

Lecture 06 - C

§8.4 Stability

T. Takahashi

Department of Mathematics and Statistics
University of New Mexico



We will answer these questions:

- Why are some systems problematic even if they are non-singular?
- What is a condition number and why is it important?
- What is a residue and what does it represent?
- How well can trust the solution to a Gaussian elimination code?



Visualizing III-conditioning



Intersecting lines

- Suppose you want the location where 2 lines intersect

$$y = -2x + 6$$

$$y = \frac{1}{2}x + 1$$

- We may write the problem in matrix form

$$\begin{bmatrix} 2 & 1 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$$

- The solution is $x = [2 \ 2]' \implies$ they intersect at (2, 2)
- Notice that $A = \begin{bmatrix} 2 & 1 \\ -1/2 & 1 \end{bmatrix}$ is nonsingular
- The lines are **not parallel** so they must intersect at exactly one point



Four Cases

- ① The lines **intersect**

$$\begin{bmatrix} 2 & 1 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$$

- ② The lines are **parallel** and **distinct**

$$\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

- ③ The lines are **parallel** and **coincident**

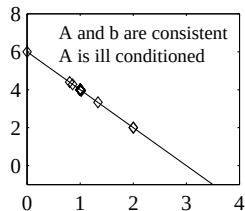
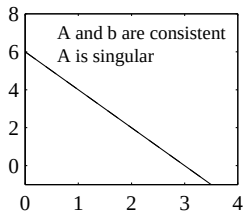
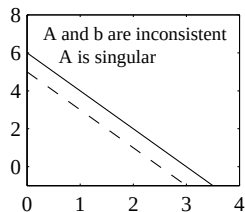
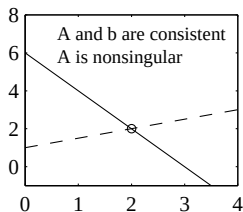
$$\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

- ④ The lines are **nearly parallel**

$$\begin{bmatrix} 2 & 1 \\ 2 + \delta & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$



Visualizing the Four Cases



The nearly parallel case is *very* problematic to handle on the computer.



Nearly Parallel Case

- Consider **nearly parallel** case

$$\begin{bmatrix} 2 & 1 \\ 2 + \delta & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

- In theory, it should have one unique solution
- In practice, we may get very inaccurate numerical answers

Question

- Are there larger $n \times n$ matrices with the same issue?
- Is there a way to tell if a matrix A has severe issues?
- Can we quantify the severity of the issue?



Nearly Parallel Case

- Consider **nearly parallel** case

$$\begin{bmatrix} 2 & 1 \\ 2 + \delta & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

- In theory, it should have one unique solution
- In practice, we may get very inaccurate numerical answers

Question

- Are there larger $n \times n$ matrices with the same issue? **yes**
- Is there a way to tell if a matrix A has severe issues? **yes**
- Can we quantify the severity of the issue? **yes**



Linear Algebra Review



Norms

Definition: Norm

The **norm** of a vector or matrix is a non-negative quantity that represents its size or magnitude. We denote it with $\|\cdot\|$.

Norms of Vectors

$$\|x\|_1 = |x_1| + |x_2| + \dots + |x_n| = \sum_{i=1}^n |x_i|$$

$$\|x\|_\infty = \max(|x_1|, |x_2|, \dots, |x_n|) = \max_i(|x_i|)$$

Norms of Matrices

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad \text{theoretical norm}$$

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \quad \text{largest column sum}$$

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}| \quad \text{largest row sum}$$



p -Norms

- Vectors:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{1/p}$$

- Matrices:

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

- p is a real number $p \geq 1$

Euclidean Norm

$p = 2$ is a special case:

$$\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \dots + |x_n|^2}$$

This is also called the **Euclidean norm**. This is default norm when you do `norm(x)` in Matlab.

Some Important Properties of Norms

Positivity

$$\|x\| \geq 0, \text{ and } \|x\| = 0 \text{ only if } x = 0$$

Homogeneity

$$\|cx\| = |c|\|x\|$$

Submultiplicative (some norms do not have this property)

$$\|Ax\| \leq \|A\|\|x\|$$

$$\|AB\| \leq \|A\|\|B\|$$

Triangle Inequality

$$\|x + y\| \leq \|x\| + \|y\|$$

useful in proofs



What is an Estimate?

Definition: Estimation

Estimation of a variable x refers to finding an approximate upper bound.

Example

Estimate the quantity $x = 47 \times 99$

- Since $47 < 50$ and $99 < 100$

$$\begin{aligned}x &= 47 \times 99 \\&< 50 \times 100 \\&= 5000\end{aligned}$$

- Our estimate is $x < 5000$



- We are mainly interested in estimating the error

$$\begin{aligned}e &= |x - 5000| \\&= |(47)(99) - 5000| \\&= 5000 \left| \frac{47}{50} \frac{99}{100} - 1 \right| \\&= 5000 |.94 \times .99 - 1| \\&\leq 5000 |.9 - 1| = 500\end{aligned}$$

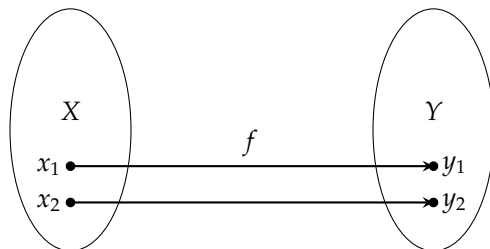
- Guaranteed my estimate is within 500 of the real solution (4653)



Perturbations



Perturbation

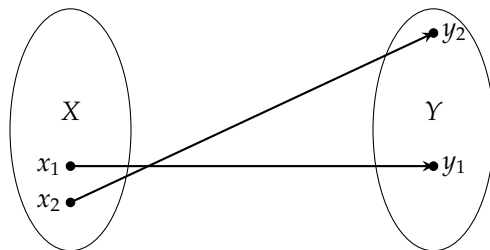


small change in input $\implies$ small change in output

- **Perturbation** means a small change in a parameter
- The mapping f has “good conditioning” if it is “stable” with respect to perturbations.



Perturbation



small change in input $\implies$ large change in output

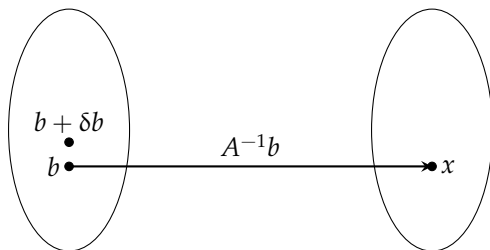
- A mapping f is “ill-conditioned” if it is “sensitive” to perturbations.



Perturbation

Right-hand Side

Solution



- Suppose we want to solve $Ax = b$.
- b gets mapped to x
- Where does $b + \delta b$ get mapped to?



Question

In the problem $Ax = b$, will a small error b lead to a big change in the solution?

Example:

- Suppose the right-hand side of a 2×2 system is

$$b = \begin{bmatrix} 1 \\ 2/3 \end{bmatrix}$$

- Consider the difference in the two problems

$$Ax = \begin{bmatrix} 1 \\ 2/3 \end{bmatrix} \quad \text{and} \quad A\hat{x} = \begin{bmatrix} 1 \\ 0.6667 \end{bmatrix}$$

- What can we say about x and $\hat{x}$?



Perturbation Notations

- For our unperturbed problem $Ax = b$,

$x =$ **exact solution**

$b =$ **exact RHS**

- We may introduce a small perturbation to the RHS so that $\hat{b} = b + \delta b$

$\delta b =$ **perturbation to RHS**

$\hat{b} =$ **perturbed RHS**

- We may write our perturbed system as $A\hat{x} = \hat{b}$

$\hat{x} = x + \delta x_b =$ **perturbed solution**

$\delta x_b =$ **perturbation to solution caused by δb**



Effect of Perturbations to b

- Suppose b is perturbed with δb such that

$$\frac{\|\delta b\|}{\|b\|} \ll 1$$

- Our solution is shifted by δx_b , i.e.

$$A(x + \delta x_b) = b + \delta b$$

- Analysis shows (see next two slides for proof) that

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$

Thus, the effect of the perturbation to b is small if $\|A\| \|A^{-1}\|$ is small.

$$\text{if } \|A\| \|A^{-1}\| \sim 1 \quad \text{then} \quad \frac{\|\delta x_b\|}{\|x\|} \ll 1$$



- Let $x + \delta x_b$ be the exact solution to the perturbed system

$$A(x + \delta x_b) = b + \delta b \quad (1)$$

- Expand

$$Ax + A\delta x_b = b + \delta b$$

- Subtract Ax from left side and b from right side since $Ax = b$

$$A\delta x_b = \delta b$$

- Left multiply by A^{-1} to get a formal solution for δx_b

$$\delta x_b = A^{-1}\delta b \quad (2)$$



- Take norm of equation (2)

$$\|\delta x_b\| = \|A^{-1} \delta b\|$$

- Utilize the **submultiplicative inequality** to get

$$\|\delta x_b\| \leq \|A^{-1}\| \|\delta b\| \quad (3)$$

- Similarly, $Ax = b$ gives $\|b\| = \|Ax\|$, and

$$\|b\| \leq \|A\| \|x\|, \quad (4)$$

or rather

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \quad (5)$$



- Multiply Equation (5) by Equation (3) to get

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|} \quad (6)$$

Theorem

If $x + \delta x_b$ is the exact solution to the perturbed system

$$A(x + \delta x_b) = b + \delta b$$

then

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$



Effect of Perturbations to b

- The quantity $\frac{\|\delta x_b\|}{\|x\|}$ is the relative error in x

$$\frac{\|\delta x_b\|}{\|x\|} = \frac{\|x - \hat{x}\|}{\|x\|} = \frac{|\text{true} - \text{approx}|}{|\text{true}|} \quad (7)$$

- Similarly, the quantity $\frac{\|\delta b\|}{\|b\|}$ is the relative error in b

Key Idea

This theorem actually gives us an error estimate of the solution in terms of the perturbation!

$$\underbrace{\frac{\|\delta x_b\|}{\|x\|}}_{\text{rel. error in } x} \leq \underbrace{\|A\| \|A^{-1}\|}_{\text{some number}} \underbrace{\frac{\|\delta b\|}{\|b\|}}_{\text{rel. error in } b}$$



Condition Number



Effect of Perturbations to A

Question

What if there is error in A with δA such that

$$\frac{\|\delta A\|}{\|A\|} \ll 1?$$

Theorem

For the perturbed system

$$(A + \delta A)(x + \delta x_A) = b,$$

we have

$$\frac{\|\delta x_A\|}{\|x + \delta x_A\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta A\|}{\|A\|}.$$

Thus, the effect of the perturbation to A is small if $\|A\| \|A^{-1}\|$ is small.

$$\text{if } \|A\| \|A^{-1}\| \sim 1 \quad \text{then} \quad \frac{\|\delta x_A\|}{\|x + \delta x_A\|} \ll 1$$



Effect of Perturbations to both A and b

Question

What if there is error in both A with δA and b with δb such that

$$\frac{\|\delta A\|}{\|A\|} \ll 1 \quad \text{and} \quad \frac{\|\delta b\|}{\|b\|} \ll 1?$$

Theorem

For the perturbed system

$$(A + \delta A)(x + \delta x) = b + \delta b,$$

we have the estimation on the relative error

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right].$$

Thus, the effect of the perturbation is small if $\|A\| \|A^{-1}\|$ is small.

$$\text{if } \|A\| \|A^{-1}\| \sim 1 \quad \text{then} \quad \frac{\|\delta x\|}{\|x\|} \ll 1$$



Common Factor

All three theorems have something in common:

- 1 Perturbation in b

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$

- 2 Perturbation in A

$$\frac{\|\delta x_A\|}{\|x + \delta x_A\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta A\|}{\|A\|}$$

- 3 Perturbation in A and b

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right]$$



Common Factor

All three theorems have something in common:

- ① Perturbation in b

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$

- ② Perturbation in A

$$\frac{\|\delta x_A\|}{\|x + \delta x_A\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta A\|}{\|A\|}$$

- ③ Perturbation in A and b

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right]$$

They all have the factor $\|A\| \|A^{-1}\|$.



Condition number of A

Definition: Condition Number

The **condition number** of a square nonsingular matrix is

$$\kappa(A) = \|A\| \|A^{-1}\|.$$

If A is singular then $\kappa(A) = \infty$.

The condition number is always in the range

$$1 \leq \kappa(A) \leq \infty$$

- $\kappa(A)$ is a mathematical property of A
- large $\kappa(A) \implies Ax = b$ is sensitive to errors
- $\kappa(A)$ indicates how close A is to being numerically singular.
- A matrix with large κ is said to be **ill-conditioned**
- Usually computed with 2-norm $\|\cdot\|_2$



Why is $\kappa \geq 1$?

- In a sense, κ numerically describes invertibility of A
- The most easily inverted matrix is the identity $A = I$

$$\kappa(I) = \|I\| \|I^{-1}\| = \|I\| \|I\| = 1$$

- You can also prove it explicitly using the submultiplicative inequality

$$\begin{aligned}\kappa(A) &= \|A\| \|A^{-1}\| \\ &\geq \|AA^{-1}\| \\ &\geq \|I\| = 1\end{aligned}$$



Significance of κ

- The condition number $\kappa(A)$ gives a sense of its severity
- If κ was small then this theorem

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$

says the error in b will not cause any issues in the computed answer

- If κ was large, small errors in b may cause drastic errors in δx_b
- In Matlab, use the command `cond(A)`

Key Idea

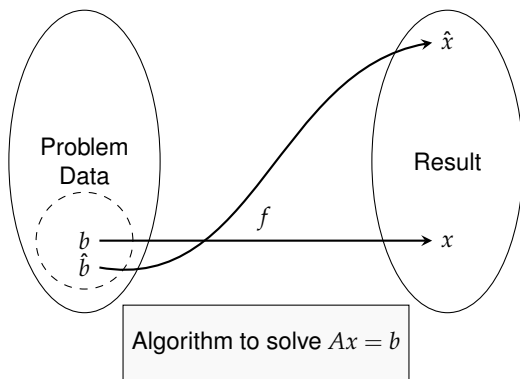
We always expect some mistakes when solving $Ax = b$ due to rounding. The condition number κ tells us how much of an impact these small errors will have on the numerical solution.



Stability



Stability of an Algorithm



- **Stability** is a property of an *algorithm*
 - Stable algorithms do not propagate error
- Do not confuse it with **conditioning** which is a property of the problem



A quote from very smart people

In Practice, applying Gaussian elimination with partial pivoting and back substitution to $Ax = b$ gives the **exact solution**, $\hat{x}$, to the **nearby problem**

$$(A + E)\hat{x} = b \quad \text{where} \quad \|E\|_{\infty} \leq \varepsilon_m \|A\|_{\infty}$$

Gaussian elimination with partial pivoting and back substitution “gives exactly the right answer to nearly the right question.”

— Trefethen and Bau



A quote from very smart people

In Practice, applying Gaussian elimination with partial pivoting and back substitution to $Ax = b$ gives the **exact solution**, $\hat{x}$, to the **nearby problem**

$$(A + E)\hat{x} = b \quad \text{where} \quad \|E\|_{\infty} \leq \varepsilon_m \|A\|_{\infty}$$

Gaussian elimination with partial pivoting and back substitution “gives exactly the right answer to nearly the right question.”

— Trefethen and Bau

The Meaning

- GE is stable because it limits propagation of rounding error
- But GE needs to have partial pivoting



The Residual

Question

Suppose we use an algorithm to get $\hat{x}$. In practice, we don't know the true solution to $Ax = b$ so we can't compute error $e = \|\hat{x} - x\|$. Then how do we know if $\hat{x}$ is accurate?

Definition: Residual

The **residual** measures how close $\hat{x}$ is to satisfying the original equation

$$r = b - A\hat{x}$$

We call the magnitude $\|r\|$ the **backward error**.

- If $x \approx \hat{x}$, then $Ax \approx A\hat{x}$. Therefore, $A\hat{x} \approx b$ should be small.
- The residue gives us a sense of error even if the true solution is not known



The Residual Theorem

Question

How reliable is $\|r\|$ as a measure of error?

Theorem

For a square linear system $Ax = b$,

$$\frac{\|\hat{x} - x\|}{\|\hat{x}\|} \leq \kappa(A) \frac{\|r\|}{\|A\| \|\hat{x}\|}$$

Key Idea

Just because the residue is small does not mean your error is small. It depends on the condition number of A .



GE Rule of Thumb



The Backslash Operator

- The formal solution to a linear problem is given by

$$Ax = b \quad \implies \quad x = A^{-1}b$$

- It is bad practice to do the following

```
1 x = inv(A)*b % solution to A*x = b
```

- Why not?
 - **Accuracy:** A can be ill-conditioned
 - **Speed:** Slower than using a specialized solver
 - **Memory:** Storing A^{-1} requires an additional $n \times n$ doubles

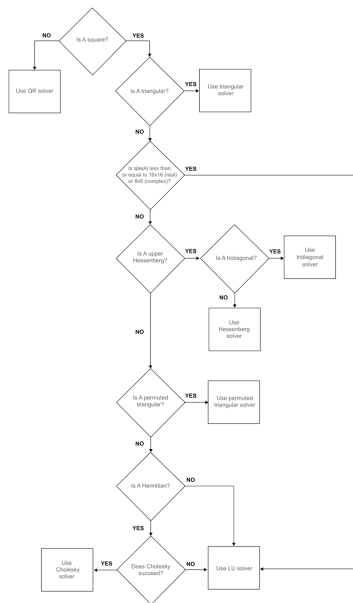
- Use this instead - **backslash operator**

```
1 x = A\b % solution to A*x = b
```



What does it do?

- The backslash operator check's for structure in the A matrix to perform optimal algorithm
- Example: if will do back substitution if A is triangular $O(n^2)$
- Last resort: GE with partial pivoting $O(n^3)$



Example Usage

If you wanted to solve

$$\begin{bmatrix} 1 & 400 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

in Matlab, you would type

```
1 >> A = [1 400; 3 1];  
2 >> b = [3; 4];  
3 >> x = A\b  
4 x =  
5     1.331943286071726  
6     0.004170141784821
```

Question

Matlab shows the answers up to 15 digits. How many of these digits can we actually trust?



How many digits can you trust?

Rule of Thumb

If A and b are stored to machine precision ε_m , then solving $Ax = b$ with Gaussian elimination with partial pivoting produces a solution that is accurate to

$$d = \lceil \log_{10}(\varepsilon_m) \rceil - \log_{10}(\kappa(A))$$

significant decimal digits.

- The value d tells you the number of significant digits you can safely quote.
- For example, if in the previous example $d = 4$ then

$x =$

1.331943286071726

0.004170141784821

All numbers after the 4th place are worthless



Rule of Thumb Example

Example

Suppose your system has a condition number $\kappa(A) \sim 10^{10}$. If you solve the system using FP64 ($\varepsilon_m \approx 2.2 \times 10^{-16}$) then how many correct decimal digits will the solution have?

We utilize the rule of thumb:

$$\begin{aligned}d &= |\log_{10}(\varepsilon_m)| - \log_{10}(\kappa(A)) \\&= |\log_{10}(2.2 \times 10^{-16})| - \log_{10}(10^{10}) \\&\approx 15 - 10 \\&= 5 \text{ correct (decimal) digits}\end{aligned}$$



Rule of Thumb Example

Example

Suppose your system has a condition number $\kappa(A) \sim 10^{10}$. If you solve the system using FP64 ($\varepsilon_m \approx 2.2 \times 10^{-16}$) then how many correct decimal digits will the solution have?

We utilize the rule of thumb:

$$\begin{aligned} d &= |\log_{10}(\varepsilon_m)| - \log_{10}(\kappa(A)) \\ &= |\log_{10}(2.2 \times 10^{-16})| - \log_{10}(10^{10}) \\ &\approx 15 - 10 \\ &= 5 \text{ correct (decimal) digits} \end{aligned}$$

Key Idea

You can initially trust up to 15 or 16 digits in double precision. After performing numerous arithmetic operations to solve for x , you lose $\log_{10}(\kappa)$ digits.

How many digits can we trust?

Question

Matlab shows the answers up to 15 digits. How many of these digits can we actually trust?

```
1 >> A = [1 400; 3 1];  
2 >> b = [3; 4];  
3 >> x = A\b  
4 x =  
5     1.331943286071726  
6     0.004170141784821  
7  
8 >> d = abs(log10(eps)) - log10(cond(A))  
9 d =  
10    13.528253505318069
```

We can trust about 13 digits out of the 15 shown. Effectively our answer is

```
x =  
1.331943286071726  
0.004170141784821
```



Summary of Limits to Numerical Solution of $Ax = b$

- 1 $\kappa(A)$ indicates how close A is to being numerically singular ($\kappa = 1$ being the nicest)
- 2 If $\kappa(A)$ is “large”, A is **ill-conditioned** and even the best numerical algorithms will produce a solution, $\hat{x}$ that cannot be guaranteed to be close to the true solution, x
- 3 In practice, Gaussian elimination with partial pivoting and back substitution produces a solution with a small residual

$$r = b - A\hat{x}$$

even if $\kappa(A)$ is large.

- 4 As a rule of thumb you expect to lose k digits of precision where $k = \log_{10}(\kappa)$

