Bias - Variance Tradeoff

Perf. of method has two comps.:

<u>Bias</u>: err from approx. complicated real world phenom w/ a simple fn.

<u>Variance</u>: err b/c our $\hat{f}$ is sens. to D_{train}

Often, are opposition.

<u>Can formulate mathy</u>:

$$Y = f(\underset{\sim}{X}) + \varepsilon, \quad \varepsilon \text{ random}$$
$$E[\varepsilon] = 0$$
$$Var(\varepsilon) = \sigma^2$$
$$\underset{\sim}{X} \perp \varepsilon$$

gen. according to this model

Let $D_{train} = \{(\underset{\sim}{X}_n, y_n)\}_{n=1}^N$ be used to

Let $D_{train} = \{(\underset{\sim}{x}_n, y_n)\}_{n=1}$ be used to fit $\hat{f} = \hat{f}_{D_{train}}$.

Let $\underset{\sim}{X}_{new}, Y_{new}$ be indep draws from this model.

Defn
$$MSE = E\left[\left(Y_{new} - \hat{f}(X_{new})\right)^2\right]$$

Assume X_{new} is fixed (not random)

Then if we define

$$Bias(\hat{f}) = E[\hat{f}(X_{new})] - f(X_{new})$$

$$Var(\hat{f}) = Var(\hat{f}_{D_{train}}(X_{new}))$$

Can show:

reducible(?)

$$MSE = Bias(\hat{f})^2 + Var(\hat{f}) + \sigma^2$$

irreducible

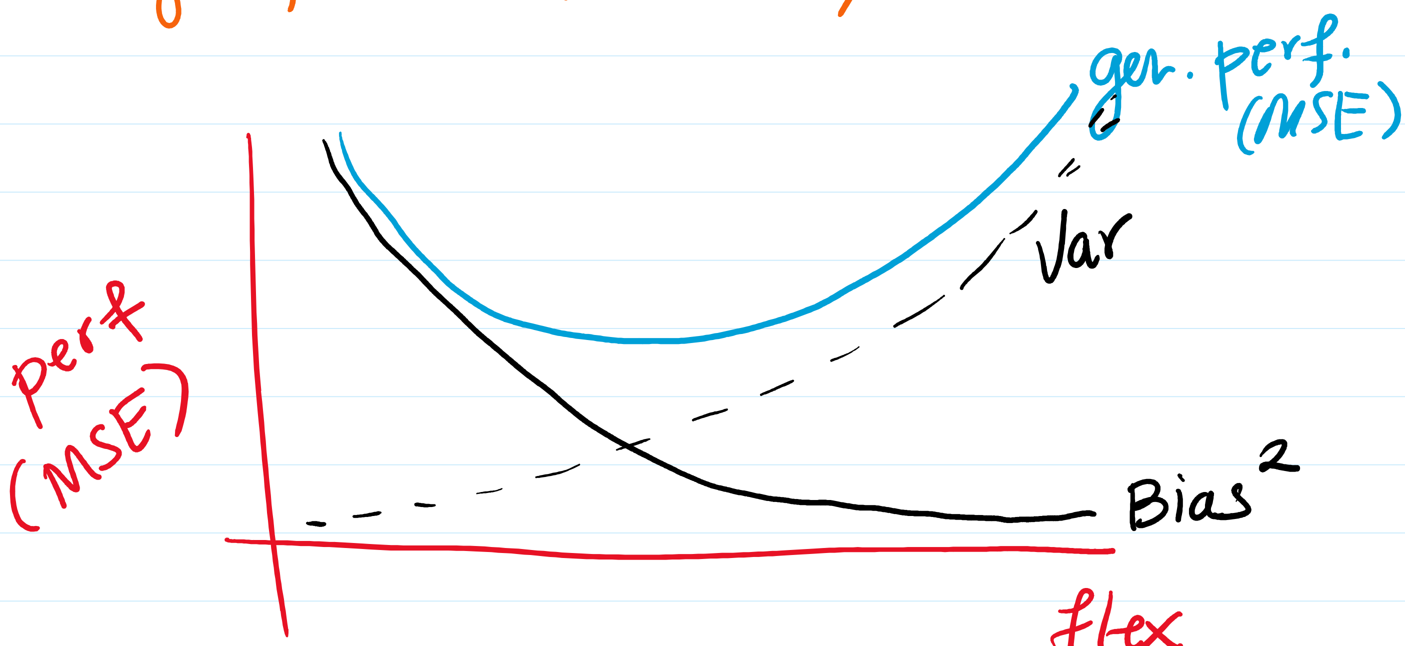
irreducible error

Called BV decomp.

Typ. bias and var are opposed.

low flex $\Leftrightarrow$ high bias / low variance

high flex $\Leftrightarrow$ low bias / high var.



Q: what's the best $\hat{f}$ theoretically?

L = loss fn

e.g. $L(y, f(x)) = (y - f(x))^2$ (sq. loss)

e.g. $L(y, f(x)) = |y - f(x)|$ (abs. loss)

$$f^* = \underset{f}{\arg\min}\, E[L(Y, f(X))]$$

$(X, Y) \sim P$ joint dist of x, y

Total Exp.

$E[A] = E_B E[A|B]$

Can get answer;

Consider $E[L(Y, f(X))]$

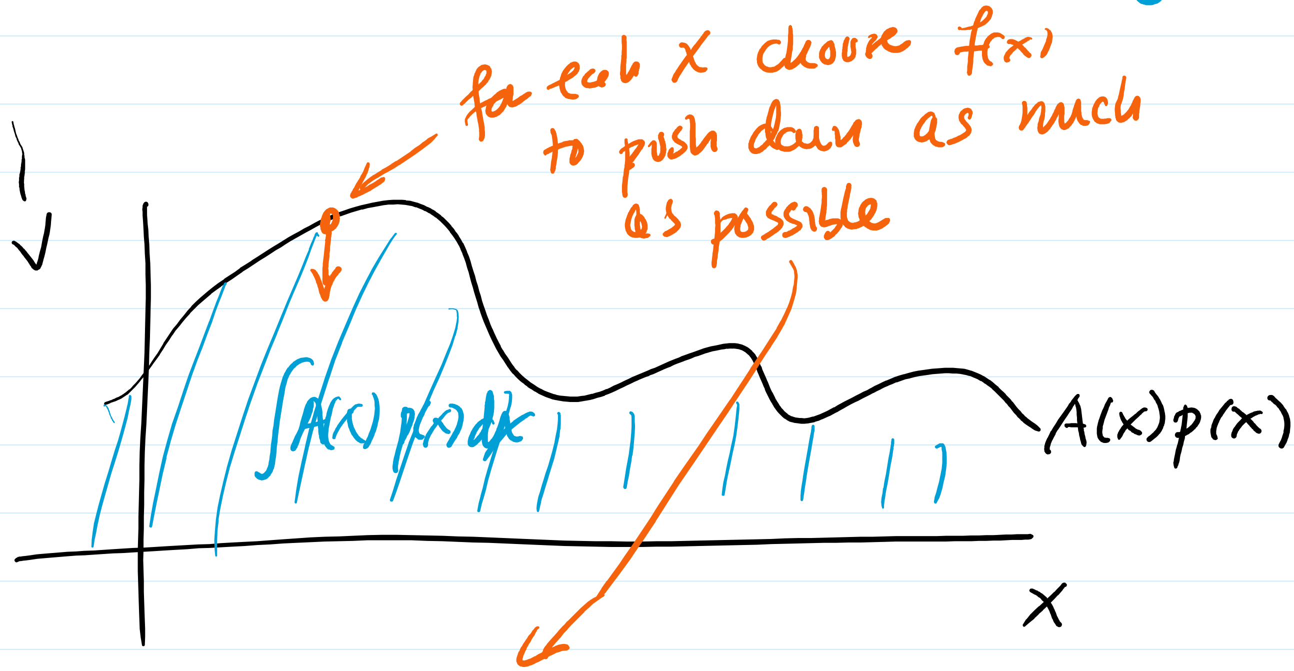
$$= E_X \underbrace{E[L(Y, f(X)) | X]}_{A(x)}$$

$$= \int A(x) p(x) dx$$

dens of x

depends on $f(x)$

Can choose $f(x)$ for each x separately.



Can choose f to do this sep. for each x

Since $p(x)$ doesn't change based on $f(x)$

all I need to look at is $A(x)$.

To find f^* just need to optimize

$$E[L(Y, f(x)) | X]$$

$E[L(Y, f(X))|X]$

sep for each x.

$$\boxed{f^*(x) = \arg\min_{f(x)} E[L(Y, f(x))|X]}$$

Consider Sq. loss $L(Y, f(x)) = (Y - f(x))^2$

$$f^*(x) = \arg\min_{c \in \mathbb{R}} E[(Y-c)^2 | X]$$

Ex. $\arg\min_c E[(Z-c)^2] = E[Z]$

pf. $E[(Z-c)^2] = E[Z^2 - 2ZC + C^2]$

$$= E[Z^2] - 2c\,E[Z] + c^2$$

take deriv wrt c:

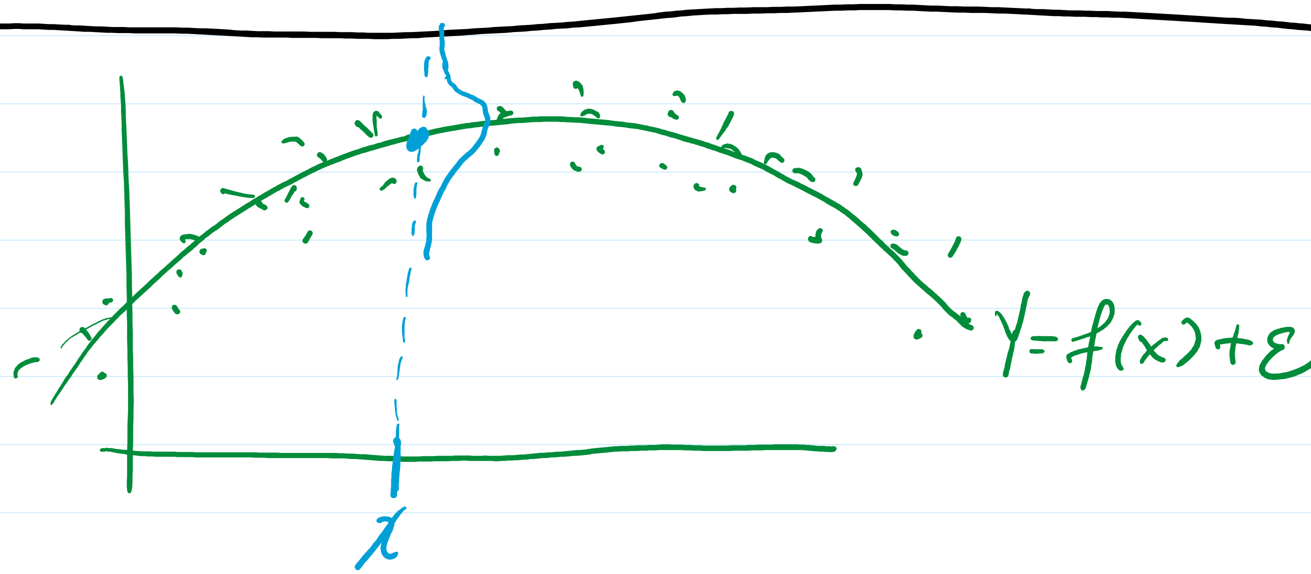
$$\frac{\partial}{\partial c}[\dots] = -2E[Z] + 2c = 0$$

$$\frac{\partial}{\partial c}\left[\ \cdots\ \right] = -2E[Z] + 2c = 0$$

Solve for c to get

$$c = E[Z].$$

So $$\boxed{f^*(x) = E[Y \mid X = x]}$$



If $L(Y, f(X)) = |Y - f(X)|$

$$f^*(x) = \text{Median}(Y \mid X = x)$$

In reality, one way to build $\hat{f}$ is as

In reality one way to build $\hat{f}$ is as

$$\hat{f}(\underset{\sim}{x}) \simeq E[Y \mid \underset{\sim}{X} = \underset{\sim}{x}]$$

using training data.

e.g. $\hat{f}(\underset{\sim}{x})$ = avg. ys for xs near $\underset{\sim}{x}$

Called KNN regression.

e.g. make some assumption about form of

$$E[Y \mid \underset{\sim}{X} = \underset{\sim}{x}]$$

maybe linear(?)

$$\hat{f}(\underset{\sim}{x}) \approx E[Y \mid \underset{\sim}{X} = \underset{\sim}{x}] = \underset{\sim}{x}^T \beta$$

Called OLS lin. regr.