Lecture 6: Least Squares + Subspaces

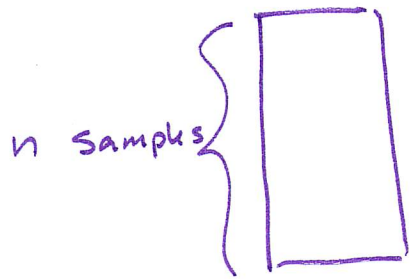
have features $\underline{x}_i \in \mathbb{R}^p$ and labels $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$

want to predict label $\hat{y}_i = \langle \underline{x}_i, \underline{w} \rangle = \underline{x}_i^T \underline{w}$

want w to minimize $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \| \underline{y} - \hat{\underline{y}} \|_2^2$

$\hat{\underline{y}} = X \underline{w}$, where $X = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]^T$

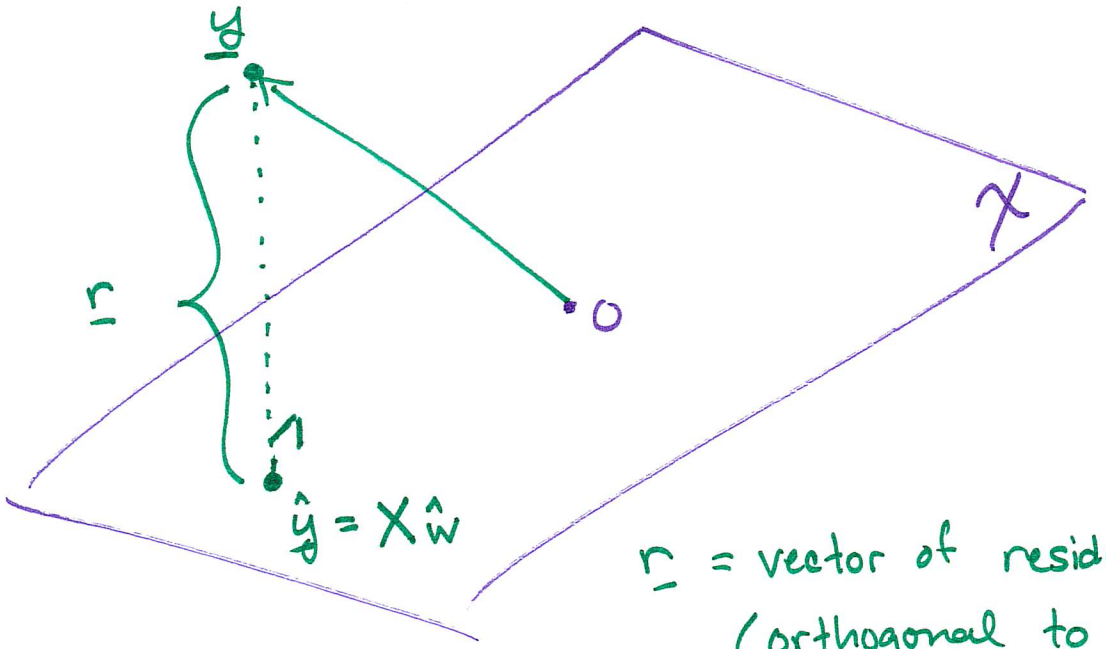
$\hat{\underline{y}}$ = weighted sum of the columns of X



$X \in \mathbb{R}^{n \times p}$

Last time, assume $n \geq p$ and p columns of X are linearly independent.

$$\hat{\underline{w}} = \arg\min_{\underline{w}} \underbrace{\|\underline{y} - X\underline{w}\|_2^2}_{\hat{\underline{y}}} \Rightarrow \hat{\underline{w}} = (X^T X)^{-1} X^T \underline{y}$$



χ = span (columns of X)

= all vectors that can be written as weighted sum of columns of X

$\underline{r}$ = vector of residuals $\in \mathbb{R}^n$

(orthogonal to χ)

$\underline{r}^T X = 0$, or $X^T \underline{r} = 0$

Review of vector calculus:

$y = f(\underline{x})$ maps $\underline{x} \in \mathbb{R}^n$ to $y \in \mathbb{R}$

$$\nabla_x f = \begin{bmatrix} df/dx_1 \\ df/dx_2 \\ \vdots \\ df/dx_n \end{bmatrix}$$

a) $f(\underline{x}) = \langle \underline{c}, \underline{x} \rangle = x^T c = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n$

$$\nabla_x f = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \underline{c}$$

b) $f(\underline{x}) = \|\underline{x}\|^2 = \underline{x}^T \underline{x}$

$= x_1^2 + x_2^2 + \cdots + x_n^2$

$$\nabla_x f = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2x$$

($f(\underline{x}) = x^T Q x$ where $Q = I$)

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

c) $f(\underline{x}) = \underline{x}^T Q \underline{x}$

assume Q is symmetric: $Q = Q^T = \frac{1}{2}(Q + Q^T)$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i Q_{ij} x_j$$

$$\left[\nabla_x f\right]_k = df/dx_k \Rightarrow \frac{d}{dx_k} x_i Q_{ij} x_j = \begin{cases} 2Q_{ij} x_i & i = j = k \\ Q_{ij} x_j & i = k,\ i \neq j \\ Q_{ij} x_i & j = k,\ i \neq j \end{cases}$$

$$\nabla_x f = (Q + Q^T) x = 2Q x$$

d) $\hat{w} = \arg\min_{w} \underbrace{\|y - Xw\|_2^2}$

$$f(w) = y^T y - 2w^T X^T y + w^T X^T X w$$

$X^T X$ = Q

$$\nabla_w f(w) = -2X^T y + 2X^T X w = 0$$

$$\Rightarrow \hat{w} = (X^T X)^{-1} X^T y$$

key $X^T X$ is positive-definite : $w^T (X^T X) w > 0$ for all $w \neq 0$

A matrix Q is positive-definite (p.d.) $\iff Q \succ 0$

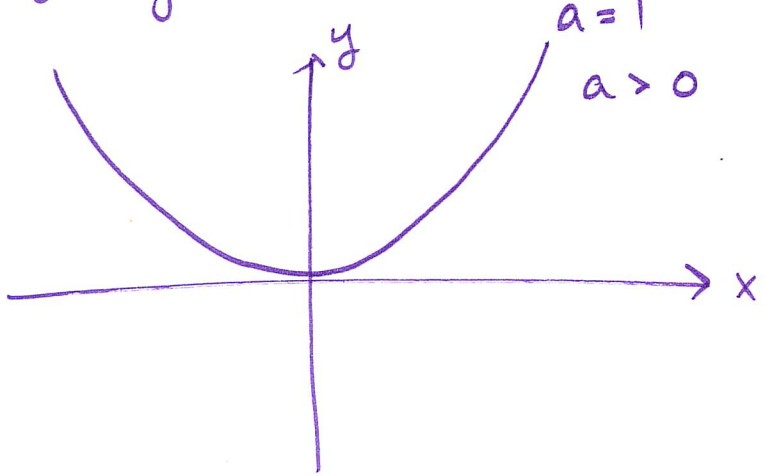
if $\underline{x}^T Q \underline{x} > 0$ for all $\underline{x} \neq 0$

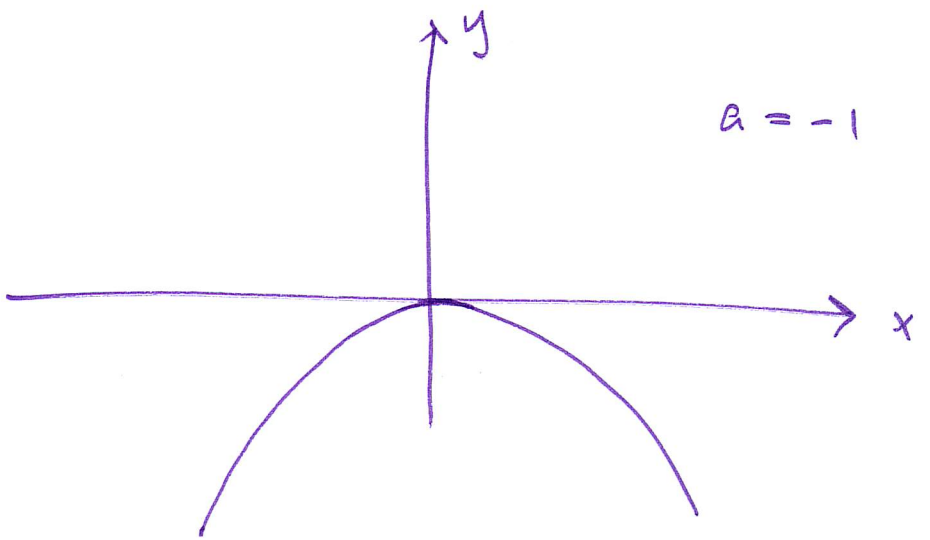
A matrix Q is positive semi-definite (psd) $\iff Q \succcurlyeq 0$

if $\underline{x}^T Q \underline{x} \geq 0$ for all $x \neq 0$

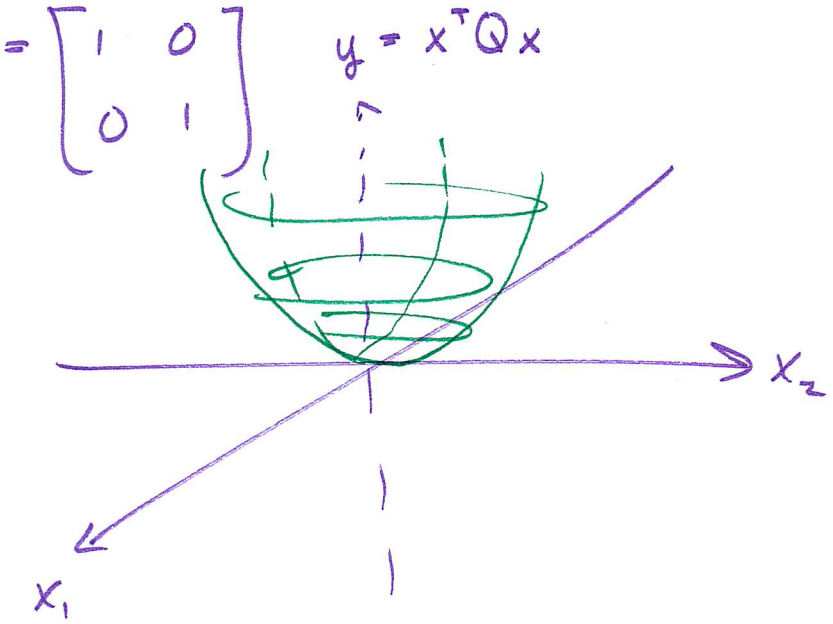
do not mean all elements of $Q \geq 0$

e.g. $y = ax^2$





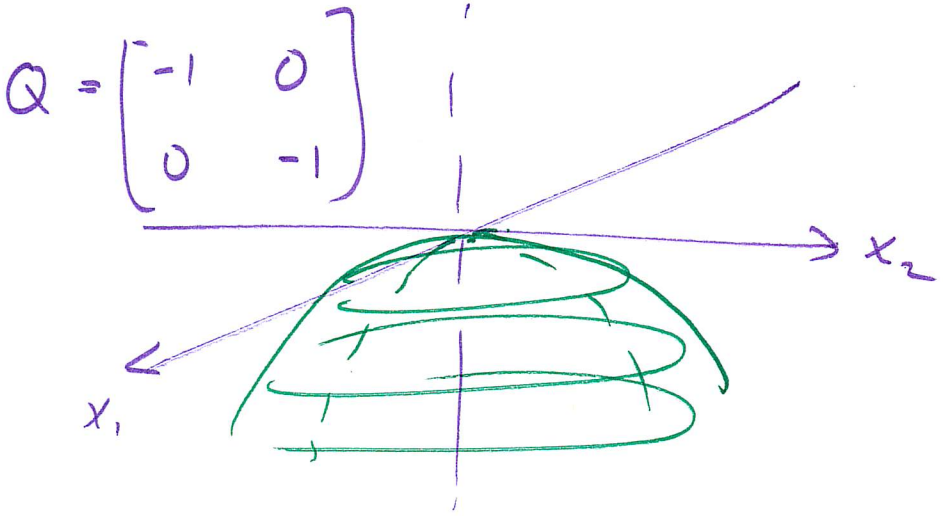
$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



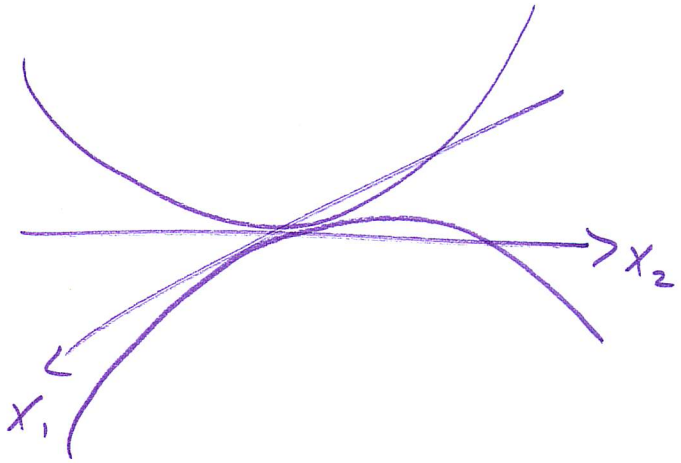
Q is pos.def.

$x^T Q x = x_1^2 + x_2^2$

$Q = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$



$Q = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$



Properties of Positive Definite Matrices:

1) if $P \succ 0$ and $Q \succ 0$, then $P + Q \succ 0$

2) if $Q \succ 0$ and $\alpha > 0$, then $\alpha Q \succ 0$

$x^T Q x > 0$, then $x^T (\alpha Q) x = \alpha (x^T Q x) > 0$

3) for <u>any</u> A, $A^T A \succeq 0$ and $A A^T \succeq 0$

if columns of A are linearly independent, then $A^T A \succ 0$

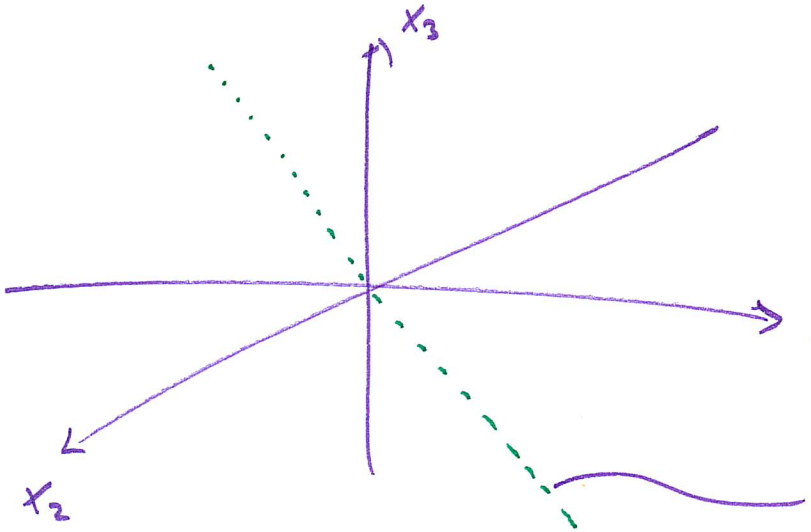
$x^T A^T A x \geq 0$

4) if $A \succ 0$ then A^{-1} exists

5) $A \succ B$ means $A - B \succ 0$

Subspaces — very common in machine learning

e.g. $\underline{x}_i \in \mathbb{R}^3$



$\underline{x}_i$'s are points on a 1-d subspace of $\mathbb{R}^3$

$$S = \{ \underline{x} \in \mathbb{R}^3 : x_1 = x_2 = x_3 \}$$

$$\underline{x} = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \Rightarrow S = \text{span}\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

$$\text{Basis for } S = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} / \sqrt{3} \right\}$$

$$\dim(S) = 1$$

A set of points $S \subseteq \mathbb{R}^n$ (every point in S is also in $\mathbb{R}^n$)

is a **subspace** if

(i) $0 \in S$ (S contains origin)

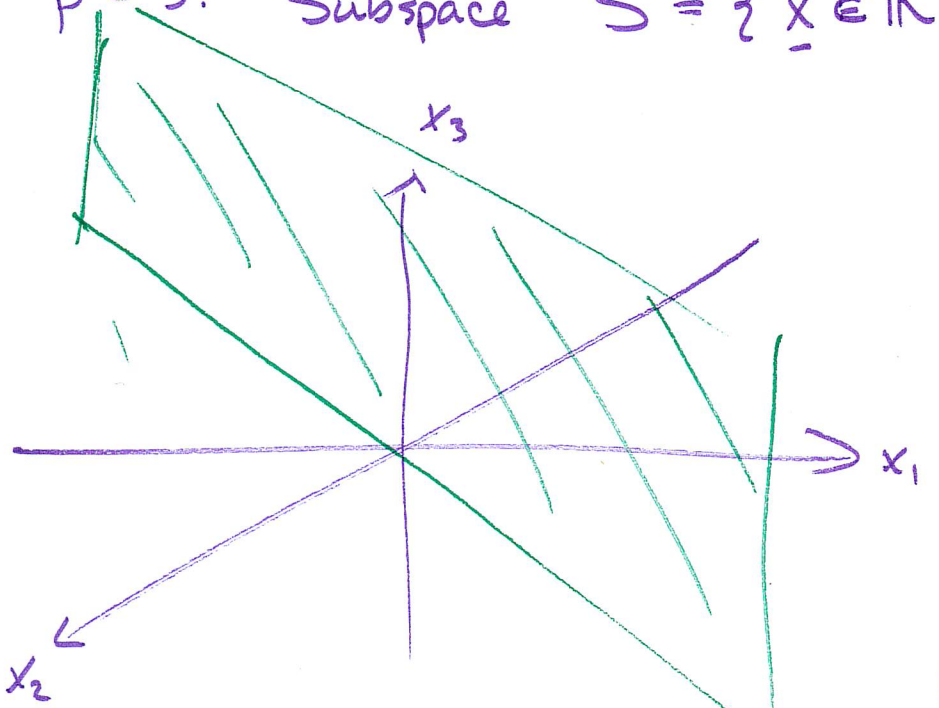
(ii) if $\underline{x}, \underline{y} \in S$, then $\underline{x} + \underline{y} \in S$

(iii) if $\underline{x} \in S$, $\alpha \in \mathbb{R}$, then $\alpha \underline{x} \in S$

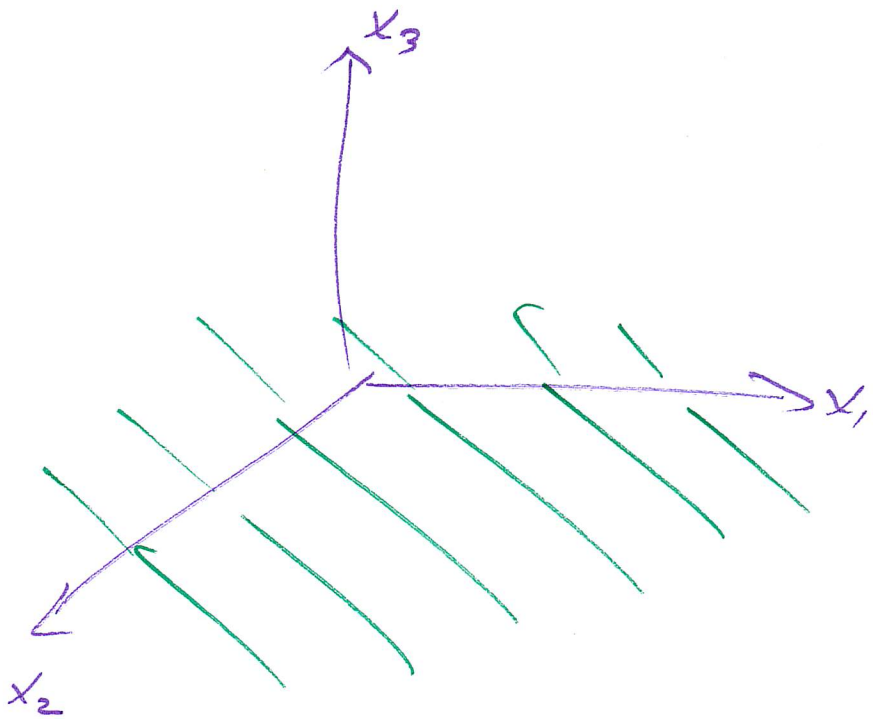
ex 1 $n = 3$. Subspace $S = \{ \underline{x} \in \mathbb{R}^3 : x_1 = x_2 = x_3 \}$

$$\underline{x} + \underline{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix} \in S \qquad \alpha \underline{x} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \alpha x_3 \end{bmatrix} \in S$$

ex 2 $p = 3$. Subspace $S = \{ \underline{x} \in \mathbb{R}^3 : x_1 = x_2 \}$



ex 3 $S = \{ \underline{x} \in \mathbb{R}^3 : x_3 = 0 \}$



ex 41

we are given points $\underline{x}_1, \underline{x}_2, \underline{x}_3 \in \mathbb{R}^n$

$$\text{Span}\{\underline{x}_1, \underline{x}_2, \underline{x}_3\} = \{ y \in \mathbb{R}^n : y = \alpha_1 \underline{x}_1 + \alpha_2 \underline{x}_2 + \alpha_3 \underline{x}_3 \text{ for some } \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R} \}$$

$$= \text{set of vecs} = \text{linear combo. of } \underline{x}_1, \underline{x}_2, \underline{x}_3$$

is a subspace

if $X = [\underline{x}_1 \ \underline{x}_2 \ \underline{x}_3]$, then $\underline{\text{range}(X)} = \text{span}(\text{cols}(X))$

eg.

$$X = \underbrace{\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}}_{\text{user}} \Big\} \text{movie} = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \underbrace{\begin{bmatrix} & & \\ & & \end{bmatrix}}_{T}$$

W = r weights for each user

T = r representative taste profiles

model: each user's taste profile lies in a subspace spanned by cols of T (representative taste profiles)

How to represent a subspace?

- as the span of a set of vectors

  ⇒ can be hard to interpret, hard to compute with, redundant

- as the span of a set of linearly indep. vectors

- as the span of a set of orthonormal vectors (BASIS of subspace)

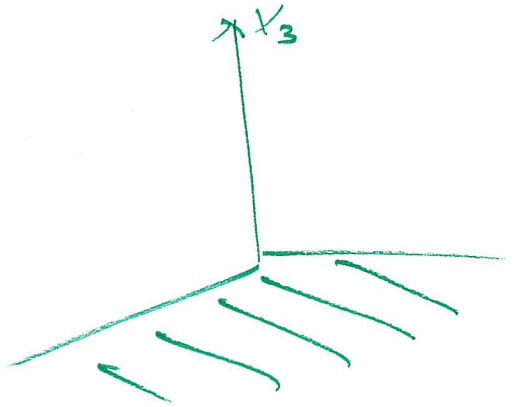
$$S = \text{span}\{\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_r\}$$

where $\underline{u}_i$'s are orthonormal

- orthogonal: $\underline{u}_i^T \underline{u}_j = 0$ if $i \neq j$
- normalized $\underline{u}_i^T \underline{u}_i = \|\underline{u}_i\| = 1$ for all i

$\Rightarrow$ dimension ("size") of $S = r$

e.g. $S = \{ \underline{x} \in \mathbb{R}^3 : x_3 = 0 \}$



all $\underline{x} \in S$ have form

$$\underline{x} = \begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\text{Basis} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

$\dim(S) = 2$ ⟷ $\text{rank}(U) = 2$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \text{basis matrix}$$