CME 302: NUMERICAL LINEAR ALGEBRA
FALL 2005/06
LECTURE 10

GENE H. GOLUB

1. Estimating the Condition Number

Consider the condition number

$$\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty.$$

Of course,

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|,$$

but how do we compute $\|A^{-1}\|_\infty$? If $A^{-1} = B$, then $\|A^{-1}\|_\infty = \max_i \sum_{j=1}^{n} |b_{ij}|$. Suppose $A\mathbf{y} = \mathbf{d}$ or $\mathbf{y} = A^{-1}\mathbf{d}$. Then $\|\mathbf{y}\|_\infty \leq \|A^{-1}\|_\infty \|\mathbf{d}\|_\infty$, and therefore

$$\|A^{-1}\|_\infty \geq \frac{\|\mathbf{y}\|_\infty}{\|\mathbf{d}\|_\infty}.$$

This suggests an algorithm for estimating the condition number: we can choose $\mathbf{d}$ to maximize $\|\mathbf{y}\|_\infty$. To illustrate the process, we let

$$A = T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & t_{nn} \end{bmatrix}$$

and examine the process of solving $T\mathbf{y} = \mathbf{d}$. Writing out this system of equations yields

$$\begin{aligned} t_{11}y_1 + t_{12}y_2 + \cdots + t_{1n}y_n &= d_1 \\ &\vdots \\ t_{nn}y_n &= d_n \end{aligned}$$

Considering the last equation $y_n = d_n/t_{nn}$, we choose $d_n = +1$ if $t_{nn} > 0$, and -1 otherwise. Next, we have

$$t_{n-1,n-1}y_{n-1} + t_{n-1,n}y_n = d_{n-1},$$

which yields

$$y_{n-1} = \frac{d_{n-1} - t_{n-1,n}y_n}{t_{n-1,n-1}}.$$

If $t_{n-1,n}y_n > 0$, we choose $d_{n-1} = -1$, otherwise, we set $d_{n-1} = +1$. We continue this process, consistently choosing $d_i = \pm 1$ depending on which choice increases $\|\mathbf{y}\|_\infty$. There are other more sophisticated strategies than this.

Date: November 25, 2005, version 1.0.
Notes originally due to James Lambers. Edited by Lek-Heng Lim.

2. Scaling and Equilibration

As we have seen, the bounds for the error depend on $\kappa(A) = \|A\|\|A^{-1}\|$. Perhaps we can re-scale the equations so that the condition number is changed. We replace the system

$$A\mathbf{x} = \mathbf{b}$$

by the equivalent system

$$DA\mathbf{x} = D\mathbf{b}$$

or possibly

$$DAE\mathbf{y} = D\mathbf{b}$$

where D and E are diagonal matrices and $y = E^{-1}\mathbf{x}$.

The answer will depend upon the norm used to compute the condition number.

Suppose A is symmetric positive definite. We want to replace A by DAD; i.e. $a_{ij} \leftarrow d_i d_j a_{ij}$. Can we choose D so that $\kappa(DAD)$ is minimized?

It turns out that for a class of symmetric matrices, this is the case. A symmetric positive definite matrix A is said to have *Property A* if there exists a permutation matrix Π such that

$$\Pi A \Pi^\top = \begin{bmatrix} D & F \\ F^\top & D \end{bmatrix}$$

where D is a diagonal matrix. All tridiagonal matrices that are symmetric positive definite have Property A.

For example, suppose

$$A = \begin{bmatrix} 50 & 7 \\ 7 & 1 \end{bmatrix}.$$

Then $\lambda_{\max} \approx 51$ and $\lambda_{\min} \approx 1/51$, which means that $\kappa(A) \approx 2500$. However,

$$DAD = \begin{bmatrix} \frac{1}{\sqrt{50}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 50 & 7 \\ 7 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{50}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{7}{\sqrt{50}} \\ \frac{7}{\sqrt{50}} & 1 \end{bmatrix}$$

and

$$\kappa = \frac{1 + 7/\sqrt{50}}{1 - 7/\sqrt{50}} \approx 200.$$

One scaling strategy is called *equilibration*. The idea is to set $A^{(0)} = A$ and compute $A^{(1/2)} = D^{(1)}A^{(0)} = \{d_i^{(1)} a_{ij}\}$, choosing the diagonal matrix D_1 so that $d_i^{(1)} \sum_{j=1}^n |a_{ij}^{(0)}| = 1$. Then, we compute $A^{(1)} = A^{(1/2)}E^{(1)} = \{a_{ij}^{(1/2)} e_j^{(1)}\}$, choosing each element of the diagonal matrix $E^{(1)}$ so that $e_j^{(1)} \sum_{i=1}^n |a_{ij}^{(1/2)}| = 1$. We then repeat this process, which yields

$$A^{(k+1/2)} = D^{(k+1)} A^{(k)}$$
$$A^{(k+1)} = A^{(k+1/2)} E^{(k+1)}$$

Under very general conditions, the $A^{(k)}$ converge to a matrix whose row and column sums are all equal.

3. The Full-rank Linear Least Squares Problem

Given an $m \times n$ matrix A, with $m \geq n$, and an m-vector $\mathbf{b}$, we consider the *overdetermined* system of equations $A\mathbf{x} = \mathbf{b}$, in the case where A has full column rank. If $\mathbf{b}$ is in the range of A, then there exists a unique solution $\mathbf{x}^*$. For example, there exists a unique solution in the case of

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

but not if $\mathbf{b} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^\top$. In such cases, when $\mathbf{b}$ is not in the range of A, then we seek to minimize $\|A\mathbf{x} - \mathbf{b}\|_p$ for some p.

Different norms give different solutions. If $p = 1$ or $p = \infty$, then the function we seek to minimize, $f(x) = \|A\mathbf{x} - \mathbf{b}\|_p$ is not differentiable, so we cannot use standard minimization techniques. However, if $p = 2$, $f(x)$ is differentiable, and thus the problem is more tractable. We now consider two methods.

The first approach is to take advantage of the fact that the 2-norm is invariant under orthogonal transformations, and seek an orthogonal matrix Q such that the transformed problem

$$\min \|A\mathbf{x} - \mathbf{b}\|_2 = \min \|Q^\top (A\mathbf{x} - \mathbf{b})\|_2$$

is "easy" to solve. Let

$$A = Q \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R.$$

Then $Q_1^\top A = R$ and

$$\begin{aligned} \min \|A\mathbf{x} - \mathbf{b}\|_2 &= \min \|Q^\top (A\mathbf{x} - \mathbf{b})\|_2 \\ &= \| \min \|(Q^\top A)\mathbf{x} - Q^\top \mathbf{b}\|_2 \\ &= \min \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} \mathbf{x} - Q^\top \mathbf{b} \right\|_2 \end{aligned}$$

If we partition

$$Q^\top \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

then

$$\min \|A\mathbf{x} - \mathbf{b}\|_2^2 = \min \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \right\|_2^2 = \min \|R\mathbf{x} - \mathbf{c}\|_2^2 + \|\mathbf{d}\|_2^2.$$

Therefore, the minimum is achieved by the vector $\mathbf{x}$ such that $R\mathbf{x} = \mathbf{c}$ and therefore

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{d}\|_2 \equiv \rho_{LS}.$$

The second method is to define $\phi(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2$, which is a differentiable function of $\mathbf{x}$. We can minimize $\phi(\mathbf{x})$ by noting that $\nabla\phi(\mathbf{x}) = A^\top(A\mathbf{x} - \mathbf{b})$, which means that $\nabla\phi(\mathbf{x}) = \mathbf{0}$ if and only if $A^\top A\mathbf{x} = A^\top \mathbf{b}$. This system of equations is called the *normal equations*, and were used by Gauss to solve the least squares problem. If $m >> n$ then $A^\top A$ is $n \times n$, which is a much smaller system to solve than $A\mathbf{x} = \mathbf{b}$, and if $\kappa(A^\top A)$ is not too large, we can use the LU factorization to solve for $\mathbf{x}$.

Which is the better method? This is not a simple question to answer. The normal equations produce an $\mathbf{x}^*$ whose relative error depends on $\kappa(A)^2$, whereas the QR factorization produces an $\mathbf{x}^*$ whose relative error depends on $\mathsf{u}(\kappa_2(A) + \rho_{LS}\kappa_2(A)^2)$. The normal equations involve much less arithmetic when $m >> n$ and they require less storage, but the QR factorization is often applicable if the normal equations break down.

4. The QR Factorization

Let A be an $m \times n$ matrix with full column rank. The QR *factorization* of A is a decomposition $A = QR$, where Q is an $m \times m$ orthogonal matrix and R is an $m \times n$ upper triangular matrix. There are two ways to compute this decomposition:

(1) Using *Householder matrices*, developed by Alston S. Householder
(2) Using *Givens rotations*, also known as *Jacobi rotations*, used by W. Givens and originally invented by Jacobi for use with in solving the symmetric eigenvalue problem in 1846.
(3) A third, less frequently used approach, is the *Gram-Schmidt orthogonalization.*

5. Orthogonalization using Givens Rotations

We illustrate the process in the case where A is a 2×2 matrix. In Gaussian elimination, we compute $L^{-1}A = U$ where L^{-1} is unit lower triangular and U is upper triangular. Specifically,

$$\begin{bmatrix} 1 & 0 \\ m_{21} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ 0 & a_{22}^{(2)} \end{bmatrix}, \quad m_{21} = -\frac{a_{21}}{a_{11}}.$$

By contrast, the QR decomposition takes the form

$$\begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}$$

where $\gamma^2 + \sigma^2 = 1$. From the relationship $-\sigma a_{11} + \gamma a_{21} = 0$ we obtain

$$\begin{aligned} \gamma a_{21} &= \sigma a_{11} \\ \gamma^2 a_{21}^2 &= \sigma^2 a_{11}^2 = (1 - \gamma^2) a_{11}^2 \end{aligned}$$

which yields

$$\gamma = \pm \frac{a_{11}}{\sqrt{a_{21}^2 + a_{11}^2}}.$$

It is conventional to choose the + sign. Then, we obtain

$$\sigma^2 = 1 - \gamma^2 = 1 - \frac{a_{11}^2}{a_{21}^2 + a_{11}^2} = \frac{a_{21}^2}{a_{21}^2 + a_{11}^2},$$

or

$$\sigma = \pm \frac{a_{21}}{\sqrt{a_{21}^2 + a_{11}^2}}.$$

Again, we choose the + sign. As a result, we have

$$r_{11} = a_{11} \frac{a_{11}}{\sqrt{a_{21}^2 + a_{11}^2}} + a_{21} \frac{a_{21}}{\sqrt{a_{21}^2 + a_{11}^2}} = \sqrt{a_{21}^2 + a_{11}^2}.$$

The matrix

$$Q^\top = \begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix}$$

is called a *Givens rotation*. It is called a rotation because it is orthogonal, and therefore length-preserving, and also because there is an angle θ such that $\sin\theta = \sigma$ and $\cos\theta = \gamma$, and its effect is to rotate a vector through the angle θ. In particular,

$$\begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \rho \\ 0 \end{bmatrix}$$

where $\rho = \sqrt{\alpha^2 + \beta^2}$, $\alpha = \rho\cos\theta$ and $\beta = \rho\sin\theta$. It is easy to verify that the product of two rotations is itself a rotation. Now, in the case where A is an $n \times n$ matrix, suppose that we have

the vector $\begin{bmatrix}\times & \cdots & \times & \alpha & \times & \cdots & \times & \beta & \times & \cdots & \times\end{bmatrix}^\top$. Then

$$\begin{bmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & \gamma & & & & \sigma & & & \\
& & & & 1 & & & & & & \\
& & & & & \ddots & & & & & \\
& & & & & & 1 & & & & \\
& & & -\sigma & & & & \gamma & & & \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{bmatrix}
\begin{bmatrix}\times\\ \vdots\\ \times\\ \alpha\\ \times\\ \vdots\\ \times\\ \beta\\ \times\\ \vdots\\ \times\end{bmatrix}
=
\begin{bmatrix}\times\\ \vdots\\ \times\\ \rho\\ \times\\ \vdots\\ \times\\ 0\\ \times\\ \vdots\\ \times\end{bmatrix}.$$

So, in order to transform A into an upper triangular matrix R, we can find a product of rotations Q such that $Q^\top A = R$. It is easy to see that $O(n^2)$ rotations are required.

6. Orthogonalization using Householder Reflections

It is natural to ask whether we can introduce more zeros with each orthogonal rotation. To that end, we examine *Householder reflections*. Consider a matrix of the form $P = I - \tau \mathbf{u}\mathbf{u}^\top$, where $\mathbf{u} \neq \mathbf{0}$ and τ is a nonzero constant. It is clear that P is a symmetric rank-1 change of I. Can we choose τ so that P is also orthogonal? From the desired relation $P^\top P = I$ we obtain

$$\begin{aligned}
P^\top P &= (I - \tau \mathbf{u}\mathbf{u}^\top)^\top (I - \tau \mathbf{u}\mathbf{u}^\top) \\
&= I - 2\tau \mathbf{u}\mathbf{u}^\top + \tau^2 \mathbf{u}\mathbf{u}^\top \mathbf{u}\mathbf{u}^\top \\
&= I - 2\tau \mathbf{u}\mathbf{u}^\top + \tau^2 (\mathbf{u}^\top \mathbf{u}) \mathbf{u}\mathbf{u}^\top \\
&= I - (\tau^2 \mathbf{u}^\top \mathbf{u} - 2\tau) \mathbf{u}\mathbf{u}^\top \\
&= I + \tau(\tau \mathbf{u}^\top \mathbf{u} - 2) \mathbf{u}\mathbf{u}^\top.
\end{aligned}$$

It follows that if $\tau = 2/\mathbf{u}^\top \mathbf{u}$, then $P^\top P = I$ for any nonzero $\mathbf{u}$. Without loss of generality, we can stipulate that $\mathbf{u}^\top \mathbf{u} = 1$, and therefore P takes the form $P = I - 2\mathbf{v}\mathbf{v}^\top$, where $\mathbf{v}^\top \mathbf{v} = 1$.

Why is the matrix P called a reflection? This is because for any nonzero vector $\mathbf{x}$, $P\mathbf{x}$ is the reflection of $\mathbf{x}$ across the hyperplane that is normal to $\mathbf{v}$. To see this, we consider the 2×2 case and set $\mathbf{v} = \begin{bmatrix}1 & 0\end{bmatrix}^\top$ and $\mathbf{x} = \begin{bmatrix}1 & 2\end{bmatrix}^\top$. Then

$$P = I - 2\mathbf{v}\mathbf{v}^\top = I - 2\begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - 2\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}$$

Therefore

$$P\mathbf{x} = \begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}-1\\2\end{bmatrix}.$$

Now, let $\mathbf{x}$ be any vector. We wish to construct P so that $P\mathbf{x} = \alpha \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^\top = \alpha \mathbf{e}_1$ for some α. From the relations

$$\|P\mathbf{x}\|_2 = \|\mathbf{x}\|_2, \qquad \|\alpha \mathbf{e}_1\|_2 = |\alpha| \|\mathbf{e}_1\|_2 = |\alpha|,$$

we obtain $\alpha = \pm \|\mathbf{x}\|_2$. To determine $\mathbf{P}$, we observe that

$$\mathbf{x} = P^{-1}(\alpha \mathbf{e}_1) = \alpha P \mathbf{e}_1 = \alpha(I - 2\mathbf{v}\mathbf{v}^\top)\mathbf{e}_1 = \alpha[\mathbf{e}_1 - 2\mathbf{v}\mathbf{v}^\top \mathbf{e}_1] = \alpha[\mathbf{e}_1 - 2\mathbf{v}v_1]$$

which yields the system of equations

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \alpha \begin{bmatrix} 1 - 2v_1^2 \\ -2v_1 v_2 \\ \vdots \\ -2v_1 v_n \end{bmatrix}.$$

From the first equation $x_1 = \alpha(1 - 2v_1^2)$ we obtain

$$v_1 = \pm\sqrt{\frac{1}{2}\left(1 - \frac{x_1}{\alpha}\right)}.$$

For $i = 2, \ldots, n$, we have

$$v_i = -\frac{x_i}{2\alpha v_1}.$$

It is best to choose α to have the opposite sign of x_1 to avoid cancellation in v_1. It is conventional to choose the + sign for α.

Note that the matrix P is never formed explicitly. For any vector $\mathbf{b}$, the product $P\mathbf{b}$ can be computed as follows:

$$P\mathbf{b} = (I - 2\mathbf{v}\mathbf{v}^\top)\mathbf{b} = \mathbf{b} - 2(\mathbf{v}^\top\mathbf{b})\mathbf{v}.$$

This process requires only $O(2n)$ operations. It is easy to see that we can represent P simply by storing only $\mathbf{v}$.

We showed how a Householder reflection of the form $P = I - 2\mathbf{u}\mathbf{u}^\top$ could be constructed so that given a vector $\mathbf{x}$, $P\mathbf{x} = \alpha\mathbf{e}_1$. Now, suppose that that $\mathbf{x} = \mathbf{a}_1$ is the first column of a matrix A. Then we construct a Householder reflection $H_1 = I - 2\mathbf{u}_1\mathbf{u}_1^\top$ such that $H\mathbf{x} = \alpha\mathbf{e}_1$, and we have

$$A^{(2)} = H_1 A = \begin{bmatrix} r_{11} & & & \\ 0 & & & \\ \vdots & \mathbf{a}_2^{(2)} & \cdots & \mathbf{a}_n^{(2)} \\ 0 & & & \end{bmatrix}.$$

where we denote the constant α by r_{11}, as it is the $(1,1)$ element of the updated matrix $A^{(2)}$. Now, we can construct H_2 such that

$$H_2\mathbf{a}^{(2)} = \begin{bmatrix} a_{12}^{(2)} \\ r_{22} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad u_{12} = 0, \quad H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \\ \vdots & h_{ij} \\ 0 & \end{bmatrix}.$$

Note that the first column of $A^{(2)}$ is unchanged by H_2. Continuing this process, we obtain

$$H_{n-1} \cdots H_1 A = A^{(n)} = R$$

where R is an upper triangular matrix. We have thus factored $A = QR$, where $Q = H_1 H_2 \cdots H_{n-1}$ is an orthogonal matrix. Note that

$$A^\top A = R^\top Q^\top Q R = R^\top R,$$

and thus R is the Cholesky factor of $A^\top A$.

7. Givens Rotations versus Household Reflections

We showed how to construct Givens rotations in order to rotate two elements of a column vector so that one element would be zero, and that approximately $n^2/2$ such rotations could be used to transform A into an upper triangular matrix R. Because each rotation only modifies two rows of A, it is possible to interchange the order of rotations that affect different rows, and thus apply sets of rotations in parallel. This is the main reason why Givens rotations can be preferable to Householder reflections. Other reasons are that they are easy to use when the QR factorization needs to be updated as a result of adding a row to A or deleting a column of A. They are also more efficient when A is sparse.

Department of Computer Science, Gates Building 2B, Room 280, Stanford, CA 94305-9025
E-mail address: golub@stanford.edu