# Lecture 14

*) Regression:

Given $x_i \in \mathbb{R}^{d}$, $y_i \in \mathbb{R}^{p}$.

for a test point $x_{test}$, we want to predict

$$y_{test} = \hat{f}(x_{test})$$

$$\hat{f}(x) = \omega_1 x + \omega_0$$

$$\omega_2 x^2 + \omega_1 x + \omega_0$$

$d=1 \quad p=1$

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & & 1 \end{bmatrix}$$

$$Y \in \mathbb{R}^{n \times 1}$$

$$W = (X'X)^{-1} X Y$$

What if you take several sub-datasets where samples Come from same distribution and then compute f^avg?

$$\hat{f}_1(x) \leftarrow w_2^1 x^2 + w_1^1 x + w_0^1 \quad \text{learnt} \quad D_1$$

$$\hat{f}_2(x) \leftarrow w_2^2 x^2 + w_1^2 x + w_0^2 \quad \text{learnt} \quad D_2$$

$$\vdots$$

$$\hat{f}_n(x) \leftarrow w_2^n x^2 + w_1^n x + w_0^n \quad \leftarrow D_n$$

$$D_1, D_2, D_n \leftarrow D$$

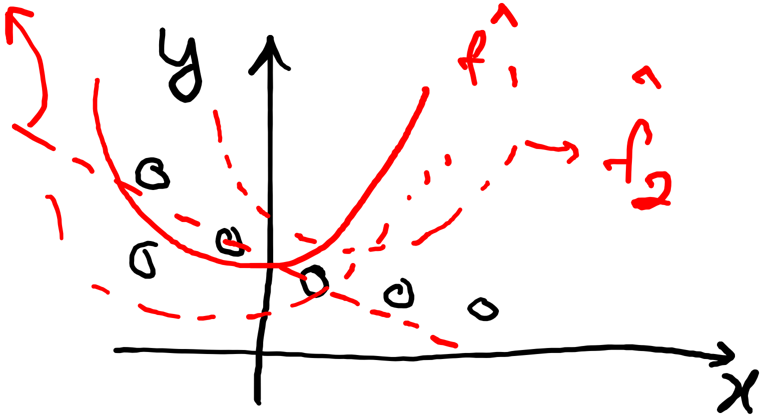
$$\hat{f}_{avg}(x) = x^2 E(w_2) + x E(w_1) + E(w_0)$$

# Bias-Variance Tradeoff

$E(\hat{f}) \approx f$

- For higher M: f^avg of E(f^) ca~~n~~ fit into data that requires degree M or less.
- Variance: (1/n-1){f^$_i$(x) – E(f^)}$^2$ is **large**
- For eg. f^$_i$(x) may be quadratic, whereas, expectation may be quad. Linear or constant.
- Bias: Will be **small** as E(f^) approx. equal to true f.

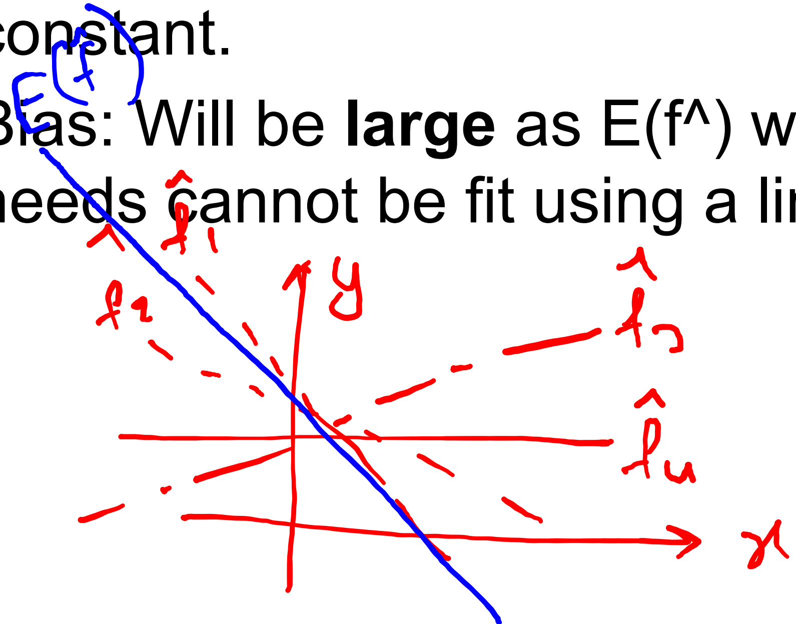
$E(\hat{f})$




true f → was a straight line

$E(\hat{f})$ can very well model straight line.

- For lower M: f^avg of E(f^) will under fit the data that requires a higher degree.
- A data that needs cannot be fit using a line.
- Variance: (1/n-1){$f^{\wedge}_i(x)$ – E(f^)}$^2$ is **small**
- For eg. $f^{\wedge}_i(x)$ may be linear, whereas, expectation may be, Linear or constant.
- Bias: Will be **large** as E(f^) wont be approx. equal to true f. A data that needs cannot be fit using a line.




$E(\hat{f})$ will be farther away from true f

# Cross-validation

- Consider degree m = 1, 2, to M
  - Divide data into K folds. K = 5, 10.
    - Hold one fold and use “remaining folds”.
    - Learn W from “remaining folds”.
    - Apply W to compute error on “remaining folds” – call this training error (k)
    - Apply W to compute error on “held out folds” – call this validation error (k)

    end

    avgTrainErr(m) = mean{train error}

    avgValErr(m) = mean{val error}
- end

$$D = \{(x_1, y_1), (x_2, y_2) \cdots\cdots (x_{10}, y_{10})\}$$

$D \rightarrow$ into 5 folds.

1st fold $\rightarrow$ $(x_1, y_1)$ $(x_2; y_2)$

2-5th fold $\rightarrow$ $(x_3, y_3) \cdots\cdots (x_{10}, y_{10})$

$$W \leftarrow (X'X)^{-1} X Y$$

$$X = \begin{bmatrix} x_3 & \cdots\cdots & x_{10} \\ 1 & & 1 \end{bmatrix} \quad Y = \begin{bmatrix} y_3 \\ \vdots \\ y_{10} \end{bmatrix}$$

apply W on 1st fold.

$(x_1, y_1)\ (x_2, y_2)$

Val. error(1)

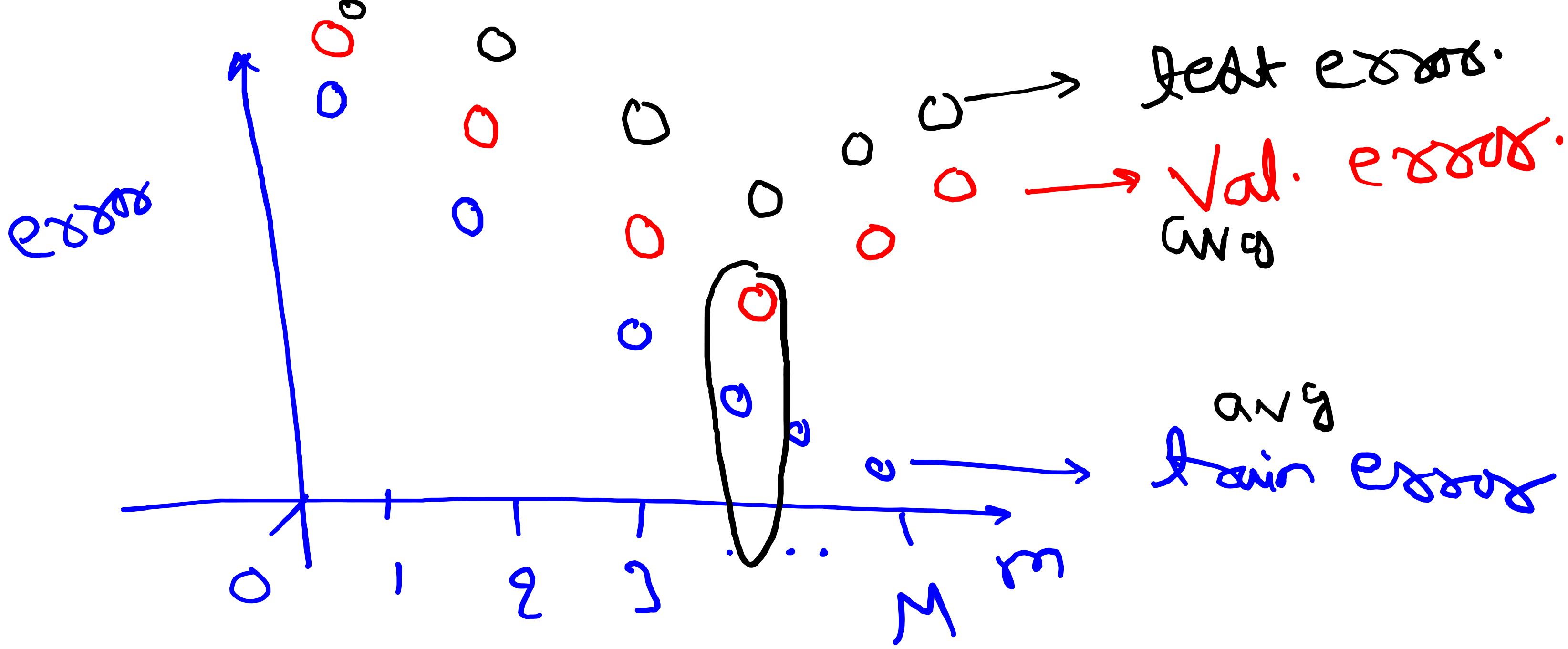
apply W on 2nd - 5th fold.

$(x_3, y_3), \ldots (x_{10}, y_{10})$

train error(1)

for K=2, $(x_1, y_1)\ (x_2, y_2), (x_5, y_5) \ldots (x_{10}, y_{10})$

W apply them on $(x_3, y_3)\ (x_4, y_4)$ → val error(2)



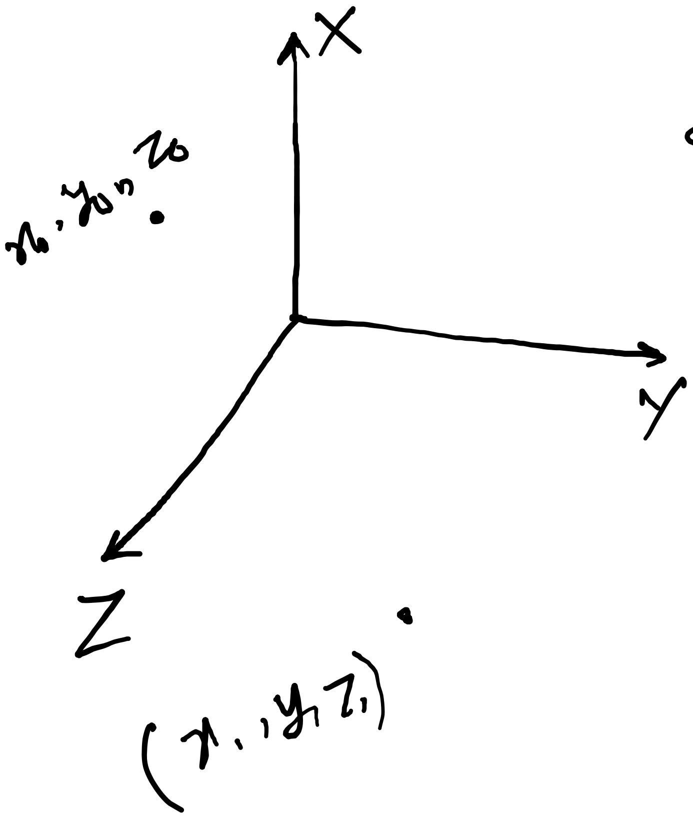

'm' that gives the least val. error. will be 'm' that you will use further.

for eg. m=4, gives least val. error.

m=4, & use the full training set.

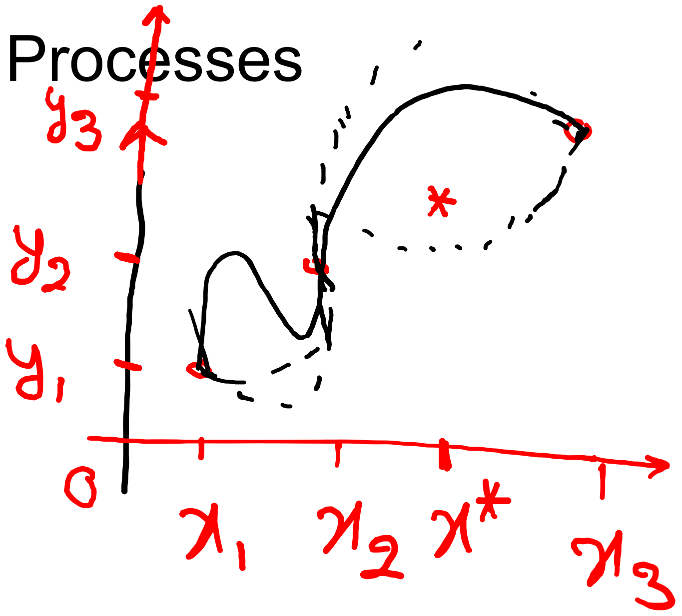
→ relearn W → testing set.

# Gaussian Processes




$S_0$

$S_1$



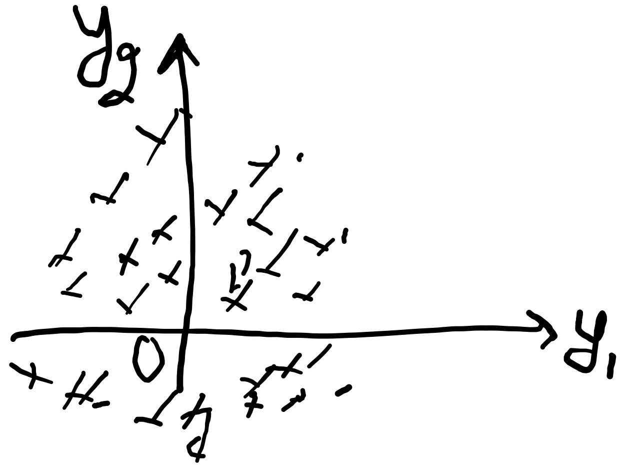

*) $(x_1, x_2)$ are closer → $(y_1, y_2)$ similar

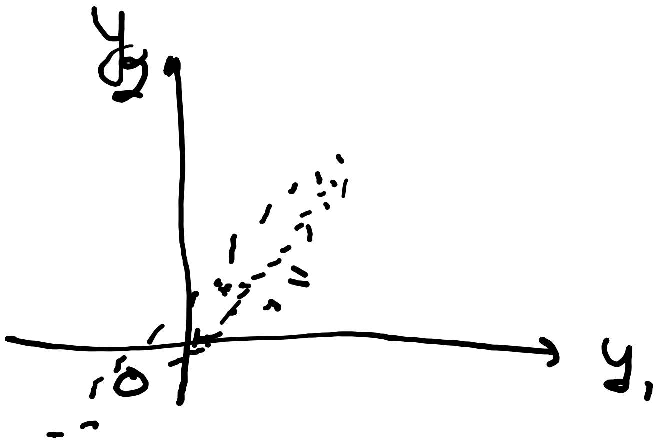
*) If $(x_1, x_3)$ are farther $(y_1, y_3)$ dis-similar

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \text{ assume coming M.V.G } \sim$$

$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

$$\text{Cov.} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \qquad \text{Cov.} \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$$



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 & 0.2 \\ 0.8 & 1 & 0.5 \\ 0.2 & 0.5 & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & K(x_1, x_3) \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix} \right)$$

Sim. function. $K(a, b)$

$$K(x_i, x_j) = K_{ij}$$

$$K(x_i, x_j) = \sigma^2 e^{-\|x_i - x_j\|_2^2 / 2l^2}$$

$l \rightarrow$ width of Gaussian Kernal

$\sigma \rightarrow$ max of Kernal.

$\sigma, l \rightarrow$ Unknown.

$\Sigma_{aa}$ $\Sigma_{ab}$

$$y_a \left\{ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_* \end{bmatrix} \right. \sim N\left( \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{1*} \\ K_{21} & K_{22} & K_{23} & K_{2*} \\ K_{31} & K_{32} & K_{33} & K_{3*} \\ K_{*1} & K_{*2} & K_{*3} & K_{**} \end{bmatrix} \right)$$

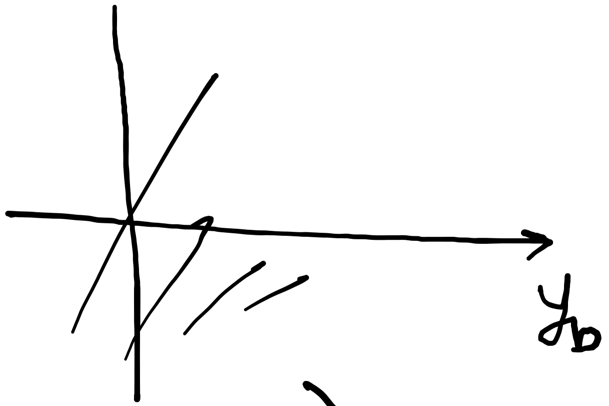
$\downarrow$ $\Sigma_{bb}$

$\Sigma_{aa}, \Sigma_{bb}, \Sigma_{ab}$

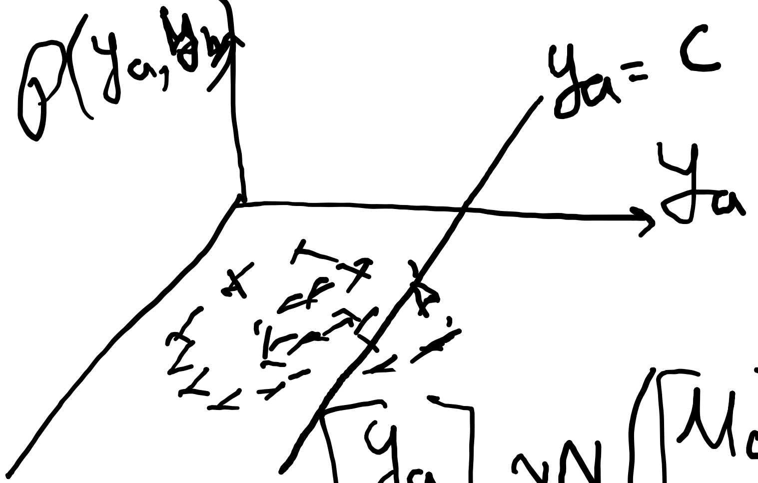
$$\left.\begin{aligned} \mu_{b|a} &= \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(y_a - \mu_a) \\ \Sigma_{b|a} &= \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{aligned}\right\} \text{ theorem.}$$

$$P(y^* \mid x, y) \sim N(\mu^*, \sigma^*)$$

Conditional distribution on $y^*$ given all other info




$P(y_b \mid y_a)$




$$\begin{bmatrix} y_a \\ y_b \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right)$$

$$P(y^* \mid y_1, y_2, y_3) \sim$$

C.I.

$$\mu^* \pm 1.96\sqrt{\Sigma_{b|a}}$$

f. dep.

$$\mu^* = \Sigma_{ba}\Sigma_{aa}^{-1} y_a$$

$$= \begin{bmatrix} K_{1*} & K_{2*} & K_{3*} \end{bmatrix} \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ & \ddots & \\ K_{31} & \cdots & K_{33} \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$\Sigma_{b|a} = K_{**} - \begin{bmatrix} K_{1*} & K_{2*} & K_{3*} \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix}^{-1} \begin{bmatrix} K_{1*} \\ K_{2*} \\ K_{3*} \end{bmatrix}$$