Numerical Methods for Data Scientists

Final Review

May 6, 2024

Contents

1 Introduction

"Unfortunately, no one can be told what the Matrix is. You have to see it for yourself."
– Morpheus

These notes may have some errors. If there are any discrepancies between these notes and the lecture notes, assume the lecture notes are correct

2 Vector Spaces

Why do we care about linear algebra? It is the only thing that your computer hardware can do.

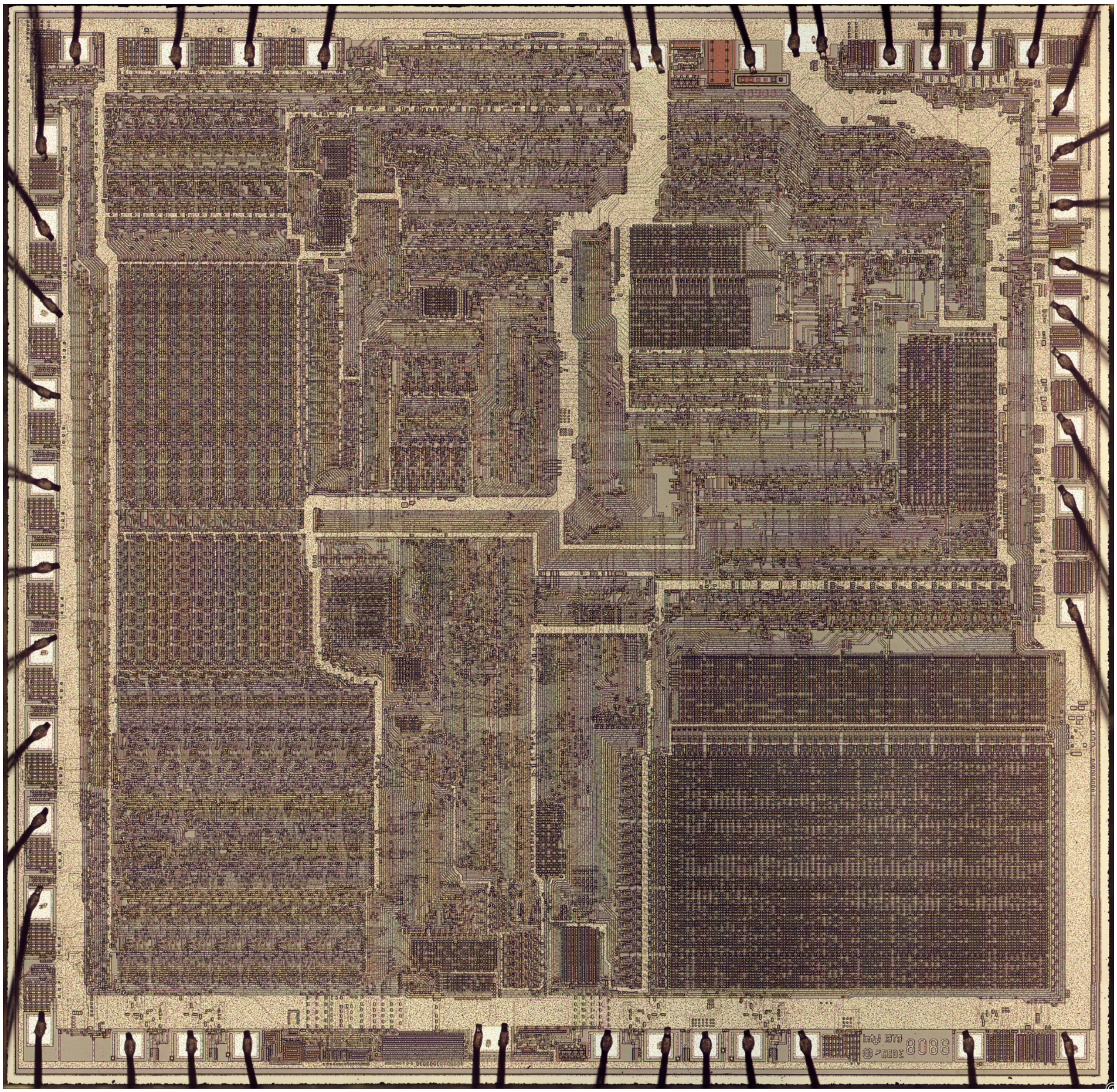

Figure 1: Intel's 8086 16-bit processor from 1979. 29,000 transistors (photo from Ken Shirriff's blog). Compare with 2.6 trillion transistors on the deep learning processor Wafer Scale Engine 2 by Cerebras.

2.1 Linearity

- **Linear Combination**: if $\mathbf{v}$ and $\mathbf{w}$ are vectors, and a and b are scalars, then $a\mathbf{v} + b\mathbf{w}$ is a linear combination of these two vectors. You MUST understand this concept to continue.
  - A non-linear combination of vectors $[x_1, x_2, x_3]$ and $[y_1, y_2, y_3]$ might look like this $[x_1^2 y_1, x_2 y_2^3, x_3 y_3]$. We don't care about this for this class.
- **Linearly independence**: a set of vectors is linearly independent if no linear combination of any subset of the vectors can produce another vector in the set.
  - example: we have a set of 3 vectors, $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$. If $a\mathbf{x} + b\mathbf{y} = \mathbf{z}$ for any a and b, then this set is not linearly independent.
- **Linearly dependent** - opposite of linearly independent.

2.2 Spaces

- **Vector Space**

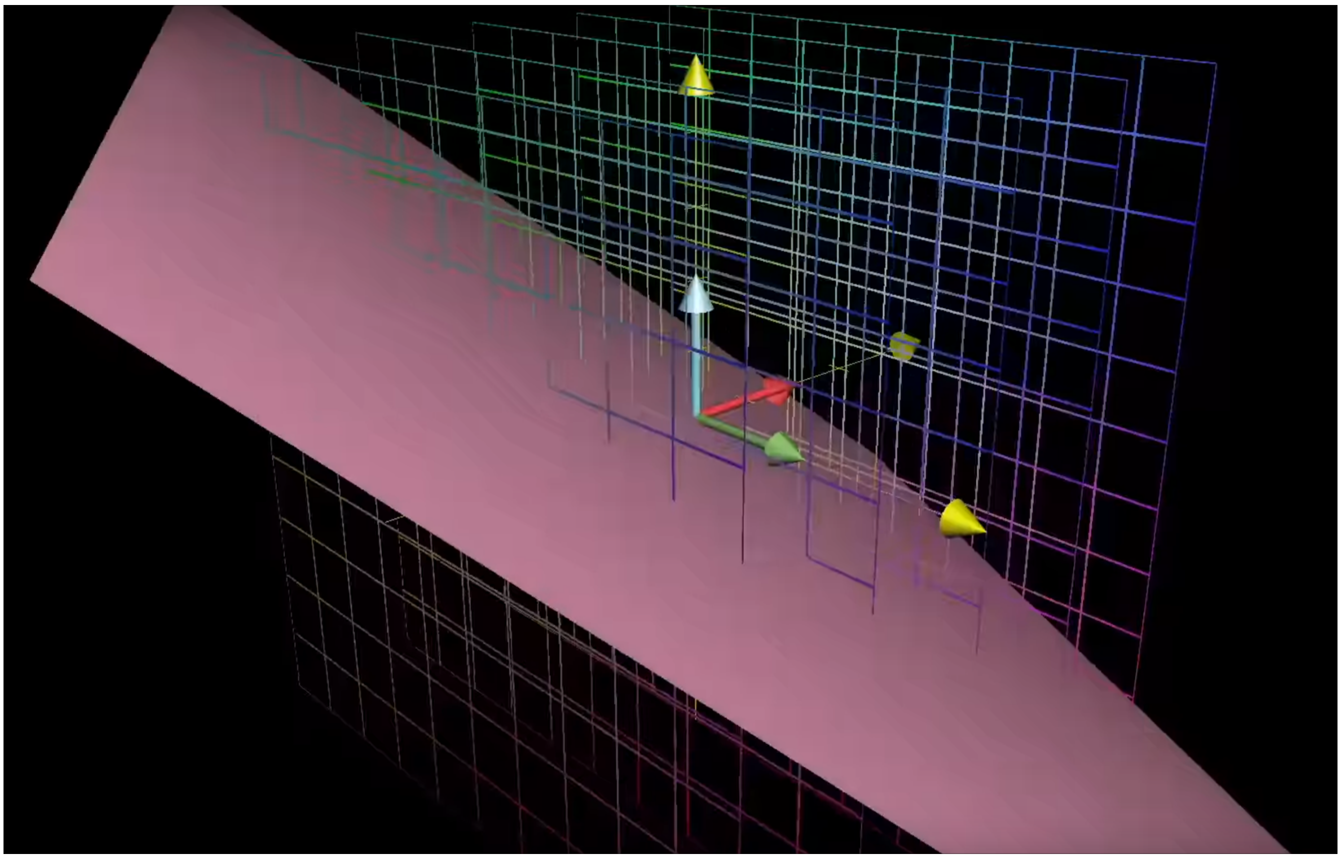

Figure 2: A planar 2D vector space embedded in a $\mathbb{R}^3$. The plane is a subspace of $\mathbb{R}^3$. Visualization from the 3Blue1Brown youtube channel by Grant Sanderson.

  - A set, whose elements are vectors, where every vector in the set must satisfy the following conditions:
    1. Commutativity of vector addition: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
    2. Associativity of vector addition: $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$
    3. Existence of a zero vector: $\mathbf{0} + \mathbf{x} = \mathbf{x} + \mathbf{0}$
    4. Existence of additive inverse for each vector: $\mathbf{x} + (-\mathbf{x}) = \mathbf{0} = (-\mathbf{x}) + \mathbf{x}$
    5. Distribution of scalars with respect to vector addition: $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$

6. Distribution with respect to scalar addition: $(a+b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$
7. Scalar associativity: $a(b\mathbf{v}) = (ab)\mathbf{v}$
8. Identity element of scalar multiplication: $1\mathbf{v} = \mathbf{v}$

  - This is the rigorous mathematical definition of a vector space.
  - The important thing to note is that with respect to all of these operations: **vectors inside the vector space, remain inside the vector space under each of these operations**.
  - Example: $\mathbb{R}^n$ - n dimensional space of real numbers.

- **Subspace** - a subset of a vector space that is also a vector space.
  - Example: $\mathbb{R}^2$ is a subspace of $\mathbb{R}^3$
- **Span** - the span of a set of vectors is the set of all linear combinations of the vectors in S.
  - I like to think of this as the space that can be reached by any linear combination of the vectors in the set.
  - $\mathbb{R}^3$ has $\{[1,0,0]\ [0,1,0]\ [0,0,1]\}$ as a spanning set.
  - Spans for set of matrices or spans of polynomials spaces also exist.
- **Basis**: a set of vectors **S** forms a basis for a vector space **V**, if
  1. **S** spans **V**
  2. The vectors in **S** are linearly independent
     - This means that **S** is a minimal set of vectors that spans **V**
- **Orthogonality**:
  - two vectors are orthogonal if their dot product is zero
  - two vector spaces are orthogonal if every vector in one space is orthogonal to every vector in the other space.

2.3 Exercises

1. Consider the following set of vectors **S**

$$\left\{ \begin{bmatrix} 1 \\ -2 \\ 1 \\ 4 \end{bmatrix}, \begin{bmatrix} -6 \\ -8 \\ -4 \\ 6 \end{bmatrix}, \begin{bmatrix} 8 \\ 4 \\ 6 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ -8 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -6 \\ -8 \\ -4 \\ 6 \end{bmatrix} \right\}$$

(a) Find the row reduced echelon form (RREF):

$$\begin{bmatrix} 1 & 0 & 2 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b) How many pivots are there?

(c) Write a set of basis vectors for the set **S**

(d) Is the dimension of the space spanned by the basis set the same as the dimension of the space spanned by **S**?

2. (T / F) Only an orthogonal set of vectors can form a basis for a vector space.

3. The space of all 3×4 matrices has what dimension?
4. All polynomials or order 5 or less occupy a space of what dimension?

3 Functions

3.1 Definitions

Everyone knows what a function is, lets formalize the definition.

- A function is a mapping from a set **X** to a **Y** that assigns one element of **Y** to each element of **X**.
- Example: $f : \mathbb{R} \to \mathbb{R} | \{f(x) = x + 5\}$
  - This is read: A function that maps the real numbers to the real numbers, such that the function applied to x equals $x + 5$
- All the parts of a function can be understood in the following diagram

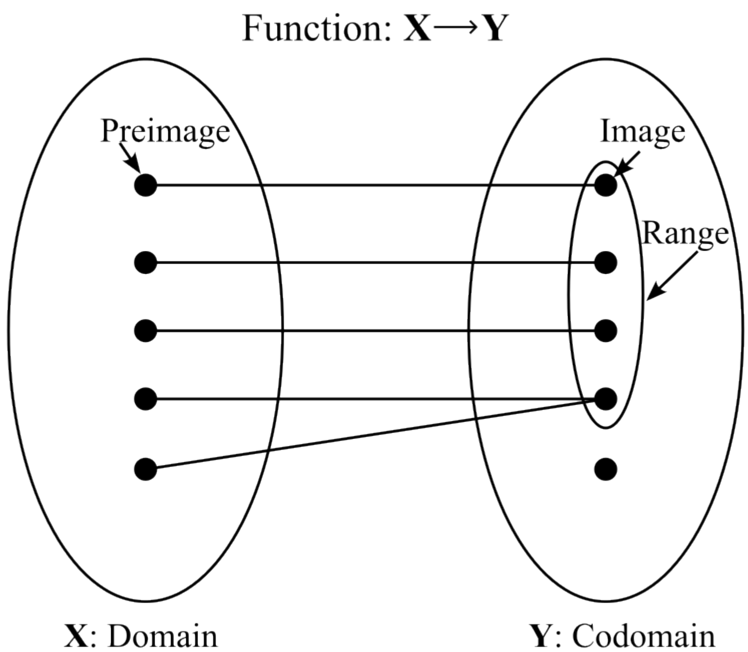

Figure 3: Diagram of a function. This function is neither one-to-one nor onto.

- **Domain**: set of possible input values that a function can take.
- **Codomain**: set of all possible output values that a function can theoretically produce.
- **Range**: The set of all actual output values that the function takes for all possible inputs in the domain.
  - Why is the codomain different from the range? The codomain is a more general statement about the "type" of output this function creates. It is not always necessary to know the exact range, as this can be a very specific subset of the codomain.
- **Onto**: A function is onto if every element in the codomain has a pre-image. Said differently, every element in the codomain has at least one element in the domain.
- **One-to-one**: every element in the domain has a unique element in the codomain.
- **Pre-image**: the pre-image of y is the set of elements in the domain that map to y.
- **Image**: the image of x is the element in the codomain that the function maps to from the pre-image.
- A function **cannot** map one element in X to two elements in Y

- **Function Composition**: sequential application of functions. The output of the first function becomes the input of the second function, and so on.

$$g(f(x)) = (g \circ f)(x)$$

- **Function Inverse**: two functions, $f : \mathbf{X} \to \mathbf{Y}$ and $g : \mathbf{Y} \to \mathbf{X}$ are inverses if and only if:
  1. $g(f(x)) = x$
  2. $f(g(y)) = y$

3.2 Exercises

1. Study the diagram at the start of this section.
2. $f : \mathbf{X} \to \mathbf{Y}$, If $f(x) = y$, x is called the what of y?
3. $f : \mathbf{X} \to \mathbf{Y}$, $\mathbf{X}$ is called what?
4. (T/F) All elements in $\mathbf{Y}$ must have a pre-image in $\mathbf{X}$
5. More in HW 21.
6. Are the following functions inverses of each other?

$$f(x) = 4x - 2$$

$$g(y) = \frac{y+2}{4}$$

7. Are the following functions inverses of each other?

$$f(x) = \frac{2}{x+3}$$

$$g(y) = \frac{-3y-2}{y}$$

4 Linear Transformations

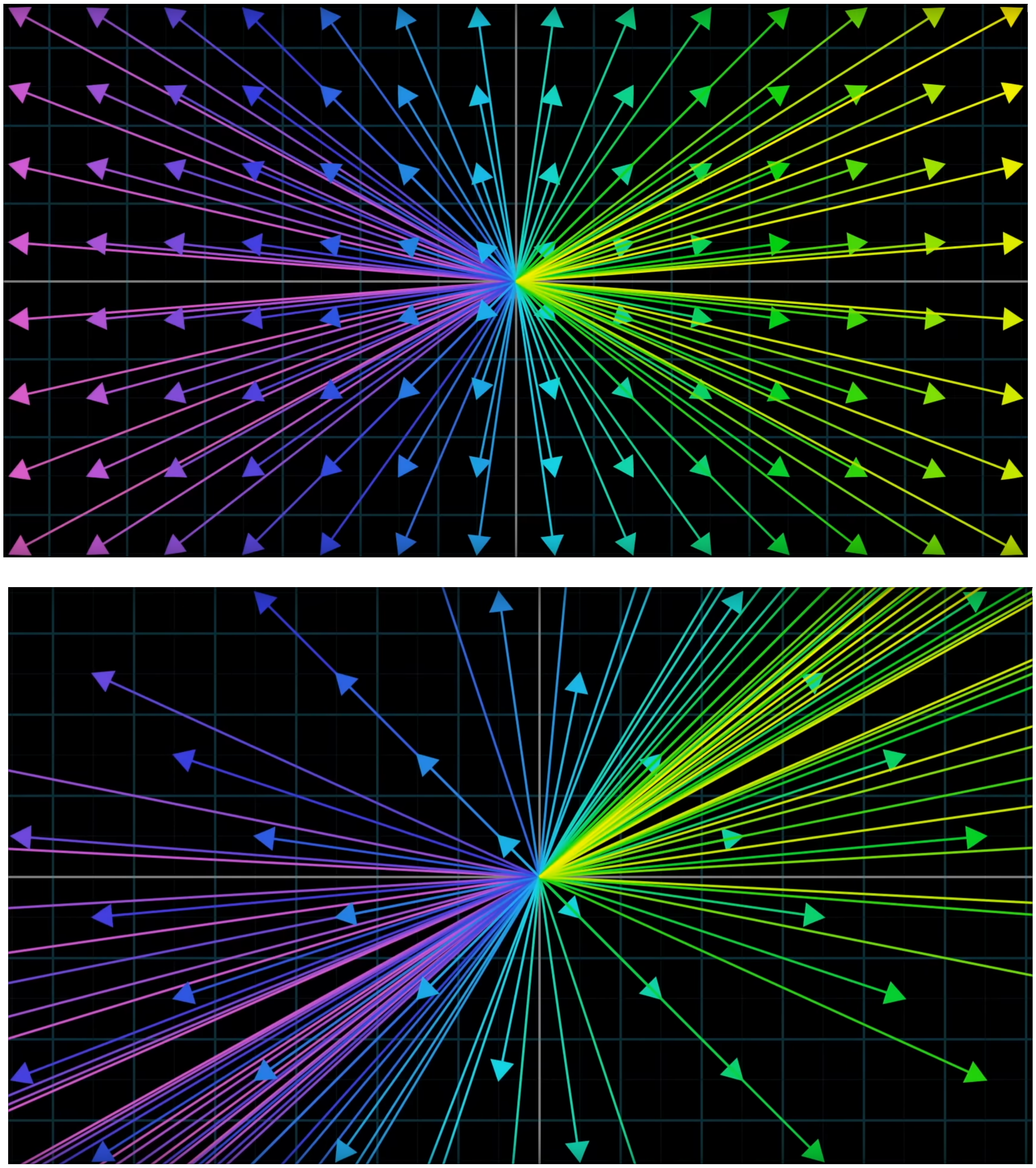

Figure 4: Vectors before a linear transformation (top), and vectors after a linear transformation (bottom). Visualization from the 3blue1brown youtube channel by Grant Sanderson.

4.1 Definitions

- A linear transformation is a mapping between two vector spaces that preserves the operations of vector addition and scalar multiplication.
- We can think of this as a function, we'll call that function L.
- A linear transformation, L, must obey the following two properties (know them):
  1. Closed under addition: $L(\mathbf{u}+\mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v})$
  2. Closed under scalar multiplication: $L(a\mathbf{u}) = aL(\mathbf{u})$
- **A matrix is a linear transformation!**
  - Think of $A\mathbf{x} = \mathbf{b}$. Does it map one vector to a different vector? Yes! It maps from the space that $\mathbf{x}$ lives in to the column space of A.

– Does it preserve vector addition and scalar multiplication? Yes! $A(\mathbf{x}+\mathbf{y}) = A\mathbf{x}+A\mathbf{y}$ and $A(a\mathbf{x}) = aA\mathbf{x}$
– So a matrix is a function? YES!
– And a function is a vector, and a matrix is a function, and derivative is a function so a derivative could be a matrix, and everything is everything, everywhere all at once? Yes!

4.2 Rank and nullity of a Linear Transformation

- Matrices have a rank and nullity, matrices are linear transformations, so linear transformations in general have a rank and a nullity.
- Let $L : \mathbf{V} \to \mathbf{W}$
  – **Nullspace** (kernel): all the vectors in $\mathbf{V}$ that map to $\mathbf{0}$
  – **Nullity** - dimension of the nullspace
  – **Rank** - dimension of the range of L

4.3 Exercises

1. Let $L : \mathbb{R}^2 \to \mathbf{V}$ be a linear transformation where $\mathbf{V}$ is a subspace of $\mathbb{R}^3$. Let $\{\mathbf{b_1}, \mathbf{b_2}\}$ form a basis for $\mathbb{R}^2$ such that

$$L(\mathbf{b_1}) = \begin{bmatrix} 3 \\ -1 \\ 0 \end{bmatrix} \quad \text{and} \quad L(\mathbf{b_2}) = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Let a vector, u, in $\mathbb{R}^2$ be given by

$$u = k_1\mathbf{b_1} + k_2\mathbf{b_2}$$

such that

$$L(\mathbf{u}) = \begin{bmatrix} 6 \\ -3 \\ -2 \end{bmatrix}$$

Find k_1 and k_2

2. Consider $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}^2$ where

$$f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x+y \\ 0 \end{bmatrix} \quad \text{and} \quad g\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} 4x+y \\ xy \end{bmatrix}$$

(a) Are f and g linear transformations?
- Are they closed under addition?
- Are they closed under scalar multiplication?

(b) Is $f \circ g$ a linear transformation?
- Is it closed under addition?
- Is it closed under scalar multiplication?

5 Matrix Subspaces

5.1 The Four fundamental subspaces

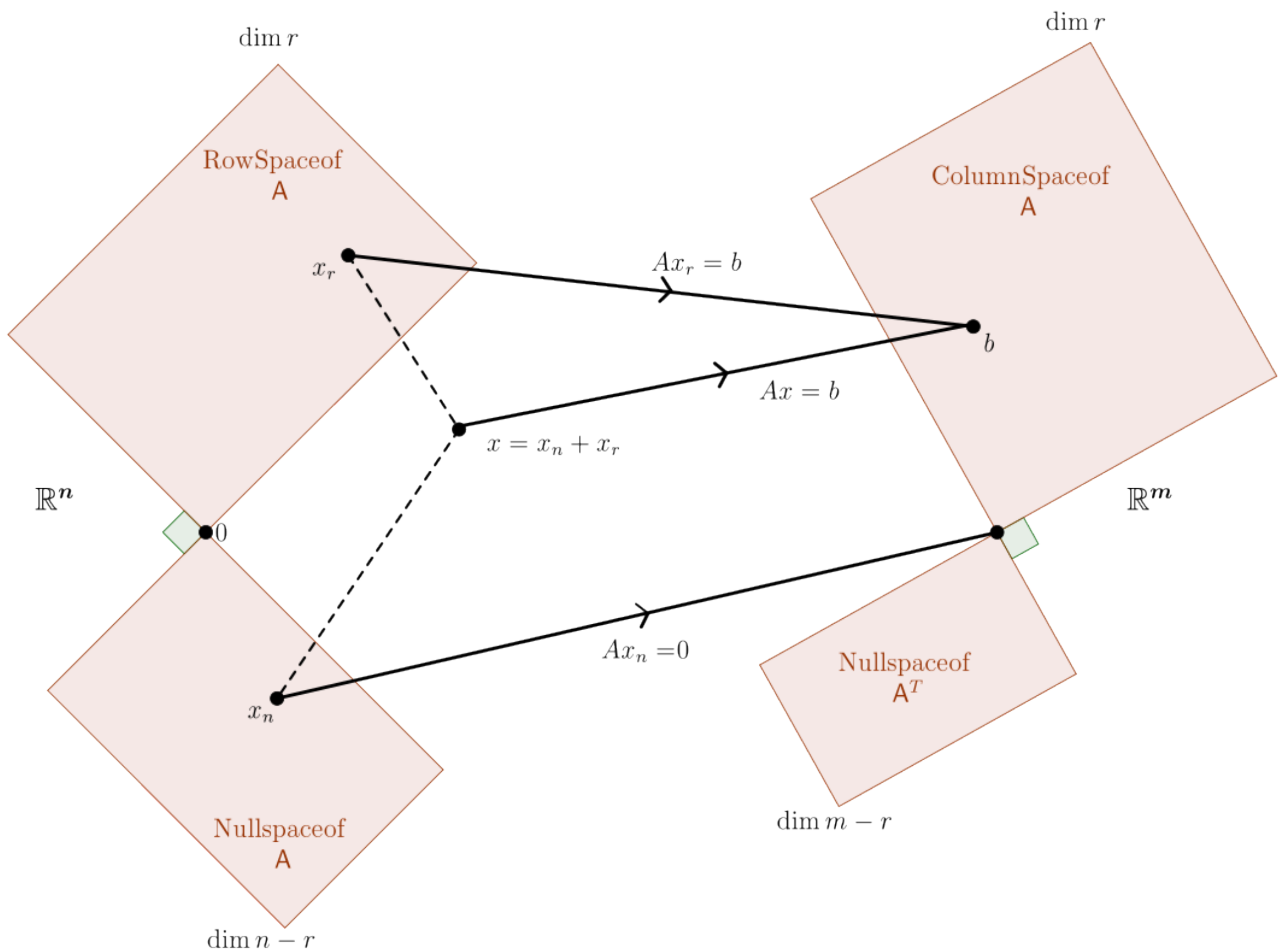

Figure 5: Diagram of the four fundamental subspaces of a matrix. Image from Wolfram MathWorld.

For each concept in this section, I'll explain the idea in three equivalent ways.

5.1.1 Column Space $C(A)$

- The space spanned by the column vectors of the matrix.
- The set of all $\mathbf{b}$ in $\mathbb{R}^m$ such that $A\mathbf{x} = \mathbf{b}$ where $\mathbf{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$
- $\{\mathbf{b} \in \mathbb{R}^m \mid \forall \mathbf{x} \in \mathbb{R}^n,\ A\mathbf{x} = \mathbf{b}\}$

5.1.2 Row Space $C(A^\intercal)$

- The space spanned by the row vectors of the matrx.
- The set of all $\mathbf{b}$ in $\mathbb{R}^n$ such that $A^\intercal\mathbf{x} = \mathbf{b}$ where $\mathbf{x} \in \mathbb{R}^m$ and $A^\intercal \in \mathbb{R}^{n \times m}$
- $\{\mathbf{b} \in \mathbb{R}^n \mid \forall \mathbf{x} \in \mathbb{R}^m,\ A^\intercal\mathbf{x} = \mathbf{b}\}$

5.1.3 Nullspace $N(A)$

- The space spanned by the vectors that are mapped to $\mathbf{0}$ by A.
- The set of all $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$ where $\mathbf{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$

- $\{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = 0\}$

5.1.4 Left Nullspace $N(A^\intercal)$

- The space spanned by the vectors that are mapped to $\mathbf{0}$ by $A^\intercal$.
- The set of all $\mathbf{x}$ such that $A^T\mathbf{x} = \mathbf{0}$ where $\mathbf{x} \in \mathbb{R}^m$ and $A^\intercal \in \mathbb{R}^{n \times m}$
- $\{\mathbf{x} \in \mathbb{R}^m \ \mid \ A^\intercal\mathbf{x} = \mathbf{0}\}$

5.2 Fundamental Theorem of Matrix Subspaces

- **The row space is orthogonal to the nullspace**
  - The space mapped by taking a linear combination of the rows of A is orthogonal to the space that a linear combination of the columns of A maps to $\mathbf{0}$
  - Only the origin is shared by the two spaces. All vector spaces emerge from the origin.
- **The column space is orthogonal to the left nullspace**
  - The space mapped by taking a linear combination of the columns of the matrix, is orthogonal to the space spanned by all the vectors that are mapped
- *Practical example*: Say you have a set of patient medical data. Each patient's data is stored in a column of a matrix. Some of these patients ended up in the hospital. We could compare the column space of different groups (age groups, geographic groups, height and weight groups, etc) with the column space of patients who ended up in the hospital to see how closely related they are. (Remember, matrix norms can give us a measure of similarity).

5.3 Rank-Nullity Theorem

- $\text{rank}(A) + \text{nullity}(A) = n$
- $\text{rank}(A^\intercal) + \text{nullity}(A^\intercal) = m$

5.4 Fundamental Theorem of Linear Algebra

Let $A \in \mathbb{R}^{n \times n}$ (We could also say let $A \in \mathbb{M}_{nn}$). All of the following are equivalent:

1. A is invertible.
2. A^{-1} exists.
3. The columns of A are linearly independent.
4. The rows of A are linearly independent.
5. $\det(A) \neq 0$
6. $A\mathbf{x} = \mathbf{0}$ only when $\mathbf{x} = \mathbf{0}$
7. $\text{rref}(A)$ has n-pivots.
8. $\text{rref}(A) = I$
9. A is full rank.
10. $\text{rank}(A) = n$
11. $C(A)$ spans all of $\mathbb{R}^n$
12. $C(A^T)$ spans all of $\mathbb{R}^n$

5.5 Exercises

1. Consider $A^\intercal \mathbf{x} = \mathbf{b}$. What matrix subspace does $\mathbf{b}$ exist in?
2. Consider $A\mathbf{x} = \mathbf{b}$. What matrix subspace does $\mathbf{b}$ exist in?
3. Consider $A^\intercal \mathbf{x} = \mathbf{0}$. What matrix subspace does $\mathbf{x}$ exist in?
4. Consider $A\mathbf{x} = \mathbf{0}$. What matrix subspace does $\mathbf{x}$ exist in?
5. Consider the following matrix:
$$\begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$
   (a) What is the rank of the matrix?

   (b) What is the nullity of the matrix?
6. (T/F) Let $A \in \mathbb{M}_{nn}$. If $\det(A) \neq 0$, then $\text{rank}(A) = n$
7. (T/F) Let $A \in \mathbb{M}_{nn}$. If $\text{rref}(A) = I$, then the columns of A are linearly independent.
8. (T/F) Let $A \in \mathbb{M}_{nn}$. If A has n pivots, then A^{-1} exists.
9. Consider the matrix:
$$A = \begin{bmatrix} 1 & 2 & 2 & -2 & -4 \\ -2 & 2 & 2 & 0 & -2 \\ -2 & 2 & 3 & 1 & -1 \\ 2 & 1 & 1 & -2 & -3 \\ 1 & -1 & -1 & 0 & 1 \end{bmatrix}$$
   The $\text{rref}(A)$ is given by:
$$\text{rref}(A) = \begin{bmatrix} 1 & 0 & 0 & -2/3 & -2/3 \\ 0 & 1 & 0 & -4/3 & -8/3 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
   (a) Find a basis set for the column space of A.

   (b) What is the nullity of A?

   (c) Find the vectors that for the basis for the nullspace of A.

   - Hint: Find the free variables and assign simple values to them, then solve for the fixed variables.

6 Projections

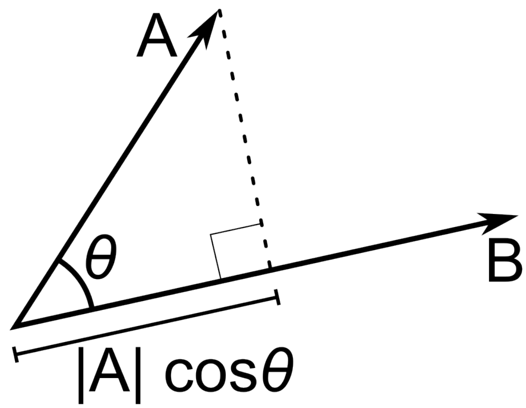

Figure 6: A is projected on B, producing a vector with length $|A|\cos(\theta)$ and in the same direction as B. Image from Wikimedia Commons.

6.1 Vector Projection

The projection of $\mathbf{a}$ onto the vector $\mathbf{b}$ is given by:

$$\text{proj}_b(\mathbf{a}) = \frac{\mathbf{a}^\intercal\mathbf{b}}{\mathbf{b}^\intercal\mathbf{b}}\mathbf{b} = \frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}}\mathbf{b}$$

6.2 Projection Matrix

Now lets find a projection matrix that can project any matrix onto $\mathbf{a}$. This means that the projection, $\mathbf{p}$, of $\mathbf{b}$ onto $\mathbf{a}$ will be given by $A\mathbf{b} = \mathbf{p}$.

$$A = \frac{\mathbf{a}\mathbf{a}^\intercal}{\mathbf{a}^\intercal\mathbf{a}}$$

Recall the difference between inner product (dot product) and outer product:

- Dot product ($\mathbf{a}^\intercal\mathbf{a}$) produces a **scalar**
- Outer product ($\mathbf{a}\mathbf{a}^\intercal$) produces a **matrix**

6.3 Exercises

1. Find the projection of $\begin{bmatrix}1\\2\\2\end{bmatrix}$ onto the vector: $\begin{bmatrix}2\\0\\1\end{bmatrix}$

2. Find the projection matrix for $\begin{bmatrix}4\\2\end{bmatrix}$

7 Eigensystems

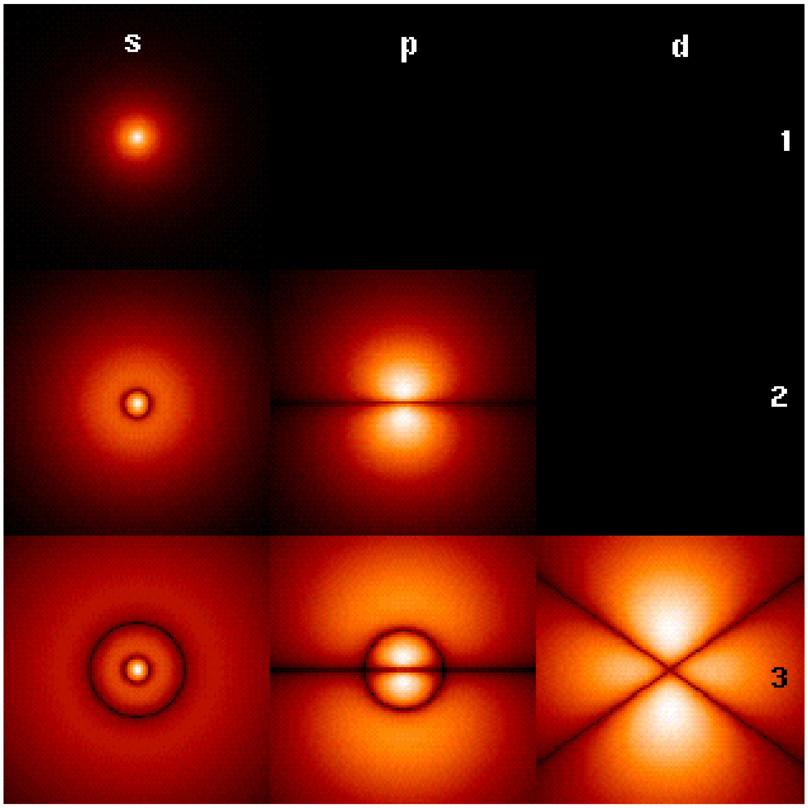

Figure 7: The eigenvalues of Schrödinger equation ($H\Psi_E = E\Psi_E$) gives us the wavefunctions of the orbitals of a hydrogen atom, where the eigenvalues are interpreted as the energy. Eigenvalues and Eigenvectors tell us about the fundamental building blocks of the given matrix. Image from Wikimedia Commons.

7.1 What are eigenvalues and eigenvectors?

- Eigenvalues and eigenvectors are special characteristics of **square matrices**.
- The eigenvalues (λ) and eigenvectors ($\mathbf{v}$) equation is give by:

$$A\mathbf{v} = \lambda\mathbf{v}$$

- Looking at this equation we see that this means that when we apply the linear transformation A (remember, matrices are linear transformations), the eigenvalues are only changed by a scale factor. They do not change dimensions (because A is square) and they do not rotate. Can you find the eigenvectors in figure 4?
- Eigenvectors are much more then just vectors that only scale when A is applied to them. Eigenvalues and eigenvectors reveal the fundamental building blocks of a matrix.

7.2 How to Find Eigenvalues

$$\begin{aligned} A\mathbf{v} &= \lambda\mathbf{v} \\ A\mathbf{v} - \lambda\mathbf{v} &= \mathbf{0} \\ (A - \lambda I)\mathbf{v} &= \mathbf{0} \end{aligned}$$

For $\mathbf{v}$ to exist and not be $\mathbf{0}$, it must be in the nullspace of $A - \lambda I$. If there exists a nullspace of $A - \lambda I$, then the matrix is not full rank, and the inverse, $(A - \lambda I)^{-1}$ doesn't exists. Therefore, the determinant of $(A - \lambda I)$ must be equal to 0.

7.2.1 characteristic equation

Therefore, by solving the following equation, we can solve for the eigenvalues.

$$\det(A - \lambda I) = 0$$

7.3 How to Find Eigenvectors

- Since we know how to find the eigenvalues, to find the eigenvectors we just need to solve for the vectors in the nullspace of $(A - \lambda I)$
- We can do this by using RREF to solve the system of equations that arises when we augment $A - \lambda I$ with $\mathbf{0}$,
- i.e. $\text{rref}([A - \lambda I \vdots \mathbf{0}])$
- example:

$$\lambda_1 = -1 \qquad A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$A - \lambda_1 I = \begin{bmatrix} 1-(-1) & 2 \\ 2 & 1-(-1) \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$$\text{rref}([A - \lambda_1 I \vdots \mathbf{0}]) = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

set $v_2 = 1$ and $v_1 = -1$

$$\mathbf{v} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

7.4 Iterative Methods for find Eigenvalues and Eigenvectors

7.4.1 Rayleigh Quotient

- If **x** is an eigenvector of A:

$$\lambda = \frac{x^\intercal A x}{x^\intercal x}$$

- Prove it! Its easy
- Rayleigh quotient iteration find the **nearest eigenvector (and eigenvalue) to the initial guess**. We coded it up. Review it!

7.4.2 Power Iteration Method

- Another iterative method for finding the **largest eigenvector and eigenvalue**.

7.5 Properties

- The **trace** of a matrix (the sum of the values along its diagonal), is equal to the sum of its eigenvalues.
- The **determinant** of a matrix is equal to the product of its eigenvalues.
- The entries on the diagonal of an **upper triangular matrix** are its eigenvalues.
- **Geometric multiplicity**: The number of eigenvectors for a given eigenvalue.
  - The GM of $\lambda = \text{nullity}(A - \lambda I)$
- **Algebraic multiplicity**: The number of times an eigenvalue occurs.
- **Defective eigenvalue**: An eigenvalue, λ, whose geometric and algebraic multiplicity are **not equal**.

- **Complete eigenvalue**: An eigenvalue, λ, whose geometric and algebraic multiplicity are **not equal**.
- **Spectrum of a matrix**: The set of eigenvalues of a matrix.

7.6 Exercises

1. Find the eigenvalues and eigenvectors of the following matrix:

$$\begin{bmatrix} -5 & 2 \\ -7 & 4 \end{bmatrix}$$

2. Let A be a diagonalizable matrix. A has an eigenvalue $\lambda = 2$ with corresponding eigenvector

$$\mathbf{v} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

   And another eigenvalue of $\lambda = 4$ with corresponding eigenvector

$$\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

   Find the matrix A

3. A matrix has eigenvalues of 3, -2, 1, and 4. What is the determinant of the matrix?

4. Consider the following matrix:

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix}$$

   (a) What are the eigenvalues of A?

   (b) What are their geometric and algebraic multiplicity?

8 Matrix Decompositions

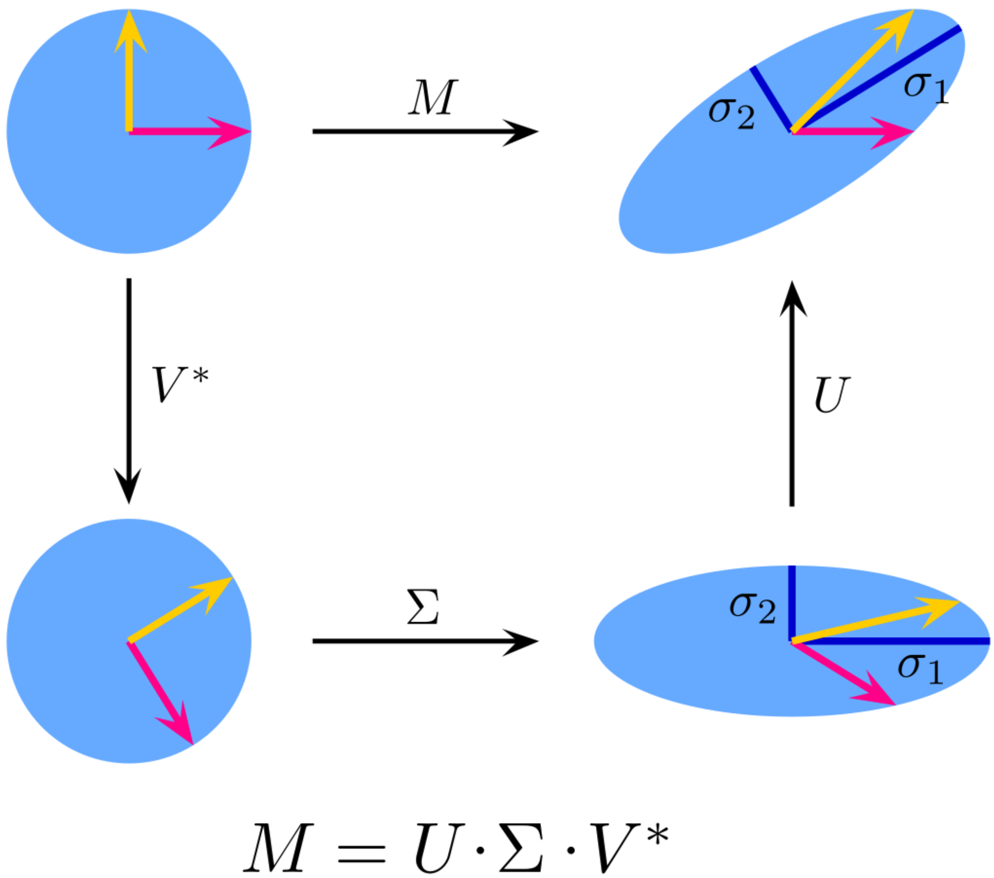

Figure 8: SVD tells us that every matrix can be decomposed into a rotaiton (V^T), a scaling (Σ), and another rotation (U). Image from Wikimedia Commons.

Why do we care? I told you at the beginning that we want to learn as much linear algebra as we can so that we can extract useful information from our systems, and build useful algorithms we can run on a computer. Matrix decompositions are fundamental techniques in linear algebra that help us with many tasks:

- Dimensionality reduction - low rank approximation (SVD), PCA.
- Easier to analyze/ vizualize/ understand large datasets
- Feature extraction
- Solving linear systems quickly
- Numerical Stability
- Optimization
- More...

8.1 LU Decomposition $A = LU$

- Very useful to solve systems of linear equations, $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ when A is fixed and you have a lot of different $\mathbf{b}$ vectors.
- Works on **square matrices.**

- L = lower triangular matrix with 1's on the diagonal.
- U = upper triangular matrix with anything on the diagonal.

$$
\begin{aligned}
A\mathbf{x} &= \mathbf{b} \\
A &= LU \\
LU\mathbf{x} &= \mathbf{b} \\
\mathbf{x} &= U^{-1}L^{-1}\mathbf{b} \\
&\text{we can also write this as} \\
L\mathbf{y} &= \mathbf{b} \\
U\mathbf{x} &= \mathbf{y}
\end{aligned}
$$

- Those last two equations can be solved with **forward substitution** and **backwards substitution**. This is what makes LU decomposition faster then finding the inverse of A.
- Finding A^{-1} is $O(n^3)$
- Finding LU is $O(n^3)$
- Solving $L\mathbf{y} = \mathbf{b}$ and $U\mathbf{x} = \mathbf{y}$ with forward and backwards substitution is $O(n^2)$

8.2 Matrix Diagonalization and Similar Matrices

8.2.1 Matrix Diagonalization $A = P\Lambda P^{-1}$

- A **square** matrix A is diagonalizable if, for some P and Λ, we get $A = P\Lambda P^{-1}$
- P is an invertible matrix.
- Λ is a diagonal matrix (only non zero on the main diagonal).

8.2.2 Similar Matrices $A = PBP^{-1}$

- Two **square** matrices, A and B, are similar if there exists an invertible matrix P such that $A = PBP^{-1}$
- Similar matrices have the same:
  - rank
  - eigenvalues with their algebraic and geometric multiplicity
  - trace
  - determinant

8.3 Shur Decomposition $A = QTQ^\intercal$

- Works for **all square** matrices.
- Q is a unitary matrix (also called an orthogonal , $Q^\intercal Q = QQ^\intercal$ and $Q^\intercal = Q^{-1}$, when Q is real).
- T is an upper triangular matrix.
- A and T are similar matrices.

8.4 Unitary Decomposition $A = Q\Lambda Q^\intercal$

- Works for **square, normal** matrices
- Q is a unitary matrix (also called an orthogonal , $Q^\intercal Q = QQ^\intercal$ and $Q^\intercal = Q^{-1}$, when Q is real).

- Λ is a diagonal matrix.

8.5 QR Decomposition $A = QR$

- Works for **all real, square** matrices (there is a version for rectangular matrices as well).
- Q is a orthonormal matrix.
- R is an upper triangular matrix (with strictly positive diagonal values).
- You can use QR to solve to solve a **linear system**.

$$\begin{aligned}
&A\mathbf{x} = \mathbf{b} \\
&A = QR \\
&QR\mathbf{x} = \mathbf{b} \\
&\mathbf{x} = R^{-1}Q^{\intercal}\mathbf{b}
\end{aligned}$$

- We can use it to solve the **normal equations**:

$$\begin{aligned}
&A^{\intercal}A\mathbf{x} = A^{\intercal}\mathbf{b} \\
&A = QR \\
&A^{\intercal} = R^{\intercal}Q^{\intercal} \\
&\text{plug the above two equations into the first equation} \\
&R^{\intercal}(Q^{\intercal}Q)R\mathbf{x} = R^{\intercal}Q^{\intercal}\mathbf{b} \\
&R^{\intercal}R\mathbf{x} = R^{\intercal}Q^{\intercal}\mathbf{b} \\
&(R^{\intercal})^{-1}R^{\intercal}R\mathbf{x} = (R^{\intercal})^{-1}R^{\intercal}Q^{\intercal}\mathbf{b} \\
&R\mathbf{x} = Q^{\intercal}\mathbf{b} \\
&\mathbf{x} = R^{-1}Q^{\intercal}\mathbf{b}
\end{aligned}$$

- Notice: **Solving a linear system using QR is the same as solving the normal equations**.
- How do we get the Q matrix? Lets talk about three common ways:

8.5.1 Classical Gram Schmidt

- **Least stable.**
- The Gram Schmidt process is used to create an orthonormal basis from a set of vectors.
- By creating an orthonormal set of basis vectors from the column space of a matrix, we have effectively created the orthonormal Q matrix for a QR decomposition.
- Finding R is trivial after we have Q.
- Algorithm:
  - Normalize the first column of the matrix A. Store it as the first column of Q.
  - From the second column of A, subtract away the the projection of the second column of A onto the first column of Q (this eliminates the portion of the two vectors that are in the same direction, thus making them orthogonal).
  - Normalize the result and store it as the second column of Q.
  - From the third column of A, subtract away the projection of the third column of A onto the first column of Q, and the projection of the third column of A onto the second column of Q. (This eliminates the portion of the first two Q columns that were in the same direction as the third column of A, thus making all three vectors orthogonal).

- Normalize the result and store it as the third column of Q.
  - Continue loop through all columns.

- The problem is that this accumulates error in every iteration, since it uses previously modified columns (with floating point, and rounding errors) to create new orthogonal vectors.

8.5.2 Modified Gram Schmidt

- **More stable than classical Gram Schmidt**
- A better way to do Gram Schmidt.
- In classical Gram-Schmidt we take each vector and make it orthogonal to all previous vectors (all columns to the left).
- In modified Gram-Schmidt we take each vector and modify all columns to the right to be orthogonal to it.
- Algorithm change:
  - Start with the left column and subtract the projection of all the columns to the right, onto the left most column, away from the columns on the right. This modifies all the columns so they are orthogonal to the leftmost vector.
  - Move on to the second leftmost column, subtract away the projection of all the columns to the right of it, onto the second leftmost column. This modifies all the columns to the right of the second leftmost vector to be orthogonal to it.
  - Continue to loop through all columns.
  - Normalize the resulting vectors

8.5.3 Householder Reflections

- **Most stable.**
- The goal is to multiply A by many unitary matrices, $Q_n Q_{n-1} ... Q_2 Q_1$, such that

$$Q_n Q_{n-1} ... Q_2 Q_1 A = R$$

  where R is an upper triangular matrix.
- Each Q_i is a unitary matrix, and the product of unitary matrices is a unitary matrix.
- Therefore we can define

$$Q^\intercal = Q_n Q_{n-1} ... Q_2 Q_1$$

  and

$$R = Q^\intercal A$$

  and we've found both Q and R.
- Each application Q_i to a matrix A will zero out all values below the i^{th} pivot, and does not change the rows above the i^{th} pivot.
- This is achieved by carefully selecting the elements in the matrix Q_i, such that the elements in the given column will zero out, with the constraint that the matrix Q_i must be unitary.

8.6 Eigendecomposition $A = S \Lambda S^{-1}$

- Works for diagonalizable, square matrices (complete matrices).
- S is a square matrix whose i^{th} column is the i^{th} eigenvector of A.
- Λ is a diagonal matrix, whose diagonal elements are the eigenvalues of A.

8.7 Singular Value Decomposition (SVD) $A = U\Sigma V^\intercal$

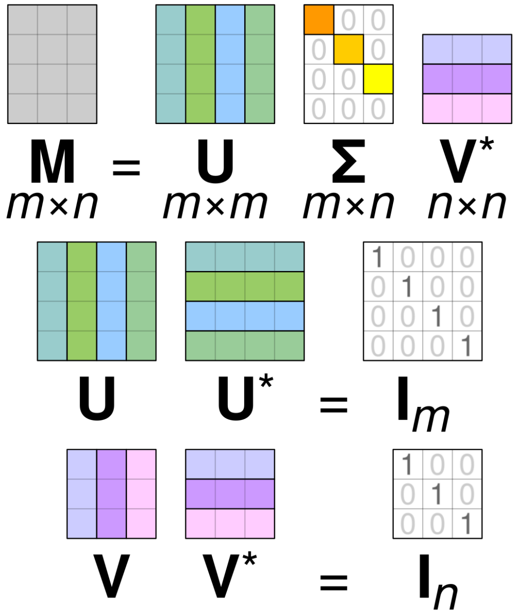

Figure 9: Visualization of the decomposed components of matrix using SVD. Image from Wikimedia Commons.

- Works for all $m \times n$ real and complex matrices (the holy grail of decompositions).
- $A \in \mathbb{R}^{m \times n}$
- U is a complex unitary matrix of size $m \times m$, whose i^{th} column is the i^{th} left-singular vector of A.
  - The left-singular vectors are the eigenvectors of $AA^\intercal$
  - $A\mathbf{v} = \sigma\mathbf{u}$
- V is a complex unitary matrix of size $n \times n$, whose i^{th} column is the i^{th} right-singular vector of A.
  - The right-singular vectors are the eigenvectors of $A^\intercal A$
  - $A^\intercal\mathbf{v} = \sigma\mathbf{u}$
- Σ is an $m \times n$ rectangular diagonal matrix which non-negative real numbers on the diagonal.
  - The diagonal entries of Σ are the singular values, σ, of A.
  - The singular values are the square roots of the eigenvalues of $A^\intercal A$ or $AA^\intercal$
- $A = \sum\limits_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\intercal$
- If $A \in \mathbb{M}_{nn}$ then $|\det(A)| = \prod_i \sigma_i$
- Has many uses such as low rank approximation, pseudoinverse, and calculating norms.

8.7.1 Low Rank Approximation

- Define $A_v = \sum\limits_{i=1}^{v} \sigma_i \mathbf{u}_i \mathbf{v}_i^\intercal$ where $A \in \mathbb{R}^{m \times n}$. This is the low rank approximation. A Rank 2 approximation of A is A_2, a rank 3 approximation of A is A_3, and so on.
- $\|A - A_v\| \leq \|A - B\|$ for any $m \times n$ matrix B with rank v.
- This means that the low rank approximation, A_v is the best approximation for A out of any possible $m \times n$, rank v matrix.

- We can use this, for example, in image compression.

8.7.2 Pseudoinverse

- The pseudoinverse of A is $A^{-1} = V\Sigma^{-1}U^\intercal$
- This is a generalization of the inverse matrix for a square matrix.
- Can be used to find the "best fit" solution to systems of linear equations that don't have a solution.

8.7.3 Norms

- 2-norm of a matrix: $\max(\sigma_i)$
- Frobenius norm: $\sqrt{\sum_i \sigma_i^2}$

8.8 Exercises

1. Markov Chain matrices are square and diagonalizable. Use that fact to find the equilibrium Markov chain transition matrix (M^∞) and equilibrium probability vector (p_∞) given the following eigendecomposition of the Markov Chain matrix M.

$$M = S\Lambda^{-1}S = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ -1/3 & -1/3 & 2/3 \\ -1/3 & 2/3 & -1/3 \end{bmatrix} \tag{1}$$

2. Why is $\sigma = \sqrt{\lambda_i}$ where σ_i are the singular values of A, and λ_i are the eigenvalues of $AA^\intercal$ or $A^\intercal A$?

3. If I double the size of a matrix, the time to perform an LU decomposition will increase by what factor?

4. Is the following a valid LU decomposition? Explain.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 5 & 1 & 1 & 0 \\ 1 & 2 & 4 & 1 \end{bmatrix} \begin{bmatrix} 4 & 1 & 4 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

5. Is the following a valid QR decomposition? Explain.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1\sqrt{2} & 1\sqrt{2} \\ 0 & 1\sqrt{2} & 1\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 4 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 2 \end{bmatrix}$$

6. Rank the three methods for performing QR decomposition from least stable to most stable.

7. $A = QTQ^\intercal$ is what type of a matrix decomposition? What type of matrices are Q and T?

8. (T/F) All matrices have a singular value decomposition.

9. (T/F) A 10×10 matrix has 8 singular values. The inverse of the matrix exists.

9 Probability and Statistics

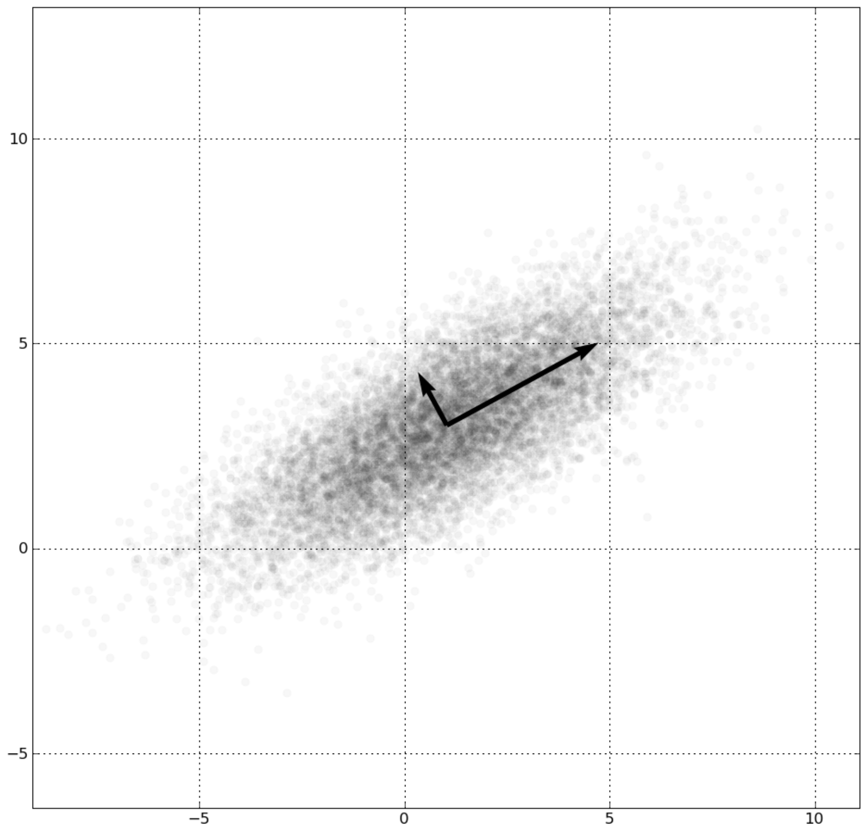

Figure 10: PCA of a distribution of data. The arrows show the eigenvectors of the covariance matrix. They are scaled by the square root of their corresponding eigenvalues. Another example of how eigenvalues and eigenvectors tell us fundamental characteristics of our matrix. Image from Wikimedia Commons.

9.1 Basics

- **Mean**: $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$
- **Expected value**: $\mu = E[x] = \sum_{i=1}^{n} p_i x_i$ where p_i is the probability of x_i
- **Sample Variance**: $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- **(True) variance**: $var(x) = \sigma^2 = E[(x - E[x])^2]$
  - We can only use this if the true population mean is known.
- **Standard Deviation**: $\sigma = \sqrt{\sigma^2}$
- **Covariance**: $\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$

9.2 Covariance Matrix

The covariance measures the linear dependencies between variables in a set of data. A simple 2×2 example is given below.

$$V = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix} \tag{2}$$

- $\sigma_{\text{xx}} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$
- $\sigma_{\text{yy}} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$
- $\sigma_{\text{xy}} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$

9.3 Principal Component Analysis (PCA)

- The principal components of a data set are the eigenvectors of the data's covariance matrix.
- Say we have a data set D where each row is a random variable and each column is a sample measurement of the random variables.
- We can turn the above covariance equation into a matrix equation as follows:

  $$C = \frac{1}{n-1} X^{\intercal} X$$

  Where X is the *mean centered data* matrix: $X = D - \bar{D}$.
- There are two ways to find the eigenvectors of the covariance matrix, C:
  1. **Eigendecomposition**:

     $$C = Q \Lambda Q^{\intercal}$$

     The columns of Q are the principal directions of the data set D.
  2. **SVD decomposition**: Start with the SVD of the mean centered data matrix X.

     $$X = U \Sigma V^{\intercal}$$

     Plug this into the equation for the covariance matrix C.

     $$C = \frac{1}{n-1} X^{\intercal} X = \frac{1}{n-1} V \Sigma (U^{\intercal} U) \Sigma V^{\intercal} = V \left(\frac{\Sigma^2}{n-1} \right) V^{\intercal}$$

     Therefore we get that the right-singular vectors V, of the mean centered data matrix, X, are the eigenvectors (principal component directions) for the covariant matrix C. And the Eigenvalues of C are on the diagonal of $\frac{\Sigma^2}{1-n}$
- We can also use PCA to project our data D onto the principal directions, and to reduce dimensionality of the problem.

9.4 Monte Carlo Methods

- Monte Carlo methods use repeated random sampling to obtain results like mean and variance.
- We can use them to create a covariance matrix if we are given a mathematical model of a given problem.
- Brute force algorithm:

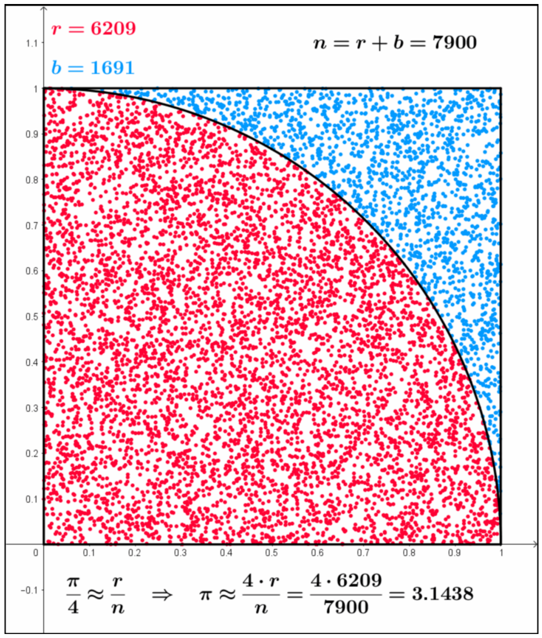

Figure 11: Plot of a standard pseudorandom Monte Carlo sampling method used to estimate the value of π. Image from Wikimedia Commons.

  - Find the probability distribution functions (PDF) of the input variables to the model.
  - Pick random variable inputs from the PDFs to run the model.
  - Get the output and include it as a data point in the data matrix D

- There are improvements over the brute force algorithm:
  - **Quasi-Monte Carlo methods**:

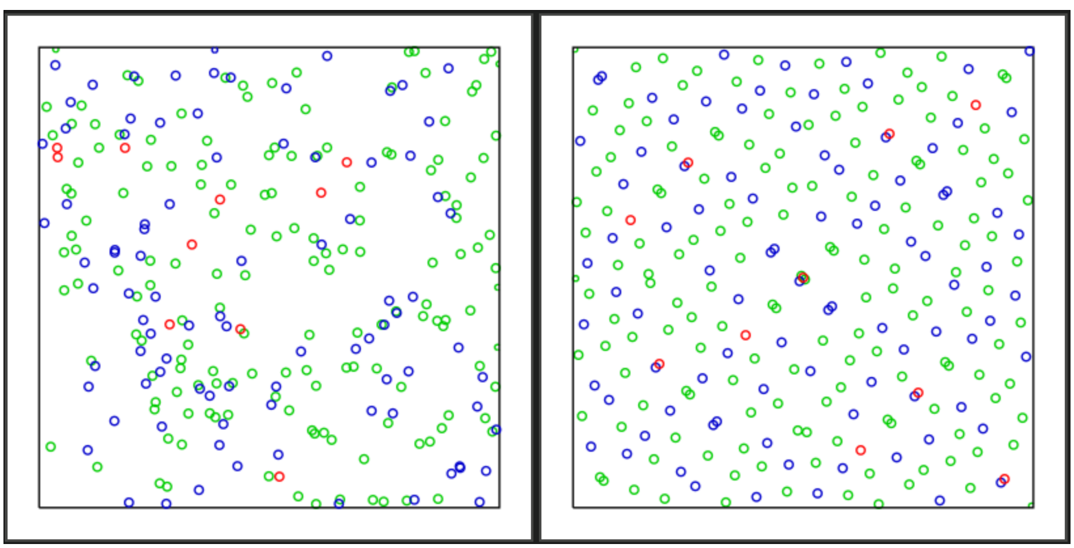

Figure 12: Standard pseudorandom sequence on the left and a quasi=monte carlo method on the right, with a modified spread of input variables. Image from Wikimedia Commons.

    * We can often achieve a faster rate of convergence by modifying the spread of the input variables.
  - **Latin Hypercube**:
    * Another sampling method where we overlay a grid on the sample domain, and require that there is one sample in each row and each column.

* For higher dimensional data, we overlay a set of axis-aligned hyperplanes as a grid.

9.5 Exercises

1. Compute the sample mean of the following data set: $[1, 3, 10, 2, 9, 6]$
2. Compute the sample variance of the following data set: $[1, 3, 10, 2, 9, 6]$
3. Consider the following data:

Trial	Data A	Data B
1	0	1
2	1	0
3	0	1
4	1	1

   Find the covariance matrix.

10 Appendix

10.1 Types of Matrices

- **Normal Matrix**: A square matrix A, such that $A^*A = AA^*$ where $*$ is the conjugate transpose. If A is real then $A^\intercal A = AA^\intercal$
- **Unitary Matrix**: A square matrix A, such that $A^*A = AA^* = I$. Where $*$ is the conjugate transpose. When A is real (as it basically always is in this class), then we call this an **orthogonal matrix.**
- **Orthogonal Matrix**: A real, square matrix A, such that $A^\intercal A = AA^\intercal = I$. This implies that $A^\intercal = A^{-1}$
- **Orthonormal Matrix**: A real, square matrix Q, that is both orthogonal ($Q^\intercal Q = QQ^\intercal$ and $Q^\intercal = Q^{-1}$), and whose columns are all normalized (the L2 norm of each column is 1). Therefore, they must satisfy the following two properties:
  1. Every column of Q is orthogonal to every other column (their dot products are all zero).
  2. The L2 norm of every column of Q is equal to 1.
- **Diagonal Matrix**: A matrix in which the entries outside the main diagonal are all zero.

$$\begin{bmatrix} 7 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

- **Positive Definite Matrix**: A matrix is positive definite if all its eigenvalues are positive.
  - This means that $\mathbf{x}^\intercal A\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}$.
- **Defective matrix**: If any of the eigenvalues of a matrix, A, are defective (algebraic multiplicity doesn't equal geometric multiplicity), then A is defective.
- **Complete matrix**: Opposite of a defective matrix. It is a matrix that has a complete set of eigenvalues. This means that all the eigenvalues of a matrix are complete (their algebraic multiplicity and geometric multiplicity are equal).
- **Semi-positive definite matrix**: All eigenvalues of the matrix are greater than or equal to 0, $\lambda \geq 0$.
- **Negative definite matrix**: All eigenvalues of the matrix are less than zero, $\lambda < 0$.