Lec 06: Recap: MAP estimate

$$w^*_{MAP} \in \arg\min_w \left\{ \frac{1}{2\sigma^2} \underbrace{\|Xw - y\|^2}_{} + \underline{\frac{\lambda}{2} \|w\|^2} \right\}$$

$$= \frac{1}{\sigma^2} \left(\underbrace{\frac{1}{2\sigma^2} X^T X + \frac{\lambda}{2} I}_{A} \right)^{-1} X^T y$$

$A_{d \times d}$ is PD if $\forall x \in \mathbb{R}^d \setminus \{0\}$

$x^T A x > 0$

$v^T A v > 0$

$= \frac{1}{2\sigma^2} v^T X^T X v + \frac{\lambda}{2} \|v\|^2$

$= \frac{1}{2\sigma^2} \underbrace{\|Xv\|^2}_{\geq 0} + \underbrace{\frac{\lambda}{2} \|v\|^2}_{} > 0$

Bias & Variance

Goal: Estimate $\hat{y}$ → not seen in our training example

$(\hat{x}, \hat{y})$ → test data point

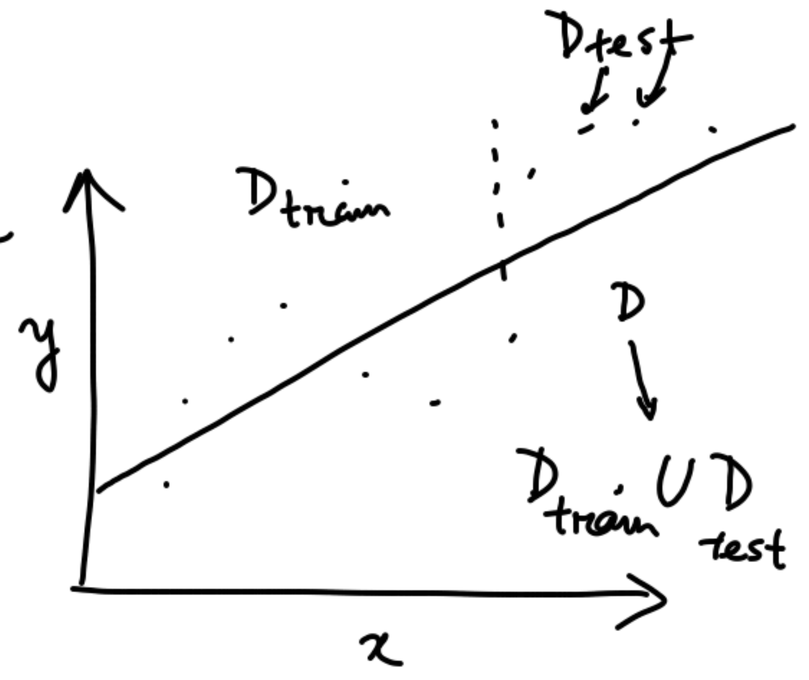
$(x_i, y_i) \in D_{train}$ → training data points.

↓

$g_D(\hat{x}) \rightarrow \hat{y}$

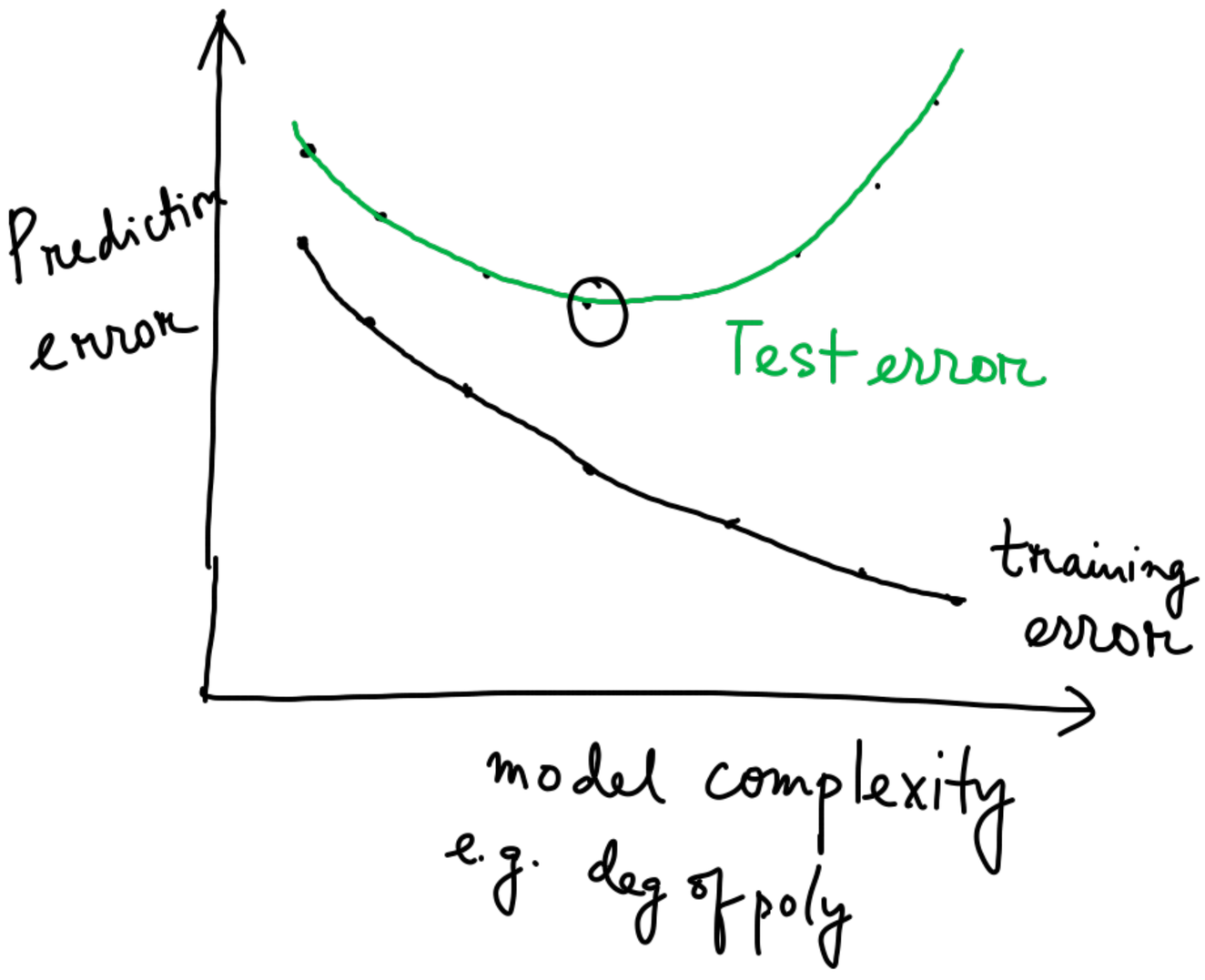
Measure of goodness: ① training error

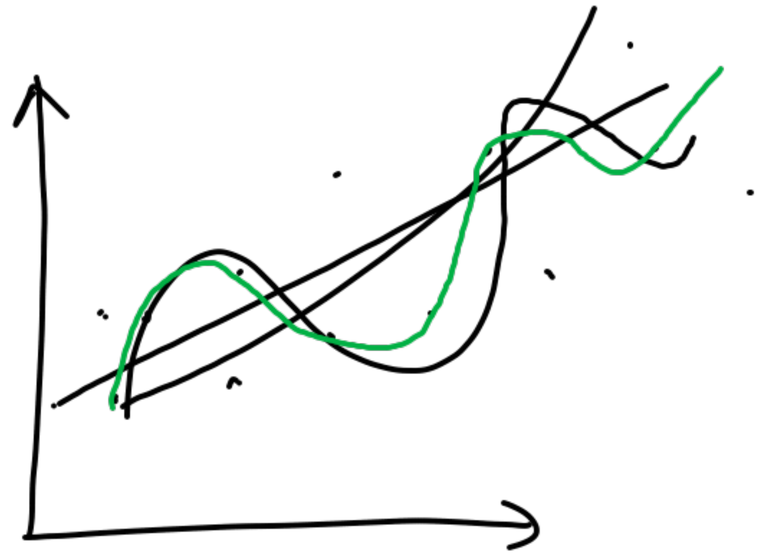
$$\sum_{i \in D} \ell\left(\underbrace{g_D(x_i)}_{\text{prediction}}, \underbrace{y_i}_{\text{actual}}\right)$$



Measure 2: **Test error** → hold out set D_{test}

$$\sum_{j \in D_{test}} \ell\left(g_D(x_j), y_j\right)$$

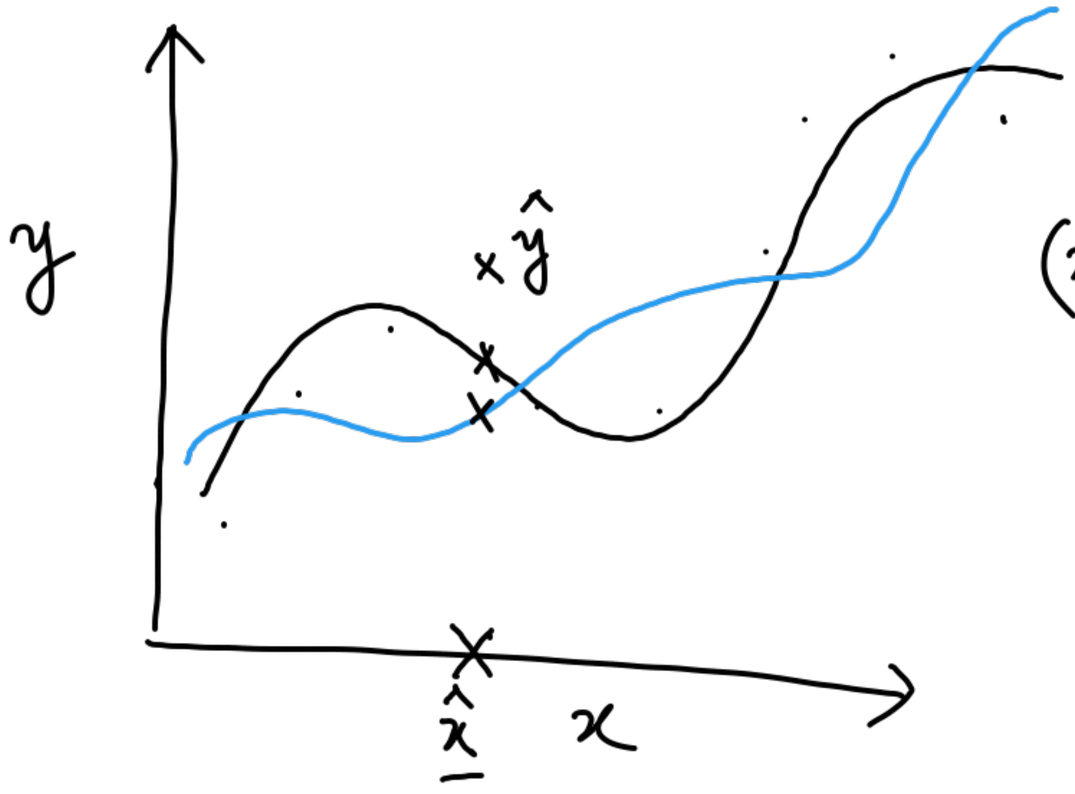




Test error

① Bias
② Variance
} model dependent
③ Noise

Q: What are Bias and Variance?

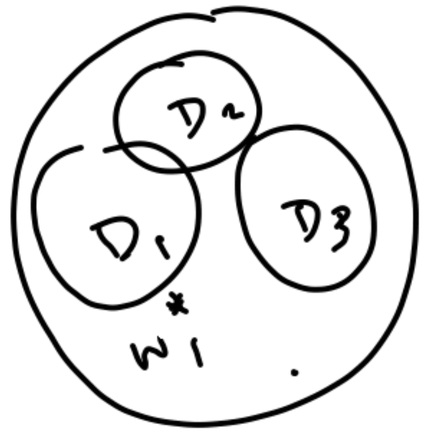


$\hat{y}^*$

$\frac{1}{3}\left(w_1^{*T}\hat{x} + w_2^{*T}\hat{x} + \right.$

$\tilde{D} \rightarrow g_{\tilde{D}}(x)$

$(x_i, y_i) \in D$



$g_D(x) = w^T x$

$y = f(x) + \epsilon \sim N(0, \sigma^2)$

Test error: $err = g_D(\hat{x}) - \hat{y}$

Test error

$\hat{x}$ fixed

$$\text{error} = g_D(\hat{x}) - \hat{y}$$

$$= \underbrace{\left(g_D(\hat{x}) - \mathbb{E}_D\left(g_D(\hat{x})\right)\right)}_{A} + \left(\mathbb{E}_D\left(g_D(\hat{x})\right) - \mathbb{E}(\hat{y})\right) \rightarrow B = \text{Bias} + \left(\mathbb{E}(\hat{y}) - \hat{y}\right) \rightarrow C$$

D distribution is independent of $\hat{y}$ dist.

\mathbb{E}

$$E(err^2) = \mathbb{E}A^2 + EB^2 + EC^2 + 2\underbrace{\left[E(AB) + E(BC) + E(CA)\right]}_{0}$$

noise

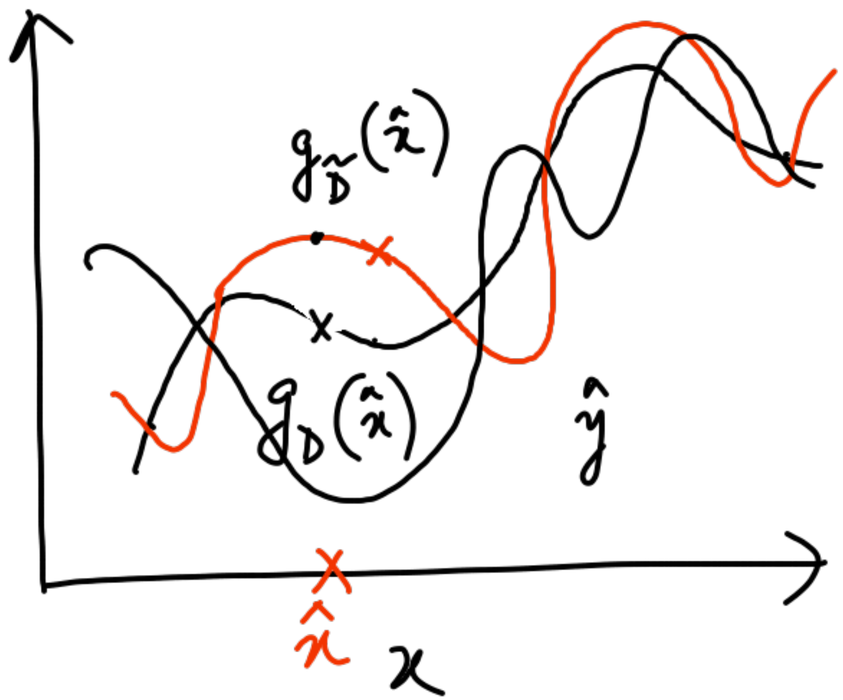
variance

Bias

$$E_D\left[\left(g_D(\hat{x}) - E_D(g_D(\hat{x}))\right)^2\right] = \text{Variance of the model}$$

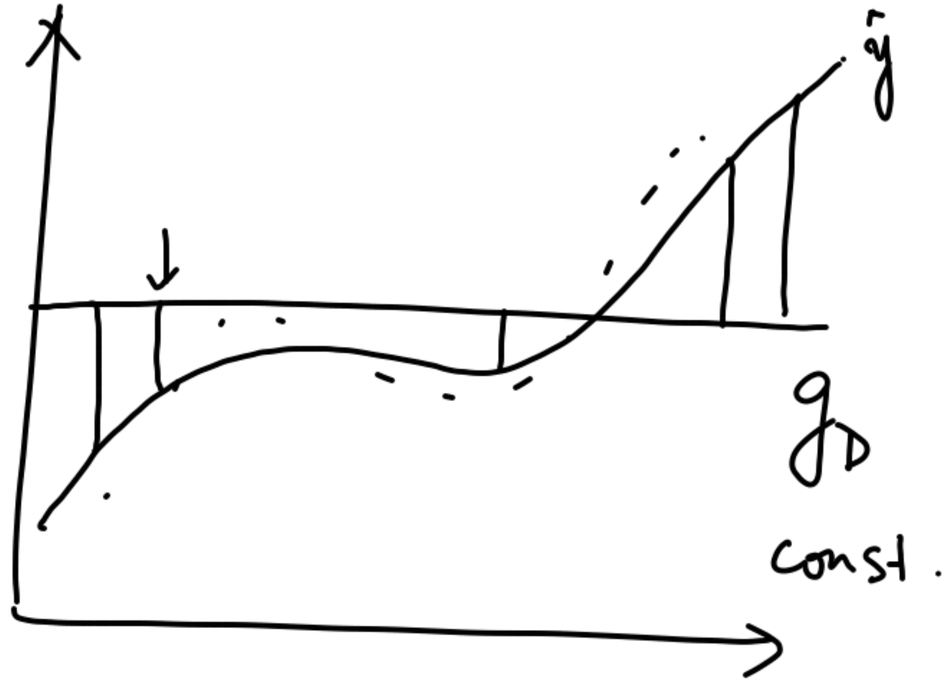
$$E_D\left(g_D(\hat{x})\right) - E(\hat{y}) = \text{Bias of the model}$$

$$E\left[(\hat{y} - E\,\hat{y})^2\right] = \text{Noise}$$



$g_{\tilde{D}}$

$g_D \rightarrow$ polynomial of deg 10



Variance high, Bias is low

Overfitting

$$var = \mathbb{E}\left(g_D(\hat{x}) - \mathbb{E}_D\left(g_D(\hat{x})\right)\right)^2$$

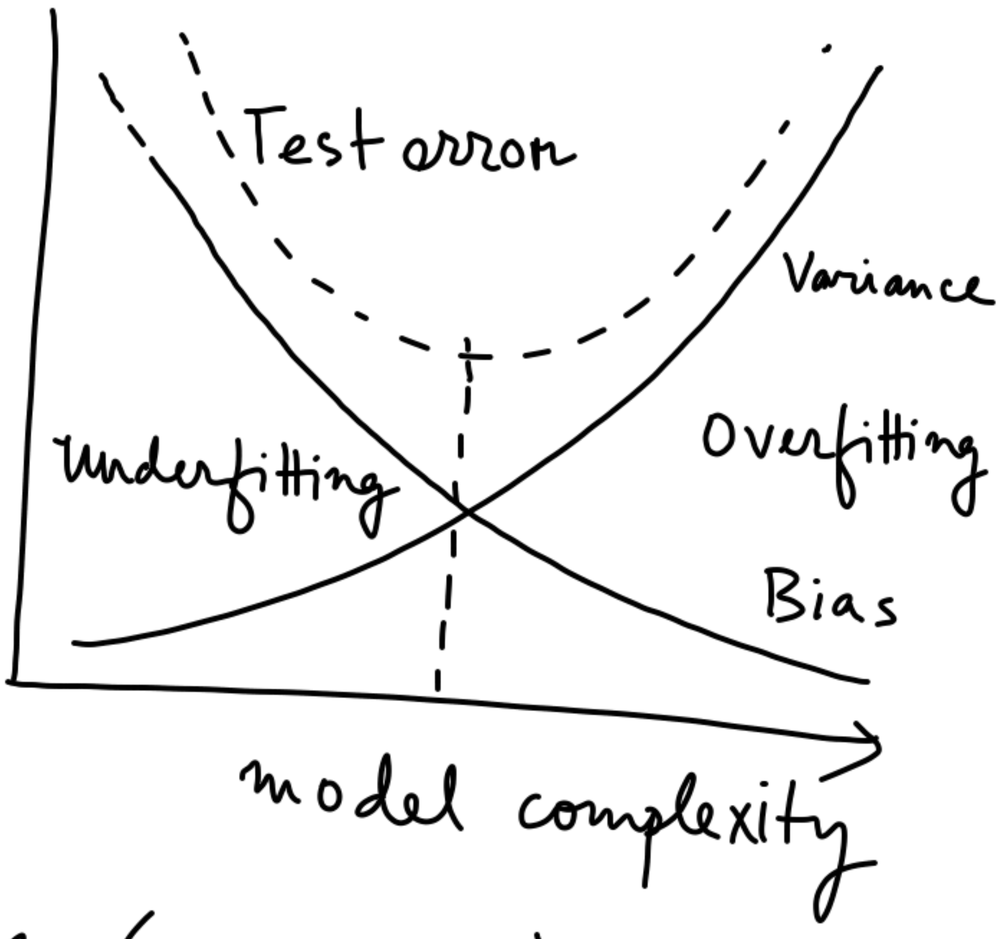
$g_D(x) = w_0$

Bias is high, variance low

Underfitting.

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

$$\mathbb{E}\, y = f(x)$$



$$\mathbb{E}_{\hat{x},\hat{y}}\, \ell\left(g_D(\hat{x}),\ \hat{y}\right)$$

Regularization for linear regression

$$w_{MLE} \in \operatorname{argmin} \frac{1}{2\sigma^2}\|Xw-y\|^2$$

$$w_{MAP} \in \operatorname{argmin}\left\{\frac{1}{2\sigma^2}\|Xw-y\|^2 + \frac{\lambda}{2}\|w\|^2\right\}$$

$\frac{1}{n}$ — hyperparameter

$$\text{Loss}(w) = \text{Loss}_D(w) + \lambda \text{Reg}(w)$$

(Regularized model)

① $Reg(w) = \|w\|_2^2 \rightarrow L_2$ norm $\sqrt{\sum_{i=1}^{d} w_i^2}$

② $Reg(w) = \|w\|_1 \rightarrow L_1$ norm $\sum_{i=1}^{d} |w_i|$

Ridge regression :

$$w^* \in \arg\min_w \left\{ \|\Phi w - y\|^2 + \lambda \|w\|_2^2 \right\}$$

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_m(x_1) \\ \vdots & & \\ \phi_0(x_n) & \cdots & \phi_m(x_n) \end{bmatrix}$$

LASSO

$$w^* \in \arg\min \left\{ \|\Phi w - y\|^2 + \lambda \|w\|_1 \right\}$$

least absolute shrinkage and selection operator (LASSO)

equivalent optimization problem

<u>LASSO</u> $\min \left(\Phi w - y\right)^T \left(\Phi w - y\right)$

s.t. $\|w\|_1 \leq c_1$ ← const. dependent on λ

<u>Ridge</u> objective is same s.t. $\|w\|_2 \leq c_2$