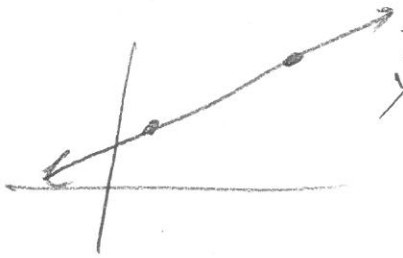
Lec 14 Math 329.4 3/21/18

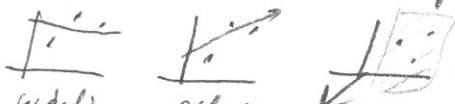
If $n = p+1 \Rightarrow X$ is square $\Rightarrow H = X(X^TX)^{-1}X^T = XX^{-1}(X^T)^{-1}X^T = I$

Imagine $n=2$, $p=1$



$\hat{\vec{y}} = \vec{y}$, $\vec{e} = \vec{0}$, $SSE = 0$, $R^2 = 1$

If $n=100$, $p=99$, there's 100 dimensional hyperplane that goes through all 100 pts.



Where is the basic problem with overfitting?

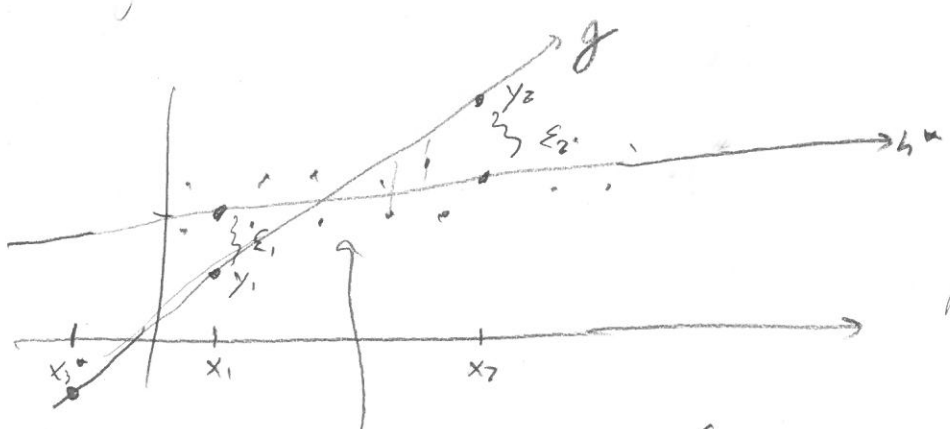
Remember $y = h^*(\vec{x}) + \varepsilon$, $h^* \in \mathcal{H}$ which is the best model in our candidate set given predictors at hand

misspecification error

→ ignorance error

y has two components
which cannot be distinguished

As an illustration, Imagine if you knew h^*...



$g - h =$ estimation error
this includes overfitting error!

How to think about this?
$n \to \infty$ $g \to h^*$
Recall $MSE = \frac{1}{n-(p+1)} SSE$

effective sample size

if n large and p large it's as if you have a small sample size!!

Why is this bad? For new points (when you predict, they will be far away from g.

The model doesn't generalize to new points.

Reason:
Overfitting is fitting ε which is determined because future prediction performance will suffer.

DEMOS