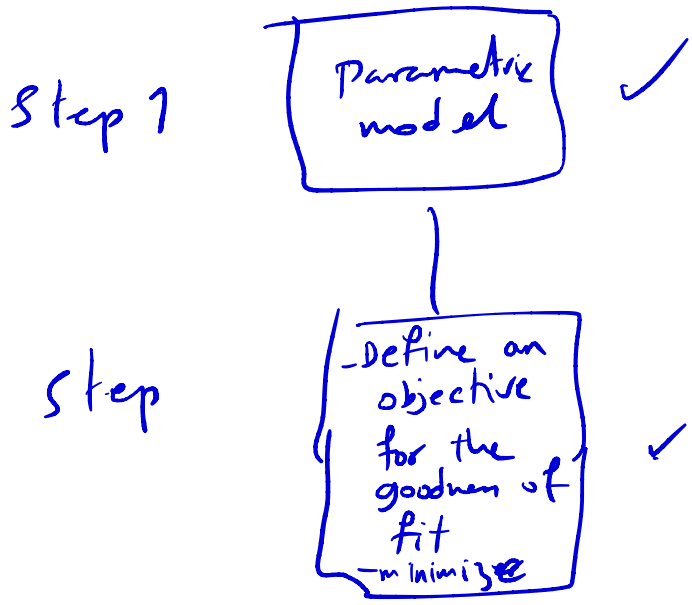
Lecture 15

P-dimensional linear regression:

Data: $\{(x_i, y_i)\}_{i=1}^{n}$ $\quad x_i \in \mathbb{R}^p$



Step 1:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p \longrightarrow y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

p+1 parameters (to be learned from data)

step 2:

$$RSS = \sum_{i=1}^{n} (\text{error}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \right) \right)^2$$

y_i: true label; $\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$: prediction

to find $\{\hat{\beta}_j\}_{j=0}^{p}$ we'll have to minimize RSS.

$$RSS = \sum_{i=1}^{n} (y_i - \tilde{x}_i^T \vec{\beta})^2$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \in \mathbb{R}^{p+1}, \quad \tilde{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^{p+1}$$

$$RSS = \| \vec{Y} - X \vec{\beta} \|_2^2$$

$\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$

$= [1, x_{i1}, x_{i2}, \ldots, x_{ip}] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$ (data i)

- $\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$ $\quad \vec{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n \times (p+1)}$

$$= \begin{pmatrix} \tilde{x}_1^T \\ \hline \tilde{x}_2^T \\ \hline \vdots \\ \hline \tilde{x}_n^T \end{pmatrix}$$

- $\|z\|_2^2 = \sum_{i=1}^{n} z_i^2$

$z_2 \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$

$$\underset{\vec{\beta}}{\text{minimize}} \; RSS(\vec{\beta})$$

$$\underset{\vec{\beta}}{\text{minimize}} \; \underbrace{\| \vec{y} - X\vec{\beta} \|_2^2}_{RSS(\vec{\beta})}$$

$$\Rightarrow \quad \underbrace{\nabla RSS(\vec{\beta}) = 0}$$

$$-2 X^T (\vec{y} - X\vec{\beta}) = 0$$

$$\Rightarrow X^T X \vec{\beta} = X^T \vec{y}$$

$$\Rightarrow \boxed{\vec{\beta} = (X^T X)^{-1} X^T \vec{y}}$$

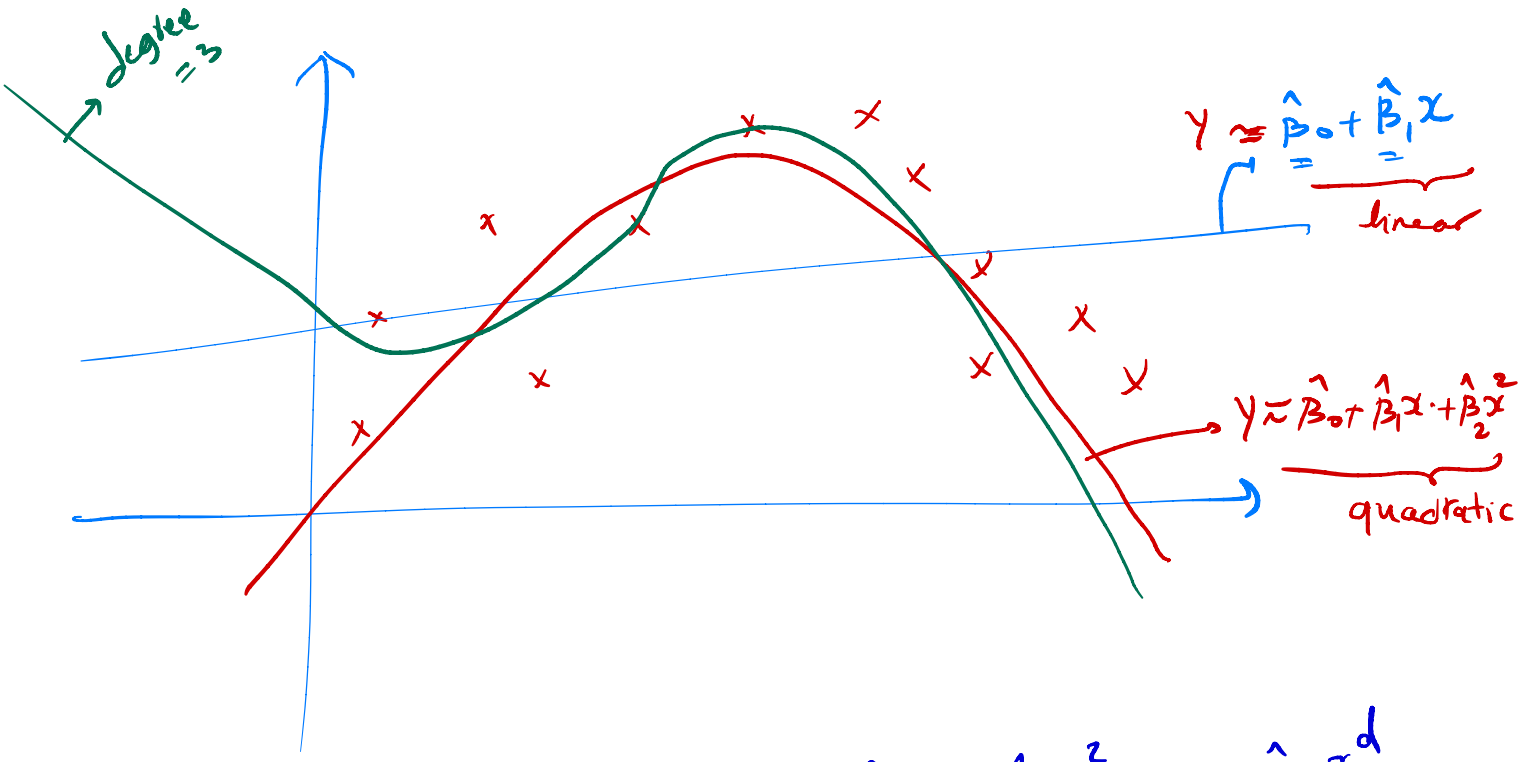
$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & & & \\ \vdots & & & \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \text{i-th} \rightarrow$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- So far, we've only talked about linear models. Let's now see how we can construct more complex (non-linear) parametric models.

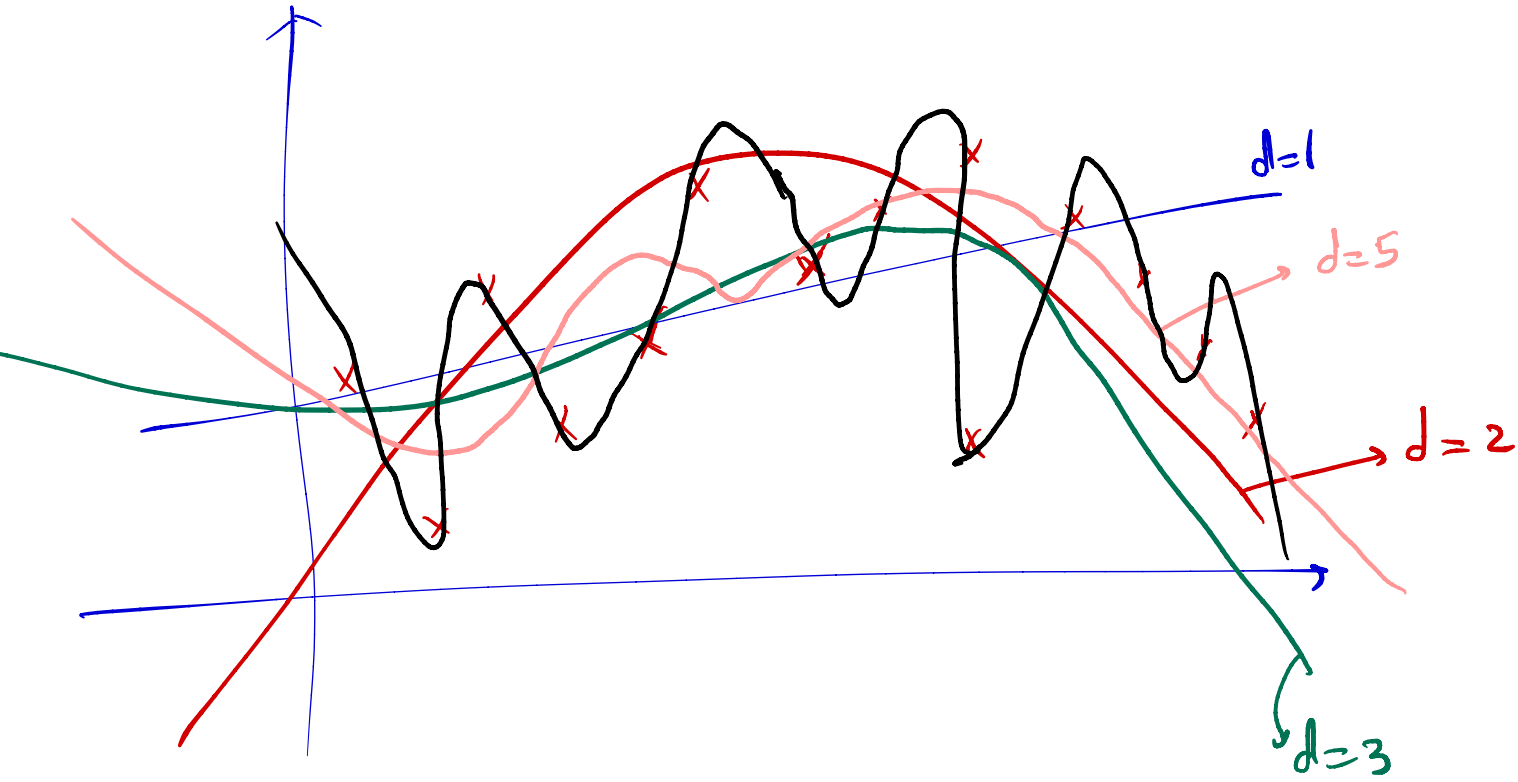
Let's assume $p=1$.



polynomial : $y = p(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_d x^d$

One way to construct non-linear parametric models is by using polynomial models:

- degree of the model : d
- parameters : $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_d$ ($d+1$ parameters)



d=1000

$$y \approx \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_{1000} x^{1000}$$

Question: Given a polynomial family of degree d how do we find the best parameters (the best fit)?

Data: $\{(x_i, y_i)\}_{i=1}^{n}$ $\quad x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$

Step 1: Parametric model: polynomials of degree d: $y = \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_d x^d$

Step: Define an objective for the goodness of fit:

$$\xrightarrow{\text{objective}} RSS = \sum_{i=1}^{n} \text{error}_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \cdots + \hat{\beta}_d x_i^d \right) \right)^2$$

this objective is analogous to the multidimensional linear regression:

$$\left(\begin{pmatrix} 1 \\ x_1 \\ x_1^2 \\ \vdots \\ x_1^d \end{pmatrix} \right)$$

$$x_i \in \mathbb{R} \longrightarrow \tilde{x}_i = \begin{pmatrix} 1 \\ x_i \\ x_i^2 \\ \vdots \\ x_i^d \end{pmatrix}, \quad \vec{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_d \end{pmatrix}$$

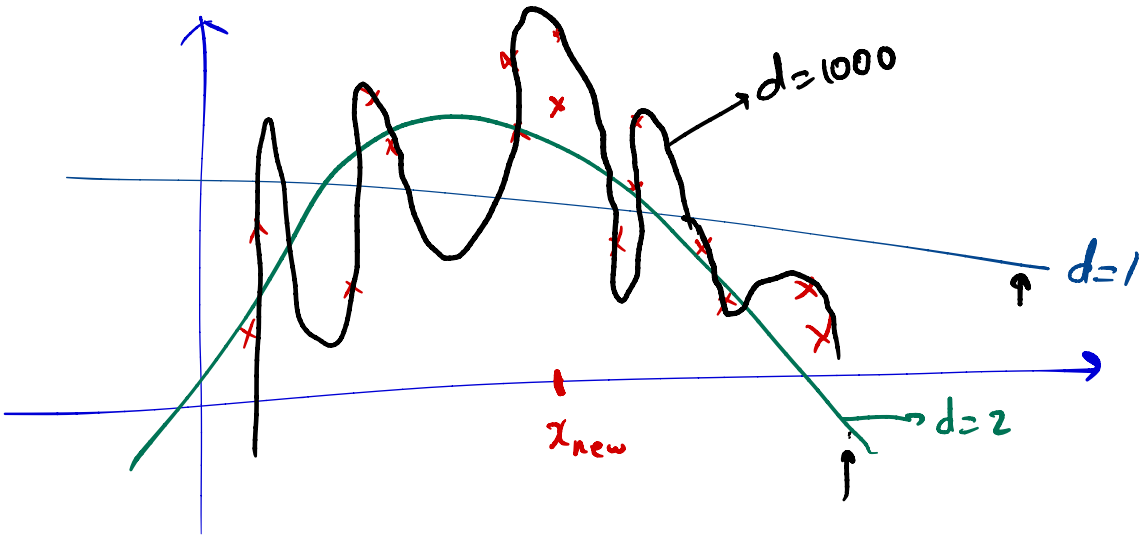
new data point #i

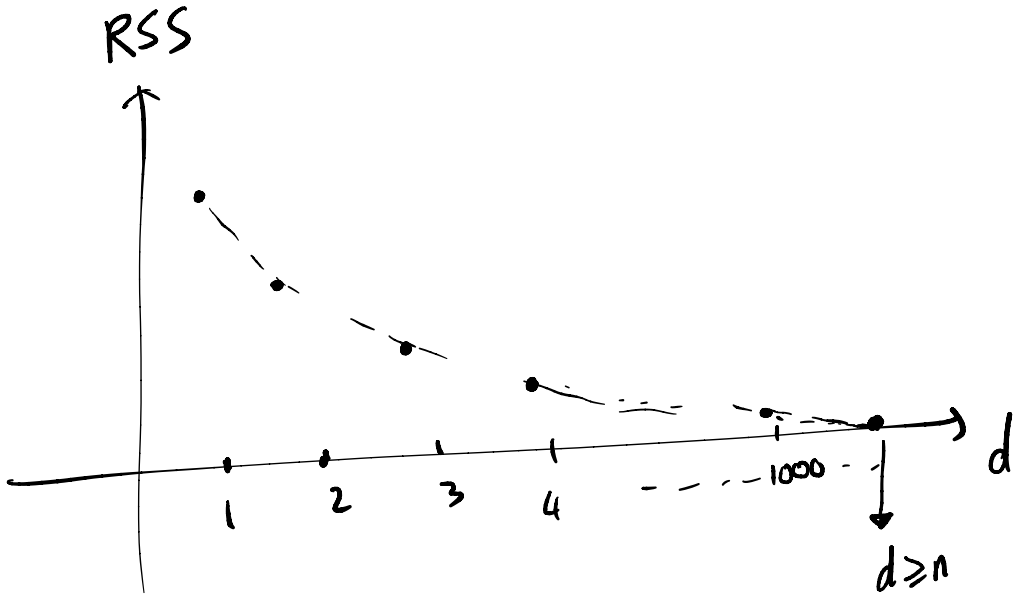
$$RSS = \sum_{i=1}^{n} \left(y_i - \tilde{x}_i^T \vec{\beta} \right)^2$$

$$\Rightarrow \vec{\beta} = (X^T X)^{-1} (X^T Y)$$

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^d \end{pmatrix}_{n \times (d+1)}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$





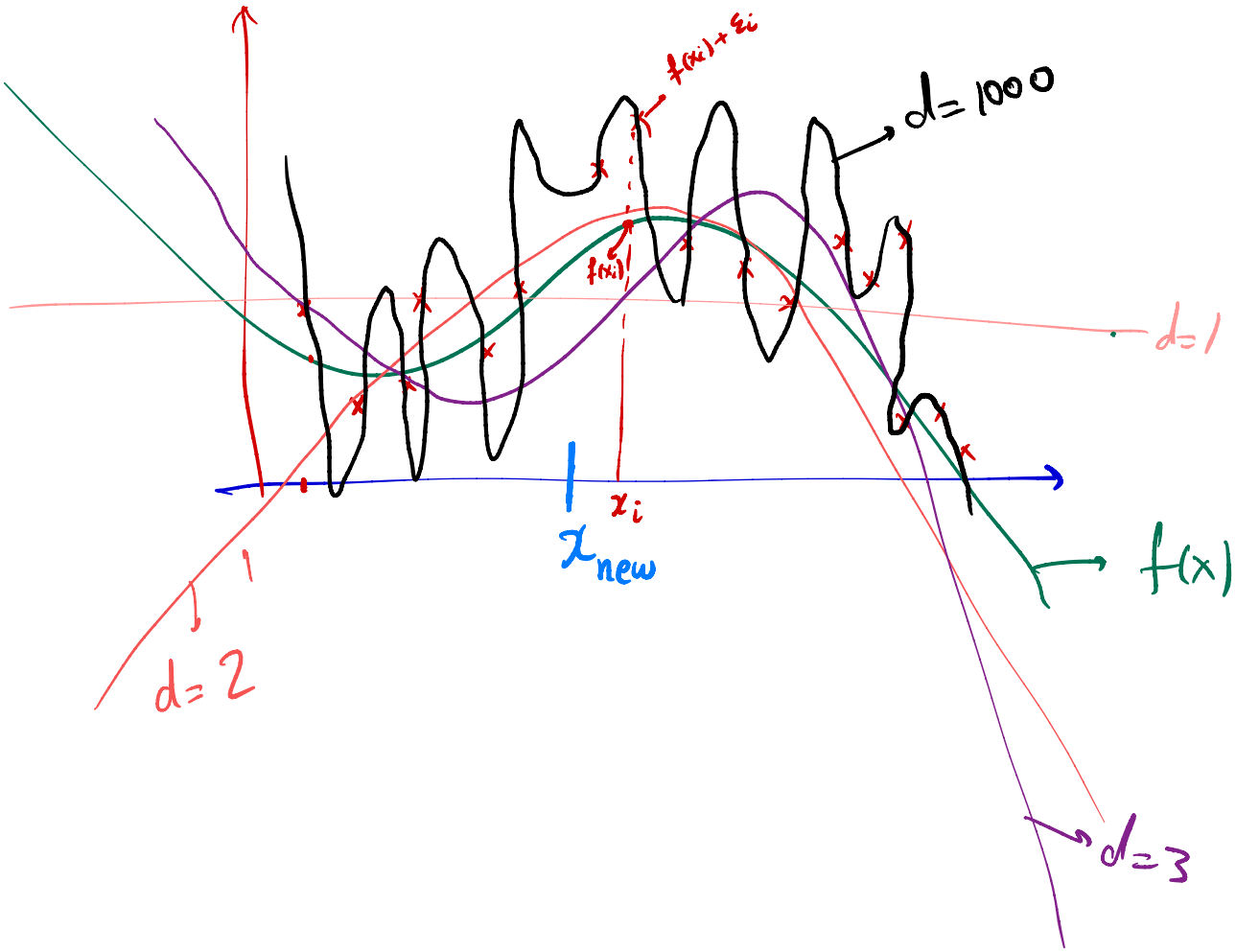
- Remember that the main goal behind learning predictive models was to be able to predict well on new and unseen data points.

- Consider the following model for data:

$$y = f(x) + \varepsilon$$

(f: how the data is generate; ε: noise)

e.g. $f(x) = x^3 + 3x^2 - 4x + 5$



- On a new data point x_{new}

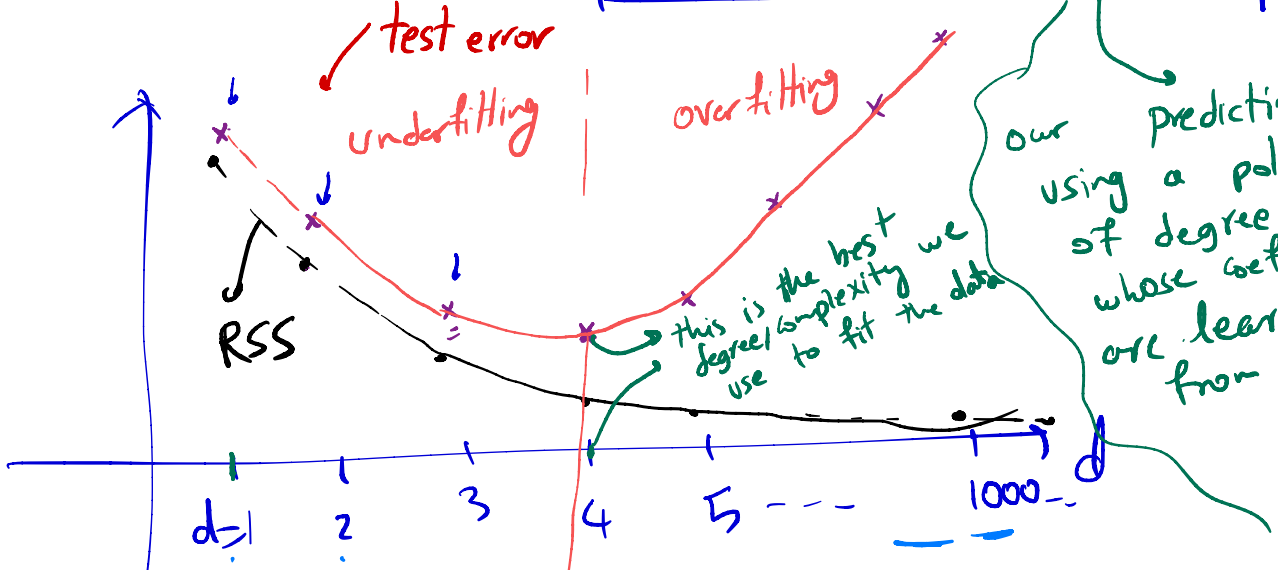
→ $\underline{\underline{y_{new} = f(x_{new}) + \varepsilon}}$

the model learned from data

w

→ Test error = $E\left[\left(y_{new} - \hat{P}_d(x_{new}, \vec{\beta}) \right)^2 \right]$

$y_{new} = f(x_{new}) + \varepsilon$



our prediction using a polynomial of degree d whose coefficients are learned from data

We have two regions depending on the complexity (degree) of the parametric model that we're fitting to the data:

- Underfitting:
  - the model is not sufficiently complex to explain the data (high bias)
  - the trained model does not change much if we add new training data points (low variance)

- Overfitting:
  - model is too complex for training the data (low bias)
  - the trained model will change significantly if we add new data points (high variance)