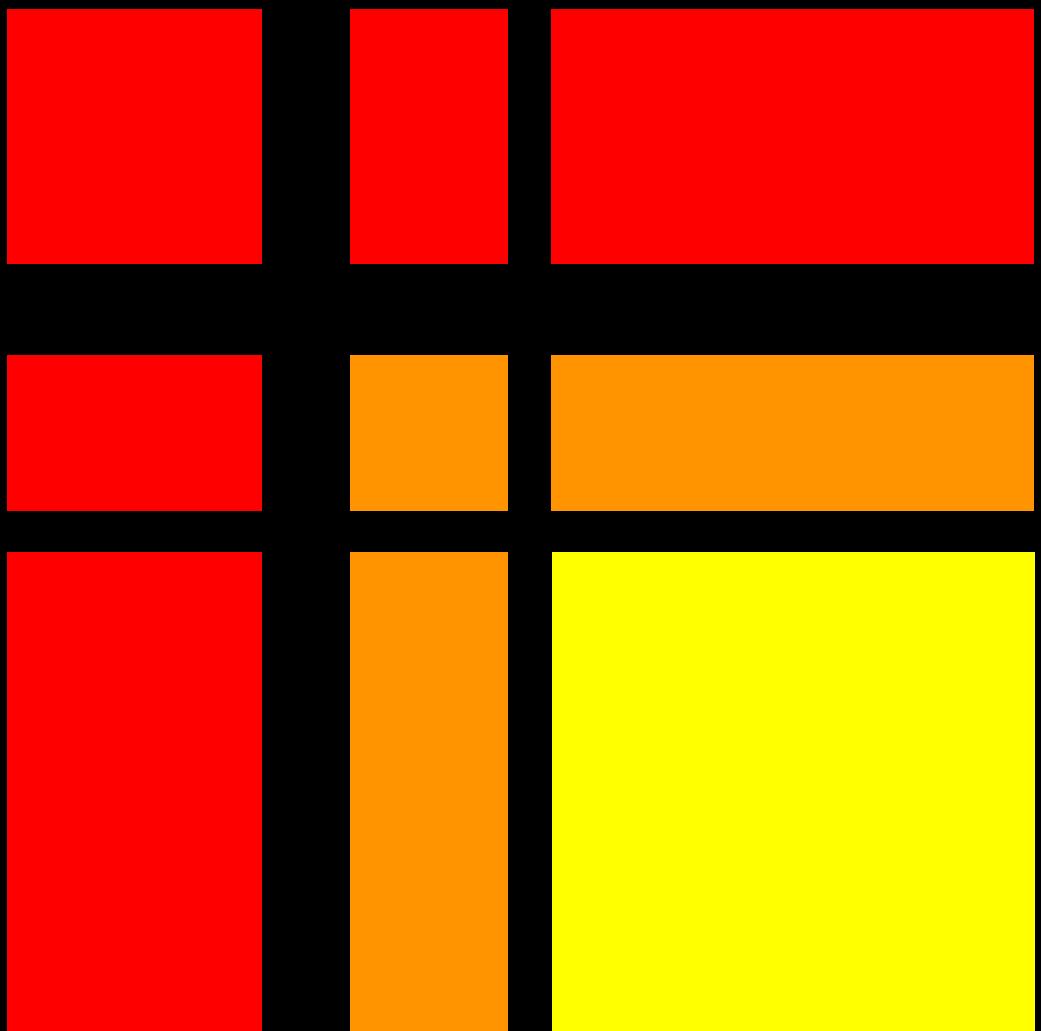


# **Advanced Linear Algebra**

## **Foundations to Frontiers**



**Robert A. van de Geijn**  
**Margaret E. Myers**

# Advanced Linear Algebra

Foundations to Frontiers



# Advanced Linear Algebra

## Foundations to Frontiers

Robert van de Geijn  
The University of Texas at Austin

Margaret Myers  
The University of Texas at Austin

January 24, 2020

**Edition:** Draft Edition 2019–2020

**Website:** [ulaff.net](http://ulaff.net)

©2019–2020 Robert van de Geijn and Margaret Myers

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the appendix entitled “GNU Free Documentation License.” All trademarks<sup>TM</sup> are the registered® marks of their respective owners.





# Acknowledgements

We would like to thank the people who created PreTeXt, the authoring system used to typeset these materials. We applaud you!

# Preface

Robert van de Geijn  
Maggie Myers  
Austin, 2019

# Contents

<b>Acknowledgements</b>	vii
<b>Preface</b>	viii
<b>0 Getting Started</b>	1
<b>I Orthogonality</b>	
<b>1 Norms</b>	12
<b>2 The Singular Value Decomposition</b>	88
<b>II Solving Linear Systems</b>	
<b>III The Algebraic Eigenvalue Problem</b>	
<b>A Notation</b>	152
<b>B Knowledge from Numerical Analysis</b>	153
<b>C GNU Free Documentation License</b>	155
<b>References</b>	163



# Week 0

## Getting Started

### 0.1 Opening Remarks

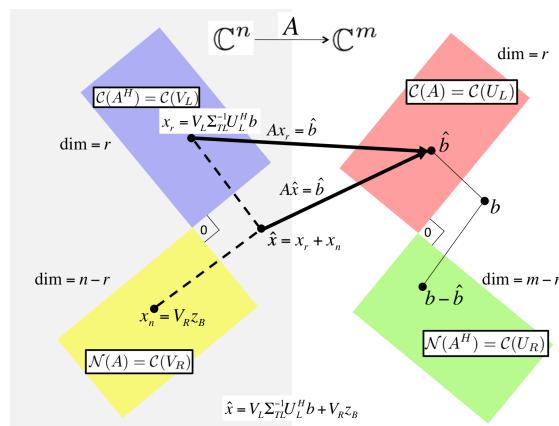
#### 0.1.1 Welcome



YouTube: <https://www.youtube.com/watch?v=KzCTMLvxtQA>

Linear algebra is one of the fundamental tools for computational and data scientists. In Advanced Linear Algebra: Foundations to Frontiers (ALAFF), you build your knowledge, understanding, and skills in linear algebra, practical algorithms for matrix computations, and how floating-point arithmetic, as performed by computers, affects correctness.

The materials are organized into Weeks that correspond to a chunk of information that is covered in a typical on-campus week. These weeks are arranged into three parts:



#### Part I: Orthogonality

The Singular Value Decomposition (SVD) is possibly the most important result in linear algebra, yet too advanced to cover in an introductory undergraduate course. To be able to get to this topic as quickly as possible, we start by focusing on orthogonality, which is at the heart of image compression, Google's page rank algorithm, and linear least-squares approximation.

### Part II: Solving Linear Systems

Solving linear systems, via direct or iterative methods, is at the core of applications in computational science and machine learning. We also leverage these topics to introduce numerical stability of algorithms: the classical study that qualifies and quantifies the "correctness" of an algorithm in the presence of floating point computation and approximation. Along the way, we discuss how to restructure algorithms so that they can attain high performance on modern CPUs.

**Algorithm:** Compute LU factorization with partial pivoting of  $A$ , overwriting  $A$  with factors  $L$  and  $U$ . The pivot vector is returned in  $p$ .

$$\text{Partition } A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), p \rightarrow \left( \begin{array}{c} p_T \\ p_B \end{array} \right).$$

where  $A_{TL}$  is  $0 \times 0$  and  $p_T$  is  $0 \times 1$

while  $n(A_{TL}) < n(A)$  do

Repartition

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ p_B \end{array} \right) \rightarrow \left( \begin{array}{c} p_0 \\ \pi_1 \\ p_2 \end{array} \right)$$

where  $\alpha_{11}, \lambda_{11}, \pi_1$  are  $1 \times 1$

$$\pi_1 = \max_i \left( \frac{\alpha_{11}}{\alpha_{21}} \right)$$

$$\left( \begin{array}{c|c|c} a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) := P(\pi_1) \left( \begin{array}{c|c|c} a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

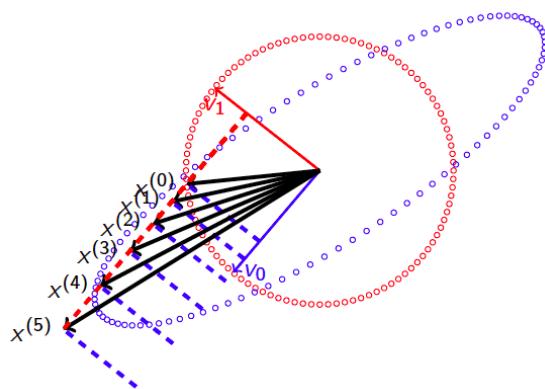
$$a_{21} := a_{21}/\alpha_{11}$$

$$A_{22} := A_{22} - a_{21}a_{12}^T$$

Continue with

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ p_B \end{array} \right) \leftarrow \left( \begin{array}{c} p_0 \\ \pi_1 \\ p_2 \end{array} \right)$$

endwhile



### Part III: Eigenvalues and Eigenvectors

Many problems in science have the property that if one looks at them in just the right way (in the right basis), they greatly simplify and/or decouple into simpler subproblems. Eigenvalue and eigenvectors are the key to discovering how to view a linear transformation, represented by a matrix, in that special way. Algorithms for computing them also are the key to practical algorithms for computing the SVD

In this week (Week 0), we walk you through some of the basic course information and help you set up for learning. The week itself is structured like future weeks, so that you become familiar with that structure.

#### 0.1.2 Outline Week 0

Each week is structured so that we give the outline for the week immediately after the "launch:"

- 0.1 Opening Remarks
  - 0.1.1 Welcome
  - 0.1.2 Outline Week 0
  - 0.1.3 What you will learn
- 0.2 Setting Up For ALAFF

- 0.2.1 Accessing these notes
- 0.2.2 Cloning the ALAFF repository
- 0.2.3 MATLAB
- 0.2.4 Setting up to implement in C (optional)
- 0.3 Enrichments
  - 0.3.1 Ten surprises from numerical linear algebra
  - 0.3.2 Best algorithms of the 20th century
- 0.4 Wrap Up
  - 0.4.1 Additional Homework
  - 0.4.2 Summary

### 0.1.3 What you will learn

The third unit of each week informs you of what you will learn. This describes the knowledge and skills that you can expect to acquire. If you return to this unit after you complete the week, you will be able to use the below to self-assess.

Upon completion of this week, you should be able to

- Navigate the materials.
- Access additional materials from GitHub.
- Track your homework and progress.
- Register for MATLAB online.
- Recognize the structure of a typical week.

## 0.2 Setting Up For ALAFF

### 0.2.1 Accessing these notes

For information regarding these and our other materials, visit [ulaff.net](http://ulaff.net).

These notes are available in a number of formats:

- As an online book authored with PreTeXt at <http://www.cs.utexas.edu/users/flame/laff/alaff/>.

- As a PDF at <http://www.cs.utexas.edu/users/flame/laff/alaff/ALAFF.pdf>.

If you download this PDF and place it in just the right folder of the materials you will clone from GitHub (see next unit), the links in the PDF to the downloaded material will work.

During Spring 2020, we will incrementally add weeks (chapters) to that material as the semester progresses. We will be updating the materials frequently as people report typos and we receive feedback from learners. Please consider the environment before you print a copy...

- Eventually, if we perceive there is demand, we may offer a printed copy of these notes from [Lulu.com](#), a self-publishing service. This will not happen until Summer 2020, at the earliest.

**Homework 0.2.1.1** If the book has chapters numbered 0 through  $n$  and you print a new copy every time a new chapter is added (first you print chapter 0, then you print chapters 0 and 1, and so forth), how many chapters (multiplicity counted) will you print?

If the book has chapters 0 through 12 and each chapter has 50 pages, how many pages do you print?

**Answer.** Number of chapters printed:  $(n + 1)(n + 2)/2$

Now prove it!

Number of pages printed:  $(12 + 1)(12 + 2)/2 \times 50 = 4550$

## 0.2.2 Cloning the ALAFF repository

We have placed all materials on GitHub, a development environment for software projects. In our case, we use it to disseminate the various activities associated with this course.

On the computer on which you have chosen to work, "clone" the GitHub repository for this course:

- Visit <https://github.com/ULAFF/ALAFF>
- Click on

 Clone or download ▾

and copy <https://github.com/ULAFF/ALAFF.git>.

- On the computer where you intend to work, in a terminal session on the command line in the directory where you would like to place the materials, execute

```
git clone https://github.com/ULAFF/ALAFF.git
```

This will create a local copy (clone) of the materials.

- Sometimes we will update some of the files from the repository. When this happens you will want to execute, in the cloned directory,

```
git stash save
```

which saves any local changes you have made, followed by

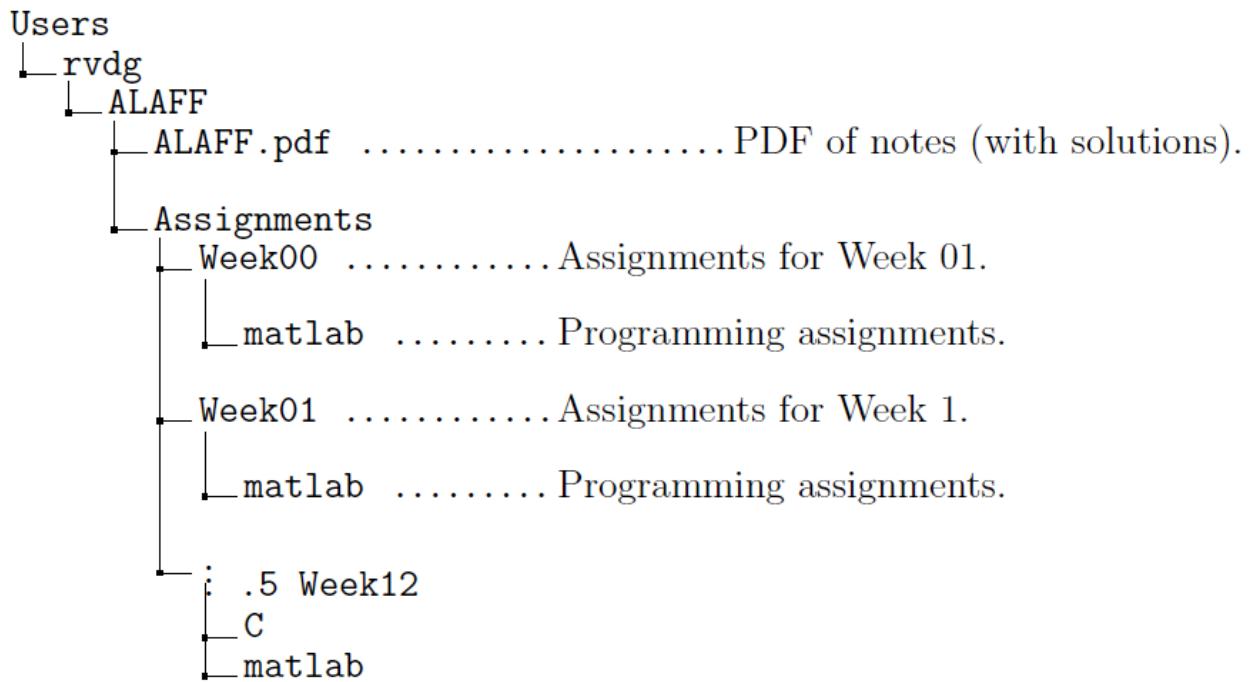
```
git pull
```

which updates your local copy of the repository, followed by

```
git stash pop
```

which restores local changes you made. This last step may require you to "merge" files that were changed in the repository that conflict with local changes.

Upon completion of the cloning, you will have a directory structure similar to that given in Figure 0.2.2.1.



**Figure 0.2.2.1** Directory structure for your ALAFF materials. In this example, we cloned the repository in Robert's home directory, rvdg.

### 0.2.3 MATLAB

We will use Matlab to translate algorithms into code and to experiment with linear algebra.

There are a number of ways in which you can use Matlab:

- Via MATLAB that is installed on the same computer as you will execute your performance experiments. This is usually called a "desktop installation of Matlab."

- Via [MATLAB Online](#). You will have to transfer files from the computer where you are performing your experiments to MATLAB Online. You could try to set up [MATLAB Drive](#), which allows you to share files easily between computers and with MATLAB Online. Be warned that there may be a delay in when files show up, and as a result you may be using old data to plot if you aren't careful!

If you are using these materials as part of an offering of the Massive Open Online Course (MOOC) titled "Advanced Linear Algebra: Foundations to Frontiers," you will be given a temporary license to Matlab, courtesy of MathWorks. In this case, there will be additional instructions on how to set up MATLAB Online, in the Unit on edX that corresponds to this section.

You need relatively little familiarity with MATLAB in order to learn what we want you to learn in this course. So, you could just skip these tutorials altogether, and come back to them if you find you want to know more about MATLAB and its programming language (M-script).

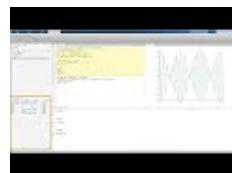
Below you find a few short videos that introduce you to MATLAB. For a more comprehensive tutorial, you may want to visit [MATLAB Tutorials](#) at MathWorks and click "Launch Tutorial".

What is MATLAB?



<https://www.youtube.com/watch?v=2sB-NMD9Qhk>

Getting Started with MATLAB  
Online



<https://www.youtube.com/watch?v=4shp284pGc8>

MATLAB Variables



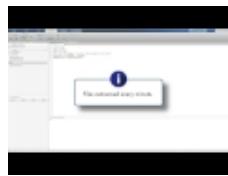
<https://www.youtube.com/watch?v=gPIsIzHJA9I>

MATLAB as a Calculator



<https://www.youtube.com/watch?v=K9xy5kQHDBo>

Managing Files with MATLAB  
Online



<https://www.youtube.com/watch?v=mqYwMnM-x5Q>

**Remark 0.2.3.1** Some of you may choose to use MATLAB on your personal computer while others may choose to use MATLAB Online. Those who use MATLAB Online will need to transfer some of the downloaded materials to that platform.

## 0.2.4 Setting up to implement in C (optional)

You may want to return to this unit later in the course. We are still working on adding programming exercises that require C implementation.

In some of the enrichments in these notes and the final week on how to attain performance, we suggest implementing algorithms that are encountered in C. Those who intend to pursue these activities will want to install a Basic Linear Algebra Subprograms (BLAS) library and our libflame library ( which not only provides higher level linear algebra functionality, but also allows one to program in a manner that mirrors how we present algorithms.)

### 0.2.4.1 Installing the BLAS

The Basic Linear Algebra Subprograms (BLAS) are an interface to fundamental linear algebra operations. The idea is that if we write our software in terms of calls to these routines and vendors optimize an implementation of the BLAS, then our software can be easily ported to different computer architectures while achieving reasonable performance.

A popular and high-performing open source implementation of the BLAS is provided by our BLAS-like Library Instantiation Software (BLIS). The following steps will install BLIS if you are using the Linux OS (on a Mac, there may be a few more steps, which are discussed later in this unit.)

- Visit the [BLIS Github repository](#).
- Click on

[Clone or download ▾](#)

and copy <https://github.com/flame/blis.git>.

- In a terminal session, in your home directory, enter

```
git clone https://github.com/flame/blis.git
```

(to make sure you get the address right, you will want to paste the address you copied in the last step.)

- Change directory to blis:

```
cd blis
```

- Indicate a specific version of BLIS so that we all are using the same release:

```
git checkout pfhp
```

- Configure, build, and install with OpenMP turned on.

```
./configure -t openmp -p ~/blis auto  
make -j8  
make check -j8  
make install
```

The -p ~/blis installs the library in the subdirectory ~/blis of your home directory, which is where the various exercises in the course expect it to reside.

- If you run into a problem while installing BLIS, you may want to consult <https://github.com/flame/blis/blob/master/docs/BuildSystem.md>.

On Mac OS-X

- You may need to install Homebrew, a program that helps you install various software on your mac. Warning: you may need "root" privileges to do so.

```
$ /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/ma
```

Keep an eye on the output to see if the "Command Line Tools" get installed. This may not be installed if you already have Xcode Command line tools installed. If this happens, post in the "Discussion" for this unit, and see if someone can help you out.

- Use Homebrew to install the gcc compiler:

```
$ brew install gcc
```

Check if gcc installation overrides clang:

```
$ which gcc
```

The output should be /usr/local/bin. If it isn't, you may want to add /usr/local/bin to "the path." I did so by inserting

```
export PATH="/usr/local/bin:$PATH"
```

into the file .bash\_profile in my home directory. (Notice the "period" before ".bash\_profile"

- Now you can go back to the beginning of this unit, and follow the instructions to install BLIS.

### 0.2.4.2 Installing libflame

Higher level linear algebra functionality, such as the various decompositions we will discuss in this course, are supported by the LAPACK library [1]. Our libflame library is an implementation of LAPACK that also exports an API for representing algorithms in code in a way that closely reflects the FLAME notation to which you will be introduced in the course.

The libflame library can be cloned from

- <https://github.com/flame/libflame>.

Instructions on how to install it are at

- <https://github.com/flame/libflame/blob/master/INSTALL>.

## 0.3 Enrichments

In each week, we include "enrichments" that allow the participant to go beyond.

### 0.3.1 Ten surprises from numerical linear algebra

You may find the following list of insights regarding numerical linear algebra, compiled by John D. Cook, interesting:

- John D. Cook. [Ten surprises from numerical linear algebra](#). 2010.

### 0.3.2 Best algorithms of the 20th century

An article published in SIAM News, a publication of the Society for Industrial and Applied Mathematics, lists the ten most important algorithms of the 20th century [6]:

1. *1946*: John von Neumann, Stan Ulam, and Nick Metropolis, all at the Los Alamos Scientific Laboratory, cook up the *Metropolis algorithm*, also known as the Monte Carlo method.
2. *1947*: George Dantzig, at the RAND Corporation, creates the *simplex method for linear programming*.
3. *1950*: Magnus Hestenes, Eduard Stiefel, and Cornelius Lanczos, all from the Institute for Numerical Analysis at the National Bureau of Standards, initiate the development of *Krylov subspace iteration methods*.
4. *1951*: Alston Householder of Oak Ridge National Laboratory formalizes the *decompositional approach to matrix computations*.
5. *1957*: John Backus leads a team at IBM in developing the *Fortran optimizing compiler*.

6. 1959–61: J.G.F. Francis of Ferranti Ltd., London, finds a stable method for computing eigenvalues, known as the *QR algorithm*.
7. 1962: Tony Hoare of Elliott Brothers, Ltd., London, presents *Quicksort*.
8. 1965: James Cooley of the IBM T.J. Watson Research Center and John Tukey of Princeton University and AT&T Bell Laboratories unveil the *fast Fourier transform*.
9. 1977: Helaman Ferguson and Rodney Forcade of Brigham Young University advance an *integer relation detection algorithm*.
10. 1987: Leslie Greengard and Vladimir Rokhlin of Yale University invent the *fast multipole algorithm*.

Of these, we will explicitly cover three: the decomposition method to matrix computations, Krylov subspace methods, and the QR algorithm. Although not explicitly covered, your understanding of numerical linear algebra will also be a first step towards understanding some of the other numerical algorithms listed.

## 0.4 Wrap Up

### 0.4.1 Additional Homework

For a typical week, additional assignments may be given in this unit.

### 0.4.2 Summary

In a typical week, we provide a quick summary of the highlights in this unit.

# Part I

## Orthogonality

# Week 1

## Norms

### 1.1 Opening

#### 1.1.1 Why norms?



YouTube: <https://www.youtube.com/watch?v=DKX3TdQWQ90>

The following exercises expose some of the issues that we encounter when computing.  
We start by computing  $b = Ux$ , where  $U$  is upper triangular.

**Homework 1.1.1.1** Compute

$$\begin{pmatrix} 1 & -2 & 1 \\ 0 & -1 & -1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} =$$

**Solution.**

$$\begin{pmatrix} 1 & -2 & 1 \\ 0 & -1 & -1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ -3 \\ 2 \end{pmatrix}$$

Next, let's examine the slightly more difficult problem of finding a vector  $x$  that satisfies  $Ux = b$ .

**Homework 1.1.1.2** Solve

$$\begin{pmatrix} 1 & -2 & 1 \\ 0 & -1 & -1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} \chi_0 \\ \chi_1 \\ \chi_2 \end{pmatrix} = \begin{pmatrix} -4 \\ -3 \\ 2 \end{pmatrix}$$

**Solution.** We can recognize the relation between this problem and [Homework 1.1.1.1](#) and hence deduce the answer without computation:

$$\begin{pmatrix} \chi_0 \\ \chi_1 \\ \chi_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$$

The point of these two homework exercises is that if one creates a (nonsingular)  $n \times n$  matrix  $U$  and vector  $x$  of size  $n$ , then computing  $b = Ux$  followed by solving  $U\hat{x} = b$  should leave us with a vector  $\hat{x}$  such that  $x = \hat{x}$ .

**Remark 1.1.1.1** We don't "teach" Matlab in this course. Instead, we think that Matlab is intuitive enough that we can figure out what the various commands mean. We can always investigate them by typing

`help <command>`

in the command window. For example, for this unit you may want to execute

```
help format
help rng
help rand
help triu
help *
help \
help diag
help abs
help min
help max
```

If you want to learn more about Matlab, you may want to take some of the tutorials offered by Mathworks at <https://www.mathworks.com/support/learn-with-matlab-tutorials.html>.

Let us see if Matlab can compute the solution of a triangular matrix correctly.

**Homework 1.1.1.3** In Matlab's command window, create a random upper triangular matrix  $U$ :

```
format long
rng( 0 );
n = 3
U = triu( rand( n,n ) )
x = rand( n,1 )
```

Report results in long format. Seed the random number generator so that we all create the same random matrix  $U$  and vector  $x$ .

<code>b = U * x;</code>	Compute right-hand side $b$ from known solution $x$ .
<code>xhat = U \ b;</code>	Solve $U\hat{x} = b$ .
<code>xhat - x</code>	Report the difference between $\hat{x}$ and $x$ .

What do we notice?

Next, check how close  $U\hat{x}$  is to  $b = Ux$ :

`b - U * xhat`

This is known as the residual.

What do we notice?

**Solution.** A script with the described commands can be found in [Assignments/Week01/matlab/Test\\_Upper\\_triangular\\_solve\\_3.m](#).

Some things we observe:

- $\hat{x} - x$  does not equal zero. This is due to the fact that the computer stores floating point numbers and computes with floating point arithmetic, and as a result roundoff error happens.
- The difference is small (notice the  $1.0e-15*$  before the vector, which shows that each entry in  $\hat{x} - x$  is around  $10^{-15}$ ).
- The residual  $b - U\hat{x}$  is small.
- Repeating this with a much larger  $n$  make things cumbersome since very long vectors are then printed.

To be able to compare more easily, we will compute the Euclidean length of  $\hat{x} - x$  instead using the Matlab command `norm( xhat - x )`. By adding a semicolon at the end of Matlab commands, we suppress output.

#### Homework 1.1.1.4 Execute

`format long`

```

rng( 0 );
n = 100;
U = triu( rand( n,n ) );
x = rand( n,1 );

b = U * x;

```

`xhat = U \ b;`

`norm( xhat - x )`

Report results in long format.

Seed the random number generator so that we all create the same random matrix  $U$  and vector  $x$ .

Compute right-hand side  $b$  from known solution  $x$ .

Solve  $U\hat{x} = b$

Report the Euclidean length of the difference between  $\hat{x}$  and  $x$ .

What do we notice?

Next, check how close  $U\hat{x}$  is to  $b = Ux$ , again using the Euclidean length:

`norm( b - U * xhat )`

What do we notice?

**Solution.** A script with the described commands can be found in [Assignments/Week01/matlab/Test\\_Upper\\_triangular\\_solve\\_100.m](#).

Some things we observe:

- $\text{norm}(\hat{x} - x)$ , the Euclidean length of  $\hat{x} - x$ , is huge. Matlab computed the wrong answer!
- However, the computed  $\hat{x}$  solves a problem that corresponds to a slightly different right-hand side. Thus,  $\hat{x}$  appears to be the solution to an only slightly changed problem.

The next exercise helps us gain insight into what is going on.

**Homework 1.1.1.5** Continuing with the  $U$ ,  $x$ ,  $b$ , and  $\hat{x}$  from [Homework 1.1.1.4](#), consider

- When is an upper triangular matrix singular?
- How large is the smallest element on the diagonal of the  $U$  from [Homework 1.1.1.4](#)? ( $\text{min}(\text{abs}(\text{diag}(U)))$  returns it!)
- If  $U$  were singular, how many solutions to  $U\hat{x} = b$  would there be? How can we characterize them?
- What is the relationship between  $\hat{x} - x$  and  $U$ ?

What have we learned?

**Solution.**

- When is an upper triangular matrix singular?

Answer:

If and only if there is a zero on its diagonal.

- How large is the smallest element on the diagonal of the  $U$  from [Homework 1.1.1.4](#)? ( $\text{min}(\text{abs}(\text{diag}(U)))$  returns it!)

Answer:

It is small in magnitude. This is not surprising, since it is a random number and hence as the matrix size increases, the chance of placing a small entry (in magnitude) on the diagonal increases.

- If  $U$  were singular, how many solutions to  $U\hat{x} = b$  would there be? How can we characterize them?

Answer:

An infinite number. Any vector in the null space can be added to a specific solution to create another solution.

- What is the relationship between  $\hat{x} - x$  and  $U$ ?

Answer:

It maps almost to the zero vector. In other words, it is close to a vector in the null space of the matrix  $U$  that has its smallest entry (in magnitude) on the diagonal changed to a zero.

What have we learned? The :"wrong" answer that Matlab computed was due to the fact that matrix  $U$  was almost singular.

To mathematically qualify and quantify all this, we need to be able to talk about "small" and "large" vectors, and "small" and "large" matrices. For that, we need to generalize the notion of length. By the end of this week, this will give us some of the tools to more fully understand what we have observed.



YouTube: <https://www.youtube.com/watch?v=2ZEtcnaynnM>

### 1.1.2 Overview

- 1.1 Opening
  - 1.1.1 Why norms?
  - 1.1.2 Overview
  - 1.1.3 What you will learn
- 1.2 Vector Norms
  - 1.2.1 Absolute value
  - 1.2.2 What is a vector norm?
  - 1.2.3 The vector 2-norm (Euclidean length)
  - 1.2.4 The vector p-norms
  - 1.2.5 Unit ball
  - 1.2.6 Equivalence of vector norms
- 1.3 Matrix Norms
  - 1.3.1 Of linear transformations and matrices
  - 1.3.2 What is a matrix norm?

- 1.3.3 The Frobenius norm
- 1.3.4 Induced matrix norms
- 1.3.5 The matrix 2-norm
- 1.3.6 Computing the matrix 1-norm and  $\infty$ -norm
- 1.3.7 Equivalence of matrix norms
- 1.3.8 Submultiplicative norms
- 1.3.9 Summary
- 1.4 Condition Number of a Matrix
  - 1.4.1 Conditioning of a linear system
  - 1.4.2 Loss of digits of accuracy
  - 1.4.3 The conditioning of an upper triangular matrix
- 1.5 Enrichments
  - 1.5.1 Condition number estimation
- 1.6 Wrap Up
  - 1.6.1 Additional homework
  - 1.6.2 Summary

### 1.1.3 What you will learn

Numerical analysis is the study of how the perturbation of a problem or data affects the accuracy of computation. This inherently means that you have to be able to measure whether changes are large or small. That, in turn, means we need to be able to quantify whether vectors or matrices are large or small. Norms are a tool for measuring magnitude.

Upon completion of this week, you should be able to

- Prove or disprove that a function is a norm.
- Connect linear transformations to matrices.
- Recognize, compute, and employ different measures of length, which differ and yet are equivalent.
- Exploit the benefits of examining vectors on the unit ball.
- Categorize different matrix norms based on their properties.
- Describe, in words and mathematically, how the condition number of a matrix affects how a relative change in the right-hand side can amplify into relative change in the solution of a linear system.
- Use norms to quantify the conditioning of solving linear systems.

## 1.2 Vector Norms

### 1.2.1 Absolute value

**Remark 1.2.1.1** Don't Panic!

In this course, we mostly allow scalars, vectors, and matrices to be complex-valued. This means we will use terms like "conjugate" and "Hermitian" quite liberally. You will think this is a big deal, but actually, if you just focus on the real case, you will notice that the complex case is just a natural extension of the real case.

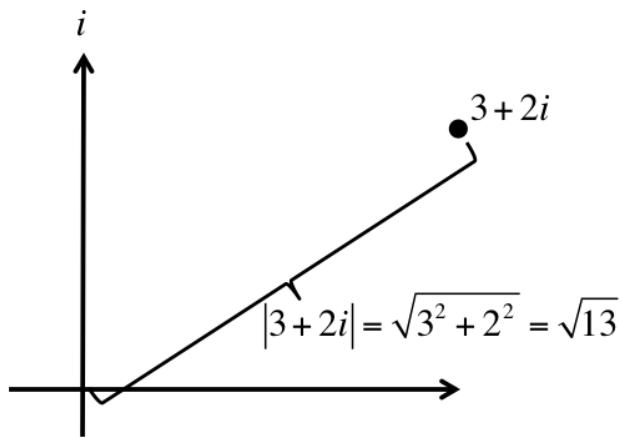


YouTube: <https://www.youtube.com/watch?v=V5ZQmR4zTeU>

Recall that  $|\cdot| : \mathbb{C} \rightarrow \mathbb{R}$  is the function that returns the absolute value of the input. In other words, if  $\alpha = \alpha_r + \alpha_c i$ , where  $\alpha_r$  and  $\alpha_c$  are the real and complex parts of  $\alpha$ , respectively, then

$$|\alpha| = \sqrt{\alpha_r^2 + \alpha_c^2}.$$

The absolute value (magnitude) of a complex number can also be thought of as the (Euclidean) distance from the point in the complex plane to the origin of that plane, as illustrated below for the number  $3 + 2i$ .



Alternatively, we can compute the absolute value as

$$\begin{aligned}
 |\alpha| &= \\
 &= \sqrt{\alpha_r^2 + \alpha_c^2} \\
 &= \sqrt{\alpha_r^2 - \alpha_c \alpha_r i + \alpha_r \alpha_c i + \alpha_c^2} \\
 &= \sqrt{(\alpha_r - \alpha_c i)(\alpha_r + \alpha_c i)} \\
 &= \sqrt{\bar{\alpha}\alpha} ,
 \end{aligned}$$

where  $\bar{\alpha}$  denotes the complex conjugate of  $\alpha$ :

$$\bar{\alpha} = \overline{\alpha_r + \alpha_c i} = \alpha_r - \alpha_c i.$$

The absolute value function has the following properties:

- $\alpha \neq 0 \Rightarrow |\alpha| > 0$  ( $|\cdot|$  is positive definite),
- $|\alpha\beta| = |\alpha||\beta|$  ( $|\cdot|$  is homogeneous), and
- $|\alpha + \beta| \leq |\alpha| + |\beta|$  ( $|\cdot|$  obeys the triangle inequality).

Norms are functions from a domain to the real numbers that are positive definite, homogeneous, and obey the triangle inequality. This makes the absolute value function an example of a norm.

The below exercises help refresh your fluency with complex arithmetic.

### Homework 1.2.1.1

$$1. (1+i)(2-i) =$$

$$2. (2-i)(1+i) =$$

$$3. \overline{(1-i)}(2-i) =$$

$$4. \overline{\overline{(1-i)}}(2-i) =$$

$$5. \overline{(2-i)}(1-i) =$$

$$6. (1-i)\overline{(2-i)} =$$

### Solution.

$$1. (1+i)(2-i) = 2 + 2i - i - i^2 = 2 + i + 1 = 3 + i$$

$$2. (2-i)(1+i) = 2 - i + 2i - i^2 = 2 + i + 1 = 3 + i$$

$$3. \overline{(1-i)}(2-i) = (1+i)(2-i) = 2 - i + 2i - i^2 = 3 + i$$

4.  $\overline{\overline{(1-i)}(2-i)} = \overline{(1+i)(2-i)} = \overline{2-i+2i-i^2} = \overline{2+i+1} = \overline{3+i} = 3-i$
5.  $\overline{(2-i)}(1-i) = (2+i)(1-i) = 2-2i+i-i^2 = 2-i+1 = 3-i$
6.  $(1-i)\overline{(2-i)} = (1-i)(2+i) = 2+i-2i-i^2 = 2-i+1 = 3-i$

**Homework 1.2.1.2** Let  $\alpha, \beta \in \mathbb{C}$ .

1. ALWAYS/SOMETIMES/NEVER:  $\alpha\beta = \beta\alpha$ .
2. ALWAYS/SOMETIMES/NEVER:  $\overline{\alpha}\beta = \overline{\beta}\alpha$ .

**Hint.** Let  $\alpha = \alpha_r + \alpha_c i$  and  $\beta = \beta_r + \beta_c i$ , where  $\alpha_r, \alpha_c, \beta_r, \beta_c \in \mathbb{R}$ .

**Answer.**

1. ALWAYS:  $\alpha\beta = \beta\alpha$ .
2. SOMETIMES:  $\overline{\alpha}\beta = \overline{\beta}\alpha$ .

**Solution.**

1. ALWAYS:  $\alpha\beta = \beta\alpha$ .

Proof:

$$\begin{aligned}
 \alpha\beta &= <\text{substitute}> \\
 (\alpha_r + \alpha_c i)(\beta_r + \beta_c i) &= <\text{multiply out}> \\
 \alpha_r\beta_r + \alpha_r\beta_c i + \alpha_c\beta_r i - \alpha_c\beta_c &= <\text{commutativity of real multiplication}> \\
 \beta_r\alpha_r + \beta_r\alpha_c i + \beta_c\alpha_r i - \beta_c\alpha_c &= <\text{factor}> \\
 (\beta_r + \beta_c i)(\alpha_r + \alpha_c i) &= <\text{substitute}> \\
 \beta\alpha.
 \end{aligned}$$

2. SOMETIMES:  $\overline{\alpha}\beta = \overline{\beta}\alpha$ .

An example where it is true:  $\alpha = \beta = 0$ .

An example where it is false:  $\alpha = 1$  and  $\beta = i$ . Then  $\overline{\alpha}\beta = 1 \times i = i$  and  $\overline{\beta}\alpha = -i \times 1 = -i$ .

**Homework 1.2.1.3** Let  $\alpha, \beta \in \mathbb{C}$ .

ALWAYS/SOMETIMES/NEVER:  $\overline{\alpha\beta} = \overline{\alpha}\beta$ .

**Hint.** Let  $\alpha = \alpha_r + \alpha_c i$  and  $\beta = \beta_r + \beta_c i$ , where  $\alpha_r, \alpha_c, \beta_r, \beta_c \in \mathbb{R}$ .

**Answer.** ALWAYS

Now prove it!

**Solution 1.**

$$\begin{aligned}
 & \overline{\alpha\beta} \\
 &= <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{(\alpha_r + \alpha_c i)(\beta_r + \beta_c i)} \\
 &= <\text{conjugate } \alpha> \\
 & \overline{(\alpha_r - \alpha_c i)(\beta_r + \beta_c i)} \\
 &= <\text{multiply out}> \\
 & \overline{(\alpha_r \beta_r - \alpha_c \beta_r i + \alpha_r \beta_c i + \alpha_c \beta_c)} \\
 &= <\text{conjugate}> \\
 & \alpha_r \beta_r + \alpha_c \beta_r i - \alpha_r \beta_c i + \alpha_c \beta_c \\
 &= <\text{rearrange}> \\
 & \beta_r \alpha_r + \beta_r \alpha_c i - \beta_c \alpha_r i + \beta_c \alpha_c \\
 &= <\text{factor}> \\
 & (\beta_r - \beta_c i)(\alpha_r + \alpha_c i) \\
 &= <\text{definition of conjugation}> \\
 & \overline{(\beta_r + \beta_c i)}(\alpha_r + \alpha_c i) \\
 &= <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{\beta\alpha}
 \end{aligned}$$

**Solution 2.** Proofs in mathematical textbooks seem to always be wonderfully smooth arguments that lead from the left-hand side of an equivalence to the right-hand side. In practice, you may want to start on the left-hand side, and apply a few rules:

$$\begin{aligned}
 & \overline{\alpha\beta} \\
 &= <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{(\alpha_r + \alpha_c i)(\beta_r + \beta_c i)} \\
 &= <\text{conjugate } \alpha> \\
 & \overline{(\alpha_r - \alpha_c i)(\beta_r + \beta_c i)} \\
 &= <\text{multiply out}> \\
 & \overline{(\alpha_r \beta_r - \alpha_c \beta_r i + \alpha_r \beta_c i + \alpha_c \beta_c)} \\
 &= <\text{conjugate}> \\
 & \alpha_r \beta_r + \alpha_c \beta_r i - \alpha_r \beta_c i + \alpha_c \beta_c
 \end{aligned}$$

and then move on to the right-hand side, applying a few rules:

$$\begin{aligned}
 & \overline{\beta\alpha} \\
 &= <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{(\beta_r + \beta_c i)}(\alpha_r + \alpha_c i) \\
 &= <\text{conjugate } \beta> \\
 & (\beta_r - \beta_c i)(\alpha_r + \alpha_c i) \\
 &= <\text{multiply out}> \\
 & \beta_r \alpha_r + \beta_r \alpha_c i - \beta_c \alpha_r i + \beta_c \alpha_c.
 \end{aligned}$$

At that point, you recognize that

$$\alpha_r \beta_r + \alpha_c \beta_r i - \alpha_r \beta_c i + \alpha_c \beta_c = \beta_r \alpha_r + \beta_r \alpha_c i - \beta_c \alpha_r i + \beta_c \alpha_c$$

since the second is a rearrangement of the terms of the first. Optionally, you then go back and presents these insights as a smooth argument that leads from the expression on the left-hand side to the one on the right-hand side:

$$\begin{aligned}
 & \overline{\alpha\beta} \\
 &= <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{(\alpha_r + \alpha_c i)(\beta_r + \beta_c i)} \\
 &= <\text{conjugate } \alpha> \\
 & \overline{(\alpha_r - \alpha_c i)(\beta_r + \beta_c i)} \\
 &= <\text{multiply out}> \\
 & \overline{(\alpha_r \beta_r - \alpha_c \beta_r i + \alpha_r \beta_c i + \alpha_c \beta_c)} \\
 &= <\text{conjugate}> \\
 & \alpha_r \beta_r + \alpha_c \beta_r i - \alpha_r \beta_c i + \alpha_c \beta_c \\
 &= <\text{rearrange}> \\
 & \beta_r \alpha_r + \beta_r \alpha_c i - \beta_c \alpha_r i + \beta_c \alpha_c \\
 &= <\text{factor}> \\
 & (\beta_r - \beta_c i)(\alpha_r + \alpha_c i) \\
 &= <\text{definition of conjugation}> \\
 & \overline{(\beta_r + \beta_c i)(\alpha_r + \alpha_c i)} \\
 &= <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{\beta}\alpha.
 \end{aligned}$$

**Solution 3.** Yet another way of presenting the proof uses an "equivalence style proof." The idea is to start with the equivalence you wish to prove correct:

$$\overline{\alpha\beta} = \overline{\beta}\alpha$$

and through a sequence of equivalent statements argue that this evaluates to TRUE:

$$\begin{aligned}
 & \overline{\alpha\beta} = \overline{\beta}\alpha \\
 & \Leftrightarrow <\alpha = \alpha_r + \alpha_c i; \beta = \beta_r + \beta_c i> \\
 & \overline{(\alpha_r + \alpha_c i)(\beta_r + \beta_c i)} = \overline{(\beta_r + \beta_c i)(\alpha_r + \alpha_c i)} \\
 & \Leftrightarrow <\text{conjugate } \times 2> \\
 & \overline{(\alpha_r - \alpha_c i)(\beta_r + \beta_c i)} = (\beta_r - \beta_c i)(\alpha_r + \alpha_c i) \\
 & \Leftrightarrow <\text{multiply out } \times 2> \\
 & \alpha_r \beta_r + \alpha_r \beta_c i - \alpha_c \beta_r i + \alpha_c \beta_c = \beta_r \alpha_r + \beta_r \alpha_c i - \beta_c \alpha_r i + \beta_c \alpha_c \\
 & \Leftrightarrow <\text{conjugate}> \\
 & \alpha_r \beta_r - \alpha_r \beta_c i + \alpha_c \beta_r i + \alpha_c \beta_c = \beta_r \alpha_r + \beta_r \alpha_c i - \beta_c \alpha_r i + \beta_c \alpha_c \\
 & \Leftrightarrow <\text{subtract equivalent terms from left-hand side and right-hand side}> \\
 & 0 = 0 \\
 & \Leftrightarrow <\text{algebra}> \\
 & \text{TRUE.}
 \end{aligned}$$

By transitivity of equivalence, we conclude that  $\overline{\alpha\beta} = \overline{\beta}\alpha$  is TRUE.

**Homework 1.2.1.4** Let  $\alpha \in \mathbb{C}$ .

ALWAYS/SOMETIMES/NEVER:  $\bar{\alpha}\alpha \in \mathbb{R}$

**Answer.** ALWAYS.

Now prove it!

**Solution.** Let  $\alpha = \alpha_r + \alpha_c i$ . Then

$$\begin{aligned} \bar{\alpha}\alpha &= <\text{ instantiate}> \\ (\alpha_r + \alpha_c i)(\alpha_r + \alpha_c i) &= <\text{conjugate}> \\ (\alpha_r - \alpha_c i)(\alpha_r + \alpha_c i) &= <\text{multiply out}> \\ \alpha_r^2 + \alpha_c^2, & \end{aligned}$$

which is a real number.

**Homework 1.2.1.5** Prove that the absolute value function is homogeneous:  $|\alpha\beta| = |\alpha||\beta|$  for all  $\alpha, \beta \in \mathbb{C}$ .

**Solution.**

$$\begin{aligned} |\alpha\beta| &= |\alpha||\beta| \\ \Leftrightarrow & <\text{squaring both sides simplifies}> \\ |\alpha\beta|^2 &= |\alpha|^2|\beta|^2 \\ \Leftrightarrow & <\text{ instantiate}> \\ |(\alpha_r + \alpha_c i)(\beta_r + \beta_c i)|^2 &= |\alpha_r + \alpha_c i|^2|\beta_r + \beta_c i|^2 \\ \Leftrightarrow & <\text{algebra}> \\ |(\alpha_r\beta_r - \alpha_c\beta_c) + (\alpha_r\beta_c + \alpha_c\beta_r)i|^2 &= (\alpha_r^2 + \alpha_c^2)(\beta_r^2 + \beta_c^2) \\ \Leftrightarrow & <\text{algebra}> \\ (\alpha_r\beta_r - \alpha_c\beta_c)^2 + (\alpha_r\beta_c + \alpha_c\beta_r)^2 &= (\alpha_r^2 + \alpha_c^2)(\beta_r^2 + \beta_c^2) \\ \Leftrightarrow & <\text{algebra}> \\ \alpha_r^2\beta_r^2 - 2\alpha_r\alpha_c\beta_r\beta_c + \alpha_c^2\beta_c^2 + \alpha_r^2\beta_c^2 + 2\alpha_r\alpha_c\beta_r\beta_c + \alpha_c^2\beta_r^2 &= \alpha_r^2\beta_r^2 + \alpha_r^2\beta_c^2 + \alpha_c^2\beta_r^2 + \alpha_c^2\beta_c^2 \\ &= \alpha_r^2\beta_r^2 + \alpha_r^2\beta_c^2 + \alpha_c^2\beta_r^2 + \alpha_c^2\beta_c^2 \\ \Leftrightarrow & <\text{subtract equivalent terms from both sides}> \\ 0 &= 0 \\ \Leftrightarrow & <\text{algebra}> \\ T & \end{aligned}$$

**Homework 1.2.1.6** Let  $\alpha \in \mathbb{C}$ .

ALWAYS/SOMETIMES/NEVER:  $|\bar{\alpha}| = |\alpha|$ .

**Answer.** ALWAYS

Now prove it!

**Solution.** Let  $\alpha = \alpha_r + \alpha_c i$ .

$$\begin{aligned}
 |\bar{\alpha}| &= <\text{ instantiate}> \\
 \overline{|\alpha_r + \alpha_c i|} &= <\text{conjugate}> \\
 |\alpha_r - \alpha_c i| &= <\text{definition of } |\cdot|> \\
 \sqrt{\alpha_r^2 + \alpha_c^2} &= <\text{definition of } |\cdot|> \\
 |\alpha_r + \alpha_c i| &= <\text{instantiate}> \\
 |\alpha|
 \end{aligned}$$

### 1.2.2 What is a vector norm?



YouTube: <https://www.youtube.com/watch?v=CTrUVfLGcNM>

A vector norm extends the notion of an absolute value to vectors. It allows us to measure the magnitude (or length) of a vector. In different situations, a different measure may be more appropriate.

**Definition 1.2.2.1 Vector norm.** Let  $\nu : \mathbb{C}^m \rightarrow \mathbb{R}$ . Then  $\nu$  is a (vector) norm if for all  $x, y \in \mathbb{C}^m$  and all  $\alpha \in \mathbb{C}$

- $x \neq 0 \Rightarrow \nu(x) > 0$  ( $\nu$  is positive definite),
- $\nu(\alpha x) = |\alpha| \nu(x)$  ( $\nu$  is homogeneous), and
- $\nu(x + y) \leq \nu(x) + \nu(y)$  ( $\nu$  obeys the triangle inequality).

◊

**Homework 1.2.2.1** TRUE/FALSE: If  $\nu : \mathbb{C}^m \rightarrow \mathbb{R}$  is a norm, then  $\nu(0) = 0$ .

**Hint.** From context, you should be able to tell which of these 0's denotes the zero vector of a given size and which is the scalar 0.

$0x = 0$  (multiplying any vector  $x$  by the scalar 0 results in a vector of zeroes).

**Answer.** TRUE.

Now prove it.

**Solution.** Let  $x \in \mathbb{C}^m$  and, just for clarity this first time,  $\vec{0}$  be the zero vector of size  $m$  so that 0 is the scalar zero. Then

$$\begin{aligned}\nu(\vec{0}) &= < 0 \cdot x = \vec{0} > \\ \nu(0 \cdot x) &= < \nu(\dots) \text{ is homogeneous} > \\ 0\nu(x) &= < \text{algebra} > \\ 0 &\end{aligned}$$

**Remark 1.2.2.2** We typically use  $\|\cdot\|$  instead of  $\nu(\cdot)$  for a function that is a norm.

### 1.2.3 The vector 2-norm (Euclidean length)



YouTube: <https://www.youtube.com/watch?v=bxDpUZEqBs>

The length of a vector is most commonly measured by the "square root of the sum of the squares of the elements," also known as the Euclidean norm. It is called the 2-norm because it is a member of a class of norms known as  $p$ -norms, discussed in the next unit.

**Definition 1.2.3.1 Vector 2-norm.** The vector 2-norm  $\|\cdot\|_2 : \mathbb{C}^m \rightarrow \mathbb{R}$  is defined for  $x \in \mathbb{C}^m$  by

$$\|x\|_2 = \sqrt{|\chi_0|^2 + \cdots + |\chi_{m-1}|^2} = \sqrt{\sum_{i=0}^{m-1} |\chi_i|^2}.$$

Equivalently, it can be defined by

$$\|x\|_2 = \sqrt{x^H x}$$

or

$$\|x\|_2 = \sqrt{\bar{\chi}_0 \chi_0 + \cdots + \bar{\chi}_{m-1} \chi_{m-1}} = \sqrt{\sum_{i=0}^{m-1} \bar{\chi}_i \chi_i}.$$

◇

**Remark 1.2.3.2** The notation  $x^H$  requires a bit of explanation. If

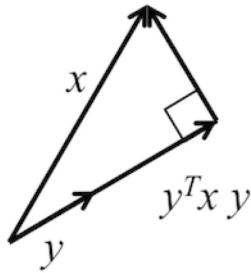
$$x = \begin{pmatrix} \chi_0 \\ \vdots \\ \chi_m \end{pmatrix}$$

then the row vector

$$x^H = \begin{pmatrix} \bar{x}_0 & \cdots & \bar{x}_m \end{pmatrix}$$

is the Hermitian transpose of  $x$  (or, equivalently, the Hermitian transpose of the vector  $x$  that is viewed as a matrix) and  $x^H y$  can be thought of as the dot product of  $x$  and  $y$  or, equivalently, as the matrix-vector multiplication of the matrix  $x^H$  times the vector  $y$ .

To prove that the 2-norm is a norm (just calling it a norm doesn't mean it is, after all), we need a result known as the Cauchy-Schwartz inequality. This inequality relates the magnitude of the dot product of two vectors to the product of their 2-norms: if  $x, y \in \mathbb{R}^m$ , then  $|x^T y| \leq \|x\|_2 \|y\|_2$ . To motivate this result before we rigorously prove it, recall from your undergraduate studies that the component of  $x$  in the direction of a vector  $y$  of unit length is given by  $(y^T x)y$ , as illustrated by



The length of the component of  $x$  in the direction of  $y$  then equals

$$\begin{aligned} & \|(y^T x)y\|_2 \\ &= <\text{definition}> \\ &= \sqrt{(y^T x)^T y^T (y^T x)y} \\ &= <z\alpha = \alpha z> \\ &= \sqrt{(x^T y)^2 y^T y} \\ &= <y \text{ has unit length}> \\ |y^T x| &= <\text{definition}> \\ &= |x^T y|. \end{aligned}$$

Thus  $|x^T y| \leq \|x\|_2$  (since a component should be shorter than the whole). If  $y$  is not of unit length (but a nonzero vector), then  $|x^T \frac{y}{\|y\|_2}| \leq \|x\|_2$  or, equivalently,  $|x^T y| \leq \|x\|_2 \|y\|_2$ .

We now state this result as a theorem, generalized to complex valued vectors:

**Theorem 1.2.3.3 Cauchy-Schwartz inequality.** *Let  $x, y \in \mathbb{C}^m$ . Then  $|x^H y| \leq \|x\|_2 \|y\|_2$ .*

*Proof.* Assume that  $x \neq 0$  and  $y \neq 0$ , since otherwise the inequality is trivially true. We can then choose  $\hat{x} = x/\|x\|_2$  and  $\hat{y} = y/\|y\|_2$ . This leaves us to prove that  $|\hat{x}^H \hat{y}| \leq 1$  since  $\|\hat{x}\|_2 = \|\hat{y}\|_2 = 1$ .

Pick

$$\alpha = \begin{cases} 1 & \text{if } x^H x = 0 \\ \hat{y}^H \hat{x} / |\hat{x}^H \hat{y}| & \text{otherwise.} \end{cases}$$

so that  $|\alpha| = 1$  and  $\alpha \hat{x}^H \hat{y}$  is real and nonnegative. Note that since it is real we also know

that

$$\begin{aligned} \alpha \hat{x}^H \hat{y} \\ = & \quad \langle \beta = \bar{\beta} \text{ if } \beta \text{ is real} \rangle \\ \frac{\alpha \hat{x}^H \hat{y}}{\alpha \hat{x}^H \hat{y}} &= \langle \text{property of complex conjugation} \rangle \\ &= \overline{\alpha \hat{y}^H \hat{x}} \end{aligned}$$

Now,

$$\begin{aligned} 0 &\leq \langle \|\cdot\|_2 \text{ is nonnegative definite} \rangle \\ \|\hat{x} - \alpha \hat{y}\|_2^2 &= \langle \|z\|_2^2 = z^H z \rangle \\ (\hat{x} - \alpha \hat{y})^H (\hat{x} - \alpha \hat{y}) &= \langle \text{multiplying out} \rangle \\ \hat{x}^H \hat{x} - \bar{\alpha} \hat{y}^H \hat{x} - \alpha \hat{x}^H \hat{y} + \bar{\alpha} \alpha \hat{y}^H \hat{y} &= \langle \text{above assumptions and observations} \rangle \\ 1 - 2\alpha \hat{x}^H \hat{y} + |\alpha|^2 &= \langle \alpha \hat{x}^H \hat{y} = |\hat{x}^H \hat{y}|; |\alpha| = 1 \rangle \\ 2 - 2|\hat{x}^H \hat{y}|. & \end{aligned}$$

Thus  $|\hat{x}^H \hat{y}| \leq 1$  and therefore  $|x^H y| \leq \|x\|_2 \|y\|_2$ . ■

The proof of [Theorem 1.2.3.3](#) does not employ any of the intuition we used to motivate it in the real valued case just before its statement. We leave it to the reader to prove the Cauchy-Schartz inequality for real-valued vectors by modifying (simplifying) the proof of [Theorem 1.2.3.3](#).

**Ponder This 1.2.3.1** Let  $x, y \in \mathbb{R}^m$ . Prove that  $|x^T y| \leq \|x\|_2 \|y\|_2$  by specializing the proof of [Theorem 1.2.3.3](#).

The following theorem states that the 2-norm is indeed a norm:

**Theorem 1.2.3.4** *The vector 2-norm is a norm.*

We leave its proof as an exercise.

**Homework 1.2.3.2** Prove [Theorem 1.2.3.4](#).

**Solution.** To prove this, we merely check whether the three conditions are met:

Let  $x, y \in \mathbb{C}^m$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $x \neq 0 \Rightarrow \|x\|_2 > 0$  ( $\|\cdot\|_2$  is positive definite):

Notice that  $x \neq 0$  means that at least one of its components is nonzero. Let's assume that  $\chi_j \neq 0$ . Then

$$\|x\|_2 = \sqrt{|\chi_0|^2 + \cdots + |\chi_{m-1}|^2} \geq \sqrt{|\chi_j|^2} = |\chi_j| > 0.$$

- $\|\alpha x\|_2 = |\alpha| \|x\|_2$  ( $\|\cdot\|_2$  is homogeneous):

$$\begin{aligned}
 \|\alpha x\|_2 &= \sqrt{|\alpha \chi_0|^2 + \dots + |\alpha \chi_{m-1}|^2} \\
 &= \sqrt{|\alpha|^2 |\chi_0|^2 + \dots + |\alpha|^2 |\chi_{m-1}|^2} \\
 &= \sqrt{|\alpha|^2 (|\chi_0|^2 + \dots + |\chi_{m-1}|^2)} \\
 &= |\alpha| \sqrt{|\chi_0|^2 + \dots + |\chi_{m-1}|^2} \\
 &= |\alpha| \|x\|_2.
 \end{aligned}$$

- $\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$  ( $\|\cdot\|_2$  obeys the triangle inequality):

$$\begin{aligned}
 \|x + y\|_2^2 &= \langle \|z\|_2^2 = z^H z \rangle \\
 (x + y)^H (x + y) &= \langle \text{distribute} \rangle \\
 x^H x + y^H x + x^H y + y^H y &= \langle \bar{\beta} + \beta = 2\text{Real}(\beta) \rangle \\
 x^H x + 2\text{Real}(x^H y) + y^H y &\leq \langle \text{algebra} \rangle \\
 x^H x + 2|\text{Real}(x^H y)| + y^H y &\leq \langle \text{algebra} \rangle \\
 x^H x + 2|x^H y| + y^H y &\leq \langle \text{algebra; Cauchy-Schwartz} \rangle \\
 \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 &= \langle \text{algebra} \rangle \\
 (\|x\|_2 + \|y\|_2)^2.
 \end{aligned}$$

Taking the square root (an increasing function that hence maintains the inequality) of both sides yields the desired result.

Throughout this course, we will reason about subvectors and submatrices. Let's get some practice:

**Homework 1.2.3.3** Partition  $x \in \mathbb{C}^m$  into subvectors:

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{pmatrix}.$$

ALWAYS/SOMETIMES/NEVER:  $\|x_i\|_2 \leq \|x\|_2$ .

**Answer.** ALWAYS

Now prove it!

**Solution.**

$$\begin{aligned}
 \|x\|_2^2 &= \left\| \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{pmatrix} \right\|_2^2 \\
 &= \left\| \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{pmatrix} \right\|_2^H \left\| \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{pmatrix} \right\|_2 \\
 &= \left\langle \text{dot product of partitioned vectors} \right\rangle \\
 x_0^H x_0 + x_1^H x_1 + \cdots + x_{M-1}^H x_{M-1} &= \left\langle \text{equivalent definition} \right\rangle \\
 \|x_0\|_2^2 + \|x_1\|_2^2 + \cdots + \|x_{M-1}\|_2^2 &\geq \left\langle \text{algebra} \right\rangle \\
 \|x_i\|_2^2
 \end{aligned}$$

so that  $\|x_i\|_2^2 \leq \|x\|_2^2$ . Taking the square root of both sides shows that  $\|x_i\|_2 \leq \|x\|_2$ .

#### 1.2.4 The vector $p$ -norms



YouTube: <https://www.youtube.com/watch?v=WGBMnmgJek8>

A vector norm is a measure of the magnitude of a vector. The Euclidean norm (length) is merely the best known such measure. There are others. A simple alternative is the 1-norm.

**Definition 1.2.4.1 Vector 1-norm.** The vector 1-norm,  $\|\cdot\|_1 : \mathbb{C}^m \rightarrow \mathbb{R}$ , is defined for  $x \in \mathbb{C}^m$  by

$$\|x\|_1 = |\chi_0| + |\chi_1| + \cdots + |\chi_{m-1}| = \sum_{i=0}^{m-1} |\chi_i|.$$

◇

**Homework 1.2.4.1** Prove that the vector 1-norm is a norm.

**Solution.** We show that the three conditions are met:

Let  $x, y \in \mathbb{C}^m$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $x \neq 0 \Rightarrow \|x\|_1 > 0$  ( $\|\cdot\|_1$  is positive definite):

Notice that  $x \neq 0$  means that at least one of its components is nonzero. Let's assume that  $\chi_j \neq 0$ . Then

$$\|x\|_1 = |\chi_0| + \dots + |\chi_{m-1}| \geq |\chi_j| > 0.$$

- $\|\alpha x\|_1 = |\alpha| \|x\|_1$  ( $\|\cdot\|_1$  is homogeneous):

$$\begin{aligned} \|\alpha x\|_1 &= < \text{scaling a vector-scales-its-components; definition} > \\ |\alpha \chi_0| + \dots + |\alpha \chi_{m-1}| &= < \text{algebra} > \\ |\alpha| |\chi_0| + \dots + |\alpha| |\chi_{m-1}| &= < \text{algebra} > \\ |\alpha| (|\chi_0| + \dots + |\chi_{m-1}|) &= < \text{definition} > \\ |\alpha| \|x\|_1. & \end{aligned}$$

- $\|x + y\|_1 \leq \|x\|_1 + \|y\|_1$  ( $\|\cdot\|_1$  obeys the triangle inequality):

$$\begin{aligned} \|x + y\|_1 &= < \text{vector addition; definition of 1-norm} > \\ |\chi_0 + \psi_0| + |\chi_1 + \psi_1| + \dots + |\chi_{m-1} + \psi_{m-1}| &\leq < \text{algebra} > \\ |\chi_0| + |\psi_0| + |\chi_1| + |\psi_1| + \dots + |\chi_{m-1}| + |\psi_{m-1}| &= < \text{commutivity} > \\ |\chi_0| + |\chi_1| + \dots + |\chi_{m-1}| + |\psi_0| + |\psi_1| + \dots + |\psi_{m-1}| &= < \text{associativity; definition} > \\ \|x\|_1 + \|y\|_1. & \end{aligned}$$

The vector 1-norm is sometimes referred to as the "taxi-cab norm". It is the distance that a taxi travels, from one point on a street to another such point, along the streets of a city that has square city blocks.

Another alternative is the infinity norm.

**Definition 1.2.4.2 Vector  $\infty$ -norm.** The vector  $\infty$ -norm,  $\|\cdot\|_\infty : \mathbb{C}^m \rightarrow \mathbb{R}$ , is defined for  $x \in \mathbb{C}^m$  by

$$\|x\|_\infty = \max(|\chi_0|, \dots, |\chi_{m-1}|) = \max_{i=0}^{m-1} |\chi_i|.$$

◊

The infinity norm simply measures how large the vector is by the magnitude of its largest entry.

**Homework 1.2.4.2** Prove that the vector  $\infty$ -norm is a norm.

**Solution.** We show that the three conditions are met:

Let  $x, y \in \mathbb{C}^m$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $x \neq 0 \Rightarrow \|x\|_\infty > 0$  ( $\|\cdot\|_\infty$  is positive definite):

Notice that  $x \neq 0$  means that at least one of its components is nonzero. Let's assume that  $\chi_j \neq 0$ . Then

$$\|x\|_\infty = \max_{i=0}^{m-1} |\chi_i| \geq |\chi_j| > 0.$$

- $\|\alpha x\|_\infty = |\alpha| \|x\|_\infty$  ( $\|\cdot\|_\infty$  is homogeneous):

$$\begin{aligned} \|\alpha x\|_\infty &= \max_{i=0}^{m-1} |\alpha \chi_i| \\ &= \max_{i=0}^{m-1} |\alpha| |\chi_i| \\ &= |\alpha| \max_{i=0}^{m-1} |\chi_i| \\ &= |\alpha| \|x\|_\infty. \end{aligned}$$

- $\|x + y\|_\infty \leq \|x\|_\infty + \|y\|_\infty$  ( $\|\cdot\|_\infty$  obeys the triangle inequality):

$$\begin{aligned} \|x + y\|_\infty &= \max_{i=0}^{m-1} |\chi_i + \psi_i| \\ &\leq \max_{i=0}^{m-1} (|\chi_i| + |\psi_i|) \\ &\leq \max_{i=0}^{m-1} |\chi_i| + \max_{i=0}^{m-1} |\psi_i| \\ &= \|x\|_\infty + \|y\|_\infty. \end{aligned}$$

In this course, we will primarily use the vector 1-norm, 2-norm, and  $\infty$ -norms. For completeness, we briefly discuss their generalization: the vector  $p$ -norm.

**Definition 1.2.4.3 Vector  $p$ -norm.** Given  $p \geq 1$ , the vector  $p$ -norm  $\|\cdot\|_p : \mathbb{C}^m \rightarrow \mathbb{R}$  is defined for  $x \in \mathbb{C}^m$  by

$$\|x\|_p = \sqrt[p]{|\chi_0|^p + \cdots + |\chi_{m-1}|^p} = \left( \sum_{i=0}^{m-1} |\chi_i|^p \right)^{1/p}.$$

◊

**Theorem 1.2.4.4** *The vector  $p$ -norm is a norm.*

The proof of this result is very similar to the proof of the fact that the 2-norm is a norm. It depends on Hölder's inequality, which is a generalization of the Cauchy-Schwartz inequality:

**Theorem 1.2.4.5 Hölder's inequality.** *Let  $1 \leq p, q \leq \infty$  with  $\frac{1}{p} + \frac{1}{q} = 1$ . If  $x, y \in \mathbb{C}^m$  then  $|x^H y| \leq \|x\|_p \|y\|_q$ .*

We skip the proof of Hölder's inequality and [Theorem 1.2.4.4](#). You can easily find proofs for these results, should you be interested.

**Remark 1.2.4.6** The vector 1-norm and 2-norm are obviously special cases of the vector  $p$ -norm. It can be easily shown that the vector  $\infty$ -norm is also related:

$$\lim_{p \rightarrow \infty} \|x\|_p = \|x\|_\infty.$$

**Ponder This 1.2.4.3** Consider [Homework 1.2.3.3](#). Try to elegantly formulate this question in the most general way you can think of. How do you prove the result?

**Ponder This 1.2.4.4** Consider the vector norm  $\|\cdot\| : \mathbb{C}^m \rightarrow \mathbb{R}$ , the matrix  $A \in \mathbb{C}^{m \times n}$  and the function  $f : \mathbb{C}^n \rightarrow \mathbb{R}$  defined by  $f(x) = \|Ax\|$ . For what matrices  $A$  is the function  $f$  a norm?

## 1.2.5 Unit ball



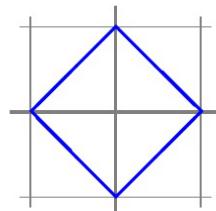
YouTube: <https://www.youtube.com/watch?v=aJgrpp7uscw>

In 3-dimensional space, the notion of the unit ball is intuitive: the set of all points that are a (Euclidean) distance of one from the origin. Vectors have no position and can have more than three components. Still the unit ball for the 2-norm is a straight forward extension to the set of all vectors with length (2-norm) one. More generally, the unit ball for any norm can be defined:

**Definition 1.2.5.1 Unit ball.** Given norm  $\|\cdot\| : \mathbb{C}^m \rightarrow \mathbb{R}$ , the unit ball with respect to  $\|\cdot\|$  is the set  $\{x \mid \|x\| = 1\}$  (the set of all vectors with norm equal to one). We will use  $\|x\| = 1$  as shorthand for  $\{x \mid \|x\| = 1\}$ .  $\diamond$

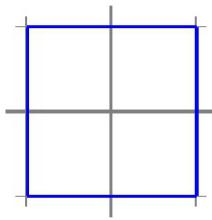
**Homework 1.2.5.1** Although vectors have no position, it is convenient to visualize a vector  $x \in \mathbb{R}^2$  by the point in the plane to which it extends when rooted at the origin. For example, the vector  $x = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$  can be so visualized with the point  $(2, 1)$ . With this in mind, match the pictures on the right corresponding to the sets on the left:

(a)  $\|x\|_2 = 1$ . (1)



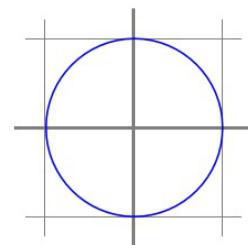
(b)  $\|x\|_1 = 1$ .

(2)



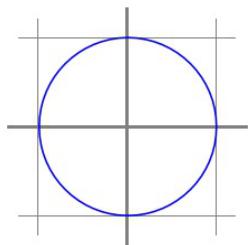
(c)  $\|x\|_\infty = 1$ .

(3)

**Solution.**

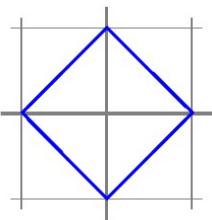
(a)  $\|x\|_2 = 1$ .

(3)



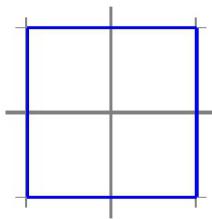
(b)  $\|x\|_1 = 1$ .

(1)



(c)  $\|x\|_\infty = 1$ .

(2)





YouTube: <https://www.youtube.com/watch?v=0v77sE90P58>

### 1.2.6 Equivalence of vector norms



YouTube: <https://www.youtube.com/watch?v=qjZyKHvL13E>

**Homework 1.2.6.1** Fill out the following table:

$x$	$\ x\ _1$	$\ x\ _\infty$	$\ x\ _2$
$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$			
$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$			
$\begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$			

**Solution.**

$x$	$\ x\ _1$	$\ x\ _\infty$	$\ x\ _2$
$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$	1	1	1
$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$	3	1	$\sqrt{3}$
$\begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$	4	2	$\sqrt{1^2 + (-2)^2 + (-1)^2} = \sqrt{6}$

In this course, norms are going to be used to reason that vectors are "small" or "large". It would be unfortunate if a vector were small in one norm yet large in another norm. Fortunately, the following theorem excludes this possibility:

**Theorem 1.2.6.1 Equivalence of vector norms.** *Let  $\|\cdot\| : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\|\cdot\||| : \mathbb{C}^m \rightarrow \mathbb{R}$  both be vector norms. Then there exist positive scalars  $\sigma$  and  $\tau$  such that for all  $x \in \mathbb{C}^m$*

$$\sigma\|x\| \leq \|\|x\||| \leq \tau\|x\|.$$

*Proof.* The proof depends on a result from real analysis (sometimes called "advanced calculus") that states that  $\sup_{x \in S} f(x)$  is attained for some vector  $x \in S$  as long as  $f$  is continuous and  $S$  is a compact (closed and bounded) set. For any norm  $\|\cdot\|$ , the unit ball  $\|x\| = 1$  is a compact set. When a supremum is an element in  $S$ , it is called the maximum instead and  $\sup_{x \in S} f(x)$  can be restated as  $\max_{x \in S} f(x)$ .

Those who have not studied real analysis (which is not a prerequisite for this course) have to take this on faith. It is a result that we will use a few times in our discussion.

We prove that there exists a  $\tau$  such that for all  $x \in \mathbb{C}^m$

$$\|\|x\||| \leq \tau\|x\|,$$

leaving the rest of the proof as an exercise.

Let  $x \in \mathbb{C}^m$  be an arbitrary vector. W.l.o.g. assume that  $x \neq 0$ . Then

$$\begin{aligned} \|\|x\||| &= <\text{algebra}> \\ \frac{\|\|x\|||}{\|x\|} \|x\| &\leq <\text{algebra}> \\ \left(\sup_{z \neq 0} \frac{\|\|z\|||}{\|z\|}\right) \|x\| &= <\text{change of variables: } y = z/\|z\|> \\ \left(\sup_{\|y\|=1} \|\|y\|||\right) \|x\| &= <\text{the set } \|y\|=1 \text{ is compact}> \\ \left(\max_{\|y\|=1} \|\|y\|||\right) \|x\| \end{aligned}$$

The desired  $\tau$  can now be chosen to equal  $\max_{\|y\|=1} \|\|y\|||$ . ■



YouTube: <https://www.youtube.com/watch?v=I1W6ErdEyoc>

**Homework 1.2.6.2** Complete the proof of Theorem 1.2.6.1.

**Solution.** We need to prove that

$$\sigma\|x\| \leq |||x|||.$$

From the first part of the proof of Theorem 1.2.6.1, we know that there exists a  $\rho > 0$  such that

$$\|x\| \leq \rho|||x|||$$

and hence

$$\frac{1}{\rho}\|x\| \leq |||x|||.$$

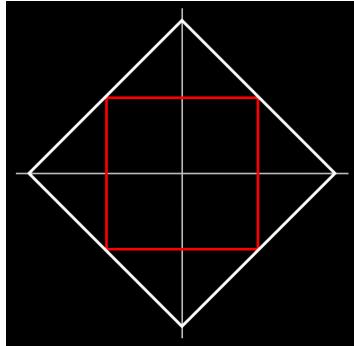
We conclude that

$$\sigma\|x\| \leq |||x|||$$

where  $\sigma = 1/\rho$ .

**Example 1.2.6.2**

- Let  $x \in \mathbb{R}^2$ . Use the picture

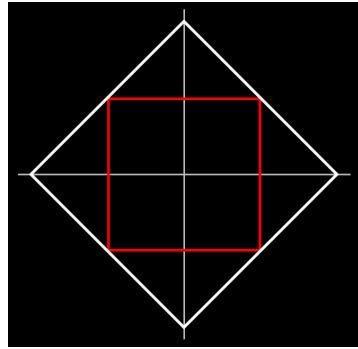


to determine the constant  $C$  such that  $\|x\|_1 \leq C\|x\|_\infty$ . Give a vector  $x$  for which  $\|x\|_1 = C\|x\|_\infty$ .

- For  $x \in \mathbb{R}^2$  and the  $C$  you determined in the first part of this problem, prove that  $\|x\|_1 \leq C\|x\|_\infty$ .
- Let  $x \in \mathbb{C}^m$ . Extrapolate from the last part the constant  $C$  such that  $\|x\|_1 \leq C\|x\|_\infty$  and then prove the inequality. Give a vector  $x$  for which  $\|x\|_1 = C\|x\|_\infty$ .

**Solution.**

- Consider the picture



- The red square represents all vectors such that  $\|x\|_\infty = 1$  and the white square represents all vectors such that  $\|x\|_1 = 2$ .
- All points on or outside the red square represent vectors  $y$  such that  $\|y\|_\infty \geq 1$ . Hence if  $\|y\|_1 = 2$  then  $\|y\|_\infty \geq 1$ .
- Now, pick any  $z \neq 0$ . Then  $\|2z/\|z\|_1\|_1 = 2$ . Hence

$$\|2z/\|z\|_1\|_\infty \geq 1$$

which can be rewritten as

$$\|z\|_1 \leq 2\|z\|_\infty.$$

Thus,  $C = 2$  works.

- Now, from the picture it is clear that  $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  has the property that  $\|x\|_1 = 2\|x\|_\infty$ . Thus, the inequality is "tight."
- We now prove that  $\|x\|_1 \leq 2\|x\|_\infty$  for  $x \in \mathbb{R}^2$ :

$$\begin{aligned}
 \|x\|_1 &= <\text{definition}> \\
 |\chi_0| + |\chi_1| &\leq <\text{algebra}> \\
 \max(|\chi_0|, |\chi_1|) + \max(|\chi_0|, |\chi_1|) &= <\text{algebra}> \\
 2 \max(|\chi_0|, |\chi_1|) &= <\text{definition}> \\
 2\|x\|_\infty.
 \end{aligned}$$

- From the last part we extrapolate that  $\|x\|_1 \leq m\|x\|_\infty$ .

$$\begin{aligned}
\|x\|_1 &= \sum_{i=0}^{m-1} |\chi_i| < \text{definition} > \\
&\leq \sum_{i=0}^{m-1} (\max_{j=0}^{m-1} |\chi_j|) < \text{algebra} > \\
&= m \max_{j=0}^{m-1} |\chi_j| < \text{algebra} > \\
&= m \|x\|_\infty. < \text{definition} >
\end{aligned}$$

Equality holds (i.e.,  $\|x\|_1 = m\|x\|_\infty$ ) for  $x = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ .

Some will be able to go straight for the general result, while others will want to seek inspiration from the picture and/or the specialized case where  $x \in \mathbb{R}^2$ .  $\square$

**Homework 1.2.6.3** Let  $x \in \mathbb{C}^m$ . The following table organizes the various bounds:

	$\ x\ _1 \leq C_{1,2}\ x\ _2$	$\ x\ _1 \leq C_{1,\infty}\ x\ _\infty$
$\ x\ _2 \leq C_{2,1}\ x\ _1$		$\ x\ _2 \leq C_{2,\infty}\ x\ _\infty$
$\ x\ _\infty \leq C_{\infty,1}\ x\ _1$	$\ x\ _\infty \leq C_{\infty,2}\ x\ _2$	

For each, determine the constant  $C_{x,y}$  and prove the inequality, including that it is a tight inequality.

Hint: look at the hint!

**Hint.**  $\|x\|_1 \leq \sqrt{m}\|x\|_2$ :

This is the hardest one to prove. Do it last and use the following hint:

Consider  $y = \begin{pmatrix} \chi_0/|\chi_0| \\ \vdots \\ \chi_{m-1}/|\chi_{m-1}| \end{pmatrix}$  and employ the Cauchy-Schwartz inequality.

**Solution 1** ( $\|x\|_1 \leq C_{1,2}\|x\|_2$ ).  $\|x\|_1 \leq \sqrt{m}\|x\|_2$ :

Consider  $y = \begin{pmatrix} \chi_0/|\chi_0| \\ \vdots \\ \chi_{m-1}/|\chi_{m-1}| \end{pmatrix}$ . Then

$$|x^H y| = \left| \sum_{i=0}^{m-1} \overline{\chi_i} \chi_i / |\chi_i| \right| = \left| \sum_{i=0}^{m-1} |\chi_i|^2 / |\chi_i| \right| = \left| \sum_{i=0}^{m-1} |\chi_i| \right| = \|x\|_1.$$

We also notice that  $\|y\|_2 = \sqrt{m}$ .

From the Cauchy-Swartz inequality we know that

$$\|x\|_1 = |x^H y| \leq \|x\|_2 \|y\|_2 = \sqrt{m} \|x\|_2.$$

If we now choose

$$x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

then  $\|x\|_1 = m$  and  $\|x\|_2 = \sqrt{m}$  so that  $\|x\|_1 = \sqrt{m}\|x\|_2$ .

**Solution 2** ( $\|x\|_1 \leq C_{1,\infty}\|x\|_\infty$ ).  $\|x\|_1 \leq m\|x\|_\infty$ :

See [Example 1.2.6.2](#).

**Solution 3** ( $\|x\|_2 \leq C_{2,1}\|x\|_1$ ).  $\|x\|_2 \leq \|x\|_1$ :

$$\begin{aligned} \|x\|_2^2 &= <\text{definition}> \\ \sum_{i=0}^{m-1} |\chi_i|^2 &\leq <\text{algebra}> \\ \left(\sum_{i=0}^{m-1} |\chi_i|\right)^2 &= <\text{definition}> \\ \|x\|_1^2. \end{aligned}$$

Taking the square root of both sides yields  $\|x\|_2 \leq \|x\|_1$ .

If we now choose

$$x = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

then  $\|x\|_2 = \|x\|_1$ .

**Solution 4** ( $\|x\|_2 \leq C_{2,\infty}\|x\|_\infty$ ).  $\|x\|_2 \leq \sqrt{m}\|x\|_\infty$ :

$$\begin{aligned} \|x\|_2^2 &= <\text{definition}> \\ \sum_{i=0}^{m-1} |\chi_i|^2 &\leq <\text{algebra}> \\ \sum_{i=0}^{m-1} \left(\max_{j=0}^{m-1} |\chi_j|\right)^2 &= <\text{definition}> \\ \sum_{i=0}^{m-1} \|x\|_\infty^2 &= <\text{algebra}> \\ m\|x\|_\infty^2. \end{aligned}$$

Taking the square root of both sides yields  $\|x\|_2 \leq \sqrt{m}\|x\|_\infty$ .

Consider

$$x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

then  $\|x\|_2 = \sqrt{m}$  and  $\|x\|_\infty = 1$  so that  $\|x\|_2 = \sqrt{m}\|x\|_\infty$ .

**Solution 5** ( $\|x\|_\infty \leq C_{\infty,1}\|x\|_1$ ):  $\|x\|_\infty \leq \|x\|_1$ :

$$\begin{aligned} & \|x\|_\infty \\ &= < \text{definition} > \\ & \max_{i=0}^{m-1} |\chi_i| \\ &\leq < \text{algebra} > \\ & \sum_{i=0}^{m-1} |\chi_i| \\ &= < \text{definition} > \\ & \|x\|_1. \end{aligned}$$

Consider

$$x = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Then  $\|x\|_\infty = 1 = \|x\|_1$ .

**Solution 6** ( $\|x\|_\infty \leq C_{\infty,2}\|x\|_2$ ):  $\|x\|_\infty \leq \|x\|_2$ :

$$\begin{aligned} & \|x\|_\infty^2 \\ &= < \text{definition} > \\ & \left( \max_{i=0}^{m-1} |\chi_i| \right)^2 \\ &= < \text{algebra} > \\ & \max_{i=0}^{m-1} |\chi_i|^2 \\ &\leq < \text{algebra} > \\ & \sum_{i=0}^{m-1} |\chi_i|^2 \\ &= < \text{definition} > \\ & \|x\|_2^2. \end{aligned}$$

Taking the square root of both sides yields  $\|x\|_\infty \leq \|x\|_2$ .

Consider

$$x = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Then  $\|x\|_\infty = 1 = \|x\|_2$ .

**Solution 7** (Table of constants).

	$\ x\ _1 \leq \sqrt{m}\ x\ _2$	$\ x\ _1 \leq m\ x\ _\infty$
$\ x\ _2 \leq \ x\ _1$		$\ x\ _2 \leq \sqrt{m}\ x\ _\infty$
$\ x\ _\infty \leq \ x\ _1$	$\ x\ _\infty \leq \ x\ _2$	

**Remark 1.2.6.3** The bottom line is that, modulo a constant factor, if a vector is "small" in one norm, it is "small" in all other norms. If it is "large" in one norm, it is "large" in all other norms.

## 1.3 Matrix Norms

### 1.3.1 Of linear transformations and matrices



YouTube: <https://www.youtube.com/watch?v=x1kiZEbYh38>

We briefly review the relationship between linear transformations and matrices, which is key to understanding why linear algebra is all about matrices and vectors.

**Definition 1.3.1.1 Linear transformations and matrices.** Let  $L : \mathbb{C}^n \rightarrow \mathbb{C}^m$ . Then  $L$  is said to be a linear transformation if for all  $\alpha \in \mathbb{C}$  and  $x, y \in \mathbb{C}^n$

- $L(\alpha x) = \alpha L(x)$ . That is, scaling first and then transforming yields the same result as transforming first and then scaling.
- $L(x + y) = L(x) + L(y)$ . That is, adding first and then transforming yields the same result as transforming first and then adding.



The importance of linear transformations comes in part from the fact that many problems in science boil down to, given a function  $F : \mathbb{C}^n \rightarrow \mathbb{C}^m$  and vector  $y \in \mathbb{C}^m$ , find  $x$  such that  $F(x) = y$ . This is known as an inverse problem. Under mild conditions,  $F$  can be locally approximated with a linear transformation  $L$  and then, as part of a solution method, one would want to solve  $Lx = y$ .

The following theorem provides the link between linear transformations and matrices:

**Theorem 1.3.1.2** *Let  $L : \mathbb{C}^n \rightarrow \mathbb{C}^m$  be a linear transformation,  $v_0, v_1, \dots, v_{k-1} \in \mathbb{C}^n$ , and  $x \in \mathbb{C}^k$ . Then*

$$L(\chi_0 v_0 + \chi_1 v_1 + \dots + \chi_{k-1} v_{k-1}) = \chi_0 L(v_0) + \chi_1 L(v_1) + \dots + \chi_{k-1} L(v_{k-1}),$$

where

$$x = \begin{pmatrix} \chi_0 \\ \vdots \\ \chi_{k-1} \end{pmatrix}.$$

*Proof.* A simple inductive proof yields the result. For details, see Week 2 of Linear Algebra: Foundations to Frontiers (LAFF) [20]. ■

The following set of vectors ends up playing a crucial role throughout this course:

**Definition 1.3.1.3 Standard basis vector.** In this course, we will use  $e_j \in \mathbb{C}^m$  to denote the standard basis vector with a "1" in the position indexed with  $j$ . So,

$$e_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j$$

◇

Key is the fact that any vector  $x \in \mathbb{C}^n$  can be written as a linear combination of the standard basis vectors of  $\mathbb{C}^n$ :

$$\begin{aligned} x &= \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} = \chi_0 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \chi_1 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \dots + \chi_{n-1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \\ &= \chi_0 e_0 + \chi_1 e_1 + \dots + \chi_{n-1} e_{n-1}. \end{aligned}$$

Hence, if  $L$  is a linear transformation,

$$\begin{aligned} L(x) &= L(\chi_0 e_0 + \chi_1 e_1 + \dots + \chi_{n-1} e_{n-1}) \\ &= \chi_0 \underbrace{L(e_0)}_{a_0} + \chi_1 \underbrace{L(e_1)}_{a_1} + \dots + \chi_{n-1} \underbrace{L(e_{n-1})}_{a_{n-1}}. \end{aligned}$$

If we now let  $a_j = L(e_j)$  (the vector  $a_j$  is the transformation of the standard basis vector  $e_j$ ) and collect these vectors into a two-dimensional array of numbers:

$$A = \left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right) \quad (1.3.1)$$

then we notice that information for evaluating  $L(x)$  can be found in this array, since  $L$  can then alternatively be computed by

$$L(x) = \chi_0 a_0 + \chi_1 a_1 + \cdots + \chi_{n-1} a_{n-1}.$$

The array  $A$  in (1.3.1) we call a **matrix** and the operation  $Ax = \chi_0 a_0 + \chi_1 a_1 + \cdots + \chi_{n-1} a_{n-1}$  we call **matrix-vector multiplication**. Clearly

$$Ax = L(x).$$

**Remark 1.3.1.4 Notation.** In these notes, as a rule,

- Roman upper case letters are used to denote matrices.
- Roman lower case letters are used to denote vectors.
- Greek lower case letters are used to denote scalars.

Corresponding letters from these three sets are used to refer to a matrix, the row or columns of that matrix, and the elements of that matrix. If  $A \in \mathbb{C}^{m \times n}$  then

$$\begin{aligned} A &= <\text{partition } A \text{ by columns and rows}> \\ \left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right) &= \begin{pmatrix} \tilde{a}_0^T \\ \tilde{a}_1^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix} \\ &= <\text{expose the elements of } A> \\ &\begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \hline \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \hline \vdots & \vdots & & \vdots \\ \hline \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix} \end{aligned}$$

We now notice that the standard basis vector  $e_j \in \mathbb{C}^m$  equals the column of the  $m \times m$  **identity matrix** indexed with  $j$ :

$$I = \left( \begin{array}{c|c|c|c} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{array} \right) = \left( \begin{array}{c|c|c|c} e_0 & e_1 & \cdots & e_{m-1} \end{array} \right) = \begin{pmatrix} \tilde{e}_0^T \\ \tilde{e}_1^T \\ \vdots \\ \tilde{e}_{m-1}^T \end{pmatrix}.$$

**Remark 1.3.1.5** The important thing to note is that a matrix is a convenient representation of a linear transformation and matrix-vector multiplication is an alternative way for evaluating that linear transformation.



YouTube: <https://www.youtube.com/watch?v=cCFAnQmwwIw>

Let's investigate matrix-matrix multiplication and its relationship to linear transformations. Consider two linear transformations

$$\begin{aligned} L_A : \mathbb{C}^k &\rightarrow \mathbb{C}^m \quad \text{represented by matrix } A \\ L_B : \mathbb{C}^n &\rightarrow \mathbb{C}^k \quad \text{represented by matrix } B \end{aligned}$$

and define

$$L_C(x) = L_A(L_B(x)),$$

as the composition of  $L_A$  and  $L_B$ . Then it can be easily shown that  $L_C$  is also a linear transformation. Let  $m \times n$  matrix  $C$  represent  $L_C$ . How are  $A$ ,  $B$ , and  $C$  related? If we let  $c_j$  equal the column of  $C$  indexed with  $j$ , then because of the link between matrices, linear transformations, and standard basis vectors

$$c_j = L_C(e_j) = L_A(L_B(e_j)) = L_A(b_j) = Ab_j,$$

where  $b_j$  equals the column of  $B$  indexed with  $j$ . Now, we say that  $C = AB$  is the product of  $A$  and  $B$  defined by

$$\left( \begin{array}{c|c|c|c} c_0 & c_1 & \cdots & c_{n-1} \end{array} \right) = A \left( \begin{array}{c|c|c|c} b_0 & b_1 & \cdots & b_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c|c} Ab_0 & Ab_1 & \cdots & Ab_{n-1} \end{array} \right)$$

and define the matrix-matrix multiplication as the operation that computes

$$C := AB,$$

which you will want to pronounce "C becomes A times B" to distinguish assignment from equality. If you think carefully how individual elements of  $C$  are computed, you will realize that they equal the usual "dot product of rows of  $A$  with columns of  $B$ ."



YouTube: [https://www.youtube.com/watch?v=g\\_9RbA5EOIc](https://www.youtube.com/watch?v=g_9RbA5EOIc)

As already mentioned, throughout this course, it will be important that you can think about matrices in terms of their columns and rows, and matrix-matrix multiplication (and other operations with matrices and vectors) in terms of columns and rows. It is also important to be able to think about matrix-matrix multiplication in three different ways. If we partition each matrix by rows and by columns:

$$C = \left( \begin{array}{c|c|c} c_0 & \cdots & c_{n-1} \end{array} \right) = \left( \begin{array}{c} \tilde{c}_0^T \\ \vdots \\ \tilde{c}_{m-1}^T \end{array} \right), A = \left( \begin{array}{c|c|c} a_0 & \cdots & a_{k-1} \end{array} \right) = \left( \begin{array}{c} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{array} \right),$$

and

$$B = \left( \begin{array}{c|c|c} b_0 & \cdots & b_{n-1} \end{array} \right) = \left( \begin{array}{c} \tilde{b}_0^T \\ \vdots \\ \tilde{b}_{k-1}^T \end{array} \right),$$

then  $C := AB$  can be computed in the following ways:

1. By columns:

$$\left( \begin{array}{c|c|c} c_0 & \cdots & c_{n-1} \end{array} \right) := A \left( \begin{array}{c|c|c} b_0 & \cdots & b_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c} Ab_0 & \cdots & Ab_{n-1} \end{array} \right).$$

In other words,  $c_j := Ab_j$  for all columns of  $C$ .

2. By rows:

$$\left( \begin{array}{c} \tilde{c}_0^T \\ \vdots \\ \tilde{c}_{m-1}^T \end{array} \right) := \left( \begin{array}{c} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{array} \right) B = \left( \begin{array}{c} \tilde{a}_0^T B \\ \vdots \\ \tilde{a}_{m-1}^T B \end{array} \right).$$

In other words,  $\tilde{c}_i^T = \tilde{a}_i^T B$  for all rows of  $C$ .

3. One you may not have thought about much before:

$$C := \left( \begin{array}{c|c|c} a_0 & \cdots & a_{k-1} \end{array} \right) \left( \begin{array}{c} \tilde{b}_0^T \\ \vdots \\ \tilde{b}_{k-1}^T \end{array} \right) = a_0 \tilde{b}_0^T + \cdots + a_{k-1} \tilde{b}_{k-1}^T,$$

which should be thought of as a sequence of rank-1 updates, since each term is an outer product and an outer product has rank of at most one.

These three cases are special cases of the more general observation that, if we can partition  $C$ ,  $A$ , and  $B$  by blocks (submatrices),

$$C = \left( \begin{array}{c|c|c} C_{0,0} & \cdots & C_{0,N-1} \\ \hline \vdots & & \vdots \\ \hline C_{M-1,0} & \cdots & C_{M-1,N-1} \end{array} \right), \left( \begin{array}{c|c|c} A_{0,0} & \cdots & A_{0,K-1} \\ \hline \vdots & & \vdots \\ \hline A_{M-1,0} & \cdots & A_{M-1,K-1} \end{array} \right),$$

and

$$\left( \begin{array}{c|c|c} B_{0,0} & \cdots & B_{0,N-1} \\ \hline \vdots & & \vdots \\ \hline B_{K-1,0} & \cdots & B_{K-1,N-1} \end{array} \right),$$

where the partitionings are "conformal", then

$$C_{i,j} = \sum_{p=0}^{K-1} A_{i,p} B_{p,j}.$$

**Remark 1.3.1.6** If the above review of linear transformations, matrices, matrix-vector multiplication, and matrix-matrix multiplication makes you exclaim "That is all a bit too fast for me!" then it is time for you to take a break and review Weeks 2-5 of our introductory linear algebra course "Linear Algebra: Foundations to Frontiers." Information, including notes [20] (optionally downloadable for free) and a link to the course on edX [21] (which can be audited for free) can be found at <http://ulaff.net>.

### 1.3.2 What is a matrix norm?



YouTube: <https://www.youtube.com/watch?v=6DsBTz1eU7E>

A matrix norm extends the notions of an absolute value and vector norm to matrices:

**Definition 1.3.2.1 Matrix norm.** Let  $\nu : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$ . Then  $\nu$  is a (matrix) norm if for all  $A, B \in \mathbb{C}^{m \times n}$  and all  $\alpha \in \mathbb{C}$

- $A \neq 0 \Rightarrow \nu(A) > 0$  ( $\nu$  is positive definite),
- $\nu(\alpha A) = |\alpha|\nu(A)$  ( $\nu$  is homogeneous), and
- $\nu(A + B) \leq \nu(A) + \nu(B)$  ( $\nu$  obeys the triangle inequality).

◊

**Homework 1.3.2.1** Let  $\nu : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  be a matrix norm.

ALWAYS/SOMETIMES/NEVER:  $\nu(0) = 0$ .

**Hint.** Review the proof on [Homework 1.2.2.1](#).

**Answer.** ALWAYS.

Now prove it.

**Solution.** Let  $A \in \mathbb{C}^{m \times n}$ . Then

$$\begin{aligned}\nu(0) &= < 0 \cdot A = 0 > \\ \nu(0 \cdot A) &= < \|\cdot\|_\nu \text{ is homogeneous} > \\ 0\nu(A) &= < \text{algebra} > \\ 0 &\end{aligned}$$

**Remark 1.3.2.2** As we do with vector norms, we will typically use  $\|\cdot\|$  instead of  $\nu(\cdot)$  for a function that is a matrix norm.

### 1.3.3 The Frobenius norm



YouTube: <https://www.youtube.com/watch?v=0ZHnGgrJXa4>

**Definition 1.3.3.1 The Frobenius norm.** The Frobenius norm  $\|\cdot\|_F : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is defined for  $A \in \mathbb{C}^{m \times n}$  by

$$\|A\|_F = \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |a_{i,j}|^2} = \sqrt{|a_{0,0}|^2 + \cdots + |a_{0,n-1}|^2 + \cdots + |a_{m-1,0}|^2 + \cdots + |a_{m-1,n-1}|^2}.$$

◇

One can think of the Frobenius norm as taking the columns of the matrix, stacking them on top of each other to create a vector of size  $m \times n$ , and then taking the vector 2-norm of the result.

**Homework 1.3.3.1** Partition  $m \times n$  matrix  $A$  by columns:

$$A = \left( \begin{array}{c|c|c} a_0 & \cdots & a_{n-1} \end{array} \right).$$

Show that

$$\|A\|_F^2 = \sum_{j=0}^{n-1} \|a_j\|_2^2.$$

**Solution.**

$$\begin{aligned}
 \|A\|_F &= <\text{definition}> \\
 &= \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2} \\
 &= <\text{commutativity of addition}> \\
 &= \sqrt{\sum_{j=0}^{n-1} \sum_{i=0}^{m-1} |\alpha_{i,j}|^2} \\
 &= <\text{definition of vector 2-norm}> \\
 &= \sqrt{\sum_{j=0}^{n-1} \|a_j\|_2^2}
 \end{aligned}$$

**Homework 1.3.3.2** Prove that the Frobenius norm is a norm.

**Solution.** Establishing that this function is positive definite and homogeneous is straight forward. To show that the triangle inequality holds it helps to realize that if  $A = (a_0 | a_1 | \cdots | a_{n-1})$  then

$$\begin{aligned}
 \|A\|_F &= <\text{definition}> \\
 &= \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2} \\
 &= <\text{commutativity of addition}> \\
 &= \sqrt{\sum_{j=0}^{n-1} \sum_{i=0}^{m-1} |\alpha_{i,j}|^2} \\
 &= <\text{definition of vector 2-norm}> \\
 &= \sqrt{\sum_{j=0}^{n-1} \|a_j\|_2^2} \\
 &= <\text{definition of vector 2-norm}> \\
 &= \sqrt{\left\| \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \right\|_2^2}.
 \end{aligned}$$

In other words, it equals the vector 2-norm of the vector that is created by stacking the columns of  $A$  on top of each other. One can then exploit the fact that the vector 2-norm obeys the triangle inequality.

**Homework 1.3.3.3** Partition  $m \times n$  matrix  $A$  by rows:

$$A = \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix}.$$

Show that

$$\|A\|_F^2 = \sum_{i=0}^{m-1} \|\tilde{a}_i\|_2^2,$$

where  $\tilde{a}_i = \tilde{a}_i^T$ .

**Solution.**

$$\begin{aligned}\|A\|_F &= \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2} \\ &= \sqrt{\sum_{j=0}^{n-1} \|\tilde{a}_j\|_2^2}.\end{aligned}$$

Let us review the definition of the transpose of a matrix (which we have already used when defining the dot product of two real-valued vectors and when identifying a row in a matrix):

**Definition 1.3.3.2 Transpose.** If  $A \in \mathbb{C}^{m \times n}$  and

$$A = \left( \begin{array}{c|c|c|c} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \hline \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \hline \vdots & \vdots & & \vdots \\ \vdots & & & \vdots \\ \hline \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{array} \right)$$

then its **transpose** is defined by

$$A^T = \left( \begin{array}{c|c|c|c} \alpha_{0,0} & \alpha_{1,0} & \cdots & \alpha_{n-1,0} \\ \hline \alpha_{0,1} & \alpha_{1,1} & \cdots & \alpha_{n-1,1} \\ \hline \vdots & \vdots & & \vdots \\ \vdots & & & \vdots \\ \hline \alpha_{0,m-1} & \alpha_{1,m-1} & \cdots & \alpha_{n-1,m-1} \end{array} \right).$$

◊

For complex-valued matrices, it is important to also define the **Hermitian transpose** of a matrix:

**Definition 1.3.3.3 Hermitian transpose.** If  $A \in \mathbb{C}^{m \times n}$  and

$$A = \left( \begin{array}{c|c|c|c} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \hline \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \hline \vdots & \vdots & & \vdots \\ \vdots & & & \vdots \\ \hline \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{array} \right)$$

then its **Hermitian transpose** is defined by

$$A^H = \overline{A}^T \left( \begin{array}{c|c|c|c} \overline{\alpha}_{0,0} & \overline{\alpha}_{1,0} & \cdots & \overline{\alpha}_{m-1,0} \\ \hline \overline{\alpha}_{0,1} & \overline{\alpha}_{1,1} & \cdots & \overline{\alpha}_{m-1,1} \\ \vdots & \vdots & & \vdots \\ \vdots & & & \vdots \\ \hline \overline{\alpha}_{0,n-1} & \overline{\alpha}_{1,n-1} & \cdots & \overline{\alpha}_{m-1,n-1} \end{array} \right),$$

where  $\overline{A}$  denotes the **conjugate of a matrix**, in which each element of the matrix is conjugated.  $\diamond$

We note that

- $\overline{A}^T = \overline{A^T}$ .
- If  $A \in \mathbb{R}^{m \times n}$ , then  $A^H = A^T$ .
- If  $x \in \mathbb{C}^m$ , then  $x^H$  is defined consistent with how we have used it before.
- If  $\alpha \in \mathbb{C}$ , then  $\alpha^H = \overline{\alpha}$ .

(If you view the scalar as a matrix and then Hermitian transpose it, you get the matrix with as only element  $\overline{\alpha}$ .)

*Don't Panic!* While working with complex-valued scalars, vectors, and matrices may appear a bit scary at first, you will soon notice that it is not really much more complicated than working with their real-valued counterparts.

**Homework 1.3.3.4** Let  $A \in \mathbb{C}^{m \times k}$  and  $B \in \mathbb{C}^{k \times n}$ . Using what you once learned about matrix transposition and matrix-matrix multiplication, reason that  $(AB)^H = B^H A^H$ .

**Solution.**

$$\begin{aligned} (AB)^H &= < X^H = \overline{X^T} > \\ \overline{(AB)^T} &= < \text{you once discovered that } (AB)^T = B^T A^T > \\ \overline{B^T A^T} &= < \text{you may check separately that } \overline{XY} = \overline{X}\overline{Y} > \\ \overline{B^T} \overline{A^T} &= < \overline{X^T} = \overline{X}^T > \\ B^H A^H & \end{aligned}$$

**Definition 1.3.3.4 Hermitian.** A matrix  $A \in \mathbb{C}^{m \times m}$  is **Hermitian** if and only if  $A = A^H$ .  $\diamond$

Obviously, if  $A \in \mathbb{R}^{m \times m}$ , then  $A$  is a Hermitian matrix if and only if  $A$  is a symmetric matrix.

**Homework 1.3.3.5** Let  $A \in \mathbb{C}^{m \times n}$ .

ALWAYS/SOMETIMES/NEVER:  $\|A^H\|_F = \|A\|_F$ .

**Answer.** ALWAYS

**Solution.**

$$\begin{aligned}
 \|A\|_F &= <\text{definition}> \\
 &= \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2} \\
 &= <\text{commutativity of addition}> \\
 &= \sqrt{\sum_{j=0}^{n-1} \sum_{i=0}^{m-1} |\alpha_{i,j}|^2} \\
 &= <\text{change of variables}> \\
 &= \sqrt{\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} |\alpha_{j,i}|^2} \\
 &= <\text{algebra}> \\
 &= \sqrt{\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} |\bar{\alpha}_{j,i}|^2} \\
 &= <\text{definition}> \\
 \|A^H\|_F
 \end{aligned}$$

Similarly, other matrix norms can be created from vector norms by viewing the matrix as a vector. It turns out that, other than the Frobenius norm, these aren't particularly interesting in practice. An example can be found in [Homework 1.6.1.6](#).

**Remark 1.3.3.5** The Frobenius norm of a  $m \times n$  matrix is easy to compute (requiring  $O(mn)$  computations). The functions  $f(A) = \|A\|_F$  and  $f(A) = \|A\|_F^2$  are also differentiable. However, you'd be hard-pressed to find a meaningful way of linking the definition of the Frobenius norm to a measure of an underlying linear transformation (other than by first transforming that linear transformation into a matrix).

### 1.3.4 Induced matrix norms



YouTube: <https://www.youtube.com/watch?v=M6ZVBRFnYcU>

Recall from [Subsection 1.3.1](#) that a matrix,  $A \in \mathbb{C}^{m \times n}$ , is a 2-dimensional array of numbers that represents a linear transformation,  $L : \mathbb{C}^n \rightarrow \mathbb{C}^m$ , such that for all  $x \in \mathbb{C}^n$  the matrix-vector multiplication  $Ax$  yields the same result as does  $L(x)$ .

The question "What is the norm of matrix  $A$ ?" or, equivalently, "How 'large' is  $A$ ?" is the same as asking the question "How 'large' is  $L$ ?" What does this mean? It suggests that what we really want is a measure of how much linear transformation  $L$  or, equivalently, matrix  $A$  "stretches" (magnifies) the "length" of a vector. This observation motivates a class of matrix norms known as induced matrix norms.

**Definition 1.3.4.1 Induced matrix norm.** Let  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  be vector norms. Define  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_{\mu,\nu} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

◊

Matrix norms that are defined in this way are said to be **induced** matrix norms.

**Remark 1.3.4.2** In context, it is obvious (from the column size of the matrix) what the size of vector  $x$  is. For this reason, we will write

$$\|A\|_{\mu,\nu} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu} \quad \text{as} \quad \|A\|_{\mu,\nu} = \sup_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

Let us start by interpreting this. How "large"  $A$  is, as measured by  $\|A\|_{\mu,\nu}$ , is defined as the most that  $A$  magnifies the length of nonzero vectors, where the length of the vector,  $x$ , is measured with norm  $\|\cdot\|_\nu$  and the length of the transformed vector,  $Ax$ , is measured with norm  $\|\cdot\|_\mu$ .

Two comments are in order. First,

$$\sup_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu} = \sup_{\|x\|_\nu=1} \|Ax\|_\mu.$$

This follows from the following sequence of equivalences:

$$\begin{aligned} & \sup_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu} \\ &= \quad < \text{homogeneity} > \\ & \sup_{x \neq 0} \left\| \frac{Ax}{\|x\|_\nu} \right\|_\mu \\ &= \quad < \text{norms are associative} > \\ & \sup_{x \neq 0} \left\| A \frac{x}{\|x\|_\nu} \right\|_\mu \\ &= \quad < \text{substitute } y = x/\|x\|_\nu > \\ & \sup_{\|y\|_\nu=1} \|Ay\|_\mu. \end{aligned}$$

Second, the "sup" (which stands for supremum) is used because we can't claim yet that there is a nonzero vector  $x$  for which

$$\sup_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu}$$

is attained or, alternatively, a vector,  $x$ , with  $\|x\|_\nu = 1$  for which

$$\sup_{\|x\|_\nu=1} \|Ax\|_\mu$$

is attained. In words, it is not immediately obvious that there is a vector for which the supremum is attained. The fact is that there is always such a vector  $x$ . The proof again depends on a result from real analysis, also employed in [Proof 1.2.6.1](#), that states that  $\sup_{x \in S} f(x)$  is attained for some vector  $x \in S$  as long as  $f$  is continuous and  $S$  is a compact set. For any norm,  $\|x\| = 1$  is a compact set. Thus, we can replace sup by max from here on in our discussion.

We conclude that the following two definitions are equivalent definitions to the one we already gave:

**Definition 1.3.4.3** Let  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  be vector norms. Define  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_{\mu,\nu} = \max_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

or, equivalently,

$$\|A\|_{\mu,\nu} = \max_{\|x\|_\nu=1} \|Ax\|_\mu.$$

◊

**Remark 1.3.4.4** In this course, we will often encounter proofs involving norms. Such proofs are much cleaner if one starts by strategically picking the most convenient of these two definitions. Until you gain the intuition needed to pick which one is better, you may have to start your proof using one of them and then switch to the other one if the proof becomes unwieldy.

**Theorem 1.3.4.5**  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is a norm.

*Proof.* To prove this, we merely check whether the three conditions are met:

Let  $A, B \in \mathbb{C}^{m \times n}$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $A \neq 0 \Rightarrow \|A\|_{\mu,\nu} > 0$  ( $\|\cdot\|_{\mu,\nu}$  is positive definite):

Notice that  $A \neq 0$  means that at least one of its columns is not a zero vector (since at least one element is nonzero). Let us assume it is the  $j$ th column,  $a_j$ , that is nonzero. Let  $e_j$  equal the column of  $I$  (the identity matrix) indexed with  $j$ . Then

$$\begin{aligned} \|A\|_{\mu,\nu} &= <\text{definition}> \\ &= \max_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu} \\ &\geq <\text{ } e_j \text{ is a specific vector}> \\ &= \frac{\|Ae_j\|_\mu}{\|e_j\|_\nu} \\ &= <\text{ } Ae_j = a_j > \\ &= \frac{\|a_j\|_\mu}{\|e_j\|_\nu} \\ &> <\text{we assumed that } a_j \neq 0> \\ &= 0. \end{aligned}$$

- $\|\alpha A\|_{\mu,\nu} = |\alpha| \|A\|_{\mu,\nu}$  ( $\|\cdot\|_{\mu,\nu}$  is homogeneous):

$$\begin{aligned}
& \|\alpha A\|_{\mu,\nu} \\
&= <\text{definition}> \\
& \max_{x \neq 0} \frac{\|\alpha Ax\|_\mu}{\|x\|_\nu} \\
&= <\text{homogeneity}> \\
& \max_{x \neq 0} |\alpha| \frac{\|Ax\|_\mu}{\|x\|_\nu} \\
&= <\text{algebra}> \\
& |\alpha| \max_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu} \\
&= <\text{definition}> \\
& |\alpha| \|A\|_{\mu,\nu}.
\end{aligned}$$

- $\|A + B\|_{\mu,\nu} \leq \|A\|_{\mu,\nu} + \|B\|_{\mu,\nu}$  ( $\|\cdot\|_{\mu,\nu}$  obeys the triangle inequality).

$$\begin{aligned}
& \|A + B\|_{\mu,\nu} \\
&= <\text{definition}> \\
& \max_{x \neq 0} \frac{\|(A+B)x\|_\mu}{\|x\|_\nu} \\
&= <\text{distribute}> \\
& \max_{x \neq 0} \frac{\|Ax+Bx\|_\mu}{\|x\|_\nu} \\
&\leq <\text{triangle inequality}> \\
& \max_{x \neq 0} \frac{\|Ax\|_\mu + \|Bx\|_\mu}{\|x\|_\nu} \\
&\leq <\text{algebra}> \\
& \max_{x \neq 0} \left( \frac{\|Ax\|_\mu}{\|x\|_\nu} + \frac{\|Bx\|_\mu}{\|x\|_\nu} \right) \\
&\leq <\text{algebra}> \\
& \max_{x \neq 0} \frac{\|Ax\|_\mu}{\|x\|_\nu} + \max_{x \neq 0} \frac{\|Bx\|_\mu}{\|x\|_\nu} \\
&= <\text{definition}> \\
& \|A\|_{\mu,\nu} + \|B\|_{\mu,\nu}.
\end{aligned}$$

■

When  $\|\cdot\|_\mu$  and  $\|\cdot\|_\nu$  are the same norm (but possibly for different sizes of vectors), the induced norm becomes

**Definition 1.3.4.6** Define  $\|\cdot\|_\mu : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

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

or, equivalently,

$$\|A\|_\mu = \max_{\|x\|_\mu=1} \|Ax\|_\mu.$$

◇

**Homework 1.3.4.1** Consider the vector  $p$ -norm  $\|\cdot\|_p : \mathbb{C}^n \rightarrow \mathbb{R}$  and let us denote the induced matrix norm by  $\|\|\cdot\||| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  for this exercise:  $\|\|A\||| = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$ .

ALWAYS/SOMETIMES/NEVER:  $\|\|y\||| = \|y\|_p$  for  $y \in \mathbb{C}^m$ .

**Answer.** ALWAYS

**Solution.**

$$\begin{aligned}
 & |||y||| \\
 &= < \text{definition} > \\
 & \max_{x \neq 0} \frac{\|yx\|_p}{\|x\|_p} \\
 &= < x \text{ is a scalar since } y \text{ is a matrix with one column, and hence } \|x\|_p = |x| > \\
 & \max_{x \neq 0} |x| \frac{\|y\|_p}{|x|} \\
 &= < \text{algebra} > \\
 & \max_{x \neq 0} \|y\|_p \\
 &= < \text{algebra} > \\
 & \|y\|
 \end{aligned}$$

This last exercise is important. One can view a vector  $x \in \mathbb{C}^m$  as an  $m \times 1$  matrix. What this last exercise tells us is that regardless of whether we view  $x$  as a matrix or a vector,  $\|x\|_p$  is the same.

We already encountered the vector  $p$ -norms as an important class of vector norms. The matrix  $p$ -norm is induced by the corresponding vector norm, as defined by

**Definition 1.3.4.7 Matrix  $p$ -norm.** For any vector  $p$ -norm, define the corresponding matrix  $p$ -norm  $\|\cdot\|_p : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \quad \text{or, equivalently,} \quad \|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p.$$

◊

**Remark 1.3.4.8** The matrix  $p$ -norms with  $p \in \{1, 2, \infty\}$  will play an important role in our course, as will the Frobenius norm. As the course unfolds, we will realize that in practice the matrix 2-norm is of great theoretical importance but difficult to evaluate, except for special matrices. The 1-norm,  $\infty$ -norm, and Frobenius norms are straightforward and relatively cheap to compute (for an  $m \times n$  matrix, computing these costs  $O(mn)$  computation).

### 1.3.5 The matrix 2-norm



YouTube: [https://www.youtube.com/watch?v=wZAlH\\_K9XeI](https://www.youtube.com/watch?v=wZAlH_K9XeI)

Let us instantiate the definition of the vector  $p$  norm for the case where  $p = 2$ , giving us a matrix norm induced by the vector 2-norm or Euclidean norm:

**Definition 1.3.5.1 Matrix 2-norm.** Define the matrix 2-norm  $\|\cdot\|_2 : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2.$$

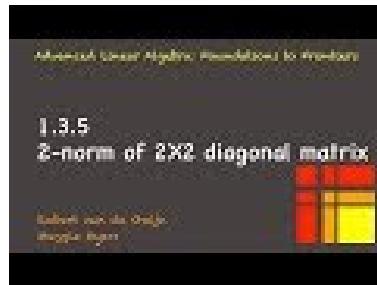
◊

**Remark 1.3.5.2** The problem with the matrix 2-norm is that it is hard to compute. At some point later in this course, you will find out that if  $A$  is a Hermitian matrix ( $A = A^H$ ), then  $\|A\|_2 = |\lambda_0|$ , where  $\lambda_0$  equals the eigenvalue of  $A$  that is largest in magnitude. You may recall from your prior linear algebra experience that computing eigenvalues involves computing the roots of polynomials, and for polynomials of degree three or greater, this is a nontrivial task. We will see that the matrix 2-norm plays an important role in the theory of linear algebra, but less so in practical computation.

**Example 1.3.5.3** Show that

$$\left\| \begin{pmatrix} \delta_0 & 0 \\ 0 & \delta_1 \end{pmatrix} \right\|_2 = \max(|\delta_0|, |\delta_1|).$$

**Solution.**



YouTube: <https://www.youtube.com/watch?v=B2rz0i5BB3A>

[slides (PDF)] [LaTeX source] □

**Remark 1.3.5.4** The proof of the last example builds on a general principle: Showing that  $\max_{x \in D} f(x) = \alpha$  for some function  $f : D \rightarrow R$  can be broken down into showing that both

$$\max_{x \in D} f(x) \leq \alpha$$

and

$$\max_{x \in D} f(x) \geq \alpha.$$

In turn, showing that  $\max_{x \in D} f(x) \geq \alpha$  can often be accomplished by showing that there exists a vector  $y \in D$  such that  $f(y) = \alpha$  since then

$$\alpha = f(y) \leq \max_{x \in D} f(x).$$

We will use this technique in future proofs involving matrix norms.

**Homework 1.3.5.1** Let  $D \in \mathbb{C}^{m \times m}$  be a diagonal matrix with diagonal entries  $\delta_0, \dots, \delta_{m-1}$ . Show that

$$\|D\|_2 = \max_{j=0}^{m-1} |\delta_j|.$$

**Solution.** First, we show that  $\|D\|_2 = \max_{\|x\|_2=1} \|Dx\|_2 \leq \max_{i=0}^{m-1} |\delta_i|$ :

$$\begin{aligned} & \|D\|_2^2 \\ &= <\text{definition}> \\ &= \max_{\|x\|_2=1} \|Dx\|_2^2 \\ &= <\text{diagonal vector multiplication}> \\ &= \max_{\|x\|_2=1} \left\| \begin{pmatrix} \delta_0 \chi_0 \\ \vdots \\ \delta_{m-1} \chi_{m-1} \end{pmatrix} \right\|_2^2 \\ &= <\text{definition}> \\ &= \max_{\|x\|_2=1} \sum_{i=0}^{m-1} |\delta_i \chi_i|^2 \\ &= <\text{homogeneity}> \\ &= \max_{\|x\|_2=1} \sum_{i=0}^{m-1} |\delta_i|^2 |\chi_i|^2 \\ &\leq <\text{algebra}> \\ &= \max_{\|x\|_2=1} \sum_{i=0}^{m-1} \left[ \max_{j=0}^{m-1} |\delta_j| \right]^2 |\chi_i|^2 \\ &= <\text{algebra}> \\ &= \left[ \max_{j=0}^{m-1} |\delta_j| \right]^2 \max_{\|x\|_2=1} \sum_{i=0}^{m-1} |\chi_i|^2 \\ &= <\|x\|_2 = 1> \\ &= \left[ \max_{j=0}^{m-1} |\delta_j| \right]^2. \end{aligned}$$

Next, we show that there is a vector  $y$  with  $\|y\|_2 = 1$  such that  $\|Dy\|_2 = \max_{i=0}^{m-1} |\delta_i|$ : Let  $j$  be such that  $|\delta_j| = \max_{i=0}^{m-1} |\delta_i|$  and choose  $y = e_j$ . Then

$$\begin{aligned} & \|Dy\|_2 \\ &= <y = e_j> \\ &= \|De_j\|_2 \\ &= < D = \text{diag}(\delta_0, \dots, \delta_{m-1})> \\ &= \|\delta_j e_j\|_2 \\ &= <\text{homogeneity}> \\ &= |\delta_j| \|e_j\|_2 \\ &= <\|e_j\|_2 = 1> \\ &= |\delta_j| \\ &= <\text{choice of } j> \\ &= \max_{i=0}^{m-1} |\delta_i| \end{aligned}$$

Hence  $\|D\|_2 = \max_{j=0}^{m-1} |\delta_j|$ .

**Homework 1.3.5.2** Let  $y \in \mathbb{C}^m$  and  $x \in \mathbb{C}^n$ .

ALWAYS/SOMETIMES/NEVER:  $\|yx^H\|_2 = \|y\|_2\|x\|_2$ .

**Hint.** Prove that  $\|yx^H\|_2 \geq \|y\|_2\|x\|_2$  and that there exists a vector  $z$  so that  $\frac{\|yx^H z\|_2}{\|z\|_2} = \|y\|_2\|x\|_2$ .

**Answer.** ALWAYS

Now prove it!

**Solution.** W.l.o.g. assume that  $x \neq 0$ .

We know by the Cauchy-Schwartz inequality that  $|x^H z| \leq \|x\|_2\|z\|_2$ . Hence

$$\begin{aligned} \|yx^H\|_2 &= <\text{definition}> \\ &= \max_{\|z\|_2=1} \|yx^H z\|_2 \\ &= <\|\cdot\|_2 \text{ is homogenius}> \\ &= \max_{\|z\|_2=1} |x^H z| \|y\|_2 \\ &\leq <\text{Cauchy-Schwartz inequality}> \\ &= \max_{\|z\|_2=1} \|x\|_2\|z\|_2\|y\|_2 \\ &= <\|z\|_2 = 1> \\ &= \|x\|_2\|y\|_2. \end{aligned}$$

But also

$$\begin{aligned} \|yx^H\|_2 &= <\text{definition}> \\ &= \max_{z \neq 0} \|yx^H z\|_2/\|z\|_2 \\ &\geq <\text{specific } z> \\ &= \|yx^H x\|_2/\|x\|_2 \\ &= <x^H x = \|x\|_2^2; \text{ homogeneity}> \\ &= \|x\|_2^2\|y\|_2/\|x\|_2 \\ &= <\text{algebra}> \\ &= \|y\|_2\|x\|_2. \end{aligned}$$

Hence

$$\|yx^H\|_2 = \|y\|_2\|x\|_2.$$

**Homework 1.3.5.3** Let  $A \in \mathbb{C}^{m \times n}$  and  $a_j$  its column indexed with  $j$ . ALWAYS/SOMETIMES/NEVER:  $\|a_j\|_2 \leq \|A\|_2$ .

**Hint.** What vector has the property that  $a_j = Ax$ ?

**Answer.** ALWAYS.

Now prove it!

**Solution.**

$$\begin{aligned}
 \|a_j\|_2 &= \\
 \|Ae_j\|_2 &\leq \\
 \max_{\|x\|_2=1} \|Ax\|_2 &= \\
 \|A\|_2.
 \end{aligned}$$

**Homework 1.3.5.4** Let  $A \in \mathbb{C}^{m \times n}$ . Prove that

- $\|A\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^H Ax|$ .
- $\|A^H\|_2 = \|A\|_2$ .
- $\|A^H A\|_2 = \|A\|_2^2$ .

**Hint.** Proving  $\|A\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^H Ax|$  requires you to invoke the Cauchy-Schwartz inequality from [Theorem 1.2.3.3](#).

**Solution.**

- $\|A\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^H Ax|$ :

$$\begin{aligned}
 &\max_{\|x\|_2=\|y\|_2=1} |y^H Ax| \\
 &\leq < \text{Cauchy-Schwartz} > \\
 &\max_{\|x\|_2=\|y\|_2=1} \|y\|_2 \|Ax\|_2 \\
 &= < \|y\|_2 = 1 > \\
 &\max_{\|x\|_2=1} \|Ax\|_2 \\
 &= < \text{definition} > \\
 &\|A\|_2.
 \end{aligned}$$

Also, we know there exists  $x$  with  $\|x\|_2 = 1$  such that  $\|A\|_2 = \|Ax\|_2$ . Let  $y = Ax/\|Ax\|_2$ . Then

$$\begin{aligned}
 |y^H Ax| &= < \text{instantiate} > \\
 \left| \frac{(Ax)^H (Ax)}{\|Ax\|_2} \right| &= < z^H z = \|z\|_2^2 > \\
 \left| \frac{\|Ax\|_2^2}{\|Ax\|_2} \right| &= < \text{algebra} > \\
 \|Ax\|_2 &= < x \text{ was chosen so that } \|Ax\|_2 = \|A\|_2 > \\
 \|A\|_2
 \end{aligned}$$

Hence the bound is attained. We conclude that  $\|A\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^H Ax|$ .

- $\|A^H\|_2 = \|A\|_2$ :

$$\begin{aligned}
 & \|A^H\|_2 \\
 &= <\text{first part of homework}> \\
 & \max_{\|x\|_2=\|y\|_2=1} |y^H A^H x| \\
 &= <|\bar{\alpha}| = |\alpha|> \\
 & \max_{\|x\|_2=\|y\|_2=1} |x^H A y| \\
 &= <\text{first part of homework}> \\
 & \|A\|_2.
 \end{aligned}$$

- $\|A^H A\|_2 = \|A\|_2^2$ :

$$\begin{aligned}
 & \|A^H A\|_2 \\
 &= <\text{first part of homework}> \\
 & \max_{\|x\|_2=\|y\|_2=1} |y^H A^H A x| \\
 &\geq <\text{restricts choices of } y> \\
 & \max_{\|x\|_2=1} |x^H A^H A x| \\
 &= <z^H z = \|z\|_2^2> \\
 & \max_{\|x\|_2=1} \|Ax\|_2^2 \\
 &= <\text{algebra}> \\
 & \left( \max_{\|x\|_2=1} \|Ax\|_2 \right)^2 \\
 &= <\text{definition}> \\
 & \|A\|_2^2.
 \end{aligned}$$

So,  $\|A^H A\|_2 \geq \|A\|_2^2$ .

Now, let's show that  $\|A^H A\|_2 \leq \|A\|_2^2$ . This would be trivial if we had already discussed the fact that  $\|\cdot\|_2$  is a submultiplicative norm (which we will in a future unit). But let's do it from scratch. First, we show that  $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$  for all (appropriately sized) matrices  $A$  and  $x$ :

$$\begin{aligned}
 & \|Ax\|_2 \\
 &= <\text{norms are homogeneous}> \\
 & \|A \frac{x}{\|x\|_2}\|_2 \|x\|_2 \\
 &\leq <\text{algebra}> \\
 & \max_{\|y\|_2=1} \|Ay\|_2 \|x\|_2 \\
 &= <\text{definition of 2-norm}> \\
 & \|A\|_2 \|x\|_2.
 \end{aligned}$$

With this, we can then show that

$$\begin{aligned}
 \|A^H A\|_2 &= \langle \text{definition of 2-norm} \rangle \\
 \max_{\|x\|_2=1} \|A^H A x\|_2 &\leq \langle \|Az\|_2 \leq \|A\|_2 \|z\|_2 \rangle \\
 \max_{\|x\|_2=1} (\|A^H\|_2 \|Ax\|_2) &= \langle \text{algebra} \rangle \\
 \|A^H\|_2 \max_{\|x\|_2=1} \|Ax\|_2 &= \langle \text{definition of 2-norm} \rangle \\
 \|A^H\|_2 \|A\|_2 &= \langle \|A^H\|_2 = \|A\| \rangle \\
 \|A\|_2^2
 \end{aligned}$$

**Homework 1.3.5.5** Partition  $A = \left( \begin{array}{c|c|c} A_{0,0} & \cdots & A_{0,N-1} \\ \vdots & & \vdots \\ A_{M-1,0} & \cdots & A_{M-1,N-1} \end{array} \right)$ .

ALWAYS/SOMETIMES/NEVER:  $\|A_{i,j}\|_2 \leq \|A\|_2$ .

**Hint.** Using [Homework 1.3.5.4](#) choose  $v_j$  and  $w_i$  such that  $\|A_{i,j}\|_2 = |w_i^H A_{i,j} v_j|$ .

**Solution.** Choose  $v_j$  and  $w_i$  such that  $\|A_{i,j}\|_2 = |w_i^H A_{i,j} v_j|$ . Next, choose  $v$  and  $w$  such that

$$v = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \hline v_j \\ 0 \\ \hline \vdots \\ 0 \end{pmatrix}, \quad w = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \hline w_i \\ 0 \\ \hline \vdots \\ 0 \end{pmatrix}, \quad \text{and} \quad w^H A v = w_i^H A_{i,j} v_j.$$

Then, by [Homework 1.3.5.4](#)

$$\|A\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^H A x| \geq |w^H A v| = |w_i^H A_{i,j} v_j| = \|A_{i,j}\|_2.$$

### 1.3.6 Computing the matrix 1-norm and $\infty$ -norm



YouTube: <https://www.youtube.com/watch?v=QTKZdGQ2C6w>

The matrix 1-norm and matrix  $\infty$ -norm are of great importance because, unlike the matrix 2-norm, they are easy and relatively cheap to compute.. The following exercises show how to practically compute the matrix 1-norm and  $\infty$ -norm.

**Homework 1.3.6.1** Let  $A \in \mathbb{C}^{m \times n}$  and partition  $A = ( a_0 \mid a_1 \mid \cdots \mid a_{n-1} )$ . ALWAYS/SOMETIMES/NEVER:  $\|A\|_1 = \max_{0 \leq j < n} \|a_j\|_1$ .

**Hint.** Prove it for the real valued case first.

**Answer.** ALWAYS

**Solution.** Let  $J$  be chosen so that  $\max_{0 \leq j < n} \|a_j\|_1 = \|a_J\|_1$ . Then

$$\begin{aligned}
 & \|A\|_1 \\
 &= \quad <\text{definition}> \\
 & \max_{\|x\|_1=1} \|Ax\|_1 \\
 &= \quad <\text{expose the columns of } A \text{ and elements of } x> \\
 & \max_{\|x\|_1=1} \left\| \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \right\|_1 \\
 &= \quad <\text{definition of matrix-vector multiplication}> \\
 & \max_{\|x\|_1=1} \|\chi_0 a_0 + \chi_1 a_1 + \cdots + \chi_{n-1} a_{n-1}\|_1 \\
 & \leq \quad <\text{triangle inequality}> \\
 & \max_{\|x\|_1=1} (\|\chi_0 a_0\|_1 + \|\chi_1 a_1\|_1 + \cdots + \|\chi_{n-1} a_{n-1}\|_1) \\
 &= \quad <\text{homogeneity}> \\
 & \max_{\|x\|_1=1} (|\chi_0| \|a_0\|_1 + |\chi_1| \|a_1\|_1 + \cdots + |\chi_{n-1}| \|a_{n-1}\|_1) \\
 & \leq \quad <\text{choice of } a_J> \\
 & \max_{\|x\|_1=1} (|\chi_0| \|a_J\|_1 + |\chi_1| \|a_J\|_1 + \cdots + |\chi_{n-1}| \|a_J\|_1) \\
 &= \quad <\text{factor out } \|a_J\|_1> \\
 & \max_{\|x\|_1=1} (|\chi_0| + |\chi_1| + \cdots + |\chi_{n-1}|) \|a_J\|_1 \\
 &= \quad <\text{algebra}> \\
 & \|a_J\|_1.
 \end{aligned}$$

Also,

$$\begin{aligned}
 & \|a_J\|_1 \\
 &= \quad <e_J \text{ picks out column } J> \\
 & \|Ae_J\|_1 \\
 & \leq \quad <e_J \text{ is a specific choice of } x> \\
 & \max_{\|x\|_1=1} \|Ax\|_1.
 \end{aligned}$$

Hence

$$\|a_J\|_1 \leq \max_{\|x\|_1=1} \|Ax\|_1 \leq \|a_J\|_1$$

which implies that

$$\max_{\|x\|_1=1} \|Ax\|_1 = \|a_J\|_1 = \max_{0 \leq j < n} \|a_j\|_1.$$

**Homework 1.3.6.2** Let  $A \in \mathbb{C}^{m \times n}$  and partition  $A = \begin{pmatrix} \overline{\tilde{a}_0^T} \\ \overline{\tilde{a}_1^T} \\ \vdots \\ \overline{\tilde{a}_{m-1}^T} \end{pmatrix}$ .

ALWAYS/SOMETIMES/NEVER:

$$\|A\|_\infty = \max_{0 \leq i < m} \|\tilde{a}_i\|_1 (= \max_{0 \leq i < m} (|\alpha_{i,0}| + |\alpha_{i,1}| + \cdots + |\alpha_{i,n-1}|))$$

Notice that in this exercise  $\tilde{a}_i$  is really  $(\tilde{a}_i^T)^T$  since  $\tilde{a}_i^T$  is the label for the  $i$ th row of matrix  $A$ .

**Hint.** Prove it for the real valued case first.

**Answer.** ALWAYS

**Solution.** Partition  $A = \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix}$ . Then

$$\begin{aligned}
& \|A\|_\infty \\
&= <\text{definition}> \\
& \max_{\|x\|_\infty=1} \|Ax\|_\infty \\
&= <\text{expose rows}> \\
& \max_{\|x\|_\infty=1} \left\| \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix} x \right\|_\infty \\
&= <\text{matrix-vector multiplication}> \\
& \max_{\|x\|_\infty=1} \left\| \begin{pmatrix} \tilde{a}_0^T x \\ \vdots \\ \tilde{a}_{m-1}^T x \end{pmatrix} \right\|_\infty \\
&= <\text{definition of } \|\cdot\|_\infty> \\
& \max_{\|x\|_\infty=1} \left( \max_{0 \leq i < m} |\tilde{a}_i^T x| \right) \\
&= <\text{expose } \tilde{a}_i^T x> \\
& \max_{\|x\|_\infty=1} \max_{0 \leq i < m} \left| \sum_{p=0}^{n-1} \alpha_{i,p} \chi_p \right| \\
&\leq <\text{triangle inequality}> \\
& \max_{\|x\|_\infty=1} \max_{0 \leq i < m} \sum_{p=0}^{n-1} |\alpha_{i,p} \chi_p| \\
&= <\|x\|_\infty = 1> \\
& \max_{\|x\|_\infty=1} \max_{0 \leq i < m} \sum_{p=0}^{n-1} (|\alpha_{i,p}| |\chi_p|) \\
&\leq <\text{algebra}> \\
& \max_{\|x\|_\infty=1} \max_{0 \leq i < m} \sum_{p=0}^{n-1} (|\alpha_{i,p}| (\max_k |\chi_k|)) \\
&= <\text{definition of } \|\cdot\|_\infty> \\
& \max_{\|x\|_\infty=1} \max_{0 \leq i < m} \sum_{p=0}^{n-1} (|\alpha_{i,p}| \|x\|_\infty) \\
&= <\|x\|_\infty = 1> \\
& \max_{0 \leq i < m} \sum_{p=0}^{n-1} |\alpha_{i,p}| \\
&= <\text{definition of } \|\cdot\|_1> \\
& \|\tilde{a}_i\|_1
\end{aligned}$$

so that  $\|A\|_\infty \leq \max_{0 \leq i < m} \|\tilde{a}_i\|_1$ .

We also want to show that  $\|A\|_\infty \geq \max_{0 \leq i < m} \|\tilde{a}_i\|_1$ . Let  $k$  be such that  $\max_{0 \leq i < m} \|\tilde{a}_i\|_1 = \|\tilde{a}_k\|_1$  and pick  $y = \begin{pmatrix} \psi_0 \\ \vdots \\ \psi_{n-1} \end{pmatrix}$  so that  $\tilde{a}_k^T y = |\alpha_{k,0}| + |\alpha_{k,1}| + \cdots + |\alpha_{k,n-1}| = \|\tilde{a}_k\|_1$ . (This is a matter of picking  $\psi_j = |\alpha_{k,j}|/\alpha_{k,j}$  if  $\alpha_{k,j} \neq 0$  and  $\psi_j = 1$  otherwise. Then  $|\psi_j| = 1$ , and

hence  $\|y\|_\infty = 1$  and  $\psi_j \alpha_{k,j} = |\alpha_{k,j}|$ .) Then

$$\begin{aligned}
 & \|A\|_\infty \\
 &= <\text{definition}> \\
 & \max_{\|x\|_1=1} \|Ax\|_\infty \\
 &= <\text{expose rows}> \\
 & \max_{\|x\|_1=1} \left\| \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix} x \right\|_\infty \\
 &\geq <\text{y is a specific } x> \\
 & \left\| \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix} y \right\|_\infty \\
 &= <\text{matrix-vector multiplication}> \\
 & \left\| \begin{pmatrix} \tilde{a}_0^T y \\ \vdots \\ \tilde{a}_{m-1}^T y \end{pmatrix} \right\|_\infty \\
 &\geq <\text{algebra}> \\
 & |\tilde{a}_k^T y| \\
 &= <\text{choice of } y> \\
 & \|\tilde{a}_k\|_1 \\
 &= <\text{choice of } k> \\
 & \max_{0 \leq i < m} \|\tilde{a}_i\|_1
 \end{aligned}$$

**Remark 1.3.6.1** The last homework provides a hint as to how to remember how to compute the matrix 1-norm and  $\infty$ -norm: Since  $\|x\|_1$  must result in the same value whether  $x$  is considered as a vector or as a matrix, we can remember that the matrix 1-norm equals the maximum of the 1-norms of the columns of the matrix: Similarly, considering  $\|x\|_\infty$  as a vector norm or as matrix norm reminds us that the matrix  $\infty$ -norm equals the maximum of the 1-norms of vectors that become the rows of the matrix.

### 1.3.7 Equivalence of matrix norms



YouTube: <https://www.youtube.com/watch?v=Csqd4AnH7ws>

**Homework 1.3.7.1** Fill out the following table:

$A$	$\ A\ _1$	$\ A\ _\infty$	$\ A\ _F$	$\ A\ _2$
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$				
$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$				
$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$				

**Hint.** For the second and third, you may want to use [Homework 1.3.5.2](#) when computing the 2-norm.

**Solution.**

$A$	$\ A\ _1$	$\ A\ _\infty$	$\ A\ _F$	$\ A\ _2$
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	1	1	$\sqrt{3}$	1
$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	4	3	$2\sqrt{3}$	$2\sqrt{3}$
$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$	3	1	$\sqrt{3}$	$\sqrt{3}$

To compute the 2-norm of  $I$ , notice that

$$\|I\|_2 = \max_{\|x\|_2=1} \|Ix\|_2 = \max_{\|x\|_2=1} \|x\|_2 = 1.$$

Next, notice that

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}.$$

and

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}.$$

which allows us to invoke the result from [Homework 1.3.5.2](#).

We saw that vector norms are equivalent in the sense that if a vector is "small" in one

norm, it is "small" in all other norms, and if it is "large" in one norm, it is "large" in all other norms. The same is true for matrix norms.

**Theorem 1.3.7.1 Equivalence of matrix norms.** *Let  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  and  $\|\|\cdot\|\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  both be matrix norms. Then there exist positive scalars  $\sigma$  and  $\tau$  such that for all  $A \in \mathbb{C}^{m \times n}$*

$$\sigma\|A\| \leq \|\|A\|\| \leq \tau\|A\|.$$

*Proof.* The proof again builds on the fact that the supremum over a compact set is achieved and can be replaced by the maximum.

We will prove that there exists a  $\tau$  such that for all  $A \in \mathbb{C}^{m \times n}$

$$\|\|A\|\| \leq \tau\|A\|$$

leaving the rest of the proof to the reader.

Let  $A \in \mathbb{C}^{m \times n}$  be an arbitrary matrix. W.l.o.g. assume that  $A \neq 0$  (the zero matrix). Then

$$\begin{aligned} \|\|A\|\| &= \langle \text{algebra} \rangle \\ \frac{\|\|A\|\|}{\|A\|} \|A\| &\leq \langle \text{algebra} \rangle \\ \left( \sup_{Z \neq 0} \frac{\|\|Z\|\|}{\|Z\|} \right) \|A\| &= \langle \text{homogeneity} \rangle \\ \left( \sup_{Z \neq 0} \|\frac{Z}{\|Z\|}\| \right) \|A\| &= \langle \text{change of variables } B = Z/\|Z\| \rangle \\ \left( \sup_{\|B\|=1} \|\|B\|\| \right) \|A\| &= \langle \text{the set } \|B\| = 1 \text{ is compact} \rangle \\ \left( \max_{\|B\|=1} \|\|B\|\| \right) \|A\| \end{aligned}$$

The desired  $\tau$  can now be chosen to equal  $\max_{\|B\|=1} \|\|B\|\|$ . ■

**Remark 1.3.7.2** The bottom line is that, modulo a constant factor, if a matrix is "small" in one norm, it is "small" in any other norm.

**Homework 1.3.7.2** Given  $A \in \mathbb{C}^{m \times n}$  show that  $\|A\|_2 \leq \|A\|_F$ . For what matrix is equality attained?

Hmmm, actually, this is really easy to prove once we know about the SVD... Hard to prove without it. So, this problem will be moved...

**Solution.** Next week, we will learn about the SVD. Let us go ahead and insert that proof here, for future reference.

Let  $A = U\Sigma V^H$  be the Singular Value Decomposition of  $A$ , where  $U$  and  $V$  are unitary and  $\Sigma = \text{diag}(\sigma_0, \dots, \sigma_{\min(m,n)})$  with  $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{\min(m,n)} \geq 0$ . Then

$$\|A\|_2 = \|U\Sigma V^H\|_2 = \sigma_0$$

and

$$\|A\|_F = \|U\Sigma V^H\|_F = \|\Sigma\|_F = \sqrt{\sigma_0^2 + \dots + \sigma_{\min(m,n)}^2}.$$

Hence,  $\|A\|_2 \leq \|A\|_F$ .

**Homework 1.3.7.3** Let  $A \in \mathbb{C}^{m \times n}$ . The following table summarizes the equivalence of various matrix norms:

	$\ A\ _1 \leq \sqrt{m}\ A\ _2$	$\ A\ _1 \leq m\ A\ _\infty$	$\ A\ _1 \leq \sqrt{m}\ A\ _F$
$\ A\ _2 \leq \sqrt{n}\ A\ _1$		$\ A\ _2 \leq \sqrt{m}\ A\ _\infty$	$\ A\ _2 \leq \ A\ _F$
$\ A\ _\infty \leq m\ A\ _1$	$\ A\ _\infty \leq \sqrt{n}\ A\ _2$		$\ A\ _\infty \leq \sqrt{m}\ A\ _F$
$\ A\ _F \leq \sqrt{n}\ A\ _1$	$\ A\ _F \leq ?\ A\ _2$	$\ A\ _F \leq \sqrt{m}\ A\ _\infty$	

For each, prove the inequality, including that it is a tight inequality for some nonzero  $A$ .

(Skip  $\|A\|_F \leq ?\|A\|_2$ : we will revisit it in Week 2.)

**Solution.**

- $\|A\|_1 \leq \sqrt{m}\|A\|_2$ :

$$\begin{aligned}
 \|A\|_1 &= <\text{definition}> \\
 \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} &\leq <\|z\|_1 \leq \sqrt{m}\|z\|_2> \\
 \max_{x \neq 0} \frac{\sqrt{m}\|Ax\|_2}{\|x\|_1} &\leq <\|z\|_1 \geq \|z\|_2> \\
 \max_{x \neq 0} \frac{\sqrt{m}\|Ax\|_2}{\|x\|_2} &= <\text{algebra; definition}> \\
 \sqrt{m}\|A\|_2 &
 \end{aligned}$$

Equality is attained for  $A = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ .

- $\|A\|_1 \leq m\|A\|_\infty$ :

$$\begin{aligned}
 \|A\|_1 &= <\text{definition}> \\
 \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} &\leq <\|z\|_1 \leq m\|z\|_\infty> \\
 \max_{x \neq 0} \frac{m\|Ax\|_\infty}{\|x\|_1} &\leq <\|z\|_1 \geq \|z\|_\infty> \\
 \max_{x \neq 0} \frac{m\|Ax\|_\infty}{\|x\|_\infty} &= <\text{algebra; definition}> \\
 m\|A\|_\infty &
 \end{aligned}$$

Equality is attained for  $A = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ .

- $\|A\|_1 \leq \sqrt{m}\|A\|_F$ :

It pays to show that  $\|A\|_2 \leq \|A\|_F$  first. Then

$$\begin{aligned} & \|A\|_1 \\ & \leq \sqrt{m}\|A\|_2 < \text{last part} > \\ & \leq \sqrt{m}\|A\|_F < \text{some other part: } \|A\|_2 \leq \|A\|_F > \\ & \leq \sqrt{m}\|A\|_F. \end{aligned}$$

Equality is attained for  $A = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ .

- $\|A\|_2 \leq \sqrt{m}\|A\|_1$ :

$$\begin{aligned} & \|A\|_2 \\ & = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} < \text{definition} > \\ & \leq \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_2} < \|z\|_2 \leq \|z\|_1 > \\ & \leq \max_{x \neq 0} \frac{\sqrt{m}\|Ax\|_1}{\|x\|_1} < \sqrt{m}\|z\|_2 \geq \|z\|_1 > \\ & = \sqrt{m}\|A\|_1. < \text{algebra; definition} > \end{aligned}$$

Equality is attained for  $A = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$ .

- $\|A\|_2 \leq \sqrt{m}\|A\|_\infty$ :

$$\begin{aligned}
 \|A\|_2 &= <\text{definition}> \\
 \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} &\leq <\|z\|_2 \leq \sqrt{m}\|z\|_\infty> \\
 \max_{x \neq 0} \frac{\sqrt{m}\|Ax\|_\infty}{\|x\|_2} &\leq <\|z\|_2 \geq \|z\|_\infty> \\
 \max_{x \neq 0} \frac{\sqrt{m}\|Ax\|_\infty}{\|x\|_\infty} &= <\text{algebra; definition}> \\
 &= \sqrt{m}\|A\|_\infty.
 \end{aligned}$$

Equality is attained for  $A = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ .

- $\|A\|_2 \leq \|A\|_F$ :  
(See Homework 1.3.7.2, which requires the SVD, as mentioned...)
- Please share more solutions!

### 1.3.8 Submultiplicative norms



YouTube: <https://www.youtube.com/watch?v=TvthvYGt9x8>

There are a number of properties that we would like for a matrix norm to have (but not all norms do have). Recalling that we would like for a matrix norm to measure by how much a vector is "stretched," it would be good if for a given matrix norm,  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$ , there are vector norms  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  such that, for arbitrary nonzero  $x \in \mathbb{C}^n$ , the matrix norm bounds by how much the vector is stretched:

$$\frac{\|Ax\|_\mu}{\|x\|_\nu} \leq \|A\|$$

or, equivalently,

$$\|Ax\|_\mu \leq \|A\| \|x\|_\nu$$

where this second formulation has the benefit that it also holds if  $x = 0$ . When this relationship between the involved norms holds, the matrix norm is said to be subordinate to the

vector norms:

**Definition 1.3.8.1 Subordinate matrix norm.** A matrix norm  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be subordinate to vector norms  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  if, for all  $x \in \mathbb{C}^n$ ,

$$\|Ax\|_\mu \leq \|A\| \|x\|_\nu.$$

If  $\|\cdot\|_\mu$  and  $\|\cdot\|_\nu$  are the same norm (but perhaps for different  $m$  and  $n$ ), then  $\|\cdot\|$  is said to be subordinate to the given vector norm.  $\diamond$

Fortunately, all the norms that we will employ in this course are subordinate matrix norms.

**Homework 1.3.8.1** ALWAYS/SOMETIMES/NEVER: The Frobenius norm is subordinate to the vector 2-norm.

**Answer.** TRUE

Now prove it.

**Solution.** W.l.o.g., assume  $x \neq 0$ .

$$\|Ax\|_2 = \frac{\|Ax\|_2}{\|x\|_2} \|x\|_2 \leq \max_{y \neq 0} \frac{\|Ay\|_2}{\|y\|_2} \|x\|_2 = \max_{\|y\|_2=1} \|Ay\|_2 \|x\|_2 = \|A\|_2 \|x\|_2.$$

So, it suffices to show that  $\|A\|_2 \leq \|A\|_F$ . But we showed that in [Homework 1.3.7.2](#).

**Theorem 1.3.8.2** *Induced matrix norms,  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$ , are subordinate to the norms,  $\|\cdot\|_\mu$  and  $\|\cdot\|_\nu$ , that induce them.*

*Proof.* W.l.o.g. assume  $x \neq 0$ . Then

$$\|Ax\|_\mu = \frac{\|Ax\|_\mu}{\|x\|_\nu} \|x\|_\nu \leq \max_{y \neq 0} \frac{\|Ay\|_\mu}{\|y\|_\nu} \|x\|_\nu = \|A\|_{\mu,\nu} \|x\|_\nu.$$

■

**Corollary 1.3.8.3** *Any matrix  $p$ -norm is subordinate to the corresponding vector  $p$ -norm.*

Another desirable property that not all norms have is that

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

This requires the given norm to be defined for all matrix sizes..

**Definition 1.3.8.4 Consistent matrix norm.** A matrix norm  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be a consistent matrix norm if it is defined for all  $m$  and  $n$ , using the same formula for all  $m$  and  $n$ .  $\diamond$

Obviously, this definition is a bit vague. Fortunately, it is pretty clear that all the matrix norms we will use in this course, the Frobenius norm and the  $p$ -norms, are all consistently defined for all matrix sizes.

**Definition 1.3.8.5 Submultiplicative matrix norm.** A consistent matrix norm  $\|\cdot\| :$

$\mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be submultiplicative if it satisfies

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

◊

**Theorem 1.3.8.6** Let  $\|\cdot\| : \mathbb{C}^n \rightarrow \mathbb{R}$  be a vector norm defined for all  $n$ . Define the corresponding induced matrix norm as

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

Then for any  $A \in \mathbb{C}^{m \times k}$  and  $B \in \mathbb{C}^{k \times n}$  the inequality  $\|AB\| \leq \|A\|\|B\|$  holds.

In other words, induced matrix norms are submultiplicative. To prove this theorem, it helps to first prove a simpler result:

**Lemma 1.3.8.7** Let  $\|\cdot\| : \mathbb{C}^n \rightarrow \mathbb{R}$  be a vector norm defined for all  $n$  and let  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  be the matrix norm it induces. Then  $\|Ax\| \leq \|A\|\|x\|$ .

*Proof.* If  $x = 0$ , the result obviously holds since then  $\|Ax\| = 0$  and  $\|x\| = 0$ . Let  $x \neq 0$ . Then

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

Rearranging this yields  $\|Ax\| \leq \|A\|\|x\|$ . ■

We can now prove the theorem:

*Proof.*

$$\begin{aligned} & \|AB\| \\ &= < \text{definition of induced matrix norm} > \\ & \max_{\|x\|=1} \|ABx\| \\ &= < \text{associativity} > \\ & \max_{\|x\|=1} \|A(Bx)\| \\ &\leq < \text{lemma} > \\ & \max_{\|x\|=1} (\|A\|\|Bx\|) \\ &\leq < \text{lemma} > \\ & \max_{\|x\|=1} (\|A\|\|B\|\|x\|) \\ &= < \|x\| = 1 > \\ & \|A\|\|B\|. \end{aligned}$$

■

**Homework 1.3.8.2** Show that  $\|Ax\|_\mu \leq \|A\|_{\mu,\nu}\|x\|_\nu$ .

**Solution.** W.l.o.g. assume that  $x \neq 0$ .

$$\|A\|_{\mu,\nu} = \max_{y \neq 0} \frac{\|Ay\|_\mu}{\|y\|_\nu} \geq \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

Rearranging this establishes the result.

**Homework 1.3.8.3** Show that  $\|AB\|_\mu \leq \|A\|_{\mu,\nu} \|B\|_\nu$ .

**Solution.**

$$\begin{aligned}
 & \|AB\|_\mu \\
 &= < \text{definition} > \\
 & \max_{\|x\|_\nu=1} \|ABx\|_\mu \\
 &\leq < \text{last homework} > \\
 & \max_{\|x\|_\nu=1} \|A\|_{\mu,\nu} \|Bx\|_\nu \\
 &= < \text{algebra} > \\
 & \|A\|_\nu \max_{\|x\|_\nu=1} \|Bx\|_\nu \\
 &= < \text{definition} > \\
 & \|A\|_{\mu,\nu} \|B\|_\nu
 \end{aligned}$$

**Homework 1.3.8.4** Show that the Frobenius norm,  $\|\cdot\|_F$ , is submultiplicative.

**Solution.**

$$\begin{aligned}
 & \|AB\|_F^2 \\
 &= < \text{partition} > \\
 & \left\| \begin{pmatrix} \tilde{a}_0^H \\ \tilde{a}_1^H \\ \vdots \\ \tilde{a}_{m-1}^H \end{pmatrix} \left( b_0 \mid b_1 \mid \cdots \mid b_{n-1} \right) \right\|_F^2 \\
 &= < \text{partitioned matrix-matrix multiplication} > \\
 & \left\| \begin{pmatrix} \tilde{a}_0^H b_0 & \tilde{a}_0^H b_1 & \cdots & \tilde{a}_0^H b_{n-1} \\ \hline \tilde{a}_0^H b_0 & \tilde{a}_0^H b_1 & \cdots & \tilde{a}_0^H b_{n-1} \\ \hline \vdots & \vdots & & \vdots \\ \hline \tilde{a}_{m-1}^H b_0 & \tilde{a}_{m-1}^H b_1 & \cdots & \tilde{a}_{m-1}^H b_{n-1} \end{pmatrix} \right\|_F^2 \\
 &= < \text{definition of Frobenius norm} > \\
 & \sum_i \sum_j |\tilde{a}_i^H b_j|^2 \\
 &\leq < \text{Cauchy-Schwartz inequality} > \\
 & \sum_i \sum_j \|\tilde{a}_i\|_2^2 \|b_j\|_2^2 \\
 &= < \text{algebra} > \\
 & (\sum_i \|\tilde{a}_i\|_2^2) (\sum_j \|b_j\|_2^2) \\
 &= < \text{previous observations about the Frobenius norm} > \\
 & \|A\|_F^2 \|B\|_F^2
 \end{aligned}$$

Hence  $\|AB\|_F^2 \leq \|A\|_F^2 \|B\|_F^2$ . Taking the square root of both sides leaves us with  $\|AB\|_F \leq \|A\|_F \|B\|_F$ .

**Homework 1.3.8.5** For  $A \in \mathbb{C}^{m \times n}$  define

$$\|A\| = \max_{i=0}^{m-1} \max_{j=0}^{n-1} |\alpha_{i,j}|.$$

1. TRUE/FALSE: This is a norm.

2. TRUE/FALSE: This is a consistent norm.

**Answer.**

1. TRUE
2. TRUE

**Solution.**

1. This is a norm. You can prove this by checking the three conditions.
2. It is a consistent norm since it is defined for all  $m$  and  $n$ .

**Remark 1.3.8.8** The important take-away: The norms we tend to use in this course, the  $p$ -norms and the Frobenius norm, are all submultiplicative.

**Homework 1.3.8.6** Let  $A \in \mathbb{C}^{m \times n}$ .

ALWAYS/SOMETIMES/NEVER: There exists a vector

$$x = \begin{pmatrix} \chi_0 \\ \vdots \\ \chi_{m-1} \end{pmatrix} \text{ with } |\chi_i| = 1 \text{ for } i = 0, \dots, n-1$$

such that  $\|A\|_\infty = \|Ax\|_\infty$ .

**Answer.** ALWAYS

Now prove it!

**Solution.** Partition  $A$  by rows:

$$A = \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix}.$$

We know that there exists  $k$  such that  $\|\tilde{a}_k\|_1 = \|A\|_\infty$ . Now

$$\begin{aligned} \|\tilde{a}_k\|_1 &= <\text{definition of 1-norm}> \\ &= |\alpha_{k,0}| + \dots + |\alpha_{k,n-1}| \\ &= <\text{algebra}> \\ &= \frac{|\alpha_{k,0}|}{\alpha_{k,0}} \alpha_{k,0} + \dots + \frac{|\alpha_{k,n-1}|}{\alpha_{k,n-1}} \alpha_{k,n-1}. \end{aligned}$$

where we take  $\frac{|\alpha_{k,j}|}{\alpha_{k,j}} = 1$  whenever  $\alpha_{k,j} = 0$ . Vector

$$x = \begin{pmatrix} \frac{|\alpha_{k,0}|}{\alpha_{k,0}} \\ \vdots \\ \frac{|\alpha_{k,n-1}|}{\alpha_{k,n-1}} \end{pmatrix}$$

has the desired property.

### 1.3.9 Summary



YouTube: <https://www.youtube.com/watch?v=DyoT2tJhxIs>

## 1.4 Condition Number of a Matrix

### 1.4.1 Conditioning of a linear system



YouTube: <https://www.youtube.com/watch?v=QwFQNAPKIwk>

A question we will run into later in the course asks how accurate we can expect the solution of a linear system to be if the right-hand side of the system has error in it.

Formally, this can be stated as follows: We wish to solve  $Ax = b$ , where  $A \in \mathbb{C}^{m \times m}$  but the right-hand side has been perturbed by a small vector so that it becomes  $b + \delta b$ .

**Remark 1.4.1.1** Notice how the  $\delta$  touches the  $b$ . This is meant to convey that this is a symbol that represents a vector rather than the vector  $b$  that is multiplied by a scalar  $\delta$ .

The question now is how a relative error in  $b$  is amplified into a relative error in the solution  $x$ .

This is summarized as follows:

$$\begin{array}{lll} Ax & = & b \\ A(x + \delta x) & = & b + \delta b \end{array} \quad \begin{array}{l} \text{exact equation} \\ \text{perturbed equation} \end{array}$$

We would like to determine a formula,  $\kappa(A, b, \delta b)$ , that gives us a bound on how much a relative error in  $b$  is potentially amplified into a relative error in the solution  $x$ :

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A, b, \delta b) \frac{\|\delta b\|}{\|b\|}.$$

We assume that  $A$  has an inverse since otherwise there may be no solution or there may be an infinite number of solutions. To find an expression for  $\kappa(A, b, \delta b)$ , we notice that

$$\begin{array}{rcl} Ax + A\delta x & = & b + \delta b \\ Ax & = & b \\ \hline A\delta x & = & \delta b \end{array}$$

and from this we deduce that

$$\begin{array}{rcl} Ax & = & b \\ \delta x & = & A^{-1}\delta b. \end{array}$$

If we now use a vector norm  $\|\cdot\|$  and its induced matrix norm  $\|\cdot\|$ , then

$$\begin{array}{rcl} \|b\| & = & \|Ax\| \leq \|A\|\|x\| \\ \|\delta x\| & = & \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\| \end{array}$$

since induced matrix norms are subordinate.

From this we conclude that

$$\frac{1}{\|x\|} \leq \|A\| \frac{1}{\|b\|}$$

and

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

so that

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

Thus, the desired expression  $\kappa(A, b, \delta b)$  doesn't depend on anything but the matrix  $A$ :

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\|A\|\|A^{-1}\|}_{\kappa(A)} \frac{\|\delta b\|}{\|b\|}.$$

**Definition 1.4.1.2 Condition number of a nonsingular matrix.** The value  $\kappa(A) = \|A\|\|A^{-1}\|$  is called the condition number of a nonsingular matrix  $A$ .  $\diamond$

A question becomes whether this is a pessimistic result or whether there are examples of  $b$  and  $\delta b$  for which the relative error in  $b$  is amplified by exactly  $\kappa(A)$ . The answer is that, unfortunately, the bound is tight.

- There is an  $\hat{x}$  for which

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \|A\hat{x}\|,$$

namely the  $x$  for which the maximum is attained. This is the direction of maximal magnification. Pick  $\hat{b} = A\hat{x}$ .

- There is an  $\hat{\delta}b$  for which

$$\|A^{-1}\| = \max_{\|x\|\neq 0} \frac{\|A^{-1}x\|}{\|x\|} = \frac{\|A^{-1}\hat{\delta}b\|}{\|\hat{\delta}b\|},$$

again, the  $x$  for which the maximum is attained.

It is when solving the perturbed system

$$A(x + \delta x) = \hat{b} + \hat{\delta b}$$

that the maximal magnification by  $\kappa(A)$  is observed.

**Homework 1.4.1.1** Let  $\|\cdot\|$  be a vector norm and corresponding induced matrix norm.

TRUE/FALSE:  $\|I\| = 1$ .

**Answer.** TRUE

**Solution.**

$$\|I\| = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1$$

**Homework 1.4.1.2** Let  $\|\cdot\|$  be a vector norm and corresponding induced matrix norm, and  $A$  a nonsingular matrix.

TRUE/FALSE:  $\kappa(A) = \|A\| \|A^{-1}\| \geq 1$ .

**Answer.** TRUE

**Solution.**

$$\begin{aligned} 1 &= <\text{last homework}> \\ \|I\| &= < A \text{ is invertible} > \\ \|AA^{-1}\| &\leq <\|\cdot\| \text{ is submultiplicative}> \\ \|A\| \|A^{-1}\|. \end{aligned}$$

**Remark 1.4.1.3** This last exercise shows that there will always be choices for  $b$  and  $\delta b$  for which the relative error is at best directly translated into an equal relative error in the solution (if  $\kappa(A) = 1$ ).

## 1.4.2 Loss of digits of accuracy



YouTube: <https://www.youtube.com/watch?v=-5l90v5RXYo>

**Homework 1.4.2.1** Let  $\alpha = -14.24123$  and  $\hat{\alpha} = -14.24723$ . Compute

- $|\alpha| =$
- $|\alpha - \hat{\alpha}| =$

- $\frac{|\alpha - \hat{\alpha}|}{|\alpha|} =$
- $\log_{10} \left( \frac{|\alpha - \hat{\alpha}|}{|\alpha|} \right) =$

**Solution.** Let  $\alpha = -14.24123$  and  $\hat{\alpha} = -14.24723$ . Compute

- $|\alpha| = 14.24123$
- $|\alpha - \hat{\alpha}| = 0.006$
- $\frac{|\alpha - \hat{\alpha}|}{|\alpha|} \approx 0.00042$
- $\log_{10} \left( \frac{|\alpha - \hat{\alpha}|}{|\alpha|} \right) \approx -3.4$

The point of this exercise is as follows:

- If you compare  $\alpha = -14.24123$  and  $\hat{\alpha} = -14.24723$  and you consider  $\hat{\alpha}$  to be an approximation of  $\alpha$ , then  $\hat{\alpha}$  is accurate to four digits:  $-14.24$  is accurate.
- Computing  $\log_{10} \left( \frac{|\alpha - \hat{\alpha}|}{|\alpha|} \right)$  tells you approximately how many decimal digits are accurate: 3.4 digits.

Be sure to read the solution to the last homework!

### 1.4.3 The conditioning of an upper triangular matrix



YouTube: <https://www.youtube.com/watch?v=LGBFyjhjt6U>

We now revisit the material from the launch for the semester. We understand that when solving  $Lx = b$ , even a small relative change to the right-hand side  $b$  can amplify into a large relative change in the solution  $\hat{x}$  if the condition number of the matrix is large.

**Homework 1.4.3.1** Change the script [Assignments/Week01/matlab/Test\\_Upper\\_triangular\\_solve\\_100.m](#) to also compute the condition number of matrix  $U$ ,  $\kappa(U)$ . Investigate what happens to the condition number as you change the problem size  $n$ .

Since in the example the upper triangular matrix is generated to have random values as its entries, chances are that at least one element on its diagonal is small. If that element were zero, then the triangular matrix would be singular. Even if it is not exactly zero, the condition number of  $U$  becomes very large if the element is small.

## 1.5 Enrichments

### 1.5.1 Condition number estimation

It has been observed that high-quality numerical software should not only provide routines for solving a given problem, but, when possible, should also (optionally) provide the user with feedback on the conditioning (sensitivity to changes in the input) of the problem. In this enrichment, we relate this to what you have learned this week.

Given a vector norm  $\|\cdot\|$  and induced matrix norm  $\|\cdot\|$ , the condition number of matrix  $A$  using that norm is given by  $\kappa(A) = \|A\| \|A^{-1}\|$ . When trying to practically compute the condition number, this leads to two issues:

- Which norm should we use? A case has been made in this week that the 1-norm and  $\infty$ -norm are candidates since they are easy and cheap to compute.
- It appears that  $A^{-1}$  needs to be computed. We will see in future weeks that this is costly:  $O(m^3)$  computation when  $A$  is  $m \times m$ . This is generally considered to be expensive.

This leads to the question "Can a reliable estimate of the condition number be cheaply computed?" In this unit, we give a glimpse of how this can be achieved and then point the interested learner to related papers.

Partition  $m \times m$  matrix  $A$ :

$$A = \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix}.$$

We recall that

- The  $\infty$ -norm is defined by

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty.$$

- From [Homework 1.3.6.2](#), we know that the  $\infty$ -norm can be practically computed as

$$\|A\|_\infty = \max_{0 \leq i < m} \|\tilde{a}_i\|_1,$$

where  $\tilde{a}_i = (\tilde{a}_i^T)^T$ . This means that  $\|A\|_\infty$  can be computed in  $O(m^2)$  operations.

- From the solution to [Homework 1.3.6.2](#), we know that there is a vector  $x$  with  $|\chi_i| = 1$ ,  $0 \leq i < m$ , such that  $\|A\|_\infty = \|Ax\|_1$ . This  $x$  satisfies  $\|x\|_\infty = 1$ .

From this we conclude that

$$\|A\|_\infty = \max_{x \in \mathcal{S}} \|Ax\|_\infty,$$

where  $\mathcal{S}$  is the set of all vectors  $x$  with  $|\chi_i| = 1$ ,  $0 \leq i < m$ .

We will illustrate the techniques that underly efficient condition number estimation by looking at the simpler case where we wish to estimate the condition number of a *real-valued*

nonsingular upper triangular  $m \times m$  matrix  $U$ , using the  $\infty$ -norm. Since  $U$  is real-valued,  $|\chi_i| = 1$  means  $\chi_i = \pm 1$ . The problem is that it appears we must compute  $\|U^{-1}\|_\infty$ . Computing  $U^{-1}$  when  $U$  is dense requires  $O(m^3)$  operations (a topic we won't touch on until much later in the course).

Our observations tell us that

$$\|U^{-1}\|_\infty = \max_{x \in \mathcal{S}} \|U^{-1}x\|_\infty,$$

where  $\mathcal{S}$  is the set of all vectors  $x$  with elements  $\chi_i \in \{-1, 1\}$ . This is equivalent to

$$\|U^{-1}\|_\infty = \max_{z \in \mathcal{T}} \|z\|_\infty,$$

where  $\mathcal{T}$  is the set of all vectors  $z$  that satisfy  $Uz = y$  for some  $y$  with elements  $\psi_i \in \{-1, 1\}$ . So, we could solve  $Uz = y$  for all vectors  $y \in \mathcal{S}$ , compute the  $\infty$ -norm for all those vectors  $z$ , and pick the maximum of those values. But that is not practical.

One simple solution is to try to construct a vector  $y$  that results in a large amplification (in the  $\infty$ -norm) when solving  $Uz = y$ , and to then use that amplification as an estimate for  $\|U^{-1}\|_\infty$ . So how do we do this? Consider

$$\underbrace{\begin{pmatrix} \vdots & \ddots & & \vdots & \\ 0 & \cdots & v_{m-2,m-2} & v_{m-2,m-1} & \\ 0 & \cdots & 0 & v_{m-1,m-1} & \end{pmatrix}}_U \underbrace{\begin{pmatrix} \vdots \\ \zeta_{m-2} \\ \zeta_{m-1} \end{pmatrix}}_z = \underbrace{\begin{pmatrix} \vdots \\ \psi_{m-2} \\ \psi_{m-1} \end{pmatrix}}_y.$$

Here is a *heuristic* for picking  $y \in \mathcal{S}$ :

- We want to pick  $\psi_{m-1} \in \{0, 1\}$  in order to construct a vector  $y \in \mathcal{S}$ . We can pick  $\psi_{m-1} = 1$  since picking it equal to  $-1$  will simply carry through negation in the appropriate way in the scheme we are describing.

From this  $\psi_{m-1}$  we can compute  $\zeta_{m-1}$ .

- Now,

$$v_{m-2,m-2}\zeta_{m-2} + v_{1,2}\zeta_{m-1} = \psi_{m-2}$$

where  $\zeta_{m-1}$  is known and  $\psi_{m-2}$  can be strategically chosen. We want to  $z$  to have a large  $\infty$ -norm and hence a *heuristic* is to now pick  $\psi_{m-2} \in \{-1, 1\}$  in such a way that  $\zeta_{m-2}$  is as large as possible in magnitude.

With this  $\psi_{m-2}$  we can compute  $\zeta_{m-2}$ .

- And so forth!

When done, the magnification equals  $\|z\|_\infty = |\zeta_k|$ , where  $\zeta_k$  is the element of  $z$  with largest magnitude. This approach provides an estimate for  $\|U^{-1}\|_\infty$  with  $O(m^2)$  operations.

The described method underlies the condition number estimator for LINPACK, developed in the 1970s [11], as described in [7]:

- A.K. Cline, C.B. Moler, G.W. Stewart, and J.H. Wilkinson, An estimate for the condition number of a matrix, SIAM J. Numer. Anal., 16 (1979).

The method discussed in that paper yields a lower bound on  $\|A^{-1}\|_\infty$  and with that on  $\kappa_\infty(A)$ .

**Remark 1.5.1.1** Alan Cline has his office on our floor at UT-Austin. G.W. (Pete) Stewart was Robert's Ph.D. advisor. Cleve Moler is the inventor of Matlab. John Wilkinson received the Turing Award for his contributions to numerical linear algebra.

More sophisticated methods are discussed in [15]:

- N. Higham, A Survey of Condition Number Estimates for Triangular Matrices, SIAM Review, 1987.

His methods underlie the LAPACK [1] condition number estimator and are remarkably accurate: most of the time they provide an almost exact estimate of the actual condition number.

## 1.6 Wrap Up

### 1.6.1 Additional homework

**Homework 1.6.1.1** For  $e_j \in \mathbb{R}^n$  (a standard basis vector), compute

- $\|e_j\|_2 =$
- $\|e_j\|_1 =$
- $\|e_j\|_\infty =$
- $\|e_j\|_p =$

**Homework 1.6.1.2** For  $I \in \mathbb{R}^{n \times n}$  (the identity matrix), compute

- $\|I\|_1 =$
- $\|I\|_\infty =$
- $\|I\|_2 =$
- $\|I\|_p =$
- $\|I\|_F =$

**Homework 1.6.1.3** Let  $D = \begin{pmatrix} \delta_0 & 0 & \cdots & 0 \\ 0 & \delta_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \delta_{n-1} \end{pmatrix}$  (a diagonal matrix). Compute

- $\|D\|_1 =$

- $\|D\|_\infty =$
- $\|D\|_p =$
- $\|D\|_F =$

**Homework 1.6.1.4** Let  $x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}$  and  $1 \leq p < \infty$  or  $p = \infty$ .

ALWAYS/SOMETIMES/NEVER:  $\|x_i\|_p \leq \|x\|_p$ .

**Homework 1.6.1.5** For

$$A = \begin{pmatrix} 1 & 2 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

compute

- $\|A\|_1 =$
- $\|A\|_\infty =$
- $\|A\|_F =$

**Homework 1.6.1.6** For  $A \in \mathbb{C}^{m \times n}$  define

$$\|A\| = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}| = \sum \begin{pmatrix} |\alpha_{0,0}|, & \cdots, & |\alpha_{0,n-1}|, \\ \vdots & & \vdots \\ |\alpha_{m-1,0}|, & \cdots, & |\alpha_{m-1,n-1}| \end{pmatrix}.$$

- TRUE/FALSE: This function is a matrix norm.
- How can you relate this norm to the vector 1-norm?
- TRUE/FALSE: For this norm,  $\|A\| = \|A^H\|$ .
- TRUE/FALSE: This norm is submultiplicative.

**Homework 1.6.1.7** Let  $A \in \mathbb{C}^{m \times n}$ . Partition

$$A = \left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right) = \begin{pmatrix} \tilde{a}_0^T \\ \tilde{a}_1^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix}.$$

Prove that

- $\|A\|_F = \|A^T\|_F$ .
- $\|A\|_F = \sqrt{\|a_0\|_2^2 + \|a_1\|_2^2 + \cdots + \|a_{n-1}\|_2^2}$ .

- $\|A\|_F = \sqrt{\|\tilde{a}_0\|_2^2 + \|\tilde{a}_1\|_2^2 + \cdots + \|\tilde{a}_{m-1}\|_2^2}$ .

Note that here  $\tilde{a}_i = (\tilde{a}_i^T)^T$ .

**Homework 1.6.1.8** Let  $x \in \mathbb{R}^m$  with  $\|x\|_1 = 1$ .

TRUE/FALSE:  $\|x\|_2 = 1$  if and only if  $x = \pm e_j$  for some  $j$ .

**Solution.** Obviously, if  $x = e_j$  then  $\|x\|_1 = \|x\|_2 = 1$ .

Assume  $x \neq e_j$ . Then  $|\chi_i| < 1$  for all  $i$ . But then  $\|x\|_2 = \sqrt{|\chi_0|^2 + \cdots + |\chi_{m-1}|^2} < \sqrt{|\chi_0| + \cdots + |\chi_{m-1}|} = \sqrt{1} = 1$ .

**Homework 1.6.1.9** Prove that if  $\|x\|_\nu \leq \beta \|x\|_\mu$  is true for all  $x$ , then  $\|A\|_\nu \leq \beta \|A\|_{\mu,\nu}$ .

## 1.6.2 Summary

If  $\alpha, \beta \in \mathbb{C}$  with  $\alpha = \alpha_r + \alpha_c i$  and  $\beta = \beta_r + i\beta_c$ , where  $\alpha_r, \alpha_c, \beta_r, \beta_c \in \mathbb{R}$ , then

- Conjugate:  $\bar{\alpha} = \alpha_r - \alpha_c i$ .
- Product:  $\alpha\beta = (\alpha_r\beta_r - \alpha_c\beta_c) + (\alpha_r\beta_c + \alpha_c\beta_r)i$ .
- Absolute value:  $|\alpha| = \sqrt{\alpha_r^2 + \alpha_c^2} = \sqrt{\alpha\bar{\alpha}}$ .

Let  $x, y \in \mathbb{C}^m$  with  $x = \begin{pmatrix} \chi_0 \\ \vdots \\ \chi_{m-1} \end{pmatrix}$  and  $y = \begin{pmatrix} \psi_0 \\ \vdots \\ \psi_{m-1} \end{pmatrix}$ . Then

- Conjugate:

$$\bar{x} = \begin{pmatrix} \bar{\chi}_0 \\ \vdots \\ \bar{\chi}_{m-1} \end{pmatrix}.$$

- Transpose of vector:

$$x^T = \begin{pmatrix} \chi_0 & \cdots & \chi_{m-1} \end{pmatrix}$$

- Hermitian transpose (conjugate transpose) of vector:

$$x^H = \bar{x}^T = \bar{x}^T = \begin{pmatrix} \bar{\chi}_0 & \cdots & \bar{\chi}_{m-1} \end{pmatrix}.$$

- Dot product (inner product):  $x^H y = \bar{x}^T y = \bar{x}^T y = \bar{\chi}_0 \psi_0 + \cdots + \bar{\chi}_{m-1} \psi_{m-1} = \sum_{i=0}^{m-1} \bar{\chi}_i \psi_i$ .

**Definition 1.6.2.1 Vector norm.** Let  $\|\cdot\| : \mathbb{C}^m \rightarrow \mathbb{R}$ . Then  $\|\cdot\|$  is a (vector) norm if for all  $x, y \in \mathbb{C}^m$  and all  $\alpha \in \mathbb{C}$

- $x \neq 0 \Rightarrow \|x\| > 0$  ( $\|\cdot\|$  is positive definite),
- $\|\alpha x\| = |\alpha| \|x\|$  ( $\|\cdot\|$  is homogeneous), and

- $\|x + y\| \leq \|x\| + \|y\|$  ( $\|\cdot\|$  obeys the triangle inequality).

◊

- 2-norm (Euclidean length):  $\|x\|_2 = \sqrt{x^H x} = \sqrt{|\chi_0|^2 + \dots + |\chi_{m-1}|^2} = \sqrt{\bar{\chi}_0 \chi_0 + \dots + \bar{\chi}_{m-1} \chi_{m-1}} = \sqrt{\sum_{i=0}^{m-1} |\chi_i|^2}$ .
- $p$ -norm:  $\|x\|_p = \sqrt[p]{|\chi_0|^p + \dots + |\chi_{m-1}|^p} = \sqrt[p]{\sum_{i=0}^{m-1} |\chi_i|^p}$ .
- 1-norm:  $\|x\|_1 = |\chi_0| + \dots + |\chi_{m-1}| = \sum_{i=0}^{m-1} |\chi_i|$ .
- $\infty$ -norm:  $\|x\|_\infty = \max(|\chi_0|, \dots, |\chi_{m-1}|) = \max_{i=0}^{m-1} |\chi_i| = \lim_{p \rightarrow \infty} \|x\|_p$ .
- Unit ball: Set of all vectors with norm equal to one. Notation:  $\|x\| = 1$ .

**Theorem 1.6.2.2 Equivalence of vector norms.** Let  $\|\cdot\| : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\|\cdot\|\| : \mathbb{C}^m \rightarrow \mathbb{R}$  both be vector norms. Then there exist positive scalars  $\sigma$  and  $\tau$  such that for all  $x \in \mathbb{C}^m$

$$\begin{array}{c} \sigma\|x\| \leq \|\|x\|\| \leq \tau\|x\|. \\ \hline \begin{array}{c|c} \|\chi_0\|_1 \leq \sqrt{m}\|\chi_0\|_2 & \|\chi_0\|_1 \leq m\|\chi_0\|_\infty \\ \hline \|\chi_0\|_2 \leq \|\chi_0\|_1 & \|\chi_0\|_2 \leq \sqrt{m}\|\chi_0\|_\infty \\ \hline \|\chi_0\|_\infty \leq \|\chi_0\|_1 & \|\chi_0\|_\infty \leq \|\chi_0\|_2 \end{array} \end{array}$$

**Definition 1.6.2.3 Linear transformations and matrices.** Let  $L : \mathbb{C}^n \rightarrow \mathbb{C}^m$ . Then  $L$  is said to be a linear transformation if for all  $\alpha \in \mathbb{C}$  and  $x, y \in \mathbb{C}^n$

- $L(\alpha x) = \alpha L(x)$ . That is, scaling first and then transforming yields the same result as transforming first and then scaling.
- $L(x + y) = L(x) + L(y)$ . That is, adding first and then transforming yields the same result as transforming first and then adding.

◊

**Definition 1.6.2.4 Standard basis vector.** In this course, we will use  $e_j \in \mathbb{C}^m$  to denote the standard basis vector with a "1" in the position indexed with  $j$ . So,

$$e_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j$$

◊

If  $L$  is a linear transformation and we let  $a_j = L(e_j)$  then

$$A = \left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right)$$

is the matrix that represents  $L$  in the sense that  $Ax = L(x)$ .

Partition  $C$ ,  $A$ , and  $B$  by rows and columns

$$C = \left( \begin{array}{c|c|c|c} c_0 & \cdots & c_{n-1} \end{array} \right) = \left( \begin{array}{c} \tilde{c}_0^T \\ \vdots \\ \tilde{c}_{m-1}^T \end{array} \right), A = \left( \begin{array}{c|c|c|c} a_0 & \cdots & a_{k-1} \end{array} \right) = \left( \begin{array}{c} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{array} \right),$$

and

$$B = \left( \begin{array}{c|c|c|c} b_0 & \cdots & b_{n-1} \end{array} \right) = \left( \begin{array}{c} \tilde{b}_0^T \\ \vdots \\ \tilde{b}_{k-1}^T \end{array} \right),$$

then  $C := AB$  can be computed in the following ways:

1. By columns:

$$\left( \begin{array}{c|c|c|c} c_0 & \cdots & c_{n-1} \end{array} \right) := A \left( \begin{array}{c|c|c|c} b_0 & \cdots & b_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c|c} Ab_0 & \cdots & Ab_{n-1} \end{array} \right).$$

In other words,  $c_j := Ab_j$  for all columns of  $C$ .

2. By rows:

$$\left( \begin{array}{c} \tilde{c}_0^T \\ \vdots \\ \tilde{c}_{m-1}^T \end{array} \right) := \left( \begin{array}{c} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{array} \right) B = \left( \begin{array}{c} \tilde{a}_0^T B \\ \vdots \\ \tilde{a}_{m-1}^T B \end{array} \right).$$

In other words,  $\tilde{c}_i^T = \tilde{a}_i^T B$  for all rows of  $C$ .

3. As the sum of outer products:

$$C := \left( \begin{array}{c|c|c|c} a_0 & \cdots & a_{k-1} \end{array} \right) \left( \begin{array}{c} \tilde{b}_0^T \\ \vdots \\ \tilde{b}_{k-1}^T \end{array} \right) = a_0 \tilde{b}_0^T + \cdots + a_{k-1} \tilde{b}_{k-1}^T,$$

which should be thought of as a sequence of rank-1 updates, since each term is an outer product and an outer product has rank of at most one.

Partition  $C$ ,  $A$ , and  $B$  by blocks (submatrices),

$$C = \left( \begin{array}{c|c|c} C_{0,0} & \cdots & C_{0,N-1} \\ \vdots & & \vdots \\ C_{M-1,0} & \cdots & C_{M-1,N-1} \end{array} \right), \left( \begin{array}{c|c|c} A_{0,0} & \cdots & A_{0,K-1} \\ \vdots & & \vdots \\ A_{M-1,0} & \cdots & A_{M-1,K-1} \end{array} \right),$$

and

$$\left( \begin{array}{c|c|c} B_{0,0} & \cdots & B_{0,N-1} \\ \hline \vdots & & \vdots \\ \hline B_{K-1,0} & \cdots & B_{K-1,N-1} \end{array} \right),$$

where the partitionings are "conformal." Then

$$C_{i,j} = \sum_{p=0}^{K-1} A_{i,p} B_{p,j}.$$

**Definition 1.6.2.5 Matrix norm.** Let  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$ . Then  $\|\cdot\|$  is a (matrix) norm if for all  $A, B \in \mathbb{C}^{m \times n}$  and all  $\alpha \in \mathbb{C}$

- $A \neq 0 \Rightarrow \|A\| > 0$  ( $\|\cdot\|$  is positive definite),
- $\|\alpha A\| = |\alpha| \|A\|$  ( $\|\cdot\|$  is homogeneous), and
- $\|A + B\| \leq \|A\| + \|B\|$  ( $\|\cdot\|$  obeys the triangle inequality).

◇

Let  $A \in \mathbb{C}^{m \times n}$  and

$$A = \begin{pmatrix} \alpha_{0,0} & \cdots & \alpha_{0,n-1} \\ \vdots & & \vdots \\ \alpha_{m-1,0} & \cdots & \alpha_{m-1,n-1} \end{pmatrix} = \left( \begin{array}{c|c|c} a_0 & \cdots & a_{n-1} \end{array} \right) = \begin{pmatrix} \tilde{a}_0^T \\ \vdots \\ \tilde{a}_{m-1}^T \end{pmatrix}.$$

Then

- Conjugate of matrix:

$$\bar{A} = \begin{pmatrix} \bar{\alpha}_{0,0} & \cdots & \bar{\alpha}_{0,n-1} \\ \vdots & & \vdots \\ \bar{\alpha}_{m-1,0} & \cdots & \bar{\alpha}_{m-1,n-1} \end{pmatrix}.$$

- Transpose of matrix:

$$A^T = \begin{pmatrix} \alpha_{0,0} & \cdots & \alpha_{m-1,0} \\ \vdots & & \vdots \\ \alpha_{0,n-1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}.$$

- Conjugate transpose (Hermitian transpose) of matrix:

$$A^H = \bar{A}^T = \bar{A}^T = \begin{pmatrix} \bar{\alpha}_{0,0} & \cdots & \bar{\alpha}_{m-1,0} \\ \vdots & & \vdots \\ \bar{\alpha}_{0,n-1} & \cdots & \bar{\alpha}_{m-1,n-1} \end{pmatrix}.$$

- Frobenius norm:  $\|A\|_F = \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2} = \sqrt{\sum_{j=0}^{n-1} \|a_j\|_2^2} = \sqrt{\sum_{i=0}^{m-1} \|\tilde{a}_i\|_2^2}$

- matrix p-norm:  $\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p.$
- matrix 2-norm:  $\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2 = \|A^H\|_2.$
- matrix 1-norm:  $\|A\|_1 = \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{0 \leq j < n} \|a_j\|_1 = \|A^H\|_\infty.$
- matrix  $\infty$ -norm:  $\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{0 \leq j < n} \|a_j\|_\infty = \|A^H\|_1.$

**Theorem 1.6.2.6 Equivalence of matrix norms.** Let  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  and  $\|\|\cdot\|\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  both be matrix norms. Then there exist positive scalars  $\sigma$  and  $\tau$  such that for all  $A \in \mathbb{C}^{m \times n}$

$$\begin{array}{c} \sigma\|A\| \leq \|\|A\|\| \leq \tau\|A\|. \\ \hline \hline \begin{array}{|c|c|c|} \hline & \|A\|_1 \leq \sqrt{m}\|A\|_2 & \|A\|_1 \leq m\|A\|_\infty \\ \hline \|A\|_2 \leq \sqrt{m}\|A\|_1 & \|A\|_2 \leq \sqrt{m}\|A\|_\infty & \|A\|_2 \leq \|A\|_F \\ \hline \|A\|_\infty \leq m\|A\|_1 & \|A\|_\infty \leq \sqrt{m}\|A\|_2 & \|A\|_\infty \leq \sqrt{m}\|A\|_F \\ \hline \|A\|_F \leq \sqrt{m}\|A\|_1 & \|A\|_F \leq \sqrt{m}\|A\|_2 & \|A\|_F \leq \sqrt{m}\|A\|_\infty \\ \hline \end{array} \end{array}$$

**Definition 1.6.2.7 Subordinate matrix norm.** A matrix norm  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be subordinate to vector norms  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  if, for all  $x \in \mathbb{C}^n$ ,

$$\|Ax\|_\mu \leq \|A\| \|x\|_\nu.$$

If  $\|\cdot\|_\mu$  and  $\|\cdot\|_\nu$  are the same norm (but perhaps for different  $m$  and  $n$ ), then  $\|\cdot\|$  is said to be subordinate to the given vector norm.  $\diamond$

**Definition 1.6.2.8 Consistent matrix norm.** A matrix norm  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be a consistent matrix norm if it is defined for all  $m$  and  $n$ , using the same formula for all  $m$  and  $n$ .  $\diamond$

**Definition 1.6.2.9 Submultiplicative matrix norm.** A consistent matrix norm  $\|\cdot\| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be submultiplicative if it satisfies

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

$\diamond$

Let  $A, \Delta A \in \mathbb{C}^{m \times m}$ ,  $x, \delta x, b, \delta b \in \mathbb{C}^m$ ,  $A$  be nonsingular, and  $\|\cdot\|$  be a vector norm and corresponding subordinate matrix norm. Then

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\|A\| \|A^{-1}\|}_{\kappa(A)} \frac{\|\delta b\|}{\|b\|}.$$

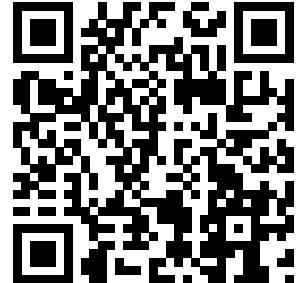
**Definition 1.6.2.10 Condition number of a nonsingular matrix.** The value  $\kappa(A) = \|A\| \|A^{-1}\|$  is called the condition number of a nonsingular matrix  $A$ .  $\diamond$

## Week 2

# The Singular Value Decomposition

## 2.1 Opening Remarks

### 2.1.1 Low rank approximation



YouTube: <https://www.youtube.com/watch?v=12K5aydB9cQ>

Consider this picture of the Gates Dell Complex that houses our Department of Computer Science:



It consists of an  $m \times n$  array of pixels, each of which is a numerical value. Think of the  $j$ th column of pixels as a vector of values,  $b_j$ , so that the whole picture is represented by columns as

$$B = ( b_0 \mid b_1 \mid \cdots \mid b_{n-1} ),$$

where we recognize that we can view the picture as a matrix. What if we want to store this picture with fewer than  $m \times n$  data? In other words, what if we want to compress the picture? To do so, we might identify a few of the columns in the picture to be the "chosen ones" that are representative of the other columns in the following sense: All columns in the picture are approximately linear combinations of these chosen columns.

Let's let linear algebra do the heavy lifting: what if we choose  $k$  roughly equally spaced columns in the picture:

$$\begin{aligned} a_0 &= b_0 \\ a_1 &= b_{n/k-1} \\ \vdots &\quad \vdots \\ a_{k-1} &= b_{(k-1)n/k-1}, \end{aligned}$$

where for illustration purposes we assume that  $n$  is an integer multiple of  $k$ . (We could instead choose them randomly or via some other method. This detail is not important as we try to gain initial insight.) We could then approximate each column of the picture,  $b_j$ , as a linear combination of  $a_0, \dots, a_{k-1}$ :

$$b_j \approx \chi_{0,j}a_0 + \chi_{1,j}a_1 + \cdots + \chi_{k-1,j}a_{k-1} = \left( \begin{array}{c|c|c} a_0 & \cdots & a_{k-1} \end{array} \right) \begin{pmatrix} \frac{\chi_{0,j}}{\chi_{k-1,j}} \\ \vdots \end{pmatrix}.$$

We can write this more concisely by viewing these chosen columns as the columns of matrix  $A$  so that

$$b_j \approx Ax_j, \quad \text{where } A = \left( \begin{array}{c|c|c} a_0 & \cdots & a_{k-1} \end{array} \right) \text{ and } x_j = \begin{pmatrix} \frac{\chi_{0,j}}{\chi_{k-1,j}} \\ \vdots \end{pmatrix}.$$

If  $A$  has linearly independent columns, the best such approximation (in the linear least squares sense) is obtained by choosing

$$x_j = (A^T A)^{-1} A^T b_j,$$

where you may recognize  $(A^T A)^{-1} A^T$  as the (left) pseudo-inverse of  $A$ , leaving us with

$$b_j \approx A(A^T A)^{-1} A^T b_j.$$

This approximates  $b_j$  with the orthogonal projection of  $b_j$  onto the column space of  $A$ . Doing this for every column  $b_j$  leaves us with the following approximation to the picture:

$$B \approx \left( \begin{array}{c|c|c} A \underbrace{(A^T A)^{-1} A^T b_0}_{x_0} & \cdots & A \underbrace{(A^T A)^{-1} A^T b_{n-1}}_{x_{n-1}} \end{array} \right),$$

which is equivalent to

$$B \approx A \underbrace{\left( A^T A \right)^{-1} A^T \begin{pmatrix} b_0 & | & \cdots & | & b_{n-1} \\ x_0 & | & \cdots & | & x_{n-1} \end{pmatrix}}_{X} = A \underbrace{\left( A^T A \right)^{-1} A^T B}_{X} = AX.$$

Importantly, instead of requiring  $m \times n$  data to store  $B$ , we now need only store  $A$  and  $X$ .

**Homework 2.1.1.1** If  $B$  is  $m \times n$  and  $A$  is  $m \times k$ , how many entries are there in  $A$  and  $X$ ?

**Solution.**

- $A$  is  $m \times k$ .
- $X$  is  $k \times n$ .

A total of  $(m + n)k$  entries are in  $A$  and  $X$ .

**Homework 2.1.1.2**  $AX$  is called a rank-k approximation of  $B$ . Why?

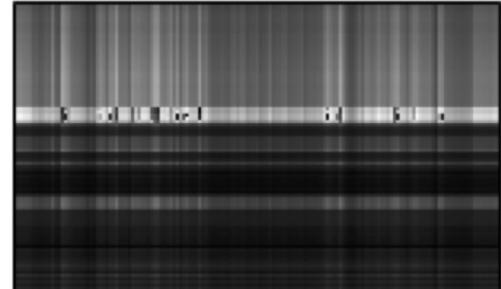
**Solution.** The matrix  $AX$  has rank at most equal to  $k$  (it is a rank-k matrix) since each of its columns can be written as a linear combinations of the columns of  $A$  and hence it has at most  $k$  linearly independent columns.

Let's have a look at how effective this approach is for our picture:

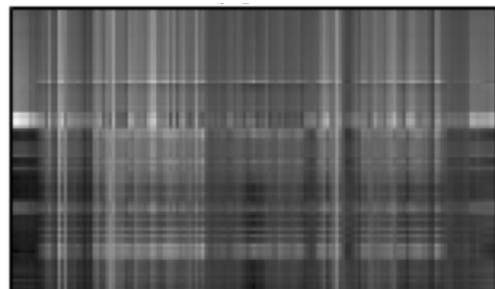
original:



$k = 1$



$k = 2$



$k = 10$



$k = 25$  $k = 50$ 

Now, there is no reason to believe that picking equally spaced columns (or restricting ourselves to columns in  $B$ ) will yield the best rank- $k$  approximation for the picture. It yields a pretty good result here in part because there is quite a bit of repetition in the picture, from column to column. So, the question can be asked: How do we find the best rank- $k$  approximation for a picture or, more generally, a matrix? This would allow us to get the most from the data that needs to be stored. It is the Singular Value Decomposition (SVD), possibly the most important result in linear algebra, that provides the answer.

**Remark 2.1.1.1** Those who need a refresher on this material may want to review Week 11 of Linear Algebra: Foundations to Frontiers [20]. We will discuss solving linear least squares problems further in (((Unresolved xref, reference "chapter04-linear-least-squares"; check spelling or use "provisional" attribute))).

## 2.1.2 Overview

- 2.1 Opening Remarks
  - 2.1.1 Low rank approximation
  - 2.1.2 Overview
  - 2.1.3 What you will learn
- 2.2 Orthogonal Vectors and Matrices
  - 2.2.1 Orthogonal vectors
  - 2.2.2 Component in the direction of a vector
  - 2.2.3 Orthonormal vectors and matrices
  - 2.2.4 Unitary matrices
  - 2.2.5 Examples of unitary matrices
  - 2.2.6 Change of orthonormal basis
  - 2.2.7 Why we love unitary matrices choice
- 2.3 The Singular Value Decomposition
  - 2.3.1 The Singular Value Decomposition Theorem
  - 2.3.2 Geometric interpretation

- 2.3.3 An "algorithm" for computing the SVD
- 2.3.4 The Reduced Singular Value Decomposition
- 2.3.5 The SVD of nonsingular matrices
- 2.3.6 Best rank-k approximation
- 2.4 Enrichments
  - 2.4.1 Principle Component Analysis (PCA)
- 2.5 Wrap Up
  - 2.5.1 Additional homework
  - 2.5.2 Summary

### 2.1.3 What you will learn

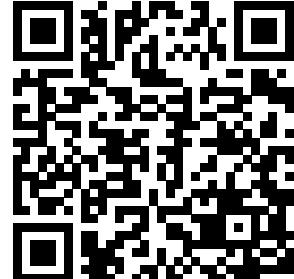
This week introduces two concepts that have theoretical and practical importance: unitary matrices and the Singular Value Decomposition (SVD).

Upon completion of this week, you should be able to

- Determine whether vectors are orthogonal.
- Compute the component of a vector in the direction of another vector.
- Relate sets of orthogonal vectors to orthogonal and unitary matrices.
- Connect unitary matrices to the changing of orthonormal basis.
- Identify transformations that can be represented by unitary matrices.
- Prove that multiplying with unitary matrices does not amplify relative error.
- Use norms to quantify the conditioning of solving linear systems.
- Prove and interpret the Singular Value Decomposition.
- Link the Reduced Singular Value Decomposition to the rank of the matrix and determine the best rank-k approximation to a matrix.
- Determine whether a matrix is close to being nonsingular by relating the Singular Value Decomposition to the condition number.

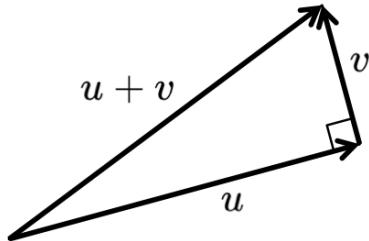
## 2.2 Orthogonal Vectors and Matrices

### 2.2.1 Orthogonal vectors

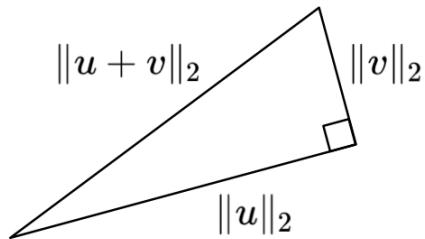


YouTube: <https://www.youtube.com/watch?v=3zpdTfwZSEo>

At some point in your education you were told that vectors are orthogonal (perpendicular) if and only if their dot product (inner product) equals zero. Let's review where this comes from. Given two vectors  $u, v \in \mathbb{R}^m$ , those two vectors, and their sum all exist in the same two dimensional (2D) subspace. So, they can be visualized as



where the page on which they are drawn is that 2D subspace. Now, if they are, as drawn, perpendicular and we consider the lengths of the sides of the triangle that they define



then we can employ the first theorem you were probably ever exposed to, the Pythagorean Theorem, to find that

$$\|u\|_2^2 + \|v\|_2^2 = \|u + v\|_2^2.$$

Using what we know about the relation between the two norm and the dot product, we find

that

$$\begin{aligned}
 u^T u + v^T v &= (u + v)^T (u + v) \\
 &\Leftrightarrow \quad < \text{multiply out} > \\
 u^T u + v^T v &= u^T u + u^T v + v^T u + v^T v \\
 &\Leftrightarrow \quad < u^T v = v^T u \text{ if } u \text{ and } v \text{ are real-valued} > \\
 u^T u + v^T v &= u^T u + 2u^T v + v^T v \\
 &\Leftrightarrow \quad < \text{delete common terms} > \\
 0 &= 2u^T v
 \end{aligned}$$

so that we can conclude that  $u^T v = 0$ .

While we already encountered the notation  $x^H x$  as an alternative way of expressing the length of a vector,  $\|x\|_2 = \sqrt{x^H x}$ , we have not formally defined the inner product (dot product), for complex-valued vectors:

**Definition 2.2.1.1 Dot product (Inner product).** Given  $x, y \in \mathbb{C}^m$  their dot product (inner product) is defined as

$$x^H y = \bar{x}^T y = \bar{x}^T y = \bar{\chi}_0 \psi_0 + \bar{\chi}_1 \psi_1 + \cdots + \bar{\chi}_{m-1} \psi_{m-1} = \sum_{i=0}^{m-1} \bar{\chi}_i \psi_i.$$

◊

The notation  $x^H$  is short for  $\bar{x}^T$ , where  $\bar{x}$  equals the vector  $x$  with all its entries conjugated. So,

$$\begin{aligned}
 x^H y &= < \text{expose the elements of the vectors} > \\
 \left( \begin{array}{c} \chi_0 \\ \vdots \\ \chi_{m-1} \end{array} \right)^H \left( \begin{array}{c} \psi_0 \\ \vdots \\ \psi_{m-1} \end{array} \right) &= < x^H = \bar{x}^T > \\
 \overline{\left( \begin{array}{c} \chi_0 \\ \vdots \\ \chi_{m-1} \end{array} \right)}^T \left( \begin{array}{c} \psi_0 \\ \vdots \\ \psi_{m-1} \end{array} \right) &= < \text{conjugate the elements of } x > \\
 \left( \begin{array}{c} \bar{\chi}_0 \\ \vdots \\ \bar{\chi}_{m-1} \end{array} \right)^T \left( \begin{array}{c} \psi_0 \\ \vdots \\ \psi_{m-1} \end{array} \right) &= < \text{view } x \text{ as a } m \times 1 \text{ matrix and transpose} > \\
 \left( \begin{array}{c|c|c} \bar{\chi}_0 & \cdots & \bar{\chi}_{m-1} \end{array} \right) \left( \begin{array}{c} \psi_0 \\ \vdots \\ \psi_{m-1} \end{array} \right) &= < \text{view } x^H \text{ as a matrix and perform matrix-vector multiplication} > \\
 \sum_{i=0}^{m-1} \bar{\chi}_i \psi_i. &
 \end{aligned}$$

**Homework 2.2.1.1** Let  $x, y \in \mathbb{C}^m$ .

ALWAYS/SOMETIMES/NEVER:  $\overline{x^H y} = y^H x$ .

**Answer.** ALWAYS

Now prove it!

**Solution.**

$$\overline{x^H y} = \overline{\sum_{i=0}^{m-1} \bar{\chi}_i \psi_i} = \sum_{i=0}^{m-1} \overline{\bar{\chi}_i \psi_i} = \sum_{i=0}^{m-1} \bar{\psi}_i \bar{\chi}_i = y^H x.$$

**Homework 2.2.1.2** Let  $x \in \mathbb{C}^m$ .

ALWAYS/SOMETIMES/NEVER:  $x^H x$  is real-valued.

**Answer.** ALWAYS

Now prove it!

**Solution.** By the last homework,

$$\overline{x^H x} = x^H x,$$

A complex number is equal to its conjugate only if it is real-valued.

The following defines orthogonality of two vectors with complex-valued elements:

**Definition 2.2.1.2 Orthogonal vectors.** Let  $x, y \in \mathbb{C}^m$ . These vectors are said to be orthogonal (perpendicular) iff  $x^H y = 0$ .  $\diamond$

## 2.2.2 Component in the direction of a vector



YouTube: <https://www.youtube.com/watch?v=CqcJ6Nh1QWg>

In a previous linear algebra course, you may have learned that if  $a, b \in \mathbb{R}^m$  then

$$\hat{b} = \frac{a^T b}{a^T a} a = \frac{aa^T}{a^T a} b$$

equals the component of  $b$  in the direction of  $a$  and

$$b^\perp = b - \hat{b} = \left( I - \frac{aa^T}{a^T a} \right) b$$

equals the component of  $b$  orthogonal to  $a$ , since  $b = \hat{b} + b^\perp$  and  $\hat{b}^T b^\perp = 0$ . Similarly, if  $a, b \in \mathbb{C}^m$  then

$$\hat{b} = \frac{a^H b}{a^H a} a = \frac{aa^H}{a^H a} b$$

equals the component of  $b$  in the direction of  $a$  and

$$b^\perp = b - \hat{b} = \left(I - \frac{aa^H}{a^H a}\right)b$$

equals the component of  $b$  orthogonal to  $a$ .

**Remark 2.2.2.1** The matrix that (orthogonally) projects the vector to which it is applied onto the vector  $a$  is given by

$$\frac{aa^H}{a^H a}$$

while

$$I - \frac{aa^H}{a^H a}$$

is the matrix that (orthogonally) projects the vector to which it is applied onto the space orthogonal to the vector  $a$ .

**Homework 2.2.2.1** Let  $a \in \mathbb{C}^m$ .

ALWAYS/SOMETIMES/NEVER>:

$$\left(\frac{aa^H}{a^H a}\right) \left(\frac{aa^H}{a^H a}\right) = \frac{aa^H}{a^H a}.$$

Interpret what this means about a matrix that projects onto a vector.

**Answer.** ALWAYS.

Now prove it.

**Solution.**

$$\begin{aligned} & \left(\frac{aa^H}{a^H a}\right) \left(\frac{aa^H}{a^H a}\right) \\ &= < \text{multiply numerators and denominators} > \\ & \frac{a^H a a a^H}{(a^H a)(a^H a)} \\ &= < \text{associativity} > \\ & \frac{a(a^H a)a^H}{(a^H a)(a^H a)} \\ &= < \text{scalar division} > \\ & \frac{aa^H}{a^H a}. \end{aligned}$$

Interpretation: orthogonally projecting the orthogonal projection of a vector yields the orthogonal projection of the vector.

**Homework 2.2.2.2** Let  $a \in \mathbb{C}^m$ .

ALWAYS/SOMETIMES/NEVER:

$$\left(\frac{aa^H}{a^H a}\right) \left(I - \frac{aa^H}{a^H a}\right) = 0$$

(the zero matrix). Interpret what this means.

**Answer.** ALWAYS.

Now prove it.

**Solution.**

$$\begin{aligned}
 & \left( \frac{aa^H}{a^H a} \right) \left( I - \frac{aa^H}{a^H a} \right) \\
 & = \quad < \text{distribute} > \\
 & \left( \frac{aa^H}{a^H a} \right) - \left( \frac{aa^H}{a^H a} \right) \left( \frac{aa^H}{a^H a} \right) \\
 & = \quad < \text{last homework} > \\
 & \left( \frac{aa^H}{a^H a} \right) - \left( \frac{aa^H}{a^H a} \right) \\
 & = \\
 & 0.
 \end{aligned}$$

Interpretation: first orthogonally projecting onto the space *orthogonal to* vector  $a$  and then orthogonally projecting the resulting vector onto that  $a$  leaves you with the zero vector.

**Homework 2.2.2.3** Let  $a, b \in \mathbb{C}^n$ ,  $\hat{b} = \frac{aa^H}{a^H a}b$ , and  $b^\perp = b - \hat{b}$ .

ALWAYS/SOMETIMES/NEVER:  $\hat{b}^H b^\perp = 0$ .

**Answer.** ALWAYS.

Now prove it.

**Solution.**

$$\begin{aligned}
 & \hat{b}^H b^\perp \\
 & = \quad < \text{substitute } \hat{b} \text{ and } b^\perp > \\
 & \left( \frac{aa^H}{a^H a} b \right)^H (b - \hat{b}) \\
 & = \quad < (Ax)^H = x^H A^H; \text{ substitute } b - \hat{b} > \\
 & b^H \left( \frac{aa^H}{a^H a} \right)^H (I - \frac{aa^H}{a^H a}) b \\
 & = \quad < (((xy^H)/\alpha)^H = yx^H/\alpha \text{ if } \alpha \text{ is real} > \\
 & b^H \frac{aa^H}{a^H a} (I - \frac{aa^H}{a^H a}) b \\
 & = \quad < \text{last homework} > \\
 & b^H 0 b \\
 & = \quad < 0x = 0; y^H 0 = 0 > \\
 & 0.
 \end{aligned}$$

## 2.2.3 Orthonormal vectors and matrices



YouTube: <https://www.youtube.com/watch?v=GFFfvDpj5dzw>

A lot of the formulae in the last unit become simpler if the length of the vector equals one: If  $\|u\|_2 = 1$  then

- the component of  $v$  in the direction of  $u$  equals

$$\frac{u^H v}{u^H u} u = u^H v u.$$

- the matrix that projects a vector onto the vector  $u$  is given by

$$\frac{u u^H}{u^H u} = u u^H.$$

- the component of  $v$  orthogonal to  $u$  equals

$$v - \frac{u^H v}{u^H u} u = v - u^H v u.$$

- the matrix that projects a vector onto the space orthogonal to  $u$  is given by

$$I - \frac{u u^H}{u^H u} = I - u u^T.$$

**Homework 2.2.3.1** Let  $u \neq 0 \in \mathbb{C}^m$ .

ALWAYS/SOMETIMES/NEVER  $u/\|u\|_2$  has unit length.

**Answer.** ALWAYS.

Now prove it.

**Solution.**

$$\begin{aligned} \left\| \frac{u}{\|u\|_2} \right\|_2 &= < \text{homogeneity of norms} > \\ \frac{\|u\|_2}{\|u\|_2} &= < \text{algebra} > \\ 1 & \end{aligned}$$

This last exercise shows that any nonzero vector can be scaled (normalized) to have unit length.

**Definition 2.2.3.1 Orthonormal vectors.** Let  $u_0, u_1, \dots, u_{n-1} \in \mathbb{C}^m$ . These vectors are said to be mutually orthonormal if for all  $0 \leq i, j < n$

$$u_i^H u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

◊

The definition implies that  $\|u_i\|_2 = \sqrt{u_i^H u_i} = 1$  and hence each of the vectors is of unit length in addition to being orthogonal to each other.

The standard basis vectors ([Definition 1.3.1.3](#))

$$\{e_j\}_{j=0}^{m-1} \subset \mathbb{C}^m,$$

where

$$e_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \text{entry indexed with } j$$

are mutually orthonormal since, clearly,

$$e_i^H e_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Naturally, any subset of the standard basis vectors is a set of mutually orthonormal vectors.

**Remark 2.2.3.2** For  $n$  vectors of size  $m$  to be mutually orthonormal,  $n$  must be less than or equal to  $m$ . This is because  $n$  mutually orthonormal vectors are linearly independent and there can be at most  $m$  linearly independent vectors of size  $m$ .

A very concise way of indicating that a set of vectors are mutually orthonormal is to view them as the columns of a matrix, which then has a very special property:

**Definition 2.2.3.3 Orthonormal matrix.** Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Then  $Q$  is said to be an orthonormal matrix iff  $Q^H Q = I$ .  $\diamond$

The subsequent exercise makes the connection between mutually orthonormal vectors and an orthonormal matrix.

**Homework 2.2.3.2** Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Partition  $Q = ( q_0 | q_1 | \cdots | q_{n-1} )$ .

TRUE/FALSE:  $Q$  is an orthonormal matrix if and only if  $q_0, q_1, \dots, q_{n-1}$  are mutually orthonormal.

**Answer.** TRUE

Now prove it!

**Solution.** Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Partition  $Q = ( q_0 | q_1 | \cdots | q_{n-1} )$ . Then

$$\begin{aligned} Q^H Q &= ( q_0 | q_1 | \cdots | q_{n-1} )^H ( q_0 | q_1 | \cdots | q_{n-1} ) \\ &= \begin{pmatrix} q_0^H \\ q_1^H \\ \vdots \\ q_{n-1}^H \end{pmatrix} ( q_0 | q_1 | \cdots | q_{n-1} ) \\ &= \begin{pmatrix} q_0^H q_0 & q_0^H q_1 & \cdots & q_0^H q_{n-1} \\ \hline q_1^H q_0 & q_1^H q_1 & \cdots & q_1^H q_{n-1} \\ \hline \vdots & \vdots & & \vdots \\ \hline q_{n-1}^H q_0 & q_{n-1}^H q_1 & \cdots & q_{n-1}^H q_{n-1} \end{pmatrix}. \end{aligned}$$

Now consider that  $Q^H Q = I$ :

$$\left( \begin{array}{c|c|c|c} q_0^H q_0 & q_0^H q_1 & \cdots & q_0^H q_{n-1} \\ \hline q_1^H q_0 & q_1^H q_1 & \cdots & q_1^H q_{n-1} \\ \vdots & \vdots & & \vdots \\ \hline q_{n-1}^H q_0 & q_{n-1}^H q_1 & \cdots & q_{n-1}^H q_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c|c} 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \hline 0 & 0 & \cdots & 1 \end{array} \right).$$

Clearly  $Q$  is orthonormal if and only if  $q_0, q_1, \dots, q_{n-1}$  are mutually orthonormal.

**Homework 2.2.3.3** Let  $Q \in \mathbb{C}^{m \times n}$ .

ALWAYS/SOMETIMES/NEVER: If  $Q^H Q = I$  then  $Q Q^H = I$ .

**Answer.** SOMETIMES.

Now explain why.

**Solution.**

- If  $Q$  is a square matrix ( $m = n$ ) then  $Q^H Q = I$  means  $Q^{-1} = Q^H$ . But then  $Q Q^{-1} = I$  and hence  $Q Q^H = I$ .
- If  $Q$  is not square, then  $Q^H Q = I$  means  $m > n$ . Hence  $Q$  has rank equal to  $n$  which in turn means  $Q Q^H$  is a matrix with rank at most equal to  $n$ . (Actually, its rank equals  $n$ ). Since  $I$  has rank equal to  $m$  (it is an  $m \times m$  matrix with linearly independent columns),  $Q Q^H$  cannot equal  $I$ .

More concretely: let  $m > 1$  and  $n = 1$ . Choose  $Q = \begin{pmatrix} e_0 \end{pmatrix}$ . Then  $Q^H Q = e_0^H e_0 = 1 = I$ . But

$$Q Q^H = e_0 e_0^H = \begin{pmatrix} 1 & 0 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

## 2.2.4 Unitary matrices



YouTube: <https://www.youtube.com/watch?v=izONEm09uqw>

**Homework 2.2.4.1** Let  $Q \in \mathbb{C}^{m \times n}$  be an orthonormal matrix.

ALWAYS/SOMETIMES/NEVER:  $Q^{-1} = Q^H$  and  $Q Q^H = I$ .

**Answer.** SOMETIMES

Now explain it!

**Solution.** If  $Q$  is unitary, then it is an orthonormal matrix and square. Because it is an orthonormal matrix,  $Q^H Q = I$ . If  $A, B \in \mathbb{C}^{m \times m}$ , the matrix  $B$  such that  $BA = I$  is the inverse of  $A$ . Hence  $Q^{-1} = Q^H$ . Also, if  $BA = I$  then  $AB = I$  and hence  $QQ^H = I$ .

However, an orthonormal matrix is not necessarily square. For example, the matrix  $Q = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$  is an orthonormal matrix:  $Q^T Q = I$ . However, it doesn't have an inverse because it is not square.

If an orthonormal matrix is square, then it is called a unitary matrix.

**Definition 2.2.4.1 Unitary matrix.** Let  $U \in \mathbb{C}^{m \times m}$ . Then  $U$  is said to be a unitary matrix if and only if  $U^H U = I$  (the identity).  $\diamond$

**Remark 2.2.4.2** Unitary matrices are always square. Sometimes the term **orthogonal matrix** is used instead of unitary matrix, especially if the matrix is real valued.

Unitary matrices have some very nice properties, as captured by the following exercises.

**Homework 2.2.4.2** Let  $Q \in \mathbb{C}^{m \times m}$  be a unitary matrix.

ALWAYS/SOMETIMES/NEVER:  $Q^{-1} = Q^H$  and  $QQ^H = I$ .

**Answer.** ALWAYS

Now explain it!

**Solution.** If  $Q$  is unitary, then it is square and  $Q^H Q = I$ . Hence  $Q^{-1} = Q^H$  and  $QQ^H = I$ .

**Homework 2.2.4.3** TRUE/FALSE: If  $U$  is unitary, so is  $U^H$ .

**Answer.** TRUE

Now prove it!

**Solution.** Clearly,  $U^H$  is square. Also,  $(U^H)^H U^H = (UU^H)^H = I$  by the last homework.

**Homework 2.2.4.4** Let  $U_0, U_1 \in \mathbb{C}^{m \times m}$  both be unitary.

ALWAYS/SOMETIMES/NEVER:  $U_0 U_1$ , is unitary.

**Answer.** ALWAYS

Now prove it!

**Solution.** Obviously,  $U_0 U_1$  is a square matrix.

Now,

$$(U_0 U_1)^H (U_0 U_1) = U_1^H \underbrace{U_0^H U_0}_I U_1 = \underbrace{U_1^H U_1}_I = I.$$

Hence  $U_0 U_1$  is unitary.

**Homework 2.2.4.5** Let  $U_0, U_1, \dots, U_{k-1} \in \mathbb{C}^{m \times m}$  all be unitary.

ALWAYS/SOMETIMES/NEVER: Their product,  $U_0 U_1 \cdots U_{k-1}$ , is unitary.

**Answer.** ALWAYS

Now prove it!

**Solution.** Strictly speaking, we should do a proof by induction. But instead we will make

the more informal argument that

$$\begin{aligned}
 (U_0 U_1 \cdots U_{k-1})^H U_0 U_1 \cdots U_{k-1} &= U_{k-1}^H \cdots U_1^H U_0^H U_0 U_1 \cdots U_{k-1} \\
 &= U_{k-1}^H \cdots U_1^H \underbrace{U_0^H U_0}_I U_1 \cdots U_{k-1} = I.
 \end{aligned}$$

(When you see a proof that involved  $\cdots$ , it would be more rigorous to use a proof by induction.)

**Remark 2.2.4.3** Many algorithms that we encounter in the future will involve the application of a sequence of unitary matrices, which is why the result in this last exercise is of great importance.

Perhaps the most important property of a unitary matrix is that it preserves length.

**Homework 2.2.4.6** Let  $U \in \mathbb{C}^{m \times m}$  be a unitary matrix and  $x \in \mathbb{C}^m$ . Prove that  $\|Ux\|_2 = \|x\|_2$ .

**Solution.**

$$\begin{aligned}
 \|Ux\|_2^2 &= < \text{alternative definition} > \\
 (Ux)^H Ux &= < (Az)^H = z^H A^H > \\
 x^H U^H Ux &= < U \text{ is unitary} > \\
 x^H x &= < \text{alternative definition} > \\
 \|x\|_2^2.
 \end{aligned}$$

The converse is true as well:

**Theorem 2.2.4.4** If  $A \in \mathbb{C}^{m \times m}$  preserves length ( $\|Ax\|_2 = \|x\|_2$  for all  $x \in \mathbb{C}^m$ ), then  $A$  is unitary.

*Proof.* We first prove that  $(Ax)^H (Ay) = x^H y$  for all  $x, y$  by considering  $\|x - y\|_2^2 = \|A(x - y)\|_2^2$ . We then use that to evaluate  $e_i^H A^H A e_j$ .

Let  $x, y \in \mathbb{C}^m$ . Then

$$\begin{aligned}
& \|x - y\|_2^2 = \|A(x - y)\|_2^2 \\
& \Leftrightarrow <\text{alternative definition}> \\
& (x - y)^H(x - y) = (A(x - y))^H A(x - y) \\
& = <(Bz) = z^H B^H> \\
& (x - y)^H(x - y) = (x - y)^H A^H A(x - y) \\
& \Leftrightarrow <\text{multiply out}> \\
& x^H x - x^H y - y^H x + y^H y = x^H A^H A x - x^H A^H A y - y^H A^H A x + y^H A^H A y \\
& \Leftrightarrow <\text{alternative definition}; \overline{x^H y} = y^H x> \\
& \|x\|_2^2 - (x^H y + \overline{x^H y}) + \|y\|_2^2 = \|Ax\|_2^2 - (x^H A^H A y + \overline{x^H A^H A y}) + \|Ay\|_2^2 \\
& \Leftrightarrow <\|Ax\|_2 = \|x\|_2 \text{ and } \|Ay\|_2 = \|y\|_2; \alpha + \bar{\alpha} = 2\operatorname{Re}(\alpha)> \\
& \operatorname{Re}(x^H y) = \operatorname{Re}((Ax)^H A y)
\end{aligned}$$

One can similarly show that  $\operatorname{Im}(x^H y) = \operatorname{Im}((Ax)^H A y)$  by considering  $A(ix - y)$ .

Conclude that  $(Ax)^H(Ay) = x^H y$ .

We now use this to show that  $A^H A = I$  by using the fact that the standard basis vectors have the property that

$$e_i^H e_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

and that the  $i, j$  entry in  $A^H A$  equals  $e_i^H A^H A e_j$ .

Note: I think the above can be made much more elegant by choosing  $\alpha$  such that  $\alpha x^H y$  is real and then looking at  $\|x + \alpha y\|_2 = \|A(x + \alpha y)\|_2$  instead, much like we did in the proof of the Cauchy-Schwartz inequality. Try and see if you can work out the details. ■

**Homework 2.2.4.7** Prove that if  $U$  is unitary then  $\|U\|_2 = 1$ .

**Solution.**

$$\begin{aligned}
& \|U\|_2 \\
& = <\text{definition}> \\
& \max_{\|x\|_2=1} \|Ux\|_2 \\
& = <\text{unitary matrices preserve length}> \\
& \max_{\|x\|_2=1} \|x\|_2 \\
& = <\text{algebra}> \\
& 1
\end{aligned}$$

(The above can be really easily proven with the SVD. Let's point that out later.)

**Homework 2.2.4.8** Prove that if  $U$  is unitary then  $\kappa_2(U) = 1$ .

**Solution.**

$$\begin{aligned}
 \kappa_2 U &= <\text{definition}> \\
 \|U\|_2 \|U^{-1}\|_2 &= <\text{both } U \text{ and } U^{-1} \text{ are unitary ; last homework}> \\
 1 \times 1 &= <\text{arithmetic}> \\
 1
 \end{aligned}$$

The preservation of length extends to the preservation of norms that have a relation to the 2-norm:

**Homework 2.2.4.9** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary and  $A \in \mathbb{C}^{m \times n}$ . Show that

- $\|U^H A\|_2 = \|A\|_2$ .
- $\|AV\|_2 = \|A\|_2$ .
- $\|U^H AV\|_2 = \|A\|_2$ .

**Hint.** Exploit the definition of the 2-norm:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

**Solution.**

- $$\begin{aligned}
 \|U^H A\|_2 &= <\text{definition of 2-norm}> \\
 \max_{\|x\|_2=1} \|U^H Ax\|_2 &= < U \text{ is unitary and unitary matrices preserve length}> \\
 \max_{\|x\|_2=1} \|Ax\|_2 &= <\text{definition of 2-norm}> \\
 \|A\|_2.
 \end{aligned}$$
- $$\begin{aligned}
 \|AV\|_2 &= <\text{definition of 2-norm}> \\
 \max_{\|x\|_2=1} \|AVx\|_2 &= < V^H \text{ is unitary and unitary matrices preserve length}> \\
 \max_{\|Vx\|_2=1} \|A(Vx)\|_2 &= <\text{substitute } y = Vx> \\
 \max_{\|y\|_2=1} \|Ay\|_2 &= <\text{definition of 2-norm}> \\
 \|A\|_2.
 \end{aligned}$$

- The last part follows immediately from the previous two:

$$\|U^H AV\|_2 = \|U^H(AV)\|_2 = \|AV\|_2 = \|A\|_2.$$

**Homework 2.2.4.10** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary and  $A \in \mathbb{C}^{m \times n}$ . Show that

- $\|U^H A\|_F = \|A\|_F$ .
- $\|AV\|_F = \|A\|_F$ .
- $\|U^H AV\|_F = \|A\|_F$ .

**Hint.** How does  $\|A\|_F$  relate to the 2-norms of its columns?

**Solution.**

- Partition

$$A = \left( \begin{array}{c|c|c} a_0 & \cdots & a_{n-1} \end{array} \right).$$

Then we saw in [Subsection 1.3.3](#) that  $\|A\|_F^2 = \sum_{j=0}^{n-1} \|a_j\|_2^2$ .

Now,

$$\begin{aligned} \|U^H A\|_F^2 &= \langle \text{partition } A \text{ by columns} \rangle \\ \|U^H \left( \begin{array}{c|c|c} a_0 & \cdots & a_{n-1} \end{array} \right) \|_2^2 &= \langle \text{property of matrix-vector multiplication} \rangle \\ \left\| \left( \begin{array}{c|c|c} U^H a_0 & \cdots & U^H a_{n-1} \end{array} \right) \right\|_2^2 &= \langle \text{exercice in Chapter 1} \rangle \\ \sum_{j=0}^{n-1} \|U^H a_j\|_2^2 &= \langle \text{unitary matrices preserve length} \rangle \\ \sum_{j=0}^{n-1} \|a_j\|_2^2 &= \langle \text{exercice in Chapter 1} \rangle \\ \|A\|_F^2. \end{aligned}$$

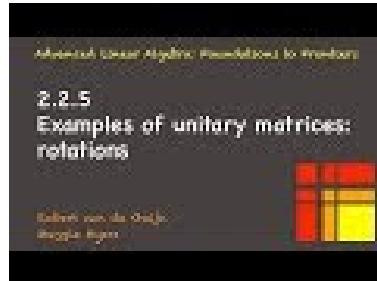
- To prove that  $\|AV\|_F = \|A\|_F$  recall that  $\|A^H\|_F = \|A\|_F$ .
- The last part follows immediately from the first two parts.

In the last two exercises we consider  $U^H AV$  rather than  $UAV$  because it sets us up better for future discussion.

## 2.2.5 Examples of unitary matrices

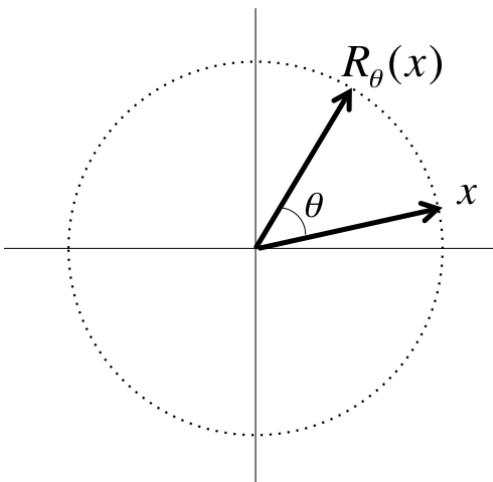
In this unit, we will discuss a few situations where you may have encountered unitary matrices without realizing. Since few of us walk around pointing out to each other "Look, another matrix!" we first consider if a transformation (function) might be a linear transformation. This allows us to then ask the question "What kind of transformations we see around us preserve length?" After that, we discuss how those transformations are represented as matrices. That leaves us to then check whether the resulting matrix is unitary.

### 2.2.5.1 Rotations



YouTube: <https://www.youtube.com/watch?v=C0mlDZ280hc>

A rotation in 2D,  $R_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , takes a vector and rotates that vector through the angle  $\theta$ :

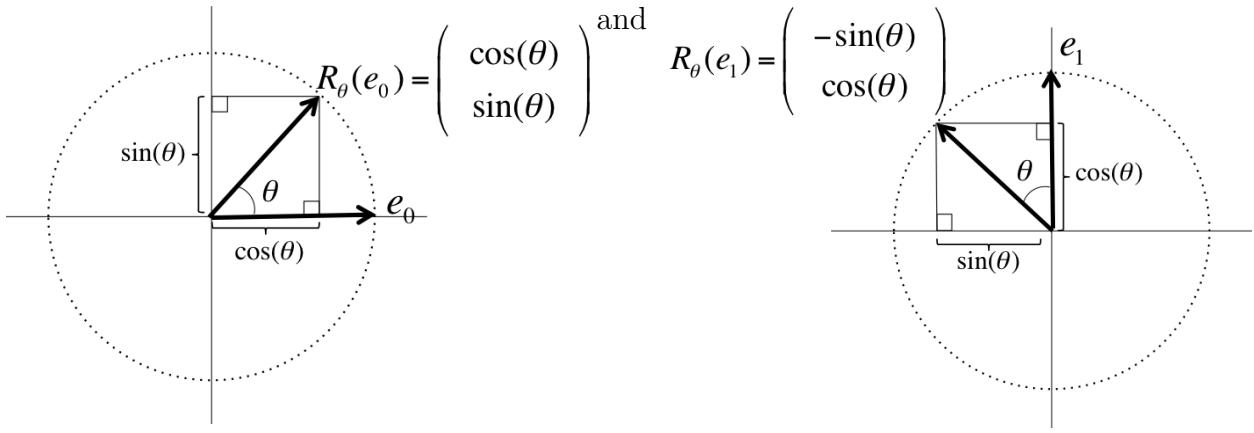


If you think about it,

- If you scale a vector first and then rotate it, you get the same result as if you rotate it first and then scale it.
- If you add two vectors first and then rotate, you get the same result as if you rotate them first and then add them.

Thus, a rotation is a linear transformation. Also, the above picture captures that a rotation preserves the length of the vector to which it is applied. We conclude that the matrix that represents a rotation should be a unitary matrix.

Let us compute the matrix that represents the rotation through an angle  $\theta$ . Recall that if  $L : \mathbb{C}^n \rightarrow \mathbb{C}^m$  is a linear transformation and  $A$  is the matrix that represents it, then the  $j$ th column of  $A$ ,  $a_j$ , equals  $L(e_j)$ . The pictures



illustrate that

$$R_\theta(e_0) = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} \quad \text{and} \quad R_\theta(e_1) = \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix}.$$

Thus,

$$R_\theta(x) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix}.$$

**Homework 2.2.5.1** Show that

$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

is a unitary matrix. (Since it is real valued, it is usually called an orthogonal matrix instead.)

**Hint.** Hint: use  $c$  for  $\cos(\theta)$  and  $s$  for  $\sin(\theta)$  to save yourself a lot of writing!

**Solution.**

$$\begin{aligned}
 & \left( \begin{array}{c|cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array} \right)^H \left( \begin{array}{c|cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array} \right) \\
 &= \text{< the matrix is real valued >} \\
 & \left( \begin{array}{c|cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array} \right)^T \left( \begin{array}{c|cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array} \right) \\
 &= \text{< transpose >} \\
 & \left( \begin{array}{cc} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{array} \right) \left( \begin{array}{c|cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array} \right) \\
 &= \text{< multiply >} \\
 & \left( \begin{array}{cc} \cos^2(\theta) + \sin^2(\theta) & -\cos(\theta)\sin(\theta) + \sin(\theta)\cos(\theta) \\ -\sin(\theta)\cos(\theta) + \cos(\theta)\sin(\theta) & \sin^2(\theta) + \cos^2(\theta) \end{array} \right) \\
 &= \text{< geometry; algebra >} \\
 & \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right)
 \end{aligned}$$

**Homework 2.2.5.2** Prove, without relying on geometry but using what you just discovered, that  $\cos(-\theta) = \cos(\theta)$  and  $\sin(-\theta) = -\sin(\theta)$

**Solution.** Undoing a rotation by an angle  $\theta$  means rotating in the opposite direction through angle  $\theta$  or, equivalently, rotating through angle  $-\theta$ . Thus, the inverse of  $R_\theta$  is  $R_{-\theta}$ . The matrix that represents  $R_{\text{theta}}$  is given by

$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

and hence the matrix that represents  $R_{-\theta}$  is given by

$$\begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix}.$$

Since  $R_{-\theta}$  is the inverse of  $R_\theta$  we conclude that

$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}^{-1} = \begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix}.$$

But we just discovered that

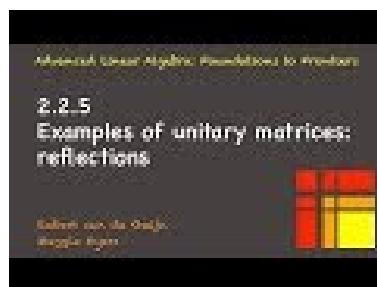
$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}^{-1} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}^T = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Hence

$$\begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}^{-1}$$

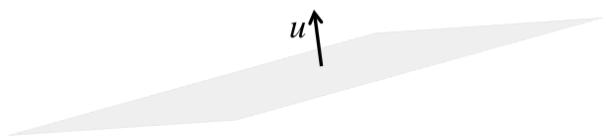
from which we conclude that  $\cos(-\theta) = \cos(\theta)$  and  $\sin(-\theta) = -\sin(\theta)$ .

### 2.2.5.2 Reflections

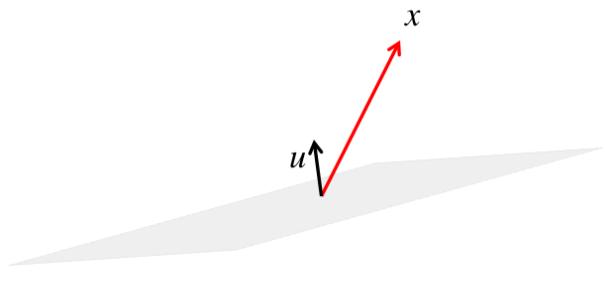


YouTube: <https://www.youtube.com/watch?v=r8S04qqcc-o>

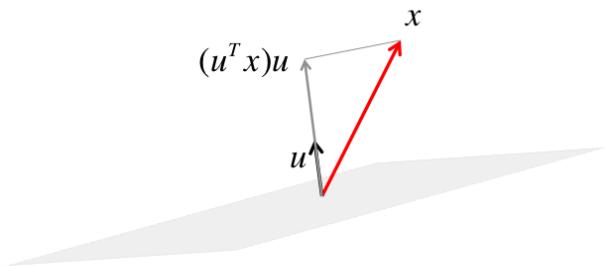
Picture a mirror with its orientation defined by a unit length vector,  $u$ , that is orthogonal to it.



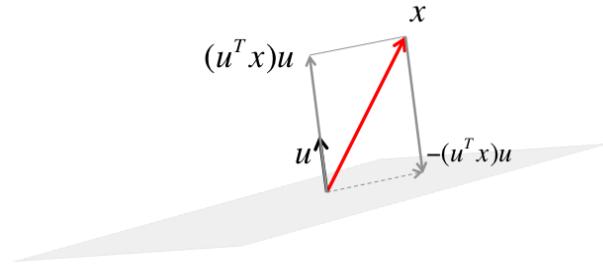
We will consider how a vector,  $x$ , is reflected by this mirror.



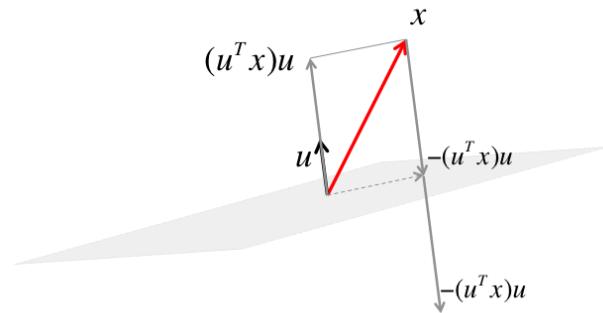
The component of  $x$  orthogonal to the mirror equals the component of  $x$  in the direction of  $u$ , which equals  $(u^T x)u$ .



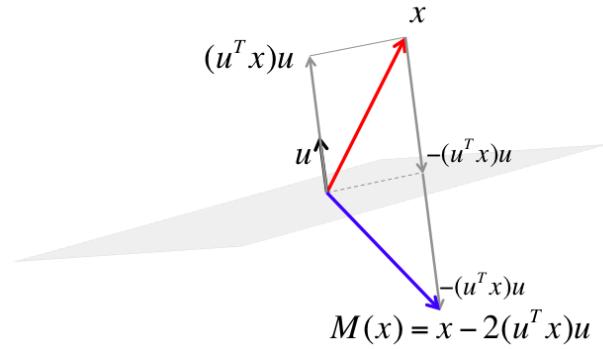
The orthogonal projection of  $x$  onto the mirror is then given by the dashed vector, which equals  $x - (u^T x)u$ .



To get to the reflection of  $x$ , we now need to go further yet by  $-(u^T x)u$ .



We conclude that the transformation that mirrors (reflects)  $x$  with respect to the mirror is given by  $M(x) = x - 2(u^T x)u$ .



The transformation described above preserves the length of the vector to which it is applied.

**Homework 2.2.5.3** (Verbally) describe why reflecting a vector as described above is a linear transformation.

**Solution.**

- If you scale a vector first and then reflect it, you get the same result as if you reflect it first and then scale it.
- If you add two vectors first and then reflect, you get the same result as if you reflect them first and then add them.

**Homework 2.2.5.4** Show that the matrix that represents  $M : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  in the above example is given by  $I - 2uu^T$ .

**Hint.** Rearrange  $x - 2(u^Tx)u$ .

**Solution.**

$$\begin{aligned} x - 2(u^Tx)u &= \langle \alpha x = x\alpha \rangle \\ x - 2u(u^Tx) &= \langle \text{associativity} \rangle \\ Ix - 2uu^Tx &= \langle \text{distributivity} \rangle \\ (I - 2uu^T)x. \end{aligned}$$

Hence  $M(x) = (I - 2uu^T)x$  and the matrix that represents  $M$  is given by  $I - 2uu^T$ .

**Homework 2.2.5.5** (Verbally) describe why  $(I - 2uu^T)^{-1} = I - 2uu^T$  if  $u \in \mathbb{R}^3$  and  $\|u\|_2 = 1$ .

**Solution.** If you take a vector,  $x$ , and reflect it with respect to the mirror defined by  $u$ , and you then reflect the result with respect to the same mirror, you should get the original vector  $x$  back. Hence, the matrix that represents the reflection should be its own inverse.

**Homework 2.2.5.6** Let  $M : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be defined by  $M(x) = (I - 2uu^T)x$ , where  $\|u\|_2 = 1$ . Show that the matrix that represents it is unitary (or, rather, orthogonal since it is in  $\mathbb{R}^{3 \times 3}$ ).

**Solution.**

$$\begin{aligned} (I - 2uu^T)^T(I - 2uu^T) &= \langle (A + B)^T = A^T + B^T \rangle \\ (I^T - (2uu^T)^T)(I - 2uu^T) &= \langle (\alpha AB^T)^T = \alpha BA^T \rangle \\ (I - 2uu^T)(I - 2uu^T) &= \langle \text{distributivity} \rangle \\ (I - 2uu^T) - (I - 2uu^T)(2uu^T) &= \langle \text{distributivity} \rangle \\ I - 2uu^T - 2uu^T + 2uu^T 2uu^T &= \langle u^T u = 1 \rangle \\ I - 4uu^T + 4uu^T &= \langle A - A = 0 \rangle \\ I. \end{aligned}$$

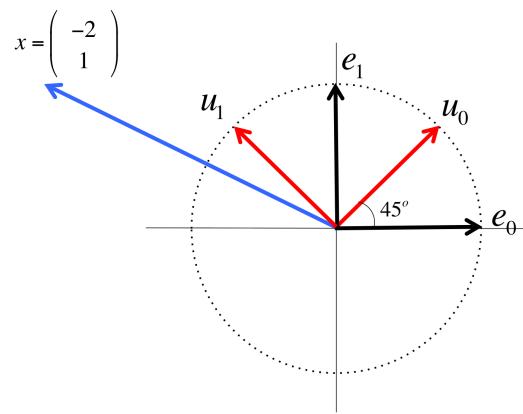
**Remark 2.2.5.1** Unitary matrices in general, and rotations and reflections in particular, will play a key role in many of the practical algorithms we will develop in this course.

## 2.2.6 Change of orthonormal basis



YouTube: <https://www.youtube.com/watch?v=DwTVkdQKJK4>

**Homework 2.2.6.1** Consider the vector  $x = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$  and the following picture that depicts a rotated basis with basis vectors  $u_0$  and  $u_1$ .



What are the coordinates of the vector  $x$  in this rotated system? In other words, find  $\hat{x} = \begin{pmatrix} \hat{x}_0 \\ \hat{x}_1 \end{pmatrix}$  such that  $\hat{x}_0 u_0 + \hat{x}_1 u_1 = x$ .

**Solution.** There are a number of approaches to this. One way is to try to remember the formula you may have learned in a pre-calculus course about change of coordinates. Let's instead start by recognizing (from geometry or by applying the Pythagorean Theorem) that

$$u_0 = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad u_1 = \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Here are two ways in which you can employ what you have discovered in this course:

- Since  $u_0$  and  $u_1$  are orthonormal vectors, you know that

$$\begin{aligned}
 x &= \underbrace{\langle u_0 \text{ and } u_1 \text{ are orthonormal} \rangle}_{\substack{(u_0^T x)u_0 \\ \text{component in the direction of } u_0}} + \underbrace{\langle u_1^T x)u_1 \rangle}_{\substack{\text{component in the direction of } u_1}} \\
 &= \langle \text{instantiate } u_0 \text{ and } u_1 \rangle \\
 &\quad \left( \frac{\sqrt{2}}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} -2 \\ 1 \end{pmatrix} \right) u_0 + \left( \frac{\sqrt{2}}{2} \begin{pmatrix} -1 \\ 1 \end{pmatrix}^T \begin{pmatrix} -2 \\ 1 \end{pmatrix} \right) u_1 \\
 &= \langle \text{evaluate} \rangle \\
 &\quad -\frac{\sqrt{2}}{2}u_0 + \frac{3\sqrt{2}}{2}u_1.
 \end{aligned}$$

- An alternative way to arrive at the same answer that provides more insight. Let  $U = (u_0 | u_1)$ . Then

$$\begin{aligned}
 x &= \langle U \text{ is unitary (or orthogonal since it is real valued)} \rangle \\
 UU^T x &= \langle \text{instantiate } U \rangle \\
 (u_0 | u_1) \left( \frac{u_0^T}{u_1^T} \right) x &= \langle \text{matrix-vector multiplication} \rangle \\
 (u_0 | u_1) \left( \frac{u_0^T x}{u_1^T x} \right) &= \langle \text{instantiate} \rangle \\
 (u_0 | u_1) \left( \begin{array}{c} \frac{\sqrt{2}}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} -2 \\ 1 \end{pmatrix} \\ \frac{\sqrt{2}}{2} \begin{pmatrix} -1 \\ 1 \end{pmatrix}^T \begin{pmatrix} -2 \\ 1 \end{pmatrix} \end{array} \right) &= \langle \text{evaluate} \rangle \\
 (u_0 | u_1) \left( \begin{array}{c} -\frac{\sqrt{2}}{2} \\ \frac{3\sqrt{2}}{2} \end{array} \right) &= \langle \text{simplify} \rangle \\
 (u_0 | u_1) \left( \frac{\sqrt{2}}{2} \begin{pmatrix} -1 \\ 3 \end{pmatrix} \right)
 \end{aligned}$$

Below we compare side-by-side how to describe a vector  $x$  using the standard basis vectors  $e_0, \dots, e_{m-1}$  (on the left) and vectors  $u_0, \dots, u_{m-1}$  (on the right):

The vector  $x = \begin{pmatrix} \chi_0 \\ \vdots \\ \chi_{m-1} \end{pmatrix}$  describes the vector  $x$  in terms of the standard basis vectors  $e_0, \dots, e_{m-1}$ :

$$\begin{aligned} x &= \langle x = Ix = IIx = II^T x \rangle \\ II^T x &= \langle \text{expose columns of } I \rangle \\ \left( e_0 \mid \dots \mid e_{m-1} \right) \begin{pmatrix} e_0^T \\ \vdots \\ e_{m-1}^T \end{pmatrix} x &= \langle \text{evaluate} \rangle \\ \left( e_0 \mid \dots \mid e_{m-1} \right) \begin{pmatrix} e_0^T x \\ \vdots \\ e_{m-1}^T x \end{pmatrix} &= \langle e_j^T x = \chi_j \rangle \\ \left( e_0 \mid \dots \mid e_{m-1} \right) \begin{pmatrix} \chi_0 \\ \vdots \\ \chi_{m-1} \end{pmatrix} &= \langle \text{evaluate} \rangle \\ \chi_0 e_0 + \chi_1 e_1 + \dots + \chi_{m-1} e_{m-1}. & \end{aligned}$$

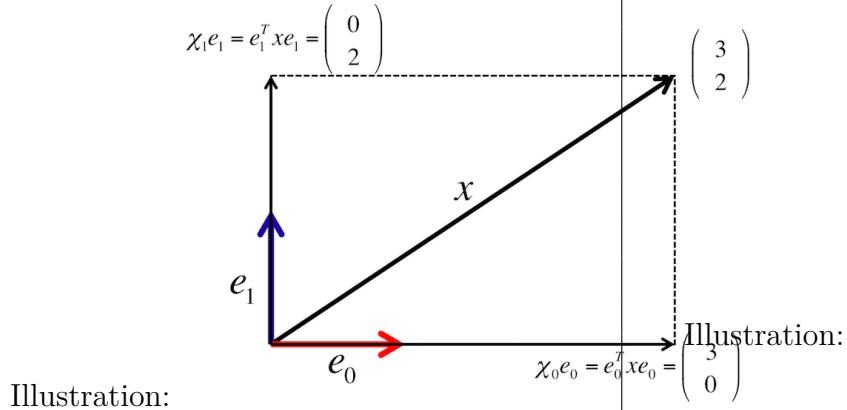


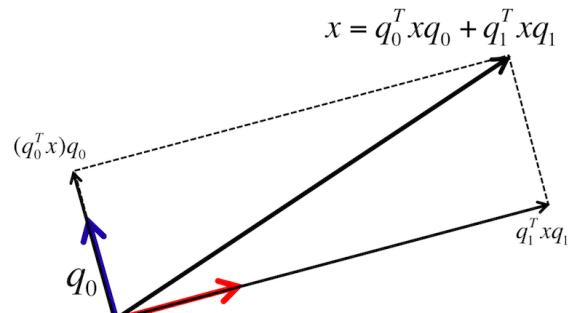
Illustration:

Another way of looking at this is that if  $u_0, u_1, \dots, u_{m-1}$  is an orthonormal basis for  $\mathbb{C}^m$ , then any  $x \in \mathbb{C}^m$  can be written as a linear combination of these vectors:

$$x = \alpha_0 u_0 + \alpha_1 u_1 + \dots + \alpha_{m-1} u_{m-1}.$$

The vector  $\hat{x} = \begin{pmatrix} u_0^T x \\ \vdots \\ u_{m-1}^T x \end{pmatrix}$  describes the vector  $x$  in terms of the orthonormal basis  $u_0, \dots, u_{m-1}$ :

$$\begin{aligned} x &= \langle x = Ix = UU^H x \rangle \\ UU^H x &= \langle \text{expose columns of } U \rangle \\ \left( u_0 \mid \dots \mid u_{m-1} \right) \begin{pmatrix} u_0^H \\ \vdots \\ u_{m-1}^H \end{pmatrix} x &= \langle \text{evaluate} \rangle \\ \left( u_0 \mid \dots \mid u_{m-1} \right) \begin{pmatrix} u_0^H x \\ \vdots \\ u_{m-1}^H x \end{pmatrix} &= \langle \text{evaluate} \rangle \\ u_0^H x u_0 + u_1^H x u_1 + \dots + u_{m-1}^H x u_{m-1}. & \end{aligned}$$



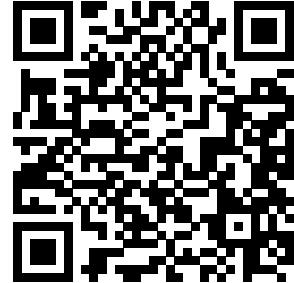
Now,

$$\begin{aligned}
 u_i^H x &= u_i^H (\alpha_0 u_0 + \alpha_1 u_1 + \cdots + \alpha_{i-1} u_{i-1} + \alpha_i u_i + \alpha_{i+1} u_{i+1} + \cdots + \alpha_{m-1} u_{m-1}) \\
 &= \underbrace{\alpha_0 u_i^H u_0}_0 + \underbrace{\alpha_1 u_i^H u_1}_0 + \cdots + \underbrace{\alpha_{i-1} u_i^H u_{i-1}}_0 \\
 &\quad + \underbrace{\alpha_i u_i^H u_i}_1 + \underbrace{\alpha_{i+1} u_i^H u_{i+1}}_0 + \cdots + \underbrace{\alpha_{m-1} u_i^H u_{m-1}}_0 \\
 &= \alpha_i.
 \end{aligned}$$

Thus  $u_i^H x = \alpha_i$ , the coefficient that multiplies  $u_i$ .

**Remark 2.2.6.1** The point is that given vector  $x$  and unitary matrix  $U$ ,  $U^H x$  computes the coefficients for the orthonormal basis consisting of the columns of matrix  $U$ . Unitary matrices allow one to elegantly change between orthonormal bases.

### 2.2.7 Why we love unitary matrices



YouTube: <https://www.youtube.com/watch?v=d8-AeC3Q8Cw>

In Subsection 1.4.1, we looked at how sensitive solving

$$Ax = b$$

is to a change in the right-hand side

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

when  $A$  is nonsingular. We concluded that

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\|A\| \|A^{-1}\|}_{\kappa(A)} \frac{\|\delta b\|}{\|b\|},$$

when an induced matrix norm is used. Let's look instead at how sensitive matrix-vector multiplication is.

**Homework 2.2.7.1** Let  $A \in \mathbb{C}^{n \times n}$  be nonsingular and  $x \in \mathbb{C}^n$  a nonzero vector. Consider

$$y = Ax \quad \text{and} \quad y + \delta y = A(x + \delta x).$$

Show that

$$\frac{\|\delta y\|}{\|y\|} \leq \underbrace{\|A\| \|A^{-1}\|}_{\kappa(A)} \frac{\|\delta x\|}{\|x\|},$$

where  $\|\cdot\|$  is an induced matrix norm.

**Solution.** Since  $x = A^{-1}y$  we know that

$$\|x\| \leq \|A^{-1}\| \|y\|$$

and hence

$$\frac{1}{\|y\|} \leq \|A^{-1}\| \frac{1}{\|x\|}. \quad (2.2.1)$$

Subtracting  $y = Ax$  from  $y + \delta y = A(x + \delta x)$  yields

$$\delta y = A\delta x$$

and hence

$$\|\delta y\| \leq \|A\| \|\delta x\|. \quad (2.2.2)$$

Combining (2.2.1) and (2.2.2) yields the desired result.

There are choices of  $x$  and  $\delta x$  for which the bound is tight.

What does this mean? It means that if as part of an algorithm we use matrix-vector or matrix-matrix multiplication, we risk amplifying relative error by the condition number of the matrix by which we multiply. Now, we saw in Section 1.4 that  $1 \leq \kappa(A)$ . So, if we there are algorithms that only use matrices for which  $\kappa(A) = 1$ , then those algorithms don't amplify relative error.

**Remark 2.2.7.1** We conclude that unitary matrices, which do not amplify the 2-norm of a vector or matrix, should be our tool of choice, whenever practical.

## 2.3 The Singular Value Decomposition

### 2.3.1 The Singular Value Decomposition Theorem



YouTube: <https://www.youtube.com/watch?v=uBo3XAGt24Q>

The following is probably the most important result in linear algebra:

**Theorem 2.3.1.1 Singular Value Decomposition Theorem.** Given  $A \in \mathbb{C}^{m \times n}$  there exist unitary  $U \in \mathbb{C}^{m \times m}$ , unitary  $V \in \mathbb{C}^{n \times n}$ , and  $\Sigma \in \mathbb{R}^{m \times n}$  such that  $A = U\Sigma V^H$ . Here  $\Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right)$  with

$$\Sigma_{TL} = \left( \begin{array}{cccc} \sigma_0 & 0 & \cdots & 0 \\ 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{r-1} \end{array} \right) \quad \text{and} \quad \sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{r-1} > 0.$$

The values  $\sigma_0, \dots, \sigma_{r-1}$  are called the singular values of matrix  $A$ . The columns of  $U$  and  $V$  are called the left and right singular vectors, respectively.

Before proving this theorem, we are going to put some intermediate results in place.

**Remark 2.3.1.2** As the course progresses, we will notice that there is a conflict between the notation that explicitly exposes indices, e.g.,

$$U = \left( \begin{array}{cccc} u_0 & u_1 & \cdots & u_{n-1} \end{array} \right)$$

and the notation we use to hide such explicit indexing, which we call the FLAME notation, e.g.,

$$U = \left( \begin{array}{c|cc} U_0 & u_1 & U_2 \end{array} \right).$$

The two linked by

$$\left( \begin{array}{cc|cc} u_0 & u_{k-1} & u_k & u_{k+1} & u_{n-1} \\ \hline U_0 & & u_1 & U_2 & \end{array} \right).$$

In algorithms that use explicit indexing,  $k$  often is the loop index that identifies where in the matrix or vector the algorithm currently has reached. In the FLAME notation, the index 1 identifies that place. This creates a conflict for the two distinct items that are both indexed with 1, e.g.,  $u_1$  in our example here. It is our experience that learners quickly adapt to this and hence have not tried to introduce even more notation that avoids this conflict. In other words: you will almost always be able to tell from context what is meant. The following lemma and its proof illustrate this further.

**Lemma 2.3.1.3** Given  $A \in \mathbb{C}^{m \times n}$ , with  $1 \leq n \leq m$  and  $A \neq 0$  (the zero matrix), there exist unitary matrices  $\tilde{U} \in \mathbb{C}^{m \times m}$  and  $\tilde{V} \in \mathbb{C}^{n \times n}$  such that

$$A = \tilde{U} \left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array} \right) \tilde{V}^H, \quad \text{where } \sigma_1 = \|A\|_2.$$

*Proof.* In the below proof, it is really important to keep track of when a line is part of the partitioning of a matrix or vector, and when it denotes scalar division.

Choose  $\sigma_1$  and  $\tilde{v}_1 \in \mathbb{C}^n$  such that

- $\|\tilde{v}_1\|_2 = 1$ ; and
- $\sigma_1 = \|A\tilde{v}_1\|_2 = \|A\|_2$ .

In other words,  $\tilde{v}_1$  is the vector that maximizes  $\max_{\|x\|_2=1} \|Ax\|_2$ .

Let  $\tilde{u}_1 = A\tilde{v}_1/\sigma_1$ . Then

$$\|\tilde{u}_1\|_2 = \|A\tilde{v}_1\|_2/\sigma_1 = \|A\tilde{v}_1\|_2/\|A\|_2 = \|A\|_2/\|A\|_2 = 1.$$

Choose  $\tilde{U}_2 \in \mathbb{C}^{m \times (m-1)}$  and  $\tilde{V}_2 \in \mathbb{C}^{n \times (n-1)}$  so that

$$\tilde{U} = \left( \begin{array}{c|c} \tilde{u}_1 & \tilde{U}_2 \end{array} \right) \text{ and } \tilde{V} = \left( \begin{array}{c|c} \tilde{v}_1 & \tilde{V}_2 \end{array} \right)$$

are unitary. Then

$$\begin{aligned} \tilde{U}^H A \tilde{V} &= \left( \begin{array}{c|c} \tilde{u}_1 & \tilde{U}_2 \end{array} \right)^H A \left( \begin{array}{c|c} \tilde{v}_1 & \tilde{V}_2 \end{array} \right) \\ &= \left( \begin{array}{c|c} \tilde{u}_1^H A \tilde{v}_1 & \tilde{u}_1^H A \tilde{V}_2 \\ \hline \tilde{U}_2^H A \tilde{v}_1 & \tilde{U}_2^H A \tilde{V}_2 \end{array} \right) \\ &= \left( \begin{array}{c|c} \tilde{u}_1^H A \tilde{v}_1 & \tilde{u}_1^H A \tilde{V}_2 \\ \hline \sigma_1 \tilde{u}_1^H \tilde{u}_1 & \tilde{U}_2^H A \tilde{V}_2 \end{array} \right) \\ &= \left( \begin{array}{c|c} \sigma_1 \tilde{u}_1^H \tilde{u}_1 & \tilde{u}_1^H A \tilde{V}_2 \\ \hline 0 & B \end{array} \right), \end{aligned}$$

< instantiate >  
< multiply out >  
<  $A\tilde{v}_1 = \sigma_1\tilde{u}_1$  >  
<  $\tilde{u}_1^H \tilde{u}_1 = 1; \tilde{U}_2^H \tilde{u}_1 = 0$ ; pick  $w = \tilde{V}_2^H A^H \tilde{u}_1$  and  $B = \tilde{U}_2^H A \tilde{V}_2$  >

where  $w = \tilde{V}_2^H A^H \tilde{u}_1$  and  $B = \tilde{U}_2^H A \tilde{V}_2$ .

We will now argue that  $w = 0$ , the zero vector of appropriate size:

$$\begin{aligned}
 \sigma_1^2 &= <\text{assumption}> \\
 \|A\|_2^2 &= <\text{2-norm is invariant under multiplication by unitary matrix}> \\
 \|\tilde{U}^H A \tilde{V}\|_2^2 &= <\text{definition of } \|\cdot\|_2> \\
 \max_{x \neq 0} \frac{\|\tilde{U}^H A \tilde{V} x\|_2^2}{\|x\|_2^2} &= <\text{see above}> \\
 \max_{x \neq 0} \frac{\left\| \begin{pmatrix} \sigma_1 & w^H \\ 0 & B \end{pmatrix} x \right\|_2^2}{\|x\|_2^2} &\geq <\text{x replaced by specific vector}> \\
 \left\| \begin{pmatrix} \sigma_1 & w^H \\ 0 & B \end{pmatrix} \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2^2 &= <\text{multiply out numerator}> \\
 \left\| \begin{pmatrix} \sigma_1^2 + w^H w \\ Bw \end{pmatrix} \right\|_2^2 &\geq <\left\| \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix} \right\|_2^2 = \|\psi_1\|_2^2 + \|y_2\|_2^2 \geq \|\psi_1\|_2^2; \left\| \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2^2 = \sigma_1^2 + w^H w > \\
 (\sigma_1^2 + w^H w)^2 / (\sigma_1^2 + w^H w) &= <\text{algebra}> \\
 \sigma_1^2 + w^H w. &
 \end{aligned}$$

Thus  $\sigma_1^2 \geq \sigma_1^2 + w^H w$  which means that  $w = 0$  (the zero vector) and  $\tilde{U}^H A \tilde{V} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & B \end{pmatrix}$  so that  $A = \tilde{U} \begin{pmatrix} \sigma_1 & 0 \\ 0 & B \end{pmatrix} \tilde{V}^H$ . ■

Hopefully you can see where this is going: If one can recursively find that  $B = U_B \Sigma_B V_B^H$ ,

then

$$\begin{aligned}
 A &= \tilde{U} \left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array} \right) \tilde{V}^H \\
 &= \tilde{U} \left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & U_B \Sigma_B V_B^H \end{array} \right) \tilde{V}^H \\
 &= \underbrace{\tilde{U} \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & U_B \end{array} \right)}_{U} \underbrace{\left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & \Sigma_B \end{array} \right)}_{\Sigma} \underbrace{\left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & V_B^H \end{array} \right) \tilde{V}^H}_{V^H} \\
 &= \underbrace{\tilde{U} \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & U_B \end{array} \right)}_{U} \underbrace{\left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & \Sigma_B \end{array} \right)}_{\Sigma} \underbrace{\left( \tilde{V} \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & V_B \end{array} \right) \right)^H}_{V^H}.
 \end{aligned}$$

The next exercise provides the insight that the values on the diagonal of  $\Sigma$  will be ordered from largest to smallest.

**Homework 2.3.1.1** Let  $A \in \mathbb{C}^{m \times n}$  with  $A = \left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array} \right)$  and assume that  $\|A\|_2 = \sigma_1$ .

ALWAYS/SOMETIMES/NEVER:  $\|B\|_2 \leq \sigma_1$ .

**Solution.** We will employ a proof by contradiction. Assume that  $\|B\|_2 > \sigma_1$ . Then there exists a vector  $z$  with  $\|z\|_2 = 1$  such that  $\|B\|_2 = \|Bz\|_2 = \max_{\|x\|_2=1} \|Bx\|_2$ . But then

$$\begin{aligned}
 \|A\|_2 &= <\text{definition}> \\
 \max_{\|x\|_2=1} \|Ax\|_2 &\geq <\text{pick a specific vector with 2-norm equal to one}> \\
 \left\| A \left( \begin{array}{c} 0 \\ z \end{array} \right) \right\|_2 &= <> \\
 \left\| \left( \begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array} \right) \left( \begin{array}{c} 0 \\ z \end{array} \right) \right\|_2 &= <> \\
 \left\| \left( \begin{array}{c} 0 \\ Bz \end{array} \right) \right\|_2 &= <> \\
 \|Bz\|_2 &= \|B\|_2 > \|A\|_2 = \sigma_1,
 \end{aligned}$$

which is a contradiction.

Hence  $\|B\|_2 \leq \sigma_1$ .

We are now ready to prove the Singular Value Decomposition Theorem.

*Proof of Singular Value Decomposition Theorem for  $n \leq m$ .* We will prove this for  $m \geq n$ , leaving the case where  $m \leq n$  as an exercise.

Proof by induction: Since  $m \geq n$ , we select  $m$  to be arbitrary and induct on  $n$ .

- Base case:  $n = 1$ .

In this case  $A = (a_1)$  where  $a_1 \in \mathbb{C}^m$  is its only column.

Case 1:  $a_1 = 0$  (the zero vector).

Then

$$A = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \underbrace{I_{m \times m}}_U \begin{pmatrix} \sigma_1 & \\ & 0 \end{pmatrix} \underbrace{I_{1 \times 1}}_{V^H}$$

so that  $U = I_{m \times m}$ ,  $V = I_{1 \times 1}$ , and  $\Sigma_{TL}$  is an empty matrix.

Case 2:  $a_1 \neq 0$ .

Then

$$A = \begin{pmatrix} a_1 \\ u_1 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 \\ \vdots \\ 0 \end{pmatrix} (\|a_1\|_2)$$

where  $u_1 = a_1 / \|a_1\|_2$ . Choose  $U_2 \in \mathbb{C}^{m \times (m-1)}$  so that  $U = \begin{pmatrix} u_1 & | & U_2 \end{pmatrix}$  is unitary. Then

$$\begin{aligned} A &= \begin{pmatrix} a_1 \\ u_1 \\ \vdots \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} u_1 \\ \vdots \\ 0 \end{pmatrix} (\|a_1\|_2) \\ &= \begin{pmatrix} u_1 & | & U_2 \end{pmatrix} \begin{pmatrix} \|a_1\|_2 & \\ 0 & \vdots \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix}^H \\ &= U \Sigma V^H, \end{aligned}$$

where

- $U = \begin{pmatrix} u_0 & | & U_1 \end{pmatrix}$ ,
- $\Sigma = \begin{pmatrix} \Sigma_{TL} & \\ 0 & \vdots \end{pmatrix}$  with  $\Sigma_{TL} = \begin{pmatrix} \sigma_1 \\ \vdots \\ 0 \end{pmatrix}$  and  $\sigma_1 = \|a_1\|_2 = \|A\|_2$
- $V = \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix}$ .

- Inductive step:

Assume the result is true for matrices with  $1 \leq k$  columns. Show that it is true for matrices with  $k+1$  columns.

Let  $A \in \mathbb{C}^{m \times (k+1)}$  with  $1 \leq k < n$ .

Case 1:  $A = 0$  (the zero matrix)

Then

$$A = I_{m \times m} \begin{pmatrix} & & \\ & \vdots & \\ 0_{m \times (k+1)} & | & \end{pmatrix} I_{(k+1) \times (k+1)}$$

so that  $U = I_{m \times m}$ ,  $V = I_{(k+1) \times (k+1)}$ , and  $\Sigma_{TL}$  is an empty matrix.

Case 2:  $A \neq 0$ .

Then  $\|A\|_2 \neq 0$ . By [Lemma 2.3.1.3](#), we know that there exist unitary  $\tilde{U} \in \mathbb{C}^{m \times m}$  and  $\tilde{V} \in \mathbb{C}^{(k+1) \times (k+1)}$  such that  $A = \tilde{U} \begin{pmatrix} \sigma_1 & 0 \\ 0 & B \end{pmatrix} \tilde{V}$  with  $\sigma_1 = \|A\|_2$ .

By the inductive hypothesis, there exist unitary  $U_B \in \mathbb{C}^{(m-1) \times (m-1)}$ , unitary  $\check{V}_B \in \mathbb{C}^{k \times k}$ , and  $\Sigma_B \in \mathbb{R}^{(m-1) \times k}$  such that  $B = \check{U} \check{\Sigma} \check{V}^H$  where  $\check{\Sigma} = \begin{pmatrix} \check{\Sigma}_{TL} & | & 0 \\ 0 & | & 0 \end{pmatrix}$ ,  $\check{\Sigma}_{TL} = \text{diag}(\sigma_2, \dots, \sigma_{r-1})$ , and  $\sigma_2 \geq \dots \geq \sigma_{r-1} > 0$ .

Now, let

$$U = \tilde{U} \begin{pmatrix} 1 & | & 0 \\ 0 & | & \check{U} \end{pmatrix}, V = \tilde{V} \begin{pmatrix} 1 & | & 0 \\ 0 & | & \check{V} \end{pmatrix}, \text{ and } \Sigma_{TL} = \begin{pmatrix} \sigma_1 & | & 0 \\ 0 & | & \check{\Sigma}_{TL} \end{pmatrix}.$$

(There are some really tough to see "checks" in the definition of  $U$ ,  $V$ , and  $\Sigma$ !!) Then  $A = UDV^H$  where  $U$ ,  $V$ , and  $D$  have the desired properties. Key here is that  $\sigma_1 = \|A\|_2 \geq \|B\|_2$  which means that  $\sigma_1 \geq \sigma_2$ .

- By the Principle of Mathematical Induction the result holds for all matrices  $A \in \mathbb{C}^{m \times n}$  with  $m \geq n$ . ■

**Homework 2.3.1.2** Let  $\Sigma = \text{diag}(\sigma_0, \dots, \sigma_{n-1})$ . ALWAYS/SOMETIMES/NEVER:  $\|\Sigma\|_2 = \max_{i=0}^{n-1} |\sigma_i|$ .

**Answer.** ALWAYS

Now prove it.

**Solution.** Yes, you have seen this before, in [Homework 1.3.5.1](#). We repeat it here because of its importance to this topic.

$$\begin{aligned} \|\Sigma\|_2^2 &= \max_{\|x\|_2=1} \|\Sigma x\|_2^2 \\ &= \max_{\|x\|_2=1} \left\| \begin{pmatrix} \sigma_0 & 0 & \cdots & 0 \\ 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{n-1} \end{pmatrix} \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \right\|_2^2 \\ &= \max_{\|x\|_2=1} \left\| \begin{pmatrix} \sigma_0 \chi_0 \\ \sigma_1 \chi_1 \\ \vdots \\ \sigma_{n-1} \chi_{n-1} \end{pmatrix} \right\|_2^2 \\ &= \max_{\|x\|_2=1} \left[ \sum_{j=0}^{n-1} |\sigma_j \chi_j|^2 \right] \\ &= \max_{\|x\|_2=1} \left[ \sum_{j=0}^{n-1} [|\sigma_j|^2 |\chi_j|^2] \right] \\ &\leq \max_{\|x\|_2=1} \left[ \sum_{j=0}^{n-1} \left[ \max_{i=0}^{n-1} |\sigma_i|^2 |\chi_j|^2 \right] \right] \\ &= \max_{\|x\|_2=1} \left[ \max_{i=0}^{n-1} |\sigma_i|^2 \sum_{j=0}^{n-1} |\chi_j|^2 \right] \\ &= \left( \max_{i=0}^{n-1} |\sigma_i| \right)^2 \max_{\|x\|_2=1} \|x\|_2^2 \\ &= \left( \max_{i=0}^{n-1} |\sigma_i| \right)^2. \end{aligned}$$

so that  $\|\Sigma\|_2 \leq \max_{i=0}^{n-1} |\sigma_i|$ .

Also, choose  $j$  so that  $|\sigma_j| = \max_{i=0}^{n-1} |\sigma_i|$ . Then

$$\|\Sigma\|_2 = \max_{\|x\|_2=1} \|\Sigma x\|_2 \geq \|\Sigma e_j\|_2 = \|\sigma_j e_j\|_2 = |\sigma_j| \|e_j\|_2 = |\sigma_j| = \max_{i=0}^{n-1} |\sigma_i|.$$

so that  $\max_{i=0}^{n-1} |\sigma_i| \leq \|\Sigma\|_2 \leq \max_{i=0}^{n-1} |\sigma_i|$ , which implies that  $\|\Sigma\|_2 = \max_{i=0}^{n-1} |\sigma_i|$ .

**Homework 2.3.1.3** Assume that  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  are unitary matrices. Let  $A, B \in \mathbb{C}^{m \times n}$  with  $B = U A V^H$ . Show that the singular values of  $A$  equal the singular values of  $B$ .

**Solution.** Let  $A = U_A \Sigma_A V_A^H$  be the SVD of  $A$ . Then  $B = U U_A \Sigma_A V_A^H V^H = (U U_A) \Sigma_A (V V_A)^H$  where both  $U U_A$  and  $V V_A$  are unitary. This gives us the SVD for  $B$  and it shows that the singular values of  $B$  equal the singular values of  $A$ .

**Homework 2.3.1.4** Let  $A \in \mathbb{C}^{m \times n}$  with  $n \leq m$  and  $A = U \Sigma V^H$  be its SVD.

ALWAYS/SOMETIMES/NEVER:  $A^H = V \Sigma^T U^H$ .

**Answer.** ALWAYS

**Solution.**

$$A^H = (U \Sigma V^H)^H = (V^H)^H \Sigma^T U^H = V \Sigma^T U^H$$

since  $\Sigma$  is real valued. Notice that  $\Sigma$  is only "sort of diagonal" (it is possibly rectangular) which is why  $\Sigma^T \neq \Sigma$ .

**Homework 2.3.1.5** Prove the Singular Value Decomposition Theorem for  $m \leq n$ .

**Hint.** Consider the SVD of  $B = A^H$

**Solution.** Let  $B = A^H$ . Since it is  $n \times m$  with  $n \geq m$  its SVD exists:  $B = U_B \Sigma_B V_B^H$ . Then  $A = B^H = V_B \Sigma_B^T U_B^H$  and hence  $A = U \Sigma V^H$  with  $U = V_B^H$ ,  $\Sigma = \Sigma_B^T$ , and  $V = U_B^H$ .

I believe the following video has material that is better presented in second video of 2.3.2.



YouTube: <https://www.youtube.com/watch?v=ZYzqTC5LeLs>

### 2.3.2 Geometric interpretation



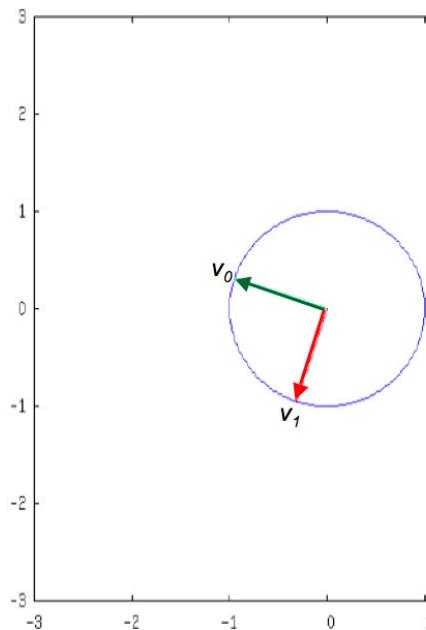
YouTube: <https://www.youtube.com/watch?v=XKhCTtX1z6A>

We will now illustrate what the SVD Theorem tells us about matrix-vector multiplication (linear transformations) by examining the case where  $A \in \mathbb{R}^{2 \times 2}$ . Let  $A = U\Sigma V^T$  be its SVD. (Notice that all matrices are now real valued, and hence  $V^H = V^T$ .) Partition

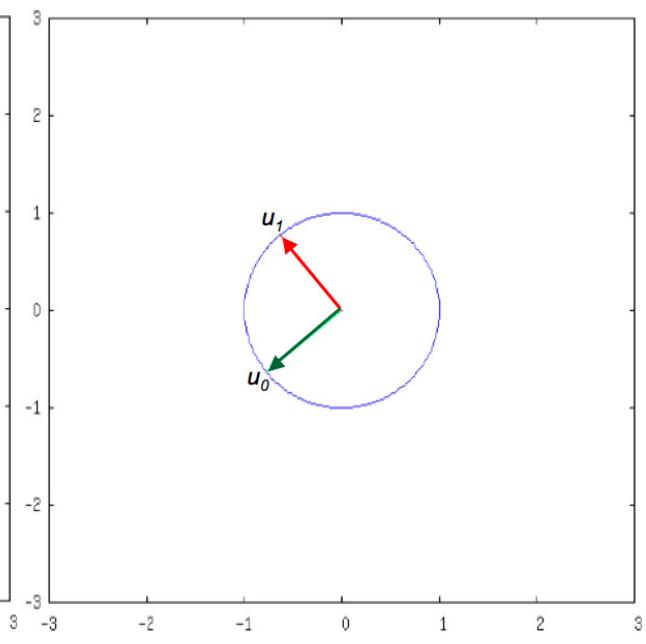
$$A = \left( \begin{array}{c|c} u_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} \sigma_0 & 0 \\ 0 & \sigma_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T.$$

Since  $U$  and  $V$  are unitary matrices,  $\{u_0, u_1\}$  and  $\{v_0, v_1\}$  form orthonormal bases for the range and domain of  $A$ , respectively:

$\mathbb{R}^2$ : Domain of  $A$ :



$\mathbb{R}^2$ : Range (codomain) of  $A$ :



Let us manipulate the decomposition a little:

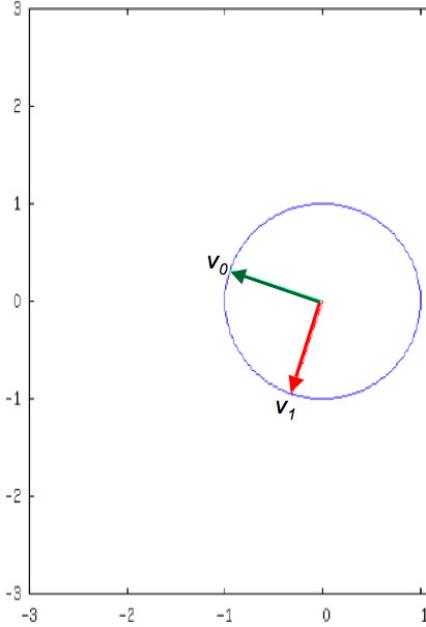
$$\begin{aligned} A &= \left( \begin{array}{c|c} u_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} \sigma_0 & 0 \\ 0 & \sigma_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T \\ &= \left[ \left( \begin{array}{c|c} u_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} \sigma_0 & 0 \\ 0 & \sigma_1 \end{array} \right) \right] \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T \\ &= \left( \begin{array}{c|c} \sigma_0 u_0 & \sigma_1 u_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T. \end{aligned}$$

Now let us look at how  $A$  transforms  $v_0$  and  $v_1$ :

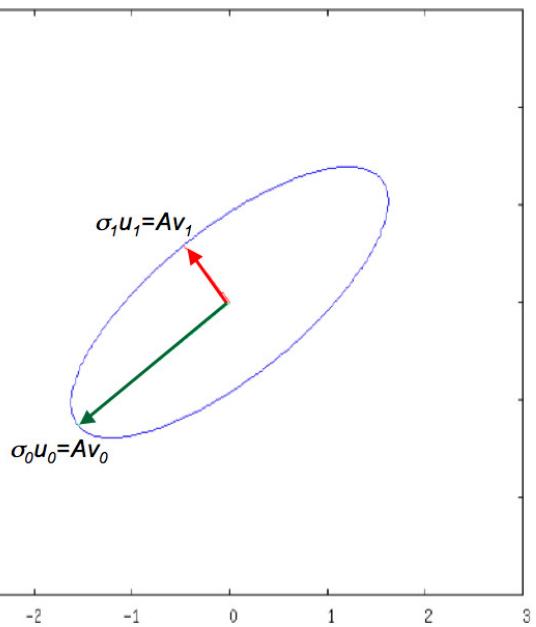
$$Av_0 = \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T v_0 = \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \sigma_0 u_0$$

and similarly  $Av_1 = \sigma_1 u_1$ . This motivates the pictures in (((Unresolved xref, reference "chapter02-figs-SVD"; check spelling or use "provisional" attribute))).

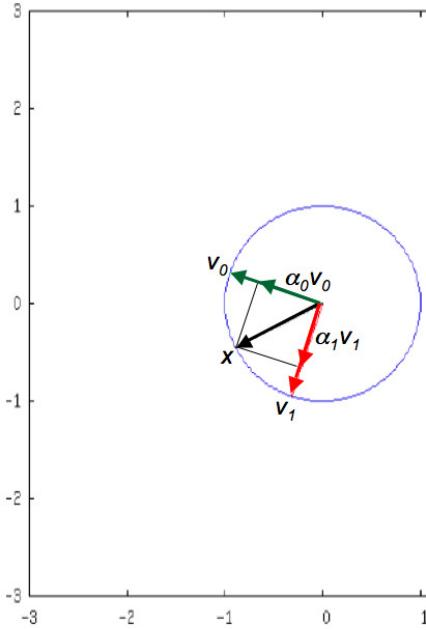
$\mathbb{R}^2$ : Domain of  $A$ :



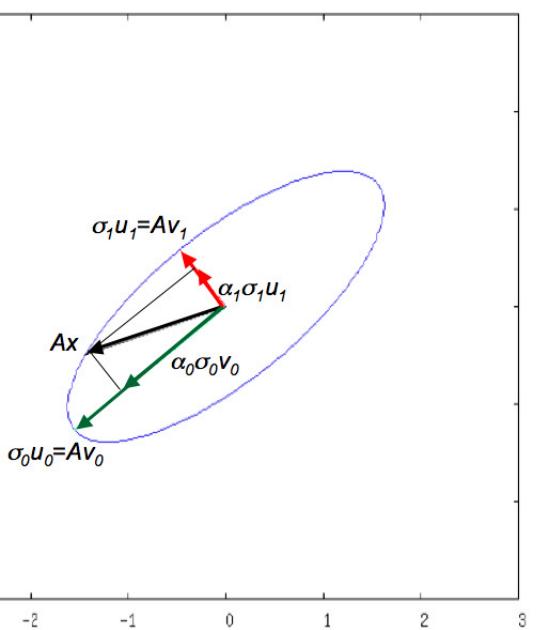
$\mathbb{R}^2$ : Range (codomain) of  $A$ :



$\mathbb{R}^2$ : Domain of  $A$ :



$\mathbb{R}^2$ : Range (codomain) of  $A$ :



**Figure 2.3.2.1** Illustration of how orthonormal vectors  $v_0$  and  $v_1$  are transformed by matrix  $A = U\Sigma V$ .

Next, let us look at how  $A$  transforms any vector with (Euclidean) unit length. Notice that  $x = \begin{pmatrix} \chi_0 \\ \chi_1 \end{pmatrix}$  means that

$$x = \chi_0 e_0 + \chi_1 e_1,$$

where  $e_0$  and  $e_1$  are the unit basis vectors. Thus,  $\chi_0$  and  $\chi_1$  are the coefficients when  $x$  is expressed using  $e_0$  and  $e_1$  as basis. However, we can also express  $x$  in the basis given by  $v_0$  and  $v_1$ :

$$\begin{aligned} x &= \underbrace{VV^T}_{I} x = \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T x = \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right) \left( \begin{array}{c} \frac{v_0^T x}{v_1^T x} \\ \frac{v_1^T x}{v_1^T x} \end{array} \right) \\ &= \underbrace{v_0^T x}_{\alpha_0} v_0 + \underbrace{v_1^T x}_{\alpha_1} v_1 = \alpha_0 v_0 + \alpha_1 v_1 = \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right) \left( \begin{array}{c} \alpha_0 \\ \alpha_1 \end{array} \right). \end{aligned}$$

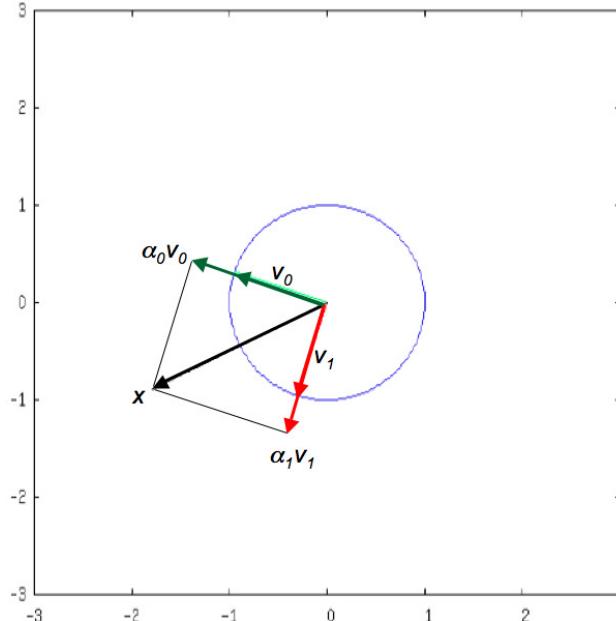
Thus, in the basis formed by  $v_0$  and  $v_1$ , its coefficients are  $\alpha_0$  and  $\alpha_1$ . Now,

$$\begin{aligned} Ax &= \left( \begin{array}{c|c} \sigma_0 u_0 & \sigma_1 u_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T x \\ &= \left( \begin{array}{c|c} \sigma_0 u_0 & \sigma_1 u_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right) \left( \begin{array}{c} \alpha_0 \\ \alpha_1 \end{array} \right) \\ &= \left( \begin{array}{c|c} \sigma_0 u_0 & \sigma_1 u_1 \end{array} \right) \left( \begin{array}{c} \alpha_0 \\ \alpha_1 \end{array} \right) = \alpha_0 \sigma_0 u_0 + \alpha_1 \sigma_1 u_1. \end{aligned}$$

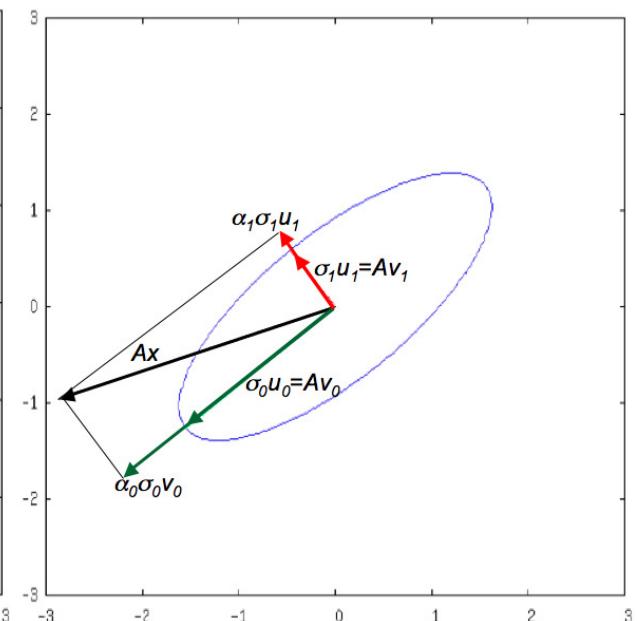
This is illustrated by the following picture, which also captures the fact that the unit ball is mapped to an oval with major axis equal to  $\sigma_0 = \|A\|_2$  and minor axis equal to  $\sigma_1$ , as illustrated in (((Unresolved xref, reference "chapter02-fig-SVD"; check spelling or use "provisional" attribute))) (bottom).

Finally, we show the same insights for general vector  $x$  (not necessarily of unit length):

$\mathbb{R}^2$ : Domain of  $A$ :



$\mathbb{R}^2$ : Range (codomain) of  $A$ :



Another observation is that if one picks the right basis for the domain and codomain, then the computation  $Ax$  simplifies to a matrix multiplication with a diagonal matrix. Let

Let us again illustrate this for nonsingular  $A \in \mathbb{R}^{2 \times 2}$  with

$$A = \underbrace{\begin{pmatrix} u_0 & u_1 \end{pmatrix}}_U \underbrace{\begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix}}_{\Sigma} \underbrace{\begin{pmatrix} v_0 & v_1 \end{pmatrix}}_V^T.$$

Now, if we chose to express  $y$  using  $u_0$  and  $u_1$  as the basis and express  $x$  using  $v_0$  and  $v_1$  as the basis, then

$$\begin{aligned} \underbrace{UU^T}_I y &= U \underbrace{U^T y}_{\hat{y}} = (u_0^T y)u_0 + (u_1^T y)u_1 \\ &= \begin{pmatrix} u_0 & u_1 \end{pmatrix} \begin{pmatrix} u_0^T y \\ u_1^T y \end{pmatrix} = U \underbrace{\begin{pmatrix} \hat{\psi}_0 \\ \hat{\psi}_1 \end{pmatrix}}_{\hat{y}} \\ \underbrace{VV^T}_I x &= V \underbrace{V^T x}_{\hat{x}} = (v_0^T x)v_0 + (v_1^T x)v_1 \\ &= \begin{pmatrix} v_0 & v_1 \end{pmatrix} \begin{pmatrix} v_0^T x \\ v_1^T x \end{pmatrix} = V \underbrace{\begin{pmatrix} \hat{\chi}_0 \\ \hat{\chi}_1 \end{pmatrix}}_{\hat{x}}. \end{aligned}$$

If  $y = Ax$  then

$$U \underbrace{U^T y}_{\hat{y}} = \underbrace{U\Sigma V^T x}_{Ax} = U\Sigma \hat{x}$$

so that

$$\hat{y} = \Sigma \hat{x}$$

and

$$\begin{pmatrix} \hat{\psi}_0 \\ \hat{\psi}_1 \end{pmatrix} = \begin{pmatrix} \sigma_0 \hat{\chi}_0 \\ \sigma_1 \hat{\chi}_1 \end{pmatrix}.$$

**Remark 2.3.2.2** The above discussion shows that if one transforms the input vector  $x$  and output vector  $y$  into the right bases, then the computation  $y := Ax$  can be computed with a diagonal matrix instead:  $\hat{y} := \Sigma \hat{x}$ . Also, solving  $Ax = y$  for  $x$  can be computed by multiplying with the inverse of the diagonal matrix:  $\hat{x} := \Sigma^{-1} \hat{y}$ .

These observations generalize to  $A \in \mathbb{C}^{m \times n}$ : If

$$y = Ax$$

then

$$U^H y = U^H A \underbrace{VV^H}_I x$$

so that

$$\underbrace{U^H y}_{\hat{y}} = \Sigma \underbrace{V^H x}_{\hat{x}}$$

( $\Sigma$  is a rectangular "diagonal" matrix.)



YouTube: <https://www.youtube.com/watch?v=1LpK0dbFX1g>

### 2.3.3 An "algorithm" for computing the SVD

In this unit, we show how the insights from the last unit can be molded into an "algorithm" for computing the SVD. We put algorithm in quotes because while the details of the algorithm mathematically exist, they are actually very difficult to compute in practice. So, this is *not* a practical algorithm. We will not discuss a practical algorithm until the very end of the course, in (((section to be determined))).

We observed that, starting with matrix  $A$ , we can compute one step towards the SVD. If we overwrite  $A$  with the intermediate results, this means that after one step

$$\begin{pmatrix} \alpha_{11} & a_{12}^T \\ a_{21} & A_{22} \end{pmatrix} = (\tilde{u}_1 \ \tilde{U}_2)^H \begin{pmatrix} \hat{\alpha}_{11} & \hat{a}_{12}^T \\ \hat{a}_{21} & \hat{A}_{22} \end{pmatrix} (\tilde{v}_1 \ \tilde{V}_2) = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & B \end{pmatrix},$$

where  $\hat{A}$  allows us to refer to the original contents of  $A$ .

In our explanation in (((Unresolved xref, reference ""; check spelling or use "provisional" attribute))), we then said that the SVD of  $B$ ,  $B = U_B \Sigma_B V_B^H = \Sigma_B$  could be computed, and the desired  $U$  and  $V$  can then be created by computing  $U = \tilde{U} U_B$  and  $V = \tilde{V} V_B$ .

Alternatively, one can accumulate  $U$  and  $V$  every time a new singular value is exposed. In this approach, you start by setting  $U = I_{m \times m}$  and  $V = I_{n \times n}$ . Upon completing the first step (which computes the first singular value), one multiplies  $U$  and  $V$  from the right with the computed  $\tilde{U}$  and  $\tilde{V}$ :

$$\begin{aligned} U &:= U \tilde{U} \\ V &:= V \tilde{V}. \end{aligned}$$

Now, every time another singular value is computed in future steps, the corresponding unitary matrices are similarly accumulated into  $U$  and  $V$ .

To explain this more completely, assume that the process has proceeded for  $k$  steps to the point where

$$\begin{aligned} U &= \left( \begin{array}{c|c} U_L & U_R \end{array} \right) \in \mathbb{C}^{m \times m} & \text{with } U_L \in \mathbb{C}^{m \times k} \\ V &= \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \in \mathbb{C}^{n \times m} & \text{with } V_L \in \mathbb{C}^{n \times k} \\ A &= \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) & \text{with } A_{TL} \in \mathbb{C}^{k \times k}, \end{aligned}$$

where the current contents of  $A$  are

$$\begin{aligned} \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) &= \left( \begin{array}{c|c} U_L & U_R \end{array} \right)^H \left( \begin{array}{c|c} \hat{A}_{TL} & \hat{A}_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \\ &= \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & B \end{array} \right). \end{aligned}$$

This means that in the current step we need to update the contents of  $A_{BR}$  with

$$\tilde{U}^H A \tilde{V} = \left( \begin{array}{c|c} \sigma_{11} & 0 \\ \hline 0 & \tilde{B} \end{array} \right)$$

and update

$$\begin{aligned} \left( \begin{array}{c|c} U_L & U_R \end{array} \right) &:= \left( \begin{array}{c|c} U_L & U_R \end{array} \right) \left( \begin{array}{c|c} I_{k \times k} & 0 \\ \hline 0 & \tilde{U} \end{array} \right) \\ \left( \begin{array}{c|c} V_L & V_R \end{array} \right) &:= \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \left( \begin{array}{c|c} I_{k \times k} & 0 \\ \hline 0 & \tilde{V} \end{array} \right), \end{aligned}$$

which simplify to

$$U_{BR} := U_{BR} \tilde{U} \text{ and } V_{BR} := V_{BR} \tilde{V}.$$

At that point,  $A_{TL}$  is expanded by one row and column, and the left-most columns of  $U_R$  and  $V_R$  are moved to  $U_L$  and  $V_L$ , respectively. If  $A_{BR}$  ever contains a zero matrix, the process completes with  $A$  overwritten with  $\Sigma = U^H \tilde{V}$ . These observations, with all details, are captured in [Figure 2.3.3.1](#). In that figure, the boxes in yellow are assertions that capture the current contents of the variables. Those familiar with proving loops correct will recognize the first and last such box as the precondition and postcondition for the operation and

$$\begin{aligned} \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) &= \left( \begin{array}{c|c} U_L & U_R \end{array} \right)^H \left( \begin{array}{c|c} \hat{A}_{TL} & \hat{A}_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \\ &= \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & B \end{array} \right) \end{aligned}$$

as the loop-invariant that can be used to prove the correctness of the loop via a proof by induction.

$A = \widehat{A}$
$U := I_{m \times m}; V := I_{n \times n}$ $A \rightarrow \left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), U \rightarrow \left( \begin{array}{c c} U_L & U_R \end{array} \right), V \rightarrow \left( \begin{array}{c c} V_L & V_R \end{array} \right)$ where $A_{TL}$ is $0 \times 0$ , $U_L$ is $m \times 0$ , $V_L$ is $n \times 0$
$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} U_L & U_R \end{array} \right)^H \left( \begin{array}{c c} \widehat{A}_{TL} & \widehat{A}_{TR} \\ \hline \widehat{A}_{BL} & \widehat{A}_{BR} \end{array} \right) \left( \begin{array}{c c} V_L & V_R \end{array} \right) = \left( \begin{array}{c c} \Sigma_{TL} & 0 \\ \hline 0 & B \end{array} \right)$ (here $\Sigma_{TL}$ is $0 \times 0$ )
<b>while</b> $\ B\ _2 \neq 0$ $\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} U_L & U_R \end{array} \right)^H \left( \begin{array}{c c} \widehat{A}_{TL} & \widehat{A}_{TR} \\ \hline \widehat{A}_{BL} & \widehat{A}_{BR} \end{array} \right) \left( \begin{array}{c c} V_L & V_R \end{array} \right) = \left( \begin{array}{c c} \Sigma_{TL} & 0 \\ \hline 0 & B \end{array} \right) \wedge \ B\ _2 \neq 0$
$\sigma_{11} = \ A_{BR}\ _2$ if $\sigma_{11} = 0$ break (exit loop) pick $\tilde{v}_1$ s.t. $\ \tilde{v}_1\ _2 = 1$ and $\ A_{BR}\tilde{v}_1\  = \ A_{BR}\ _2 (= \sigma_{11})$ $\tilde{u}_1 := A_{BR}\tilde{v}_1/\sigma_{11}$ pick $\tilde{V}_2$ and $\tilde{U}_2$ s.t. $\tilde{V} = \left( \begin{array}{c c} \tilde{v}_1 & \tilde{V}_2 \end{array} \right)$ and $\tilde{U} = \left( \begin{array}{c c} \tilde{u}_1 & \tilde{U}_2 \end{array} \right)$ are unitary $V_R := V_R\tilde{V}; U_R := U_R\tilde{U}$ $\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c cc} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c c} U_L & U_R \end{array} \right) \rightarrow \left( \begin{array}{c cc} U_0 & u_1 & U_2 \end{array} \right), \left( \begin{array}{c c} V_L & V_R \end{array} \right) \rightarrow \left( \begin{array}{c cc} V_0 & v_1 & V_2 \end{array} \right)$ $\left( \begin{array}{cc} \alpha_{11} & a_{12}^T \\ a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{cc} \sigma_{11} & 0 \\ 0 & \tilde{U}_2^H A_{BR} \tilde{V}_2 \end{array} \right)$ $\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c cc} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c c} U_L & U_R \end{array} \right) \leftarrow \left( \begin{array}{c cc} U_0 & u_1 & U_2 \end{array} \right), \left( \begin{array}{c c} V_L & V_R \end{array} \right) \leftarrow \left( \begin{array}{c cc} V_0 & v_1 & V_2 \end{array} \right)$
$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} U_L & U_R \end{array} \right)^H \left( \begin{array}{c c} \widehat{A}_{TL} & \widehat{A}_{TR} \\ \hline \widehat{A}_{BL} & \widehat{A}_{BR} \end{array} \right) \left( \begin{array}{c c} V_L & V_R \end{array} \right) = \left( \begin{array}{c c} \Sigma_{TL} & 0 \\ \hline 0 & B \end{array} \right)$
<b>endwhile</b>
$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} U_L & U_R \end{array} \right)^H \left( \begin{array}{c c} \widehat{A}_{TL} & \widehat{A}_{TR} \\ \hline \widehat{A}_{BL} & \widehat{A}_{BR} \end{array} \right) \left( \begin{array}{c c} V_L & V_R \end{array} \right) = \left( \begin{array}{c c} \Sigma_{TL} & 0 \\ \hline 0 & B \end{array} \right) \wedge \ B\ _2 = 0$
$\underbrace{\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)}_A = \underbrace{\left( \begin{array}{c c} U_L & U_R \end{array} \right)}_{U^H} \underbrace{\left( \begin{array}{c c} \widehat{A}_{TL} & \widehat{A}_{TR} \\ \hline \widehat{A}_{BL} & \widehat{A}_{BR} \end{array} \right)}_{\widehat{A}} \underbrace{\left( \begin{array}{c c} V_L & V_R \end{array} \right)}_V = \underbrace{\left( \begin{array}{c c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right)}_{\Sigma}$

**Figure 2.3.3.1** Algorithm for computing the SVD of  $A$ , overwriting  $A$  with  $\Sigma$ . In the yellow boxes are assertions regarding the contents of the various matrices.

The reason this algorithm is not practical is that many of the steps are easy to state mathematically, but difficult (computationally expensive) to compute in practice. In particular:

- Computing  $\|A_{BR}\|_2$  is tricky and as a result, so is computing  $\tilde{v}_1$ .
- Given a vector, determining a unitary matrix with that vector as its first column is computationally expensive.
- Assuming for simplicity that  $m = n$ , even if all other computations were free, computing the product  $A_{22} := \tilde{U}_2^H A_{BR} \tilde{V}_2$  requires  $O((m - k)^3)$  operations. This means that the entire algorithm requires  $O(m^4)$  computations, which is prohibitively expensive when  $n$  gets large. (We will see that most practical algorithms discussed in this course cost  $O(m^3)$  operations or less.)

Later in this course, we will discuss an algorithm that has an effective cost of  $O(m^3)$  (when  $m = n$ ).

**Ponder This 2.3.3.1** An implementation of the "algorithm" in Figure 2.3.3.1, using our FLAME API for Matlab (FLAME@lab) [2] that allows the code to closely resemble the algorithm as we present it, is given in mySVD.m (Assignments/Week02/matlab/mySVD.m). This implementation depends of routines in subdirectory Assignments/flameatlab being in the path. Examine this code. What do you notice? Execute it with

```
m = 5;
n = 4;
A = rand( m, n );      % create m x n random matrix
[ U, Sigma, V ] = mySVD( A )
```

Then check whether the resulting matrices form the SVD:

```
norm( A - U * Sigma * V' )
```

and whether  $U$  and  $V$  are unitary

```
norm( eye( n,n ) - V' * V );
norm( eye( m,m ) - U' * U );
```

## 2.3.4 The Reduced Singular Value Decomposition



YouTube: <https://www.youtube.com/watch?v=HAAh4IsIdsY>

**Corollary 2.3.4.1 Reduced Singular Value Decomposition.** Let  $A \in \mathbb{C}^{m \times n}$  and  $r = \text{rank}(A)$ . There exist orthonormal matrix  $U_L \in \mathbb{C}^{m \times r}$ , orthonormal matrix  $V_L \in \mathbb{C}^{n \times r}$ , and matrix  $\Sigma_{TL} \in \mathbb{R}^{r \times r}$  with  $\Sigma_{TL} = \text{diag}(\sigma_0, \dots, \sigma_{r-1})$  and  $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{r-1} > 0$  such that  $A = U_L \Sigma_{TL} V_L^H$ .

**Homework 2.3.4.1** Prove the above corollary.

**Solution.** Let  $A = U \Sigma V^H = \left( \begin{array}{c|c} U_L & U_R \end{array} \right) \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ 0 & 0 \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right)^H$  be the SVD of  $A$ , where  $U_L \in \mathbb{C}^{m \times r}$ ,  $V_L \in \mathbb{C}^{n \times r}$  and  $\Sigma_{TL} \in \mathbb{C}^{r \times r}$  with  $\Sigma_{TL} = \text{diag}(\sigma_0, \sigma_1, \dots, \sigma_{r-1})$  and  $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{r-1} > 0$ . Then

$$\begin{aligned} A &= <\text{SVD of } A> \\ U \Sigma V^T &= <\text{Partitioning}> \\ \left( \begin{array}{c|c} U_L & U_R \end{array} \right) \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ 0 & 0 \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right)^H \\ &= <\text{partitioned matrix-matrix multiplication}> \\ U_L \Sigma_{TL} V_L^H. \end{aligned}$$

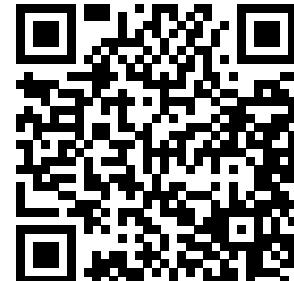
**Corollary 2.3.4.2** Let  $A = U_L \Sigma_{TL} V_L^H$  be the Reduced SVD with  $U_L = \left( \begin{array}{c|c|c|c} u_0 & \cdots & | & u_{r-1} \end{array} \right)$ ,  $V_L = \left( \begin{array}{c|c|c|c} v_0 & \cdots & | & v_{r-1} \end{array} \right)$ , and  $\Sigma_{TL} = \left( \begin{array}{c|c|c|c} \sigma_0 & & & \\ & \ddots & & \\ & & \sigma_{r-1} & \end{array} \right)$ . Then

$$A = \sigma_0 u_0 v_0^H + \cdots + \sigma_{r-1} u_{r-1} v_{r-1}^H.$$

**Remark 2.3.4.3** This last result establishes that any matrix  $A$  with rank  $r$  can be written as a linear combination of  $r$  outer products:

$$A = \underbrace{\frac{\sigma_0 u_0 v_0^H}{\sigma_0 | \rule[1.5ex]{0pt}{0pt} \cdots}}_{\sigma_0 | \rule[1.5ex]{0pt}{0pt}} + \underbrace{\frac{\sigma_1 u_1 v_1^H}{\sigma_1 | \rule[1.5ex]{0pt}{0pt} \cdots}}_{\sigma_1 | \rule[1.5ex]{0pt}{0pt}} + \cdots + \underbrace{\frac{\sigma_{r-1} u_{r-1} v_{r-1}^H}{\sigma_{r-1} | \rule[1.5ex]{0pt}{0pt}}}_{\sigma_{r-1} | \rule[1.5ex]{0pt}{0pt}}.$$

### 2.3.5 SVD of nonsingular matrices



YouTube: <https://www.youtube.com/watch?v=5Gvmtll5T3k>

**Homework 2.3.5.1** Let  $A \in \mathbb{C}^{m \times m}$  and  $A = U\Sigma V^H$  be its SVD.

TRUE/FALSE:  $A$  is nonsingular if and only if  $\Sigma$  is nonsingular.

**Answer.** TRUE

**Solution.**  $\Sigma = U^H A V$ . The product of square matrices is nonsingular if and only if each individual matrix is nonsingular. Since  $U$  and  $V$  are unitary, they are nonsingular.

**Homework 2.3.5.2** Let  $A \in \mathbb{C}^{m \times m}$  and  $A = U\Sigma V^H$  be its SVD with

$$\Sigma = \begin{pmatrix} \sigma_0 & 0 & \cdots & 0 \\ 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{m-1} \end{pmatrix}$$

TRUE/FALSE:  $A$  is nonsingular if and only if  $\sigma_{m-1} \neq 0$ .

**Answer.** TRUE

**Solution.** By the last homework,  $A$  is nonsingular if and only if  $\Sigma$  is nonsingular. A diagonal matrix is nonsingular if and only if its diagonal elements are all nonzero.  $\sigma_0 \geq \cdots \geq \sigma_{m-1} \geq 0$ . Hence the diagonal elements of  $\Sigma$  are nonzero if and only if  $\sigma_{m-1} \neq 0$ .

**Homework 2.3.5.3** Let  $A \in \mathbb{C}^{m \times m}$  be nonsingular and  $A = U\Sigma V^H$  be its SVD.

ALWAYS/SOMETIMES/NEVER: The SVD of  $A^{-1}$  equals  $V\Sigma^{-1}U^H$ .

**Answer.** SOMETIMES

Explain it!

**Solution.** It would seem that the answer is ALWAYS:  $A^{-1} = (U\Sigma V^H)^{-1} = (V^H)^{-1}\Sigma^{-1}U^{-1} =$

$V\Sigma^{-1}U^H$  with

$$\begin{aligned}\Sigma^{-1} &= \left\langle \begin{array}{c|ccccc} \sigma_0 & 0 & \cdots & 0 \\ \hline 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & \sigma_{m-1} \end{array} \right\rangle^{-1} \\ &= \left\langle \begin{array}{c|ccccc} 1/\sigma_0 & 0 & \cdots & 0 \\ \hline 0 & 1/\sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & 1/\sigma_{m-1} \end{array} \right\rangle.\end{aligned}$$

However, the SVD requires the diagonal elements to be positive and ordered from largest to smallest.

So, only if  $\sigma_0 = \sigma_1 = \cdots = \sigma_{m-1}$  is it the case that  $V\Sigma^{-1}V^H$  is the SVD of  $A^{-1}$ . In other words, when  $\Sigma = \sigma_0 I$ .

**Homework 2.3.5.4** Let  $A \in \mathbb{C}^{m \times m}$  be nonsingular and

$$\begin{aligned}A &= U\Sigma V^H \\ &= \left( \begin{array}{c|cc|c} u_0 & \cdots & u_{n-1} \end{array} \right) \left( \begin{array}{c|cc|c} \sigma_0 & \cdots & 0 \\ \hline \vdots & \ddots & \vdots \\ \hline 0 & \cdots & \sigma_{m-1} \end{array} \right) \left( \begin{array}{c|cc|c} v_0 & \cdots & v_{m-1} \end{array} \right)^H\end{aligned}$$

be its SVD.

The SVD of  $A^{-1}$  is given by (indicate all correct answers):

1.  $V\Sigma^{-1}U^H$ .

2.  $\left( \begin{array}{c|cc|c} v_0 & \cdots & v_{m-1} \end{array} \right) \left( \begin{array}{c|cc|c} 1/\sigma_0 & \cdots & 0 \\ \hline \vdots & \ddots & \vdots \\ \hline 0 & \cdots & 1/\sigma_{m-1} \end{array} \right) \left( \begin{array}{c|cc|c} u_0 & \cdots & u_{m-1} \end{array} \right)^H$

3.  $\left( \begin{array}{c|cc|c} v_{m-1} & \cdots & v_0 \end{array} \right) \left( \begin{array}{c|cc|c} 1/\sigma_{m-1} & \cdots & 0 \\ \hline \vdots & \ddots & \vdots \\ \hline 0 & \cdots & 1/\sigma_0 \end{array} \right) \left( \begin{array}{c|cc|c} u_{m-1} & \cdots & u_0 \end{array} \right)^H$ .

4.  $(VP^H)(P\Sigma^{-1}P^H)(UP^H)^H$  where  $P = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{pmatrix}$

**Answer.** 3. and 4.

Explain it!

**Solution.** This question is a bit tricky.

1. It is the case that  $A^{-1} = V\Sigma^{-1}U^H$ . However, the diagonal elements of  $\Sigma^{-1}$  are ordered from smallest to largest, and hence this is not its SVD.
2. This is just Answer 1. but with the columns of  $U$  and  $V$ , and the elements of  $\Sigma$ , exposed.
3. This answer corrects the problems with the previous two answers: it reorders columns of  $U$  and  $V$  so that the diagonal elements of  $\Sigma$  end up ordered from largest to smallest.
4. This answer is just a reformulation of the last answer.

**Homework 2.3.5.5** Let  $A \in \mathbb{C}^{m \times m}$  be nonsingular. TRUE/FALSE:  $\|A^{-1}\|_2 = 1 / \min_{\|x\|_2=1} \|Ax\|_2$ .

**Answer.** TRUE

**Solution.**

$$\begin{aligned}
 & \|A^{-1}\|_2 \\
 &= <\text{definition}> \\
 & \max_{x \neq 0} \frac{\|A^{-1}x\|_2}{\|x\|_2} \\
 &= <\text{algebra}> \\
 & \max_{x \neq 0} \frac{\frac{1}{\|x\|_2}}{\frac{\|A^{-1}x\|_2}{\|x\|_2}} \\
 &= <\text{algebra}> \\
 & \frac{1}{\min_{x \neq 0} \frac{\|x\|_2}{\|A^{-1}x\|_2}} \\
 &= <\text{substitute } z = A^{-1}x> \\
 & \frac{1}{\min_{Az \neq 0} \frac{\|Az\|_2}{\|z\|_2}} \\
 &= <\text{A is nonsingular}> \\
 & \frac{1}{\min_{z \neq 0} \frac{\|Az\|_2}{\|z\|_2}} \\
 &= <x = z/\|z\|_2> \\
 & \frac{1}{\min_{\|x\|_2=1} \|Ax\|_2}
 \end{aligned}$$

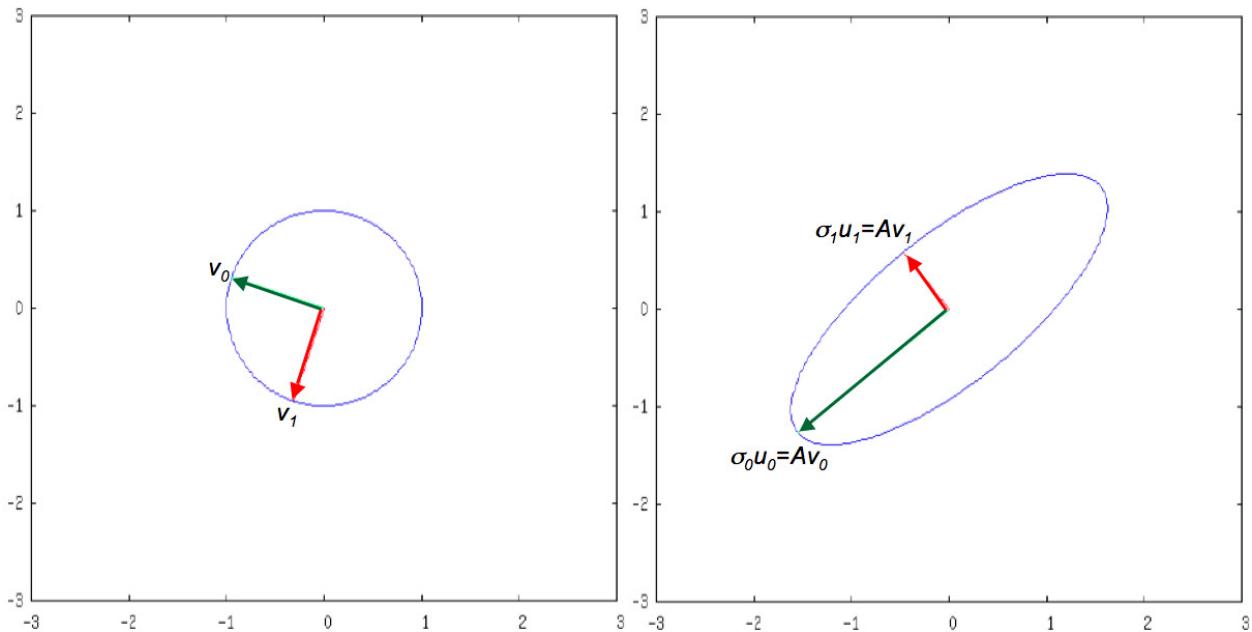
In Subsection 2.3.2, we discussed the case where  $A \in \mathbb{R}^{2 \times 2}$ . Letting  $A = U\Sigma V^T$  and partitioning

$$A = \left( \begin{array}{c|c} u_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} \sigma_0 & 0 \\ 0 & \sigma_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T$$

yielded the pictures

$\mathbb{R}^2$ : Domain of  $A$ :

$\mathbb{R}^2$ : Range (codomain) of  $A$ :



This captures what the condition number  $\kappa_2(A) = \sigma_0/\sigma_{n-1}$  captures: how elongated the oval that equals the image of the unit ball is. The more elongated, the greater the ratio  $\sigma_0/\sigma_{n-1}$ , and the worse the condition number of the matrix. In the limit, when  $\sigma_{n-1} = 0$ , the unit ball is mapped to a lower dimensional set, meaning that the transformation cannot be "undone."

**Ponder This 2.3.5.6** For the 2D problem discussed in this unit, what would the image of the unit ball look like as  $\kappa_2(A) \rightarrow \infty$ ? When is  $\kappa_2(A) = \infty$ ?

## 2.3.6 Best rank-k approximation



YouTube: <https://www.youtube.com/watch?v=sN0DKG8vPhQ>

We are now ready to answer the question "How do we find the best rank-k approximation for a picture (or, more generally, a matrix)? " posed in [Subsection 2.1.1](#).

**Theorem 2.3.6.1** Given  $A \in \mathbb{C}^{m \times n}$ , let  $A = U\Sigma V^H$  be its SVD. Assume the entries on the main diagonal of  $\Sigma$  are  $\sigma_0, \dots, \sigma_{\min(m,n)-1}$  with  $\sigma_0 \geq \dots \geq \sigma_{\min(m,n)-1} \geq 0$ . Given  $k$  such that  $0 \leq k \leq \min(m, n)$ , partition

$$U = \left( \begin{array}{c|c} U_L & U_R \end{array} \right), V = \left( \begin{array}{c|c} V_L & V_R \end{array} \right), \text{ and } \Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ 0 & \Sigma_{BR} \end{array} \right),$$

where  $U_L \in \mathbb{C}^{m \times k}$ ,  $V_L \in \mathbb{C}^{n \times k}$ , and  $\Sigma_{TL} \in \mathbb{C}^{k \times k}$ . Then

$$B = U_L \Sigma_{TL} V_L^H$$

is the matrix in  $\mathbb{C}^{m \times n}$  closest to  $A$  in the following sense:

$$\|A - B\|_2 = \min_{\substack{C \in \mathbb{C}^{m \times n} \\ \text{rank}(C) \leq k}} \|A - C\|_2.$$

In other words,  $B$  is the matrix with rank at most  $k$  that is closest to  $A$  as measured by the 2-norm. Also, for this  $B$ ,

$$\|A - B\|_2 = \begin{cases} \sigma_k & \text{if } k < \min(m, n) \\ 0 & \text{otherwise.} \end{cases}$$

The proof of this theorem builds on the following insight:

**Homework 2.3.6.1** Given  $A \in \mathbb{C}^{m \times n}$ , let  $A = U\Sigma V^H$  be its SVD. Show that

$$Av_j = \sigma_j u_j \text{ for } 0 \leq j < \min(m, n),$$

where  $u_j$  and  $v_j$  equal the columns of  $U$  and  $V$  indexed by  $j$ , and  $\sigma_j$  equals the diagonal element of  $\Sigma$  indexed with  $m$ .

**Solution.** W.l.o.g. assume  $n \leq m$ . Rewrite  $A = U\Sigma V^H$  as  $AV = \Sigma V$ . Then

$$\begin{aligned} AV &= U\Sigma = <\text{partition}> \\ &A \left( \begin{array}{c|c|c} v_0 & \cdots & v_{n-1} \end{array} \right) \\ &= \left( \begin{array}{c|c|c|c|c} u_0 & \cdots & u_{n-1} & u_n & \cdots & u_{m-1} \end{array} \right) \left( \begin{array}{c|c|c} \sigma_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{n-1} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & & 0 \end{array} \right) \\ &= <\text{multiply out}> \\ &\left( \begin{array}{c|c|c} Av_0 & \cdots & Av_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c} \sigma_0 u_0 & \cdots & \sigma_{n-1} u_{n-1} \end{array} \right). \end{aligned}$$

Hence  $Av_j = \sigma_j u_j$  for  $0 \leq j < n$ .

*Proof of Theorem 2.3.6.1.* First, if  $B$  is as defined, then  $\|A - B\|_2 = \sigma_k$ :

$$\begin{aligned}
 & \|A - B\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \|U^H(A - B)V\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \|U^H A V - U^H B V\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \left\| \Sigma - \left( \begin{array}{c|c} U_L & U_R \end{array} \right)^H B \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \right\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \left\| \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ 0 & \Sigma_{BR} \end{array} \right) - \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ 0 & 0 \end{array} \right) \right\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \left\| \left( \begin{array}{c|c} 0 & 0 \\ 0 & \Sigma_{BR} \end{array} \right) \right\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \|\Sigma_{BR}\|_2 \\
 &= \quad \text{ $\Leftrightarrow$ } \\
 & \sigma_k
 \end{aligned}$$

Next, assume that  $C$  has rank  $r \leq k$  and  $\|A - C\|_2 < \|A - B\|_2$ . We will show that this leads to a contradiction.

- The null space of  $C$  has dimension at least  $n - k$  since  $\dim(\mathcal{N}(C)) = n - r$ .
- If  $x \in \mathcal{N}(C)$  then

$$\|Ax\|_2 = \|(A - C)x\|_2 \leq \|A - C\|_2 \|x\|_2 < \sigma_k \|x\|_2.$$

- Partition  $U = \left( \begin{array}{c|c|c|c} u_0 & \cdots & | & u_{m-1} \end{array} \right)$  and  $V = \left( \begin{array}{c|c|c|c} v_0 & \cdots & | & v_{n-1} \end{array} \right)$ . Then  $\|Av_j\|_2 = \|\sigma_j u_j\|_2 = \sigma_j \geq \sigma_k$  for  $j = 0, \dots, k$ .
- Now, let  $x$  be any linear combination of  $v_0, \dots, v_k$ :  $x = \alpha_0 v_0 + \cdots + \alpha_k v_k$ . Notice that

$$\|x\|_2^2 = \|\alpha_0 v_0 + \cdots + \alpha_k v_k\|_2^2 \leq |\alpha_0|^2 + \cdots + |\alpha_k|^2.$$

Then

$$\begin{aligned}
 \|Ax\|_2^2 &= \langle x = \alpha_0 v_0 + \cdots + \alpha_k v_k \rangle \\
 \|A(\alpha_0 v_0 + \cdots + \alpha_k v_k)\|_2^2 &= \langle \text{distributivity} \rangle \\
 \|\alpha_0 A v_0 + \cdots + \alpha_k A v_k\|_2^2 &= \langle A v_j = \sigma_j u_j \rangle \\
 \|\alpha_0 \sigma_0 u_0 + \cdots + \alpha_k \sigma_k u_k\|_2^2 &= \langle \rangle \\
 \|\alpha_0 \sigma_0 u_0\|_2^2 + \cdots + \|\alpha_k \sigma_k u_k\|_2^2 &= \langle \rangle \\
 |\alpha_0|^2 \sigma_0^2 + \cdots + |\alpha_k|^2 \sigma_k^2 &\geq \langle \rangle \\
 (|\alpha_0|^2 + \cdots + |\alpha_k|^2) \sigma_k^2 &= \langle \rangle \\
 \sigma_k^2 \|x\|_2.
 \end{aligned}$$

so that  $\|Ax\|_2 \geq \sigma_k \|x\|_2$ . In other words, vectors in the subspace of all linear combinations of  $\{v_0, \dots, v_k\}$  satisfy  $\|Ax\|_2 \geq \sigma_k \|x\|_2$ . The dimension of this subspace is  $k+1$  (since  $\{v_0, \dots, v_k\}$  form an orthonormal basis).

- Both these subspaces are subspaces of  $\mathbb{C}^n$ . Since their dimensions add up to more than  $n$  there must be at least one nonzero vector  $z$  that satisfies both  $\|Az\|_2 < \sigma_k \|z\|_2$  and  $\|Az\|_2 \geq \sigma_k \|z\|_2$ , which is a contradiction.

■

[Theorem 2.3.6.1](#) tells us how to pick the best approximation to a given matrix of a given desired rank. In Section [Subsection 2.1.1](#) we discussed how a low rank matrix can be used to compress data. The SVD thus gives the best such rank-k approximation. Let us revisit this.

Let  $A \in \mathbb{R}^{m \times n}$  be a matrix that, for example, stores a picture. In this case, the  $i, j$  entry in  $A$  is, for example, a number that represents the grayscale value of pixel  $(i, j)$ .

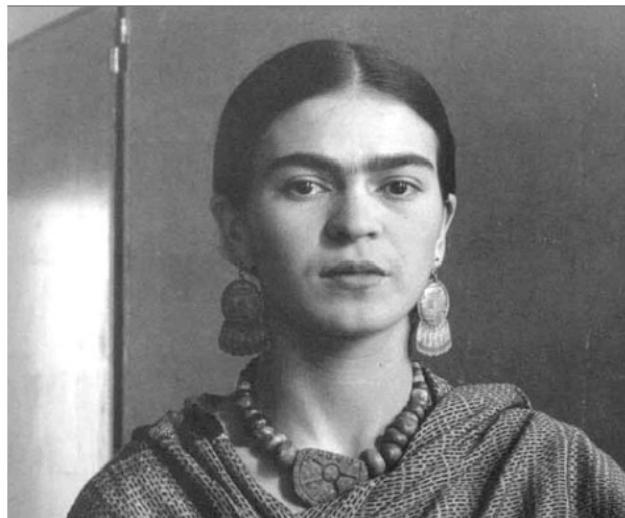
**Homework 2.3.6.2** In Assignments/Week02/matlab execute

```

IMG = imread( 'Frida.jpg' );
A = double( IMG( :, :, 1 ) );
imshow( uint8( A ) )
size( A )

```

to generate the picture of Mexican artist Frida Kahlo



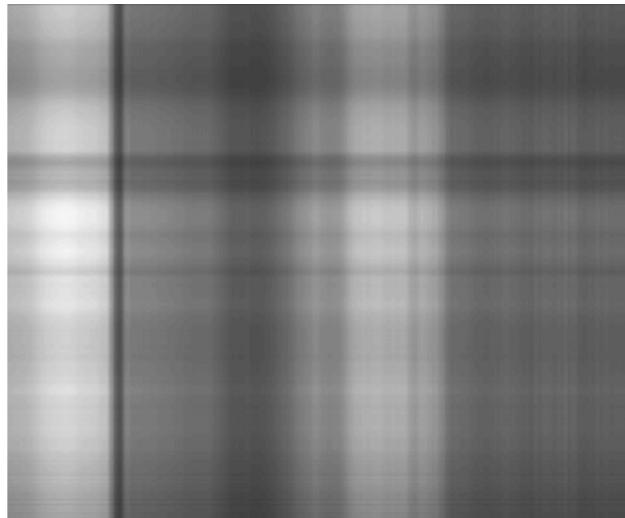
Although the picture is black and white, it was read as if it is a color image, which means a  $m \times n \times 3$  array of pixel information is stored. Setting  $A = \text{IMG}(:, :, 1)$  extracts a single matrix of pixel information. (If you start with a color picture, you will want to approximate  $\text{IMG}(:, :, 1)$ ,  $\text{IMG}(:, :, 2)$ , and  $\text{IMG}(:, :, 3)$  separately.)

Next, compute the SVD of matrix  $A$

```
[ U, Sigma, V ] = svd( A );
```

and approximate the picture with a rank- $k$  update, starting with  $k = 1$ :

```
k = 1
B = uint8( U( :, 1:k ) * Sigma( 1:k, 1:k ) * V( :, 1:k )' );
imshow( B );
```



Repeat this with increasing  $k$ .

```
r = min( size( A ) )
for k=1:r
```

```

imshow( uint8( U( :, 1:k ) * Sigma( 1:k,1:k ) * V( :, 1:k )' ) );
input( strcat( num2str( k ), "    press return" );
end

```

To determine a reasonable value for  $k$ , it helps to graph the singular values:

```

figure
r = min( size( A ) );
plot( [ 1:r ], diag( Sigma ), 'x' );

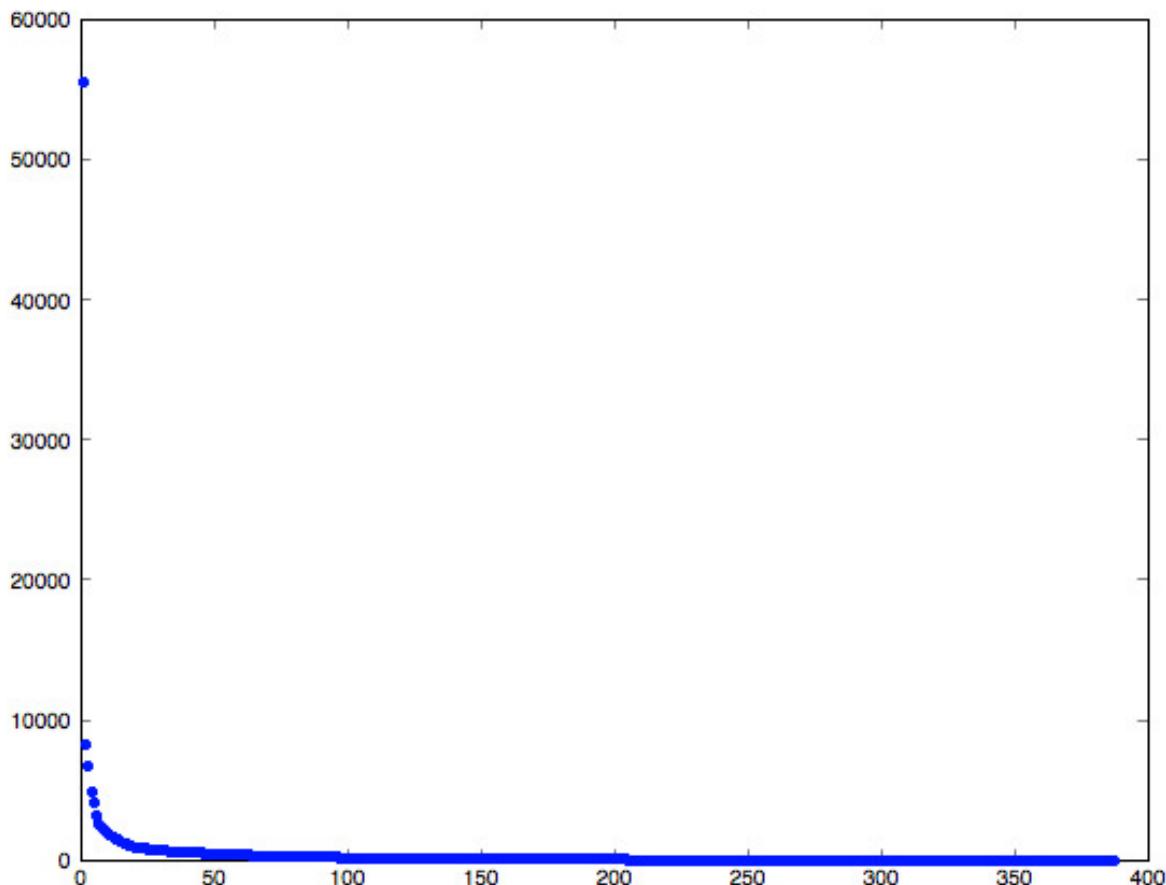
```

Since the singular values span a broad range, we may want to plot them with a log-log plot

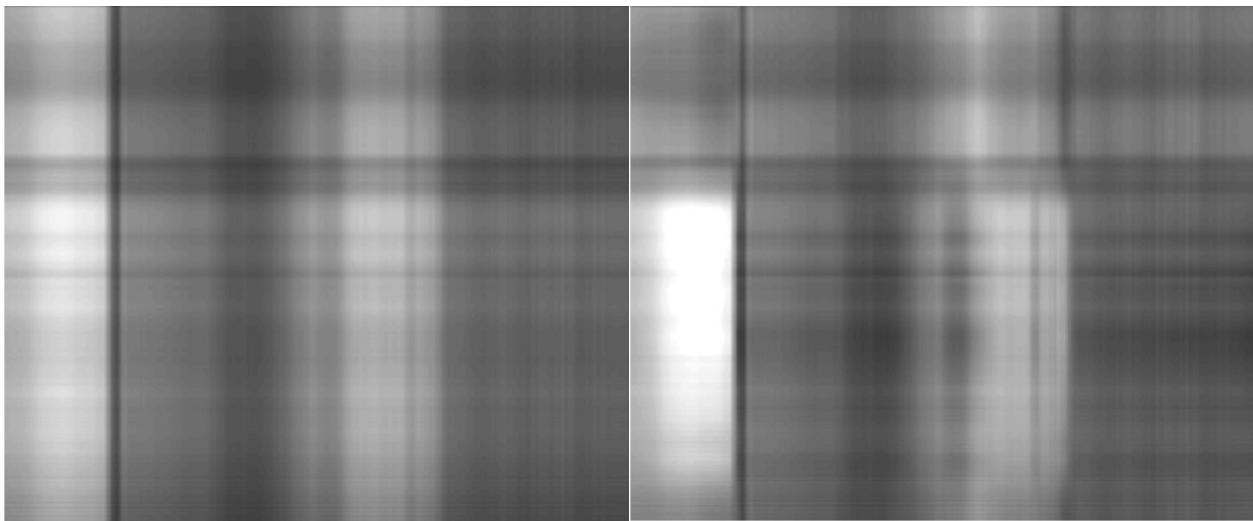
```
loglog( [ 1:r ], diag( Sigma ), 'x' );
```

For this particular matrix (picture), there is no dramatic drop in the singular values that makes it obvious what  $k$  is a natural choice.

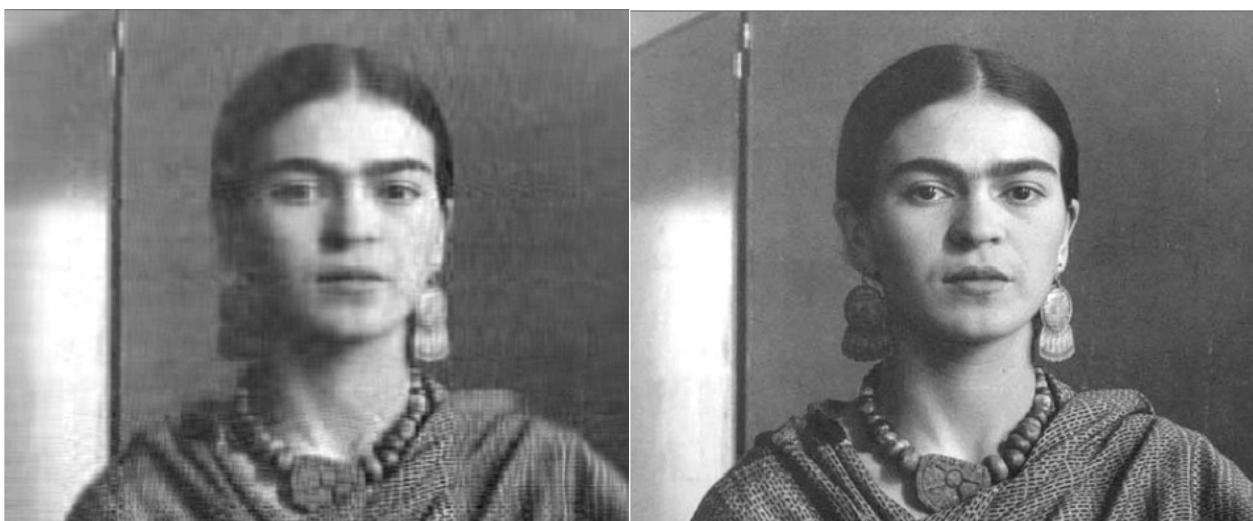
**Solution.**



**Figure 2.3.6.2** Distribution of singular values for the picture.

$k = 1$  $k = 2$  $k = 5$  $k = 10$  $k = 25$ 

Original picture



**Figure 2.3.6.3** Multiple pictures as generated by the code.

## 2.4 Enrichments

### 2.4.1 Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is a standard technique in data science related to the SVD. You may enjoy the article

- [22] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.. Nelson, M. Stephens, C.D. Bustamante, , Nature, 2008.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/>.

In that article, PCA is cast as an eigenvalue problem rather than a singular value problem. Later in the course, in Week 11, we will link these.

## 2.5 Wrap Up

### 2.5.1 Additional homework

**Homework 2.5.1.1**  $U \in \mathbb{C}^{m \times m}$  is unitary if and only if  $(Ux)^H(Uy) = x^H y$  for all  $x, y \in \mathbb{C}^m$ .

**Hint.** Revisit the proof of [Homework 2.2.4.6](#).

**Homework 2.5.1.2** Let  $A, B \in \mathbb{C}^{m \times n}$ . Furthermore, let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary.

TRUE/FALSE:  $UAV^H = B$  iff  $U^H BV = A$ .

**Answer.** TRUE

Now prove it!

**Homework 2.5.1.3** Prove that nonsingular  $A \in \mathbb{C}^{n \times n}$  has condition number  $\kappa_2(A) = 1$  if and only if  $A = \sigma Q$  where  $Q$  is unitary and  $\sigma \in \mathbb{R}$  is positive.

**Hint.** Use the SVD of  $A$ .

**Homework 2.5.1.4** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary.

ALWAYS/SOMETIMES/NEVER: The matrix  $\begin{pmatrix} U & | & 0 \\ 0 & | & V \end{pmatrix}$  is unitary.

**Answer.** ALWAYS

Now prove it!

**Homework 2.5.1.5** Matrix  $A \in \mathbb{R}^{m \times m}$  is a stochastic matrix if and only if it is nonnegative (all its entries are nonnegative) and the entries in its columns sum to one:  $\sum_{0 \leq i < m} \alpha_{i,j} = 1$ . Such matrices are at the core of Markov processes. Show that a matrix  $A$  is both unitary matrix and a stochastic matrix if and only if it is a permutation matrix.

**Homework 2.5.1.6** Pretty sure this is true: Let  $Q$  be a unitary matrix and nonnegative. Show that  $Q$  must be a permutation matrix.

**Homework 2.5.1.7** Show that if  $\|\cdot\|$  is a norm and  $A$  is nonsingular, then  $\|\cdot\|_{A^{-1}}$  defined by  $\|x\|_{A^{-1}} = \|A^{-1}x\|$  is a norm.

Interpret this result in terms of the change of basis of a vector.

**Homework 2.5.1.8** Let  $A \in \mathbb{C}^{m \times m}$  be nonsingular and  $A = U\Sigma V^H$  be its SVD with

$$\Sigma = \begin{pmatrix} \sigma_0 & 0 & \cdots & 0 \\ 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{m-1} \end{pmatrix}$$

The condition number of  $A$  is given by (mark all correct answers):

1.  $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$ .
2.  $\kappa_2(A) = \sigma_0 / \sigma_{m-1}$ .
3.  $\kappa_2(A) = u_0^H A v_0 / u_{m-1}^H A v_{m-1}$ .
4.  $\kappa_2(A) = \max_{\|x\|_2=1} \|Ax\|_2 / \min_{\|x\|_2=1} \|Ax\|_2$ .

(Mark all correct answers.)

**Answer.** All of the above.

**Homework 2.5.1.9** [Theorem 2.2.4.4](#) stated: If  $A \in \mathbb{C}^{m \times m}$  preserves length ( $\|Ax\|_2 = \|x\|_2$  for all  $x \in \mathbb{C}^m$ ), then  $A$  is unitary. Give an alternative proof using the SVD.

**Homework 2.5.1.10** In [Homework 1.3.7.2](#) you were asked to prove that  $\|A\|_2 \leq \|A\|_F$  given  $A \in \mathbb{C}^{m \times n}$ . Give an alternative proof that leverages the SVD.

**Homework 2.5.1.11** In [Homework 1.3.7.3](#), we skipped how the 2-norm bounds the Frobenius norm. We now have the tools to do so elegantly: Prove that, given  $A \in \mathbb{C}^{m \times n}$ ,

$$\|A\|_F \leq \sqrt{r} \|A\|_2,$$

where  $r$  is the rank of matrix  $A$ .

## 2.5.2 Summary

Given  $x, y \in \mathbb{C}^m$

- their dot product (inner product) is defined as

$$x^H y = \bar{x}^T y = \bar{x}^T y = \bar{\chi}_0 \psi_0 + \bar{\chi}_1 \psi_1 + \cdots + \bar{\chi}_{m-1} \psi_{m-1} = \sum_{i=0}^{m-1} \bar{\chi}_i \psi_i.$$

- These vectors are said to be orthogonal (perpendicular) iff  $x^H y = 0$ .

- The component of  $y$  in the direction of  $x$  is given by

$$\frac{x^H y}{x^H x} x = \frac{xx^H}{x^H x} y.$$

The matrix that projects a vector onto the space spanned by  $x$  is given by

$$\frac{xx^H}{x^H x}.$$

- The component of  $y$  in the orthogonal to  $x$  is given by

$$y - \frac{x^H y}{x^H x} x = \left( I - \frac{xx^H}{x^H x} \right) y.$$

The matrix that projects a vector onto the space spanned by  $x$  is given by

$$I - \frac{xx^H}{x^H x}.$$

Given  $u, v \in \mathbb{C}^m$  with  $u$  of unit length

- The component of  $v$  in the direction of  $u$  is given by

$$u^H vu = uu^H v.$$

- The matrix that projects a vector onto the space spanned by  $u$  is given by

$$uu^H$$

- The component of  $v$  orthogonal to  $u$  is given by

$$v - u^H vu = \left( I - uu^H \right) v.$$

- The matrix that projects a vector onto the space spanned by  $x$  is given by

$$I - uu^H$$

Let  $u_0, u_1, \dots, u_{n-1} \in \mathbb{C}^m$ . These vectors are said to be mutually orthonormal if for all  $0 \leq i, j < n$

$$u_i^H u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Then  $Q$  is said to be

- an orthonormal matrix iff  $Q^H Q = I$ .
- a unitary matrix iff  $Q^H Q = I$  and  $m = n$ .
- an orthogonal matrix iff it is a unitary matrix and is real-valued.

Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Then  $Q = \left( q_0 \mid \cdots \mid q_{n-1} \right)$  is orthonormal iff  $\{q_0, \dots, q_{n-1}\}$  are mutually orthonormal.

**Definition 2.5.2.1 Unitary matrix.** Let  $U \in \mathbb{C}^{m \times m}$ . Then  $U$  is said to be a unitary matrix if and only if  $U^H U = I$  (the identity).  $\diamond$

If  $U, V \in \mathbb{C}^{m \times m}$  are unitary, then

- $U^H U = I$ .
- $UU^H = I$ .
- $U^{-1} = U^H$ .
- $U^H$  is unitary.
- $UV$  is unitary.

If  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  are unitary,  $x \in \mathbb{C}^m$ , and  $A \in \mathbb{C}^{m \times n}$ , then

- $\|Ux\|_2 = \|x\|_2$ .
- $\|U^H A\|_2 = \|UA\|_2 = \|AV\|_2 = \|AV^H\|_2 = \|U^H AV\|_2 = \|UAV^H\|_2 = \|A\|_2$ .
- $\|U^H A\|_F = \|UA\|_F = \|AV\|_F = \|AV^H\|_F = \|U^H AV\|_F = \|UAV^H\|_F = \|A\|_F$ .
- $\|U\|_2 = 1$
- $\kappa_2(U) = 1$

Examples of unitary matrices:

- Rotation in 2D:  $\begin{pmatrix} c & -s \\ s & c \end{pmatrix}$ .
- Reflection:  $I - 2uu^H$  where  $u \in \mathbb{C}^m$  and  $\|u\|_2 = 1$ .

Change of orthonormal basis: If  $x \in \mathbb{C}^m$  and  $U = \left( \begin{array}{c|c|c} u_0 & \cdots & u_{m-1} \end{array} \right)$  is unitary, then

$$x = (u_0^H x)u_0 + \cdots + (u_{m-1}^H x)u_{m-1} = \left( \begin{array}{c|c|c} u_0 & \cdots & u_{m-1} \end{array} \right) \underbrace{\begin{pmatrix} u_0^H x \\ \vdots \\ u_{m-1}^H x \end{pmatrix}}_{U^H x} = UU^H x.$$

Let  $A \in \mathbb{C}^{n \times n}$  be nonsingular and  $x \in \mathbb{C}^n$  a nonzero vector. Consider

$$y = Ax \quad \text{and} \quad y + \delta y = A(x + \delta x).$$

Then

$$\frac{\|\delta y\|}{\|y\|} \leq \underbrace{\|A\| \|A^{-1}\|}_{\kappa(A)} \frac{\|\delta x\|}{\|x\|},$$

where  $\|\cdot\|$  is an induced matrix norm.

**Theorem 2.5.2.2 Singular Value Decomposition Theorem.** Given  $A \in \mathbb{C}^{m \times n}$  there exist unitary  $U \in \mathbb{C}^{m \times m}$ , unitary  $V \in \mathbb{C}^{n \times n}$ , and  $\Sigma \in \mathbb{R}^{m \times n}$  such that  $A = U\Sigma V^H$ . Here  $\Sigma = \begin{pmatrix} \Sigma_{TL} & 0 \\ 0 & 0 \end{pmatrix}$  with

$$\Sigma_{TL} = \begin{pmatrix} \sigma_0 & 0 & \cdots & 0 \\ 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{r-1} \end{pmatrix} \quad \text{and} \quad \sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{r-1} > 0.$$

The values  $\sigma_0, \dots, \sigma_{r-1}$  are called the singular values of matrix  $A$ . The columns of  $U$  and  $V$  are called the left and right singular vectors, respectively.

Let  $A \in \mathbb{C}^{m \times n}$  and  $A = U\Sigma V^H$  its SVD with

$$U = \left( \begin{array}{c|c} U_L & U_R \end{array} \right) = \left( \begin{array}{c|c|c} u_0 & \cdots & u_{m-1} \end{array} \right),$$

$$V = \left( \begin{array}{c|c} V_L & V_R \end{array} \right) = \left( \begin{array}{c|c|c} v_0 & \cdots & v_{n-1} \end{array} \right),$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{TL} & 0 \\ 0 & 0 \end{pmatrix}, \text{ where } \Sigma_{TL} = \begin{pmatrix} \sigma_0 & 0 & \cdots & 0 \\ 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{r-1} \end{pmatrix} \quad \text{and} \quad \sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{r-1} > 0.$$

Here  $U_L \in \mathbb{C}^{m \times r}$ ,  $V_L \in \mathbb{C}^{n \times r}$  and  $\Sigma_{TL} \in \mathbb{R}^{r \times r}$ . Then

- $\|A\|_2 = \sigma_0$ . (The 2-norm of a matrix equals the largest singular value.)
- $\text{rank}(A) = r$ .
- $\mathcal{C}(A) = \mathcal{C}(U_L)$ .
- $\mathcal{N}(A) = \mathcal{C}(V_R)$ .
- $\mathcal{R}(A) = \mathcal{C}(V_L)$ .
- Left null-space of  $A = \mathcal{C}(V_R)$ .
- $A^H = V\Sigma^T U^H$ .
- SVD:  $A^H = V\Sigma U^H$ .
- Reduced SVD:  $A = U_L \Sigma_{TL} V_L^H$ .
- 

$$A = \underbrace{\sigma_0 u_0 v_0^H}_{\sigma_0 \mid \text{——}} + \underbrace{\sigma_1 u_1 v_1^H}_{\sigma_1 \mid \text{——}} + \cdots + \underbrace{\sigma_{r-1} u_{r-1} v_{r-1}^H}_{\sigma_{r-1} \mid \text{——}}.$$

- Reduced SVD:  $A^H = V_L \Sigma U_L^H$ .
- If  $m \times m$  matrix  $A$  is nonsingular:  $A^{-1} = V \Sigma^{-1} U^H$ .
- If  $A \in \mathbb{C}^{m \times m}$  then  $A$  is nonsingular if and only if  $\sigma_{m-1} \neq 0$ .
- If  $A \in \mathbb{C}^{m \times m}$  is nonsingular then  $\kappa_2(A) = \sigma_0/\sigma_{m-1}$ .
- (Left) pseudo inverse: if  $A$  has linearly independent columns, then  $A^\dagger = A(A^H A)^{-1} A^H = V \Sigma_{TL} U_L^H$ .
- $v_0$  is the direction of maximal magnification.
- $v_{n-1}$  is the direction of minimal magnification.
- If  $n \leq m$ , then  $A v_j = \sigma_j u_j$ , for  $0 \leq j < n$ .

**Theorem 2.5.2.3** Given  $A \in \mathbb{C}^{m \times n}$ , let  $A = U \Sigma V^H$  be its SVD. Assume the entries on the main diagonal of  $\Sigma$  are  $\sigma_0, \dots, \sigma_{\min(m,n)-1}$  with  $\sigma_0 \geq \dots \geq \sigma_{\min(m,n)-1} \geq 0$ . Given  $k$  such that  $0 \leq k \leq \min(m, n)$ , partition

$$U = \left( \begin{array}{c|c} U_L & U_R \end{array} \right), V = \left( \begin{array}{c|c} V_L & V_R \end{array} \right), \text{ and } \Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & \Sigma_{BR} \end{array} \right),$$

where  $U_L \in \mathbb{C}^{m \times k}$ ,  $V_L \in \mathbb{C}^{n \times k}$ , and  $\Sigma_{TL} \in \mathbb{C}^{k \times k}$ . Then

$$B = U_L \Sigma_{TL} V_L^H$$

is the matrix in  $\mathbb{C}^{m \times n}$  closest to  $A$  in the following sense:

$$\|A - B\|_2 = \min_{\substack{C \in \mathbb{C}^{m \times n} \\ \text{rank}(C) \leq k}} \|A - C\|_2.$$

In other words,  $B$  is the matrix with rank at most  $k$  that is closest to  $A$  as measured by the 2-norm. Also, for this  $B$ ,

$$\|A - B\|_2 = \begin{cases} \sigma_k & \text{if } k < \min(m, n) \\ 0 & \text{otherwise.} \end{cases}$$

# Part II

## Solving Linear Systems

## **Part III**

### **The Algebraic Eigenvalue Problem**

# Appendix A

## Notation

### A.0.1 Householder notation

Alston Householder introduced the convention of labeling matrices with upper case Roman letters ( $A$ ,  $B$ , etc.), vectors with lower case Roman letters ( $a$ ,  $b$ , etc.), and scalars with lower case Greek letters ( $\alpha$ ,  $\beta$ , etc.). When exposing columns or rows of a matrix, the columns of that matrix are usually labeled with the corresponding Roman lower case letter, and the individual elements of a matrix or vector are usually labeled with "the corresponding Greek lower case letter," which we can capture with the triplets  $\{A, a, \alpha\}$ ,  $\{B, b, \beta\}$ , etc.

$$A = \left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right) = \left( \begin{array}{cc|cc|c} \alpha_{0,0} & & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m-1,0} & & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{array} \right)$$

and

$$x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{m-1} \end{pmatrix},$$

where  $\alpha$  and  $\chi$  is the lower case Greek letters "alpha" and "chi," respectively. You will also notice that in this course we start indexing at zero. We mostly adopt this convention (exceptions include  $i$ ,  $j$ ,  $p$ ,  $m$ ,  $n$ , and  $k$ , which usually denote integer scalars.)

# Appendix B

## Knowledge from Numerical Analysis

Typically, an undergraduate numerical analysis course is considered a prerequisite for a graduate level course on numerical linear algebra. There are, however, relatively few concepts from such a course that are needed to be successful in such a course. In this appendix, we very briefly discuss some of these concepts.

### B.0.1 Cost of basic linear algebra operations

### B.0.2 Catastrophic cancellation

Recall that if

$$\chi^2 + \beta\chi + \gamma = 0$$

then the quadratic formula gives the largest root of this quadratic equation:

$$\chi = \frac{-\beta + \sqrt{\beta^2 - 4\gamma}}{2}.$$

**Example B.0.2.1** We use the quadratic equation in the exact order indicated by the parentheses in

$$\chi = \left[ \frac{[-\beta + [\sqrt{[\beta^2] - [4\gamma]}]]}{2} \right],$$

truncating every expression within square brackets to three significant digits, to solve

$$\chi^2 + 25\chi + \gamma = 0$$

$$\begin{aligned} \chi &= \left[ \frac{[-25 + [\sqrt{[25^2] - [4]}]]}{2} \right] = \left[ \frac{[-25 + [\sqrt{625 - 4}]]}{2} \right] \\ &= \left[ \frac{[-25 + [\sqrt{621}]]}{2} \right] = \left[ \frac{[-25 + 24.9]}{2} \right] = \left[ \frac{-0.1}{2} \right] = -0.05. \end{aligned}$$

Now, if you do this to the full precision of a typical calculator, the answer is instead approximately  $-0.040064$ . The relative error we incurred is, approximately,  $0.01/0.04 = 0.25$ .

What is going on here? The problem comes from the fact that there is error in the 24.9 that is encountered after the square root is taken. Since that number is close to in magnitude, but of opposite sign to the  $-25$  to which it is added, the result of  $-25 + 24.9$  is mostly error.

This is known as catastrophic cancellation: adding two nearly equal numbers of opposite sign, at least one of which has some error in it related to roundoff, yields a result with large relative error.

Now, one can use an alternative formula to compute the root:

$$\chi = \frac{-\beta + \sqrt{\beta^2 - 4\gamma}}{2} = \frac{-\beta + \sqrt{\beta^2 - 4\gamma}}{2} \times \frac{-\beta - \sqrt{\beta^2 - 4\gamma}}{-\beta - \sqrt{\beta^2 - 4\gamma}},$$

which yields

$$\chi = \frac{2\gamma}{-\beta - \sqrt{\beta^2 - 4\gamma}}.$$

Carrying out the computations, rounding intermediate results, yields  $-.0401$ . The relative error is now  $0.00004/0.040064 \approx .001$ . It avoids catastrophic cancellation because now the two numbers of nearly equal magnitude are added instead.  $\square$

**Remark B.0.2.2** The point is: if possible, avoid creating small intermediate results that amplify into a large relative error in the final result.

Notice that in this example it is not inherently the case that a small relative change in the input is amplified into a large relative change in the output (as is the case when solving a linear system with a poorly conditioned matrix). The problem is with the standard formula that was used. Later we will see that this is an example of an unstable algorithm.

# Appendix C

## GNU Free Documentation License

Version 1.3, 3 November 2008

Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc. <<http://www.fsf.org/>>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

**0. PREAMBLE.** The purpose of this License is to make a manual, textbook, or other functional and useful document “free” in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of “copyleft”, which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

**1. APPLICABILITY AND DEFINITIONS.** This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “Document”, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “you”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A “Modified Version” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A “Secondary Section” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “Invariant Sections” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “Cover Texts” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “Transparent” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “Opaque”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “Title Page” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

The “publisher” means any person or entity that distributes copies of the Document to the public.

A section “Entitled XYZ” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “Acknowledgements”, “Dedications”, “Endorsements”, or “History”.) To “Preserve the Title” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

**2. VERBATIM COPYING.** You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

**3. COPYING IN QUANTITY.** If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document’s license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated lo-

cation until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

**4. MODIFICATIONS.** You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the “History” section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled “Acknowledgements” or “Dedications”, Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled “Endorsements”. Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled “Endorsements” or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version’s license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties — for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

**5. COMBINING DOCUMENTS.** You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of

the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

**6. COLLECTIONS OF DOCUMENTS.** You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

**7. AGGREGATION WITH INDEPENDENT WORKS.** A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

**8. TRANSLATION.** Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and

disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

**9. TERMINATION.** You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

**10. FUTURE REVISIONS OF THIS LICENSE.** The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy’s public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

**11. RELICENSING.** “Massive Multiauthor Collaboration Site” (or “MMC Site”) means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A “Massive Multiauthor Collaboration” (or “MMC”) contained in the site means any set of copyrightable works thus published on the MMC site.

“CC-BY-SA” means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

“Incorporate” means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is “eligible for relicensing” if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

**ADDENDUM: How to use this License for your documents.** To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright (C) YEAR YOUR NAME.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the “with... Texts.” line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

# References

- [1] Ed Anderson, Zhaojun Bai, James Demmel, Jack J. Dongarra, Jeremy DuCroz, Ann Greenbaum, Sven Hammarling, Alan E. McKenney, Susan Ostrouchov, and Danny Sorensen, *LAPACK Users' Guide*, SIAM, Philadelphia, 1992.
- [2] Paolo Bientinesi, Enrique S. Quintana-Orti, Robert A. van de Geijn, *Representing linear algebra algorithms in code: the FLAME application program interfaces*, ACM Transactions on Mathematical Software (TOMS), 2005
- [3] Paolo Bientinesi, Robert A. van de Geijn, *Goal-Oriented and Modular Stability Analysis*, SIAM Journal on Matrix Analysis and Applications , Volume 32 Issue 1, February 2011.
- [4] Paolo Bientinesi, Robert A. van de Geijn, *The Science of Deriving Stability Analyses*, FLAME Working Note #33. Aachen Institute for Computational Engineering Sciences, RWTH Aachen. TR AICES-2008-2. November 2008.
- [5] Christian Bischof and Charles Van Loan, *The WY Representation for Products of Householder Matrices*, SIAM Journal on Scientific and Statistical Computing, Vol. 8, No. 1, 1987.
- [6] Barry A. Cipra, *The Best of the 20th Century: Editors Name Top 10 Algorithms*, SIAM News, Volume 33, Number 4, 2000. Available from <https://archive.siam.org/pdf/news/637.pdf>.
- [7] A.K. Cline, C.B. Moler, G.W. Stewart, and J.H. Wilkinson, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979).
- [8] Jack J. Dongarra, Jeremy DuCroz, Ann Greenbaum, Sven Hammarling, Alan E. McKenney, Susan Ostrouchov, and Danny Sorensen, *LAPACK Users' Guide*, SIAM, Philadelphia, 1992.
- [9] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Iain Duff, *A Set of Level 3 Basic Linear Algebra Subprograms*, ACM Transactions on Mathematical Software, Vol. 16, No. 1, pp. 1-17, March 1990.
- [10] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Richard J. Hanson, *An Extended Set of {FORTRAN} Basic Linear Algebra Subprograms*, ACM Transactions

- on Mathematical Software, Vol. 14, No. 1, pp. 1-17, March 1988.
- [11] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G. W. Stewart, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, 1979.
  - [12] Leslie V. Foster, *Gaussian elimination with partial pivoting can fail in practice*, SIAM Journal on Matrix Analysis and Applications, 15, 1994.
  - [13] Gene H. Golub and Charles F. Van Loan, *Matrix Computations, Fourth Edition*, Johns Hopkins Press, 2013.
  - [14] Brian C. Gunter, Robert A. van de Geijn, *Parallel out-of-core computation and updating of the QR factorization*, ACM Transactions on Mathematical Software (TOMS), 2005.
  - [15] N. Higham, *A Survey of Condition Number Estimates for Triangular Matrices*, SIAM Review, 1987.
  - [16] C. G. J. Jacobi, *Über ein leichtes Verfahren, die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen*, Crelle's Journal 30, 51-94, 1846.
  - [17] Thierry Joffrain, Tze Meng Low, Enrique S. Quintana-Orti, Robert van de Geijn, Field G. Van Zee, *Accumulating {H}ouseholder transformations, revisited*, ACM Transactions on Mathematical Software, Vol. 32, No 2, 2006.
  - [18] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh, *Basic Linear Algebra Subprograms for Fortran Usage*, ACM Transactions on Mathematical Software, Vol. 5, No. 3, pp. 308-323, Sept. 1979.
  - [19] Per-Gunnar Martinsson, Gregorio Quintana-Orti, Nathan Heavner, Robert van de Geijn, *Householder QR Factorization With Randomization for Column Pivoting (HQRRP)*, SIAM Journal on Scientific Computing, Vol. 39, Issue 2, 2017.
  - [20] Margaret E. Myers, Pierce M. van de Geijn, and Robert A. van de Geijn, *Linear Algebra: Foundations to Frontiers - Notes to LAFF With*, self-published at [ulaff.net](http://ulaff.net), 2014.
  - [21] Margaret E. Myers and Robert A. van de Geijn, *Linear Algebra: Foundations to Frontiers*, a Massive Open Online Course offered on edX.
  - [22] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.. Nelson, M. Stephens, C.D. Bustamante, *Genes mirror geography within Europe*, Nature, 2008
  - [23] C. Puglisi, *Modification of the Householder method based on the compact WY representation*, SIAM Journal on Scientific Computing, Vol. 13, 1992.
  - [24] Gregorio Quintana-Orti, Xioabai Sun, and Christof H. Bischof, *A BLAS-3 version of the QR factorization with column pivoting*, SIAM Journal on Scientific Computing, 19, 1998.
  - [25] Robert Schreiber and Charles Van Loan, *A Storage-Efficient WY Representation for Products of Householder Transformations*, SIAM Journal on Scientific and Statistical

- Computing, Vol. 10, No. 1, 1989.
- [26] G.W. Stewart, *Matrix Algorithms, Volume I: Basic Decompositions*, SIAM Press, 2001.
  - [27] Robert van de Geijn and Kazushige Goto, *BLAS (Basic Linear Algebra Subprograms)*, Encyclopedia of Parallel Computing, Part 2, pp. 157-164, 2011. If you don't have access, you may want to read an [advanced draft](#).
  - [28] Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, *Restructuring the Tridiagonal and Bidiagonal QR Algorithms for Performance*, ACM Transactions on Mathematical Software (TOMS), Vol. 40, No. 3, 2014. Available free from <http://www.cs.utexas.edu/~flame/web/FLAMEPublications.html> Journal Publication #33. Click on the title of the paper.
  - [29] Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, G. Joseph Elizondo, *Families of Algorithms for Reducing a Matrix to Condensed Form*. ACM Transactions on Mathematical Software (TOMS) , Vol, No. 1, 2012. Available free from <http://www.cs.utexas.edu/~flame/web/FLAMEPublications.html> Journal Publication #26. Click on the title of the paper.
  - [30] H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM Journal on Scientific and Statistical Computing, Vol. 9, No. 1, 1988.

# Index

- (Euclidean) length, 84
- $I$ , 43
- $\infty$ -norm (vector), 84
- $\infty$ -norm, vector, 30
- $\kappa(A)$ , 76, 87
- $\overline{A}$ , 50
- $\bar{x}$ , 94
- $-$ , 19
- $|\cdot|$ , 18
- $e_j$ , 99
- $p$ -norm (vector), 84
- $p$ -norm, matrix, 55
- $p$ -norm, vector, 31
- 1-norm (vector), 84
- 1-norm, vector, 29
- 2-norm (vector), 84
- 2-norm, matrix, 56
- 2-norm, vector, 25
- absolute value, 18, 83
- catastrophic cancellation, 153
- Cauchy-Schwartz inequality, 26
- complex conjugate, 19
- complex product, 83
- condition number, 76, 87
- conjugate, 19, 83
- conjugate (of matrix), 86
- conjugate (of vector), 83
- conjugate of a matrix, 50
- conjugate transpose (of matrix), 86
- conjugate transpose (of vector), 83
- consistent matrix norm, 71, 87
- cost of basic linear algebra operations, 153
- direction of maximal magnification, 76
- distance, 18
- dot product, 83, 94
- equivalence style proof, 22
- Euclidean distance, 18
- FLAME notation, 117
- Frobenius norm, 47, 86
- Hermitian, 50
- Hermitian transpose, 26, 49
- Hermitian transpose (of matrix), 86
- Hermitian transpose (of vector), 83
- homogeneity (of absolute value), 19
- homogeneity (of matrix norm), 46, 86
- homogeneity (of vector norm), 24, 83
- identity matrix, 43
- induced matrix norm, 51, 52
- infinity norm, 30
- inner product, 83, 94
- left pseudo-inverse, 89
- left singular vector, 117, 148
- linear transformation, 41
- magnitude, 18
- matrix, 41, 43
- matrix 1-norm, 87
- matrix 2-norm, 56, 87

- matrix  $\infty$ -norm, 87  
matrix  $p$ -norm, 55  
matrix norm, 46, 86  
matrix norm, 2-norm, 56  
matrix norm,  $p$ -norm, 55  
matrix norm, consistent, 71, 87  
matrix norm, Frobenius, 47  
matrix norm, induced, 51, 52  
matrix norm, submultiplicative, 70, 71, 87  
matrix norm, subordinate, 71, 87  
matrix  $p$ -norm, 87  
matrix-vector multiplication, 43  
  
norm, 12  
norm, Frobenius, 47  
norm, infinity, 30  
norm, matrix, 46, 86  
norm, vector, 24, 83  
  
orthogonal matrix, 101  
orthogonal projection, 89  
orthogonal vectors, 95  
orthonormal matrix, 99  
orthonormal vectors, 98  
  
positive definiteness (of absolute value), 19  
positive definiteness (of matrix norm), 46, 86  
positive definiteness (of vector norm), 24, 83  
pseudo-inverse, 89  
residual, 14  
  
right singular vector, 117, 148  
rotation, 106  
  
Singular Value Decomposition, 88, 91  
singular vector, 117, 148  
standard basis vector, 42, 84  
submultiplicative matrix norm, 70, 71, 87  
subordinate matrix norm, 71, 87  
SVD, 88, 91  
  
transpose, 49  
transpose (of matrix), 86  
transpose (of vector), 83  
triangle inequality (for absolute value)), 19  
triangle inequality (for matrix norms)), 46, 86  
triangle inequality (for vector norms)), 24, 83  
  
unit ball, 32, 84  
unitary matrix, 101, 147  
  
vector 1-norm, 29, 84  
vector 2-norm, 25, 84  
vector  $\infty$ -norm, 30, 84  
vector  $p$ -norm, 84  
vector  $p$ -norm, 31  
vector norm, 24, 83  
vector norm, 1-norm, 29  
vector norm, 2-norm, 25  
vector norm,  $\infty$ -norm, 30  
vector norm,  $p$ -norm, 31

## **Colophon**

This book was authored in MathBook XML.