April 2, 2025 - lecture #1 -

Notation - end of Chapter 1.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$\mathbf{X}$: capital letter for matrices. (bold in textbook)

$n \times p$: n = # of obs, p = # of features

$i = 1, \ldots, n$

$j = 1, \ldots, p$

$\mathbf{X}_i$ = i^{th} row of $\mathbf{X} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$

$\mathbf{X}_j = \vec{X}_j$ = j^{th} column of $\mathbf{X} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$

* all vector are column vectors *

* vectors of length n are bold (or have arrow on top) *

$\vec{a}$ or $\mathbf{a}$: vector of length n.

a : could be a vector not of length n, or it could be a scalar.

$\vec{a} \in \mathbb{R}^n$ } vector of length n w/ elts in $\mathbb{R}$

$A \in \mathbb{R}^{w \times v}$ } matrix of dim. $w \times v$

$c \in \mathbb{R}$ } scalar

$a \in \mathbb{R}^q$ } vector of length q.

$a \in \{0, 1\}$] a can equal 0 or 1

$A \in \{1, 2, 3, 4, 5\}^{173 \times 12}$] 173×12 matrix, w/ elts that are 1, 2, 3, 4, or 5.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix}_{2 \times 2}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix}_{2 \times 1}$$

Chapter 2 - What is statistical learning?

Response $Y \in \mathbb{R}$: what we want to predict.
↳ outcome, ...

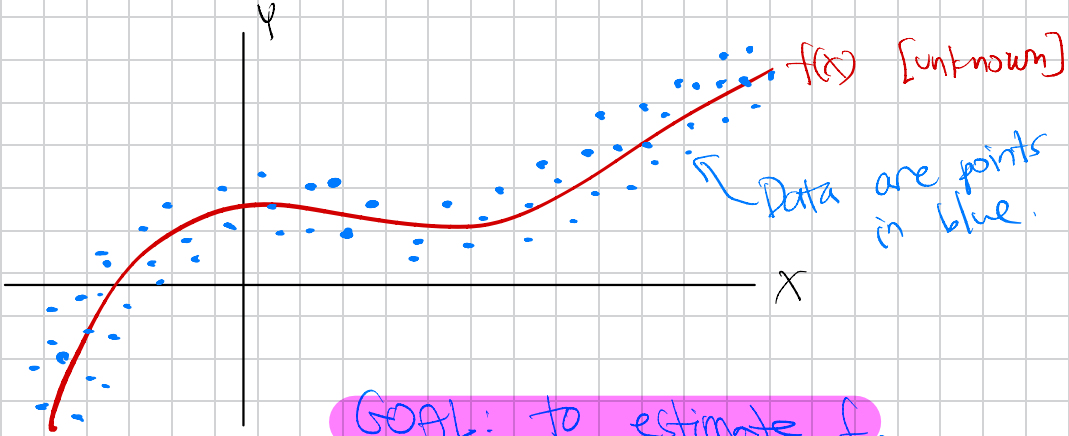
Predictors $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^p$
↳ features, covariates

mean-zero noise term $E(\varepsilon) = 0$.

$$Y = f(X) + \varepsilon = f(X_1, \ldots, X_p) + \varepsilon$$

f is UNKNOWN.
our goal is to est f.

$f : \mathbb{R}^p \to \mathbb{R}$



GOAL: to estimate f.

($\hat{f}$ is our estimate of f.)

Why do I want to est. f?

Prediction

Inference

PREDICTION: we want a "good" est. of f, so that we can compute $\hat{Y} = \hat{f}(x)$ for a new X.

e.g. I want to predict value of house based on # BR's, proximity to light rail, etc

INFERENCE: we want to understand the relationship btw $X_1, \ldots, X_p$ & Y, b/c we are interested in it

e.g. how much would this house's value increase if we added another BR?

How to estimate $f(\cdot)$?

Training data: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad y \in \mathbb{R} \text{ (usually)}$$

Goal: $\hat{f}(\cdot)$ such that $\hat{f}(x) \approx y$ for all (x, y).

→ parametrically (Option 1)

→ non-parametrically (Option 2)

Option 1: Estimate f parametrically

Step 1: "assume" some functional form for relationship btw X & y. For example

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}_{f(x)} + \varepsilon$$

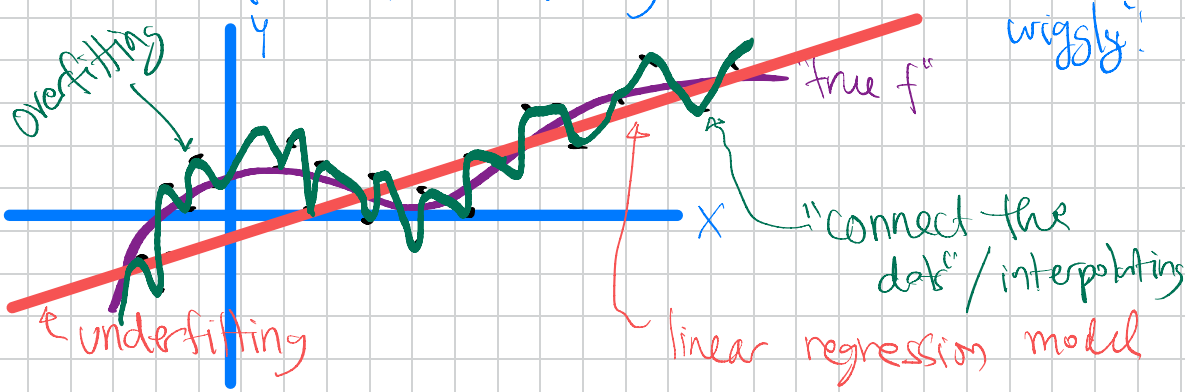
Step 2: Use $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ to "fit the model", i.e. estimate $\beta_0, \beta_1, \ldots, \beta_p$ (call the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$).

Examples: linear regression (L2)
logistic regression (L3)
⋮

Option 2: Estimate f non-parametrically.

⇒ Do NOT make explicit assumptions about the form of f.

⇒ Instead, find $\hat{f}$ that is close to the training data & "not too wiggly".



Why would we prefer parametric versus non-parametric?

Parametric:

(+) simple → easy to interpret -- useful for inference.
→ easy computationally (?)

(+) guard against overfitting: simplifies estimation task, which helps us re: bias-variance trade-off. (B-V T-O)
especially when sample size (n) small.

(+) we might "believe" the parametric model

(−) strong assumption about f, & it might be wrong!

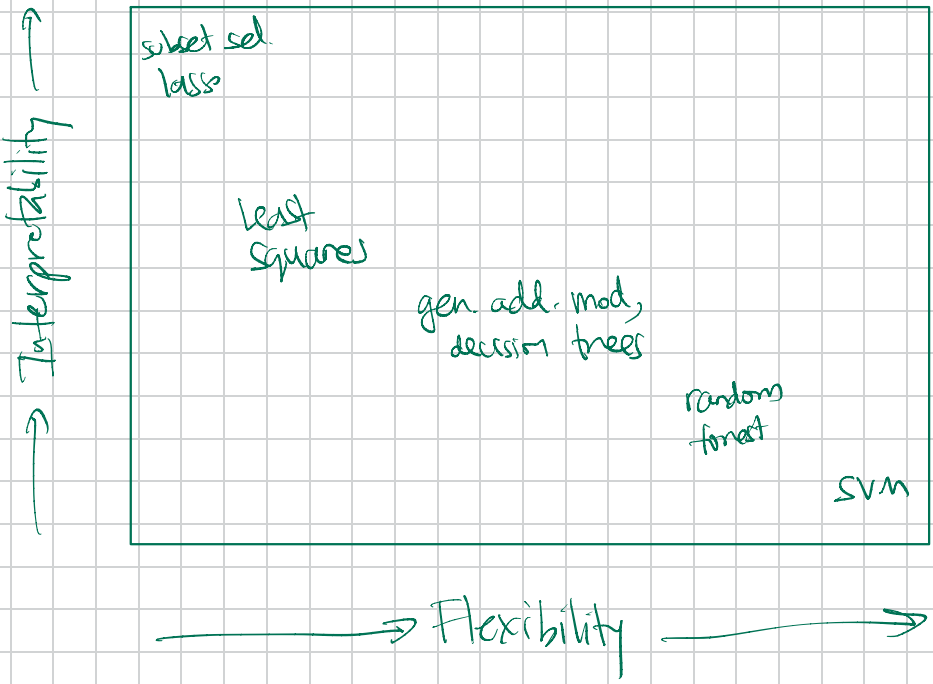
Non-Parametric

⊕ super flexible!! no assumption on $f(\cdot)$ is needed.

⊖ super flexible!! get hurt by B/V T/O. (overfitting).

⊖ LARGE SAMPLE SIZE.
(+ relatively few features)
$n >>> p$.

⊖ not great for inference/interpretation.


subset sel.
lasso
least squares
gen. add. mod, decision trees
random forest
SVM
Interpretability
Flexibility

DL

2 Types of Statistical Learning

Supervised

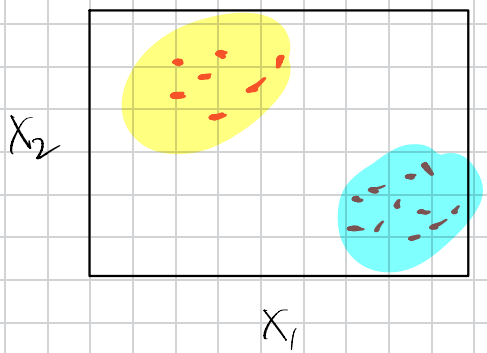
(X, Y) pairs.

$Y = f(X) + \varepsilon$

e.g., regression, classification, etc., DL, xgboost, RFs, ...

Unsupervised

No Y! Just X's.



e.g. clustering, dimension reduction, etc

Supervised Learning

Regression: $Y \in \mathbb{R}$

quantitative cont real-valued

Classification

$Y \in \{\text{blue, brown, green, black, hazel}\}$

$Y \in \{0, 1\}$

discrete, qualitative, categorical

$Y = f(X) + \varepsilon$, $E(\varepsilon) = 0$

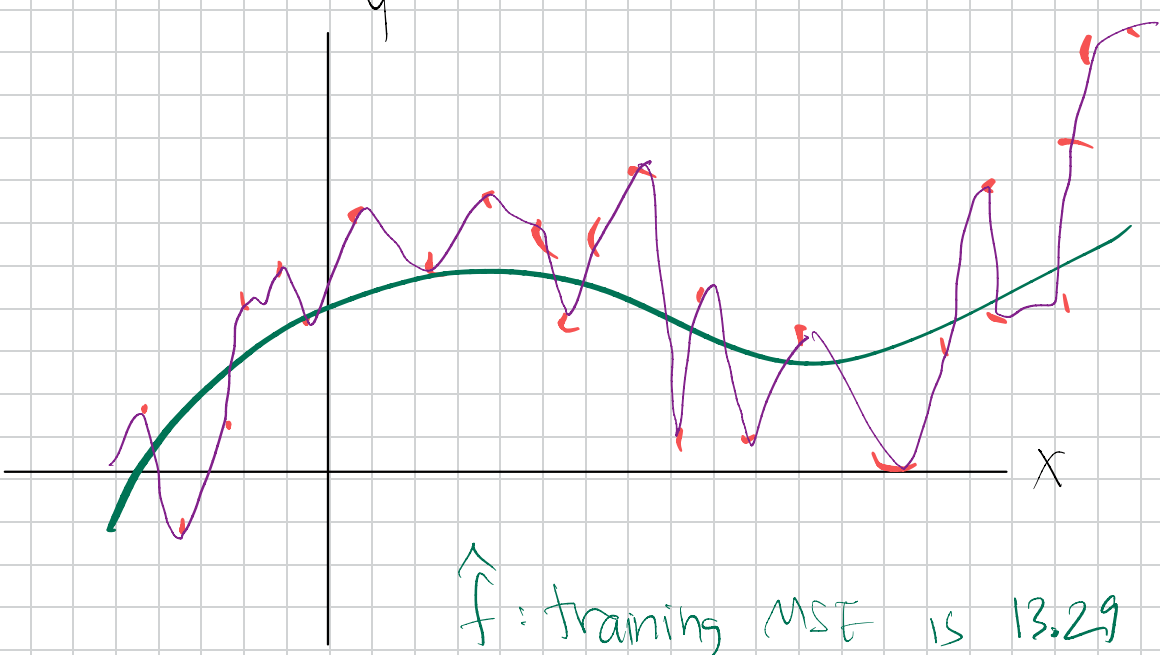
unknown.

How do we assess model accuracy?

(Training) $MSE := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$

↑ evaluated on same n obs. used to est $\hat{f}$ (training set).

We want (training) MSE to be small, i.e., $y_i \approx \hat{f}(x_i)$, $i = 1, \ldots, n$



$\hat{f}$: training MSE is 13.29

$\hat{f}$: training MSE is 0.000000

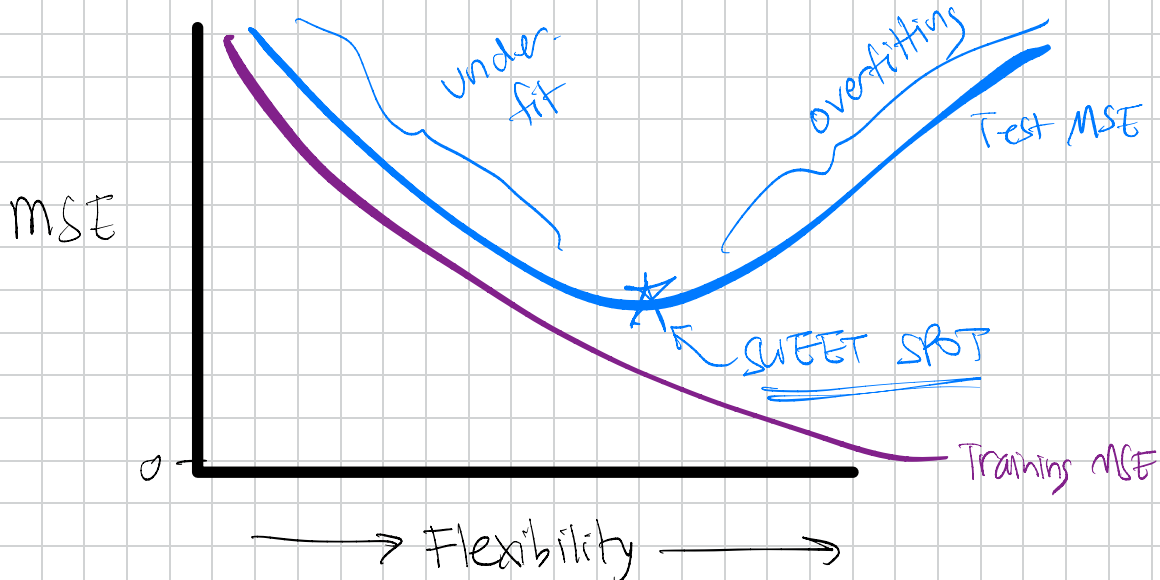
What I actually care about is ~~naught~~ not

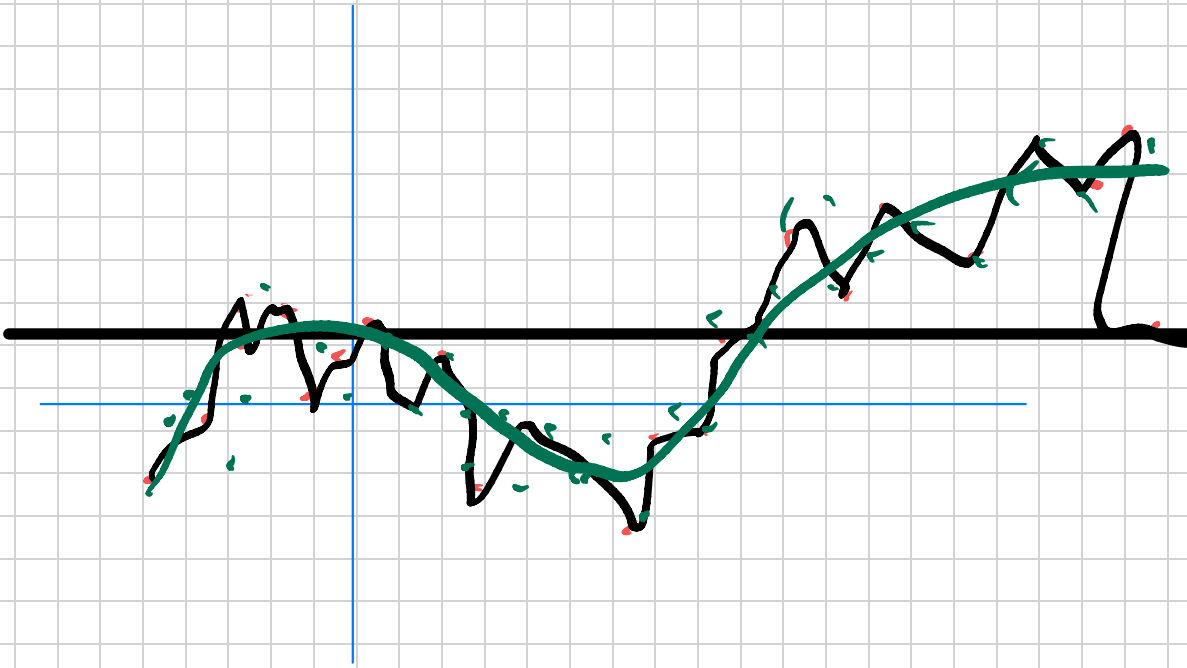
test MSE: for a "test" obs (x_0, y_0) that wasn't in training set, & that's drawn from same distribution as training set,

$$\text{Test MSE} = \text{Average}\left[(y_0 - \hat{f}(x_0))^2\right]$$

average over a whole bunch of (x_0, y_0)'s.

I want THIS to be small!





Bias-Variance Trade-off

$Y = f(X) + \varepsilon, \quad E(\varepsilon) = 0$

f: unknown; est w/ training set

Test obs. (x_0, y_0)

In this class: X fixed (not random).

ε random,

so Y is random b/c of ε

Expected prediction error (exp test MSE)

$$EPE := E\left[\left[y_0 - \hat{f}(x_0)\right]^2\right]$$

$$EA^2 = E(A^2) \neq (EA)^2 = [E(A)]^2$$

expectation over (x_0, y_0)'s AND over training set

$$E\left[(y_0 - \hat{f}(x_0))^2\right]$$

$y = f(x) + \varepsilon$, $E(\varepsilon) = 0$

$$= E\left[\left(\underbrace{y_0 - f(x_0)}_{A} + \underbrace{f(x_0) - E\hat{f}(x_0)}_{B} + \underbrace{E\hat{f}(x_0) - \hat{f}(x_0)}_{C}\right)^2\right]$$

Note: $(A+B+C)^2 = A^2 + B^2 + C^2 + 2AB + 2AC + 2BC$

$$= E\left[A^2 + B^2 + C^2 + 2AB + 2AC + 2BC\right]$$

$$= E[A^2] + E[B^2] + E[C^2] + 2E[AB] + 2E[AC] + 2E[BC]$$

$$E(A^2) = E\left[\underbrace{(y_0 - f(x_0))^2}_{\varepsilon_0}\right] = E(\varepsilon^2) = E(\varepsilon^2) - (E\varepsilon)^2 = \mathrm{Var}(\varepsilon)$$

$$E(B^2) = E\left[\underbrace{(f(x_0)}_{\text{not random}} - \underbrace{E\hat{f}(x_0))}_{\text{not random}}^2\right] = \underbrace{\left[f(x_0) - E\hat{f}(x_0)\right]^2}_{= \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2}$$

recall: parameter θ, estimator $\hat{\theta}$

$$(E\hat{\theta} - \theta)^2 = \left(\mathrm{BIAS}(\hat{\theta})\right)^2$$

$$E(C^2) = E\left[\left(\hat{f}(x_0) - E\hat{f}(x_0)\right)^2\right] = \mathrm{Var}\left[\hat{f}(x_0)\right]$$

recall: param θ, estimator $\hat{\theta}$,

$$E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2\right] = \mathrm{Var}(\hat{\theta})$$

$$E\left[\underbrace{(Y_0 - f(x_0))}_{A} + \underbrace{f(x_0) - E\hat{f}(x_0)}_{B} + \underbrace{E\hat{f}(x_0) - \hat{f}(x_0)}_{C}\right]^2$$

$(E(3x) = 3E(x))$

$$E(AB) = E\left[(Y_0 - f(x_0))\underbrace{(f(x_0) - E\hat{f}(x_0))}_{\text{not random}}\right]$$

$$= (f(x_0) - E\hat{f}(x_0))\, E\underbrace{[Y_0 - f(x_0)]}_{\varepsilon_0}$$

$$= (f(x_0) - E\hat{f}(x_0))\, \underbrace{E(\varepsilon)}_{=0} = 0$$

$$E(AC) = E\left[\underbrace{(Y_0 - f(x_0))}_{\varepsilon_0}\underbrace{(E\hat{f}(x_0) - \hat{f}(x_0))}_{\text{not random}}\right]$$

$\hat{f}(x_0)$ random b/c of $\varepsilon_1, \ldots, \varepsilon_n$

$Y_0 = Y_0 - f(x_0) + f(x_0)$

$$= E(\varepsilon_0) \cdot E\left[E\hat{f}(x_0) - \hat{f}(x_0)\right]$$

$$= 0$$

$$E\left[\underbrace{(y_0 - f(x_0))}_{A} + \underbrace{f(x_0) - E\hat{f}(x_0)}_{B} + \underbrace{E\hat{f}(x_0) - \hat{f}(x_0)}_{C})^2\right]$$

$$E(BC) = E\left[(f(x_0) - E\hat{f}(x_0)) \cdot (E\hat{f}(x_0) - \hat{f}(x_0))\right]$$

$(f(x_0) - E\hat{f}(x_0))$: not random

$$= (f(x_0) - E\hat{f}(x_0)) \cdot E\left[E\hat{f}(x_0) - \hat{f}(x_0)\right]$$

$$= E\hat{f}(x_0) - E\hat{f}(x_0) = 0$$

$$= 0$$

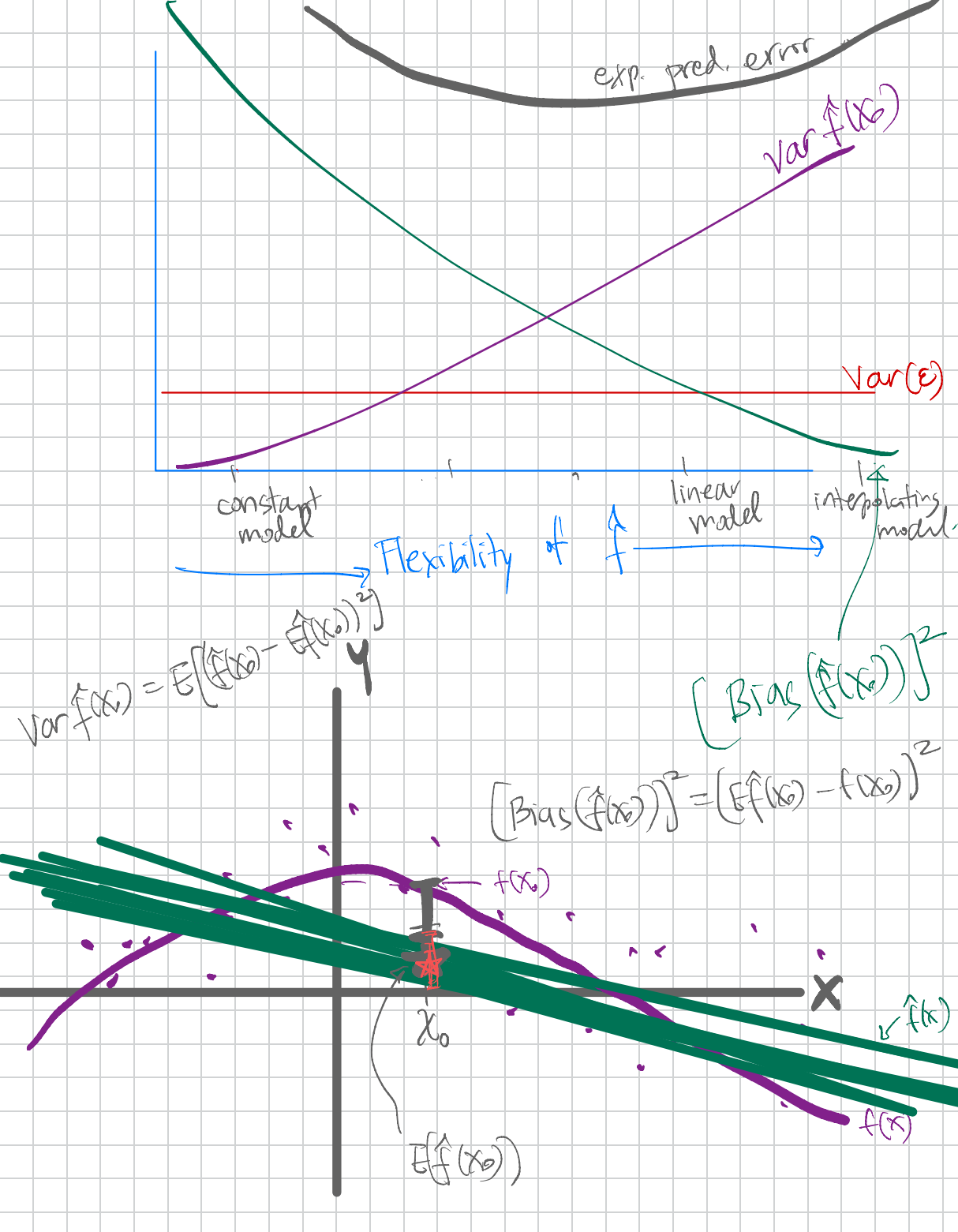
$$EPE(x_0) := E\left[(y_0 - \hat{f}(x_0))^2\right]$$

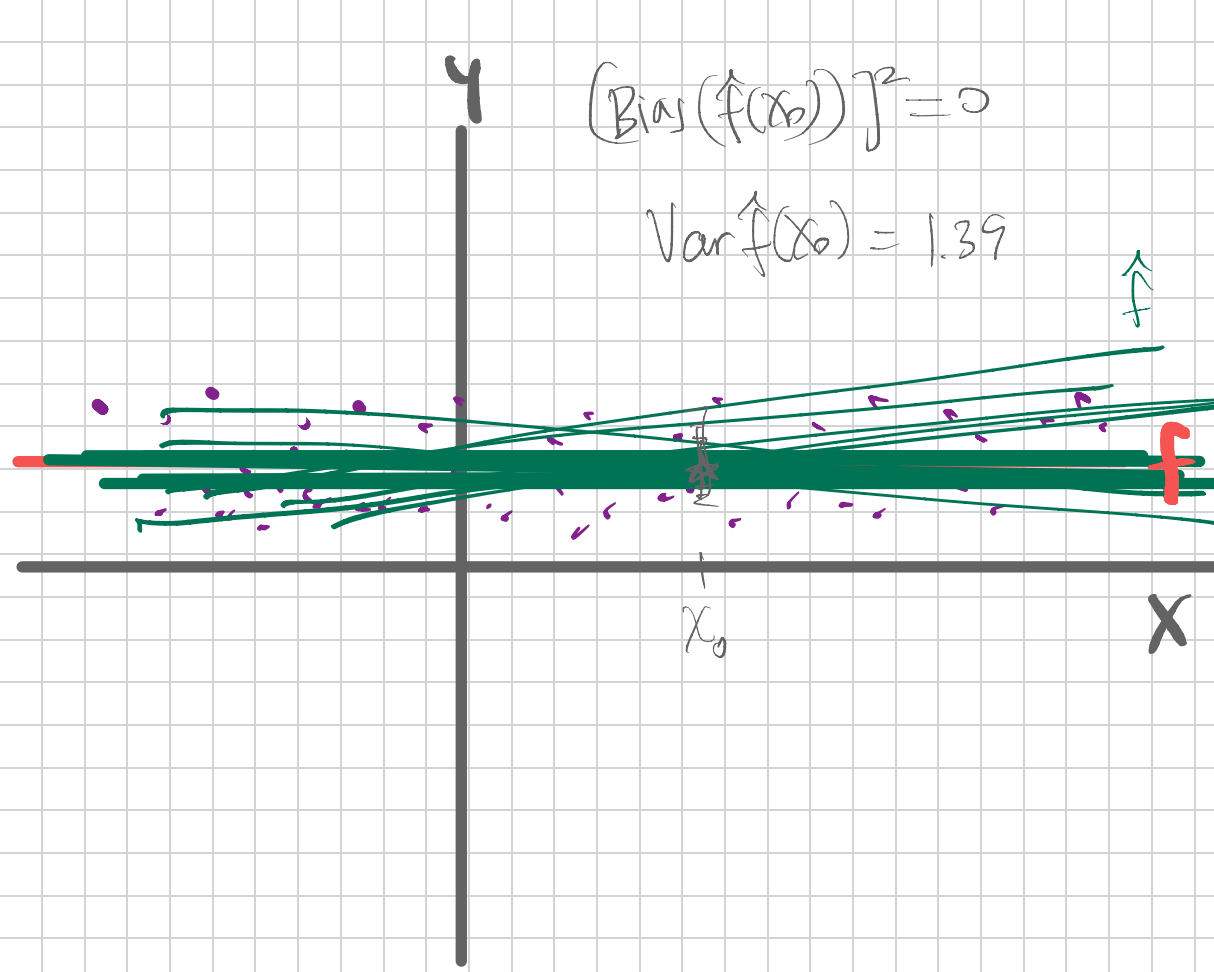
$$= \boxed{Var(\varepsilon)} + \boxed{\left[Bias(\hat{f}(x_0))\right]^2 + Var\left[\hat{f}(x_0)\right]}$$

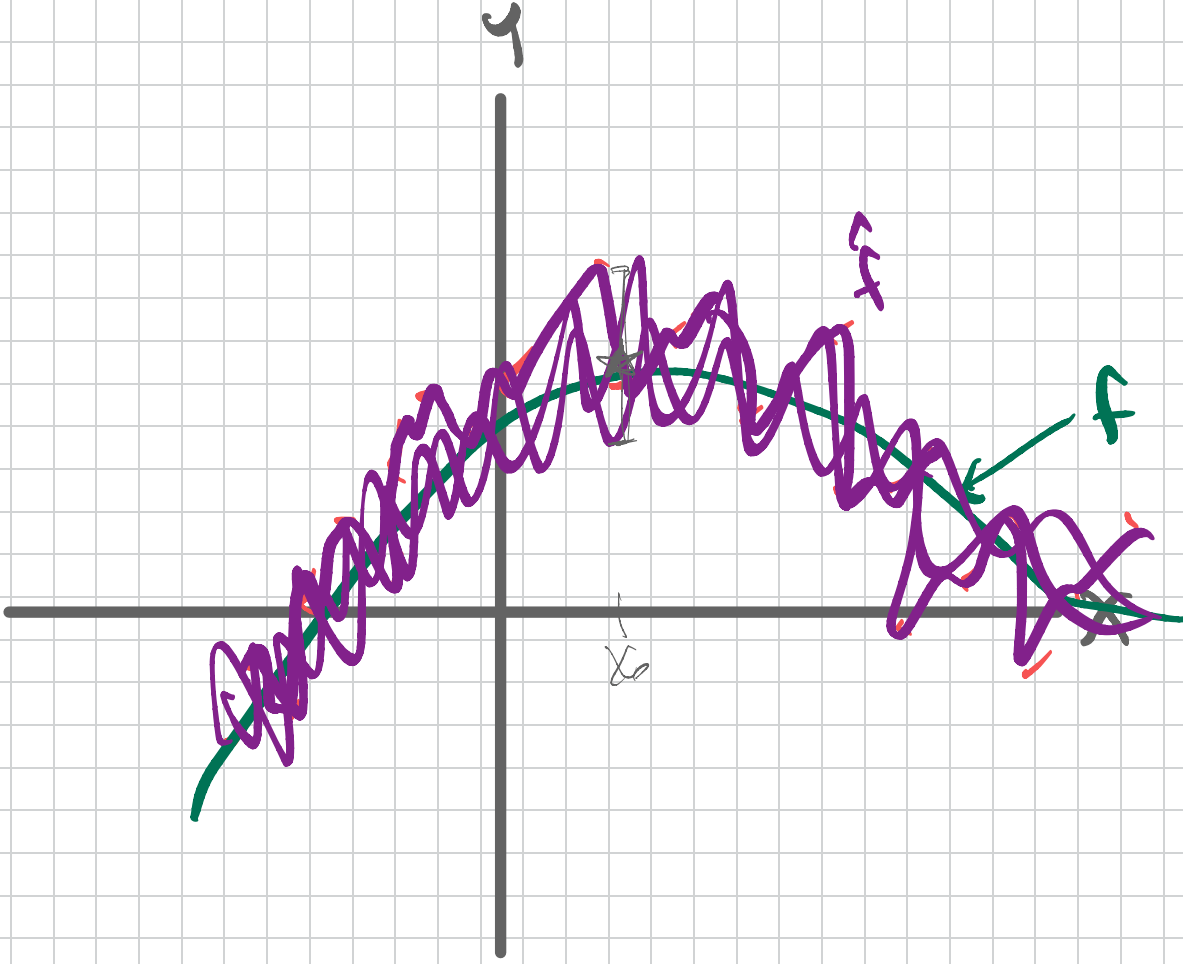
BIAS-VARIANCE TRADE-OFF

irreducible error.

reducible error.


exp. pred. error
Var $\hat{f}(x_0)$
Var(ε)
constant model
linear model
interpolating model
Flexibility of $\hat{f}$
$\text{Var}\hat{f}(x_0) = E\left[(\hat{f}(x_0) - E\hat{f}(x_0))^2\right]$
$\left[\text{Bias}(\hat{f}(x_0))\right]^2$
$\left[\text{Bias}(\hat{f}(x_0))\right]^2 = (E\hat{f}(x_0) - f(x_0))^2$
Y
X
$f(x)$
x_0
$\hat{f}(x)$
$f(x)$
$E\hat{f}(x_0)$


$[\text{Bias}(\hat{f}(x_0))]^2 = 0$
$\text{Var}\,\hat{f}(x_0) = 1.39$
$\hat{f}$
f
Y
X
x_0



$$[\text{Bias}\{\hat{f}(x_0)\}]^2 \approx 0$$

$\text{Var}(\hat{f}(x_0))$ is big

Classification:

$y \in \{1, 2, \ldots, K\}$

~~$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$~~

(training) (mis) classification error:

$$\frac{1}{n} \sum_{i=1}^{n} 1(y_i \neq \hat{y}_i) \rightarrow \hat{f}(x_i)$$

$$1_A = \begin{cases} 1 & \text{if } A \text{ holds} \\ 0 & \text{else} \end{cases}$$

$$1(y_i \neq \hat{y}_i) = \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \\ 0 & \text{otherwise} \end{cases}$$

test (mis) classification error

Average $(1(y_0 \neq \hat{y}_0)) \rightarrow \hat{f}(x_0)$

for test obs y_0.

Bayes classifier ← unattainable dream

↳ assigns a test obs. x_0 to the class for which

$$\Pr(Y=k \mid X=x_0)$$ is largest.

don't know this!

need to est. this

The Bayes classifier minimizes the expected test (mis)classification error

"Bayes error"

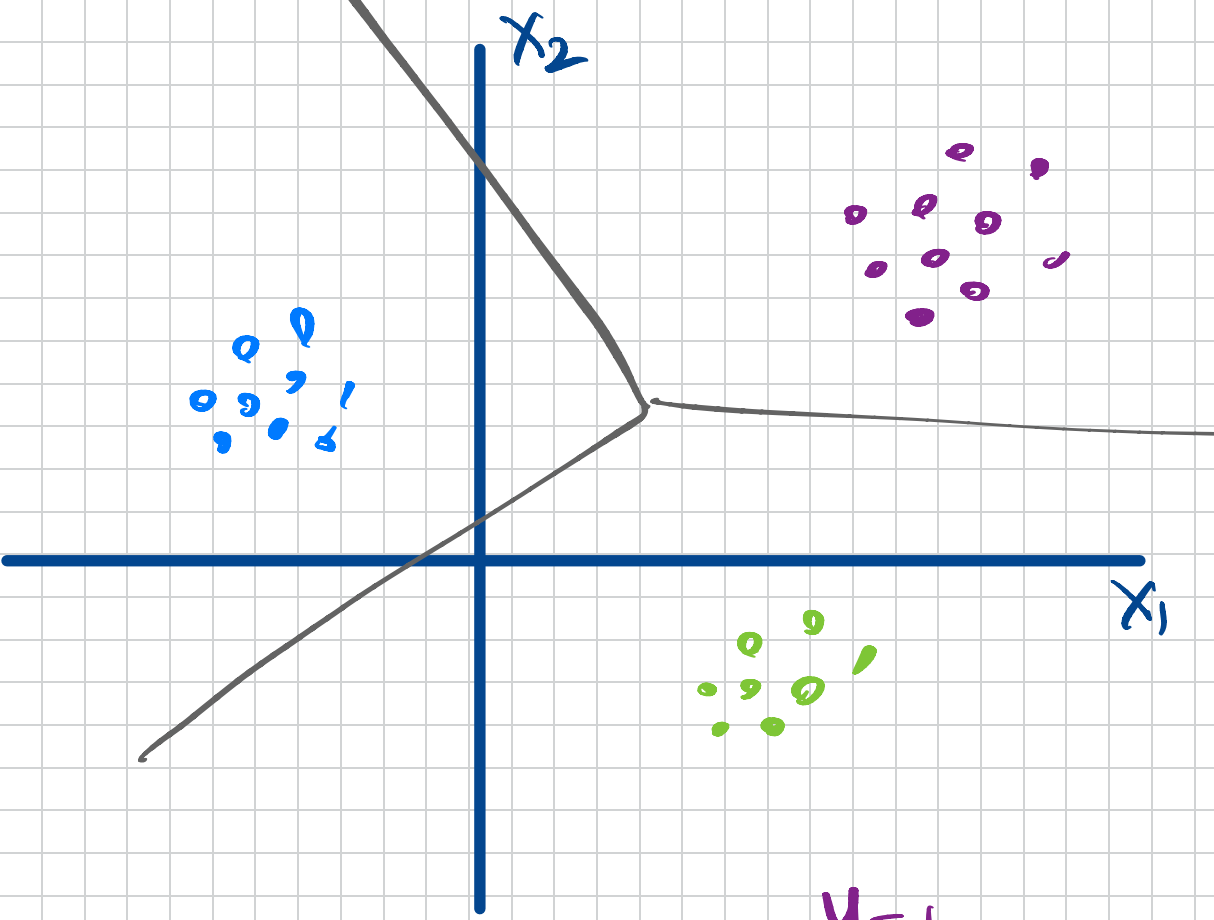
Error of Bayes classifier:

$$1 - E\left[\max_k \Pr(Y=k \mid X=x_0)\right]$$

$\Pr(Y=\text{cat} \mid X=x_0) = 0.7$

$\Pr(Y=\text{dog} \mid X=x_0) = 0.2$

$\Pr(Y=\text{monkey} \mid X=x_0) = 0.1$

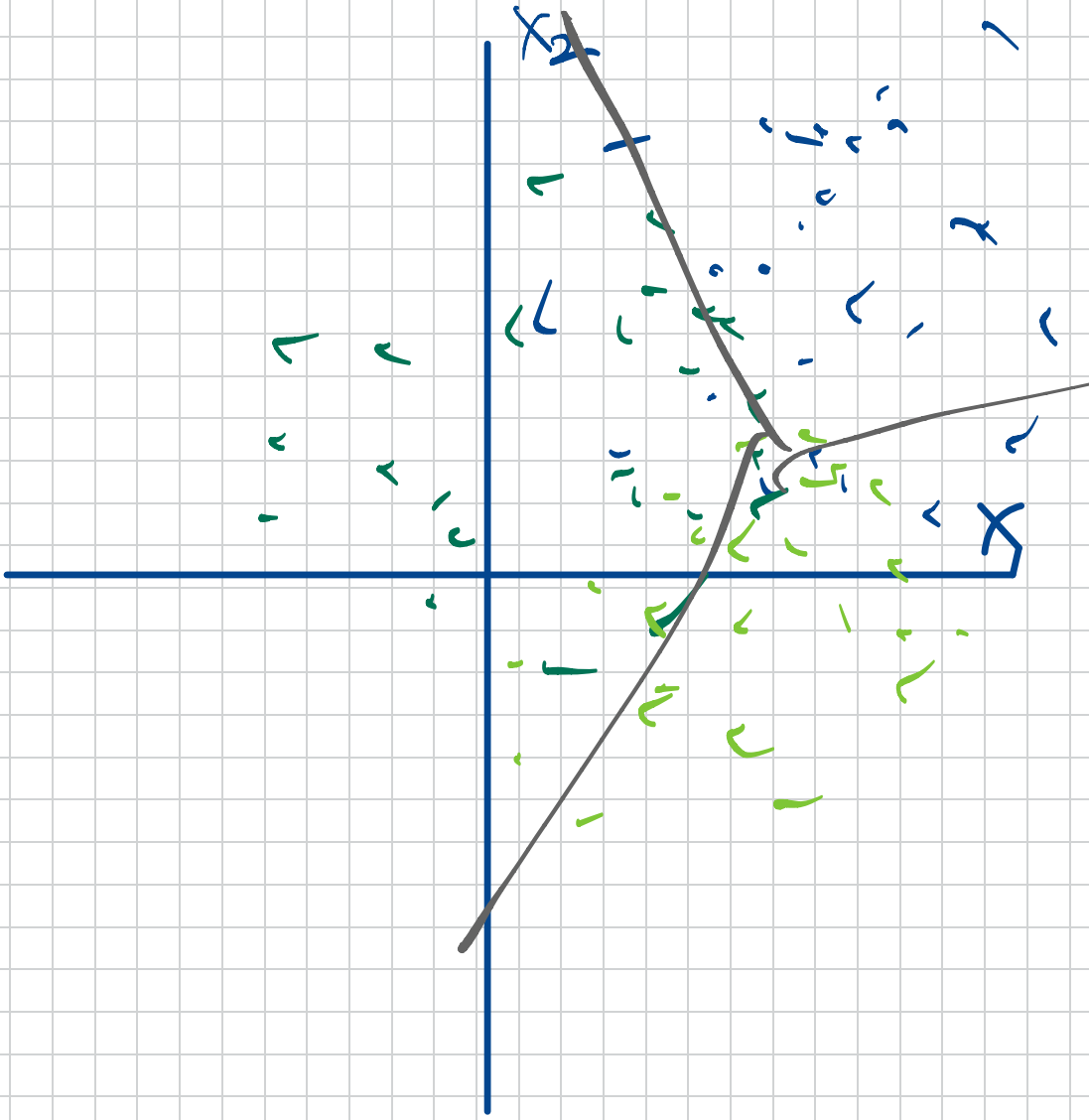


$Y=1$

$Y=2$

$Y=3$

Bayes error is TINY



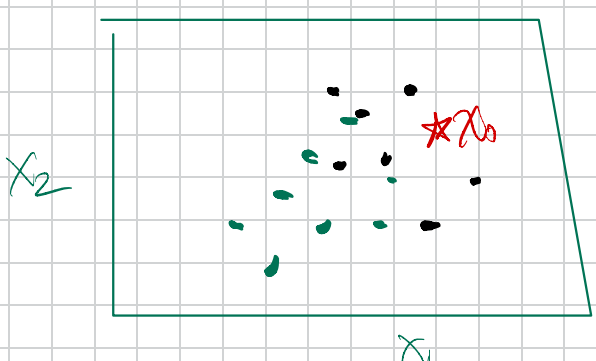
K-nearest neighbors.

Non-parametric approach to est.

$$\Pr(Y=j \mid X=x_0)$$

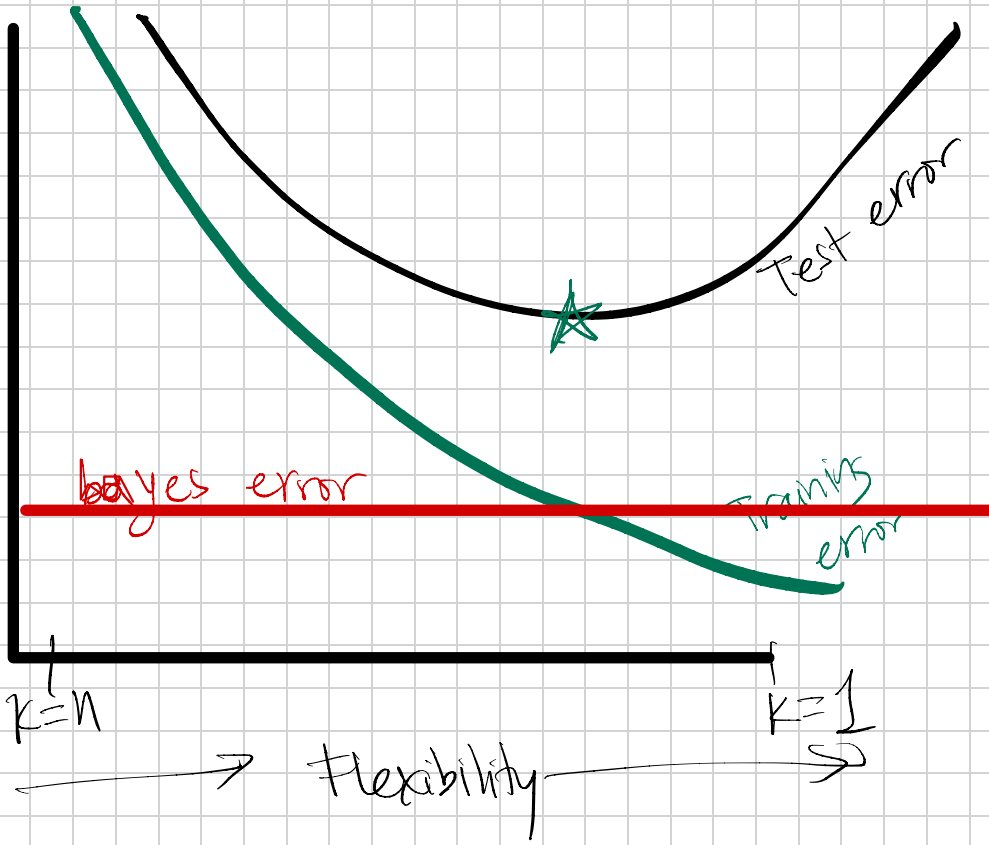
$$\approx \frac{1}{K} \sum_{i \in N(x_0)} 1(y_i = j)$$

where $N(x_0)$ are the indices of the K "nearest neighbors" to x_0 in the training set.



K=1: zero misclass. error on training set

K=5: nonzero misclass error on test.


test error
bayes error
training error
K=N
K=1
Flexibility