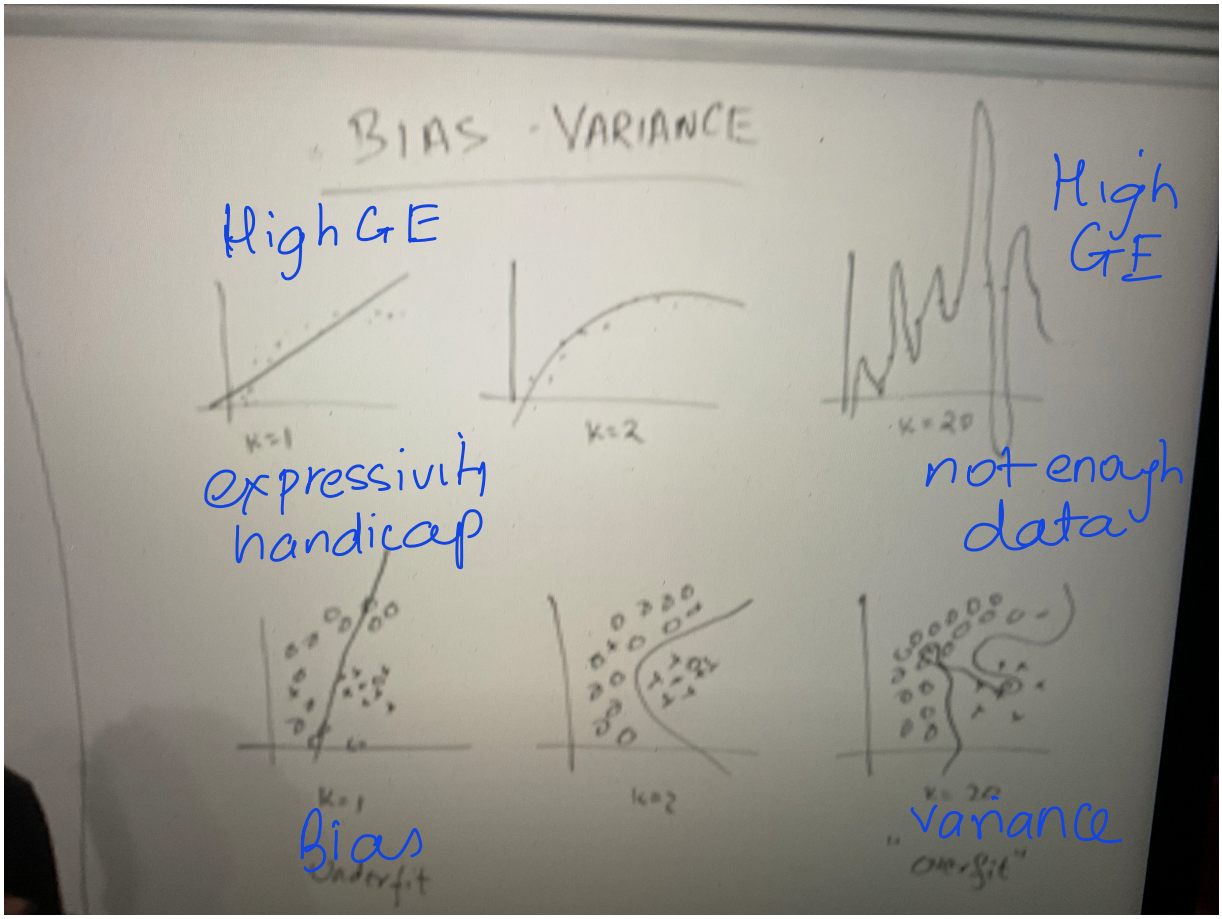
BIAS-VARIANCE



Generalisation Error : loss on unseen data, possibly ∞, from whose distribution we got our training set.

Two components

- Bias — component of GE due to "expressivity handicap"
- Variance - component of GE due to finite sample of training set.

Squared Error

$y^{(i)} = f(x^{(i)}) + \varepsilon^{(i)}$ → Data Generating Process

$E[\varepsilon] = 0 \quad V[\varepsilon] = \sigma^2$

Test Error [G.E.] $= \mathbb{E}[(y - \hat{f}_n(x))^2]$

[expection over all ε in training set & test eg]

↑ [function obtained by training n egs from Data-Gen.

Process

$$= \sigma^2 + \mathbb{E}\left[f(x) - \hat{f}_n(x)\right]^2 + \mathrm{Var}\left[\hat{f}_n(x)\right]$$

Irreducible error

(Bias)^2

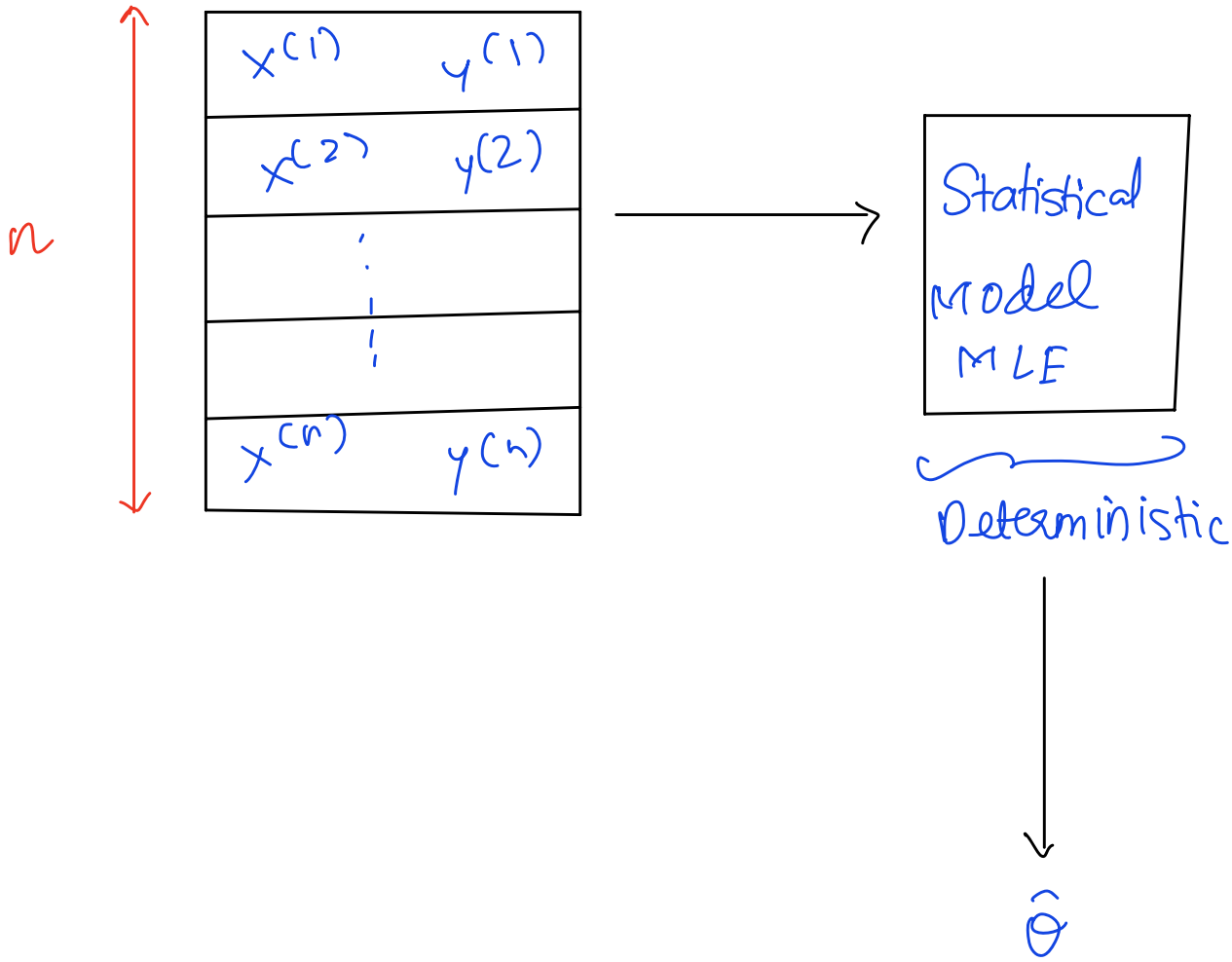
Variance

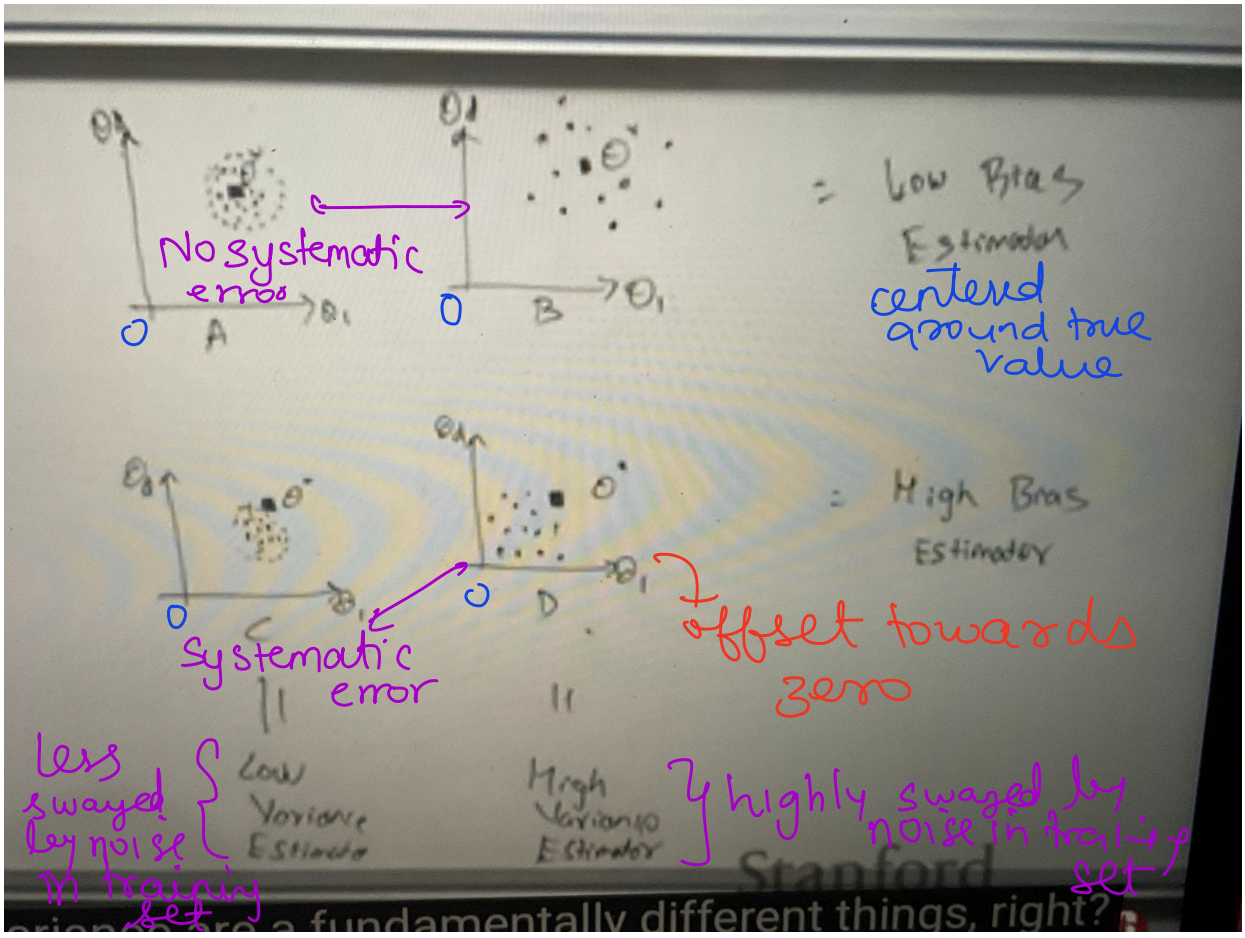
Irreducible error → due to noisiness of data [There is noise in data ε, so for same x too, you can have diff y]

(Bias)^2 → tells us how systematically wrong are we.

Variance → Repeat by taking new n eqs, fit on it, & then prediction you will get a diff $f_n(x)$

$(x, y) \sim \text{Dist}(\theta^*)$

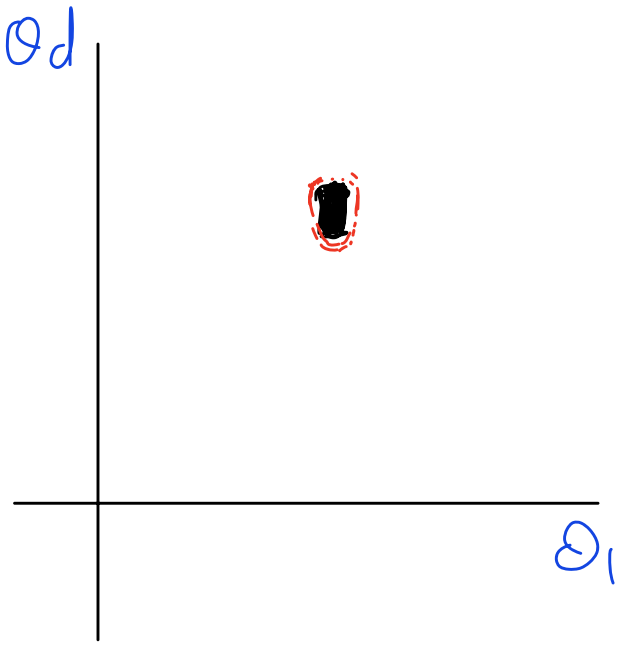


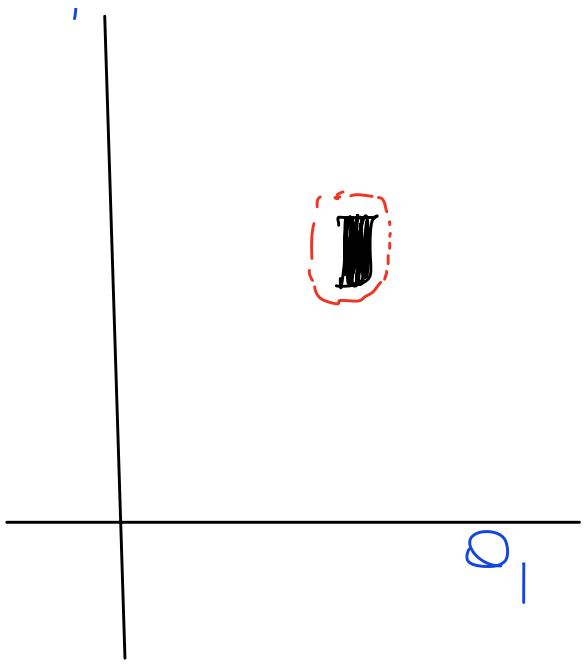


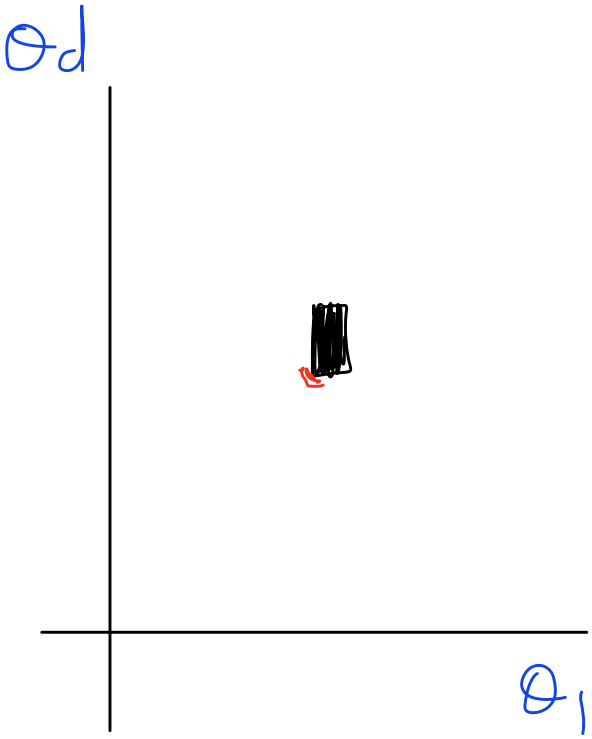
Bias doesn't tell us about variance & vice-versa

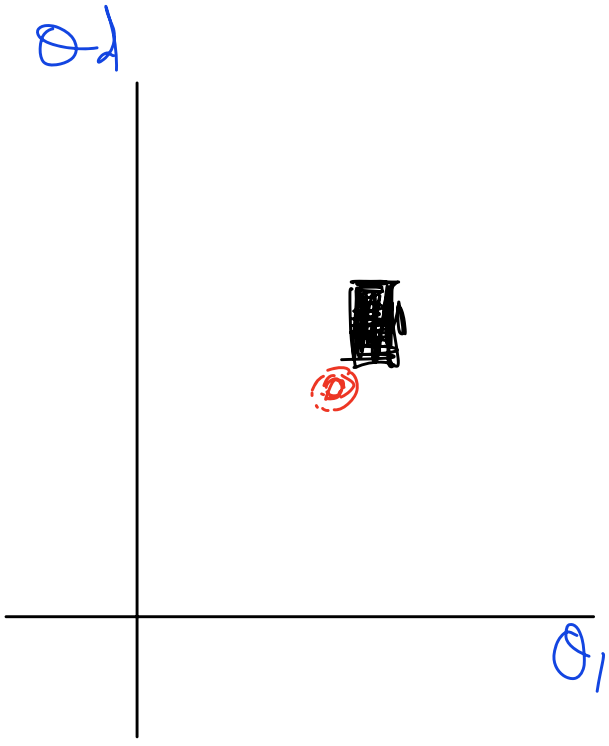
Suppose we take a larger data set with N examples (1000x bigger)

θ_d









In all 4 → variance has come down although bias still remains

$$E[\hat{\theta} - \theta^*] = \text{Bias}$$

$\hat{\theta}$ ↑ Predicted

θ^* ↳ True signal [unknown]

$$\text{Var}[\hat{\theta}] = \text{Variance}$$

$n \to \infty$

If $\underset{n \to \infty}{\text{Bias}} \to 0 \Rightarrow$ Consistent Estimation

Gaussian

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \hat{\mu})^2$$

↑ MLE

Biased Estimator

Underestimated bias

$n \to \infty$, bias $\to 0$.

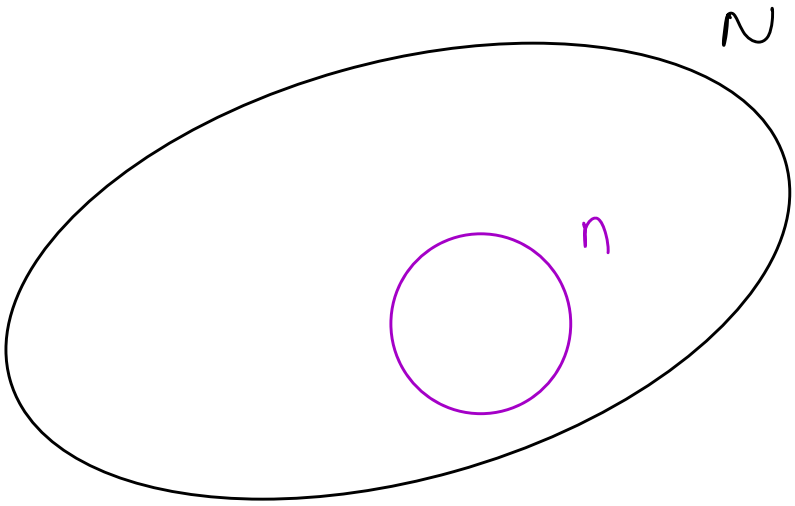
[As you are using $\hat{\mu}$ → estimated mean

Unbiased

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x^{(i)} - \hat{\mu})^2$$

OR $\frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu)^2$

↳ Actual μ



Mean

	Population (parameter)	Sample (statistic)
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

N n

Variance $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$ $S_n = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

↑ biased underestimate

unbiased $S_{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

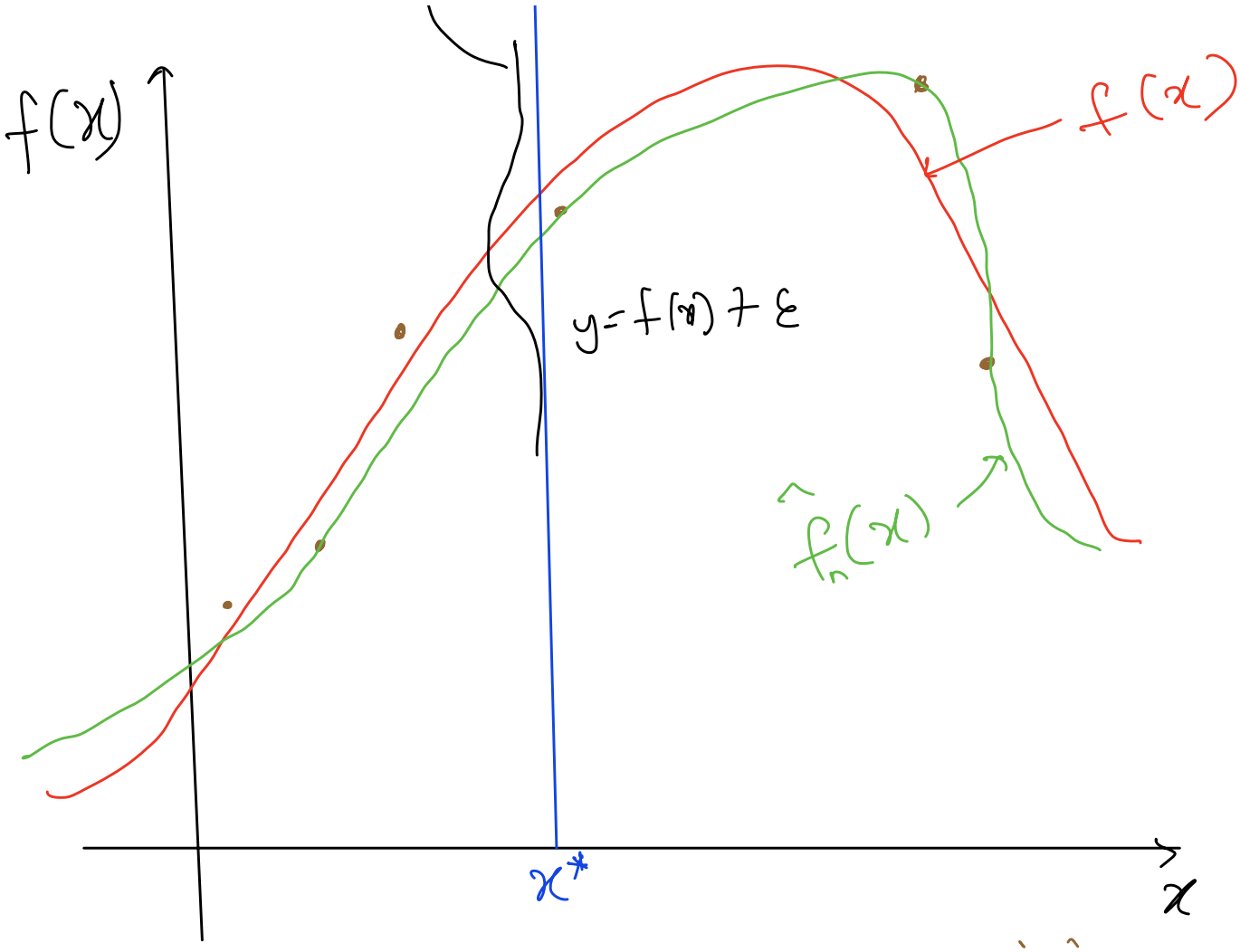
$\lim_{n \to \infty} var[\hat{\theta}] \xrightarrow[rate]{} 0$

↳ statistical efficiency

↳ $\frac{1}{n}, \frac{1}{\sqrt{n}}$

↑ mean → red

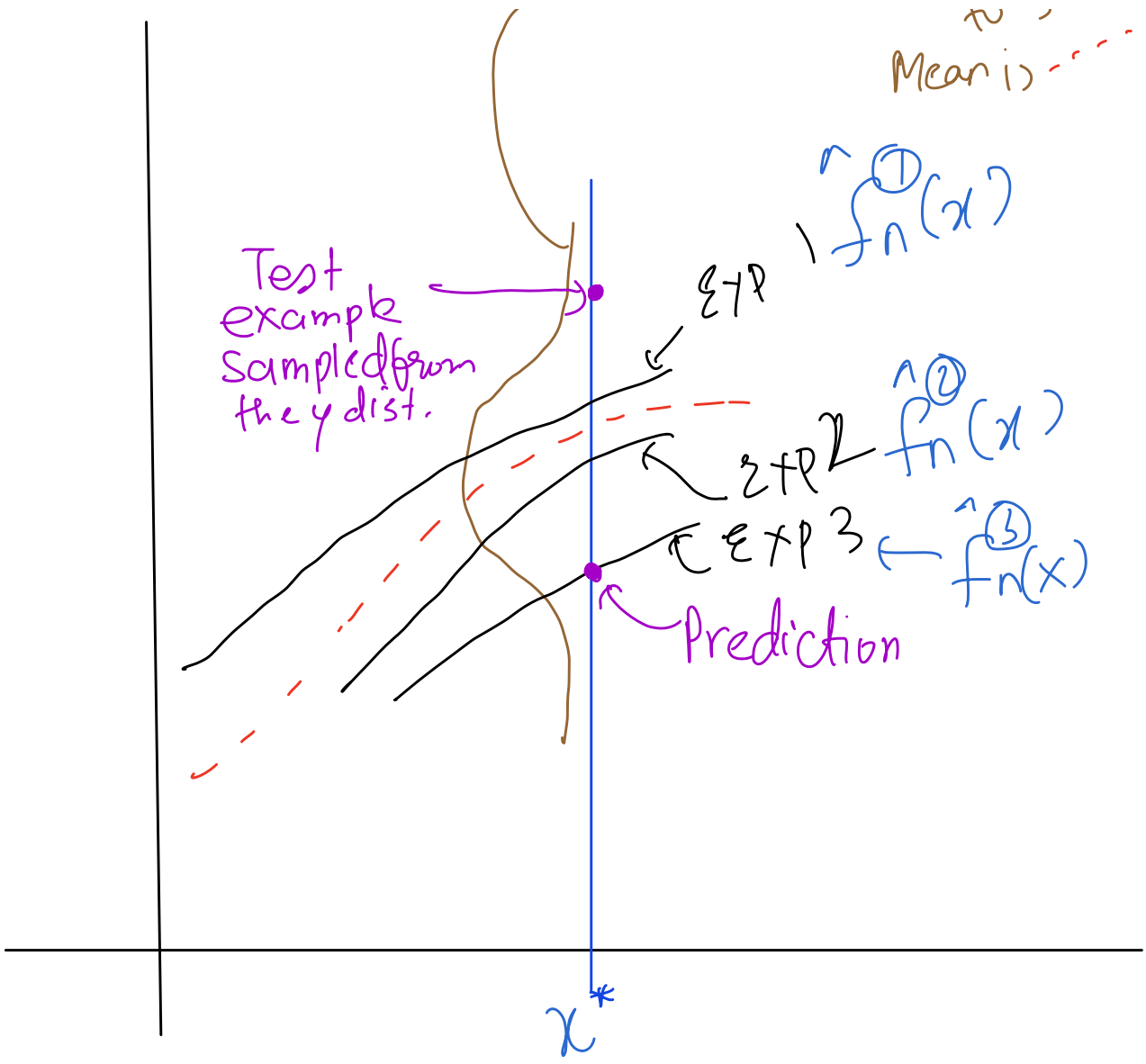
Variance → σ^2



• → training set points (noise added)

∩ → invisible to us

→ distribution which tells us which value of x ($x \in y$) of sample



y is a sample for

$E[(\text{distance})^2]$ = test error

$= \sigma^2 + E[f(x) - \hat{f}_n(x)]^2 + \text{Var}[\hat{f}_n(x)]$

Irreducible error (due to variance of ( ))

(bias error)^2

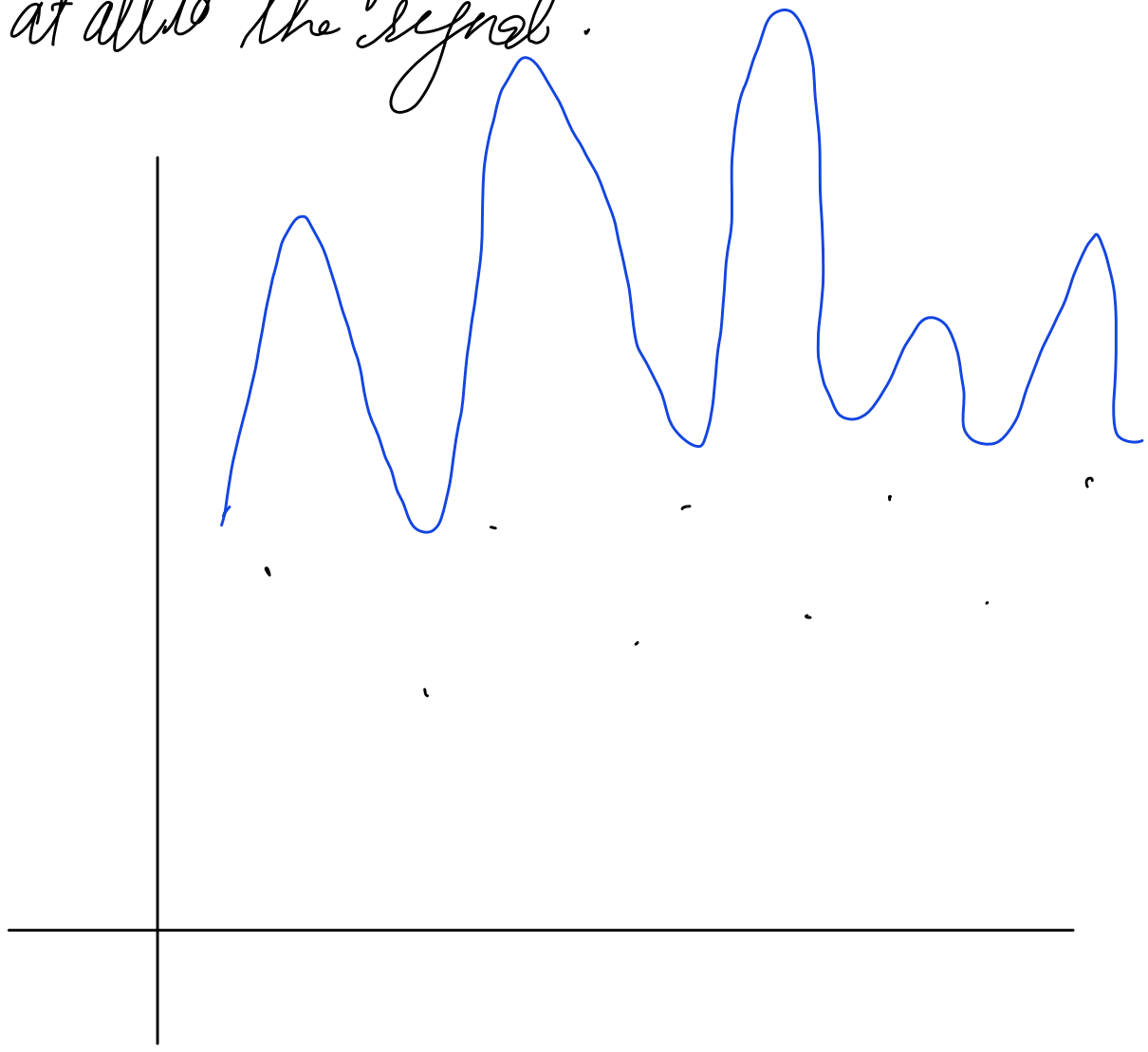
If for our diff. experiments, the predicted y for x^* is below only/ above only or is well centered

how far apart are our predicted y in diff exp.

Underfitting ≈ High bias

Overfitting ≈ High Variance

You can have both at same time like if your model is overfitting/underfitting at same time if model fits to the noise & doesn't fit at all to the signal.



Goal: Do well on Generalization Error

- Cross Validation

Do well on training set/Min. loss

Split into

Train	70	60	80
Valid/Dev	20	20	10
Test	10	20	10

Do well on gen. error

↳ Estimate generalization error

We check loss on valid set. Then do hyperparameter tuning

By repeatedly doing cross validation, the model is looking at valid set, it might rote it!
To counter that we have test set

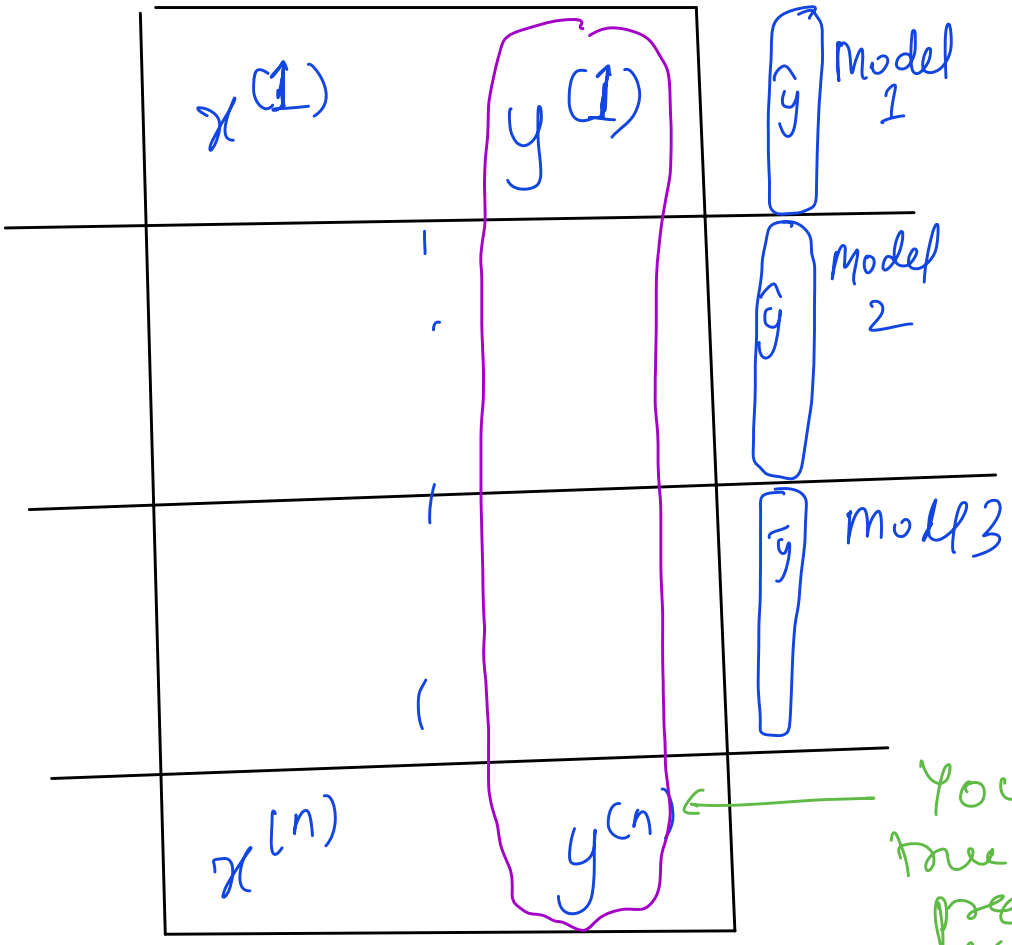
The above method is Hold Out Cross validation

K-Fold Cross Validation

Small Datasets

leave one out Cross Validation

$K = n$

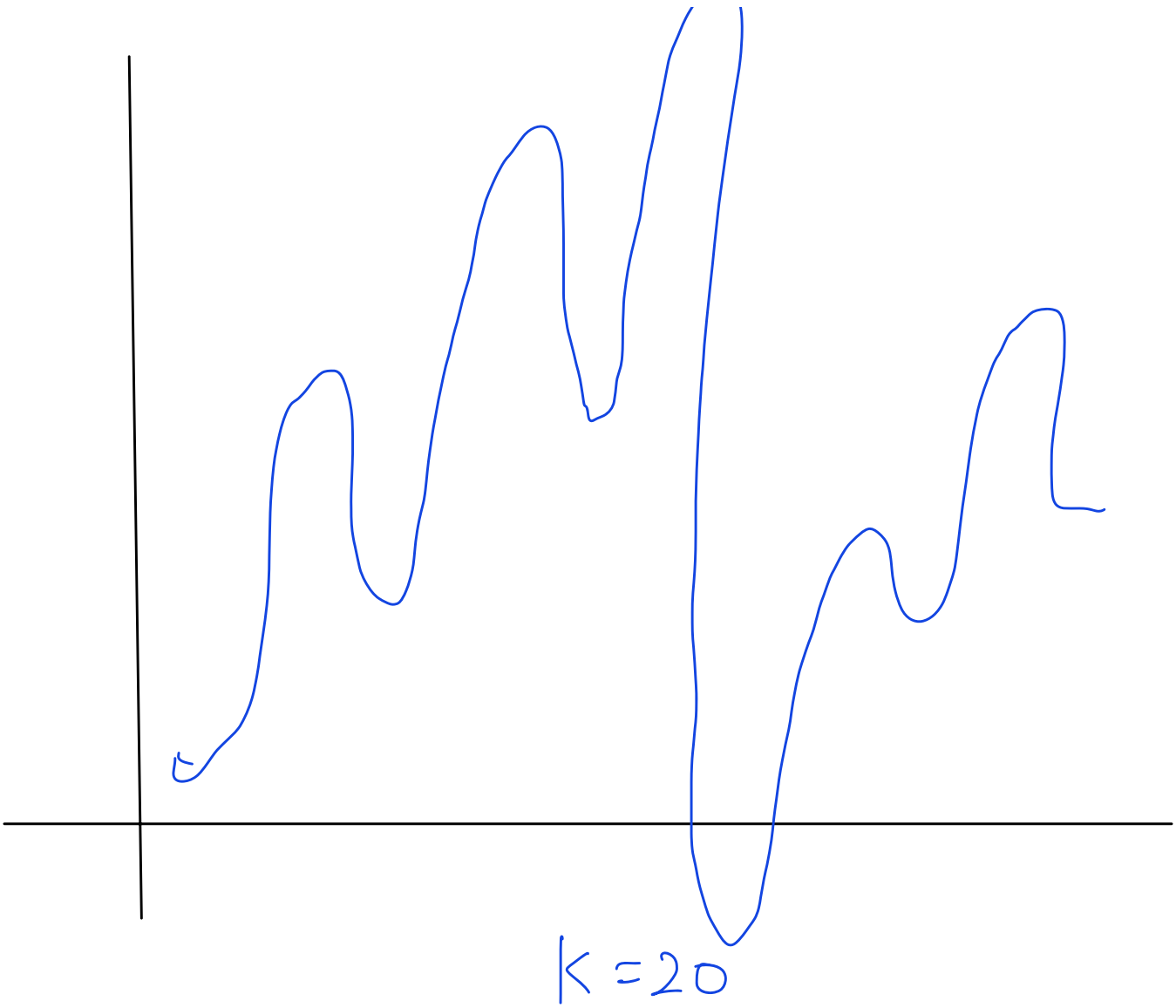


n ~~parts~~ division of data set

Model 1, fold 1 is val set
2-n → train set
Prediction on fold 1

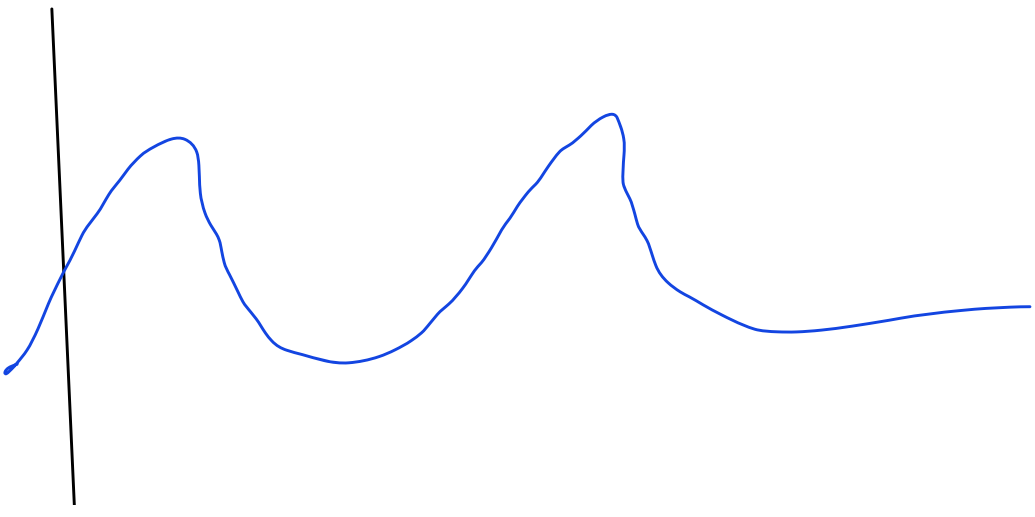
Model 2, fold 2 is val. set
1, 3-n → trainset
pred. on fold 2
⋮ So on

Then, you can either ensemble all k-model & average their pred. at test time
Use this technique for hyperpar. tuning & then refit on entire set



$K = 20$

$|\theta| >> 0 \rightarrow bad$



You want less wigly but still want it to be expressive

||

Regularization

Encouraging small $||\theta||$ is Regularization

$$J(\theta) = \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right)^2 + \lambda ||\theta||_2^2$$

$\lambda \rightarrow$ relative weighing factor

$h_\theta(x) \rightarrow \theta^T x$