

For the three most widely used norms:

$$\begin{aligned}\kappa_2 &= \text{Cond}_2(A) \\ \kappa_1 &= \text{Cond}_1(A) \\ \kappa_\infty &= \text{Cond}_\infty(A)\end{aligned}$$

## Geometric Interpretation of $\|\cdot\|$ , $\kappa_2$

### Singular Value Decomposition

It is a proven mathematical identity that any matrix can be written as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where,  $\mathbf{U}$  and  $\mathbf{V}$  are orthogonal matrices and  $\mathbf{\Sigma}$  is a diagonal matrix such that  $\sigma_i \geq 0$ .

$$\begin{aligned}\mathbf{U}\mathbf{U}^T &= \mathbf{U}^T\mathbf{U} = \mathbf{I} \\ \mathbf{V}\mathbf{V}^T &= \mathbf{V}^T\mathbf{V} = \mathbf{I} \\ \mathbf{\Sigma} &= \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n \end{bmatrix}\end{aligned}$$

Then  $\sigma_i$  are called the singular values of  $\mathbf{A}$ . Since the matrices  $\mathbf{V}$  and  $\mathbf{U}$  are orthogonal matrices the magnitude of the matrix  $\mathbf{A}$  is contained in the matrix  $\sigma$ . The orthogonal matrices are rotating the coordinate system such that  $\sigma$  is the magnitude of matrix along each of the direction of the coordinate system.

$$\mathbf{A}_{m \times n} = \mathbf{U}_{m \times m} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \mathbf{V}_{n \times n}^T$$

Assume  $m = n$  (only to simplify the case). In that case

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

$$\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$$

$$\mathbf{A} [v_1 \dots v_i \dots v_n] = [u_1 \dots u_i \dots u_n] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}$$

Stable algorithm implies that for a small perturbation in data the numerical result does not suffer large errors. Hence,

$$|y_{\text{num}} - \hat{y}| \text{ is small}$$

where  $\hat{y}$  is the exact result of  $\hat{x}$ . Hence a stable algorithm produces nearly the exact result.

$$|\hat{x} - x| \text{ is small}$$

where  $\hat{x}$  nearly the exact problem.

**Question** How accurate (measure of forward error) is the numerical solution produced by a stable algorithm?

The forward error is given by the expression  $\frac{|y_{\text{num}} - y|}{\max(|y_{\text{num}}|, |y|)}$ . By adding and subtracting  $\hat{y}$  in the numerator and separating the terms we can write

$$\frac{|y_{\text{num}} - y|}{\max(|y_{\text{num}}|, |y|)} = \frac{|y_{\text{num}} - \hat{y} + \hat{y} - y|}{\max(|y_{\text{num}}|, |y|)}$$

An upper bound for the above expression can be found by applying a different denominator for the two terms in the above relation and can be written as

$$\frac{|y_{\text{num}} - y|}{\max(|y_{\text{num}}|, |y|)} \leq \frac{|y_{\text{num}} - \hat{y}|}{|y_{\text{num}}|} + \frac{|\hat{y} - y|}{|y|}$$

$$\frac{|y_{\text{num}} - y|}{\max(|y_{\text{num}}|, |y|)} \leq \frac{|y_{\text{num}} - \hat{y}|}{|y_{\text{num}}|} + \kappa_{\text{problem}} \frac{|\hat{x} - x|}{|x|}$$

Note that an accurate numerical result implies that the value of the forward error is small. This achieved by applying a stable algorithm to a well conditioned problem. There is no guarantee for accuracy if we apply

- Unstable algorithm to an ill-conditioned problem
- Stable algorithm to an ill-conditioned problem

- Unstable algorithm to a well-conditioned problem

It would be interesting to see the outcome of applying an unstable algorithm to an ill-conditioned problem.

Lets look at some common problems are analyze whether they are ill-conditioned or well-conditioned.

## Case Study I: Solving Linear Systems

Consider a linear system  $\mathbf{Ax} = \mathbf{b}$ , where  $\epsilon \ll 1$  and

$$\mathbf{A} = \begin{bmatrix} 1 + \epsilon & 1 \\ 1 & 1 \end{bmatrix}$$

$$\mathbf{b} = \begin{Bmatrix} b_1 \\ b_2 \end{Bmatrix}$$

Lets look at the change in the result  $\mathbf{x}$  for small changes in the system by changing  $\epsilon$ . The solution of the system  $\mathbf{x}$  is given by  $\mathbf{A}^{-1}\mathbf{b}$ . Consider the case where  $\epsilon$  is changed to  $2\epsilon$ . The change in the system is given by  $\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}$  and the change in the solution is given by  $\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}$ . The change in the system is evaluated as

$$\Delta\mathbf{A} = \begin{bmatrix} \epsilon & 0 \\ 0 & 0 \end{bmatrix}$$

Hence the relative change in the norm of the system  $\mathbf{A}$  is given by

$$\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \sim \epsilon$$

Similarly the change in the solution can be evaluated as follows

$$\mathbf{A}^{-1} = \frac{1}{\epsilon} \begin{bmatrix} +1 & -1 \\ -1 & 1 + \epsilon \end{bmatrix}$$

$$\mathbf{x} = \frac{1}{\epsilon} \begin{Bmatrix} b_1 - b_2 \\ -b_1 + (1 + \epsilon)b_2 \end{Bmatrix}$$

Hence the relative change in the norm of the solution  $\mathbf{x}$  is given by,

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \sim \frac{1}{\epsilon}$$

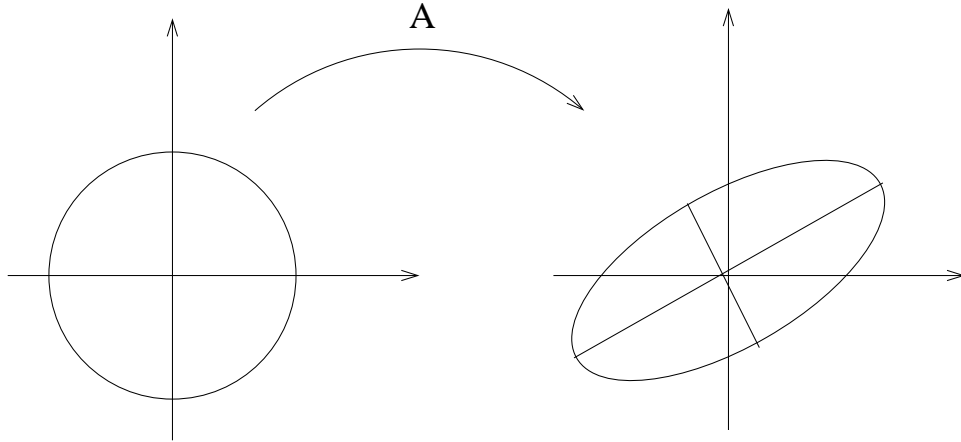


Figure 1: The transformation of the unit circle by  $\mathbf{A}$  into an ellipse

For a small change in the system of the order of  $\epsilon$  the solution changes by an amount  $\frac{1}{\epsilon}$ , which is a very large number.

Consider again the equation

$$\begin{aligned} [\mathbf{A}v_1 \dots \mathbf{A}v_i \dots \mathbf{A}v_n] &= [\sigma_1 u_1 \dots \sigma_i u_i \dots \sigma_n u_n] \\ \mathbf{A}v_i &= \sigma_i u_i \end{aligned}$$

From linear algebra we know that operating a matrix on a vector produces another vector different length and orientation and that the stretch and the rotation can be separated. As seen in the picture, the matrix  $\mathbf{A}$  transforms the unit circle made of two unit vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$  into an ellipse of semi axis lengths  $\sigma_1$  and  $\sigma_2$ .

Recall,

$$\begin{aligned} \|\mathbf{A}\|_2 &= \sigma_{\max} \\ &= \text{Maximum stretch when a vector is operated on by } \mathbf{A} \end{aligned}$$

Consider the condition number of the system defined by

$$\begin{aligned} \kappa_2(\mathbf{A}) &= \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 \\ &= \sigma_{\max}(\mathbf{A}) \sigma_{\max}(\mathbf{A}^{-1}) \end{aligned}$$

Lets try to evaluate the maximum SVD of  $\mathbf{A}^{-1}$ . The matrix  $\mathbf{A}$  can be written

in the form

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \mathbf{V}^T \quad \text{such that} \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$$

Hence, if  $\mathbf{A}$  is non-singular then

$$\mathbf{A}^{-1} = \mathbf{V} \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_n \end{bmatrix} \mathbf{U}^T$$

So the maximum SVD of  $\mathbf{A}^{-1}$  is

$$\sigma_{\max}(\mathbf{A}^{-1}) = \frac{1}{\sigma_{\min}(\mathbf{A})}$$

Hence the condition number for the system can be evaluated in terms of the SVD values of the matrix  $\mathbf{A}$  alone.

$$\kappa_2(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})} = \frac{\text{Largest semi axis of the ellipse}}{\text{Smallest semi axis of the ellipse}}$$

Hence the condition number is a measure of how skewed the ellipse is after transformation.

Lets look at some special cases of the matrix  $\mathbf{A}$  and find how the transformed ellipse looks like.

### Case I: $\mathbf{A}$ is an Orthogonal Matrix

Since  $\mathbf{A}$  is an orthogonal matrix,

$$\mathbf{A}^T \mathbf{A} = \mathbf{I}$$

Writing  $\mathbf{A}$  in terms of its the singular value decomposition,

$$\begin{aligned} \mathbf{I} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^2 \mathbf{V}^T \end{aligned}$$

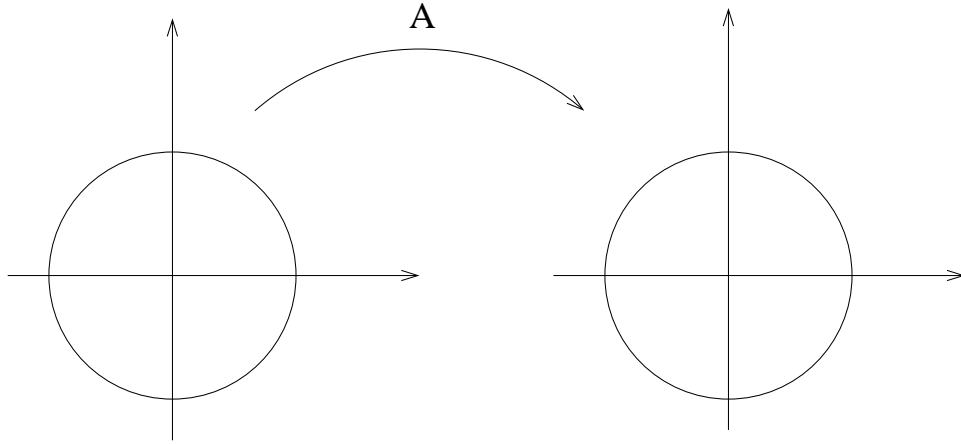


Figure 2: The matrix  $\mathbf{A}$  is orthogonal

Pre-multiplying the above relation by  $\mathbf{V}^T$  and post-multiplying the above relation by  $\mathbf{V}$ , we get

$$\begin{aligned}
 \text{pre-multiplying by } \mathbf{V}^T & \quad \mathbf{V}^T \mathbf{I} = \mathbf{V}^T (\mathbf{V} \boldsymbol{\Sigma}^2 \mathbf{V}^T) \\
 & \Rightarrow \mathbf{V}^T = \boldsymbol{\Sigma}^2 \mathbf{V}^T \\
 \text{post-multiplying by } \mathbf{V} & \quad \mathbf{V}^T \mathbf{V} = \boldsymbol{\Sigma}^2 \mathbf{V}^T \mathbf{V} \\
 & \Rightarrow \mathbf{I} = \boldsymbol{\Sigma}^2 \\
 & \Rightarrow \sigma_1 = \sigma_2 = \dots = \sigma_n = 1
 \end{aligned}$$

Hence all the SVD values of an orthogonal matrix are unity. Hence the condition number of the matrix  $\kappa(\mathbf{A})$  is unity. Hence an orthogonal matrix operated on a system of vectors forming a circle is transformed to another unit circle.



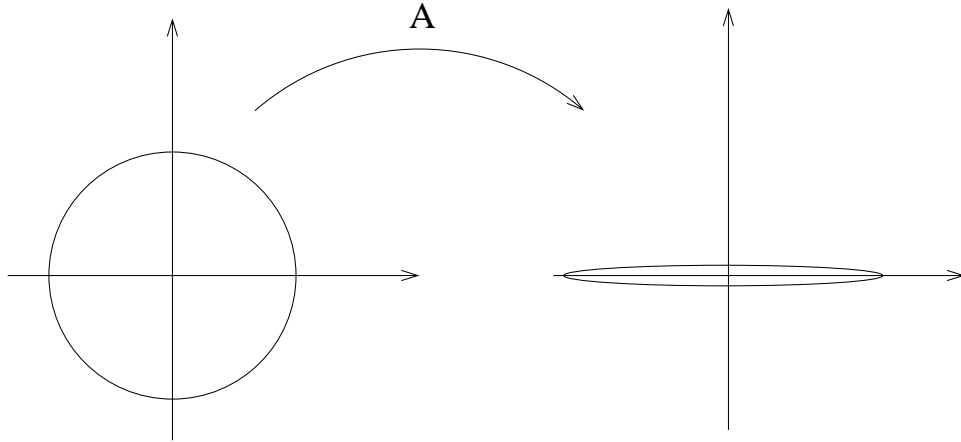


Figure 3: The matrix  $\mathbf{A}$  is singular

Hence the eigenvalues  $\mathbf{A}^T \mathbf{A}$  is contained in the diagonal matrix  $\Sigma^2$ . But  $\Sigma$  contains the singular decomposition values of the matrix  $\mathbf{A}$ . Hence SVD of the matrix  $\mathbf{A}$  is the square root of the eigen values of  $\mathbf{A} \mathbf{A}^T$ .

$$\lambda_i(\mathbf{A} \mathbf{A}^T) = (\sigma_i(\mathbf{A}))^2$$

It should be notes the eigenvalues of  $\mathbf{A}$  may be complex or negative real numbers but the singular values of  $\mathbf{A}$  are always positive.

**Question** What is the relation between the singular values, condition number of  $\mathbf{A}$  and the impact of data errors in the result?

Consider a two dimensional situation for the ease in physical explanation. Lets look at two situations of a linear system  $\mathbf{A} \mathbf{x} = \mathbf{b}$  where  $\mathbf{A}$  is orthogonal and singular. The physical representation of the solution in both the cases are shown in Fig. 4 and Fig. 5.

**Case I: A is Orthogonal** In both the cases the solid line is the exact line corresponding to the equations of the system. The dotted lines show the errors in the system due to a small change in the vector  $\mathbf{b}$ . It can be seen from both the figures that a small change in input data  $\delta b_1, \delta b_2$  produce a small change and a large change in the output data in case one and two respectively.

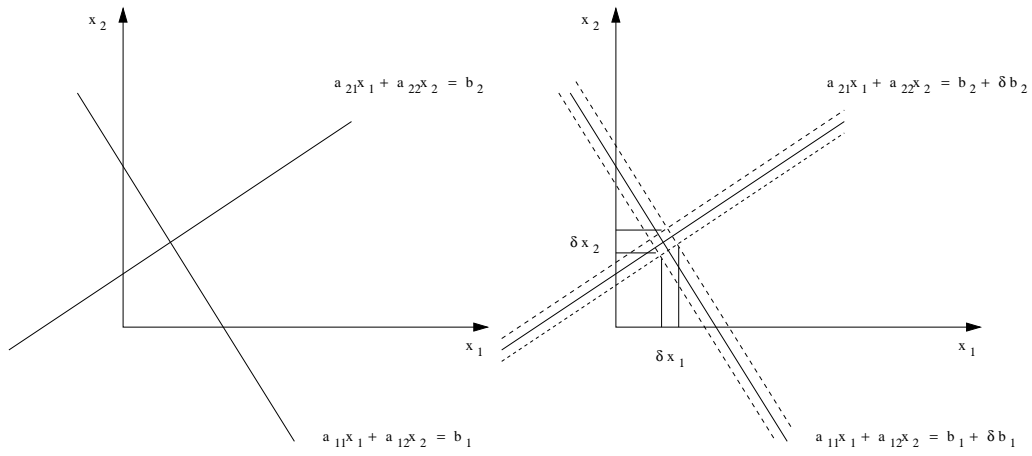


Figure 4: The matrix  $\mathbf{A}$  is orthogonal

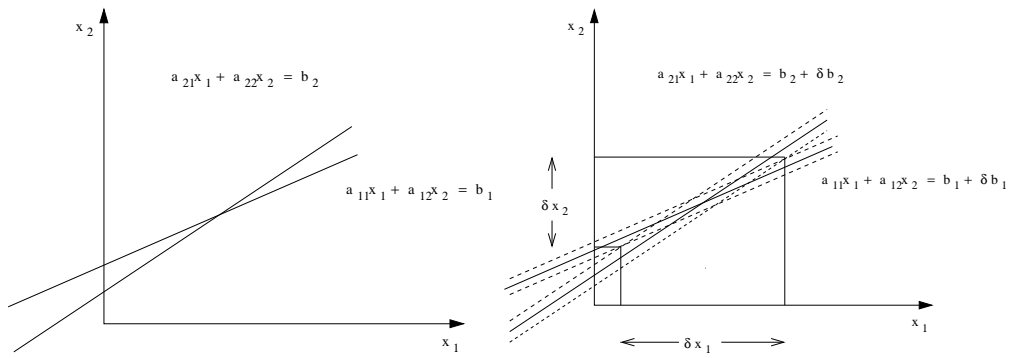


Figure 5: The matrix  $\mathbf{A}$  is singular