MA110
Lecture 19

Saurav Bhaumik
Department of Mathematics
IIT Bombay

Spring 2025

Orthogonal Projection

Let $\mathbb{K}$ denote either $\mathbb{R}$ or $\mathbb{C}$. Recall that in Lecture 13, we have defined the (perpendicular) projection of $\mathbf{x} \in \mathbb{K}^{n\times 1}$ in the direction of nonzero $\mathbf{y} \in \mathbb{K}^{n\times 1}$ as follows:

$$P_{\mathbf{y}}(\mathbf{x}) := \frac{\langle \mathbf{y}, \mathbf{x}\rangle}{\langle \mathbf{y}, \mathbf{y}\rangle}\,\mathbf{y}.$$

In particular, if $\mathbf{y}$ is a unit vector, then $P_{\mathbf{y}}(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x}\rangle \mathbf{y}$.

We noted that the vector $P_{\mathbf{y}}(\mathbf{x})$ is a scalar multiple of the vector $\mathbf{y}$, and proved the important relation

$$\big(\mathbf{x} - P_{\mathbf{y}}(\mathbf{x})\big) \perp \mathbf{y}.$$

As a consequence,

$$\begin{aligned} \|\mathbf{x} - P_{\mathbf{y}}(\mathbf{x})\|^2 &= \langle \mathbf{x} - P_{\mathbf{y}}(\mathbf{x}),\, \mathbf{x} - P_{\mathbf{y}}(\mathbf{x})\rangle \\ &= \langle \mathbf{x},\, \mathbf{x} - P_{\mathbf{y}}(\mathbf{x})\rangle \\ &= \|\mathbf{x}\|^2 - \langle \mathbf{x},\, P_{\mathbf{y}}(\mathbf{x})\rangle. \end{aligned}$$

More generally, let Y be a nonzero subspace of $\mathbb{K}^{n\times 1}$. We would like to find a (perpendicular) projection of $\mathbf{x} \in \mathbb{K}^{n\times 1}$ into Y, that is, we want to find $\mathbf{y} \in Y$ such that $(\mathbf{x} - \mathbf{y}) \in Y^{\perp}$. (This $\mathbf{y}$ is 'the foot of the perpendicular' from $\mathbf{x}$ into Y.)

If $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is an orthonormal basis for the subspace Y, then a vector belongs to $Y^{\perp}$ if and only if it is orthogonal to each $\mathbf{u}_j$ for $j = 1, \ldots, k$. As we saw while studying G-S OP, the vector

$$\tilde{\mathbf{y}} := \mathbf{x} - P_{\mathbf{u}_1}(\mathbf{x}) - \cdots - P_{\mathbf{u}_k}(\mathbf{x}) = \mathbf{x} - \langle \mathbf{u}_1, \mathbf{x}\rangle \mathbf{u}_1 - \cdots - \langle \mathbf{u}_k, \mathbf{x}\rangle \mathbf{u}_k$$

is orthogonal to each $\mathbf{u}_j$ for $j = 1, \ldots, k$, and so $\tilde{\mathbf{y}} \in Y^{\perp}$.

Since the vector $\mathbf{y} := \langle \mathbf{u}_1, \mathbf{x}\rangle \mathbf{u}_1 + \cdots + \langle \mathbf{u}_k, \mathbf{x}\rangle \mathbf{u}_k$ belongs to Y, it is a (perpendicular) projection of $\mathbf{x}$ in Y.

The following result shows that this is the only vector in Y that works!

Proposition (Projection Theorem)

Let Y be a subspace of $\mathbb{K}^{n\times 1}$. Then for every $\mathbf{x} \in \mathbb{K}^{n\times 1}$, there are unique $\mathbf{y} \in Y$ and $\tilde{\mathbf{y}} \in Y^{\perp}$ such that $\mathbf{x} = \mathbf{y} + \tilde{\mathbf{y}}$, that is, $\mathbb{K}^{n\times 1} = Y \oplus Y^{\perp}$. The map $P_Y : \mathbb{K}^{n\times 1} \to \mathbb{K}^{n\times 1}$ given by $P_Y(\mathbf{x}) = \mathbf{y}$ is linear and satisfies $(P_Y)^2 = P_Y$.

In fact, if $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is an orthonormal basis for Y, then

$$P_Y(\mathbf{x}) = \langle \mathbf{u}_1, \mathbf{x}\rangle \mathbf{u}_1 + \cdots + \langle \mathbf{u}_k, \mathbf{x}\rangle \mathbf{u}_k.$$

Proof. If $Y = \{\mathbf{0}\}$, then every $\mathbf{x} \in Y^{\perp}$, and so $\mathbf{x} = \mathbf{0} + \mathbf{x}$.

Suppose $Y \neq \{\mathbf{0}\}$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_k$ be an orthonormal basis for Y. For $\mathbf{x} \in \mathbb{K}^{n\times 1}$, define

$$P_Y(\mathbf{x}) := \langle \mathbf{u}_1, \mathbf{x}\rangle \mathbf{u}_1 + \cdots + \langle \mathbf{u}_k, \mathbf{x}\rangle \mathbf{u}_k.$$

Clearly, $\mathbf{y} := P_Y(\mathbf{x}) \in Y$. Define $\tilde{\mathbf{y}} := \mathbf{x} - P_Y(\mathbf{x})$. Then

$$\begin{aligned}\langle \mathbf{u}_j, \tilde{\mathbf{y}} \rangle &= \langle \mathbf{u}_j, \mathbf{x} - P_Y(\mathbf{x}) \rangle \\ &= \langle \mathbf{u}_j, \mathbf{x} \rangle - \sum_{\ell=1}^{k} \langle \mathbf{u}_\ell, \mathbf{x} \rangle \langle \mathbf{u}_j, \mathbf{u}_\ell \rangle \\ &= \langle \mathbf{u}_j, \mathbf{x} \rangle - \langle \mathbf{u}_j, \mathbf{x} \rangle = 0\end{aligned}$$

for $j = 1, \ldots, k$ by orthonormality. Hence $\tilde{\mathbf{y}} \in Y^\perp$. Thus $\mathbf{x} = \mathbf{y} + \tilde{\mathbf{y}}$ with $\mathbf{y} \in Y$ and $\tilde{\mathbf{y}} \in Y^\perp$. This proves existence.

To prove uniqueness, let $\mathbf{z} \in Y$ and $\tilde{\mathbf{z}} \in Y^\perp$ be such that $\mathbf{x} = \mathbf{z} + \tilde{\mathbf{z}}$. Then $(\mathbf{y} - \mathbf{z}) \in Y$ and also $\mathbf{y} - \mathbf{z} = (\tilde{\mathbf{z}} - \tilde{\mathbf{y}}) \in Y^\perp$, so that $(\mathbf{y} - \mathbf{z}) \in Y \cap Y^\perp$. Hence $(\mathbf{y} - \mathbf{z}) \perp (\mathbf{y} - \mathbf{z})$, and so $\mathbf{y} - \mathbf{z} = \mathbf{0}$. Thus $\mathbf{z} = \mathbf{y}$, and in turn, $\tilde{\mathbf{z}} = \tilde{\mathbf{y}}$.

The map P_Y is linear since the inner product is linear in the second variable. Also, $P_Y(\mathbf{u}_j) = \mathbf{u}_j$ for all $j = 1, \ldots, k$. Hence $P_Y(\mathbf{x}) = \mathbf{x}$ if and only if $\mathbf{x} \in Y$. As a consequence, $P_Y^2(\mathbf{x}) = P_Y(P_Y(\mathbf{x})) = P_Y(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{K}^{n \times 1}$. □

Let Y be a subspace of $\mathbb{K}^{n\times 1}$. Then the linear map $P_Y : \mathbb{K}^{n\times 1} \to \mathbb{K}^{n\times 1}$ whose existence and uniqueness is proved in the above result is called the **orthogonal projection map** of $\mathbb{K}^{n\times 1}$ onto the subspace Y.

Given $\mathbf{x} \in \mathbb{K}^{n\times 1}$, we shall show that its orthogonal projection $P_Y(\mathbf{x})$ is the unique vector in Y which is closest to $\mathbf{x}$.

Definition

Let E be a nonempty subset of $\mathbb{K}^{n\times 1}$ and let $\mathbf{x} \in \mathbb{K}^{n\times 1}$.
A **best approximation** *to $\mathbf{x}$ from E is an element $\mathbf{y}_0 \in E$ such that $\|\mathbf{x} - \mathbf{y}_0\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{y} \in E$.*

Simple examples show that a best approximation to a vector $\mathbf{x}$ from a nonempty subset E of $\mathbb{K}^{n\times 1}$ may not exist, and if it exists, it may not be unique. However, if E is in fact a subspace of $\mathbb{K}^{n\times 1}$, then the following noteworthy result holds.

Proposition

Let Y be a subspace of $\mathbb{K}^{n\times 1}$ and let $\mathbf{x} \in \mathbb{K}^{n\times 1}$. Then there is a unique best approximation to $\mathbf{x}$ from Y, namely, $P_Y(\mathbf{x})$.

Further, $P_Y(\mathbf{x})$ is the unique element of Y such that $\mathbf{x} - P_Y(\mathbf{x})$ is orthogonal to Y. Also, the square of the distance from $\mathbf{x}$ to its best approximation from Y is

$$\|\mathbf{x} - P_Y(\mathbf{x})\|^2 = \|\mathbf{x}\|^2 - \langle \mathbf{x}, P_Y(\mathbf{x})\rangle.$$

Proof. We note that $P_Y(\mathbf{x}) \in Y$ and $\big(\mathbf{x} - P_Y(\mathbf{x})\big) \in Y^\perp$.

Let $\mathbf{y} \in Y$. Then $P_Y(\mathbf{x}) - \mathbf{y}$ also belongs to Y. Hence by the Pythagorus theorem,

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|^2 &= \big\|\big(\mathbf{x} - P_Y(\mathbf{x})\big) + \big(P_Y(\mathbf{x}) - \mathbf{y}\big)\big\|^2 \\ &= \|\mathbf{x} - P_Y(\mathbf{x})\|^2 + \|P_Y(\mathbf{x}) - \mathbf{y}\|^2 \\ &\geq \|\mathbf{x} - P_Y(\mathbf{x})\|^2, \end{aligned}$$

where equality holds if and only if $\mathbf{y} = P_Y(\mathbf{x})$. This shows that

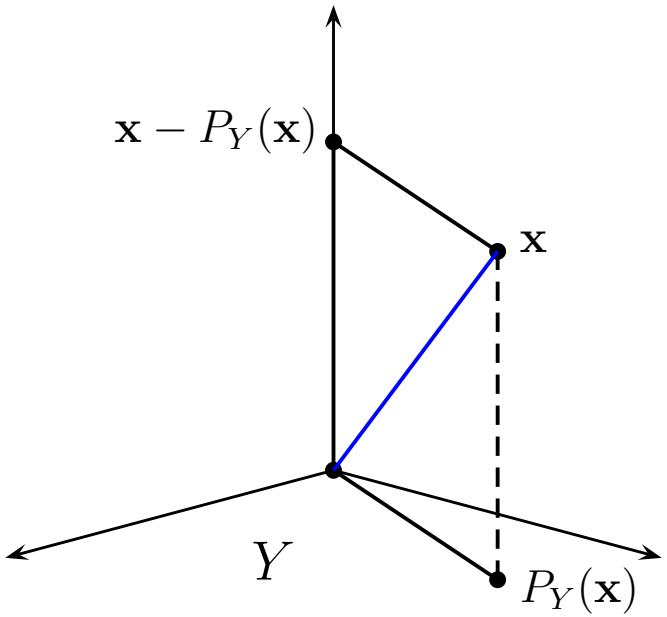
$P_Y(\mathbf{x})$ is the unique best approximation to $\mathbf{x}$ from Y.

Further, let $\mathbf{y} \in Y$ be such that $(\mathbf{x} - \mathbf{y}) \in Y^\perp$. Then $\mathbf{x} = \mathbf{y} + (\mathbf{x} - \mathbf{y})$, and since $\mathbb{K}^{n\times 1} = Y \oplus Y^\perp$, it follows that $\mathbf{y} = P_Y(\mathbf{x})$.

Finally, since $P_Y(\mathbf{x}) \in Y$ and $\big(\mathbf{x} - P_Y(\mathbf{x})\big) \in Y^\perp$,

$$\begin{aligned} \|\mathbf{x} - P_Y(\mathbf{x})\|^2 &= \langle \mathbf{x} - P_Y(\mathbf{x}),\, \mathbf{x} - P_Y(\mathbf{x}) \rangle \\ &= \langle \mathbf{x},\, \mathbf{x} - P_Y(\mathbf{x}) \rangle \\ &= \|\mathbf{x}\|^2 - \langle \mathbf{x},\, P_Y(\mathbf{x}) \rangle. \end{aligned}$$

□



Remark

As we have seen before, if $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is an orthonormal basis for Y, then $P_Y(\mathbf{x}) = \langle \mathbf{u}_1, \mathbf{x} \rangle \mathbf{u}_1 + \cdots + \langle \mathbf{u}_k, \mathbf{x} \rangle \mathbf{u}_k$, and so

$$\left\| \mathbf{x} - P_Y(\mathbf{x}) \right\|^2 = \|\mathbf{x}\|^2 - \langle \mathbf{x}, P_Y(\mathbf{x}) \rangle = \|\mathbf{x}\|^2 - \sum_{j=1}^{k} |\langle \mathbf{x}, \mathbf{u}_j \rangle|^2.$$

We now give an application of the above considerations to a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{K}^{m \times n}$ and $\mathbf{b} \in \mathbb{K}^{m \times 1}$.

Let $\mathbf{A} = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_n \end{bmatrix}$ in terms of its n columns. Now the above system has a solution $\mathbf{x} := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^{\mathsf{T}}$ if and only if $x_1\mathbf{c}_1 + \cdots + x_n\mathbf{c}_n = \mathbf{b}$, that is, $\mathbf{b}$ belongs to the column space $\mathcal{C}(\mathbf{A})$ of $\mathbf{A}$.

Suppose $\mathbf{b} \notin \mathcal{C}(\mathbf{A})$. We would like to find the best approximation to $\mathbf{b}$ from the subspace $\mathcal{C}(\mathbf{A})$ of $\mathbb{K}^{m \times 1}$.

Starting with the first column $\mathbf{c}_1$ of $\mathbf{A}$, and using the G-S OP, we find an ordered orthonormal basis $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ of $\mathcal{C}(\mathbf{A})$, where $k \leq n$. The best approximation $\mathbf{a}$ to $\mathbf{b}$ from $\mathcal{C}(\mathbf{A})$ is

$$\mathbf{a} := P_{\mathcal{C}(\mathbf{A})}(\mathbf{b}) = \sum_{j=1}^{k} \langle \mathbf{u}_j, \mathbf{b} \rangle \mathbf{u}_j = \sum_{j=1}^{k} (\mathbf{u}_j^* \mathbf{b}) \mathbf{u}_j.$$

Example

Let $\mathbf{A} := \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{b} := \begin{bmatrix} 1 \\ 0 \\ -5 \end{bmatrix}$. Then $\mathcal{C}(\mathbf{A})$ is the span of the 2 column vectors $\begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^{\mathsf{T}}$ and $\begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^{\mathsf{T}}$ of $\mathbf{A}$. Then $\mathbf{u}_1 := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix}^{\mathsf{T}}$ and $\mathbf{u}_2 := \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \end{bmatrix}^{\mathsf{T}}$ form an orthonormal basis for $\mathcal{C}(\mathbf{A})$. Hence the best approximation to $\mathbf{b}$ from $\mathcal{C}(\mathbf{A})$ is

$$(\mathbf{u}_1^* \mathbf{b})\mathbf{u}_1 + (\mathbf{u}_2^* \mathbf{b})\mathbf{u}_2 = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^{\mathsf{T}} - \frac{3}{2} \begin{bmatrix} 1 & -1 & 2 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} -1 & 2 & -3 \end{bmatrix}^{\mathsf{T}}.$$

Let us calculate the distance from $\mathbf{b}$ to its best approximation $\mathbf{a}$ from $\mathcal{C}(\mathbf{A})$. Clearly, it is equal to

$$\begin{aligned} \|\mathbf{b}-\mathbf{a}\| &= \|\begin{bmatrix}1 & 0 & -5\end{bmatrix}^{\mathsf{T}} - \begin{bmatrix}-1 & 2 & -3\end{bmatrix}^{\mathsf{T}}\| \\ &= \|\begin{bmatrix}2 & -2 & -2\end{bmatrix}^{\mathsf{T}}\| = 2\sqrt{3}. \end{aligned}$$

Also, according to an earlier formula, the square of this distance is equal to

$$\|\mathbf{b}\|^2 - |\langle \mathbf{b}, \mathbf{u}_1\rangle|^2 - |\langle \mathbf{b}, \mathbf{u}_2\rangle|^2 = 26 - \frac{1}{2} - \frac{27}{2} = 12,$$

and so the desired distance is equal to $\sqrt{12} = 2\sqrt{3}$, as obtained above.

We now explain an alternative approach for finding the best approximation $\mathbf{y}_0$ to $\mathbf{b}$ from $\mathcal{C}(\mathbf{A})$. It avoids the G-S OP.

By the previous proposition, $\mathbf{y}_0$ is the unique element of $\mathcal{C}(\mathbf{A})$ such that $(\mathbf{b} - \mathbf{y}_0) \in \mathcal{C}(\mathbf{A})^\perp$, that is, $\mathbf{c}_j^*(\mathbf{b} - \mathbf{y}_0) = 0$, where $\mathbf{c}_j$ is the jth column of $\mathbf{A}$, for $j = 1, \ldots, n$, which means $\mathbf{A}^*(\mathbf{b} - \mathbf{y}_0) = \mathbf{0}$.

Thus we need to find $\mathbf{x}_0 \in \mathbb{K}^{n\times 1}$ such that $\mathbf{A}^*(\mathbf{b} - \mathbf{A}\mathbf{x}_0) = \mathbf{0}$. Hence $\mathbf{y}_0$ can be obtained by finding a solution $\mathbf{x}_0$ of the **normal equations** $\boxed{\mathbf{A}^*\mathbf{A}\,\mathbf{x} = \mathbf{A}^*\mathbf{b}}$, and then letting $\mathbf{y}_0 := \mathbf{A}\mathbf{x}_0$. Since $\mathbf{y}_0 \in \mathcal{C}(\mathbf{A})$, this system of n equations in n unknowns always has a solution.

Suppose rank $\mathbf{A} = n$, that is, the columns of the matrix $\mathbf{A}$ are linearly independent. Then rank $\mathbf{A}^*\mathbf{A} = n$, and the unique solution of the normal equations is given by $\mathbf{x}_0 := (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}$. In this case, the best approximation to $\mathbf{b}$ from $\mathcal{C}(\mathbf{A})$ is $\mathbf{y}_0 := \mathbf{A}\mathbf{x}_0 = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}$. In particular,

if $m = n$, then $\mathbf{A}$ is invertible, and $\mathbf{y}_0 = \mathbf{A}\mathbf{A}^{-1}(\mathbf{A}^*)^{-1}\mathbf{A}^*\mathbf{b} = \mathbf{b}$.

Next, suppose rank $\mathbf{A} < n$, that is, the columns of $\mathbf{A}$ are linearly dependent. Then rank $\mathbf{A}^*\mathbf{A} < n$, and the normal equations have infinitely many solutions, but if $\mathbf{x}_0$ is any one of them, then $\mathbf{y}_0 = \mathbf{A}\mathbf{x}_0$.

Example

Let $\mathbf{A} := \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{K}^{3\times 2}$ and $\mathbf{b} := \begin{bmatrix} 1 \\ 0 \\ -5 \end{bmatrix} \in \mathbb{K}^{3\times 1}$. We find the best approximation to $\mathbf{b}$ from the column space of $\mathbf{A}$ using normal equations. Here $\mathbf{A}^*\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and $\mathbf{A}^*\mathbf{b} = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$.

Now the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$, that is, $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$ have a unique solution, namely $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$. Hence $\mathbf{A}\mathbf{x} = \begin{bmatrix} -1 & 2 & -3 \end{bmatrix}^{\mathsf{T}}$ as before.

Next, let $\mathbf{B} := \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \end{bmatrix}$ and $\mathbf{b} := \begin{bmatrix} 1 \\ 0 \\ -5 \end{bmatrix}$. Then $\mathcal{C}(\mathbf{B}) = \mathcal{C}(\mathbf{A})$ since the third column of $\mathbf{B}$ is the average of the first two columns of $\mathbf{A}$, and so the best approximation to $\mathbf{b}$ from $\mathcal{C}(\mathbf{B})$ is the same column vector $\begin{bmatrix} -1 & 2 & -3 \end{bmatrix}^{\mathsf{T}}$ as before. But here

$$\mathbf{B}^*\mathbf{B} = \begin{bmatrix} 2 & 1 & 3/2 \\ 1 & 2 & 3/2 \\ 3/2 & 3/2 & 3/2 \end{bmatrix} \quad \text{and} \quad \mathbf{B}^*\mathbf{b} = \begin{bmatrix} 1 \\ -4 \\ -3/2 \end{bmatrix}.$$

Now the normal equations $\mathbf{B}^*\mathbf{B}\mathbf{x} = \mathbf{B}^*\mathbf{b}$ have many solutions such as $\begin{bmatrix} 2 & -3 & 0 \end{bmatrix}^{\mathsf{T}}$ or $\begin{bmatrix} 3 & -2 & -2 \end{bmatrix}^{\mathsf{T}}$, since the columns of $\mathbf{B}$ are linearly dependent. However,

$$\mathbf{B}\begin{bmatrix} 2 & -3 & 0 \end{bmatrix}^{\mathsf{T}} = \mathbf{B}\begin{bmatrix} 3 & -2 & -2 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} -1 & 2 & -3 \end{bmatrix}^{\mathsf{T}} \text{ as before.}$$

Least Squares Approximation

Suppose a large number of data points $(a_1, b_1), \ldots, (a_n, b_n)$ in $\mathbb{R}^2$ are given, and we want to find a straight line passing through these points. If all these points are collinear, then we may join any two of them by a straight line, which will work for us. If, however, these points are not collinear (which is most often the case), then we want to find a straight line $t = x_1 + s\,x_2$ in the st-plane which is most suitable in the following sense.

Consider the 'error' $|x_1 + a_j x_2 - b_j|$ caused because of the straight line $t = x_1 + s\,x_2$ not passing through the data point (a_j, b_j) for $j = 1, \ldots, n$. We collect these errors and attempt to find $x_1, x_2 \in \mathbb{R}$ such that the **least squares error**

$$\left(\sum_{j=1}^{n} |x_1 + a_j x_2 - b_j|^2 \right)^{1/2}.$$

is minimized. This is known as the **least squares problem**.

To solve this problem, let

$$\mathbf{A} := \begin{bmatrix} 1 & a_1 \\ 1 & a_2 \\ \vdots & \vdots \\ 1 & a_n \end{bmatrix}, \quad \mathbf{x} := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{b} := \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then $\mathbf{Ax} = \begin{bmatrix} x_1 + a_1 x_2 & x_1 + a_2 x_2 & \cdots & x_1 + a_n x_2 \end{bmatrix}^{\mathsf{T}}$.

The least squares problem is to find $\mathbf{x} \in \mathbb{R}^{2\times 1}$ such that

$$\|\mathbf{Ax} - \mathbf{b}\|^2 = \sum_{j=1}^{n} |x_1 + a_j x_2 - b_j|^2$$

is minimised, that is, to find the best approximation to the vector $\mathbf{b}$ from the column space $\mathcal{C}(\mathbf{A}) = \{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^{2\times 1}\}$ of $\mathbf{A}$.

As we have seen, this is done either by orthonormalizing the columns of $\mathbf{A}$, or by finding a vector $\mathbf{x}_0 := \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathsf{T}}$ such that $\mathbf{b} - \mathbf{Ax}_0$ is orthogonal to $\mathcal{C}(\mathbf{A})$ using the normal equations.

Instead of fitting a straight line $t = x_1 + s\,x_2$ to the data points, we may attempt to fit a curve given by a polynomial $t = x_1 + s\,x_2 + s^2 x_3 + \cdots + s^k x_{k+1}$ of degree $k \in \mathbb{N}$, to the data points $(a_1, b_1), \ldots, (a_n, b_n)$. To solve this problem, let

$$\mathbf{A} := \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^k \\ 1 & a_2 & a_2^2 & \cdots & a_2^k \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^k \end{bmatrix}, \quad \mathbf{x} := \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k+1} \end{bmatrix} \quad \text{and} \quad \mathbf{b} := \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then $\mathbf{Ax}$ is equal to

$$\begin{bmatrix} x_1 + a_1 x_2 + \cdots + a_1^k x_{k+1} & \cdots & x_1 + a_n x_2 + \cdots + a_n^k x_{k+1} \end{bmatrix}^{\mathsf{T}}.$$

As before, the least squares problem is to minimize

$$\|\mathbf{Ax} - \mathbf{b}\|^2 = \sum_{j=1}^{n} |x_1 + a_j x_2 + \cdots + a_j^k x_{k+1} - b_j|^2$$

by finding the best approximation to the vector $\mathbf{b}$ from the column space $\mathcal{C}(\mathbf{A}) = \{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^{(k+1)\times 1}\}$ of $\mathbf{A}$.

Example

Let us find a straight line $t = x_1 + s\,x_2$ that best fits the data points $(-1, 1), (1, 1), (2, 3)$ in the least squares sense. For this purpose, we find the orthogonal projection of the vector $\mathbf{b} := \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$ into the column space of the matrix $\mathbf{A} := \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$.

Then $\mathbf{A}^*\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ and $\mathbf{A}^*\mathbf{b} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$.

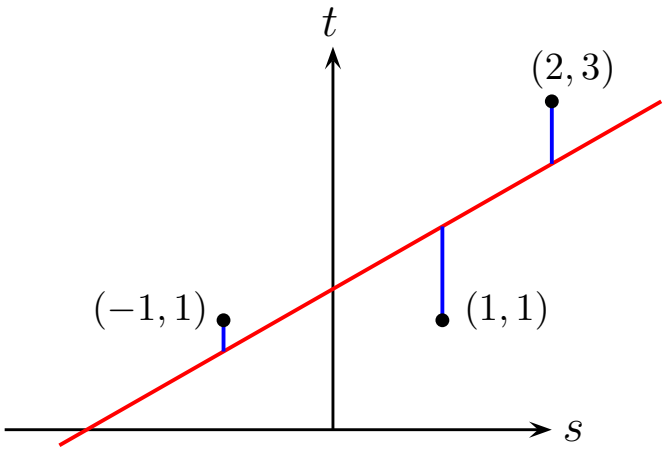
Hence the unique solution of the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ is $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathsf{T}} := \begin{bmatrix} 9/7 & 4/7 \end{bmatrix}^{\mathsf{T}}$. Thus the straight line which best fits the data is $7t = 9 + 4s$.

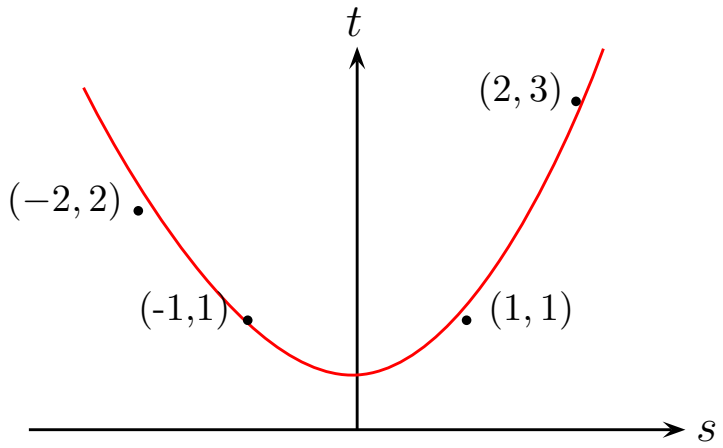
Next, we find a parabola $t = x_1 + s\,x_2 + s^2 x_3$ that best fits the data points $(-1, 1), (1, 1), (2, 3), (-2, 2)$ in the least squares sense.

Let $\mathbf{A} := \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & -2 & 4 \end{bmatrix}$ and $\mathbf{b} := \begin{bmatrix} 1 \\ 1 \\ 3 \\ 2 \end{bmatrix}$.

Then $\mathbf{A}^*\mathbf{A} = \begin{bmatrix} 4 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix}$ and $\mathbf{A}^*\mathbf{b} = \begin{bmatrix} 7 & 2 & 22 \end{bmatrix}$.

The unique solution of the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ is $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^{\mathsf{T}} := \begin{bmatrix} 1/2 & 1/5 & 1/2 \end{bmatrix}^{\mathsf{T}}$. Thus the parabola which best fits the data is $10t = 5 + 2s + 5s^2$.





Abstract Vector Spaces

Throughout the course so far, we have dealt with row vectors, column vectors and matrices. We have introduced many interesting concepts such as linear independence of vectors, span of a set of vectors, a subspace of vectors, a basis for a subspace, the dimension of a subspace, the nullity and the rank of a matrix, linear transformations induced by matrices, inner product of two vectors, orthogonality of vectors, orthonormal basis for a subspace, orthogonal projection onto a subspace, and so on.

Based on these concepts, we have proved some important theorems like the Rank-Nullity theorem, the Fundamental Theorem for Linear Systems, the Spectral Theorem, the Projection Theorem.

Now we discuss these concepts in a more abstract setting.

Abstract Vector Space

Definition Let $\mathbb{K}$ denote $\mathbb{R}$ or $\mathbb{C}$ as usual. A **vector space** over $\mathbb{K}$ is a nonempty set V together with the operation of **addition** (i.e., a map $V \times V \to V$ given by $(u, v) \mapsto u + v$) and of **scalar multiplication** (i.e., a map $\mathbb{K} \times V \to V$ given by $(\alpha, v) \mapsto \alpha v$) satisfying the following properties.

I Closure axioms

1. $u + v \in V$ for all $u, v \in V$.
2. $\alpha \cdot v \in V$ for all $\alpha \in \mathbb{K}$ and $v \in V$.

(We shall write αv instead of $\alpha \cdot v \in V$ hence onward.)

II Axioms for addition

1. $u + v = v + u$ for all $u, v \in V$. (commutativity)
2. $u+(v+w)=(u+v)+w$ for all $u, v, w \in V$. (associativity)
3. There is (unique) $0 \in V$ such that $v + 0 = v$ for all $v \in V$.
4. For $v \in V$, there is (unique) $u \in V$ such that $v + u = 0$.

(We shall write this element u as $-v$.)

III Axioms for scalar multiplication

1. $\alpha(\beta v) = (\alpha\beta)v$ for all $\alpha, \beta \in \mathbb{K}$ and $v \in V$.
2. $\alpha(u + v) = \alpha u + \alpha v$ for all $\alpha \in \mathbb{K}$ and $u, v \in V$.
3. $(\alpha + \beta)v = \alpha v + \beta v$ for all $\alpha, \beta \in \mathbb{K}$ and $v \in V$.
4. $1v = v$ for all $v \in V$.

An element of a vector space is called a **vector**.

Examples: 1 $\mathbb{K}^n = \mathbb{K}^{n\times 1}$ is a vector space over $\mathbb{K}$. More generally, every vector subspace of $\mathbb{K}^n$ is a vector space over $\mathbb{K}$. Likewise for $\mathbb{K}^{1\times n}$.

2 $\mathbb{K}^{m\times n}$, the space of all $m \times n$ matrices with entries in $\mathbb{K}$, is a vector space over $\mathbb{K}$ with respect to the addition and scalar multiplication of matrices.

3 The set $\mathbb{K}[x]$ of all polynomials in the indeterminate x with coefficients in $\mathbb{K}$ is a vector space over $\mathbb{K}$ with respect to the usual addition and scalar multiplication of polynomials.

4 The set $\mathcal{P}_n$ of polynomials in $\mathbb{K}[x]$ of degree $\leq n$ is a vector space with addition and scalar multiplication as in 3.

5 Let $a, b \in \mathbb{R}$ with $a < b$. Then the space $C[a, b]$ of all continuous functions from $[a, b]$ to $\mathbb{R}$ a vector space over $\mathbb{R}$ with respect to pointwise addition and pointwise scalar multiplication of functions.

6 Let $a, b \in \mathbb{R}$ with $a < b$. Then the space $C^1[a, b]$ of all continuously differentiable functions from $[a, b]$ to $\mathbb{R}$ a vector space over $\mathbb{R}$ with addition and scalar multiplication as in 5 .

7 The space c of all convergent sequences of real numbers is a vector space over $\mathbb{R}$ with respect to pointwise addition and pointwise scalar multiplication of sequences.

Exercise: Verify that the above spaces are indeed vector spaces, i.e., all the axioms in the definition are satisfied.

Definition

Let V be a vector space (over $\mathbb{K}$). A nonempty subset W of V is called a **subspace** *of V if $v + w \in W$ for all $v, w \in W$ and $\alpha w \in W$ for all $\alpha \in \mathbb{K}$ and $w \in W$.*

Definition

Let V be a vector space (over $\mathbb{K}$), and let $n \in \mathbb{N}$. Given $v_1, \ldots, v_n \in V$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{K}$, the element

$$\alpha_1 v_1 + \cdots + \alpha_n v_n$$

of V is called the **linear combination** *of the vectors $v_1, \ldots, v_n$ with* **coefficients** *$\alpha_1, \ldots, \alpha_n$.*

Let V be a vector space. If W is a subspace of V, then clearly every linear combination of the elements of W belongs W.

Let W_1. W_2 be subspaces of V. Then it is easy to see that:

- $W_1 \cap W_2$ is a subspace of V. In fact it is the largest subspace of V which is contained in both W_1 and W_2.
- $W_1 \cup W_2$ need not be a subspace of V.
- $W_1 + W_2 := \{w_1 + w_2 : w_1 \in W_1 \text{ and } w_2 \in W_2\}$ is a subspace of V. In fact it is the smallest subspace of V containing both W_1 and W_2.

The notions in $\mathbb{R}^n$ of span of a set of vectors, linear dependence and independence of vectors, the dimension of a subspace of vectors carry over to an abstract vector space without any difficulty. Let V be a vector space (over $\mathbb{K}$).

Definition

Let $S \subset V$. The set of all linear combinations of elements of S is called the **span** *of S, and we denote it by* span S .

Definition

A subset S of V is called **linearly dependent** *if there are $v_1, \ldots, v_m$ in S and there are $\alpha_1, \ldots, \alpha_m \in \mathbb{K}$, not all zero, satisfying*

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0.$$

It can be seen that S is linearly dependent $\iff$ either $\mathbf{0} \in S$ or a vector in S is a linear combination of other vectors in S.

Definition

A subset S of V is called **linearly independent** *if it is not linearly dependent, that is,*

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0 \implies \alpha_1 = \cdots = \alpha_m = 0,$$

whenever $v_1, \ldots, v_m \in S$ and $\alpha_1, \ldots, \alpha_m \in \mathbb{K}$.

Clearly, S is linearly independent if and only if $\mathbf{0} \notin S$ and no elements of S is a linear combination of other elements of S.

The following crucial result was proved for the case $V := \mathbb{R}^{n\times 1}$. Exactly the same proof works in the general case.

Proposition

Let S be a subset of s elements and R be a set of r elements of V. If $S \subset \operatorname{span} R$ and $s > r$, then S is linearly dependent.

Examples

1. Let $m, n \in \mathbb{N}$, and let $V := \mathbb{K}^{m \times n}$ be set of all $m \times n$ matrices with entries in $\mathbb{K}$ with entry-wise addition and scalar multiplication. For $j = 1, \ldots, m$ and $k = 1, \ldots, n$, let $\mathbf{E}_{jk}$ denote the $m \times n$ matrix whose (j, k)th entry is equal to 1 and all other entries are equal to zero. Then the set $S := \left\{\mathbf{E}_{jk} : 1 \leq j \leq m,\ 1 \leq k \leq, n\right\}$ is linearly independent.

2. Let $V := c_0$ denote the set of all sequences in $\mathbb{K}$ which converge to 0. For $j \in \mathbb{N}$, let e_j denote the element of S whose j-th term is equal to 1 and all other terms are equal to 0. Then the set $S := \left\{e_j : j \in \mathbb{N}\right\}$ is linearly independent. Next, let $S_1 := S \cup \{e\}$, where the nth entry of the sequence e is equal to $1/n$ for $n \in \mathbb{N}$. Then the set S_1 is also linearly independent since e is not a (finite) linear combination of elements of S.

3. Let $V := \mathbb{K}[x]$ denote the set of all polynomials in the indeterminate x with coefficients in $\mathbb{K}$. Then the set $S := \{x^j : j = 0, 1, 2, \ldots\}$ is linearly independent.
Next, let $S_1 := S \cup \{p\}$, where $p \in \mathbb{K}[x]$. Then the set S_1 is linearly dependent since $p \in \operatorname{span} S$.

4. Let $V := C[-\pi, \pi]$. For $n \in \mathbb{N}$, let

$$u_n(t) := \cos nt \quad \text{and} \quad v_n(t) := \sin nt \quad \text{for } t \in [-\pi, \pi].$$

Then the set $S := \{u_1, u_2, \ldots\} \cup \{v_1, v_2, \ldots\}$ is linearly independent. (Note that the zero element of this vector space is the function having all its values on $[-\pi, \pi]$ equal to 0.) To prove this, use the idea that if $\alpha \cos t + \beta \sin t = 0$, then by differentiating, we also have $-\alpha \sin t + \beta \cos t = 0$.

Next, let $S_1 := S \cup \{w\}$, where $w(t) := t$ for $t \in [-\pi, \pi]$. Then the set S_1 is also linearly independent, since $w(\pi) \neq w(-\pi)$, and so $w \notin \operatorname{span} S$.

Definition

A vector space V is said to be **finite dimensional** *if there is a finite subset S of V such that $V = \operatorname{span} S$; otherwise the vector space V is said to be* **infinite dimensional**.

If a vector space V is infinite dimensional, then V is larger than the span of any finite subset of V, and so V must contain an infinite linearly independent subset. Conversely, if V contains an infinite linearly independent subset, then V must be infinite dimensional.

Examples: Let $n, m \in \mathbb{N}$. The vector spaces $\mathbb{K}^{n\times 1}$, $\mathbb{K}^{1\times n}$ and $\mathbb{K}^{m\times n}$ are finite dimensional, and so is the vector space $\mathcal{P}_n$ of all polynomials in the indeterminate x having degree less than or equal to n. But the vector spaces $\mathbb{K}[x]$, $C[-\pi, \pi]$, c, and c_0 are infinite dimensional.