Lecture 7: Orthogonal Bases

$$\underline{y}^T X \underline{w} = \underline{w}^T X^T \underline{y} \quad \Longleftarrow \quad \langle \underline{b}, \underline{y} \rangle = \underline{y}^T \underline{b} = \sum_{i=1}^{n} y_i b_i = \underline{b}^T \underline{y}$$

Let $\underline{b} = X\underline{w}$

$$\underline{y}^T X \underline{w} = (X\underline{w})^T \underline{y} = \underline{w}^T X^T \underline{y}$$

$$b = Ax$$

$$\Downarrow$$

$$\underset{\text{calories}}{\underset{\uparrow}{\underline{y}}} = \underset{\text{fat, carbs, etc}}{\underset{\uparrow}{X}} \, \underline{w} \leftarrow \text{weights}$$

$$X = \begin{bmatrix} | & & | \\ \underline{x}_1 & \cdots & \underline{x}_p \\ | & & | \end{bmatrix}$$

model $\hat{\underline{y}} = \underline{\underline{w_1 \underline{x}_1}} + w_2 \underline{x}_2 + \cdots + w_p \underline{x}_p$

what if $\underline{x}_1 = \underline{x}_2 + \underline{x}_3$? $\quad \underline{x}_2 = \underline{x}_1 - \underline{x}_3$

$w_1 \underline{x}_1 = w_1 \underline{x}_2 + w_1 \underline{x}_3 \qquad w_2 \underline{x}_2 = w_2 \underline{x}_1 - w_2 \underline{x}_3$

model : $y = w_1 \underline{x}_1 + w_2 \underline{x}_2 + w_3 \underline{x}_3 + \cdots$

$$= \underline{0 \cdot \underline{x}_1} + \underline{(w_1 + w_2)}\, \underline{x}_2 + \underline{(w_1 + w_3)}\, \underline{x}_3 + \cdots$$

$$= \underline{\underline{(w_1 + w_2)}}\, \underline{x}_1 + 0\, \underline{x}_2 + (w_3 - w_2)\, \underline{x}_3 + \cdots$$

$\hat{\underline{w}} = (X^T X)^{-1} X^T \underline{y}$ ← if cols of X are linearly dep. then $X^T X$ is not invertible

impossible to tell which features are "significant" predictors of $\underline{y}$.

what if $\tilde{X} = \begin{bmatrix} \vert & \vert & \vert & & \vert \\ \underline{x}_1/10^3 & \underline{x}_2 & \underline{x}_3 & \cdots & \underline{x}_p \\ \vert & \vert & \vert & & \vert \end{bmatrix}$

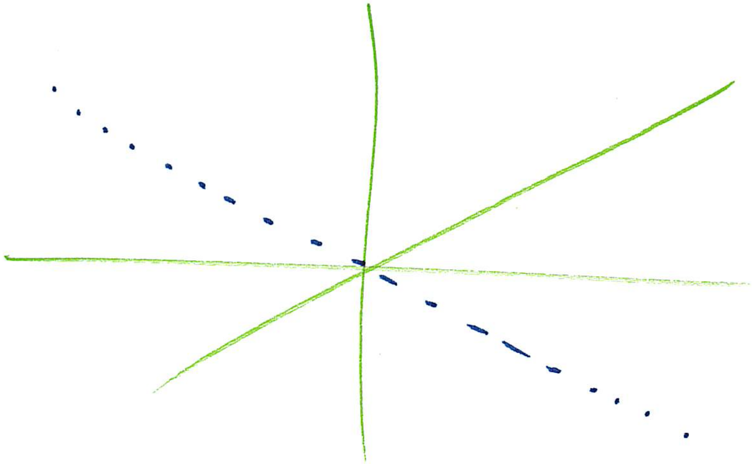
when choose weights $\underline{w}$, w_1 to be 10^3 times as big

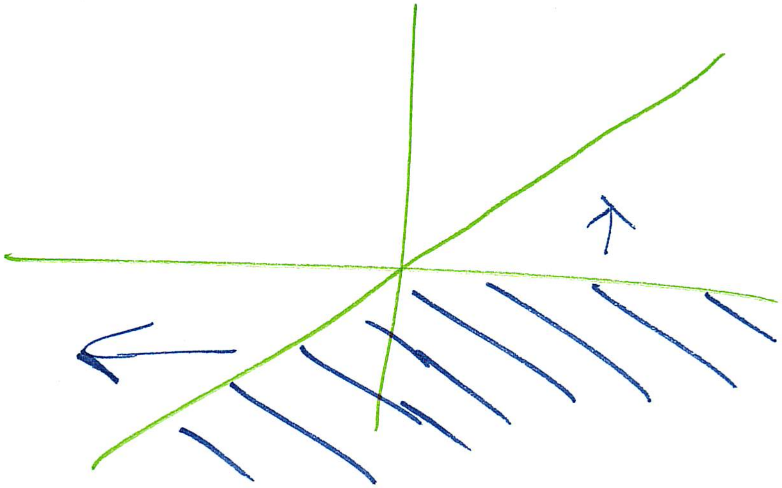
$$\underline{x}_1 w_1 = \frac{\underline{x}_1}{10^3}(w_1 \cdot 10^3)$$

If we are handed a matrix $X \in \mathbb{R}^{n \times p}$, how can we prevent this scenario?

Recall, $\hat{y}$ = weighted sum of cols of X

$$\in \underbrace{\text{span}(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p)}_{\text{this is a subspace}}$$





Give $X = [\underline{x}_1 , \ldots \ \underline{x}_p]$, how can we represent the corresponding subspace ?

Option 1 : use X $\Longrightarrow$ LS can be hard to interpret

Option 2 : use an $\boxed{\text{orthonormal basis}}$ for subspace.

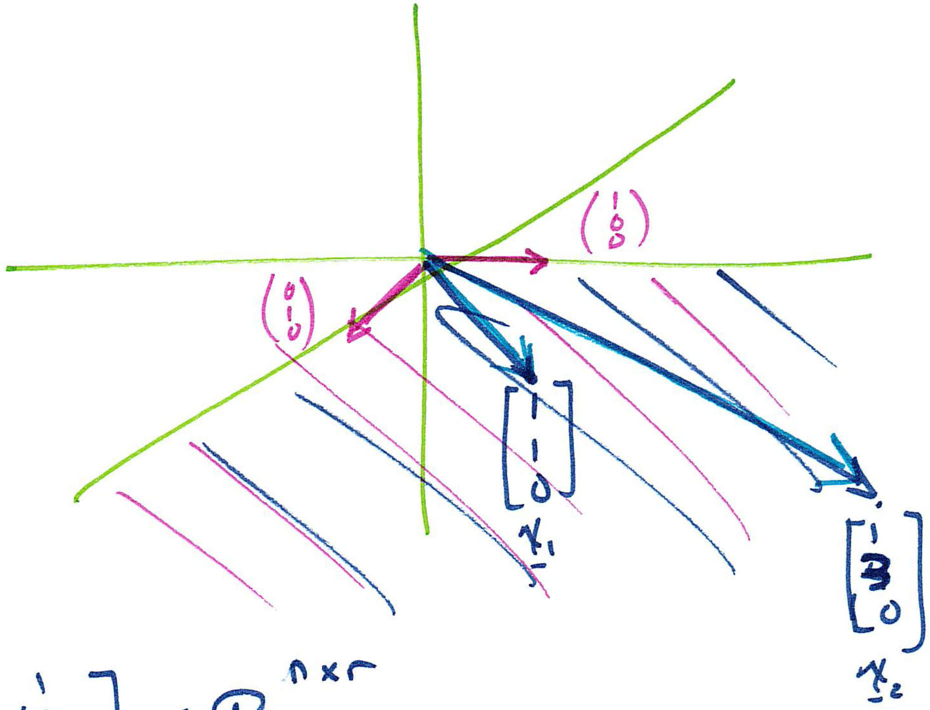
o.n. basis $= \{ \underline{u}_1 , \ldots , \underline{u}_r \}$ s.t. subspace $S = \text{span} \{ \underline{u}_1 , \ldots , \underline{u}_r \}$

and $$\underline{u}_i^T \underline{u}_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

$$\text{span} (\underline{u}_1 , \ldots , \underline{u}_r) = \text{span} (\underline{x}_1 , \ldots , \underline{x}_p)$$

$r \leq \min(n,p)$

↑ rank of subspace
= # Lin. Inp. cols of **X**



Let $U = \text{basis matrix} = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_r \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times r}$

Before

$$\hat{\underline{w}} = \arg\min_{\underline{w}} \| y - X\underline{w} \|_2^2$$

$\hat{\underline{w}}$ can be hard to interpret

if we get a new sample $\underline{x} \in \mathbb{R}^p$ and want to predict y, then

$$\hat{y} = \langle \underline{x}, \hat{\underline{w}} \rangle$$

$$\hat{w} = \underbrace{(X^TX)^{-1}}_{} X^T y$$

↳ can be hard to compute

Now

$$\hat{\underline{v}} = \arg\min_{v} \| y - U\underline{v} \|_2^2$$

$\hat{\underline{v}}$ can be easier to interpret

- how to compute $\hat{\underline{v}}$?
- how to get U give X
- if we get a new sample $\underline{x} \in \mathbb{R}^p$, how do we predict y?

Some properties of basis matrix $U = [\underline{u}_1, \dots, \underline{u}_r] \in \mathbb{R}^{n \times r}$

i) $U^T U = I$

$$TW = \begin{bmatrix} t_1^T w_1 & t_1^T w_2 & \cdots \\ t_2^T w_1 & \vdots & \\ \vdots & & \ddots \end{bmatrix}$$

$$\boxed{\underline{u}_i^T \underline{u}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}}$$

ii) if $\underbrace{U}_{n \times r}$ and $\underbrace{V}_{r \times p}$ are both orthogonal, then

UV is also orthogonal

iii) "length preserving" $\|Uv\|_2 = \|v\|_2$

$$\|Uv\|_2^2 = (Uv)^T(Uv) = v^T U^T U v = v^T v = \|v\|_2^2$$

⑦

$$\hat{\underline{v}} = \arg\min_{v} \| \underline{y} - U\underline{v} \|_2^2$$

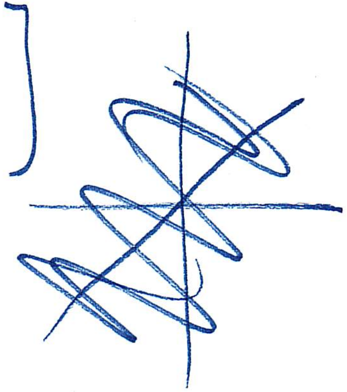
need $\hat{\underline{v}}$ to satisfy $\underbrace{U^T U}_{=I} \hat{\underline{v}} = U^T \underline{y} \Rightarrow \hat{\underline{v}} = U^T \underline{y}$

easy!

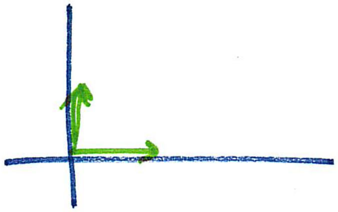
how to take X and get orthonormal basis U?

Gram-Schmidt Orthogonalization

ex. $\underline{x}_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\underline{x}_2 = \begin{bmatrix} a \\ b \end{bmatrix}$ $\quad U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$$\underline{u}_1 = \frac{\underline{x}_1}{\|\underline{x}_1\|_2} = \frac{\underline{x}_1}{2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$



write $\underline{x}_2$ as weighted sum of $\underline{u}_1$ + residual

$$\underline{x}_2 = a \cdot \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{\underline{u}_1} + \begin{bmatrix} 0 \\ b \end{bmatrix} \Rightarrow \text{resid } \underline{x}_2' = \begin{bmatrix} 0 \\ b \end{bmatrix} \Rightarrow \underline{u}_2 = \frac{\underline{x}_2'}{\|\underline{x}_2'\|_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

ex. $\underline{x}_1 = \begin{bmatrix}1\\1\\0\end{bmatrix}, \quad \underline{x}_2 = \begin{bmatrix}1\\3\\0\end{bmatrix}$

$$\underline{u}_1 = \frac{\underline{x}_1}{\|\underline{x}_1\|_2} = \frac{\underline{x}_1}{\sqrt{2}} = \begin{bmatrix}1/\sqrt{2}\\1/\sqrt{2}\\0\end{bmatrix}$$

write $\underline{x}_2$ = weighted $\underline{u}_1$ + $\underbrace{\text{residual}}_{\text{make as small as possible}}$

what is the least-squares fit of $\underline{u}_1$ to $\underline{x}_2$?

$$\underset{w}{\arg\min} \| \underline{x}_2 - \underline{u}_1 w \|_2^2$$

$$\hat{w} = \underline{u}_1^T \underline{x}_2$$

$$\underline{x}_2 = \hat{w}\,\underline{u}_1 + \text{resid}$$

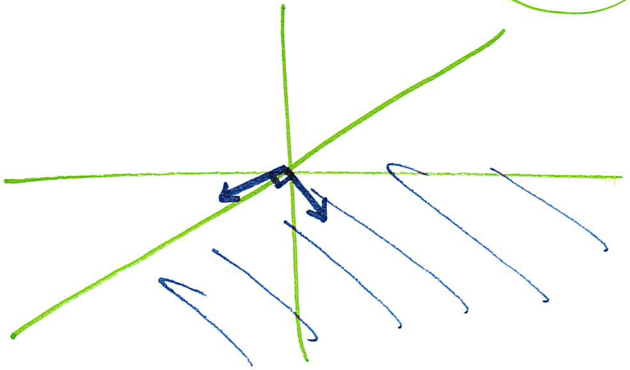
$$= \underline{u}_1(\underline{u}_1^T \underline{x}_2) + \text{resid}$$

$$\text{resid} = \underline{x}_2' = \underline{x}_2 - \underline{u}_1(\underline{u}_1^T \underline{x}_2)$$

$$\underline{u}_1^T \underline{x}_2 = 1/\sqrt{2} + 3/\sqrt{2} = 4/\sqrt{2}$$

$$\underline{x}_2' = \begin{bmatrix}1\\3\\0\end{bmatrix} - 4/\sqrt{2}\begin{bmatrix}1/\sqrt{2}\\1/\sqrt{2}\\0\end{bmatrix} = \begin{bmatrix}-1\\1\\0\end{bmatrix}$$

$$\underline{u}_2 = \frac{\underline{x}_2'}{\|\underline{x}_2'\|_2} = \begin{bmatrix}-1/\sqrt{2}\\1/\sqrt{2}\\0\end{bmatrix}$$



Gram-Schmidt Orthogonalization Algorithm

input: $X = [\underline{x}_1 \quad \underline{x}_2 \quad \cdots \quad \underline{x}_p] \in \mathbb{R}^{n \times p}$

output: $U = [\underline{u}_1 \quad \underline{u}_2 \quad \cdots \quad \underline{u}_r] \in \mathbb{R}^{n \times r}$

$r = \text{rank}(X)$
$\leq \min(n, p)$

initialize: $\underline{u}_1 = \underline{x}_1 / \|\underline{x}_1\|_2$

for $j = 2, 3, \ldots, p$

$\underline{x}_j' =$ all ~~inf~~ components of $\underline{x}_j$ $\underline{\underline{\text{not}}}$ represented by

$\underline{u}_1, \ldots, \underline{u}_{j-1}$

$$= \underline{x}_j - \sum_{i=1}^{j-1} \underbrace{(\underline{u}_i^T \underline{x}_j)}_{\text{least squares weight for } \underline{u}_i} \underline{u}_i$$

$$\underline{u}_j = \begin{cases} \dfrac{\underline{x}_j'}{\|x_j'\|_2} & \text{if} \quad x_j' \neq 0 \\ 0 & \text{if} \quad x_j' = 0 \end{cases}$$

~~end for~~

end

ex.
$$X = \begin{array}{c} \begin{bmatrix} -1 & 1 & 2 \\ 1 & -1 & 2 \end{bmatrix} \\ \begin{array}{ccc} \underline{x}_1 & \underline{x}_2 & \underline{x}_3 \end{array} \end{array}$$

$$\underline{u}_1 = \frac{x_1}{\|x_1\|_2} = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$\underline{x}_2' = \underline{x}_2 - (u_1^T x_2)\, u_1$$

$$= \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \left(+2/\sqrt{2}\right)\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \underline{u}_2 = 0$$

$$\underline{x}_3' = x_3 - (u_1^T x_3)u_1 - (u_2^T x_3)u_2$$

$$= x_3 - 0 \cdot u_1 \qquad 0$$

$$= x_3$$

$$u_3 = \frac{x_3}{\|x_3\|_2} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$