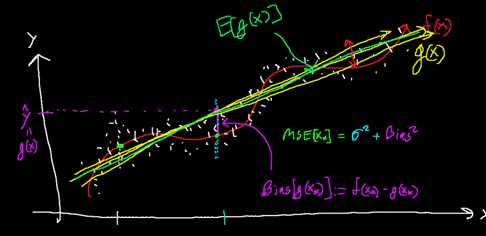


$$\delta_1, \ldots, \delta_n \overset{iid}{\sim} P(\Delta)$$

$$Y_* = f(x_*) + \Delta$$

$$Var[W] := E\left[(W - E[W])^2\right]$$

$$E[Y \mid x = x_0] = f(x_0)$$

Last class, everything was constant except for the δ's at x*.
Let's step it up a notch. Let's allow for many different datasets boldD where X (n x p) matrix of feature values is constant but y1,...,yn are drawn from the rv's Y1,..., Yn meaning δ1, ..., δn are draws from Δ1, ..., Δn. We still want to find the MSE just at x*.

=> Because the D's are random => g becomes random (g is fixed only with a fixed D).

$E_X[17] = 17$

$$MSE(x_*) = E_{\Delta_1, \Delta_2, \ldots, \Delta_n, \Delta_*}\left[(Y_* - g(\vec{x}_*))^2 \mid X, \vec{x}_*\right]$$

$$= E_{\Delta_1, \ldots, \Delta_n, \Delta_*}[Y_*^2] - 2E_{\Delta_1, \ldots, \Delta_n, \Delta_*}[Y_* g(x_*)] + E_{\Delta_1, \ldots, \Delta_n, \Delta_*}[g(x_*)^2]$$

see lec 22

$$= E_{\Delta_*}[Y_*^2] - 2E_{\Delta_*}[Y_*] E_{\Delta_1, \ldots, \Delta_n}[g(x_*)] + E_{\Delta_1, \ldots, \Delta_n}[g(x_*)^2]$$

$$= f(x_*)^2 + \sigma^2 - 2f(x_*) E[g(x_*)] + Var[g(x_*)] + E[g(x_*)]^2$$

$$= \sigma^2 + \underbrace{(f(x_*) - E[g(x_*)])}_{Bias[g(x_*)]}{}^2 + Var[g(x_*)]$$

$$= \sigma^2 + Bias[g(x_*)]^2 + Var[g(x_*)]$$

Let's step it up two more notches in one shot (a) Let x1, ..., xn be random realizations from P(X), thus X is random and (b) x* is a random realization from P(X) as well.

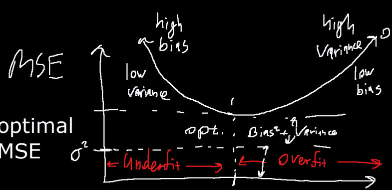
$$MSE = E_{X_1, \ldots, X_n, X_*}[MSE(x_*)]$$ previous result

$$= E_X\left[\sigma^2 + Bias[g]^2 + Var[g]\right]$$

$$= \sigma^2 + E_X\left[Bias[g]^2\right] + E_X[Var[g]]$$

This is the general bias-variance decomposition formula.
This is not computable in practice since you only have one D and one g. It is a theoretical formula.

Is there a "bias-variance tradeoff"? Yes and no...
It is not a zero sum game. If you make smart modeling decisions both bias AND variance decrease.



For one A, varying complexity of H

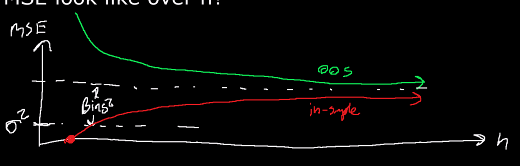
optimal MSE

complexity

optimal complexity

Overfit models have low bias since they are flexible and can fit f very snugly. Overfit models have high variance since they fit the y's snugly which means they fit the δ's snugly and the δ's change dataset-dataset. Underfit models have high bias since they cannot express f's complexity. Underfit models have low variance since they have few parameters whose estimates are stable dataset-dataset.

For any A, pick a large H (to intentionally overfit). What does MSE look like over n?



As n-->infinity, Var[g] --> 0. We are only limited by Bias which we can reduce by increasing complexity in curlyH (sigsq we can't do anything about).

Let's consider a metaalgorithm called "model averaging". You fit g_1, g_2, ..., g_M and then you ship their average,

$$g_{avg} := \frac{g_1 + g_2 + \cdots + g_M}{M}$$

What is the MSE of g_avg?

$$MSE = \sigma^2 + E_X\left[Bias[g_{avg}]^2\right] + E_X\left[Var[g_{avg}]\right]$$

$$= \sigma^2 + E_X\left[\Big(\underbrace{f}_{= \frac{f + f + \cdots + f}{M}} - \frac{g_1 + \cdots + g_M}{M}\Big)^2\right] + E_X\left[Var\left[\frac{g_1 + \cdots + g_M}{M}\right]\right]$$

$$= \sigma^2 + E_X\left[\left(\frac{1}{M}\Big(\overbrace{(f - g_1)}^{Bias[g_1]} + \cdots + \overbrace{(f - g_M)}^{Bias[g_M]}\Big)\right)^2\right] + \frac{1}{M^2} E_X\left[Var[g_1 + \cdots + g_M]\right]$$

Assume (a) Bias of each of the g_m's are the same and assume (b) g_1, ..., g_M are independent models and (c) The variance of each are the same too.

$$= \sigma^2 + E_X\left[Bias[g_1]^2\right] + \frac{1}{M} E_X\left[Var[g_1]\right]$$

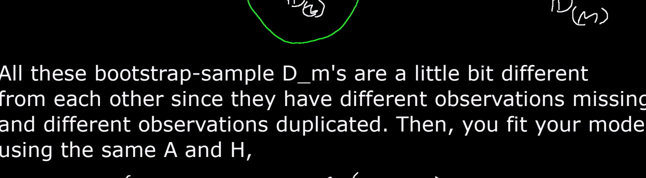
$$\overset{\text{let } M \to \infty}{=} \sigma^2 + E_X\left[Bias[g_1]^2\right] \overset{\text{overfit a lot}}{=} \sigma^2,$$ the theoretical minimum

Is this practical? No because assumption (b) is never met since g_1, ..., g_M are all built using D and thus must have a dependence on one another. (a) can be made true in practice since we can just use algorithms that overfit. (c) doesn't matter if (b) is true since that term --> 0 as M --> infinity. The problem is (b)!!!! And it seems to be impossible to solve!!

Although it's impossible, can we decrease their dependence? Yes. Enter Leo Breiman in 1994. Here was his big breakthrough.

Consider taking samples of n rows of D with replacement (this is called a "nonparametric bootstrap sample". You can prove (in Math 241) that approximately 2/3 of the rows of D appear in each sample and 1/3 are missing. Do this M times.

$$\mathbb{D}_{(1)} = sample(\mathbb{D}), \quad \mathbb{D}_{(2)} = sample(\mathbb{D}), \ldots, \mathbb{D}_{(M)} = sample(\mathbb{D})$$



All these bootstrap-sample D_m's are a little bit different from each other since they have different observations missing and different observations duplicated. Then, you fit your models using the same A and H,

$$g_1 = \mathcal{A}(\mathbb{D}_{(1)}, \mathcal{H}), \quad g_2 = \mathcal{A}(\mathbb{D}_{(2)}, \mathcal{H}), \ldots, g_M = \mathcal{A}(\mathbb{D}_{(M)}, \mathcal{H})$$

these models will be less dependent on one another than if you didn't use boostrap samples of D. So when you average,

$$g_{avg} = \frac{1}{M} \sum_{b=1}^{M} g_b$$

You should be able to bring MSE down. This meta-procedure is called boostrap aggregation = "bagging". We will calculate the reduction from bagging in MSE next class.