Math 5610 Fall 2021

Notes of 9/22/21

Study Session

• Study Session tomorrow, at 10:45, on Zoom
• Login Link on Canvas
• Also: https://utah.zoom.us/j/96915702991
• pass code is 5610
• Available for consultation by appointment
Sensitivity Analysis for \( Ax = b \)

- Consider the square linear system

\[ Ax = b \] (1)

- How do errors
  - in \( A \)
  - in \( b \)
  - in the solution process affect the accuracy of the computed solution?

- There are two main types of error analysis. In \textbf{Forward Error Analysis} we examine how errors in the data (\( A \) or \( b \)) propagate through the specific method used for solving \( Ax = b \) and effect the solution. A classic example of this technique is \textbf{Interval Analysis} (available, for example, in the Matlab INTLAB package) where numbers are represented as intervals (containing the number being represented).

- The second type, which is less specific, but more generally applicable, is \textbf{Backward Error Analysis} where the numerical solution is thought of \textit{not as the approximate solution of an exactly given problem}, but as the \textit{exact solution of an approximately given problem}. The results so obtained are \textit{independent} of the numerical method used to solve the mathematical problem.
• In these notes we explore this idea in the context of the linear system (1).

• Suppose that, by some means or other, we obtain an approximate solution \( \hat{x} \) of (1) that contains an error \( e \), i.e.,

\[
\hat{x} = x - e.
\] (2)

• We know neither \( x \) nor \( e \), but we can compute the residual

\[
r = b - A\hat{x}.
\] (3)

• It is crucial that \( e \) and \( r \) are related in the same way as \( x \) and \( b \):

\[
Ae = A(x - \hat{x}) = Ax - A\hat{x} = b - A\hat{x} = r.
\] (4)

This illustrates a general principle: for any linear problem, the error satisfies the same equation as the solution, except that the right hand side is replaced by the residual. This applies not just to linear equations but also to linear operator equations like linear differential and integral equations.

• This principal is so important that I’ll derive it in a different, perhaps more intuitive, way. Start with

\[
A\hat{x} = A(x - e) = Ax - r = b - r.
\] (5)
Then we have 
\[ Ax = b \]
and 
\[ A\hat{x} = b - r \]
Take the difference: 
\[ Ae = r. \] (6)

- The essence of backward error analysis is that we consider \( \hat{x} \) the exact solution of the perturbed problem \( A\hat{x} = b - r \) rather than the approximate solution of the problem \( Ax = b \). So we are addressing the question of how perturbations of the right hand side effect the solution of the original problem (1).

- So how do we get a handle on this? We are interested in the error, or its norm. We can afford a large error if the solution itself is large. So we are asking about the relative error \( \|e\|/\|x\| \) (which we don’t know) and how it relates to the relative residual \( \|r\|/\|b\| \) (which we can compute).

- One way to approach the matter is simply to write down what we know, and then go from there. A creative step is to invoke the inverse of \( A \), even though of course we never actually invert a matrix.

- Using any vector norm and the induced ma-
t otherwise:

\[ Ax = b \quad \| b \| \leq \| A \| \| x \| \quad [1] \]

\[ A^{-1} b = x \quad \| x \| \leq \| A^{-1} \| \| b \| \quad [2] \]

\[ Ae = r \quad \| r \| \leq \| A \| \| e \| \quad [3] \]

\[ A^{-1} r = e \quad \| e \| \leq \| A^{-1} \| \| r \| \quad [4] \]

(7)

- We can combine two inequalities in (7) by dividing the smaller of one by the larger of the other, and the larger of the one by the smaller of the other. In particular, if we divide the larger side of [4] by the smaller of [1], and the smaller of [4] by the larger of [1], we obtain

\[
\frac{\| e \|}{\| A \| \| x \|} \leq \frac{\| A^{-1} \| \| r \|}{\| b \|}
\]

which can be rewritten as

\[
\frac{\| e \|}{\| x \|} \leq \| A \| \| A^{-1} \| \frac{\| r \|}{\| b \|}.
\]

(9)

- Similarly, we obtain by combining [2] and [3]:

\[
\frac{1}{\| A \| \| A^{-1} \|} \frac{\| r \|}{\| b \|} \leq \frac{\| e \|}{\| x \|}.
\]

(10)

- Combining (9) and (10) we obtain

\[
\frac{1}{\| A \| \| A^{-1} \|} \frac{\| r \|}{\| b \|} \leq \frac{\| e \|}{\| x \|} \leq \| A \| \| A^{-1} \| \frac{\| r \|}{\| b \|}
\]

(11)
These inequalities merit deep study!

**Notes:**

1. The expression $\|A\|\|A^{-1}\|$ is called the **condition number of** $A$ with the respect to the underlying vector norm. Unless it’s specified otherwise the norm is usually the 2-norm.

2. The matrix $A$, and the linear system (1), are said to be **ill conditioned** if $\|A\|\|A^{-1}\|$ is large, and **well conditioned** if $\|A\|\|A^{-1}\|$ is small. The meaning of “large” and “small” depends on the context and will become clearer in the following notes.

3. The relative residual will usually be greater than the round-off unit, i.e., the smallest positive number $\tau$ for which the machine recognizes that $1 + \tau$ is greater than 1. On a Unix system, $\tau \approx 10^{-16}$. Thus a solution cannot be expected to have more than 16 correct digits. If the condition number is $10^p$ then this means that the solution may have no more than $16 - p$ correct digits. If the condition number is $10^{16}$ then the relative error may be as large as 1, the error will be as large as the solution, and therefore we won’t have a solution.

4. We never compute a matrix inverse, and so it’s not trivial to compute or approximate $\|A\|\|A^{-1}\|$. An effective approximation, implemented in LAPACK and MATLAB is described in the beautiful paper

5. While the condition number does depend on the norm being used, it can be large regardless of the norm. Let $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$ the largest and smallest of the absolute values of the eigenvalues of $A$. Since for any induced matrix norm, $\|A\| \geq \lambda_{\text{max}}$ and $\|A^{-1}\| \geq 1/\lambda_{\text{min}}$, we have

\[
\|A\|\|A^{-1}\| \geq \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}. \tag{12}
\]

6. It’s clear (e.g., from (12)), that

\[
\|AA^{-1}\| \leq \|A\|\|A^{-1}\| \geq 1. \tag{13}
\]

A matrix with condition number 1 is as well conditioned as it can be.

7. The condition number (with respect to the 2-norm) of an orthogonal matrix is 1. In other words, orthogonal matrices are as well conditioned as possible.

8. A matrix is singular if and only if zero is one of its eigenvalues. Consider a non-singular matrix $A$ where one eigenvalue remains constant, and another approaches zero. In that sense, $A$ approaches singularity. As it does so, according to (12), the condition number approaches infinity.

9. This illustrates a general point: If there is a mathematical singularity you expect numeri-
cal difficulties if you are close to that singularity.

10. The inequalities (12) are sharp. It is possible, for any vector norm and its corresponding induced matrix norm, and for either of the inequalities in (12), to find vectors $x$ and $e$ such that the inequality is satisfied with equality. It’s a good exercise to verify this statement.

11. The inequalities (11) say that the relative error may be as large as the relative residual multiplied with the condition number, or as small as the relative residual divided by the condition number.

12. However, one can do a probabilistic analysis that shows that with a high probability the right hand inequality in (11) is satisfied with close to equality. In other words, the quotient

$$\frac{\|e\|}{\|x\|} \leq \frac{\|A\|\|A^{-1}\|}{\|r\|/\|b\|}$$

will be less than, but close to 1. For details see the above mentioned paper by Cline et al.

13. The inequalities (11) are independent of the method that is used to solve (1). This means that once you have an ill-conditioned linear system, there is little you can do. The key to handling ill-conditioning is to avoid it, not to fight it. We will see various ways of doing this in various contexts as we proceed through the semester. The example below illustrates this idea.
14. The condition number of a $1 \times 1$ “matrix” is of course 1 (why?). The formal view of a diagonal matrix is that it may be ill-conditioned. For example, the 2-norm condition number of the matrix
\[
A^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 10^p \end{bmatrix}
\]
is $10^p$ and can be arbitrarily large. However, linear equations with a diagonal matrix do not interact at all. It is more fruitful to think of a diagonal “system” $Dx = b$ as a set of $n$ separate single equations, all of which have condition number 1.

Exercise. In the above analysis we interpret the error $e$ as being caused by a perturbation $r$ of the right hand side $e$. How about a perturbation of the coefficient matrix? Consider the system
\[
(A - E)(x - e) = b
\]
and show that
\[
\frac{\|e\|}{\|x - e\|} \leq \|A\|\|A^{-1}\|\|E\|/\|A\|.
\]
Thus we have the same sort of result: the relative error in the solution may be as large as the condition number multiplied with the relative error in the coefficient matrix. Again, the inequality is sharp.
An Example

Ill-conditioning can easily occur when doing things that come naturally.

Suppose we want to approximate a function $f$ with a polynomial $p$ such that

$$\int_0^1 (f(x) - p(x))^2 \, dx = \min.$$  \hfill (18)

Writing

$$p(x) = \sum_{j=0}^{n} \alpha_j x^j$$  \hfill (19)

we obtain the minimization problem

$$F(\alpha_0, \alpha_1, \ldots, \alpha_n) = \int_0^1 \left( f(x) - \sum_{j=0}^{n} \alpha_j x^j \right)^2 \, dx = \min$$  \hfill (20)

To solve this problem we proceed as usual, by differentiating with respect to variables, and setting the gradient equal to zero. This gives

$$\frac{\partial}{\partial \alpha_i} F = -2 \int_0^1 \left( f(x) - \sum_{j=0}^{n} \alpha_j x^j \right) x^i \, dx = 0, \quad i = 0, \ldots, n$$  \hfill (21)

This can be written as the $(n+1) \times (n+1)$ linear system
where $H_{n+1}$ is the $(n + 1) \times (n + 1)$ Hilbert Matrix with entries

$$h_{ij} = \int_0^1 x^i x^j dx = \frac{1}{i + j + 1}, \quad i, j = 0, \ldots, n$$

Thus, for example,

$$H_4 = \begin{bmatrix} 1/1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \end{bmatrix}$$

The Hilbert matrix is obviously symmetric, and it’s a good exercise to show that it is positive definite. It also happens to have an inverse where all entries are integer. However, floating point computations with the Hilbert matrix are essentially impossible, due to its poor conditioning, as illustrated in this Table:

$$\begin{array}{cccccc}
  n : & 2 & 3 & 6 & 10 & 15 \\
  \|H_n\|_2\|H_n^{-1}\|_2 : & 19 & 524 & 1.5 \times 10^6 & 1.6 \times 10^{13} & 2.5 \times 10^{17} \\
\end{array}$$

Thus we’ll lose about 17 of 16 digits when computing with the $15 \times 15$ Hilbert matrix!
Figure 1. Monomials $x^n$, $n = 0, 1, \ldots, 10$.

It is not surprising that the Hilbert matrix is ill-conditioned, because its rows all look the same! If they were identical the matrix would be singular, if they all look similar the matrix is ill-conditioned (despite being positive definite).
Figure 2. Legendre Polynomials $P_n$, $n = 0, 1, \ldots, 10$.

It's also no surprise that our basic approach gives rise to an ill-conditioned system, because the monomials in (19) all have the same shape. To the seasoned numerical analysts it makes no sense to express a polynomial, that can have many oscillations, in terms of a bunch of basis functions that all have the same shape, and no oscillations at all, as in Figure 1.
So what can we do? The idea is to use polynomial basis functions that do not have these deficiencies. In this case one should use **shifted Legendre Polynomials**. Letting

\[ <f,g> = \int_0^1 f(x)g(x)dx \]  

(26)

these are defined as follows

\[
\begin{align*}
P_0(x) &= 1 \\
P_1(x) &= x - a_1 \\
P_n(x) &= (x - a_n)P_{n-1} - b_n P_{n-2} \\
\end{align*}
\]

where

\[
\begin{align*}
a_n &= \frac{<xP_{n-1}, P_{n-1}>}{<P_{n-1}, P_{n-1}>} \quad \text{and} \quad b_n = \frac{<xP_{n-1}, P_{n-2}>}{<P_{n-2}, P_{n-2}>}
\end{align*}
\]

(27)

This gives (exercise)

\[
\begin{align*}
P_0(x) &= 1 \\
P_1(x) &= 2x - 1 \\
P_2(x) &= 6x^2 - 6x + 1 \\
P_3(x) &= 20x^3 - 30x^2 + 12x - 1 \\
P_4(x) &= 70x^4 - 140x^3 + 90x^2 - 20x + 1 \\
P_5(x) &= 252x^5 - 630x^4 + 560x^3 - 210x^2 + 30x - 1 \\
\end{align*}
\]

\[ \ldots \]  

(28)

---

see, e.g., Cheney, Introduction to Approximation Theory, p. 107
With these polynomials one does not need to solve a linear system at all, since the resulting matrix, corresponding to the Hilbert matrix, is diagonal (exercise). In fact, the solution of the problem (18) is, simply,

\[ p(x) = \sum_{j=0}^{n} \frac{<f, P_j>}{<P_j, P_j>} P_j(x). \]  

(29)

The basis functions \( P_i \) all look very different, as illustrated in Figure 2.

So here is the moral of our story:

\[ \textbf{Don’t fight ill-conditioning, avoid it!} \]  

(30)