STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2015
LECTURE 10

1. FULL RANK LEAST SQUARES VIA NORMAL EQUATIONS

- the second approach is to define
$$\varphi(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2$$
which is a differentiable function of $\mathbf{x}$
- we can minimize $\varphi(\mathbf{x})$ by noting that $\nabla\varphi(\mathbf{x}) = A^*(A\mathbf{x} - \mathbf{b})$ which means that $\nabla\varphi(\mathbf{x}) = \mathbf{0}$ if and only if
$$A^*A\mathbf{x} = A^*\mathbf{b} \tag{1.1}$$
- this system of equations is called the *normal equations*, and were used by Gauss to solve least squares problems
- we saw at least two other ways to derive (1.1) in the homeworks
- while I said that it's generally a bad idea to solve the normal equations to get the least squares solution, this is not always the case
- for example, if $n \ll m$ then A^*A is $n \times n$, which is a much smaller system to solve than solving $\min \|A\mathbf{x} - \mathbf{b}\|_2^2$ via finding QR of A, and if $\kappa(A^*A)$ is not too large, we can indeed solve (1.1)
- for A of full column rank, the matrix A^*A is positive definite and one should apply Cholesky factorization (to be discussed later) to the matrix A^*A in order to solve (1.1)
- 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 $\kappa_2(A^\mathsf{T}A) = \kappa_2(A)^2$, whereas the QR factorization produces an $\mathbf{x}^*$ whose relative error depends on $\kappa_2(A) + \rho_{\text{LS}}(\mathbf{x}^*)\kappa_2(A)^2$ where
$$\rho_{\text{LS}}(\mathbf{x}) := \frac{\|\mathbf{b} - A\mathbf{x}\|_2}{\|A\|_2\|\mathbf{x}\|_2}$$
is called the *relative residual* at $\mathbf{x}$
- so the QR factorization method in the previous section is appealing if $\rho_{\text{LS}}(\mathbf{x}^*)$ is small, i.e., $\mathbf{b}$ is very close to $\text{im}(A)$, the span of the columns of A — which is more often than not the case (e.g. in linear regression) as the most common reason for wanting to solve $\min\|A\mathbf{x} - \mathbf{b}\|_2$ is when we expect $A\mathbf{x}^* \approx \mathbf{b}$
- the normal equations involve much less arithmetic when $n \ll m$ and the $n \times n$ matrix A^*A requires less storage

2. QR FACTORIZATION VERSUS NORMAL EQUATIONS

- assuming a dense A, the following table compares the relative merits of normal equations (NE) method, QR method, and the SVD method discussed a few lectures ago

accuracy:	NE	$<$	QR	$<$	SVD
speed:	NE	$>$	QR	$>$	SVD

Date: October 27, 2015, version 1.0. Comments, bug reports: lekheng@galton.uchicago.edu.

2.1. **Conditioning of least squares.**

Theorem 1 (Wedin)**.** *Let* $A, \widehat{A} \in \mathbb{R}^{m\times n}$ *where* $\operatorname{rank}(A) = \operatorname{rank}(\widehat{A}) = n \leq m$*. Suppose* $\mathbf{x}$ *and* $\widehat{\mathbf{x}} \in \mathbb{R}^n$ *are solutions to the respective least squares problems*

$$\min\|A\mathbf{x} - \mathbf{b}\|_2 \quad \textit{and} \quad \min\|\widehat{A}\widehat{\mathbf{x}} - \widehat{\mathbf{b}}\|_2,$$

and let $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ *and* $\widehat{\mathbf{r}} = \widehat{\mathbf{b}} - \widehat{A}\widehat{\mathbf{x}}$ *be the respective residuals. If* $\epsilon > 0$ *is such that*

$$\frac{\|A - \widehat{A}\|_2}{\|A\|_2} \leq \epsilon, \quad \frac{\|\mathbf{b} - \widehat{\mathbf{b}}\|_2}{\|\mathbf{b}\|_2} \leq \epsilon, \quad \kappa_2(A)\epsilon < 1,$$

then

$$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\kappa_2(A)\epsilon}{1 - \kappa_2(A)\epsilon}\left(2 + (\kappa_2(A) + 1)\frac{\|\mathbf{r}\|_2}{\|A\|_2\|\mathbf{x}\|_2}\right), \tag{2.1}$$

and

$$\frac{\|\mathbf{r} - \widehat{\mathbf{r}}\|_2}{\|\mathbf{r}\|_2} \leq 1 + 2\kappa_2(A)\epsilon.$$

- recall that for singular or rectangular matrices, $\kappa_2(A) = \|A\|_2\|A^\dagger\|_2$
- note that if $\mathbf{r} = \mathbf{0}$, i.e., the least squares problem becomes a linear system, (2.1) reduces to the bound we obtained in Homework **1**, Problem **4**(e)
- in other words, for a linear system, the term involving $\kappa_2(A)^2$ vanishes
- a simplification of (2.1) is to assume that $\widehat{\mathbf{b}} = \mathbf{b}$ and get

$$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\kappa_2(A)\epsilon}{1 - \kappa_2(A)\epsilon}\left(1 + \kappa_2(A)\frac{\|\mathbf{r}\|_2}{\|A\|_2\|\mathbf{x}\|_2}\right)$$

- if we expand the right hand side, we get

$$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \leq \kappa_2(A)\left(1 + \kappa_2(A)\frac{\|\mathbf{r}\|_2}{\|A\|_2\|\mathbf{x}\|_2}\right)\epsilon + O(\epsilon^2) \tag{2.2}$$

- the coefficient of ϵ above is sometimes called the *least squares condition number*

2.2. **Accuracy.**

- the QR method, if properly implemented (say, using Householder or Givens), is backward stable in the following sense: when we use the method to solve

  $$\min\|A\mathbf{x} - \mathbf{b}\|_2,$$

  we get the *exact* solution to a perturbed problem

  $$\min\|\widehat{A}\widehat{\mathbf{x}} - \widehat{\mathbf{b}}\|_2,$$

  that is near to our original problem in the sense that

  $$\frac{\|A - \widehat{A}\|_2}{\|A\|_2} \leq \epsilon, \quad \frac{\|\mathbf{b} - \widehat{\mathbf{b}}\|_2}{\|\mathbf{b}\|_2} \leq \epsilon$$

  for some small ϵ
- in practice, the value of ϵ depends on m, n and the unit roundoff u of the computer/program[1] you use and is typically very small, roughly $mn\mathsf{u}/(1 - mn\mathsf{u})$
- this, combined with Theorem 1 allows us to get a bound on the relative error (as long as $\kappa_2(A) < 1/\epsilon$)
- if we use (2.2), we see that the relative error is bounded by $(\kappa_2(A) + \rho_{\text{LS}}(\mathbf{x}^*)\kappa_2(A)^2)\epsilon$
- so if $\rho_{\text{LS}}(\mathbf{x}^*)$ is small, then QR is good for accuracy

[1] usually u is around 10^{-16} for double precision, 10^{-19} for extended precision, 10^{-35} for quadruple precision

- the normal equations method, given that it relies on solving $A^\mathsf{T}A\mathbf{x} = A^\mathsf{T}\mathbf{b}$, cannot avoid the condition number $\kappa_2(A^\mathsf{T}A) = \kappa_2(A)^2$ no matter which version of Homework **1**, Problem **4** we use
- the relative error in this case is therefore always bounded by $\kappa_2(A)^2\epsilon$, never just $\kappa_2(A)\epsilon$
- as long as A is well-conditioned, it is alright to use the normal equations method, especially if you want to save on computational cost, the QR method is generally preferable
- for very ill-conditioned problem, one would have to use the SVD method discussed a few lectures ago but this is the most expensive

2.3. **Computational costs.**

- assuming that our matrix $A \in \mathbb{R}^{m\times n}$ is dense (most or all entries nonzero), then the exact flop counts for the two methods described earlier for computing the least squares solution $\mathbf{x}$ are:
  - QR factorization ($A = Q\left[\begin{smallmatrix} R \\ 0 \end{smallmatrix}\right]$) + orthogonal transformation ($\mathbf{c} = Q^\mathsf{T}\mathbf{b}$) + backsolve ($R\mathbf{x} = \mathbf{c}$):
  $$2n^2\left(m - \frac{n}{3}\right) \tag{2.3}$$
  - normal equations ($C = A^\mathsf{T}A$, $\mathbf{c} = A^\mathsf{T}\mathbf{b}$) + Cholesky factorization ($C = R^\mathsf{T}R$) + two backsolves ($R^\mathsf{T}\mathbf{y} = \mathbf{c}$, $R\mathbf{x} = \mathbf{y}$):
  $$n^2\left(m + \frac{n}{3}\right) \tag{2.4}$$
- so both methods have similar computation cost if $m \approx n$ but the normal equations method is up to twice as fast for $m \gg n$
- the flop count in (2.3) assumes that we do Householder QR (discussed later) since the matrix is dense
- the flop count in (2.4) assumes that we do Cholesky factorization (discussed later)

2.4. **Roundoff errors.**

- another issue with the normal equations is the loss of information when we roundoff
- for example, if
$$A = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \end{bmatrix}, \qquad A^\mathsf{T}A = \begin{bmatrix} 1+\epsilon^2 & 1 \\ 1 & 1 \end{bmatrix},$$
and ϵ is so small that your computer rounds off $1 + \epsilon^2$ to 1, then you end up with a rank-deficient matrix
$$\mathrm{fl}(A^\mathsf{T}A) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$
- note that for the QR method, we work directly with A and do not need to form $A^\mathsf{T}A$ so we don't face this problem
- statisticians often use the normal equations because in many statistical problems, the measurement errors in A are much larger than the roundoff errors and so the latter type of errors are relatively insignificant

3. Full rank least squares via augmented system

- we can cast the normal equation in another form
- let $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ be the residual
- now by the normal equations
$$A^*\mathbf{r} = A^*\mathbf{b} - A^*A\mathbf{x} = \mathbf{0}$$

- and so we obtain the system

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

- in matrix form, we get

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

- this is often a large system since the coefficient matrix has dimension $(m+n) \times (m+n)$, but it preserves the sparsity of A