STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2023
LECTURE 8

1. introduction

- if $\mathbf{x}$ is the exact solution and $\widehat{\mathbf{x}}$ is the computed solution, errors like $\|\mathbf{x} - \widehat{\mathbf{x}}\|$ can never be known in reality since we do not know the exact solution $\mathbf{x}$ but the important point is that:

  errors can be bounded
- say we want to determine bounds on the error in a computed solution to $A\mathbf{x} = \mathbf{b}$ where $A \in \mathbb{R}^{n\times n}$ is nonsingular
- let $\mathbf{x}$ be exact solution, i.e., $\mathbf{x} = A^{-1}\mathbf{b}$ analytically,[1] and $\widehat{\mathbf{x}}$ be solution computed via floating-point arithmetic — therefore there will be rounding error in $\widehat{\mathbf{x}}$
- *backward error analysis* means we view $\widehat{\mathbf{x}}$ as the exact solution of the "nearby" system
  $$(A + \Delta A)\widehat{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$$
  - if
  $$\frac{\|\Delta A\|}{\|A\|} \le \varepsilon, \quad \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|} \le \varepsilon \tag{1.1}$$
  - then
  $$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{2\varepsilon}{1-\rho}\kappa(A) \tag{1.2}$$
    where
  $$\rho = \|\Delta A\|\|A^{-1}\| = \|\Delta A\|\kappa(A)/\|A\|$$
  - we will see how to derive this in (2.5)
- it is really relative error that we bound
  - absolute error $\|\mathbf{x} - \widehat{\mathbf{x}}\|$ is difficult to bound and is dependent on the choice of units of measurement
  - relative error $\|\mathbf{x} - \widehat{\mathbf{x}}\|/\|\mathbf{x}\|$ can be more readily bounded and is independent of units
- as was pointed out earlier, in either case we can't compute the error (absolute or relative) exactly since we don't know $\mathbf{x}$
- but it's enough to be able to *bound* errors: e.g. if we know that the error is less than 10^{-6}, we know our answer has at least 5 digits of accuracy
- the number
  $$\kappa(A) = \|A\|\|A^{-1}\|$$
  is the *condition number* of A — a singularly important notion
- why important? $\kappa(A)$ measures how an error in the system $A\mathbf{x} = \mathbf{b}$ is amplified in the solution
- even if ε is small, the computed solution can be useless if $\kappa(A)$ is large
- a system $A\mathbf{x} = \mathbf{b}$ where $\kappa(A)$ is large is an example of an *ill-conditioned* problem
- no algorithm, no matter how accurate, will be an effective tool for solving such an ill-conditioned problem
- it is important to distinguish between ill-conditioned problems from unstable algorithms

Date: October 28, 2023, version 1.0. Comments, bug reports: `lekheng@uchicago.edu`.

[1] you should never ever compute inverse explicitly but using it in mathematical expressions is OK

- informally, a problem or an algorithm is *stable* if a small change in its input yields a small change in its output
- ensuring that a problem is well-conditioned is the responsibility of the modeller, who formulates the mathematical problem from the original application
- ensuring the stability of an algorithm is the responsibility of the numerical analyst
- for a problem, the output is the exact solution, whereas for an algorithm, the output is the computed solution

2. simple perturbation theory

- in homework 2, you will be asked to do a more accurate version of this analysis
- as an illustration, we will do a simplified version where we assume that the error occurs only in $\mathbf{b} \in \mathbb{R}^n$ but $A \in \mathbb{R}^{n\times n}$ is known exactly and is nonsingular
- let $\mathbf{x} \in \mathbb{R}^n$ be the unique exact solution to
$$A\mathbf{x} = \mathbf{b} \tag{2.1}$$
- $\mathbf{x}$ is the 'true solution' we seek and it's unique because A is nonsingular
- taking norms, we get
$$\|\mathbf{b}\| \le \|A\|\|\mathbf{x}\|$$
where the norm on A is the operator norm
- hence
$$\frac{1}{\|\mathbf{x}\|} \le \|A\|\frac{1}{\|\mathbf{b}\|} \tag{2.2}$$
- suppose the solution to (2.1) with the right-hand side perturbed to $\mathbf{b} + \Delta\mathbf{b}$ is given by[2] $\mathbf{x} + \Delta\mathbf{x}$
$$A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$$
- then $A\Delta\mathbf{x} = \Delta\mathbf{b}$ and so $\Delta\mathbf{x} = A^{-1}\Delta\mathbf{b}$
- taking norms, we get
$$\|\Delta\mathbf{x}\| \le \|A^{-1}\|\|\Delta\mathbf{b}\| \tag{2.3}$$
- combining (2.2) and (2.3), we get
$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|A^{-1}\|\|A\|\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$$
or
$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \kappa(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$$
- in this simple case, the relative error in $\mathbf{x}$ is bounded by the relative error in $\mathbf{b}$ scaled by the condition number of A
- suppose the error is only in A and $\mathbf{b}$ is known perfectly, i.e., the case
$$(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$$
- we can show that
$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\kappa(A)\dfrac{\|\Delta A\|}{\|A\|}}{1 - \kappa(A)\dfrac{\|\Delta A\|}{\|A\|}} \tag{2.4}$$
under some mild assumptions
- if the error is in both A and $\mathbf{b}$, i.e.,
$$(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$$

[2] note that this always works: if the solution is $\widehat{\mathbf{x}}$, then we just set $\Delta\mathbf{x} := \widehat{\mathbf{x}} - \mathbf{x}$

- we can show that
$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\kappa(A)\left(\dfrac{\|\Delta A\|}{\|A\|} + \dfrac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right)}{1 - \kappa(A)\dfrac{\|\Delta A\|}{\|A\|}} \tag{2.5}$$
under some mild assumptions
- (2.4) and (2.5) will be in homework 2
- what we did above is usually called perturbation analysis, which assumes nothing about the error
- if we assume that all errors come from rounding errors with bounds like (1.1) on the input, then we can deduce error bounds on the output like (1.2) from (2.5), this is called *backward error analysis*

3. condition number of a matrix

- we defined the 2-norm condition number for a nonsingular square matrix $A \in \mathbb{C}^{n\times n}$ as
$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 \tag{3.1}$$
- what if A is singular? one way is to set $\kappa_2(A) = \infty$
- this is natural though not very useful — the only information it conveys is what you already know, namely, A is singular
- if we apply SVD of A and the unitary invariance of the 2-norm, then an alternative expression for (3.1) is
$$\kappa_2(A) = \frac{\sigma_1(A)}{\sigma_n(A)} = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$
- note that $\operatorname{rank}(A) = n$ and we could have written
$$\kappa_2(A) = \frac{\sigma_1(A)}{\sigma_{\operatorname{rank}(A)}(A)} = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \tag{3.2}$$
where $\sigma_{\min}(A)$ denotes the smallest non-zero singular value of A
- this last expression extends to any singular and even rectangular $A \in \mathbb{C}^{m\times n}$ as long as $A \ne O$
- note that $\sigma_{\operatorname{rank}(A)}(A)$ is the smallest non-zero singular value of A
- we call (3.2) the *generalized condition number* to distinguish it from (3.1)
- another expression for (3.2) is
$$\kappa_2(A) = \|A\|_2\|A^{\dagger}\|_2 \tag{3.3}$$
- proof: use SVD to see that $\|A\|_2 = \sigma_1(A)$ and $\|A^\dagger\|_2 = 1/\sigma_{\operatorname{rank}(A)}(A)$
- (3.3) can be used to extend generalized condition number to any matrix norm, for example
$$\kappa_p(A) = \|A\|_p\|A^\dagger\|_p, \quad \kappa_{\mathsf{F}}(A) = \|A\|_{\mathsf{F}}\|A^\dagger\|_{\mathsf{F}}$$

4. condition number as distance to ill-posedness

- a *well-posed* instance of a problem is one whose existence and uniqueness of solutions hold, otherwise the instance is *ill-posed*
- for example for the problem $A\mathbf{x} = \mathbf{b}$,
$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

is an ill-posed instance because a solution does not exist; and

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{4.1}$$

is an ill-posed instance because solutions are not unique; but

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{4.2}$$

is a well-posed instance

- you will often hear of people speaking of "ill-posed problem" and "well-posed problem" — what they really mean are "ill-posed instance of a problem" and "well-posed instance of a problem"
- the notion of conditioning is a further refinement of the notion of well-posedness, for example, the instance in (4.2) and the instance

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-1000} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{4.3}$$

are both well-posed but intuitively we know that (4.3) is worse because the coefficient matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 10^{-1000} \end{bmatrix} \approx \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

is nearly singular and so this instance is very close to the ill-posed instance (4.1)

- in fact, in floating point arithmetic of any realistic precision,

$$\mathrm{fl}\left(\begin{bmatrix} 1 & 0 \\ 0 & 10^{-1000} \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

- the respective condition numbers of the coefficient matrices are

$$\kappa\left(\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right) = 2, \qquad \kappa\left(\begin{bmatrix} 1 & 0 \\ 0 & 10^{-1000} \end{bmatrix}\right) = 10^{1000}, \qquad \kappa\left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right) = \infty$$

- the first is *well-conditioned*, the second is *ill-conditioned*, both are *well-posed*; but the third is *ill-posed*
- more generally, the condition number of an instance of a problem is the reciprocal of the normalized distance to the nearest ill-posed instance
- for example, if the problem is solving a linear system with a *fixed* right-hand side **b**, or, equivalently, matrix inversion , i.e.,

$$f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}, \qquad f(X) = X^{-1},$$

then the condition number of a problem instance $A \in \mathbb{R}^{n\times n}$ is

$$\kappa(A) = \begin{cases} \|A\|\|A^{-1}\| & A \text{ invertible;} \\ \infty & \text{otherwise} \end{cases}$$

- why? because the set of ill-posed problem is the set of singular matrices $\mathcal{M} = \{X \in \mathbb{R}^{n\times n} : \det(X) = 0\}$ and the distance of any nonsingular A to this set is (exercise: verify this for the matrix 2-norm)

$$\mathrm{dist}(A, \mathcal{M}) := \min_{X\in\mathcal{M}} \|A - X\| = \frac{1}{\|A^{-1}\|}$$

and so the normalized distance is

$$\frac{\mathrm{dist}(A, \mathcal{M})}{\|A\|} = \frac{1}{\|A\|\|A^{-1}\|} = \frac{1}{\kappa(A)},$$

the reciprocal of the usual condition number

- in other words, the "condition number of a matrix" that we defined earlier really comes from the condition number of the problem of matrix inversion or the problem of solving linear systems with fixed right-hand side
- the reciprocal condition number tells us how close a matrix is to singularity
- you should never compute $\det(A)$ to check whether $A \in \mathbb{R}^{n\times n}$ is singular/inverticle, instead you should compute $1/\kappa(A)$, given by `rcond` in MATLAB

> anyone who computes the determinant of a matrix to check invertibility in their actual or pseudo codes fails this class instantly

- we can also do this for the problem of solving for a minimum norm least squares problem with a fixed $\mathbf{b}$, which is equivalent to finding pseudoinverse
$$f : \mathbb{R}^{m\times n} \to \mathbb{R}^{n\times m}, \qquad f(X) = X^{\dagger}$$
- in this case, the set of ill-posed problem is the set of rank-deficiency matrices $\mathcal{M} = \{A \in \mathbb{R}^{m\times n} : \operatorname{rank}(A) < \min(m,n)\}$
- the condition number of a problem instance $A \in \mathbb{R}^{m\times n}$ is
$$\frac{\operatorname{dist}_{\mathsf{F}}(A,\mathcal{M})}{\|A\|_{\mathsf{F}}} = \min_{X\in\mathcal{M}} \|A - X\|_{\mathsf{F}} = \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)} = \frac{1}{\kappa_{\mathsf{F}}(A)},$$
- in other words, the "generalized condition number of a matrix" that we defined earlier really comes from the condition number of the problem of computing pseudoinverse or the problem of solving minimum norm least squares problem with a fixed $\mathbf{b}$
- there are many others: linear system, least squares, linear programming, eigenvalue problems, polynomial eigenvalue problems
- for example, for a linear programming problem
$$\begin{array}{ll} \text{minimize} & \mathbf{c}^{\mathsf{T}}\mathbf{x} \\ \text{subject to} & A\mathbf{x} \le \mathbf{b} \end{array}$$
where inequality between two vectors is interpreted in an entrywise sense, the condition number is given by
$$\frac{1}{\kappa_2(A,\mathbf{b})} = \frac{\operatorname{dist}_2([A,\mathbf{b}],\mathcal{M})}{\|[A,\mathbf{b}]\|_2}$$
where $\mathcal{M}$ = boundary of feasible pairs $(A,\mathbf{b}) \in \mathbb{R}^{m\times(n+1)}$
- later we will derive the condition number for an eigenvalue problem and in a future lecture, we will derive that for a least squares problem
- the condition number of a matrix is in some sense a misnomer — the same matrix A can have different condition numbers depending on which problem it appears in, whether linear system, least squares, eigenvalue problem, singular value problem, linear programming, etc

5. CONDITION NUMBER AS DERIVATIVE

- the definition of a condition number in the previous section as distance to ill-posedness can sometimes be difficult to use
- there is a slightly less general version that is often easier to use
- suppose we can formulate the problem in the form of evaluating a function $f : \mathbb{R}^n \to \mathbb{R}^m$, where the domain and codomain can be replaced by open subsets of more general spaces
- for example solving a nonsingular linear system $A\mathbf{x} = \mathbf{b}$ may be written as
$$f : \mathrm{GL}(n) \times \mathbb{R}^n \to \mathbb{R}^n, \qquad (A,\mathbf{b}) \mapsto A^{-1}\mathbf{b}$$
solving a minimum norm least squares $\min\|A\mathbf{x} - \mathbf{b}\|$ as
$$f : \mathbb{R}^{m\times n} \times \mathbb{R}^m \to \mathbb{R}^n, \qquad (A,\mathbf{b}) \mapsto A^{\dagger}\mathbf{b}$$

computing singular decomposition $A = U\Sigma V^{\mathsf{T}}$ as

$$f : \mathbb{R}^{m\times n} \to \mathrm{O}(m) \times \mathbb{R}^{\min(m,n)} \times \mathrm{O}(n), \qquad A \mapsto (U, \boldsymbol{\sigma}, V) \tag{5.1}$$

where $\Sigma = \operatorname{diag}(\boldsymbol{\sigma})$, and so on

- write $\operatorname{RelError}(\mathbf{x})$ the relative error in a quantity $\mathbf{x}$
- then condition number of input $\mathbf{a} \in \mathbb{R}^n$ with respect to the problem f is

$$\kappa_f(\mathbf{a}) \coloneqq \lim_{\delta\to 0} \sup_{\operatorname{RelError}(\mathbf{a}\le\delta} \frac{\operatorname{RelError}(f(\mathbf{a}))}{\operatorname{RelError}(\mathbf{a})} \tag{5.2}$$

- in other words, it quantifies the worst possible magnification of the output error with respect to a small input perturbation
- for small values of δ, we expect

$$\kappa_f(\mathbf{a}) \approx \sup_{\operatorname{RelError}(\mathbf{a}\le\delta} \frac{\operatorname{RelError}(f(\mathbf{a}))}{\operatorname{RelError}(\mathbf{a})}$$

and therefore

$$\operatorname{RelError}(f(\mathbf{a})) \lesssim \kappa_f(\mathbf{a}) \cdot \operatorname{RelError}(\mathbf{a}) \tag{5.3}$$

where '$\lesssim$' means 'roughly bounded by'

- or more precisely

$$\operatorname{RelError}(f(\mathbf{a})) \le \kappa_f(\mathbf{a}) \cdot \operatorname{RelError}(\mathbf{a}) + o(\operatorname{RelError}(\mathbf{a})) \tag{5.4}$$

as $\operatorname{RelError}(\mathbf{a}) \to 0$

- if we set $\operatorname{RelError}(\mathbf{x}) = \|\Delta\mathbf{x}\|/\|\mathbf{x}\|$ in the usual way, then

$$\kappa_f(\mathbf{a}) \coloneqq \lim_{\delta\to 0} \sup_{\|\Delta\mathbf{a}\|\le\delta} \left(\frac{\|\Delta f(\mathbf{a})\|}{\|f(\mathbf{a})\|} \Bigg/ \frac{\|\Delta\mathbf{a}\|}{\|\mathbf{a}\|} \right) = \frac{\|Df(\mathbf{a}\|}{\|f(\mathbf{a})\|/\|\mathbf{a}\|} \tag{5.5}$$

if f is differentiable at $\mathbf{a}$

- note the this depends on a choice of norms used to quantify the relative error
- for example, the condition number of computing square root $f : (0, \infty) \to \mathbb{R}$, $a \mapsto \sqrt{a}$ is

$$\kappa_f(a) = \frac{|f'(a)|}{|f(a)|/|a|} = \frac{1/(2\sqrt{a})}{\sqrt{a}/a} = \frac{1}{2}$$

a very well-conditioned problem

- more generally for any univariate differentiable $f : \mathbb{R} \to \mathbb{R}$

$$\kappa_f(a) = \left| \frac{af'(a)}{f(a)} \right|$$

- for another example, the ∞-norm condition number of subtraction $f : \mathbb{R}^2 \to \mathbb{R}$, $(a, b) \mapsto a-b$ is

$$\kappa_f(a,b) = \frac{\|\mathrm{J}\, f(a,b)\|_\infty}{|f(a,b)|/\|\left[\begin{smallmatrix}a\\b\end{smallmatrix}\right]\|_\infty} = \frac{\left\|\begin{bmatrix}1 & -1\end{bmatrix}\right\|_\infty}{|a-b|/\max(|a|,|b|)} = \frac{2\max(|a|,|b|)}{|a-b|}$$

where $\mathrm{J}\, f = \begin{bmatrix}\partial f/\partial a & \partial f/\partial b\end{bmatrix}$ is the Jacobian of f

- the condition number of subtraction is large when $a \approx b$, i.e., subtraction of nearly equal numbers is an ill-conditioned problem
- this is the root cause of cancellation error that we discussed in the last lecture
- the condition number of matrix-vector product for a fixed matrix $A \in \mathbb{R}^{m\times n}$ is given by $f : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \mapsto A\mathbf{x}$ can be derived from

$$\sup_{\|\Delta\mathbf{x}\|\le\delta} \left(\frac{\|A(\mathbf{x}+\Delta\mathbf{x}) - A\mathbf{x}\|}{\|A\mathbf{x}\|} \Bigg/ \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \right) = \left(\sup_{\|\Delta\mathbf{x}\|\le\delta} \frac{\|A\Delta\mathbf{x}\|}{\|\Delta\mathbf{x}\|} \right) \Bigg/ \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|A\|\|\mathbf{x}\|}{\|A\mathbf{x}\|}$$

where we have used the operator norm on A and there is no need to take $\lim_{\delta\to 0}$ since the right-hand side does not depend on δ

- therefore the condition number of matrix-vector multiplication by a fixed A is given by
$$\kappa_f(\mathbf{x}) = \frac{\|A\|\|\mathbf{x}\|}{\|A\mathbf{x}\|}$$
- if $A \in \mathrm{GL}(n)$, then $\|\mathbf{x}\| = \|A^{-1}A\mathbf{x}\| \le \|A^{-1}\|\|A\mathbf{x}\|$ so $\|\mathbf{x}\|/\|A\mathbf{x}\| \le \|A^{-1}\|$ and thus
$$\kappa_f(\mathbf{x}) \le \|A\|\|A^{-1}\| = \kappa(A)$$
so the matrix condition number plays a role in matrix-vector multiplication too
- for a challenge, find the condition number of matrix-vector product $f : \mathbb{R}^{m\times n} \times \mathbb{R}^n \to \mathbb{R}^m$, $(A, \mathbf{x}) \mapsto A\mathbf{x}$ without fixing A
- strictly speaking the condition number we defined above is *relative condition number*, there is also a notion of *absolute condition number* where relative error RelError is replaced by absolute error AbsError in (5.2) and if set $\mathrm{AbsError}(\mathbf{a}) = \|\Delta\mathbf{a}\|$, then (5.5) becomes
$$\kappa_f^{\mathrm{ab}}(\mathbf{a}) \coloneqq \lim_{\delta\to 0} \sup_{\|\Delta\mathbf{a}\|\le\delta} \frac{\|\Delta f(\mathbf{a})\|}{\|\Delta\mathbf{a}\|} = \|Df(\mathbf{a}\|$$
- in other words, absolute condition number is exactly the norm of the derivative of f
- in this case we also the absolute error analogues of (5.3) and (5.4)
$$\begin{aligned} \mathrm{AbsError}(f(\mathbf{a})) &\lesssim \kappa_f^{\mathrm{ab}}(\mathbf{a}) \cdot \mathrm{AbsError}(\mathbf{a}) \\ \mathrm{AbsError}(f(\mathbf{a})) &\le \kappa_f^{\mathrm{ab}}(\mathbf{a}) \cdot \mathrm{AbsError}(\mathbf{a}) + o(\mathrm{AbsError}(\mathbf{a})) \end{aligned} \tag{5.6}$$
- it turns out the relations (5.3), (5.4), (5.6) are much more useful than exact definitions like distance to ill-posedness or (5.2) in the practical calculations of condition number
- remember this: condition numbers are never needed to any degree of accuracy
- all we need to know is the order of magnitude of our condition number to gauge the level of accuracy of our results

anyone who gives condition number to more than two significant digits fails this class instantly

6. Condition number for eigenvalue and singular value problems

- basically what we did in section 2 is applying (5.3) and (5.4) to get the (relative) condition number for the problem of solving a linear system $A\mathbf{x} = \mathbf{b}$
- here we will use (5.6) to obtain the (absolute) condition number for eigenvalue decomposition, singular value decomposition, and eigenvalue problem
- suppose $A \in \mathbb{C}^{n\times n}$ is diagonalizable and so it has an eigenvalue decomposition
$$A = X\Lambda X^{-1}$$
- because of all kinds of errors, the matrix of eigenvalues we actually computed is not Λ but $\Lambda + \Delta\Lambda$ for some $\Delta\Lambda$, not necessarily a diagonal matrix
- so we assume that with $\Lambda + \Delta\Lambda$, we get an exact decomposition
$$A + \Delta A = X(\Lambda + \Delta\Lambda)X^{-1} \tag{6.1}$$
for some matrix $A + \Delta A$ with the same eigenvectors
- so this gives us
$$\Delta\Lambda = X^{-1}(\Delta A)X$$
- taking norms, we get
$$\|\Delta\Lambda\| \le \|X\|\|X^{-1}\|\|\Delta A\| = \kappa(X)\|\Delta A\| \tag{6.2}$$

- note that the condition number appears again but in this case, it is the condition number of the matrix of eigenvectors X and not the original matrix A
- this is a very crude analysis — since $\Delta\Lambda$ is not a diagonal matrix, it actually doesn't tell us how badly the eigenvalues of A are affected by an error ΔA
- but the general idea is correct: sensitivity of the eigenvalues is estimated by the condition number of the matrix of eigenvectors
- if we try to do the same thing for singular value decomposition $A = U\Sigma V^*$, the analogue of (6.1) is

  $$A + \Delta A = U(\Sigma + \Delta\Sigma)V^*$$

  and we deduce the analogue of (6.2)

  $$\|\Delta\Sigma\| = \|\Delta A\|$$

  for any unitarily invariant norm $\|\cdot\|$ (e.g., 2- or F-norm)
- in other words, (5.1) is always a perfectly conditioned problem regardless of what matrix A you are given
- we never hear of "condition number of singular value decomposition" because it is always 1
- in particular computing singular value decomposition is independent of the matrix condition number $\kappa(A)$, any respectable numerical software like MATLAB always computes singular value decomposition to full floating point accuracy
- we will do a more precise analysis for a single eigenvalue $A\mathbf{x} = \lambda\mathbf{x}$
- the surprising thing is that even if we only interested in in right eigenvector, left eigenvector will make an appearance
- suppose λ is eigenvalue of A with right eigenvector $\mathbf{x}$ and left eigenvector $\mathbf{y}$, i.e.,

  $$A\mathbf{x} = \lambda\mathbf{x}, \qquad \mathbf{y}^*A = \lambda\mathbf{y}^*$$

- backward analysis starts from

  $$(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = (\lambda + \Delta\lambda)(\mathbf{x} + \Delta\mathbf{x})$$

- ignoring second order terms (i.e., terms involving two Δ's) and using $A\mathbf{x} = \lambda\mathbf{x}$, we get

  $$\Delta A\mathbf{x} + A\Delta\mathbf{x} = \Delta\lambda\mathbf{x} + \lambda\Delta\mathbf{x}$$

- multiplying by left eigenvector and using $\mathbf{y}^*A = \lambda\mathbf{y}^*$, we get

  $$\mathbf{y}^*\Delta A\mathbf{x} + \lambda\mathbf{y}^*\Delta\mathbf{x} = \Delta\lambda\mathbf{y}^*\mathbf{x} + \lambda\mathbf{y}^*\Delta\mathbf{x}$$

  and so

  $$\Delta\lambda = \frac{\mathbf{y}^*\Delta A\mathbf{x}}{\mathbf{y}^*\mathbf{x}}$$

- taking absolute value, applying Cauchy–Schwartz and using submultiplicativity of matrix 2-norm give

  $$|\Delta\lambda| \le \frac{\|\mathbf{y}\|_2\|\mathbf{x}\|_2}{|\mathbf{y}^*\mathbf{x}|}\|\Delta A\|_2$$

- the number

  $$\kappa(\lambda, A) := \frac{\|\mathbf{y}\|_2\|\mathbf{x}\|_2}{|\mathbf{y}^*\mathbf{x}|}$$

  is called the *eigenvalue condition number* of A and λ
- note that

  $$\kappa(\lambda, A) \ge 1 \tag{6.3}$$

- note also that $\kappa(\lambda, A)$ depends only on the directions of the right/left eigenvectors $\mathbf{x}$, $\mathbf{y}$ and is independent of how we normalize them

- we could do the standard thing and normalize them to unit vectors
$$\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2 = 1 \tag{6.4}$$
and in which case we see that
$$\kappa(\lambda, A) = \frac{1}{|\mathbf{y}^*\mathbf{x}|}$$
i.e., the eigenvalue condition number depends on the angle between the left and right eigenvectors
- but there is another interesting normalization
- suppose A is diagonalizable, recall from Homework **0**, Problem **5**(b) that we may always choose the matrix of left eigenvectors Y so that
$$Y^* = X^{-1} \tag{6.5}$$
where Y is the matrix of left eigenvectors
- this is easy to deduce from $A = X\Lambda X^{-1}$ iff $A^* = X^{-*}\Lambda X^* = Y\Lambda Y^{-1}$
- if we choose the left eigenvectors so that (6.5) holds, then we have
$$Y^*X = I$$
which implies that $\mathbf{y}^*\mathbf{x} = 1$ for the left/right eigenvectors corresponding to the same λ and so
$$\kappa(\lambda, A) = \|\mathbf{y}\|_2\|\mathbf{x}\|_2$$
- now since $\|\mathbf{x}\|_2 \le \|X\|_2$ and $\|\mathbf{y}\|_2 \le \|Y\|_2 = \|X^{-1}\|_2$ (why?)
$$\kappa(\lambda, A) = \|\mathbf{y}\|_2\|\mathbf{x}\|_2 \le \|X\|_2\|X^{-1}\|_2 = \kappa(X)$$
- so the individual eigenvalue condition numbers are bounded above by the condition of the eigenvector matrix, i.e., which is consistent with (6.2)
- in general it is not possible to choose $\mathbf{x}$ and $\mathbf{y}$ so that both (6.4) and (6.5) hold
- but for a normal matrix A, this is certainly possible by the spectral theorem, i.e., if A is a normal matrix, then
$$\kappa(\lambda, A) = 1$$
- the eigenvalue problem for a normal matrix is perfectly conditioned
- a characterization of the eigenvalue condition number in terms of distance to ill-posedness is more involved and beyond the scope of our course

7. ASIDE: NUMERICAL RANK

- just as the condition number (or rather its reciprocal) gives us a continuous measure of how singular a matrix is, we can do the same with the rank of a matrix
- rounding errors makes the exact rank of a matrix difficult to determine and far less useful than in pure math
- note that matrix rank is a discrete notion that is sometimes too imprecise, for example both
$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 \\ 0 & 10^{-1000} \end{bmatrix}$$
have rank 2
- another example: take a randomly generated vector $\mathbf{x} \sim N(\mathbf{0}, I_n)$ and consider the $n \times n$ matrices
$$X = [\mathbf{x}, 2\mathbf{x}, \dots, n\mathbf{x}] \quad \text{and} \quad \mathrm{fl}(X) = [\mathrm{fl}(\mathbf{x}), \mathrm{fl}(2\mathbf{x}), \dots, \mathrm{fl}(n\mathbf{x})]$$
- in the presence of rounding error, we will get
$$\operatorname{rank}(X) = 1 \quad \text{and} \quad \operatorname{rank}(\mathrm{fl}(X)) = n$$

- the singular values are much more informative
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$
- the profile or decay rate of these can often tell us the 'true rank' of a matrix
- exercise: plot the singular value profile of $\operatorname{fl}(X)$ in MATLAB
- this is the notion of what is often called *numerical rank*
- for us, numerical rank of $A \in \mathbb{C}^{m\times n}$ *is* simply the singular values
$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_r, 0, \dots, 0) \in \mathbb{R}^{\min(m,n)}$$
which contains more information than a single integer r that counts how many nonzero singular values there are but completely ignores their magnitudes
- but some folks insist that numerical rank of a matrix must be a single number like rank, not a list of numbers
- there are several proposals on how it could be defined, three of the most common ones are defined as follows
- let $\tau > 0$ be some predetermined tolerance level (in practice a small number ≈ 0.1 or $\tau = \max(m,n) \times \varepsilon_{\text{machine}}$) and $A \in \mathbb{C}^{m\times n}$ be a non-zero matrix
- the term *numerical rank* of A have variously been given to
  - the positive integer
  $$\sigma \operatorname{rank}(A) \coloneqq \min\left\{ r \in \mathbb{N} : \frac{\sigma_r(A)}{\sigma_1(A)} \geq \tau \right\}$$
  - the positive integer
  $$\rho \operatorname{rank}(A) \coloneqq \min\left\{ r \in \mathbb{N} : \frac{\sigma_{r+1}(A)}{\sigma_r(A)} \leq \tau \right\}$$
  - or the positive integer
  $$\mu \operatorname{rank}(A) \coloneqq \min\left\{ r \in \mathbb{N} : \frac{\sum_{i\geq r+1} \sigma_i(A)^2}{\sum_{i\geq 1} \sigma_i(A)^2} \leq \tau \right\}$$
  - or the positive real number
  $$\nu \operatorname{rank}(A) = \frac{\|A\|_{\mathsf{F}}^2}{\|A\|_2^2} = \frac{\sum_{i=1}^{\min(m,n)} \sigma_i(A)^2}{\sigma_1(A)^2}$$

  depending on the application