Week 02 - Part 3

Review:
- Deriving $\underline{w}_{LS}$ by finding the gradient an setting it to zero
- Deriving $\underline{w}_{LS}$ by (pseudo)-solving the system of linear equations $\underline{y} = X\,\underline{w}$.

Today:
- Deriving $\underline{w}_{LS}$ with Geometric interpretations
- Regularized least squares
- Non-linear transformation

Recall:

- Least square solution: $\underline{w}_{ls} = X^{\dagger}\underline{y} = (X^T X)^{-1} X^T \underline{y}$

- prediction by $\underline{w}_{ls}$: $\hat{\underline{y}}_{ls} = X\underline{w}_{ls} = XX^{\dagger}\underline{y}$

  - It's like we take $\underline{y}$ and with a projection matrix transforming it into $\hat{\underline{y}}_{ls}$.

  - $XX^{\dagger}$ is a projection matrix.

- This observation leads us into geometric interpretation of Least squares.

Geometric Interpretation of Least Squares

- Observe that $\hat{y} = X\underline{w} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1d} \\ x_{20} & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{N0} & x_{N1} & \cdots & x_{Nd} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$

$$= ?$$

- So, $\underline{\hat{y}}$ is linear combination of ...

- Thus, $\underline{\hat{y}}$ is in the space of all possible linear Combinations of columns of X

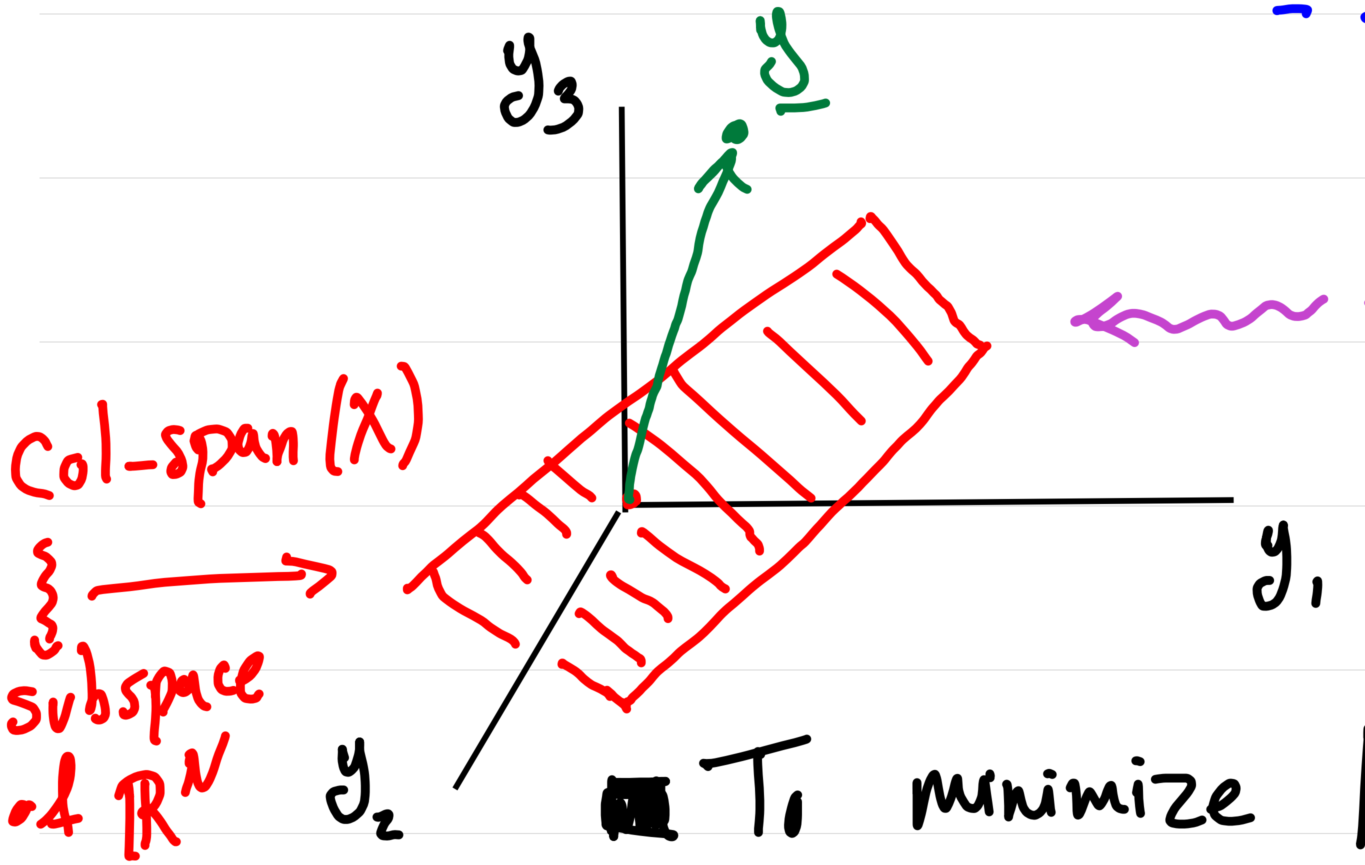
- The space of all possible linear Combination of Columns of X is Called Col-span$\{X\}$

- Let's illustrate Col-span(X) for $N=3$, $d=1$, and $X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$.

Col-span$(X) =$

- $\text{Col-span}(X) = \left\{ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} : \underbrace{-y_1 + 2y_2 - y_3 = 0}_{} \right\}$

- A plane that goes through origin.
- It is a subspace of $\mathbb{R}^N$.



- To minimize $\|\underline{y} - \underline{\hat{y}}\|$: Euclidean distance b/w $\underline{y}$ & $\underline{\hat{y}}$

- must find $\underline{\hat{y}}$ on Col-span$\{X\}$ that is closest to $\underline{y}$.

- The best $\underline{\hat{y}}$ (i.e. $\hat{y}_{LS}$) is the projection of $\underline{y}$ onto Col-span$\{X\}$.

- That means $(\underline{y} - \underline{\hat{y}}_{LS})$ must be orthogonal to any vector in Col-span$\{X\}$.

- Thus, $(\underline{y} - \underline{\hat{y}}_{LS})$ is orthogonal to every column of X.

Reminder: $\underline{a} \perp \underline{b} \iff \underline{a}^T \underline{b} = 0$

- Thus, $X^T(\underline{y} - \underline{\hat{y}}_{ls}) = \underline{0} \Longrightarrow$

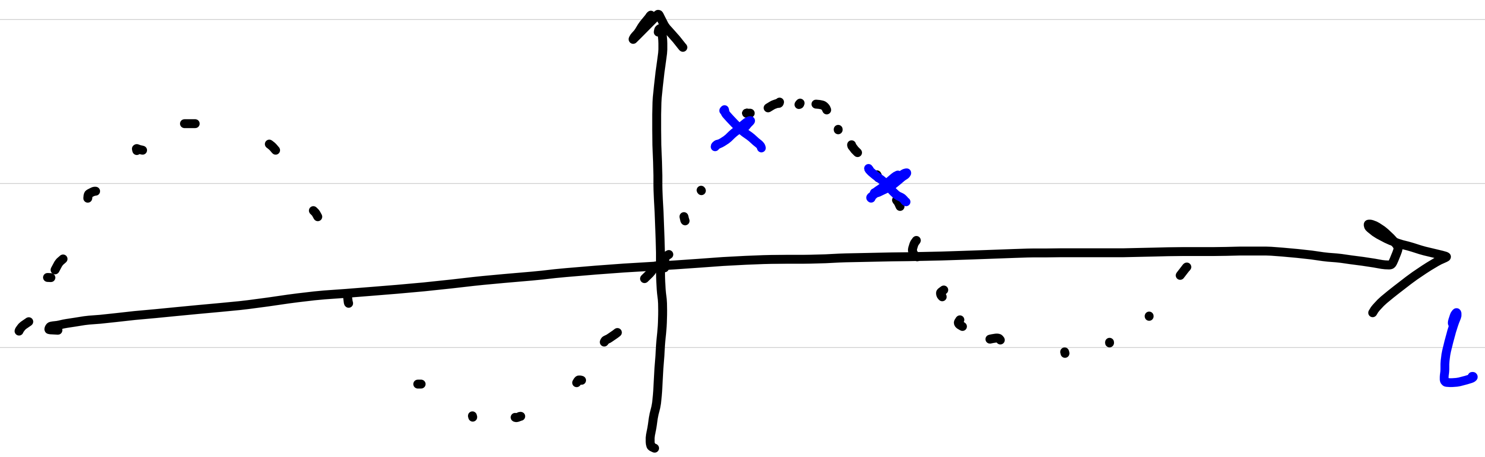
Regularized Linear Regression / Least squares

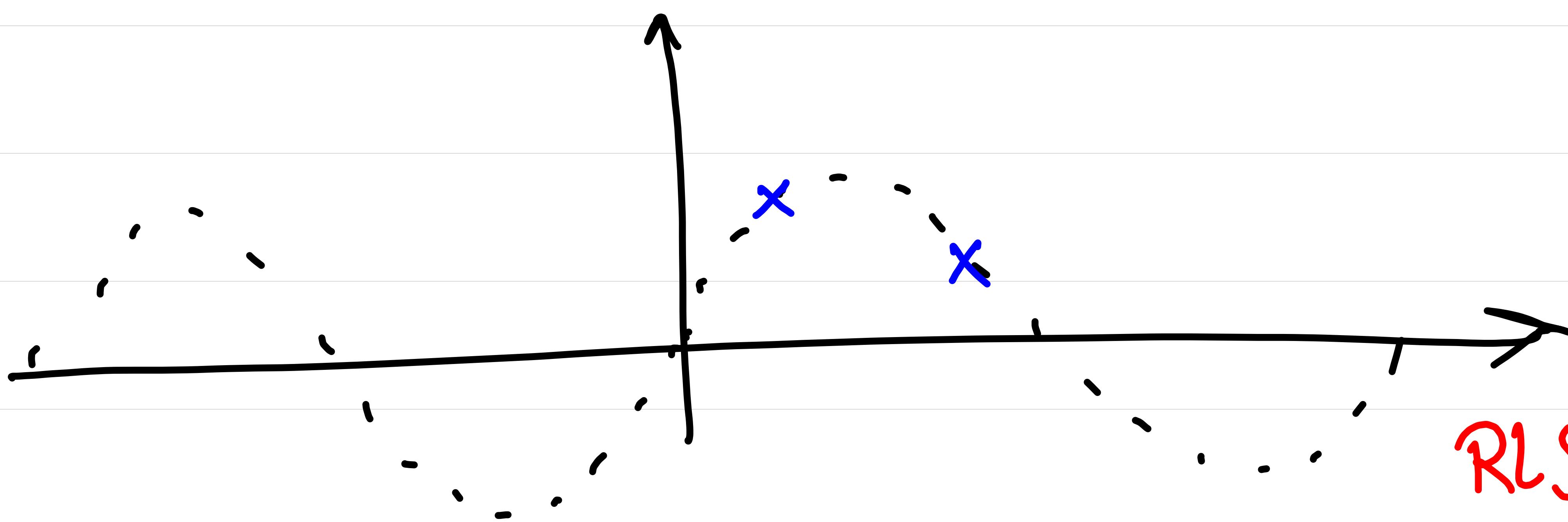
- Previously, we tried to minimize $\| X\underline{w} - \underline{y} \|^2$

- In regularized version, we minimize $\| X\underline{w} - \underline{y} \|^2 + \underbrace{\lambda \|\underline{w}\|^2}_{\text{Penalty function (against large weight)}}$

- The motivation is to avoid overfitting
  - your data is noisy
  - you do not have enough data (compared to the complexity of the target function)

■ E.g. target: $f(x) = \sin(\pi x)$





Note: ① $\lambda = 0 \Longrightarrow$ LS

② How to choose λ? validation

How do we Solve this Problem?

- We want to $\min_{\underline{w}} \|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|^2$

- Let $f(\underline{w}) = \|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|^2$

- Observe that $\nabla_{\underline{w}} f(\underline{w}) = 2X^T(X\underline{w} - \underline{y}) + 2\lambda \underline{w}$

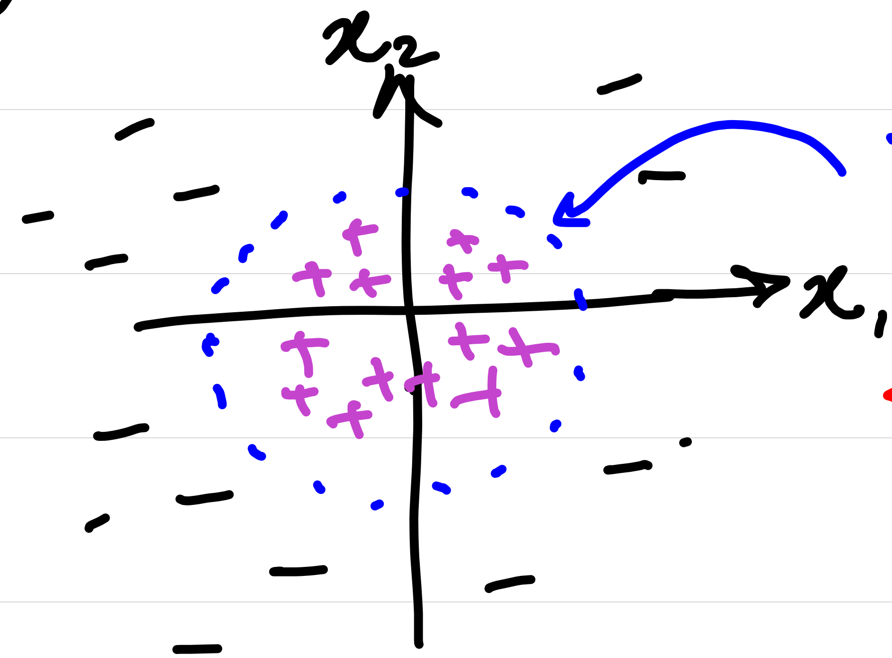
- We want $\nabla_{\underline{w}} f(\underline{w}) = 0 \Rightarrow (X^TX + \lambda I)\underline{w} = X^T\underline{y}$

$$\Rightarrow \underline{w}_{RLS} = (X^TX + \lambda I)^{-1} X^T \underline{y}$$

So far: We studied linear Models

- But in many cases linear Models are not good enough.

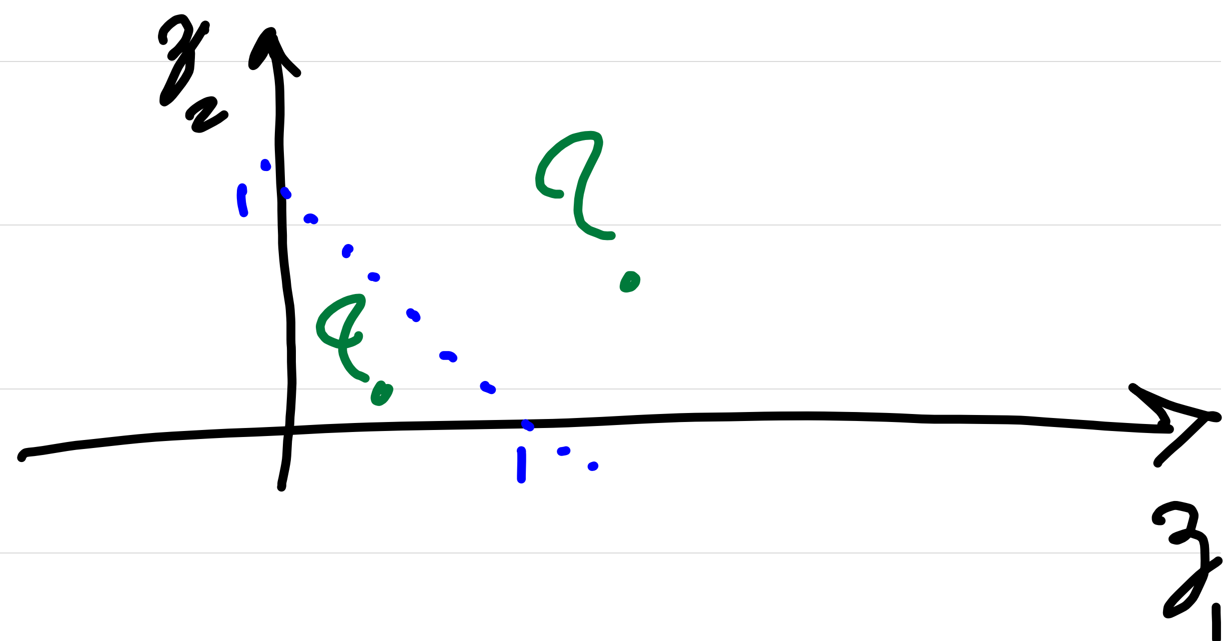
E.g.



true decision boundary $x_1^2 + x_2^2 = 1$

linear decision boundaries won't perform well.

Then define $z_1 = x_1^2$ and $z_2 = x_2^2$



The points are linearly separable in Z-space.

Suppose PLA gives you $h(\underline{z}) = \text{sign}(z_1 + z_2 - 1)$. Then, we know $g(\underline{x}) = \text{sign}(x_1^2 + x_2^2 - 1)$

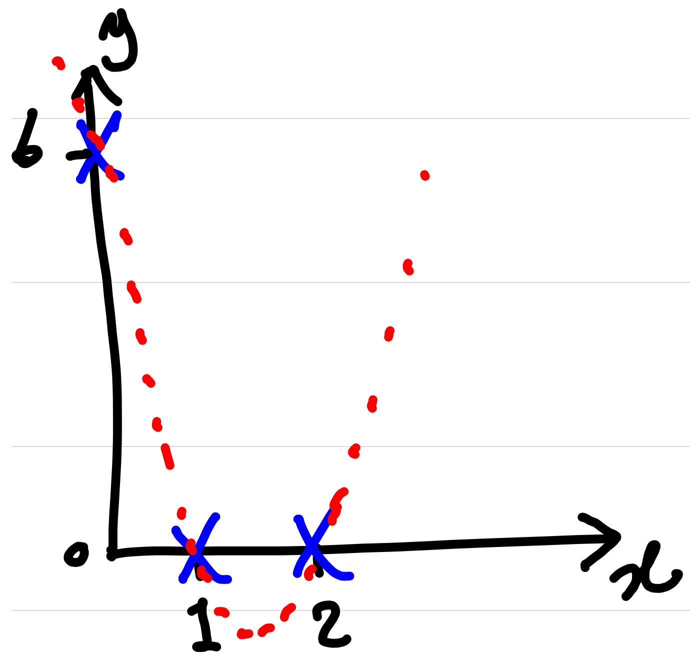
In general:

- Let $\underline{z} = \Phi(\underline{x})$ be non-linear transformation ("feature transformation")

- Let $h(\underline{z})$ be a linear classifier / regression function in $\underline{z}$ space $\left(h(\underline{z}) = \text{Sign}(\underline{w}^T \underline{z}) \text{ or } h(\underline{z}) = \underline{w}^T \underline{z} \right)$

- Then $g(\underline{x}) = h(\Phi(\underline{x}))$ is non-linear classifier in $\underline{x}$ space

E.g. Quadratic Regression



- Define $\underline{z} = (z_0 = 1,\ z_1 = x,\ z_2 = x^2)$

- $\underline{y} = \underline{w}^T \underline{z} = w_0 + w_1 z_1 + w_2 z_2$ linear in $\underline{z}$

  $= w_0 + w_1 x + w_2 x^2$ Quadratic in $\underline{x}$

- Let's find the $\underline{w}_{LS}$: $\underline{x}_1 =$, $\underline{x}_2 =$, $\underline{x}_3 =$

  $\underline{z}_1 =$, $\underline{z}_2 =$, $\underline{z}_3 =$ $\rightarrow Z =$

  $\underline{y} =$

  $\hat{y} = \underline{w}_{LS}^T \underline{z} =$

$\underline{w}_{LS} =$

$= \cdots = \Rightarrow$