the spectrum $S(\omega)$ as follows:

$$S_{\min} \leq \lambda_i \leq S_{\max}, \quad i = 1, \dots, n. \quad (64)$$

Further, the condition number $\kappa(\mathbf{R})$ is bounded as

$$\kappa(\mathbf{R}) \leq \frac{S_{\max}}{S_{\min}}. \quad (65)$$

□

A consequence of (65) is that if a covariance matrix \mathbf{R} is rank deficient, then there exist values of $\omega \in [-\pi, \pi]$ such that the power spectrum is zero.

7 Massively Parallel Systolic Array Architectures for Matrix Computations

Consider the matrix-matrix multiplication operation given by

$$\begin{array}{ccccc} \mathbf{C} & = & \mathbf{A} & \mathbf{B} & \\ m \times n & & m \times k & k \times n & \end{array}$$

It is easily seen \mathbf{C} requires mkn floating-point operations (flops) to evaluate. If τ is the time for one flop on a conventional machine, then $mkn\tau$ is the time to evaluate \mathbf{C} .

Typical processors execute a matrix multiply sequentially. That is, for each element c_{ij} in turn, the corresponding dot product operation $\mathbf{A}_i^T \mathbf{b}_j$ is executed one term at a time. This results in the execution time $mkn\tau$ seconds.

It is easy to see however, that for the matrix multiply operation, there exist many opportunities for exploiting *concurrency*. Concurrency involves the elements of *parallelism* and *pipelining*. Parallelism involves many processors working separately on different non-overlapping parts of the same problem. An example of where parallelism may be used in matrix multiplication is to have one processor independently evaluating each element of the product as shown in Fig. 2a. Thus in principle k processors working together can evaluate the complete matrix product k times faster than a single processor.

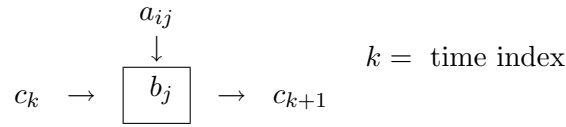
Pipelining involves many serial processors each performing a small part of the complete problem. A good example of pipelining is an assembly line in an automotive plant. Within each time period, each processor computes a small “chunk” of a given operation, and the result is passed on to the subsequent processor, as shown in Fig. 2b. Over many operations, the pipelining configuration provides a speedup factor of $nk/(n+k)$ over a single processor, where n is the number of operations and k is the number of processors.

Systolic arrays are computational structures which exploit both parallelism and pipelining to provide maximum throughput capability. We will soon see that this type of structure “beats” in synchronism with a common clock, and thus is somewhat analogous to the pumping action of a heart. This is the origin of the name of the structure. We will look at two examples to illustrate the technique.

7.1 Matrix-Vector Multiplication using Systolic Arrays

The basis of the systolic array is the *processing element* (pe). The pe is a simple computational device capable of performing the basic multiply-accumulate operation $c_{k+1} = c_k + a_{ij}b_j$. It may be represented by the

following diagram:



$$c_{k+1} = c_k + a_j b_j \quad (\text{one mult/accumulate operation})$$

At the beginning of each clock cycle, the pe reads in the values a_{ij} and c_k , performs the necessary arithmetic using the value b which is stored internally, and outputs the result c_{k+1} . Lets see how these simple processing elements can be combined together to perform matrix operations.

First, consider matrix-vector multiplication:

$$\begin{array}{c} \mathbf{c} \\ m \times 1 \end{array} = \begin{array}{c} \mathbf{A} \\ m \times n \end{array} \begin{array}{c} \mathbf{b} \\ n \times 1 \end{array}$$

A systolic array for evaluating the first element of the matrix product is shown in Fig. 3. The structure consists of a linearly-connected array of pe's as shown in the figure. Each pe is activated by a common clock. On each clock pulse, each of the "a" values fall down one position:

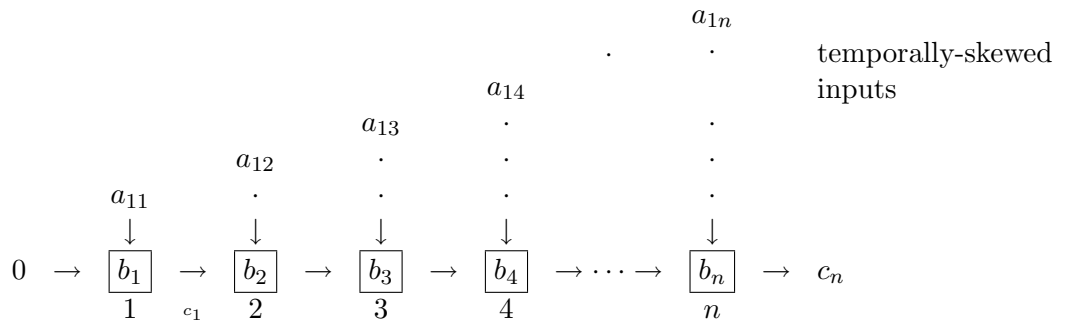


Fig. 3.

The dots in Fig. 3 represent temporary storage elements. At the first clock pulse, the a 's fall down one position, and pe 1 computes $c_1 = a_{1j}b_1$ and passes result to output of pe 1. On second clock pulse, pe 2 computes

$c_2 = c_1 + a_{2j}b_2$ and passes result to its output. (after the a 's fall one further position). After n clock pulses, the result $c_n = \sum_{i=1}^n a_{ij}b_i$ is available at the output of pe n . This value corresponds to the j th element of the product.

However, with above scheme, only one pe is busy at a time, and only one row of the matrix \mathbf{A} has been considered. It is possible to evaluate all elements of the product vector \mathbf{c} concurrently, with all processors busy almost all the time. All that is necessary is to place additional rows of \mathbf{A} immediately above the first row shown in Fig. 3. Then the computation of the second inner product involving the second row of \mathbf{A} follows directly behind the computation of the first element of the product. Similarly with third row, etc. After m clock periods, the m^{th} row begins to accumulate, and after $m + n$ periods, all elements of the product have appeared at the output.

If \mathbf{A} is $m \times n$, then mn flops are required for matrix-vector multiplication on a uni-processor. With a systolic array, only $m + n$ time periods are required, with n simple processors. Hence, significant speedups can be obtained using the systolic array concept.

Further advantages of systolic arrays:

1. The pe's only talk to nearest neighbours.
2. Each pe is exactly the same.

The above points make the silicon VLSI layout of this computational structure relatively simple. Only one cell need be designed; the entire array is formed by repeating this design many times, which is a simple process in VLSI design. The interconnections between processors are simple because they talk only to nearest neighbours.

7.2 Matrix-Matrix Multiplication by Systolic Arrays

First, we consider the evaluation of a single outer product. Extension to full matrix-matrix multiplication follows easily from the outer product rep-

representation of matrix multiplication:

$$\mathbf{c} = \mathbf{A} \mathbf{B}^T = \sum_{i=1}^k \mathbf{A}_i \cdot \mathbf{b}_i^T$$

$n \times n$ $n \times k$ $k \times n$

Consider the following structure (*Whitehouse, 1985*) for the 3×3 case:

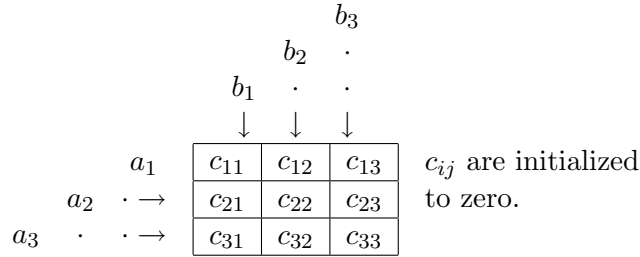
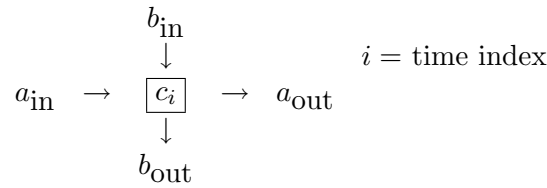


Fig. 4.

The operation of each cell is as follows:



$$c_{i+1} = c_i + a_{in} b_{in}; \quad c_0 = 0$$

$$a_{out} = a_{in}$$

$$b_{out} = b_{in}$$

It is easily seen with the above configuration (because of the temporal skew on the input data) all elements to form a specific outer product arrive in synchronism at the appropriate cell so that the correct term may be evaluated. Other outer products are formed by placing additional rows of \mathbf{B} above the first row shown in Fig. 4, and by placing corresponding additional columns of \mathbf{A} to the left of the \mathbf{A} - column which is shown. At the end of the operation, all outer product terms are accumulated in the appropriate way in each respective pe.

With this structure, we evaluate the complete matrix product in $k + \max(n, m)$ time, using nm processors. This compares to kmn time using a uni-processor.

Other systolic structures will be discussed later in the course.