STAT 24300

Anthony Yoon and Martin Pestana

October 2024

Abstract

This are my notes for STAT 24300 taught by Anjali Nair, typed up because god forbid, no one can read my own handwriting. That includes me. These also are based on the lecture notes of Martin Pestana. In all honesty, if you studied 183/184 or 19620 well enough, the first half of this is review. All the proofs you need should be in the very last section. **There are a lot of typos, use common sense to fill in the gaps**

Contents

1 Lecture 1

1.1 Properties of vector space

Refer to my linear algebra notes (MATH 19620) notes for this for more in-depth explanations. The only thing to note is that we had a rigorous definition of a vector space. It goes as follows:

- **Closure under addition:**
  For all $u, v \in V$, their sum $u + v$ is also in V.
  $$u + v \in V$$
- **Closure under scalar multiplication:**
  For any scalar $a \in F$ and any vector $v \in V$, the scalar multiple $a \cdot v$ is in V.
  $$a \cdot v \in V$$
- **Associativity of vector addition:**
  For all $u, v, w \in V$, the following holds:
  $$u + (v + w) = (u + v) + w$$
- **Commutativity of vector addition:**
  For all $u, v \in V$, the following holds:
  $$u + v = v + u$$
- **Existence of an additive identity:**
  There exists an element $0 \in V$ (called the zero vector) such that for every vector $v \in V$, the following holds:
  $$v + 0 = v$$
- **Existence of additive inverses:**
  For every $v \in V$, there exists a vector $-v \in V$ such that:
  $$v + (-v) = 0$$
- **Distributivity of scalar multiplication with respect to vector addition:**
  For all $a \in F$ and $u, v \in V$, the following holds:
  $$a \cdot (u + v) = a \cdot u + a \cdot v$$
- **Distributivity of scalar multiplication with respect to scalar addition:**
  For all $a, b \in F$ and any vector $v \in V$, the following holds:
  $$(a + b) \cdot v = a \cdot v + b \cdot v$$
- **Compatibility of scalar multiplication with field multiplication:**
  For all $a, b \in F$ and any vector $v \in V$, the following holds:
  $$(a \cdot b) \cdot v = a \cdot (b \cdot v)$$
- **Identity element of scalar multiplication:**
  For every vector $v \in V$, multiplying by the scalar 1 (the multiplicative identity in F) gives:
  $$1 \cdot v = v$$
- **Closure under linear combinations:**
  For every vector $v, w \in V$ and any scalar α, β gives[1]:
  $$\alpha v + \beta w \in V$$

[1] This is just a combination of the closure under addition and multiplication, but I thought it would be good to add.

We also can be dealing with the complex plane. This is denoted as $\mathbb{C}$, and each entry is a complex number. This idea is extended in $\mathbb{C}^n$, so it is okay. We should also note the concept of what the transpose is:

$$v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \qquad v^t = \begin{bmatrix} v_1, v_2, v_3 \end{bmatrix}$$

Here are some nice properties to know about the transpose:

- For any matrix A, the transpose of the transpose returns the original matrix: $(A^T)^T = A$.
- The transpose of a sum is the sum of the transposes: $(A + B)^T = A^T + B^T$.
- The transpose of a scalar multiple of a matrix is the scalar multiple of the transpose: $(\alpha A)^T = \alpha A^T$, where α is a scalar.
- The transpose of a product of two matrices reverses the order of multiplication: $(AB)^T = B^T A^T$.
- For a square matrix A, A is symmetric if $A = A^T$.
- The transpose of an identity matrix is itself: $(I_n)^T = I_n$, where I_n is the $n \times n$ identity matrix.
- The rank of a matrix is equal to the rank of its transpose: $\text{rank}(A) = \text{rank}(A^T)$.
- If A is an invertible matrix, then the transpose of the inverse is the inverse of the transpose: $(A^{-1})^T = (A^T)^{-1}$.

You should know what the dot product and cross product is as well. We didn't learn what the cross product is, but it is a way to find a vector that is orthogonal (know what this means, dot product = 0) to both of the 2 vectors.

We also defined dot products as the following

$$\langle v, w \rangle = v^t w$$

but any other definition of dot product works, this is just the one we are used to using. We can also define magnitudes as:

$$\|v\| = \sqrt{\langle v, v \rangle}$$

We can think of magnitudes as lengths for vectors in $\mathbb{R}^2$.

1.2 Properties of the Dot Product

Let $\mathbf{a}$ and $\mathbf{b}$ be vectors in $\mathbb{R}^n$, and let c be a scalar. The dot product of two vectors, denoted $\mathbf{a} \cdot \mathbf{b}$, has the following properties:

Commutativity

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$$

The order of the vectors does not affect the result of the dot product.

Distributivity over Vector Addition

$$\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$$

The dot product distributes over vector addition.

Scalar Multiplication

$$(c\mathbf{a}) \cdot \mathbf{b} = c(\mathbf{a} \cdot \mathbf{b})$$

Multiplying a vector by a scalar before taking the dot product is equivalent to taking the dot product first and then multiplying by the scalar.

Relationship with Magnitudes (Self Dot Product)

$$\mathbf{a} \cdot \mathbf{a} = \|\mathbf{a}\|^2$$

The dot product of a vector with itself is the square of its magnitude, where $\|\mathbf{a}\|$ denotes the Euclidean norm of $\mathbf{a}$.

Orthogonality If $\mathbf{a} \cdot \mathbf{b} = 0$, then $\mathbf{a}$ and $\mathbf{b}$ are orthogonal (perpendicular).

Angle Between Vectors

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos\theta$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$. This equation shows that the dot product is related to the magnitudes of the vectors and the cosine of the angle between them.

1.3 Matrix Vector Multiplication

Assume that we are given the matrix:

$$A_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

and the vector

$$v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix}$$

We can define matrix vector multiplication as:

$$A_{m\times n}v_{m\times n} = \begin{bmatrix} a_{11}v_1 & a_{12}v_2 & a_{13}v_3 & \dots & a_{1n}v_n \\ a_{21}v_1 & a_{22}v_2 & a_{23}v_3 & \dots & a_{2n}v_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn}v_n \end{bmatrix}$$

1.4 Gaussian Elimination

Again, refer to my MATH 19620 Notes on this topic, but the gist of it is

- If we are given a system of equations, we can write an matrix that represents the coefficients of each equation in the system of equations.
- We can then setup the **Augmented Matrix**, where we add an additional column such that it represents what each equations equal
- We can then use **row operations**, which represents basic operations we can do matrices (I.E adding two equations together, scalar multiplication, and switching rows)
- Then we can do back substitution. [2]

This is an example of this

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 4 \\ 2 & 3 & 1 & 7 \\ -1 & -1 & 2 & -3 \end{array}\right]$$

Step 1: $R_2 - 2R_1 \to R_2,\ R_3 + R_1 \to R_3$

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 4 \\ 0 & -1 & 3 & -1 \\ 0 & 1 & 1 & 1 \end{array}\right]$$

Step 2: $-R_2 \to R_2$

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 4 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right]$$

Step 3: $R_3 - R_2 \to R_3$

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 4 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 4 & 0 \end{array}\right]$$

[2] If you know how to do things RREF, just do it. It's a lot a easier that back substitution. Higher chances of making mistakes, but like might as well do it you know?.

Step 4: $\frac{1}{4}R_3 \to R_3$

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 4 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

Step 5: $R_2 + 3R_3 \to R_2$

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 4 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

Step 6: $R_1 + R_3 \to R_1$

$$\left[\begin{array}{ccc|c} 1 & 2 & 0 & 4 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

Step 7: $R_1 - 2R_2 \to R_1$

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

And yes, this example is GPTed. I am not LaTeXing all this. Note for the purposes of this problem, you really only need to go to step 4. But, if you want to go to RREF, complete it all the way through.

Determinant of a matrix

The determinant of a square matrix is a special scalar value you can derive that provides insights into the factors of the matrix: such as if it is invertible, nature of eigenvalues, etc. I don't think the conditions of a invertible matrix are covered, so I'll cover them later.

Note a matrix is invertible if and only if $\det A \neq 0$. We will cover the two cases [3] of calculating the determinant.

Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

then the determinant of A is calculated as follows:

$$\det(A) = ad - bc$$

This works for all 2 by 2 matrices. For 3 by 3 matrices, we can see that if we let A be

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

Then, we can calculate the determinant as follows:

$$\det(A) = a \det\left(\begin{bmatrix} e & f \\ h & i \end{bmatrix}\right) - b \det\left(\begin{bmatrix} d & f \\ g & i \end{bmatrix}\right) + c \det\left(\begin{bmatrix} d & e \\ g & h \end{bmatrix}\right)$$

1.5 Singular vs Non-singular

A **Non-singular matrix** (or invertible) matrix is a matrix that has a non-zero determinant. This means that it has an unique inverse A^{-1} and an unique solution

A singular matrix is a matrix that with a zero determinant. This means that the matrix has no inverse nor an unique solution. To find systems that are singular, we can first find cases where there is a row with all zeros when we do row reduction, then the system is singular. [4]. We can also note something about the consistency of the matrix:

- If in the 0 row, we have a corresponding 0 in the "b" column. Then the system is **Consistent** because we have infinitely many solutions because we have a free variable.
- If there lies a contradiction in one of the row's of the matrix, this means that the system i **inconsistent** and thus means that there are no solutions to this system.

[3] There are **A LOT** more than this, but it really doesn't matter. This class is weird

[4] Technically Speaking, if there is a column of 0s, that also shows that the matrix is singular.

1.6 Linear Dependence

If at least one of the vectors in a set can be expressed as a linear combination (meaning vectors multiplied by a scalar added together), then the set of vectors is linearly dependent. This means that at least one vector is redundant, and the set can be represented more succinctly.

We now introduce the idea of linear independence. This means that there is no possible way of representing any vector in a span by any linear combination of vectors. I.E, let $v_i \in \mathbb{R}^n$, and $c_i \in \mathbb{R}$. The vectors are linearly independent if

$$c_1v_1 + c_2v_2 + c_3v_3 + \dots c_nv_n = 0$$

In other words, the only way to get all the vectors equal to 0 if all c_is are equal to 0.

1.7 Null Space

We note that the null space [5] is the set of vectors that when multiplied by A, gives us the 0 vector. The nullspace of a matrix A can tell us whether the columns of A are linearly dependent or independent.

- If the Nullspace contains only the zero vector (I.E $Null(A) = \{0\}$, it means there are no non-trivial combinations of the columns that result in the zero vector. AKA, this implies that the columns are linearly independent.
- If the nullspace contains non-zero vectors, this implies that there are some combinations of vectors that lead to the 0 vector. Thus, this means that the columns themselves are linearly dependent.

We should note something about the nullspace, all vectors in the null space are orthogonal to the rows of A. So if $z \in Null(A)$, then z is orthogonal to all rows in A.

1.8 Free variabiity in linear systems

Essentially, what this boils down to is the idea that if there are **fewer independent equations than unknowns**, then free variables come up. A free variable is a variable that we can set any value to.

2 Lecture 2

Not much really to cover here? I'll go through some proofs that were brought up in class.

2.1 Analyzing existence and uniqueness of linear solutions

Singularity is the property of A. However, the solution is determined whether b exists. Of course, this only works in the form:

$$Ax = b$$

Also, the null space is defined as the set of all vectors such that $Ax = 0$. This implies that the x vector is orthogonal to all rows in A.

We can also claim that following is a subspace: For a fixed vector in $V \in \mathbb{R}^n$, the set $\{z \in \mathbb{R}^n | z \perp v\}$

Some definitions to clear up on: **Subspace (W)**: This is a vector space contained in a larger vector space (v). You can think of this as a subset of an another vector space.

Linear Dependence versus Linear Independence This concept uses the same concepts as earlier. *Linear Independence* relies on the idea that we cannot represent any vector in the vector space as a linear combination of the others. I.E, if we we are given a set of vectors:

$$A = \begin{bmatrix} | & | & | & \dots & | \\ a_1 & a_2 & a_3 & \dots & a_n \\ | & | & | & \dots & | \end{bmatrix}$$

$$\sum_{i=1}^{n} \alpha_j a_j = 0 \text{ if and only if } \alpha_j = 0 \, \forall j$$

[5] Or Kernel. Same thing

And conversely, *Linear Dependence* is the idea that we *can* represent a vector in the vector space as a linear combination of other vectors in the space. We can represent this in the following

$$\sum_{j=1}^{n}$$

2.2 Nullspaces

Here are some things to know about null spaces.

- These are subspaces consisting of vectors z such that Az = 0. This also means that every row in A will be orthogonal to every vector in the null space.
- The **0** vector is in every null space. Note that if the null space only has the **0** vector, then the null space is deemed **trivial**
- If the null space is not trivial, then solutions to Ax = b are not unique.

We can note that for a system Ax = b, we can rewrite it as

$$y = x^* + z$$

This is the proof on why this holds.

Proof. Suppose there exists an x^* such that $Ax^* = b$. Now we claim that there also exists an y such that $Ay = b$. We can then do some manipulation to do the following:

$$\begin{aligned} Ax^* &= Ay \\ Ax^* - Ay &= 0 \\ A(x^* - y) &= 0 \end{aligned}$$

Let $z = x^* - y$. This implies that $Az = 0$. This means that z is an element of the Null Space. But, given our construction of z, we can note that we can write the following:

$$x^* = y + z$$

This means that given a solution to a system, we can represent the solution as a solution to the same system combined with an element to the null space. □

3 Lecture 3

We now introduce the idea of a span. A span is *a collection of all possible linear combinations of a certain collection (set) of vectors.* We can interpret this geometrically:

- One vector in a span means that it is just a line. When we multiply this singular vector by a scalar value, we can note that we will get a line.
- 2 vectors in a span means that it is a plane. This means that when we added the two vectors in the plane with linear combinations of each other, we can note that we will get a plane. We will only stay the directions directed by the vectors, but then it doesn't take any other direction.
- 3 vectors in a span means that using these 3 vectors, we can recreate the entirety 3d space. The intuition for this is that we have 3 vectors, which will create the whole 3d space as there are 3 unique directions we can take that we can use to create the whole space.

3.1 Basis of a vector space

Given a span for a vector space, we know that it is a basis if and only if the the vectors in this vector space are all linearly independent. It also be ensured that the span can represent all vectors within the vector space. Note that there is no **one** basis for a vector space. Rather, the basis of choice is determined by the vectors within the set that creates the span.

3.2 Dimension of a Vector Space

The dimension of a vector space is the number of the vectors in the basis hat represents the entire vector space. Note that this concept applies to subspaces. We do have some things we can derive from the dimension of a space.

Let us be given two vector spaces, V and W. We can reach the following conclusions if we do the following:

- $\dim(V) = \dim(W)$, this really doesn't tell us anything. The two vector spaces are only equal if and only if they operate under the same fields (I.E Same types of scalar multiplication, addition, multiplication) Only if you guys know that property, then this holds true.
- $\dim(V) \leq \dim(W)$, this implies that V is a subspace of W
- If we know that W is a subspace of V, we know that $\dim(V) \leq \dim(W)$.

3.3 Range of a matrix

The range of a matrix

- intuitively is every linear combination of the columns of A.
- You can also think of it as when we multiply any arbitrary vector by A, it is the set of all possible outcomes.

Note something important here. $Rank(A) = \dim(Range(A))$ if and only if the $Range(A)$ is a basis. This means that columns are linearly independent. Note that this subspace must be a subspace of $\mathbb{R}$ to the number of columns in the matrix. We also must note that if $b \in Range(A.$ this implies that the there exists at least one solution to Ax = b. A similar logic applies to he rows as well.

$$\text{Row Rank}(A) = dim(Range(A^T)) = \text{Column Rank}(A)$$

3.4 Nullity

Using the same definition of the null space as before, we define Nullity as the *dimension of the null space* and this tells us *the number of free variables in the homogeneous system.*. Rigoursly, we can say that:

$$\dim(\text{nullspace}) := \text{nullity} A$$

If the nullity of a matrix A is greater than 0, then there is exists a non-trivial null space However, if the nullity of the matrix is equal to 0, then this implies that we have a trivial null space, which means that the only vector in the null space is the 0 vector. Here is an example of finding the nullspaace:

3.5 Example of finding the null space

Consider the matrix A:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 1 & 1 \end{bmatrix}$$

We want to find the null space of A, which consists of all vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ such that $A\mathbf{x} = 0$.

To find this, we solve the equation:

$$A\mathbf{x} = 0$$

This expands to the following system of linear equations:

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 0 \\ 2x_1 + 4x_2 + 6x_3 = 0 \\ x_1 + x_2 + x_3 = 0 \end{cases}$$

We can write this system in augmented matrix form:

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 2 & 4 & 6 & 0 \\ 1 & 1 & 1 & 0 \end{array}\right]$$

Using Gaussian elimination, we start by eliminating x_1 from the second and third rows. Subtracting 2 times the first row from the second row gives:

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{array}\right]$$

Next, we subtract the first row from the third row:

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & -2 & 0 \end{array}\right]$$

Now we have a simplified system:

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 0 \\ -x_2 - 2x_3 = 0 \end{cases}$$

From the second equation, we can express x_2 in terms of x_3:

$$x_2 = -2x_3$$

Substituting $x_2 = -2x_3$ into the first equation gives:

$$x_1 + 2(-2x_3) + 3x_3 = 0$$

$$x_1 - 4x_3 + 3x_3 = 0$$

$$x_1 = x_3$$

Thus, the solutions to this system can be written in terms of the free variable x_3 as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_3 \\ -2x_3 \\ x_3 \end{bmatrix} = x_3 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

This shows that any vector in the null space of A is a scalar multiple of $\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. Therefore, the null space of A is:

$$\text{Null}(A) = \text{span}\left\{ \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \right\}$$

This non-trivial null space contains all scalar multiples of $\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. [6]

3.6 Rank-Nullity Theorem

This theorem is important because this tells us the relation between the nullity and rank of a matrix. We can denote this theorem as the following. If we were to express any matrix with n columns, we can see that

$$\text{Nullity(A)} + \text{Rank(A)} = \text{n}$$

If we were to take A^T, this logic would also apply to the rows of the matrix. This is important.

3.7 Behaviors of matrices given their dimensions

We are given a matrix $A_{m \times n}$, where m denotes the numbers of rows and n denotes the number of columns.

- $\mathbf{m > n}$. This means that there are more rows than columns. This means that there are more unknowns than equations, thus meaning that this is a overdetermined and over-constrained system. This often leads to no solutions.
- $\mathbf{m < n}$ This means that there are more columns that rows. This means that the system is underdetermined of underconstrained. Thus, this means that by the rank nullity theoremn (I encourage you to do the proof by yourself), $nullity(A) > 0$
- $\mathbf{m = n}$ This means that there all the columns and rows are linearly independent, where rank m = n. Thus, we can conclude that for the system $Ax = b$, for any b, we can find an unique solution to it. This is known as a full rank system.

[6] Yes this is GPTed. I am not willingly LaTeXing all this

4 Lecture 4

4.1 The 4 fundamental subspaces

There are 4 fundamental subspaces given a matrix $A_{m\times n}$ They are as follows:

- $range(A)$. This is the basis of the columns of A. This is by the definition of the range as we defined earlier. Note that $\dim(range(a)) = rank(A) = r$
- $null(A)$. These are all the vectors that we multiply A by to get the zero vector. Thus, every vector in this space is orthogonal to every row in A. We can denote this as a set in the following manner: $\{z : Az = 0\}$.
- $range(A^t)$. This is the span of the rows of A. This is also the column space of A^t. Note that $dim(range(A^t)) = r$.
- $null(A^t)$ = These are all the vectors we multiply A^t by to get the zero vector. Note that every elemenet in this subspace must be ortogonal to every column in A. We can write this in set notation as follows: $\{z : A^tz = 0\}$.

We also know the fact that $null(A)$ and row space $C(A^t)$ are subspaces of $\mathbb{R}^n$ whearas the nullspace $N(A^t)$ and the column spaces $C(A)$ are subsapces of $\mathbb{R}^m$.

Here are soome definitions to know of. $rank(A)$ is the number of linearly independent vectors that span the column space. We will also have to note that **row rank = column rank**. We can also define $rank(A) = dim(range(A))$.

4.2 Fundamental Theorem of Linear Algebra Part 1

There are 4 parts to this. Given that $rank(A) = r$ and $null(A) = n$, we can see that

- $C(A)$ column space of A, dimension = r
- $N(A)$ = nullspace of A, dimension = n - r
- $C(A^T)$ = rowspace of A, dimension = r
- $N(A^t)$ = left nullspace of A, dimension = m - r

We also have defined the direct sum of vectors as the following, Let us consider the case where $V \subset W$. We can define the direct sum, denoted by $\oplus$ as

$$v \oplus w = \{\alpha v + \beta w, v \in V, w \in W\}$$

where α, β are scalars. We can also denote this if we are given the following. If we let $v = \text{span}\{v_1, v_2, \ldots, v_m\}$ and $w = \text{span}\{w_1, w_2, \ldots, w_n\}$, then we can say that

$$v \oplus w = \text{span}\{v_1, v_2, v_3, \ldots, v_m, w_1, w_2, \ldots, w_n\}$$

We continue our discussion of the fundamental theorem of linear algebra. For a matrix $A_{m\times n}$, we can find that the following are true:

$$range(A) \oplus null(A^T) = \mathbb{R}^m$$
$$range(A^T) \oplus null(A) = \mathbb{R}^n$$

This is just an extension of the rank nullity theore.

4.3 Orthogonality

This is when things get slightly confusing. Let V and W be two different vector spaces. We can claim that they are orthogonal if we for any vector in V, denoted by v and any vector in W, denoted by w, we get v is perpendicular to w. However, this is different to the concept of an orthogonal complement. Think of this as a set of vectors that is orthorgonal to every vector in a space. So if we conisder the subspace W such that it is an orthogonal complement of V, we get

$$W = \{w | v \perp w \forall v \in V\}$$

We must note that **these two concepts are not the same.** There is a difference. Orthogonal complements are a one way thing; vectors in the orthogonal to each other in one way, $V \to W$, but not $W \not\to V$. Orthogonal sets are both ways, so $V \iff W$.

If we know that 2 vectors are orthogonal to each other, we can conlcude some things about the vectors themselves. Let x, y be two vectors that are orthogonal to each other. Then the following holds:

$$\|x\|^2 + \|y\|^2 = \|x + y\|^2 = \|x + y\|^2$$

We also note that the inner product is symmetric, which means that:

$$x^t y = y^t x$$

4.4 Fundamental Theorem of Linear Algebra Part 2

The $range(A)$ and $null(A^T)$ are orthogonal complements to each other. And similarly $range(A^T)$ and $null(A)$ are orthogonal complements to each other.

Remark: $Ax = b$ is solveable if and only if $y^T b = 0$ whenever $y^T A = 0$.

4.5 Least Squares

Within statsitics, there are inconsistent equations that arise all the time. Sometimes, it is impossible to fit solutions to some equations. However, to circumvent this we can also minimize the error that comes as a result from with it. Let us begin with the following system for simplicity sake.

$$2x = b_1$$
$$3x = b_2$$
$$4x = b_3$$

We choose an x such that the average error E in the m equations. We choose to represent this using the squared error form:

$$E^2 = (2x - b_1)^2 + (3x - b_2)^2 + (4x - b_3)^2$$

if there is an exact solution, the minimum error is $E = 0$ This means that the graph of E^2 will be a parabola with it's vertex intersecting with the x axis. In order to find th minimum error, we can differentiate the error expression in the following manner.

$$\frac{\partial E^2}{\partial x} = 2\left((2x - b_1)2 + (3x - b_2)3 + (4x - b_4)4\right) = 0$$

When we simplify this equation down, we can get the least squares solution:

$$\bar{x} = \frac{2b_1 + 3b_2 + 4b_3}{2^2 + 3^2 + 4^2} = \frac{a^T b}{a^T a}$$

we can generalize this following form. Given the system $A\bar{x} = b$, the solution the system is

$$\hat{x} = \frac{A^T b}{A^T a} \iff A^T b\hat{x} = A^T b$$

4.6 These concepts applied to least squares

Full discliamer. The professor did a really bad job explaining this. Like. Really bad. **The main intution behind this proof is that we can break up any vector into any two components and that b is not in the range of (A)**. So for example, let us say we have a vector b, and the system $A = b$ We can break this vector up in the folowing manner.

$$b = b_1 + b_2$$

b_1 can be described as one leg and b_2 as the other. However, we can be smart about this. We can let $b_1 \in range(A)$ and $b_2 \in null(A^T)$. We are doing so because we are aware of the fact that $range(A)$ and $null(A^t)$ are orthogonal complements to each other. From here, note the following:

$$Ax = b \iff Ax - b = 0 \iff Ax - b \in \mathbb{R}^m$$

So thus, we have already proven that $b = b_1 + b_2$, we can break up the vector in the folllowing fashion:

$$Ax - b = Ax - b_1 - b_2$$

however, since we know that $Ax - b_1$ and b_2 are orthogonal to each other, we can see that we can do the following the via the Pythagorean Theorem.

$$\|Ax - b\|^2 = \|Ax - b_1\|^2 + \|b_2\|^2$$

However, we can choose an x^* such that $Ax^* = b_1$. And when we do so, we can note that $\|Ax - b_1\|^2 = 0$. Thus, we can call x^* the argumenet that minimizes the sqaured error of the problem. We are then interested in. Thus, we can do the following:

$$Ax^* - b = (Ax^* - b)_{range(A)} + (Ax^* - b)_{null(A)}$$

Future Anthony here: I did the proof in the proof section in the back. Check it out there!

5 Lecture 5

This section of covers many of the properties of matrices. A lot of this is pretty trivial but, we can generalize this to the following properties:

- **Addition for matrix**: $[A + B]_{ij} = a_{ij} + b_{ij}$
- **Scalar multiplication**: For a scalar λ we can write the following: $[\lambda A]_{ij} = \lambda a_{ij}$
- **Tranposes**: $[A^T]_{ij} = a_{ji}$
- **Matrix-Matrix product**: $C = A_{m\times n}B_{p\times q}$ if and only if $n = p, C_{m\times q}$. We can also view this from a entry wise interpretation, where: $[C]_{ij} = [AB]_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$
- There is also a column wise interpretation of his. It is as follows:

$$A\begin{bmatrix} | & | & \dots & | \\ b_1 & b_2 & \dots & b_v \\ | & | & \dots & | \end{bmatrix} = \begin{bmatrix} | & | & \dots & | \\ Ab_1 & Ab_2 & \dots & Ab_n \\ | & | & \dots & | \end{bmatrix}$$

  and the row wise interpretation would be:

$$\begin{bmatrix} -a_1- \\ -a_2- \\ \vdots \\ -a_n- \end{bmatrix} B = \begin{bmatrix} (-a_1-)b_1 \\ (-a_2-)b_2 \\ \vdots \\ (-a_n-)b_n \end{bmatrix}$$

- **Outer Product Definition**: The outer product definition of matrix multiplication involves expressing the product of two matrices A and B as a sum of outer products. Let A be an $m \times n$ matrix and B be an $n \times p$ matrix. Then the matrix product $C = AB$ can be written as:

$$C = AB = \sum_{k=1}^{n} \mathbf{a}_k \mathbf{b}_k^T$$

  where $\mathbf{a}_k$ is the k-th column of A (an m-dimensional vector), and $\mathbf{b}_k^T$ is the k-th row of B (a p-dimensional vector). Each outer product $\mathbf{a}_k\mathbf{b}_k^T$ is an $m \times p$ matrix.

  Expanding this in element-wise form, we have:

$$C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$$

  where C_{ij} represents the element in the i-th row and j-th column of the resulting matrix C.

We can also find inverses. Inverse matrices are those who simplify sastify the conditions:

$$AA^{-1} = I$$

where I denotes the identity matrix. This is true if and only if the determinant of the matrix is not 0. Here is an example of this:

Step 1: Set Up the Augmented Matrix

We start with the augmented matrix $(A|I)$:

$$\left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array}\right)$$

Step 2: Make the First Pivot Equal to 1

The pivot element in the first row and first column is already 1, so we proceed without any changes.

Step 3: Make the Elements Below the Pivot Zero

To make the element below the pivot (3 in the second row, first column) equal to zero, we perform the row operation $R_2 \to R_2 - 3 \times R_1$:

$$\left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array}\right)$$

Step 4: Make the Second Pivot Equal to 1

Now, divide the second row by -2 to make the pivot in the second row and second column equal to 1:

$$\left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 1 & \frac{3}{2} & -\frac{1}{2} \end{array}\right)$$

Step 5: Make the Elements Above the Second Pivot Zero

To make the element above the second pivot (2 in the first row, second column) equal to zero, we perform the row operation $R_1 \to R_1 - 2 \times R_2$:

$$\left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & 1 & \frac{3}{2} & -\frac{1}{2} \end{array}\right)$$

Conclusion

We have transformed $(A|I)$ into $(I|A^{-1})$, so the inverse of A is:

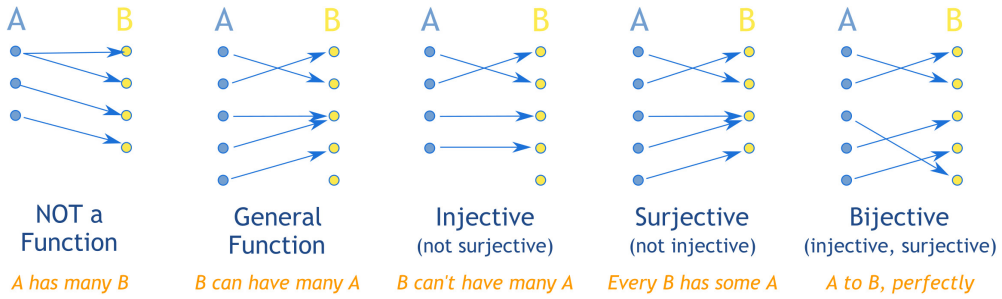
$$A^{-1} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$$

5.1 Injectivity, Surjectivity, Bijectivity

There are 3 main ideas within set theory. They go as follows:

- **Injectiviy**: Between two sets, you can map one element to another element in the other set, with no overlap. Mathematically, between if we have a function that takes us between the set A and B, $f(x_1) = f(x_2)$ if and only if $x_1 = x_2$.
- **Surjective**: Between two sets, you can map an element in one set to the other such that there exists a mapping between every element.
- **Bijectivity**: This means that a function is both injective and surjective.

This is a diagram that I ripped off of google that explains this well



Here are some things that it means in regards to linear algebra.

- **Injective (One-to-One) Function:** A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is injective if different inputs map to different outputs, meaning $T(\mathbf{x}) = T(\mathbf{y}) \Rightarrow \mathbf{x} = \mathbf{y}$. An example of an injective transformation is the mapping $T : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $T(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}$. This transformation is injective because no two different vectors in $\mathbb{R}^2$ are mapped to the same vector in $\mathbb{R}^3$.

- **Surjective (Onto) Function:** A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is surjective if for every element $\mathbf{y}$ in $\mathbb{R}^m$, there exists an $\mathbf{x}$ in $\mathbb{R}^n$ such that $T(\mathbf{x}) = \mathbf{y}$. For example, the transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $T(\mathbf{x}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x}$ is surjective because every vector in $\mathbb{R}^2$ has a pre-image under T.
- **Bijective (One-to-One and Onto) Function:** A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is bijective if it is both injective and surjective, meaning there is a one-to-one correspondence between elements of the domain and the codomain. An example of a bijective transformation is the mapping $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $T(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \mathbf{x}$. This transformation is bijective because each vector in $\mathbb{R}^2$ has a unique pre-image, and all vectors in $\mathbb{R}^2$ are covered.

These are in regards to linear transformations [7]. Next up, we have things regarding null spaces and ranges.

- **Range of A:** The range (or column space) of a matrix A is the set of all possible linear combinations of its columns. It represents the subspace of the codomain that A maps onto. If A is an $m \times n$ matrix, then the range of A, denoted as Range(A), is a subspace of $\mathbb{R}^m$. For example, if A is a surjective transformation, then Range(A) $= \mathbb{R}^m$, meaning that A maps onto the entire codomain. This occurs if A has rank m, meaning it has m linearly independent columns.
- **Null Space of A:** The null space (or kernel) of a matrix A, denoted Null(A), is the set of all vectors $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = 0$. It represents the subspace of the domain that is mapped to the zero vector in $\mathbb{R}^m$. If A is injective, then Null(A) $= \{0\}$, meaning that the only solution to $A\mathbf{x} = 0$ is the zero vector. For an injective transformation, A must have rank n, so there are no non-trivial solutions in the null space.
- **Range of A^T:** The range of the transpose A^T, denoted Range(A^T), is the subspace spanned by the rows of A. If A is an $m \times n$ matrix, then A^T is an $n \times m$ matrix, and Range(A^T) is a subspace of $\mathbb{R}^n$. This space can be important for understanding the solvability of $A\mathbf{x} = \mathbf{b}$. If A is surjective, then Range(A^T) $= \mathbb{R}^n$.
- **Null Space of A^T:** The null space of A^T, Null(A^T), is the set of all vectors $\mathbf{y} \in \mathbb{R}^m$ such that $A^T\mathbf{y} = 0$. This null space is orthogonal to the range of A, and its dimension provides insight into the rank-nullity theorem, which states that rank(A) + nullity(A) $= n$, where n is the number of columns of A. If A is an injective transformation, then Null(A^T) $= \{0\}$, meaning that the rows of A are linearly independent.

And these are some things we can note about[8]

6 Reduced Row Eichleon Form

Within linear algebra, there is a really important concept that is called **RREF, or Reduced Row Eichleon Form**. I don't know why we didn't cover this till now. But, here we are. A matrix is in RREF form if

- it's in eichleon form.
- All pivots are equal to 1
- each pivot is the only non-zero entry in its columns

But what is a pivot and being in eicheleon form? A pivot is the first non-zero element in a row. Eicheleon form means that a matrix has a staircase pattern of zeros and no zero values. An example of eichleon form can be seen below:

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 6 & 8 \\ 0 & 0 & 1 \end{bmatrix}$$

This is very important because of how we can use this notation to easily find the $rank(A)$, $nullity(A)$, $basis(range(A))$, $basis(null(A))$. But we continue, there are some terms that has to be noted.

- **Pivot Columns**: A column that contains a pivot point
- **Free Columns**: Column with no pivots.

[7] Weirdly, we haven't talked about linear transformations ...
[8] I gpted like all of this.

6.1 How to use RREF to gain crucial information

6.1.1 Finding the range

The main idea is that the **the pivot columns of the RREF form of the matrix corresponds to the columns in the original matrix that make up the range(A)**We begin with how to find the $range(A)$. Let us be given the matrix:

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 0 \\ 3 & 6 & 1 \end{bmatrix}$$

The RREF version of this matrix is:

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

Note that columns 1 and 3 are pivot columns. This means that the basis that makes up the range of the matrix is:

$$basis(range(A)) = \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

6.1.2 Finding the null space

Let us work with the same matrix. Note that it is really easy to find the null space with the RREF from of the matrix:

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

We can easily findn that the null space here is

$$null(A) = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

6.2 CR Decomposition

With this information, we can begin our first matrix decomposition. If we are given a matrix A, this can be decomposed with $A = CR$ such that

- C contains all non-zero columns of the RREF(A). We also remove any all 0 rows.
- R conatins all the columns that make up the range of (A). Or, we can think of it as all the pivot columns that are replaced with their original columns

The reason why this works is the most evident when we look at this from the outer product view of matrix-matrix multiplication. Because each column in A can be expressed by a linear combination of the columns in range(A). Here is an example of this.

Let us be given the following:

$$A = \begin{bmatrix} -1 & 0 & 2 & -2 & -10 \\ -1 & 1 & 3 & -3 & -13 \\ 1 & 0 & -2 & 4 & 16 \\ 0 & 2 & -2 & 2 & -6 \end{bmatrix}$$

the RREf from of this matrix is:

$$rref(A) = \begin{bmatrix} 1 & 0 & -2 & 2 & 10 \\ 0 & 1 & 1 & -1 & 3 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We can then write the C and R matrix as follows:

$$C = \begin{bmatrix} 1 & 0 & -2 & 0 & 4 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \qquad R = \begin{bmatrix} -1 & 0 & -2 \\ -1 & 1 & -3 \\ 1 & 0 & 4 \\ 0 & 2 & 2 \end{bmatrix}$$

6.3 Orthogonal Matrices

A matrix $Q_{m\times n}$ is **orthogonal if it's columns are all perpendicular to each other.** A matrix $Q_{m\times n}$ is **orthonormal** if it the matrix itself is orthogonal and each column has a magnitude of 1. Orthonormal matrices are very crucial in the study of linear algebra; you really can't go anywhere without seeing them.

Note that there are some interesting properties about the Q matrix. We first note that

$$Q^TQ = I$$

which implies that

$$Q^T = Q^{-1}$$

Note that

$$QQ^T \neq I$$

6.4 Change of basis (And permutation/rotation matrices covered at the same time)

This was very fastly covered, by here is my way of explaining things. Basically, we can think of a basis as a coordinate system. With each basis system, we can construct a space where every element of that space is a combination of the basis vectors. However, we can trasnform these basis vectors to essentially change the coordinate system that we are working in. So take the elementary basis in 2 dimensions:

$$\left\{\begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}\right\}$$

We note that every vector in 2 dimensions can be constructed from these 2. Howeverm let's say we want to rotate every vector in this space. We can do so by rotating the basis vectors by multiplying the vectors by some matrix. So for example, let's say we want to multiply every basis vector by the following:

$$\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}$$

what we have here is the **rotation matrix**, where when we multiply a vector by this matrix, we inherently rotate that vector θ degrees counterclockwise. Thus, any vector that was in the trivial space, will be rotated as well. To boil it down simply: **A change of basis means that we are changing the axises themselves**.

7 Lecture 7

So a little rant about notation. Professor Nair uses this really weird notation for the dot product and the projection formula. I'm going to be using notation that I am more commonly familar with (and frankly, I think everyone is more comfortable with) So we can write the **Projection of a onto b**, denoted by

$$proj_b(a) = \frac{a\cdot b}{\|b\|^2}b$$

INSERT IMAGE HERE LATER We can also construct this by doing the following;

$$proj_b(a) = (\hat{b}\hat{b}^T)a$$

7.1 Properties of projection matrices

We can also construct vector spaces onto a vector space. We [9] begin with the properties of the projection matrices. A projection matrix onto the vector space v if

- $P_Vx \in V$
- $P_Vv = v \qquad \forall v \in V$

What this means in english is that P_V projects things into the vector space V, which means that if the vector is already in it, then the vector does not change. We can also note that

$$P_V(P_Vx) = P_Vx \iff P_V^2 = P_Vx \iff P_V = P_V$$

[9] for some reason

We can also call a matrix an orthogonal projection onto v if its is a projection onto V. We can denote this as

$$P_V x \perp x - P_V x$$

From here we can note that matheatically, we can write this as:

$$\begin{aligned}(x - P_V x)^T P_V x &= 0\\ x^T P_V v - x^T P_V^T P_V x &= 0\\ x^T (P_V v - P_V^T P_V) x &= 0\end{aligned}$$

which implies

$$\begin{aligned}P_V &= P_V^T P_V\\ P_V^T &= (P_V^T P_V)^T\\ P_V^T &= P_V^T P_V\\ P_V &= P_V^T\end{aligned}$$

So for P_V to be an orthogonal projector, the following must hold:

- $P_V^2 = P_V$
- $P_V = P_V^T$

However, to construct an orthogonal projector, we can use the following for certain cases.

If we are trying to consruct the projection onto the line v. We can calculate it by:

$$P_v = \hat{v}\hat{v}^T$$

where $\hat{v}$ denotes the unit vector. We also can do this for subspaces, where we can note that:

$$P_V = QQ_T$$

where Q is the matrix that has columns that are orthonormal to each that is based on the vectors in the basis. [10]

7.2 Application to the least squares problem

This is a similar arguememt to that was said earlier. I'll elaborate this a bit more in detail. Let's say that we are given a system $Ax = b$. However, assume that $b \notin range(A)$. Thus, there is no x that sastifies this equation. However, we can force this problem to be solveable once we force the b into the column space of A, or in other words minimizing the square error. Once we do that, we know that there exists a x^* such that $Ax^* = b_{range(A)}$. So let us think of it in this manner.

- We are given a system $Ax = b$ such that b is not in the range of A
- We can make an estimate, but there will be some error assocaited with it.
- Note that we can break up the vector b into two components: one that lies on range(A) and one that lies perpendicular to that, which implies that in lies in $null(A^T)$. We can denote this as:

  $$b = b_1 + b_2, b_1 \in range(A), b_2 \in null(A^T)$$

- We also should set up an error term:

  $$\|Ax - b\|^2$$

  and seek to minimize this error.
- Note that we can do the following:

  $$Ax - b = Ax - b_1 - b_2$$

- Thus, by the Pythagorean theorem, we know thar

  $$\|Ax - b_1\|^2 + \|b_2\|^2 = \|Ax - b\|^2$$

- We can note that if let x^* be such that $Ax^* - b_1$, we minimize the error.

[10] I skipped the proof because I doubt it is relevant.

- So when we do this, we then know that $Ax^* - b_1 = 0 \iff Ax^* = b_1$.
- Then, we can do the following:
$$Ax^* - b_1 - b_2 = Ax^* - b$$
But notice that we cooked the problem such that $Ax^* - b_1 = 0$, once we substitute that, we get:
$$Ax^* - b = -b_2$$
This implies that $Ax^* - b$ is in $null(A^T)$, which implies that
$$A^T(Ax^* - b) = 0$$
- Solving this yields $A^T(Ax - b) = 0 \iff A^TAx^* = A^Tb$.

8 Lecture 8

Instead of me typing all this out, look at these two links, they do a much better job of explaining it than the professor:

- Gram Schmidt
- QR Decomposition - Using the Gram-Schmidt Process

We can note that we can use the QR Decomposition approach, we can do the following:

$$\begin{aligned} Ax^* &= b_{range(A)} \\ QRx^* &= QQ^Tb \\ Q^TQx^* &= Q^TQQ^Tb \\ Rx^* &= Q^Tb \end{aligned}$$

So in general, when we want to do a least squares problem, there are two approaches to it:

- Solve $A^TAx^* = A^Tb$
- Solve using QR

9 Lecture 6

9.1 Linear Transformation

A linear transformation is to show transformation T sastifies the linearity condition:

$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$$

where x and y are vectors and α and β. This transformation can be also defined in terms of a matrix such that:

$$T(x) = Ax$$

meaning that T acts on any vector x by multiplying it with the matrix A. A present a fixed matrix defines the transformation.

9.2 Proof of Linearity

We begin with the value $T(\alpha x + \beta y)$. Note that we can rearrange this in the following:

$$\begin{aligned} T(\alpha x + \beta y) &= A(\alpha x + \beta y) \\ &= A(\alpha x) + A(\beta y) \\ &= \alpha Ax + \beta Ay \\ &= \alpha T(x) + \beta T(y) \end{aligned}$$

And here are some examples of linear transformation:

Identity matrix

This is the matrix:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

This means that the vector will be returned to itself.

Diagonal Matrix

Another example of a linear transformation is the **Diagonal Matrix**, which is the following:

$$\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$$

where $\alpha, \beta \in \mathbb{R}$. This matrix just takes vectors and stretchs it's components respectively.

Rotation matrix

We can rotate the basis vectors. It is in the form:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

we rotate the basis vectors counterclockwise by θ degrees.

Orthonormal transformation

These are linear transformations that rotate the vectors, preserve the angles between them, and preserves the lengths of them. In other words:

$$\|Qx\| = \|x\|$$

9.3 Determinant

A determinant, notes as $\det(A)$ is a volume of a parallelopiped. This sides of the parallelipied are represented by the columns of the matrix A. Note that

- Positive determinant indicates preservation of oreintation
- Negative determinant indicates reversing of oreintation

and these are some properties of the determinant are:

- $\det(A)$ can be positive and Negative
- $\det(A)$ is only defined for squared matrices
- If A is square and non-invertible if $\det(A) = 0$
- If $\det(A) = 0$, then A^{-1} does not exist
- $\det(AB) = \det(A)\det(B)$

We can use the last property to note that:

$$\det(Q) = \pm 1$$

And thus

$$\det(Q) = \pm\det(R) = \Pi_{j=1}^{n} r_{jj}$$

Eigenvalue

An Eigenvalue is a scalar that provides insight into how a matrix or transformation acts on specific vectors. These vectors, known as eigenvectors, remain in the same directioon after transformation, though they may be scaled by the eigenvalue. So, an eigenvector is something in the form:

$$Av = \lambda v$$

where A denotes the matrix representing the linear transformation, v is the eigenvector associated with the corresponding to the eigenvalue λ, and λ is thh eigenvalue associated with v. To find the vector assigned to each eigenvaue, note the following proof:

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

Note that v should be in $null(A - \lambda I)$, which implies that the matrix cannot be invertible. Thus, λ is an eigenvalue of A if and only if $A - \lambda I$ is singular, or:

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

and the spectrum of A is the set of all eigenvalues of A denoted by $\sigma(A)$ or spec(A).

10 Lecture 10

This is just more eigenvector stuff. This is a general method of finding eigenvectors:

1. Form the characteritsic polynomial:
$$P_A(A) = \det(A - \lambda I)$$
2. Find the roots of P_A
3. For each λ_i, solve $(A - \lambda_i I)v_i = 0$ for all eigenvectors $v \neq 0$

and note that these eigenvalues essentially scale eigenvectors. We can also be concerned about the degree of the characteritsic polynomial. If A is $n \times n$, the characteritsic polynomial $P_A(A)$ is degree n. This degree n means that there can be up to n eigenvalues, but not all of them may be distinct. Note that if all the eigenvalues are distinct, that means that are the corresponding eigenvectors must e linearly independent.

10.1 Diagonalization of a matrix A

If the matrix A has a complete set of linearly independent eigenvectors, it can be decomposed (diagnolized) also

$$A = V \Lambda V^{-1}$$

where

- V is a matrix whose column are eigenvectors of A
- Λ is the diagnoal matrix that contains the eigenvalues of A along its diagnoal in the order they are in V
- V^{-1} is the inverse of V

So formally: Given an $n \times n$ matrix A, suppose A has n distinct linearly independent eigenvectors $\{v_1, v_2, \ldots, v_n\}$ corresponding to eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$,

$$V = [v_1, v_2, \ldots, v_n]$$

an $n \times n$ matrix whose columns are the eignevectors of A and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Note that since all eigenvectors are linearly independent to each other, we know that V is a basis for $\mathbb{R}^n$, which implies that any vector $x \in \mathbb{R}^n$ can be decomposed into a linear combination of eigenvectors. Note that the folowing is equivalent:

$$Ax = V \Lambda V^{-1} x$$

We can also solve the system $Ax = b$ using diagnoalization. Note that we can do the operations[11]:

$$\begin{aligned} Ax &= b \\ V\Lambda V^{-1}x &= b \\ V^{-1}V\Lambda V^{-1}x &= V^{-1}b \\ \Lambda V^{-1}x &= V^{-1}b \\ \Lambda^{-1}\Lambda V^{-1}x &= \Lambda^{-1}V^{-1}b \\ V^{-1}x &= \Lambda^{-1}V^{-1}b \\ VV^{-1}x &= V\Lambda^{-1}V^{-1}b \\ x &= V\Lambda^{-1}V^{-1}b \end{aligned}$$

One thing we can note about the eigenvalue decomposition is that the following proof of the idea that:

$$A^n = V\Lambda^n V^{-1}$$

Proof. Base case: Consider the case AA. Note the following:

$$\begin{aligned} AA &= V\Lambda V^{-1}V\Lambda V^{-1} \\ A^2 &= V\Lambda^2 V^{-1} \end{aligned}$$

we have verified the base case. Now inductively assume that $A^n = V\Lambda^n V^{-1}$ is true. Then, we can perform the following operations:

$$\begin{aligned} A^n \cdot A &= V\Lambda^n V^{-1} \cdot A \\ A^{n+1} &= V\Lambda^n V^{-1}V\Lambda V^{-1} \\ A^{n+1} &= V\Lambda^{n+1} V^{-1} \end{aligned}$$

thus, we have proven inductively that $A^n = V\Lambda^n V^{-1}$ is true. □

11 Lecture 11

We now expand the ideas of diagnoalizable matrices. We begin with the notion that if A is diagnoalizable, then A can be diagnoalized by a change of coordinates associated to the eigenvalue.

11.1 Similarity Tranform

We define a similarlity tranform of $T(x) = Ax$ as a mapping $T(x) = Ax \to \hat{T}(y) = \hat{A}y$. We are then given these special characteritsics about the function itself. Note that $\hat{A}$ denotes the tranformational matrix associated with the new basis and B is the matrix whose columns are the eigenvectors A. So we can begin with the following idea. Let $x \in R^n$. Now, assume that $x = By$, then we know that:

$$T(x) = B\hat{T}(y)$$

which simplifes down to:

$$Ax = B\hat{A}y$$

now we can see that the following are equivalent:

$$Ax = B\hat{A}y = B\hat{A}B^{-1}x$$

Which simplifies down to:

$$A = B\hat{A}B^{-1}$$

Or equivalently

$$\hat{A} = B^{-1}AB$$

Thus we can define a similarity transform of A with respect to basis B as

$$\hat{A} = B^{-1}AB$$

and if there exsist a B that makes the above operation true, then we say that A and $\hat{A}$ are similar. Or $A \sim \hat{A}$ Notye that we know that when we perform an eigenvalue decomposition, we know that $A \sim \Sigma$.

[11] There is a complicated method that was brought up during lecture, but I think this is a lot more straightforward

11.2 Geometric Interpretation of the Eigenvalue decomposition

Now we look at the geometric interpretation of the eigenvalue decomposition [12]. Given the equation $Ax = V\Lambda V^{-1}$,

- $V^{-1}x$ changes x to be on the eigenbasis.
- Λ stretchs, relects, or scales the vector
- V changes x onto the original coordinate system

Howver, *not all matrices can be diagnoalizable*, however, most applications have matrices that can be decomposed.

11.3 Algebraic and Geometric Multiplicity

The **Algebraic** multipicity of λ corresponds to the number of times the certain λ appears as the root of the characteritsic polynomial. The **Geometric multipicity** of λ is the dimension of $null(A - \lambda I)$. Here are some facts associated with this:

- Geometric Multiplicity $\leq$ Algebraic Multiplicity
- If there exists a λ such that geometric multipicity $<$ algebraic multipicity, then the matrix in question *cannot* be diagnoalizable.

Note that there are some shortcuts we can use to see if a matrix is diagnoalizable or not. A is diagnoalizable if any of the folllowing are true:

- A is normal: $A^TA = AA^T$
- A is symmetric: $A^T = A$
- A is skeer symmetric $A^T = -A$

12 Lecture 12

Now we expand to harder topics.

12.1 Power method

Before we begin, here's just a basic understanding on what this method does. This method essentially helps find the largest eigenvalue. We already know that:

$$A^K = V\Lambda^K V^{-1}$$

So we know that:

$$A^K x = V\Lambda^K V^{-1}x$$

and let $y = V^{-1}x$, where y are the coordinates of x in the eigenbasis. Thus, we know that

$$\begin{aligned} A^K x &= V\Lambda^K V^{-1}x \\ &= V\Lambda^K y \\ &= \sum_{j=1}^{n} (\lambda_j^K y_j) v_j \end{aligned}$$

and note that when we repeat this summation, this summation will be dominated by the largest eigenvalue. However, note something about the values of eigenvalues. Note that when $k \to \infty$, if $|\lambda| > 1$, $\lambda \to \infty$ and if $|\lambda| < 1$, then $\lambda \to 0$. To prevent this issue, we want to normalize the vector every time. So, to put it simply, given $W = Ax_k$,

$$x_{k+1} = \frac{W}{\|W\|}$$

that way, we prevent the vector diverging or converging and see exactly which eigenvalue dominates.

[12] or spectral decomposition, whatever fits your fancy

12.2 Singular Value Decomposition

To account for the limitations in the eigenvalue decompoistion, we have this method. A warning before we continue, ***This method take FOREVER to do by hand. This will definetly be tested in some capacity.***. Note that this decomposition can be done for any matrix. So given a matrix $A_{m\times n}$, we get that:

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$$

and there are some important observations to note. U and V are orthonormal, which means that all orthonormal properties hold. There are also some special names we associate with the columns of these matrices.

- columns of U are called the left singular vectors of A
- columns of V are called the right singular vectors of A

Note that Σ we almost diagnoal with entires with entires $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ arranged along the diagonal. Note there are some interesting behaviors of the SVD that are associated with $m > n$ and $m < n$. If $m > n$, the matrix Σ will have the following structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \text{where} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

If $m < n$, the matrix Σ will have the following structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_m & \cdots & 0 \end{bmatrix}$$

12.3 SVD and the relation to the fundamental subspaces

We know by the fundamental theorem of linear algebra that

$$range(A) \oplus null(A^T) = \mathbb{R}^m$$

and

$$range(A^T) \oplus null(A) = \mathbb{R}^n$$

Thus, given the SVD from above we can see that:

- $\begin{bmatrix} | & | & \dots & | \\ U_1 & U_2 & \dots & U_r \\ | & | & \dots & | \end{bmatrix}$ forms an orthonormal basis for the range of A
- $\begin{bmatrix} | & | & \dots & | \\ U_r & U_{r+1} & \dots & U_m \\ | & | & \dots & | \end{bmatrix}$ forms an orthonormal basis for $null(A^T)$
- $\begin{bmatrix} | & | & \dots & | \\ V_1 & V_2 & \dots & V_r \\ | & | & \dots & | \end{bmatrix}$ forms an orthonornal basis for range of A^T
- $\begin{bmatrix} | & | & \dots & | \\ V_r & V_{r+1} & \dots & V_m \\ | & | & \dots & | \end{bmatrix}$ forms an orthonormal basis for $null(A)$

The proofs of these will be given in the proofs section.

12.4 Pseudo-Inverse

Given a matrix A with the standard SVD decompoistion, we see that the pseudo-inverse is the following:

$$A^t = V\Sigma^{-1}U^T$$

and $A^{-1} = A^t$ if A is invertible.

13 Lecture 13

We are now intereted in the least squares problem but in the context of finding the minimum norm solutions. We are interested in solving the following optimization problem:

$$\min_x \|Ax - b\|^2$$

Note that we already know that from the proof of the least squares solutions in regards to the fundamental subspaces approach that $Ax^* = b_{\|\text{range}(A)}$. is the solution that minimizes the quantity demanded. Now, note that:

$$Q_A Q_A^T b = b_{\|\text{range}(A)}$$

where Q_A is the orthonormal basis of the range of A. Thus, given that $A = U\Sigma V^T$ from the SVD, we know that:

$$U\Sigma V^T x^* = Q_A Q_A^T b$$

However, let us multiply both sides by U^T, thus, we get the following:

$$\begin{aligned} U\Sigma V^T x^* &= Q_A Q_A^T b \\ U^T U\Sigma V^T x^* &= U^T Q_A Q_A^T b \\ \Sigma V^T x^* &= U^T Q_A Q_A^T b \end{aligned}$$

Now let $V^T x^* = y$. Additionally, note that:

$$U^T Q_A = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \ddots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where the 1 is on the diagnoal r times. Thus, we can see that:

$$U^T Q_A Q_A^T b = \begin{bmatrix} U_1^T b \\ U_2^T \\ \vdots \\ U_r^T \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

So we are left with:

$$\Sigma y = \begin{bmatrix} U_1^T b \\ U_2^T b \\ \vdots \\ U_r^T b \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

expanding, Σy yields:

$$\Sigma y = \begin{bmatrix} \sigma_1 y_1 \\ \sigma_2 y_2 \\ \vdots \\ \sigma_r y_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} U_1^T b \\ U_2^T b \\ \vdots \\ U_r^T b \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus, for any i, we can see that

$$\sigma_i y_i = U_i^T b$$
$$y_i = \frac{U_i^T b}{\sigma_i}$$

We can denote the above quantity as a piecewise function:

$$y_i \begin{cases} \frac{U_i^T}{\sigma_i} & i \leq r \\ 0 & i > r \end{cases}$$

The reason why this provides the minimum norm is because when we decompose the vector magnitude as follows:

$$\|y\|^2 = \sum_{i=1}^{r} y_i^2 + \sum_{i>r} y_i^2$$

Note that for all $i > r$, the norm is 0. Thus, we know that $\sum_{i=1}^{r} y_i^2$ is the only contributer to the norm of y. Therefore, this is the minimum norm solution.

13.1 Relation to the Psuedoinverse

The **Moore - Penrose Psuedoinverse** has an important relationship to the proof above. Note that the relationship $y = V^T x^* \iff x^* = V^T y$. [13]. We can see that;

$$x^* = V \begin{bmatrix} \frac{U_1^T b}{\sigma_1} \\ \frac{U_2^T b}{\sigma_2} \\ \vdots \\ \frac{U_r^T b}{\sigma_r} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$x^* = V\Sigma^- 1 U^T$$

Thus,

$$x^* = A^t b$$

provides the minimum norm solution.

13.2 Special cases of least squares minimization

We have two cases we need to worry about: *Overdetermined System* amd *Underdetermined System.*

13.2.1 Overdetermined System

This is when $m \geq n$ and $rank(A) = n$. This means there are more equations than unknowns.We also have know that $A^T A$ is $n \times n$ and invertible. Thus, we can use the normal equation apprpoach that we derived earlier,

$$A^T A x = A^T b \quad x^* = (A^T A)^{-1} A^T b$$

both ways are good ways of approaching this problem.

13.2.2 Underdetermined System

The conditions here are that $m \leq n$ and $rank(A) = m$. We can observe that since $rank(A) = m$, the range of A is $\mathbb{R}^m$, meaning that b must lie within this range for a solution to exist. Thus we should use the pseudoinverse approach to this. Thus,,

$$x^* = A^t b$$

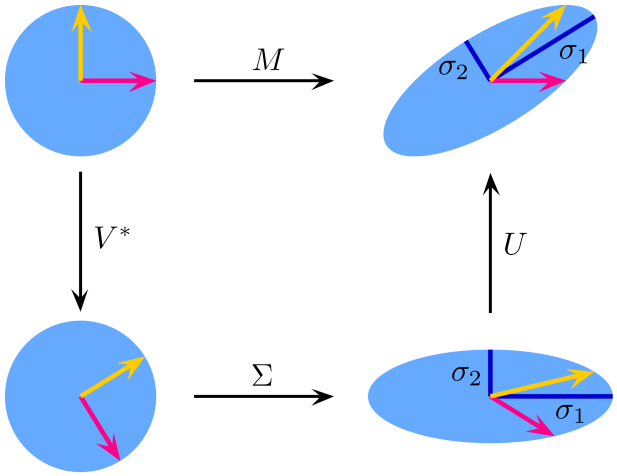
[13] Math shorthand for if and only if

14 Lecture 14

We now move to SVD decomposition interpreted geometrically. We know that $A = U\Sigma V^T$ and $Ax = U\Sigma V^T x$. So we can use the following the diagram to see how we can get to each quantity in $Ax = U\Sigma V^T x$.

$$\begin{array}{lll} x & A \rightarrow & Ax \\ V^T \downarrow & & \uparrow U \\ V^T x & \Sigma \rightarrow & \Sigma V^T x \end{array}$$

Geometrically, given a 2×2 matrix A with the trivial basis, we can see that the following is a geoemtric interpretation.



$$M = U \cdot \Sigma \cdot V^*$$

I ripped this off Wikipedia, so courtesy to them. Let me explain what each step does.

- When we multiply the basis vectors by U, a orthonormal matrix, we rotate the basis vectors in some fashion.
- When we multiply the basis vectors by Σ, we stretch the ones repsectively.
- Finally, the V^T is still a orthonormal matrix, so we rotate the basis vectors again.

We look at the basis vectors, as these create any vector we want.

14.1 Frobenius Norm

This is just the idea of expanding vector norms to matrices. Intuitively, the Frobenius norm is just every entry squared, added together, and then square rooted. Or if we wanted to be fancy, it is

$$\|A\|_F = \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}$$

Or

$$\|A\|_F = \max_x \left(\frac{\|Ax\|}{\|x\|} \right)$$

a proof of this will be in the proof section.

15 Lecture 15

We explore the applications of the SVD now.

15.0.1 Stability and Conditioning

Given a guess to a solution:

$$Ax^* = b^* + \epsilon$$

where ϵ denotes the error. Note the following cases:

- if $\epsilon = 0$, this means that $Ax = b^*$, which means that $b \in range(A)$ and solution is $x^* + z$ where $z \in null(A)$
- if $\epsilon \neq 0$, we have a different question to analyze. Assume that x_o is a solution. We are now interested in $\|x_o - x\|$

We have two cases

- If $\|x_o - x\|$ is small, then we have a stable or well conditioned system
- If $\|x_o - x\|$ is large, then we have an unstable of ill conditioned system.

However, if $y \notin range(A)$ we know that we can now rewrite this in a least squares format. Now we are interested in what happens when we change y a little bit. When we do this and try to go back to our original solution, there is no gaurantee that we can d o so. We could be close, but there could be some sort of error [14]. Let us look at the minimum norm solution for the least squares problem. Note that the minimum norm solution is defined as :

$$x_o = A^t y = V\Sigma^{-1}U^T y$$

and σ_j are small, then $\frac{1}{\sqrt{\sigma)j}}$ will be big and vice versa. Note that we can conduct the same kind of ϵ analysis as before.

- If $\epsilon = 0$, this means that $y = b^*$ so $x_o = x^*$.
- If $\epsilon \neq 0$, this means that we can do the following:

$$\begin{aligned} x_o &= A^t(b^* + \epsilon) \\ &= A^t b^* + A^t\epsilon \\ &= x^* + A^t\epsilon \end{aligned}$$

which implies that when we solve for the solution that solves the least squares problem, we get the pseudo-inverse.

16 Lecture 16

This is just lecture 15, just more organized.

16.1 Stability analysis

We are interested in the solution to $Ax = b$. Note that if $b \notin range(A)$, this means that $Ax = b$ has no solution. Thus, we turn the least squares problem. Now, we can use an x^* to force the equation to be solved, or similarly $Ax^* = b^*$. However, suppose that there is a b_1 such that $b_1 = b^* + \epsilon$, where ϵ is a vector with a small magnitude. We can see that there are 2 cases associated with this, $\epsilon = 0$ and $\epsilon \neq 0$.

- If $\epsilon = 0$, we see that $x_o = A^t b^*$
- If $\epsilon \neq 0$, we can see that

$$\begin{aligned} x_o &= A^t(b^* + \epsilon) \\ x_o &= A^t b^* + A^t\epsilon \\ x_o - x^* &= A^t\epsilon \end{aligned}$$

where $A^t\epsilon$ is the error associated with the approximation. We can also see that A^t has the singular values $\{\sigma_r^{-1}, \sigma_{r-1}^{-1}, \ldots, \sigma_1^{-1}\}$. So intuively, if σ_r are large, so $|\sigma_r| > 1$, the error will be small. If $|\sigma_r| < 1$, we can see that the error will be amplified. Thus, we should be interested in the maximum relative error, so we can see how severe our error would be.

[14] This is probably what numerical stability

16.2 Relative Error Proofs

We know that error is defined as $\|x_o - x^*\| = \|A^t\epsilon\|$, and thus relative error should be

$$\frac{\|x_o - x^*\|}{\|x^*\|}$$

and thus we should be interested in maximumizing this quantity. Note that

$$\frac{\|x_o - x^*\|}{\|x^*\|} = \frac{\|A^t\epsilon\|}{\|A^t b^*\|}$$

and note that:

$$\frac{\|A^t\epsilon\|}{\|A^t b^*\|} \implies \frac{\frac{\|A^t\epsilon\|}{\|\epsilon\|}}{\frac{\|A^t b^*\|}{\|b^*\|}}$$

Thus, we are interesting this quantity, or more specifically:

$$\max_{b,\epsilon}\left\{\frac{\frac{\|A^t\epsilon\|}{\|\epsilon\|}}{\frac{\|A^t b^*\|}{\|b^*\|}}\right\}$$

But note that since the numerator are not related, we can see that we can seperate this into two seperate quantities, or more specifically:

$$\max_{b,\epsilon}\left\{\frac{\frac{\|A^t\epsilon\|}{\|\epsilon\|}}{\frac{\|A^t b^*\|}{\|b^*\|}}\right\} = \max_{b}\left\{\frac{\|b^*\|}{\|A^t b^*\|}\right\}\max_{\epsilon}\left\{\frac{\|A^t\epsilon\|}{\|\epsilon\|}\right\} = \max_{x}\left\{\frac{\|Ax^*\|}{\|x^*\|}\right\}\max_{\epsilon}\left\{\frac{\|A^t\epsilon\|}{\|\epsilon\|}\right\} = \frac{\sigma_1}{\sigma_r}$$

We can also note that a well conditioned system has a small condition number and an ill conditioned system has a large condition number.

16.3 Low Rank Approximations

We should know by that $A_{m\times m} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$ by the SVD. When we view this from the outer product definition of matrix multiplication, we see that

$$A = \sum_{j=1}^{n}\sigma_j u_j v_j^T$$

where we are indexing based on column for U, row for V^T, and diagnol entry in Σ. Thus, essentially we have a summation of Rank 1 matrices. And since we are doing this summation, we cann actually cut this summation at some index k, like the following:

$$A^k = \sum_{j=1}^{k}\sigma_j u_j v_j^T$$

And the whole idea of the Low Rank Approximation is that we have this matrix A^k that is **close enough** to the original matrix A. But, how do we test the similarlity to this? We can use the Frobenius Norm to see how similar these matrices are to each other, or more specifically:

$$\|A - A^k\|_{Fro}$$

Note that we have already proven the notion that

$$\|A\|_{fro}^2 = \sum_{j=1}^{r}\sigma_j^2$$

And also

$$A - A^k = \sum_{j=1}^{n}\sigma_j u_j v_j^T - \sum_{j=1}^{k}\sigma_j u_j v_j^T = \sum_{j=k+1}^{n}\sigma_j u_j v_j^T$$

Thus, we can see that:

$$\|A - A^k\|_{Fro}^2 = \sum_{j=k+1}^{r}\sigma_j^2$$

17 Lecture 17

We define A as an $m \times n$ matrix and $M = A^T A$. Now we introduce the Raleigh Quotient:

$$\frac{x^T H x}{\|x\|^2}$$

where this helps determine the behavior of random matrices. Consider the matrix A_ω, whose entries are randomized. Then, when we multiply this matrix by x, we get the system $A_\omega x = y_\omega$. We are now ineterested in measuring $\mathbb{E}\|y_\omega\|^2$, which is the variance of $\mathbb{E}y_\omega = 0$. Note that the following:

$$\mathbb{E}\|A_\omega x\|^2 = \mathbb{E}(x^T A_\omega^T A_\omega x) = x^T \mathbb{E}(A\omega^T A_\omega)x$$

Now we see that

$$x^T \mathbb{E}(A\omega^T A_\omega)x = x^T M x = cov(y)$$

Normalized this to make sure that represents probabilities means that we are interested in the quantity

$$\frac{x^T M x}{\|x\|^2}$$

and we want to maximize this value because we want to see what inputs maximize the variance of y.

Proof. Note that

$$\frac{x^T M x}{\|x\|^2} = \frac{x^T A^T A x}{\|x\|^2} = \frac{x^T V \Sigma^T \Sigma V^T x}{\|x\|^2} = \frac{x^T V \Sigma^2 V^T x}{\|x\|^2}$$

Now we see that $Mx = V\Sigma^2 V^T x$. When we analyze this from the outer product view of matrix vector multiplication, we see that

$$\begin{aligned} Mx &= V\Sigma^T \Sigma x \\ &= \sum_{j=1}^{r} v_j \sigma_j^2 v_j^T x \\ &= \sum_{j=1}^{r} \sigma_j^2 v_j v_j^T x \end{aligned}$$

However note that any vector x can be represneted as any combination of it's basis vectors, hence

$$x = \sum_{j=1}^{n} y_j \alpha_j$$

where α_j deontes a scalar value. Thus, we see that

$$\begin{aligned} Mx &= \sum_{j=1}^{r} \sigma_j^2 v_j v_j^T x \\ &= \sum_{j=1}^{r} \sigma_j^2 v_j v_j^T v_j \alpha_j \\ &= \sum_{j=1}^{r} \alpha_j \sigma_j^2 v_j \end{aligned}$$

Thus, we see by a similar logic that:

$$x^T M x = \sum_{j=1}^{r} \alpha_j^2 \sigma_j^2$$

Since $\|x\|^2 = \sum_{j=1}^{n} \alpha_j^2$, we are interesting in maximizing

$$\max_{\alpha} \left(\frac{\sum_{j=1}^{r} \alpha_j^2 \sigma_j^2}{\sum_{j=1}^{r} \alpha_j^2} \right)$$

and if we let $\hat{\alpha}$ be a vector that contains all α values, we can nomralize this such that $\|\hat{\alpha}\| = 1$, thus:

$$\max_{\|\hat{\alpha}\|=1} \sum_{j=1}^{r} \alpha_j^2 \sigma_j^2$$

and now, if we wish to maximize this, we want to put more weight on the largest singular value. Hence,

$$\max_{\|\hat{\alpha}\|=1} \sum_{j=1}^{r} \alpha_j^2 \sigma_j^2 = \sigma_r^2$$

□

18 Proofs

A little section that complies all the proofs in this class. Highly recommend to study these. Sections with an $*$ denote proofs that were tested on the midterm. *These are how I did the proofs, reading over it is not enough to understand it. Please write it out, and try to understand why each step works. I try to handholding every step in these proofs to help with that*

18.1 Least Squares Solution (calculus)*

Proof. When we are interested in proving the least squares, we are interested in solving the following optimization problem:

$$\min \|Ax - b\|^2$$

Let us denote this quantity as E^2. Thus, we are interested in the equation

$$E^2 = \|Ax - b\|^2$$

But note that:

$$\|Ax - b\|^2 = (Ax - b)^T(Ax - b)$$

Thus, we can do the following:

$$\|Ax - b\|^2 = (Ax - b)^T(Ax - b) = (b^T - x^T A^T)(Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$

since we are optimizing based on x, we must take the partial deriative with respect with to x, hence we must do the following.

$$\frac{\partial}{\partial x}(x^T A^T Ax - 2b^T Ax + b^T b) = 2A^T Ax - 2A^T b$$

Since we are optimizing, we want the above quantity to equal 0. Thus, we solve the following system of equations.

$$\begin{aligned} 2A^T Ax - 2A^T b &= 0 \\ 2A^T Ax &= 2A^T b \\ A^T Ax &= A^T b \end{aligned}$$

□

18.2 Least Squares solution (Fundamental Subspaces)*

This[15] is a similar arguememt to that was said earlier. I'll elaborate this a bit more in detail. Let's say that we are given a system $Ax = b$. However, assume that $b \notin range(A)$. Thus, there is no x that sastifies this equation. However, we can force this problem to be solveable once we force the b into the column space of A, or in other words minimizing the square error. Once we do that, we know that there exists a x^* such that $Ax^* = b_{range(A)}$. So let us think of it in this manner.

- We are given a system $Ax = b$ such that b is not in the range of A
- We can make an estimate, but there will be some error assocaited with it.
- Note that we can break up the vector b into two components: one that lies on range(A) and one that lies perpendicular to that, which implies that in lies in $null(A^T)$. We can denote this as:

  $$b = b_1 + b_2, b_1 \in range(A), b_2 \in null(A^T)$$

- We also should set up an error term:

  $$\|Ax - b\|^2$$

  and seek to minimize this error.
- Note that we can do the following:

  $$Ax - b = Ax - b_1 - b_2$$

- Thus, by the Pythagorean theorem, we know thar

  $$\|Ax - b_1\|^2 + \|b_2\|^2 = \|Ax - b\|^2$$

[15] I copy pasted the proof from above lmao

- We can note that if let x^* be such that $Ax^* - b_1$, we minimize the error.
- So when we do this, we then know that $Ax^* - b_1 = 0 \iff Ax^* = b_1$.
- Then, we can do the following:
$$Ax^* - b_1 - b_2 = Ax^* - b$$
But notice that we cooked the problem such that $Ax^* - b_1 = 0$, once we substitute that, we get:
$$Ax^* - b = -b_2$$
This implies that $Ax^* - b$ is in $null(A^T)$, which implies that
$$A^T(Ax^* - b) = 0$$
- Solving this yields $A^T(Ax - b) = 0 \iff A^TAx^* = A^Tb$.

18.3 Repeated multiplication of A using Eigendecomposition

One thing we can note about the eigenvalue decomposition is that the following proof of the idea that:

$$A^n = V\Lambda^nV^{-1}$$

Proof. Base case: Consider the case AA. Note the following:

$$AA = V\Lambda V^{-1}V\Lambda V^{-1}$$
$$A^2 = V\Lambda^2V^{-1}$$

we have verified the base case. Now inductively assume that $A^n = V\Lambda^nV^{-1}$ is true. Then, we can perform the following operations:

$$A^n \cdot A = V\Lambda^nV^{-1} \cdot A$$
$$A^{n+1} = V\Lambda^nV^{-1}V\Lambda V^{-1}$$
$$A^{n+1} = V\Lambda^{n+1}V^{-1}$$

thus, we have proven inductively that $A^n = V\Lambda^nV^{-1}$ is true. □

18.4 Finding the SVD by hand

Note that:

$$A = U\Sigma V^T \quad A^T = V\Sigma U^T$$

and thus:

$$AA^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma^2U^T$$

and

$$A^TA = (A^T = V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2V^T$$

Note that the above form are in the form of the eigendecomposition of the respective matrices getting multiplied together. So all you have to do is solve the eigendecomposition twice to get your answer. (It takes forever)

18.5 Fundamental Subspaces and the relation to the SVDs

We are given the SVD:

$$A = U\Sigma V$$

where Σ is defined in the following way: If $m > n$, the matrix Σ will have the following structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \text{where} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

If $m < n$, the matrix Σ will have the following structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_r & \cdots & 0 \end{bmatrix}$$

the key idea here is to note that if $r < m$ or $r < n$, there will be a "lack" of singular values. So we begin with the proof that Q_{1-r} represents the range of A.

18.5.1 range of (A)

We know that:

$$A = U\Sigma V^T$$

thus we can do the following. Let x be any vector you want (in math terms, any $x \in \mathbb{R}^n$):

$$A = U\Sigma V^T$$
$$Ax = U\Sigma V^T$$

Let $y = \Sigma V^T$. Note that in this form, we get a vector in the form:

$$y = \begin{bmatrix} \sigma_1 v_1 \\ \sigma_2 v_2 \\ \vdots \\ \sigma_r v_r \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Where each $\sigma_i v_i$ represents a vector. When we look at this from this perspective of matrx-matrix maultiplication, we get that:

$$Ax = Uy = U_1(\sigma_1 v_1) + U_2(\sigma_2 v_2) + \ldots U_r(\sigma_r v_r)$$

where U_i denotes the respective column in U Or if you prefer this notation:

$$Ax = U \begin{bmatrix} \Sigma_r y_r \\ 0 \end{bmatrix}$$

Thus, we can construct any element in the range of A by a linear combination of the first r columns of U. Therefore, by construction, we know that all the columns in U are orthogonal to each other, thus meaning that U_{1-r} is the basis of range(A).

18.5.2 Range of A transpose

We know that

$$A^T = V\Sigma U^T$$

We do this to analyze the column space of A. Let x be any vector you want (in math terms, any $x \in \mathbb{R}^n$). Let $y = \Sigma U^T$. By a similar logic (or without a loss of generality) to the proof of the Range of A [16], we find that V_{1-r} are the columns that make up the basis of range(A^T).

18.5.3 Finding Null(A)

This is my own proof. Not Professor Nair's proof. It's a similar structure to the **15.5.1**, so remembering this would be easier. Refer to your lecture notes for the original proof. This one is a bit more intutive. We start by letting any vector $z \in \mathbb{R}^n$ and doing:

$$Az = 0 \iff U\Sigma V^T z = 0$$

[16] This depends on prof. This basically means in english "The proof is the same with some values switched around and as a Mathematician I am too lazy to rewrite the proof". Ask your professor before doing this. I am clearly invoking the lazy part here, but seriously the proof is identical. Do it yourself and see.

Multiply both sides by U^T to get:

$$\Sigma V^T z = 0$$

However if $m > n$, the matrix Σ will have the following structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \text{where} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

If $m < n$, the matrix Σ will have the following structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_r & \cdots & 0 \end{bmatrix}$$

point being that, to sastify the null space conditions, we are only interested in values $r+1 < n$ given that $r < n$, so when we multiply the matrix V^T by Σ, we will get the zero vector. As all these singular values with be 0. So, we can deduce that $V_{r+1\to n}$ is the orthonormal basis for $null(A)$. With similar logic, we can extend this to $U_{r+1\to m}$ beinng the orthonromal basis for $null(A^T)$.

18.5.4 Using the Fundamental Theorem of Linear Algebra to prove the null spaces

By the fundamental theorem of linear algebra, we know that $null(A^T)$ must be orthogonal to $range(A)$ and $null(A)$ must be orthogonal to $range(A^T)$. Thus, by construction, we know that set of column vectors U_{r+1-m} must be orthogonal to U_{1-r}. The same logic applies to $V_{r+1-\to n}$ must be orthogo nal to $V_{1\to r}$. Therefore, $U_{r+1\to m}$ is the orthonormal basis for $null(A^T)$ and $V_{r+1\to n}$ is the orthonormal basis for $null(A)$ respectively.

18.6 Minimum norm applied to LSQ

See Lecture 13

18.7 Proof that Qx = x

We see that

$$\begin{aligned} \|Qx\| &= \sqrt{(Qx)^T(Qx)} \\ &= \sqrt{x^T Q^T Q x} \\ &= \sqrt{x^T x} \\ &= \|x\| \end{aligned}$$

18.8 Proof of Ax over x

Proof. We know that

$$\|A\|_{fro} = \max_{x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$$

By the singular value decomposition, we see that $A = U\Sigma V^T$. Thus,

$$\|Ax\| = \|U\Sigma V^T x\| = \|\Sigma V^T x\|$$

Let $y = V^T x$. Therefore,

$$\frac{\|\Sigma V^T x\|}{\|x\|} = \frac{\|\Sigma y\|}{\|Vy\|} = \frac{\|\Sigma y\|}{\|y\|}$$

And when we expand these terms out, we get

$$\frac{\left(\sum_{i=1}^r \sigma_i^2 y_i^2\right)^{\frac{1}{2}}}{\left(\sum_{i=1}^2 y_i^2\right)^{\frac{1}{2}}}$$

Let $y = [1, 0, 0, 0, 0 \ldots, 0]$ we see that we have no proven that $\|A\|_{fro} = \sigma_1$ □

18.9 Proof that Trace(AB) = Trace (BA)

Proof.

$$Trace(AB) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ji}b_j i = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ij}a_{ji} = Trace(BA)$$

□

18.10 Prof that frobenius norm is really just the singular values squared

Proof. We want to show that

$$\|A\|_{fro}^2 = \sum_{j=1}^{r} \sigma_j^2$$

By the definition of Frobenius Norm, we see that

$$\|A\|_{fro}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ji}^T A_{ij} = \sum_{i=1}^{n} (A^T A)_i i = Trace(A^T A)$$

Therefore, using thne proff from above, we see that

$$\begin{aligned}
\|A\|_{Fro}^2 &= Trace(A^T A) \\
&= Trace((U\Sigma V^T)^T (U\Sigma V^T)) \\
&= Trace(V\Sigma^T U^T U\Sigma V^T) \\
&= Trace(V\Sigma^T \Sigma V^T) \\
&= Trace(VV^T \Sigma^T \Sigma) \\
&= Trace(\Sigma^T \Sigma) \\
&= Trace(\Sigma^2) \\
&= \sum_{i=1}^{r} \sigma_i^2
\end{aligned}$$

□

18.11 Proof of conditioning number

See Lecture 16

18.12 Proof of Rayleigh

See Lecture 17