CME 302: NUMERICAL LINEAR ALGEBRA
FALL 2005/06
LECTURE 11

GENE H. GOLUB

1. Using the Normal Equations

We can solve the linear least squares problem using the normal equations

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

as follows: first, we solve the above system to obtain an approximate solution $\hat{\mathbf{x}}$, and compute the residual vector $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$. Now, because

$$A^\top \mathbf{r} = A^\top \mathbf{b} - A^\top A\hat{\mathbf{x}} = \mathbf{0},$$

we obtain the system

$$\begin{aligned} \mathbf{r} + A\hat{\mathbf{x}} &= \mathbf{b} \\ A^\top \mathbf{r} &= \mathbf{0} \end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} I & A \\ A^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

This is a large system, but it preserves the sparsity of A. It can be used in connection with iterative refinement, but unfortunately this procedure does not work well because it is very sensitive to the residual.

2. Hilbert Matrices

A *Hilbert matrix* has the form

$$H = \begin{bmatrix} 1 & 1/2 & 1/3 & \cdots & 1/n \\ 1/2 & 1/3 & \cdots & & 1/(n+1) \\ 1/3 & \cdots & & & \vdots \\ \vdots & & & & \vdots \\ 1/n & \cdots & \cdots & \cdots & 1/(2n-1) \end{bmatrix}, \quad h_{ij} = \frac{1}{i+j-1}.$$

It is very ill-conditioned, but H^{-1} is known, and its entries are all integers.

3. Complete Orthogonal Decomposition

We seek a decomposition of the form $A = Q^\top R\Pi$ where Π is chosen so that the diagonal elements of R are maximized at each stage. Specifically, suppose

$$H_1 A = \begin{bmatrix} r_{11} & \\ 0 & \\ \vdots & * \\ 0 & \end{bmatrix}, \quad r_{11} = \|\mathbf{a}_1\|_2.$$

Date: November 28, 2005, version 1.0.

Notes originally due to James Lambers. Edited by Lek-Heng Lim.

So, we choose Π_1 so that $\|\mathbf{a}_1\|_2 \geq \|\mathbf{a}_j\|_2$ for $j \geq 2$. For Π_2, look at the lengths of the columns of the submatrix. We don't need to recompute the lengths each time, because we can update by subtracting the square of the first component from the square of the total length. Eventually, we get

$$Q^\top \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \Pi_1 \cdots \Pi_r = A$$

where R is upper triangular. Using this decomposition, we can solve the linear least squares problem $A\mathbf{x} = \mathbf{b}$ by observing that

$$\begin{aligned} \|\mathbf{b} - A\mathbf{x}\|_2^2 &= \left\| \mathbf{b} - Q^\top \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \Pi\mathbf{x} \right\|_2^2 \\ &= \left\| Q\mathbf{b} - \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \right\|_2^2 \\ &= \left\| \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} - \begin{bmatrix} R\mathbf{u} + S\mathbf{v} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 \\ &= \|\mathbf{c} - R\mathbf{u} - S\mathbf{v}\|_2^2 + \|\mathbf{d}\|_2^2. \end{aligned}$$

Thus $\min \|\mathbf{b} - A\mathbf{x}\|_2^2 = \|\mathbf{d}\|_2^2$ provided that $R\mathbf{u} + S\mathbf{v} = \mathbf{c}$. A basic solution is obtained by choosing $\mathbf{v} = \mathbf{0}$. A second solution is to choose $\mathbf{u}$ and $\mathbf{v}$ so that $\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$ is minimized. This criterion is related to the pseudoinverse of A.

Suppose

$$A = Q^\top \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \Pi$$

where R is upper triangular. Then

$$A^\top = \Pi^\top \begin{bmatrix} R^\top & 0 \\ S^\top & 0 \end{bmatrix} Q$$

where $R^\top$ is lower triangular. We apply Householder reflections so that

$$H_i \cdots H_2 H_1 \begin{bmatrix} R^\top & 0 \\ S^\top & 0 \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix}.$$

Then

$$A^\top = Z^\top \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix} Q$$

where $Z = H_i \cdots H_1 \Pi$. In other words,

$$A = Q^\top \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} Z$$

where L is a lower triangular matrix of size $r \times r$, where r is the rank of A. This is the *complete orthogonal decomposition* of A.

Recall that X is the *pseudoinverse* of A if

(1) $AXA = A$
(2) $XAX = X$
(3) $(XA)^\top = XA$
(4) $(AX)^\top = AX$

Given the above complete orthogonal decomposition of A, the pseudoinverse of A, denoted A^+, is given by

$$A^+ = Z^\top \begin{bmatrix} L^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q.$$

Let $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{b} - A\mathbf{x}\|_2 = \min\}$. If $\mathbf{x} \in \mathcal{X}$ and we desire $\|\mathbf{x}\|_2 = \min$, then $\mathbf{x} = A^+\mathbf{b}$. Note that in this case,

$$\mathbf{r} = \mathbf{b} - A\mathbf{x} = \mathbf{b} - AA^+\mathbf{b} = (I - AA^+)\mathbf{b}$$

where the matrix $(I - AA^+)$ is a projection matrix $P^\perp$. To see that $P^\perp$ is a projection, note that

$$\begin{aligned} P &= AA^+ \\ &= Q^\top \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} ZZ^\top \begin{bmatrix} L^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q \\ &= Q^\top \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q. \end{aligned}$$

Suppose that we perturb the data, so that we are solving $(A + \epsilon E)\mathbf{x}(\epsilon) = A^\top \mathbf{b}$. Then what is $\|\mathbf{x} - \mathbf{x}(\epsilon)\|_2$ or $\|\mathbf{r} - \mathbf{r}(\epsilon)\|_2$? Using the fact that $PA = AA^+A = A$, we differentiate with respect to ϵ and obtain

$$P\frac{dA}{d\epsilon} + \frac{dP}{d\epsilon}A = \frac{dA}{d\epsilon}.$$

It follows that

$$\frac{dP}{d\epsilon}A = (I - P)\frac{dA}{d\epsilon} = P^\perp \frac{dA}{d\epsilon}.$$

Multiplying through by A^+, we obtain

$$\frac{dP}{d\epsilon}P = P^\perp \frac{dA}{d\epsilon}A^+.$$

Because P is a projection,

$$\frac{d(P^2)}{d\epsilon} = P\frac{dP}{d\epsilon} + \frac{dP}{d\epsilon}P = \frac{dP}{d\epsilon},$$

so, using the relationship $A^\top P = A^\top$,

$$\frac{dP}{d\epsilon} = P^\perp \frac{dA}{d\epsilon}A^+ + (A^+)^\top \frac{dA^\top}{d\epsilon}P^\perp.$$

4. More Perturbation Theory

Suppose that we are solving the perturbed least squares problem

$$A(\epsilon)\mathbf{x}(\epsilon) = \mathbf{b}, \quad A(\epsilon) = A + \epsilon E.$$

How does the residual vector $\mathbf{r}(\epsilon) = \mathbf{b} - A\mathbf{x}(\epsilon)$ and the solution $\mathbf{x}(\epsilon)$ change as a function of ϵ?

From last time, recall that the computed solution $\hat{\mathbf{x}} = A^+\mathbf{b}$ is very sensitive to the residual. To see this, suppose that $\mathbf{b}$ is replaced by $\mathbf{b} + \alpha\mathbf{r}$, where α is a constant. Then

$$\begin{aligned} A^+(\mathbf{b} + \alpha\mathbf{r}) &= \hat{\mathbf{x}} + \alpha A^+\mathbf{r} \\ &= \hat{\mathbf{x}} + \alpha A^+(I - AA^+)\mathbf{b} \\ &= \hat{\mathbf{x}} + \alpha[A^+\mathbf{b} - A^+AA^+\mathbf{b}] \\ &= \hat{\mathbf{x}} \end{aligned}$$

so the computed solution is unchanged, even if α is large.

Recall that $P = AA^+$ is a projection, with orthogonal complement $P^\perp = I - AA^+$. Furthermore, recall that

$$\frac{dP}{d\epsilon} = P^\perp \frac{dA}{d\epsilon}A^+ + (A^+)^\top \frac{dA^T}{d\epsilon}P^\perp.$$

Now, using a Taylor expansion around $\epsilon = 0$, we obtain

$$\begin{aligned}\mathbf{r}(\epsilon) &= \mathbf{r}(0) + \epsilon \frac{dP^{\perp}}{d\epsilon}\mathbf{b} + O(\epsilon^2) \\ &= \mathbf{r}(0) - \epsilon \frac{dP}{d\epsilon}\mathbf{b} + O(\epsilon^2) \\ &= \mathbf{r}(0) - \epsilon[P^{\perp}E\hat{\mathbf{x}}(0) + (A^+)^T E^T \mathbf{r}(0)] + O(\epsilon^2)\end{aligned}$$

from the relations $\hat{\mathbf{x}} = A^+\mathbf{b}$ and $\mathbf{r} = P^{\perp}\mathbf{b}$. Taking norms, we obtain

$$\frac{\|\mathbf{r}(\epsilon) - \mathbf{r}(0)\|_2}{\|\hat{\mathbf{x}}\|_2} = |\epsilon|\|E\|_2 \left(1 + \|A^+\|_2 \frac{\|\mathbf{r}(0)\|_2}{\|\hat{\mathbf{x}}(0)\|_2}\right) + O(\epsilon^2).$$

Note that if A is scaled so that $\|A\|_2 = 1$, then the second term above involves the condition number $\kappa_2(A)$. We also have

$$\frac{\|\mathbf{x}(\epsilon) - \mathbf{x}(0)\|_2}{\|\hat{\mathbf{x}}\|_2} = |\epsilon|\|E\|_2 \left(2\kappa(A) + \kappa_2(A)^2 \frac{\|\mathbf{r}(0)\|_2}{\|\hat{\mathbf{x}}(0)\|_2}\right) + O(\epsilon^2).$$

Note that a small perturbation residual does not imply a small perturbation in the solution.

5. Gram-Schmidt Orthogonalization

Consider the QR factorization

$$A = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{bmatrix} = \begin{bmatrix}\mathbf{q}_1 & \cdots & \mathbf{q}_n\end{bmatrix} \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn}\end{bmatrix}.$$

From the above matrix product we can see that $\mathbf{a}_1 = r_{11}\mathbf{q}_1$, from which it follows that

$$r_{11} = \pm\|\mathbf{a}_1\|_2, \quad \mathbf{q}_1 = \frac{1}{\|\mathbf{a}_1\|_2}\mathbf{a}_1.$$

Next, from $\mathbf{a}_2 = r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2$ we obtain

$$r_{12} = \mathbf{q}_1^\top \mathbf{a}_2, \quad r_{22} = \pm\|\mathbf{a}_2 - r_{12}\mathbf{q}_1\|_2, \quad \mathbf{q}_2 = \frac{1}{r_{22}}(\mathbf{a}_2 - r_{12}\mathbf{q}_1).$$

In general, we use the relation

$$\mathbf{a}_k = \sum_{j=1}^{k} r_{jk}\mathbf{q}_j$$

to obtain

$$\mathbf{q}_k = \frac{1}{r_{kk}}\left(\mathbf{a}_k - \sum_{j=1}^{k-1} r_{jk}\mathbf{q}_j\right), \quad r_{jk} = \mathbf{q}_j^\top \mathbf{a}_k.$$

Note that $\mathbf{q}_k$ can be rewritten as

$$\mathbf{q}_k = \frac{1}{r_{kk}}\left(\mathbf{a}_k - \sum_{j=1}^{k-1}(\mathbf{q}_j^\top \mathbf{a}_k)\mathbf{q}_j\right) = \frac{1}{r_{kk}}\left(\mathbf{a}_k - \sum_{j=1}^{k-1}\mathbf{q}_j\mathbf{q}_j^\top \mathbf{a}_k\right) = \frac{1}{r_{kk}}\left(I - \sum_{j=1}^{k-1}\mathbf{q}_j\mathbf{q}_j^\top\right)\mathbf{a}_k.$$

If we define $P_i = \mathbf{q}_i\mathbf{q}_i^\top$, then P_i is a *symmetric projector* that satisfies $P_i^2 = P_i$, and $P_iP_j = \delta_{ij}$. Thus we can write

$$\mathbf{q}_k = \frac{1}{r_{kk}}\left(I - \sum_{j=0}^{k-1} P_j\right)\mathbf{a}_k = \frac{1}{r_{kk}}\prod_{j=1}^{k-1}(I - P_j)\mathbf{a}_k.$$

Why doesn't Gram-Schmidt work? If $\mathbf{a}_1$ and $\mathbf{a}_2$ are almost parallel, then $\mathbf{a}_2 - r_{12}\mathbf{q}_1$ is almost zero and roundoff error becomes significant.

6. Modified Gram-Schmidt

Although the classical Gram-Schmidt process is numerically unstable, the *Modified Gram-Schmidt* method alleviates this difficulty. Recall

$$A = QR = \begin{bmatrix} r_{11}\mathbf{q}_1 & r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2 & \cdots \end{bmatrix}$$

We define

$$A^{(k)} = \sum_{i=1}^{k-1} \mathbf{q}_i \mathbf{r}_i^\top, \quad \mathbf{r}_i^\top = \begin{bmatrix} r_{i1} & r_{i2} & \cdots & r_{ii} \end{bmatrix}$$

which means

$$A - \sum_{i=1}^{k-1} \mathbf{q}_i \mathbf{r}_i^\top = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & A^{(k)} \end{bmatrix}.$$

If we write

$$A^{(k)} = \begin{bmatrix} \mathbf{z} & B \end{bmatrix}$$

then

$$r_{kk} = \|\mathbf{z}\|_2, \quad \mathbf{q}_k = \frac{1}{r_{kk}}\mathbf{z}.$$

We then compute

$$\begin{bmatrix} r_{k,k+1} & \cdots & r_{k,n} \end{bmatrix} = \mathbf{q}_k^\top B$$

which yields

$$A^{(k+1)} = B - \mathbf{q}_k \begin{bmatrix} r_{1k} & \cdots & r_{kk} \end{bmatrix}$$

This process is numerically stable.

We can show

$$\hat{Q}_1^\top \hat{Q}_1 = I + E_{MGS}, \quad \|E_{MGS}\| \approx \mathsf{u}\kappa_2(A),$$

and $\hat{Q}_1$ can be computed in approximately $2mn^2$ flops, whereas with Householder QR,

$$\hat{Q}_1^\top \hat{Q}_1 = I + E_n, \quad \|E_n\| \approx \mathsf{u},$$

with $\hat{Q}_1$ being computed in approximately $2mn^{-}2n^2/3$ flops to factor A and an additional $2mn^2 - 2n^2/3$ flops to obtain the n columns of Q.

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