12 | January 2021
Dr.Rakesh Nigam- ML Lecture 12
Scribe:Sahana Seetharaman

0.1 Basic Definitions and Results

We state the following definitions and theorems for completeness. They have been covered in greater detail in earlier lectures. Let A be a matrix of order $m \times n$.

Definition 0.1.1 (Row Space)**.** The row space of matrix A is the set of linear combinations of the rows of the matrix A. If $\overrightarrow{x}_i^T$ are rows of A, then the row space is written as

$$R(A) = \{\sum_{i=1}^{m} a_i \overrightarrow{x}_i^T : a_i \in \mathbb{R}\}$$

.

Example 0.1.1. The row space of the matrix $\begin{bmatrix} 1 & 0 & -1 & 5 \\ 2 & 7 & 9 & 4 \end{bmatrix}$ is all linear combinations of the rows. That is, $\{a(1, 0, -1, 5) + b(2, 7, 9, 4)\}$ for all $a, b \in \mathbb{R}$.

Definition 0.1.2 (Column Space)**.** The column space of matrix A is the set of linear combinations of the columns of A. If $\overrightarrow{y_i}$ are columns of A, then column space is written as

$$C(A) = \{\sum_{i=1}^{n} a_i \overrightarrow{y_i} : a_i \in \mathbb{R}\}$$

.

Example 0.1.2. The column space of the matrix $\begin{bmatrix} 1 & 0 & -1 & 5 \\ 2 & 7 & 9 & 4 \end{bmatrix}$ is all linear combinations of the columns. That is $\{a \begin{bmatrix} 1 \\ 2 \end{bmatrix} + b \begin{bmatrix} 0 \\ 7 \end{bmatrix} + c \begin{bmatrix} -1 \\ 9 \end{bmatrix} + d \begin{bmatrix} 5 \\ 4 \end{bmatrix}\}$, for all $a, b, c, d \in \mathbb{R}$

Definition 0.1.3 (Null Space)**.** The nullspace of a matrix A is the set of all vectors that equal $\overrightarrow{0}$ when operated upon by A.

$$N(A) = \{\overrightarrow{x} \in \mathbb{R}^n : A\overrightarrow{x} = \overrightarrow{0}_m\}$$

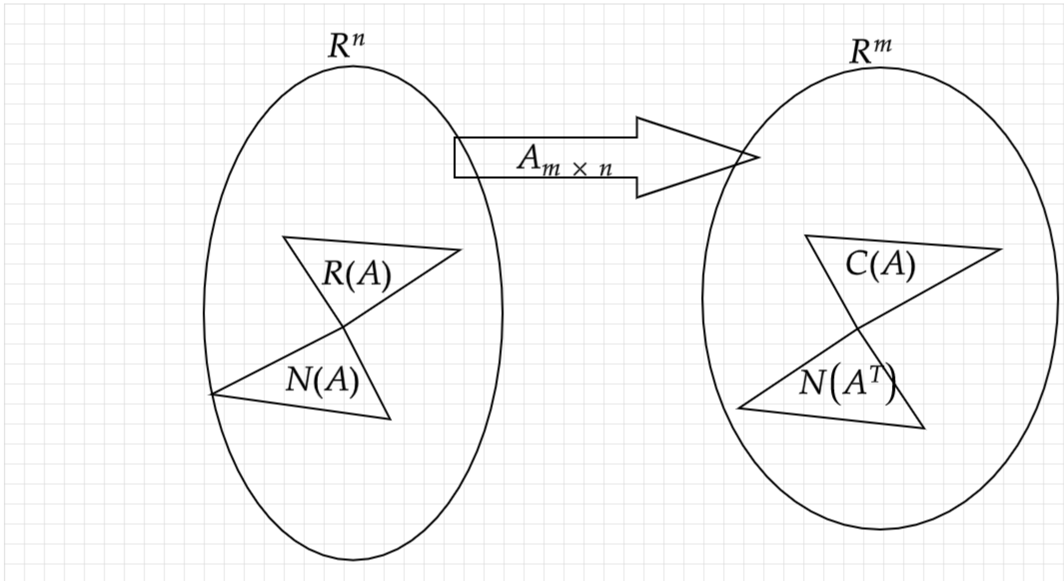

Figure 1: **The diagram shows the four fundamental subspaces of a matrix $A_{m\times n}$. The spaces $R(A)$ and $N(A)$ are in $\mathbb{R}^n$ while the spaces $C(A)$ and $N(A^T)$ are in $\mathbb{R}^m$. The dimension of $R(A)$ and $C(A)$ are equal. Further, the spaces $R(A)$, $N(A)$ and $C(A)$, $N(A^T)$ are orthogonal complements of each other respectively.**

Definition 0.1.4 (Null Space of A^T)**.** The nullspace of A^T is the set of all vectors that equal $\overrightarrow{0}$ when operated upon by A^T.

$$N(A^T) = \{\overrightarrow{x} \in \mathbb{R}^m : A\overrightarrow{x} = \overrightarrow{0}_n\}$$

Remark 0.1.1. $R(A)$, $N(A)$ are subspaces of $\mathbb{R}^n$ and $C(A)$, $N(A^T)$ are subspaces of $\mathbb{R}^m$. They are called the Four Fundamental Subspaces.

Definition 0.1.5 (Rank of Matrix)**.** The rank r of a matrix A is the number of linearly independent columns.

We state the following important theorem without proof. They have been covered in more detail in earlier lectures.

Theorem 1. *1. The rank of a matrix is also equal to the number of linearly independent rows. Further, the rank is the dimension of the column space and the dimension of the row space. Thus, $dim(C(A)) = dim(R(A)) = r$.*

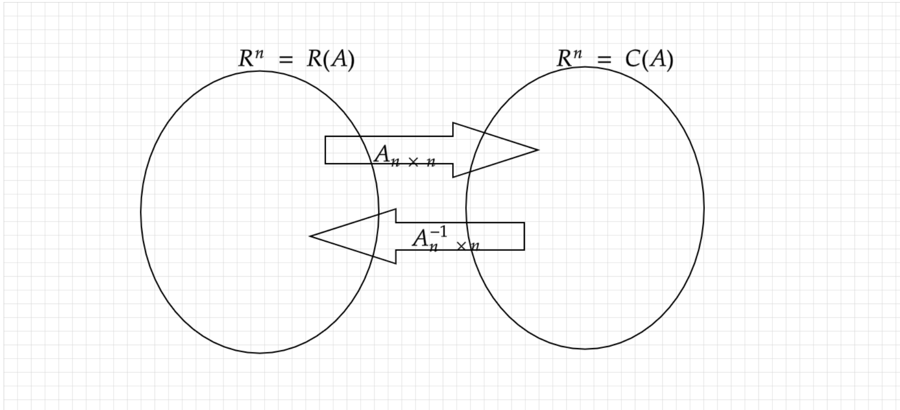

Figure 2: **A matrix A^{-1} is said to be an inverse of the matrix A if $AA^{-1} = I$. It is unique.The inverse of a matrix exists if and only if $m = n = r$. That is, the number of rows, the number of columns and the rank of the matrix are all equal.** $\mathbb{R}^n = \mathbb{R}^m = R(A) = C(A)$. **The spaces $N(A)$ and $N(A^T)$ become the zero subspaces.**

2. *$R(A)$ and $N(A)$ are orthogonal complements of each other.*

3. *Similarly, $C(A)$ and $N(A^T)$ are orthogonal complements of each other.*

0.2 Left and Right Inverses

Definition 0.2.1 (Inverse of a Matrix)**.** The inverse of a square matrix A is the matrix A^{-1} such that $AA^{-1} = I$ and $A^{-1}A = I$, where I is the identity matrix.

Remark 0.2.1. 1. The inverse of a square matrix A exists only when $m = n = r$. That is, the number of rows is equal to the number of columns and is equal to the rank of the matrix. Such matrices are called full rank matrices.

2. $N(A) = \overrightarrow{0}_n$ and $N(A^T) = \overrightarrow{0}_n$. Thus, $\mathbb{R}^n = C(A) = R(A)$.

Definition 0.2.2 (Left Inverse Function)**.** Given a map $f : S \to T$ between sets S and T, the map $g : T \to S$ is called a left inverse of f provided that $g \circ f = id_S$, that is composing f with g from the left gives the identity on S.

Remark 0.2.2. A function f can have a left inverse only if it is injective. A matrix A of order $m \times n$ is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$.

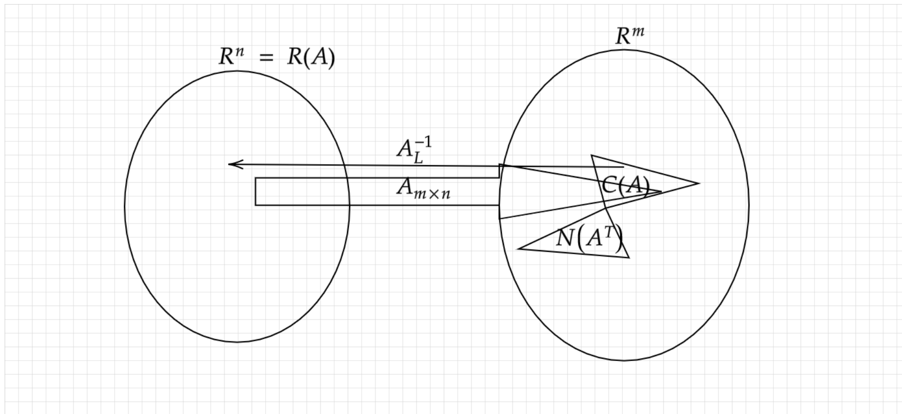

Figure 3: **A matrix $A_L{}^{-1}$ is a left inverse of the matrix A if $A_L{}^{-1}A = I$. A left inverse exists if and only if $N(A) = \overrightarrow{0}$ or, equivalently the columns of A are linearly independent.**

Definition 0.2.3 (Left Inverse of a Matrix)**.** Given a matrix A of order $m \times n$, a matrix $A_L{}^{-1}$ is said to be a left inverse if $A_L{}^{-1}A = I_n$.

Remark 0.2.3. A left inverse exists if and only if the following conditions are satisfied:

1. $N(A) = \overrightarrow{0}_n$
2. The columns of A are linearly independent. In other words, the matrix has full column rank, $n = r$.

Theorem 2. *A^TA is invertible if the columns of A are independent.*

Proof. Let $A^TA\overrightarrow{x} = \overrightarrow{0}_n$. Then $A\overrightarrow{x} \in N(A^T)$. But by definition, $A\overrightarrow{x} \in C(A)$. Since $N(A^T) \cap C(A) = \overrightarrow{0}_m$, we have $A\overrightarrow{x} = \overrightarrow{0}$.
Now, since columns of A are independent, $N(A) = \overrightarrow{0}_n$, we have $\overrightarrow{x} = \overrightarrow{0}_n$. This implies that A^TA is injective.
Now, as A^TA is a square matrix or order $n \times n$. By the rank nullity theorem, $dim(R(A)) = dim(C(A)) = n$ and thus A^TA is invertible. $\square$

We now give a particular left inverse.

Lemma 1. *If the columns of A are linearly independent, $(A^TA)^{-1}A^T$ is a left inverse of A.*

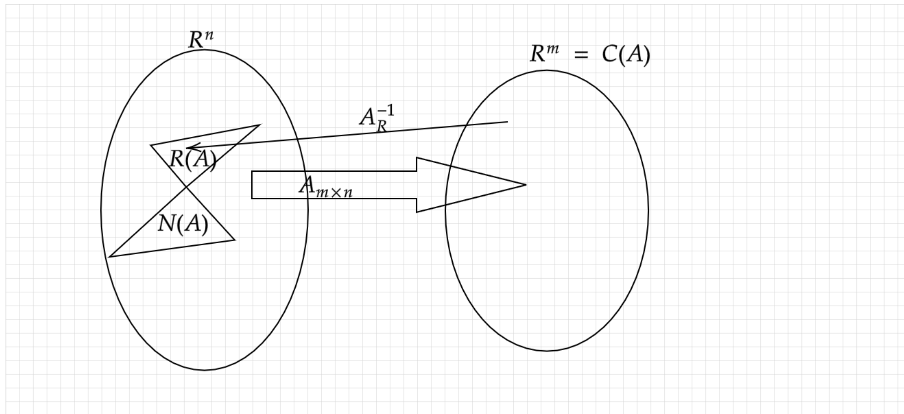

Figure 4: **A matrix $A_R{}^{-1}$ is said to be a right inverse of a matrix A if $AA_R{}^{-1} = I$. A right inverse exists if and only if $N(A^T) = \overrightarrow{0}$ and the rows of A are linearly independent.**

Proof. We have proved the existence of $(A^TA)^{-1}$ in the above theorem. By associativity, it is clear that $(A^TA)^{-1}A^TA = I$. □

Definition 0.2.4 (Function Right Inverse)**.** Given a map $f : S \to T$ between sets S and T, the map $g : T \to S$ is called a right inverse to f provided that $f \circ g = id_T$, that is, composing f with g from the right gives the identity on T.

Remark 0.2.4. If f has a right inverse, then f is surjective.

Definition 0.2.5. Given a matrix A of order $m \times n$, a matrix A_R^{-1} is said to be a right inverse if $AA_R{}^{-1} = I_m$.

Remark 0.2.5. A right inverse exists if and only if the following conditions are satisfied:

1. $N(A^T) = \overrightarrow{0}_m$
2. The rows of A are linearly independent. In other words, the matrix has full row rank, or $m = r$.

Theorem 3. *AA^T is invertible if the rows of A are independent.*

Proof. Let $AA^T\overrightarrow{x} = \overrightarrow{0}$. Then $A^T\overrightarrow{x} \in N(A)$. But, by definition, $A^T\overrightarrow{x} \in R(A)$. Since $N(A) \cap R(A) = \overrightarrow{0}$, we have $A^T\overrightarrow{x} = \overrightarrow{0}$, we have $A^T\overrightarrow{x} = \overrightarrow{x}$. Now since the rows are linearly independent, $N(A^T) = \overrightarrow{0}$, we have $\overrightarrow{x} = \overrightarrow{0}$. This implies that AA^T is invertible. □

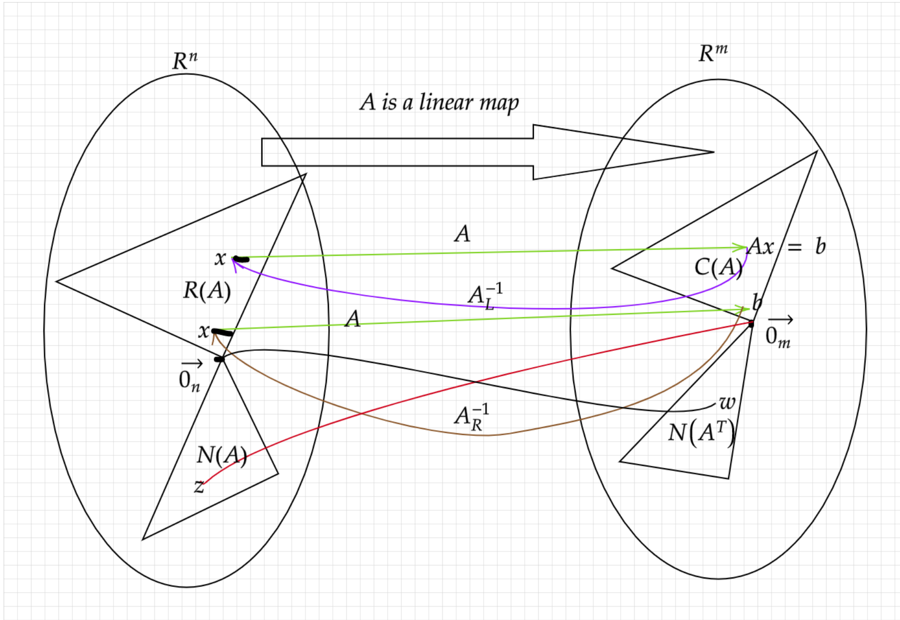

Figure 5: **This diagram shows the four fundamental subspaces of the matrix $A_{m\times n}$. The subspaces $R(A)$ and $N(A)$ are in $\mathbb{R}^n$ and are orthogonal complements of each other. Similarly, the subspaces $C(A)$ and $N(A^T)$ are in $\mathbb{R}^m$ and are orthogonal complements of each other. A left inverse exists iff $N(A) = \overrightarrow{0}$ and columns of A are linearly independent. A right inverse exixts iff $N(A^T) = \overrightarrow{0}$ and the rows are linearly independent.**

We now give a particular right inverse.

Lemma 2. *If the rows of A are linearly independent, $A^T(AA^T)^{-1}$ is a right inverse.*

Proof. We have proved the existence of $AA^{T^{-1}}$ in the above theorem. By associativity, it is clear that $AA^T(AA^T)^{-1} = I$ □

0.3 Projection Matrix

Let us look at the 1-D case. What is the point closest to $\overrightarrow{b}$ on the line $\overrightarrow{a}$ in the diagram below? We call this closest point the projection of $\overrightarrow{b}$ on $\overrightarrow{a}$.
We can see that the dot product of $\overrightarrow{a}$ and $\overrightarrow{b} - x\overrightarrow{a}$ is 0 because of orthogonality. The dot product can also be expressed as $\overrightarrow{a}^T(\overrightarrow{b} - x\overrightarrow{a}) = 0$. We omit the $\rightarrow$ as it

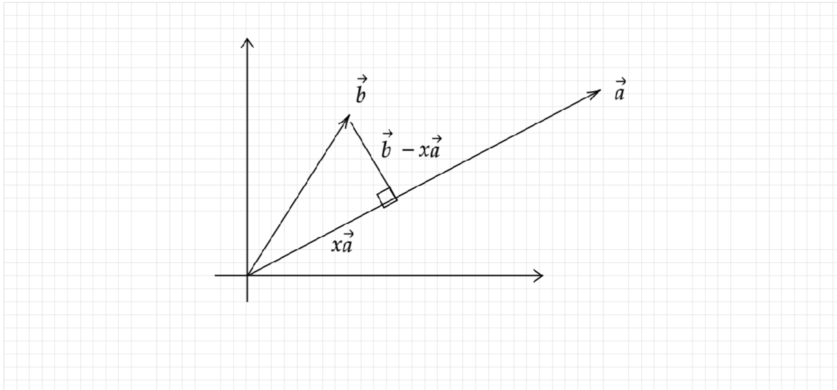

Figure 6: **Given two vectors $\overrightarrow{a}$ and $\overrightarrow{b}$, we want to find the point closest to $\overrightarrow{b}$ on the line spanned by $\overrightarrow{a}$. This point is obtained by projecting $\overrightarrow{b}$ onto $\overrightarrow{a}$.**

is clear from context.

$$\begin{aligned} a^T(b - xa) &= 0 \\ \implies xa^Ta &= a^Tb \\ \implies x &= \frac{a^Tb}{a^Ta} \end{aligned}$$

and

$$p = a\frac{a^Tb}{a^Ta}$$

Thus, the projection p is given by

$$p = Pb \tag{1}$$

where

$$P = \frac{aa^T}{a^Ta}. \tag{2}$$

Observe that P is a matrix. The numerator term is a matrix while the denominator term is a scalar.

Lemma 3. *For P defined as above*

1. $P^T = P$
2. $P^2 = P$.

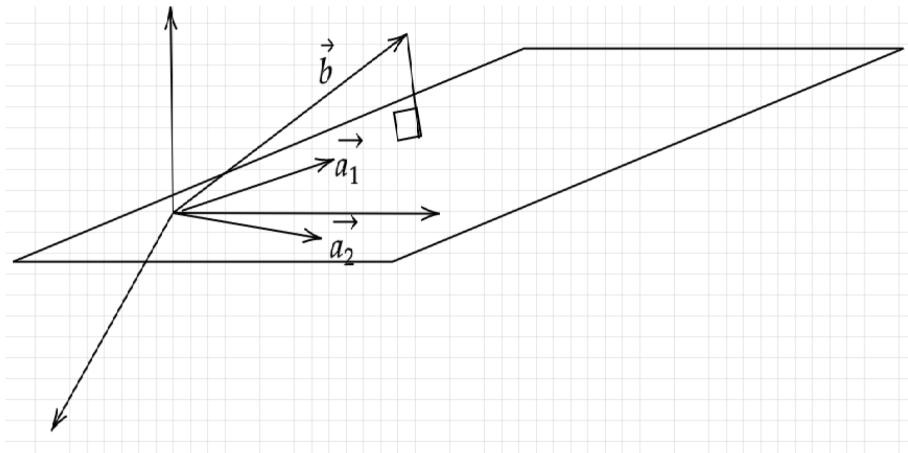

Figure 7: **What is the closest point on the plane spanned by vectors $\overrightarrow{a}_1$ and $\overrightarrow{a}_2$ to the vector $\overrightarrow{b}$? The answer is obtained by projecting $\overrightarrow{b}$ onto the plane spanned by $\overrightarrow{a}_1$ and $\overrightarrow{a}_2$. We multiply $\overrightarrow{b}$ by the projection matrix P in order to get this point.**

Proof. 1. We can see that

$$P^T = (\frac{aa^T}{a^Ta})^T = \frac{(aa^T)^T}{a^Ta} = \frac{aa^T}{a^Ta} = P.$$

2. And

$$P^2 = (\frac{aa^T}{a^Ta})^2 =$$

$$(\frac{aa^T}{a^Ta})(\frac{aa^T}{a^Ta})$$

$$= \frac{a(a^Ta)a^T}{a^Taa^Ta} = \frac{aa^T}{a^Ta} = P.$$

The term a^Ta is a scalar and is cancelled from both numerator and denominator.

□

Why do we have to project? Because $A\overrightarrow{x} = \overrightarrow{b}$ may have no solution. We instead solve for $A\hat{x} = p$ instead, where p is the projection of $\overrightarrow{b}$ onto the column space.
In the diagram above, the plane of a_1, a_2 is the column space of a_1 and a_2. Here, $A = [a_1, a_2]$, that is the columns of A are the vectors a_1 and a_2. Then, $e = b - p$

is perpendicular to the plane. Since p is on the plane, it is in the column space of a_1, a_2. That is, $p = x_1 a_1 + x_2 a_2$. Using more compact notation, $p = A\hat{x}$. Our goal is to find $\hat{x}$. The key is that $b - A\hat{x}$ is perpendicular to the plane.

We get the following equations:

$$a_1^T(b - A\hat{x}) = 0 \tag{3}$$

$$a_2^T(b - A\hat{x}) = 0 \tag{4}$$

equivalently,

$$\begin{bmatrix} a_1^T \\ a_2^T \end{bmatrix} (b - A\hat{x}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{5}$$

or,

$$A^T(b - A\hat{x}) = 0. \tag{6}$$

which implies

$$A^T A\hat{x} = A^T b \tag{7}$$

thus,

$$\hat{x} = (A^T A)^{-1} A^T b \tag{8}$$

The projection p is given by

$$p = A\hat{x} = A(A^T A)^{-1} A^T b \tag{9}$$

and the projection matrix is

$$P = A(A^T A)^{-1} A^T b \tag{10}$$

The above expression for P is well-defined as $A^T A$ is invertible as the columns of the matrix A are linearly independent.

Lemma 4. *For P the projection matrix, defined as above*

1. *P is symmetric, that is, $P = P^T$.*

2. *$P^2 = P$.*

Proof. 1. $P^T = (A(A^T A)^{-1} A^T)^T = (A^T)^T((A^T A)^{-1})^T A^T = A((A^T A)^T)^{-1} A^T = A(A^T A)^{-1} A^T = P$

2. $P^2 = (A(A^T A)^{-1} A^T)^2 = (A(A^T A)^- 1 A^T)(A(A^T A)^{-1} A^T) = A(A^T A)^{-1}(A^T A)(A^T A)^{-1} A^T = A(A^T A)^{-1} A^T$

$\square$

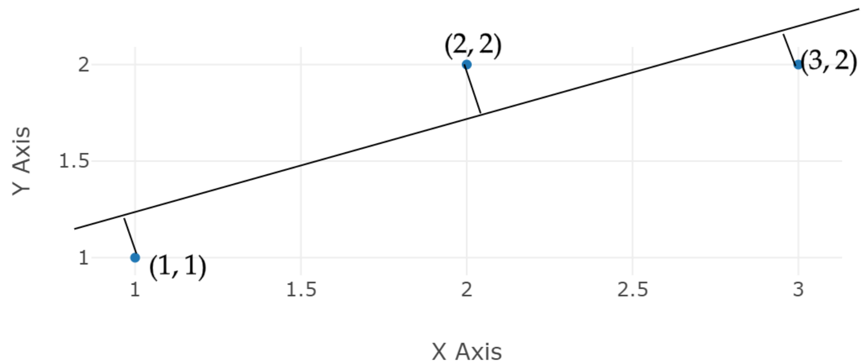

Figure 8: **The regression problem is to fit a straight line to the given points** $(1,1)$**,** $(2,2)$ **and** $(3,2)$**.Since no line passes through all the given points, we find the best possible line that fits the points. We want to minimize the errors.**

0.4 Least Squares

The problem is to fit a line through the given points.

Problem 1. *Fit the points* $(1,1)$*,* $(2,2)$ *and* $(3,2)$ *by a straight line.*

Solution 1. *We want to solve the equations*

$$C + D = 1$$

$$C + 2D = 2$$

$$C + 3D = 2$$

In matrix notation, we write $\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$ *Clearly, this system has no solution. However, the system* $A^T A\hat{x} = A^T b$ *has a solution.*

Since we cannot fit a line through the points, we solve the closest problem, we solve for $A^T A\hat{x} = A^T b$ such that error is minimized. **This problem is called regression in statistics.**

Problem 2. *Minimize*

$$\|Ax - b\|^2 = \|e\|^2 = e_1^2 + e_2^2 + e_3^2$$

We get the same solution when we use calculus and linear algebra.

Solution 2 (Linear Algebra). *We solve the normal equations*

$$A^T A\hat{x} = A^T b$$

The LHS can be written as

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$$

and the RHS can be written as

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}$$

We write the augmented matrix

$$\left[\begin{array}{cc|c} 3 & 6 & 5 \\ 6 & 14 & 11 \end{array}\right]$$

Using ERO $R_2 \to R_2 - 2R_1$, we get

$$\left[\begin{array}{cc|c} 3 & 6 & 5 \\ 0 & 2 & 1 \end{array}\right]$$

Using ERO, $R_2 \to \frac{1}{2}R_2$, we get

$$\left[\begin{array}{cc|c} 3 & 6 & 5 \\ 0 & 1 & 1/2 \end{array}\right]$$

Solving we get $C = 2/3$ and $D = 1/2$.

Solution 3 (Calculus). *We minimize the error term $\|Ax - b\|^2$ using calculus.*

$\|Ax - b\|^2 = \|error^2\| = e_1^2 + e_2^2 + e_3^2 = (C + D - 1)^2 + (C + 2D - 2)^2 + (C + 3D - 2)^2.$

We differentiate and set the first partial derivatives to 0.

$$\frac{\partial(error)}{\partial C} = 2(C + D - 1) + 2(C + 2D - 2) + 2(C + 3D - 2) = 0$$

. *And,*

$$\frac{\partial(error)}{\partial D} = 2(C + D - 1) + 4(C + 2D - 2) + 6(C + 3D - 2) = 0$$

. *The equations above simplify to*

$$6C + 12D = 10$$

and

$$12C + 28D = 28$$

which further simplify to

$$3C + 6D = 5$$

$$6C + 14D = 11$$

which gives $D = 1/2$ *and* $C = 2/3$. *We skipped the second derivative test as the error is a sum of squares and cannot have a maximum*

Thus we get the same solutions in both the cases, that is solving the normal equations in linear algebra and minimizing the errors using calculus.

Bibliography

[1] Prof. Gilbert Strang, MIT 18.06 OpenCourseware,
https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010

[2] Dr. Eric Weisstein, Wolfram MathWorld
https://mathworld.wolfram.com/