Lecture 18. Regression Basics

1. Regression tasks

1) Data: $X = \{x_1, x_2, \cdots x_n\} = \{x_i\}_{i=1}^n$ a set n data samples

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix} \text{ feature vector. (ith)}$$

$y = \{\underline{y_i}\}_{i=1}^n$ targets / ground truth

y_i : continuous value, e.g., $y_i \in \underline{\mathbb{R}}$

$(x^2)' = 2x$

$(\frac{1}{2}x^2)' = \underline{x}$

$$x_i \longrightarrow \boxed{f} \longrightarrow \hat{y}_i \text{ (estimation)} \longleftrightarrow y_i$$

2) model: f: if f is a linear function, f is a linear model.

3) Loss function. $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$, $\frac{1}{2n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

$\sum_{i=1}^{n}(\hat{y}_i - y)^2$, $\frac{1}{2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$,

2. Linear Regression

Model: $f = w^T \cdot x_i + w_0 * 1 = 1 * w_0 + w_1 x_{i1} + w_2 \cdot x_{i2} + w_3 \cdot x_{i3} + \cdots w_m x_{im}$

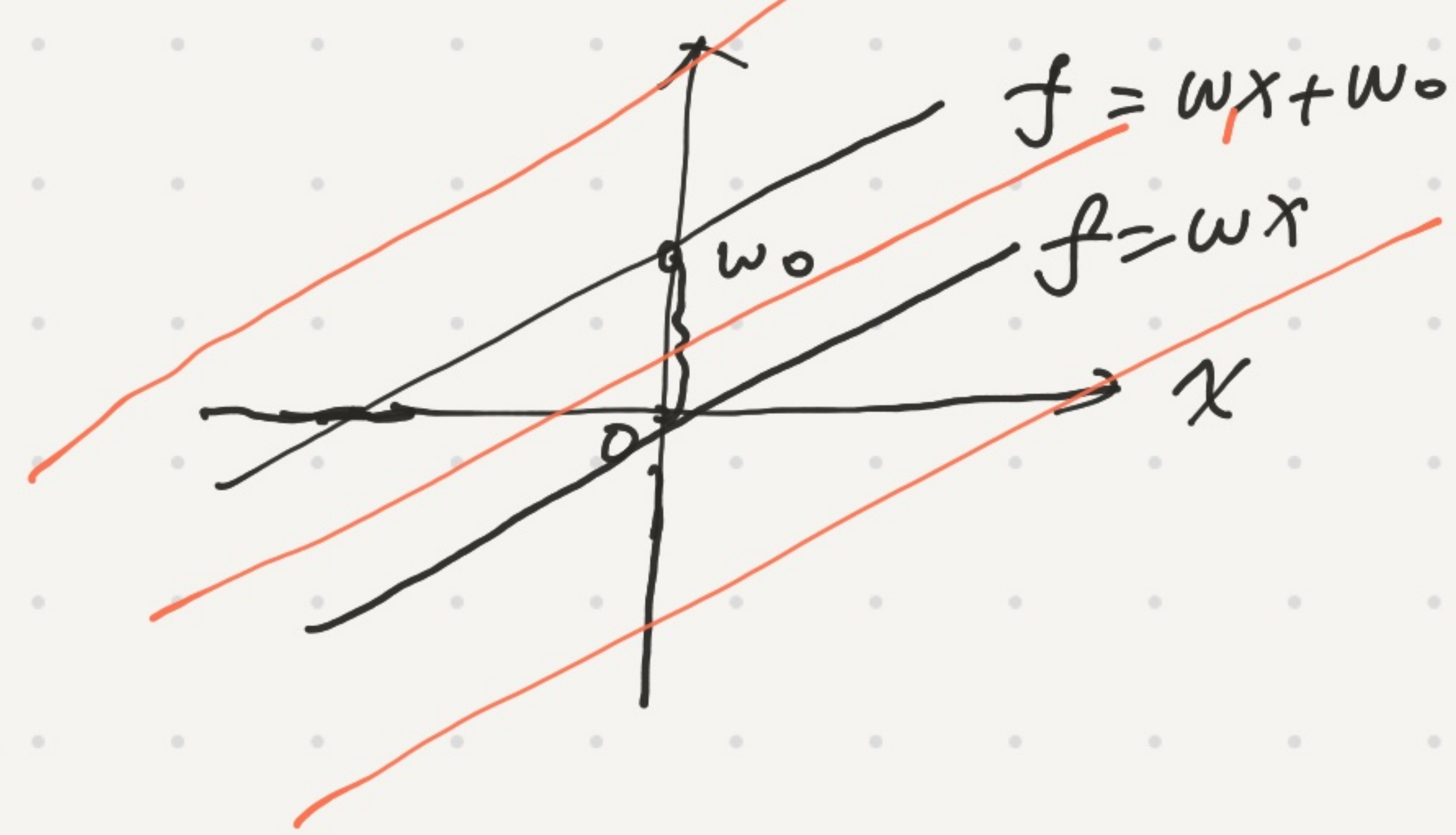
$w_0 \rightarrow$ bias

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix} \qquad w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix}$$

$\Rightarrow$ defines the contribution of each feature.

General form: $$w = \begin{pmatrix} w_0 \\ \vdots \\ w_m \end{pmatrix}_{(m+1)} \qquad x_i = \begin{pmatrix} x_{i0} = 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix}_{(m+1)}$$

$$f = w^T \cdot x_i$$

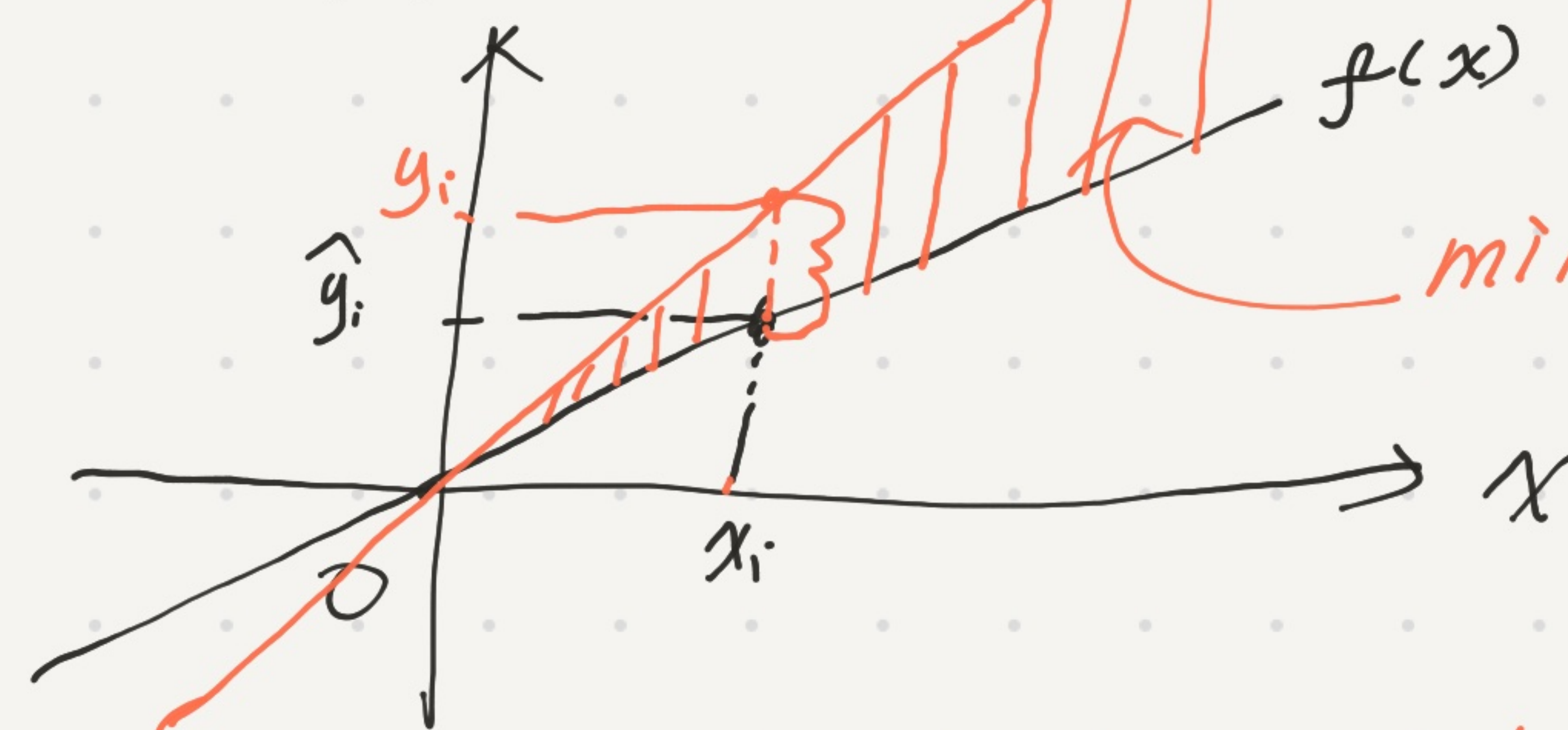


w_0: bias / intercept

$w_1, w_2, \cdots w_m$: Coef

LOSS:

$$L = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (\underline{f}(x_i) - y_i)^2$$



minimize this area.

true / actual model. (y_i)

2.1. Ordinary Least Squares (OLS) (standard Linear Regression)

Loss : $J_1 = \sum_{i=1}^{n} \underline{(\hat{y}_i - y_i)^2}$

optimization: $w^* = \arg\min_{w} J_1$

All distances $\underline{(\hat{y}_i - y_i)^2}$ contribute equally.

↳ Cause a major challenge in OLS: OLS sensitive to outliers / noise ⇒

- → Remove outliers
- → Regularization
  - Ridge
  - LASSO

3. Regularization: strategy used to model user preference

No-free-Lunce theorem: there is no machine learning algorithm that can generalize well to all application.

good test performance.

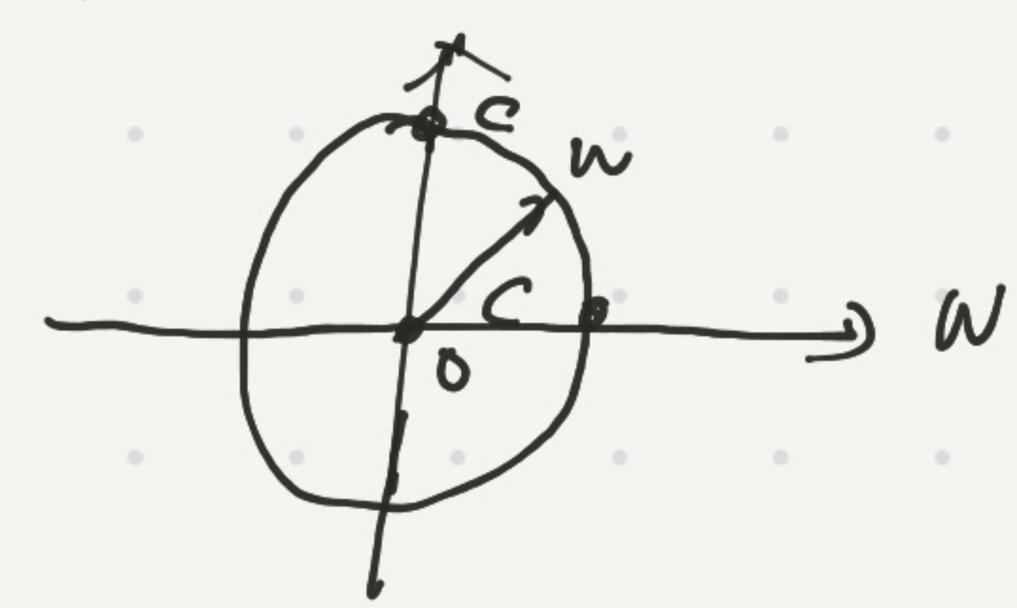
OLS: $\min J_1 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

define constraints to add preference.

$$\min J_1 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

subject to (s.t.): $\|w\|_2^2 \le C$

constraints

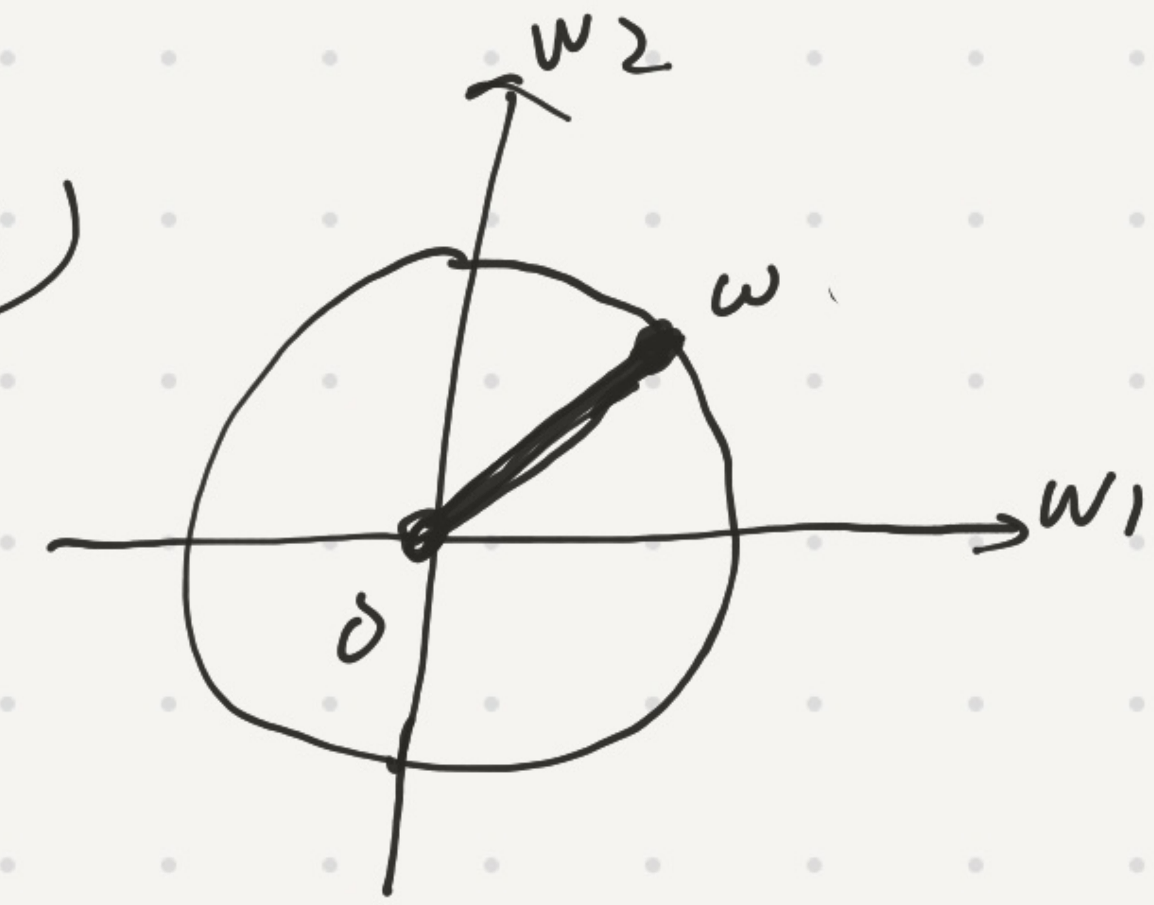


In ML, we transform all constraints to Regularizers which are terms in the loss function.

Ridge Regression

A) $J_2 = J_1 + \lambda \underbrace{\|w\|_2^2}_{\text{regularizer; length of } w} \rightarrow (L_2 \text{ norm})$

(OLS)



Our preference in this example is small w.

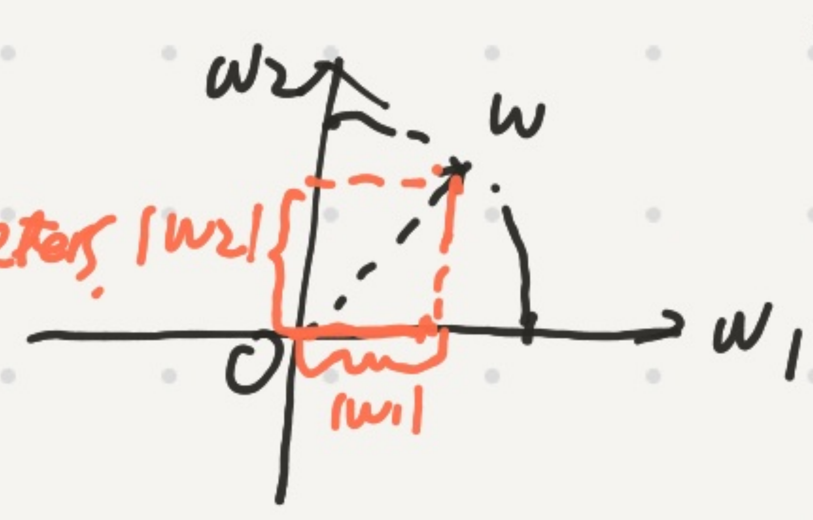
with this $\|w\|_2^2$, ridge regression will shrink the model parameters.

B) Least absolute shrinkage and selection operator (LASSO)

$$J_3 = J_1 + \lambda \|w\|_1 \rightarrow L_1 \text{ norm.}$$

$$\|w\|_1 = |w_0| + |w_1| + |w_2| + \cdots + |w_m| = \sum_{i=1}^{m} |w_i|$$

LASSO will produce more zero parameters.



c) Elastic Net

$$J_4 = J_1 + \underbrace{\lambda_1 \|W\|_2^2}_{L_2} + \underbrace{\lambda_2 \|W\|_1}_{L_1}$$

λ_1, λ_2 : constants, define the contribution of each term.

In a Loss function, if we have M terms, we will M-1 constants. (e.g., $\lambda_1, \lambda_2, \cdots \lambda_{M-1}$)

$$\begin{cases} \text{Model parameters: } w_0, w_1, \cdots, w_m \quad (f = w^T x) \\ \text{Hyperparameters: } \lambda_1, \lambda_2. \quad (\text{not related to } f) \end{cases}$$

4. Problems in Linear Regression

$$f(x_i; w) = w_0 + w_1 \cdot x_{i1} + w_2 \cdot x_{i2} + \cdots + w_m \cdot x_{im}$$

1). the independent assumption: All features are independent

$$x_i = \begin{pmatrix} x_{i1} = \text{height} \\ x_{i2} = \text{weight} \end{pmatrix}$$

the standard linear model (f) cannot model the dependence.

We can add interaction terms in model to model dependency

$$f = w_0 + x_{i1} \cdot w_1 + w_{i2} \cdot x_{i2}$$

$$\downarrow$$

$$f_1 = w_0 + x_{i1} \cdot w_1 + w_{i2} \cdot x_{i2} + \boxed{w_3 \cdot x_{i1} \cdot x_{i2}}$$

interaction term between x_{i1} and x_{i2}.

If we have 3 features: x_{i1}, x_{i2}, x_{i3}.

$x_{i1} \cdot x_{i2}$, $x_{i2} \cdot x_{i3}$, $x_{i1} \cdot x_{i3}$, , $x_{i1} \cdot x_{i2} \cdot x_{i3}$

2) Non-linear relationship between x and y.

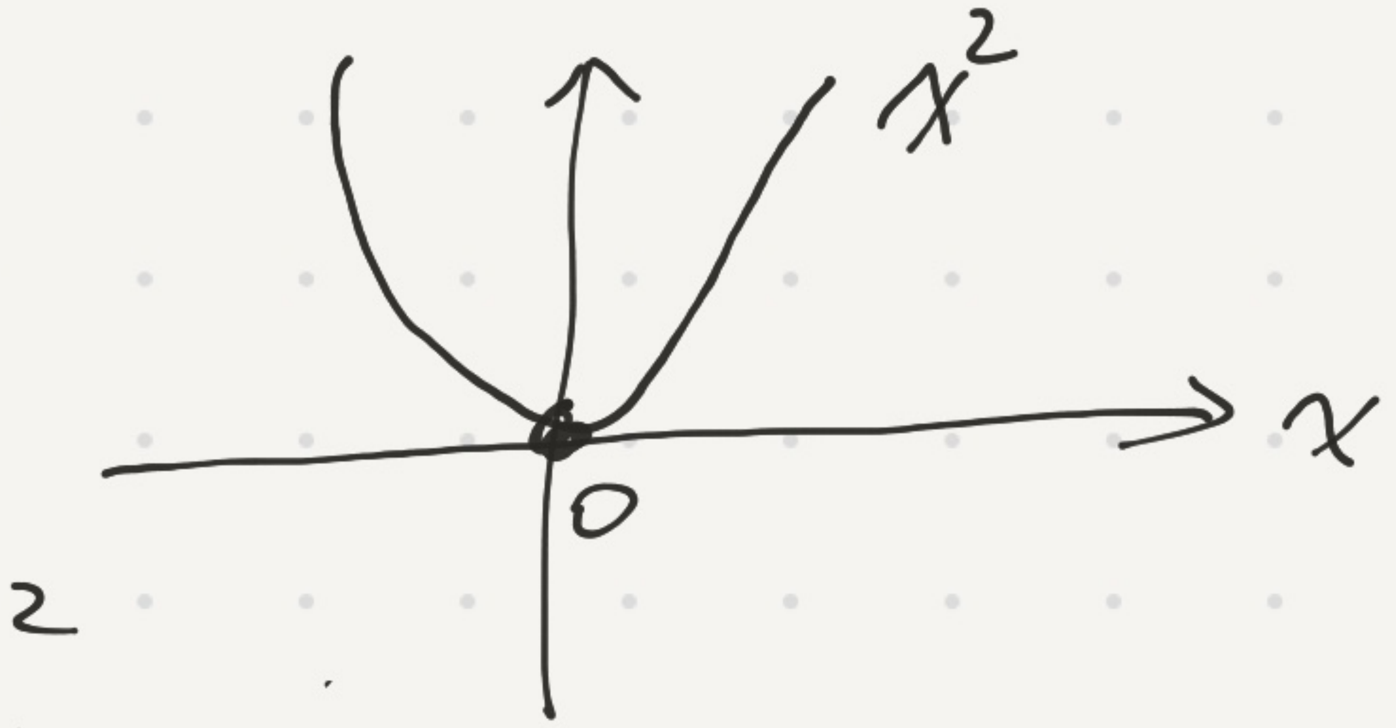
add non-linear terms into model function.

polynomial model:

Linear model: $f = w_0 + w_1 \cdot \underline{x_{i1}} + w_2 \underline{x_{i2}}$.

$\downarrow$

polynomial model: $f = \underbrace{w_0 + w_1 x_{i1} + w_2 x_{i2}}_{\text{Linear terms}} + \underbrace{w_3 x_{i1}^2 + w_4 x_{i2}^2 + w_3 x_{i1} x_{i2}}_{\text{non-linear terms}}$ with degree 2



Polynomial function / model with degree 3:

degree 3: x_{i1}^3, x_{i2}^3, $x_{i1}^2 \cdot x_{i2}$, $x_{i1} \cdot x_{i2}^2$

degree 2:

degree 1. and 0: