Bias Variance Trade-off

Agenda

- Loss function for regression
- Bias variance trade off
- Debugging variance and bias

Potential Problem – Outliers

How does outlier influence the regression model?

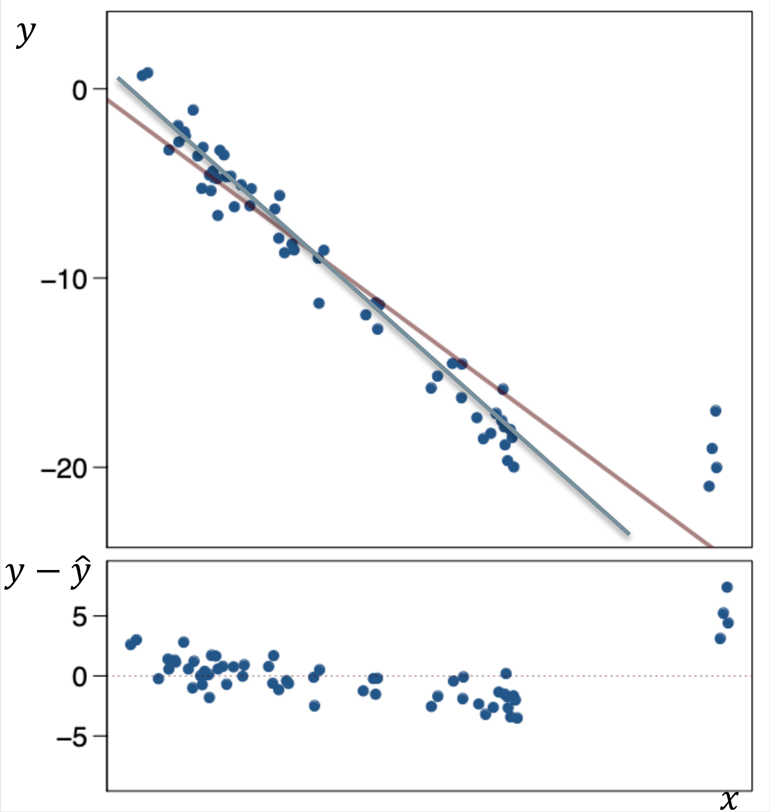
How do we identify outliers?

Residual plot, Data distribution

How do we deal with outliers?

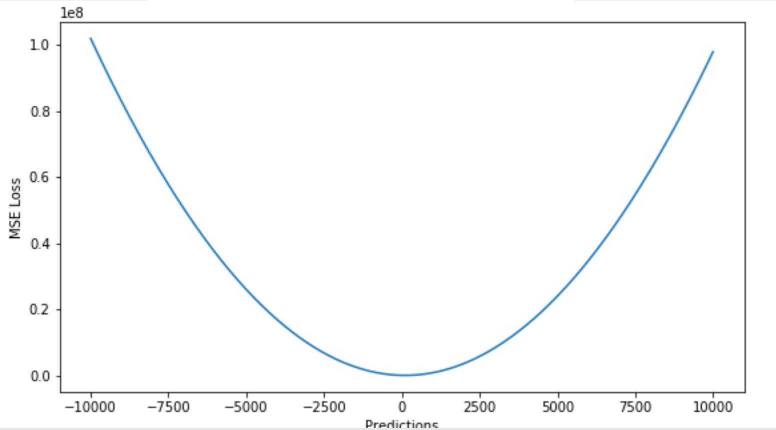
Delete the outlier

More robust models

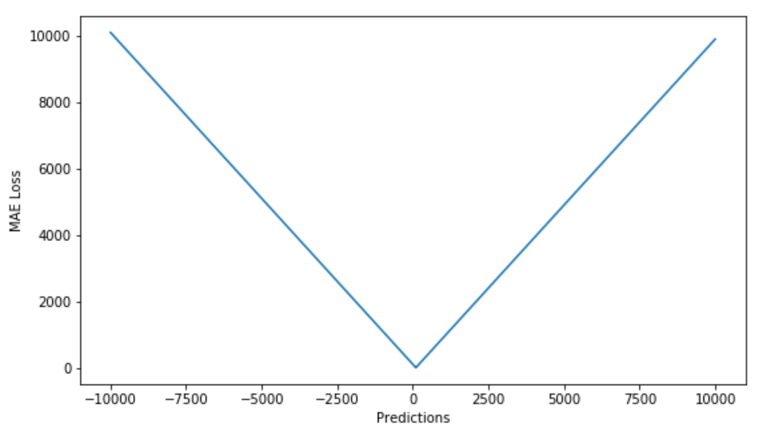


Loss function for regression

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_i^p)^2}{n}$$



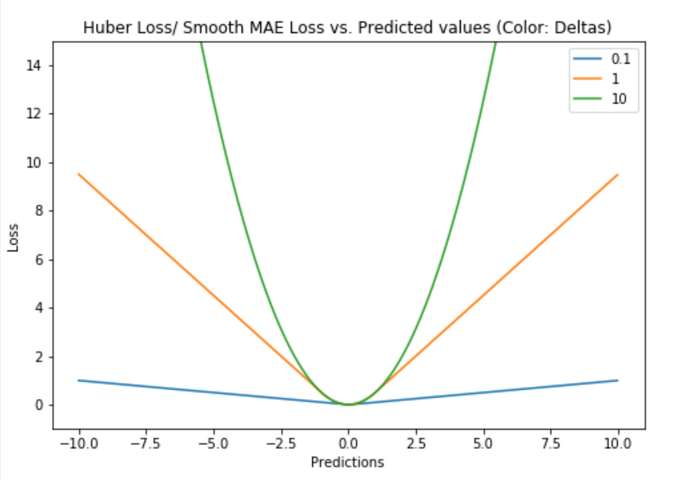
$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i^p|}{n}$$



Loss function for regression

Huber Loss (Smooth Mean Absolute Error)

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \le \delta, \\ \delta\, |y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$



Bias variance trade-off

Linear Regression

Model: $Y = w_0 + w_1 X_1 + \cdots + w_p X_p + \varepsilon$

Given the training data: $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\quad X = \begin{pmatrix} — & x_1 & — \\ & \vdots & \\ — & x_2 & — \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$

The model can be written as $y = Xw + \varepsilon$

Least square estimation for the parameters:

$$\hat{w} = (X^T X)^{-1} X^T y$$

$\hat{w}$ is a random variable! Even though ground truth w^* is not

Bias and variance in parameter estimation

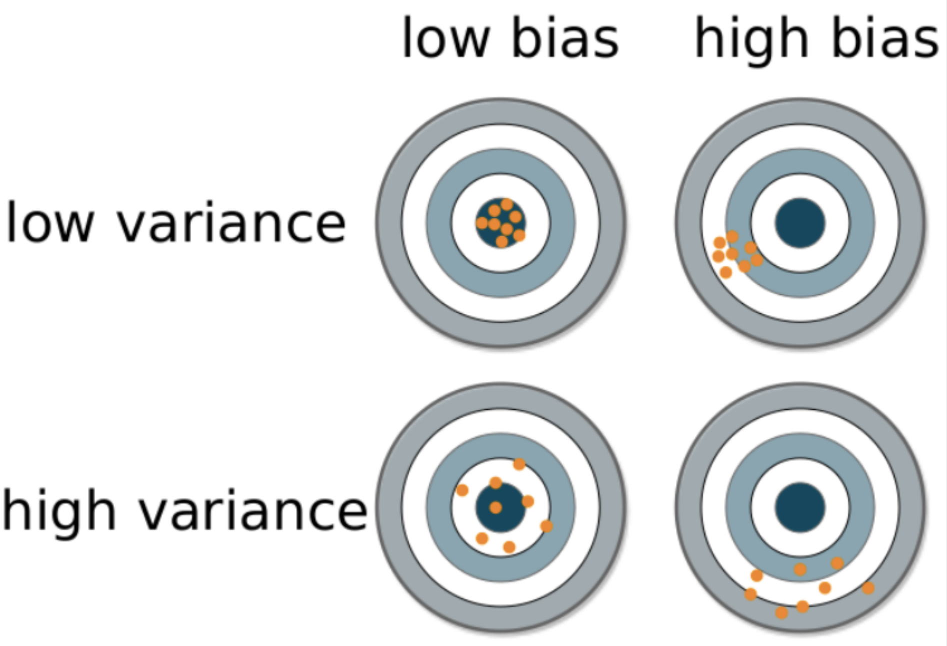
In general given a model with parameter θ, we get an estimator $\hat{\theta}$

- Bias of the estimator: $Bias(\hat{\theta}) = E[\hat{\theta} - \theta^*]$
- Variance of the estimator: $Var(\hat{\theta}) = Cov(\hat{\theta})$

If $Bias(\hat{\theta}) = 0$, then $\hat{\theta}$ is an unbiased estimation for θ

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}_n) &= \mathbb{E}\left[\|\hat{\theta}_n - \theta^*\|^2\right] \\
&= \mathbb{E}\left[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta^*\|^2\right] \\
&= \mathbb{E}\left[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2 + \underbrace{\|\mathbb{E}[\hat{\theta}_n] - \theta^*\|^2}_{\text{Constant}} + 2\underbrace{(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])}_{\text{Zero Mean}}^T(\mathbb{E}[\hat{\theta}_n] - \theta^*)\right] \\
&= \mathbb{E}\left[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2\right] + \|\mathbb{E}[\hat{\theta}_n] - \theta^*\|^2 \\
&= \mathbb{E}\left[\mathrm{tr}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^T\right]\right] + \|\mathbb{E}[\hat{\theta}_n] - \theta^*\|^2 \\
&= \mathrm{tr}\left[\mathrm{Var}(\hat{\theta}_n)\right] + \|\mathrm{Bias}(\hat{\theta}_n)\|^2.
\end{aligned}
$$

Bias-Variance Decomposition



Bias Variance in Prediction

We have a dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ drawn from a distribution over random variables X and Y, which following the following relation

$$Y = f(X) + \varepsilon$$

$$E(\varepsilon) = 0$$

$$X \perp \varepsilon$$

Suppose $\hat{f}(x)$ based on the training data D, consider a test point (x_*, y_*), the squared loss is

$$E_{D,\varepsilon_*}\left[\left(y_* - \hat{f}(x_*)\right)^2\right]$$

$$= E_{D,\varepsilon_*}\left[\left(f(x_*) + \varepsilon_* - \hat{f}(x_*)\right)^2\right]$$

$$= \underbrace{Var(\varepsilon_*)}_{\text{Irreducible error}} + \underbrace{\left\{E\left[\hat{f}(x_*) - f(x_*)\right]\right\}^2}_{\text{Bias of } \hat{f}} + \underbrace{Var\left[\hat{f}(x_*)\right]}_{\text{Variance of of } \hat{f}}$$

Bias Variance in Prediction

$$E_{D,\varepsilon_*}\left[\left(y_* - \hat{f}(x_*)\right)^2\right] = Var(\varepsilon_*) + \left\{E\left[\hat{f}(x_*) - f(x_*)\right]\right\}^2 + Var\left[\hat{f}(x_*)\right]$$

$$E_{D,\varepsilon_*}\left[y_* - \hat{f}(x^*)\right]^2$$

$$= E_{D,\varepsilon^*}\left[f(x^*) + \varepsilon^* - \hat{f}(x^*)\right]^2$$

$$= \underbrace{E_{D\varepsilon^*}\left[f(x^*) - \hat{f}(x^*)\right]^2}_{\text{unrelated with } \varepsilon^*} + E\left[\varepsilon^{*2}\right] + \underbrace{E\left[2\varepsilon^*\left(f(x^*) - \hat{f}(x^*)\right)\right]}_{\downarrow\, 0}$$

$$= E_D\left[f(x^*) - \hat{f}(x^*)\right]^2 + Var(\varepsilon^*)$$

$$= \left[E\left(f(x^*) - \hat{f}(x^*)\right)\right]^2 + Var\left[\underbrace{f(x^*)}_{\text{constant}} - \hat{f}(x^*)\right] + Var(\varepsilon^*)$$

$$= \left[E\left(f(x^*) - \hat{f}(x^*)\right)\right]^2 + Var\left[\hat{f}(x^*)\right] + Var(\varepsilon).$$

Bias Variance Trade off

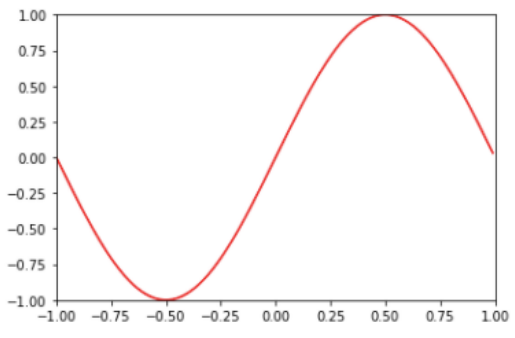
$$E_{D,\varepsilon_*}\left[\left(y_* - \hat{f}(x_*)\right)^2\right] = Var(\varepsilon_*) + \{E[\hat{f}(x_*) - f(x_*)]\}^2 + Var[\hat{f}(x_*)]$$

$$E_{X\in D_{test}}(squared\ loss) = E_X\left\{\underbrace{\left[E_D\left(f(X) - \hat{f}(X)\right)\right]^2}_{\text{Bias of } \hat{f}} + \underbrace{Var_D\left[\hat{f}(X)\right]}_{\text{Variance of of } \hat{f}} + \underbrace{Var_\varepsilon(\varepsilon)}_{\text{Irreducible error}}\right\}$$

Illustrated Example

Example: Approximate a sine function

True function $f(x) = \sin(\pi x)$, $f : [-1, 1] \to R$



You are given two hypotheses of the function to fit:

$$H_0 : f(x) = c$$

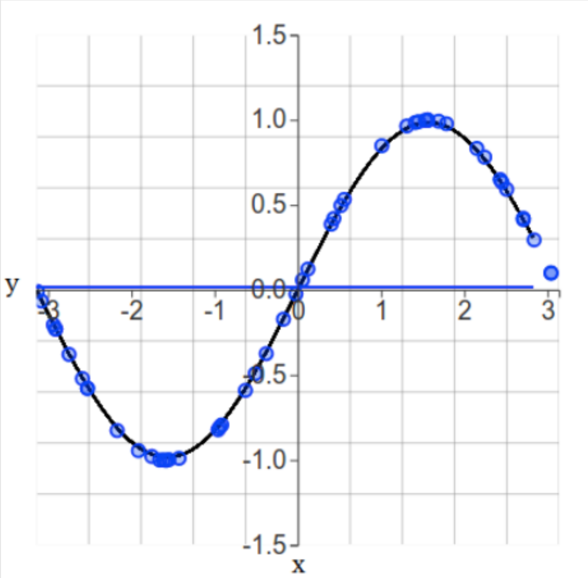
$$H_1 : f(x) = w_0 + w_1 x$$

Which leads to better results?

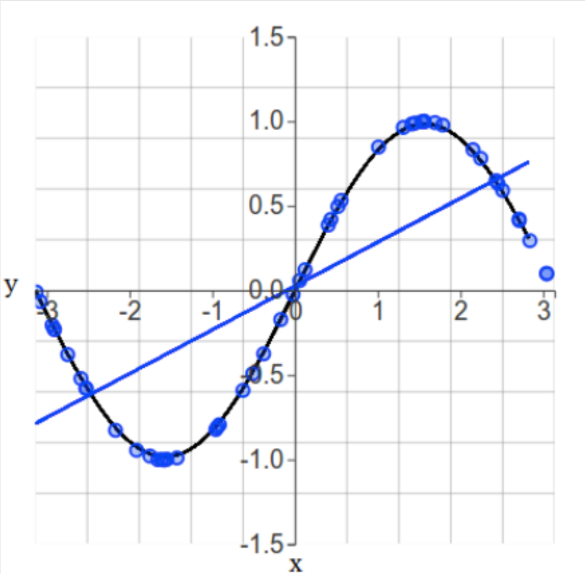
Learning H_0 vs H_0

Data simulation with M=50

$H_0: f(x) = c$

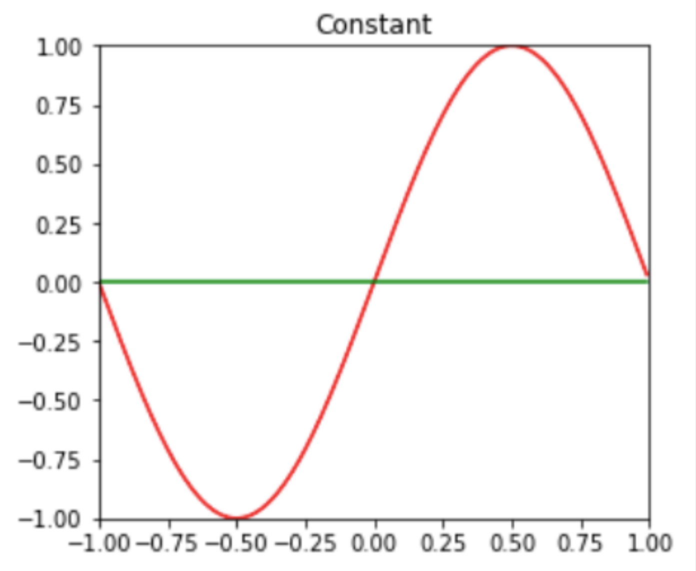


$H_1: f(x) = w_0 + w_1 x$

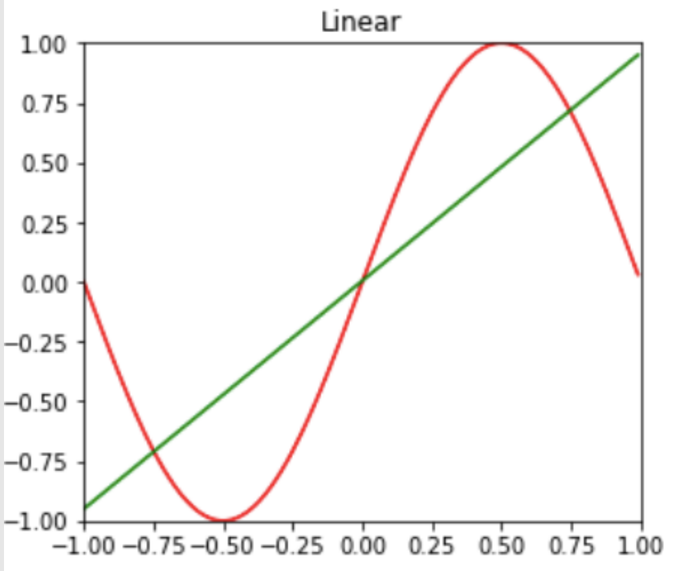


Learning H_0 vs H_0

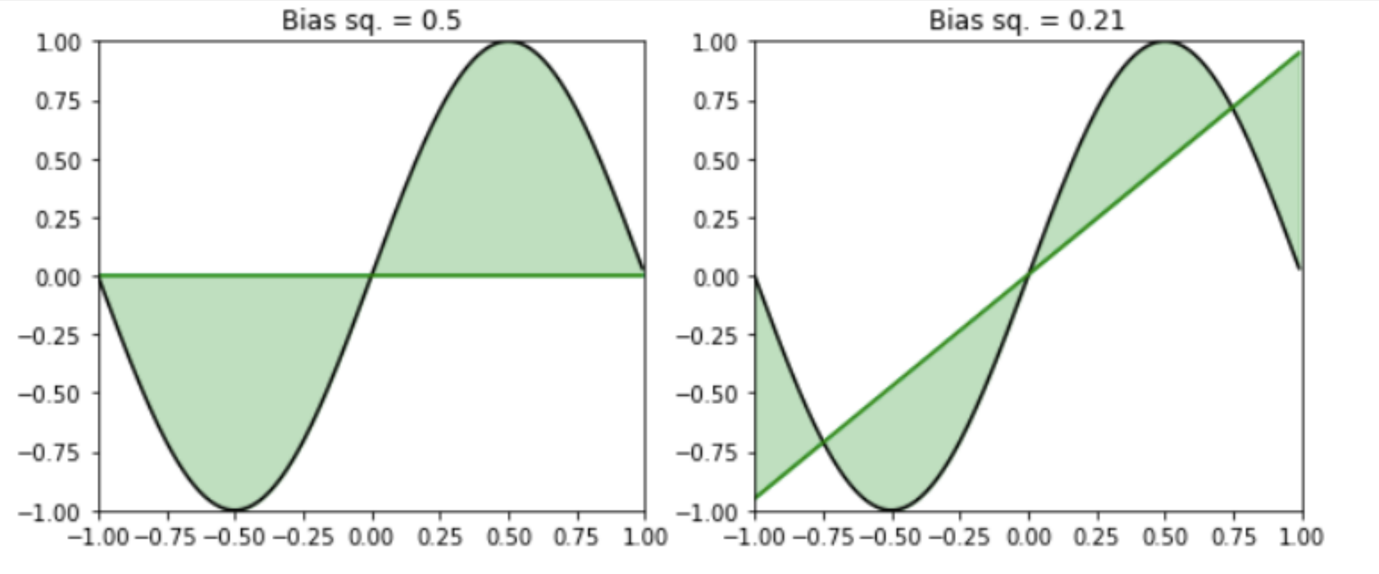
$H_0: f(x) = c$



$H_1: f(x) = w_0 + w_1 x$



Bias

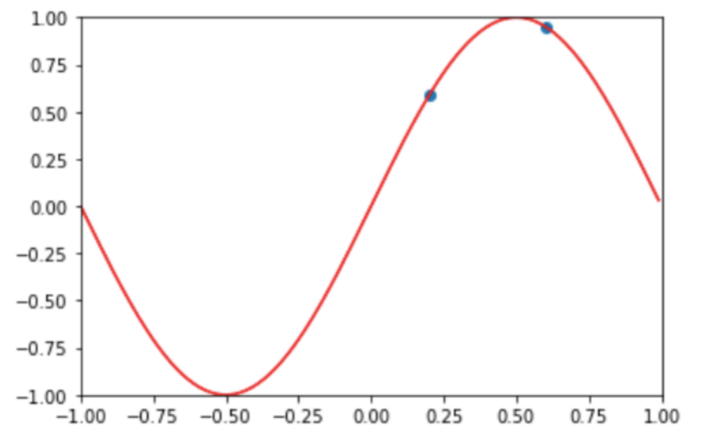


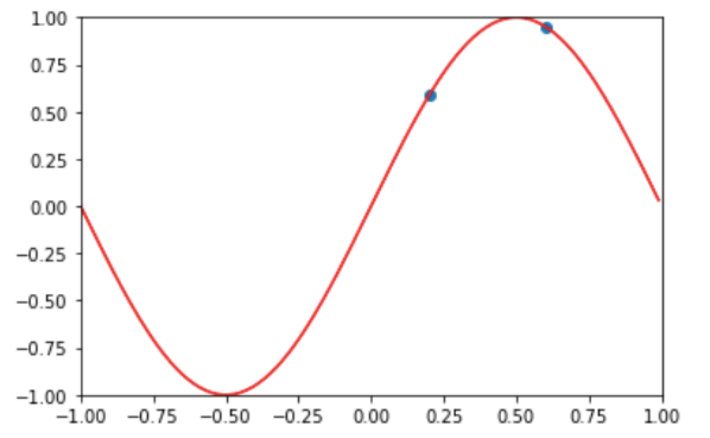
$$E_X(squared\ loss) = E_X\left\{\underbrace{\boxed{\left[E_D\left(f(X)-\hat{f}(X)\right)\right]}^2}_{\text{Bias of } \hat{f}} + \underbrace{Var_D\left[\hat{f}(X)\right]}_{\text{Variance of of } \hat{f}} + \underbrace{Var_{\varepsilon}(\varepsilon)}_{\text{Irreducible error}}\right\}$$

What if you are only given two points?

$H_0 : f(x) = c$

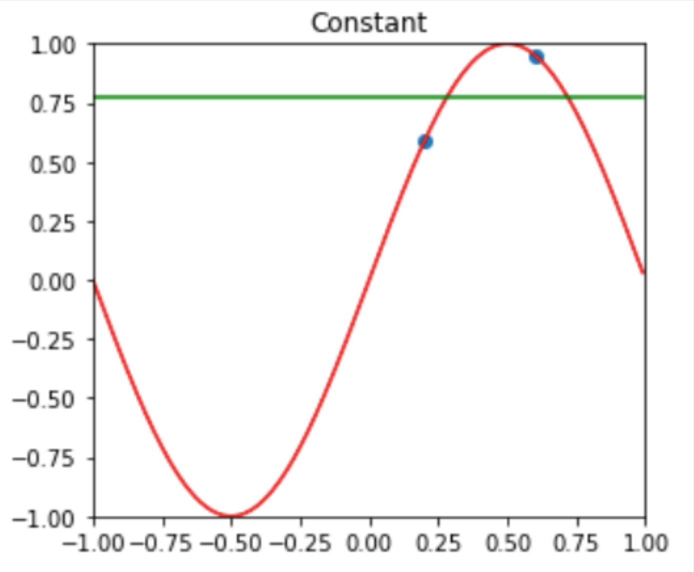
$H_1 : f(x) = w_0 + w_1 x$



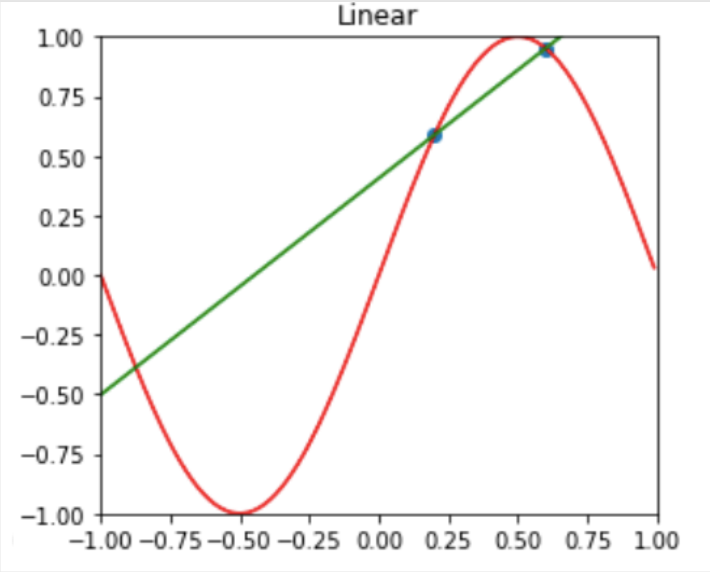


What if you are only given two points?

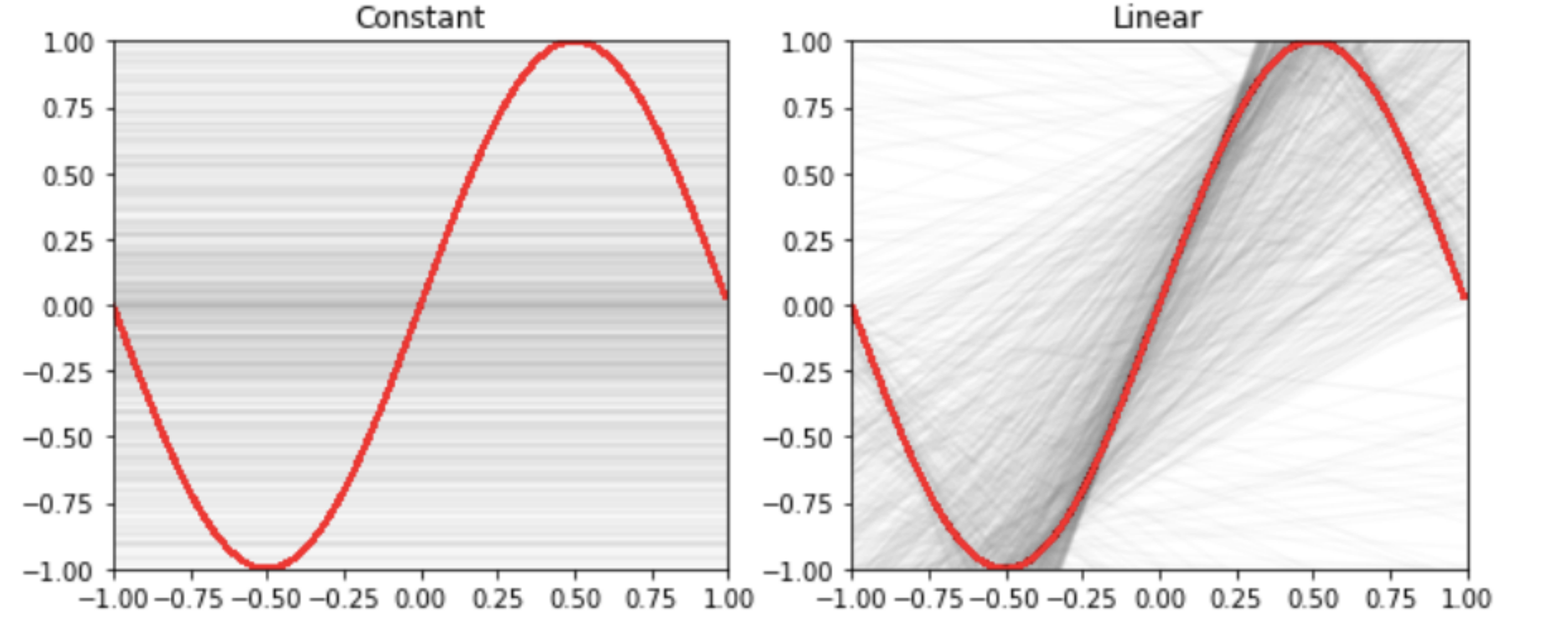
$H_0\colon f(x) = c$



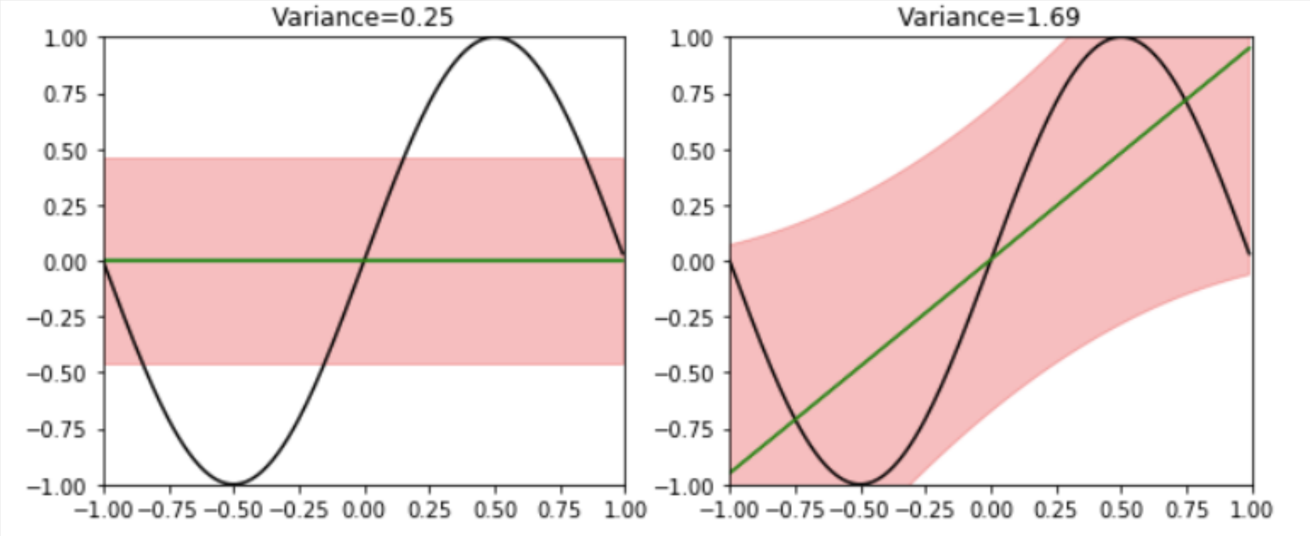
$H_1\colon f(x) = w_0 + w_1 x$



Let us repeat the experiment



Variance

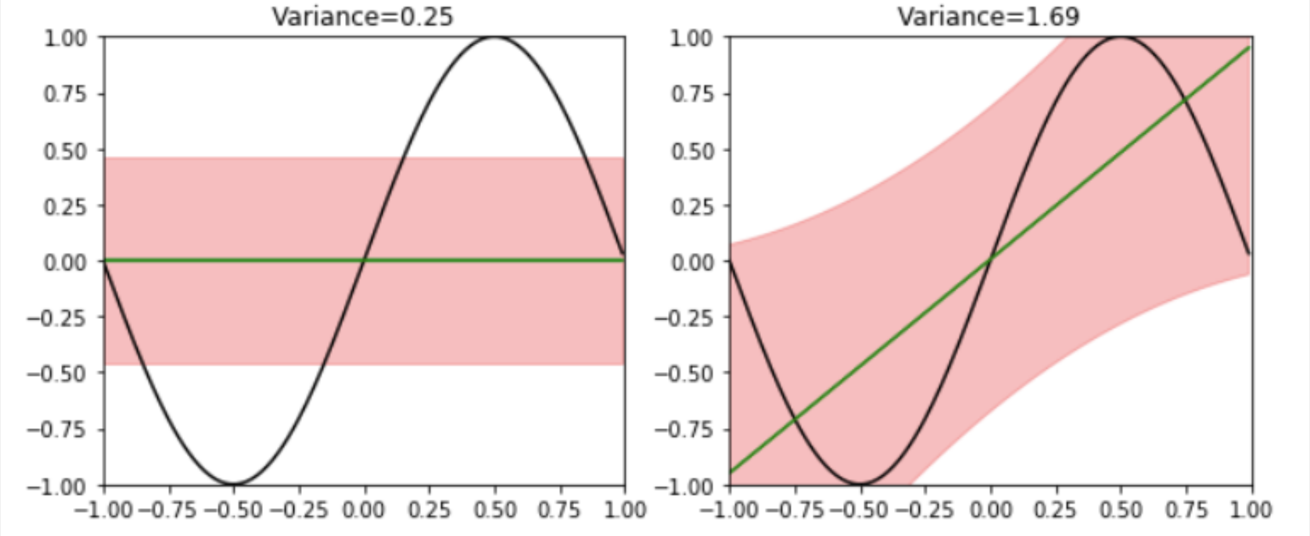


$$E_X(squared\ loss) = E_X\left\{\underbrace{\left[E_D\left(f(X) - \hat{f}(X)\right)\right]^2}_{\text{Bias of } \hat{f}} + \underbrace{\boxed{Var_D\left[\hat{f}(X)\right]}}_{\text{Variance of of } \hat{f}} + \underbrace{Var_\varepsilon(\varepsilon)}_{\text{Irreducible error}}\right\}$$

The winner is?

$H_0: f(x) = c$

$H_1: f(x) = w_0 + w_1 x$



Bias = 0.50
Variance = 0.25

Bias = 0.21
Variance = 1.69

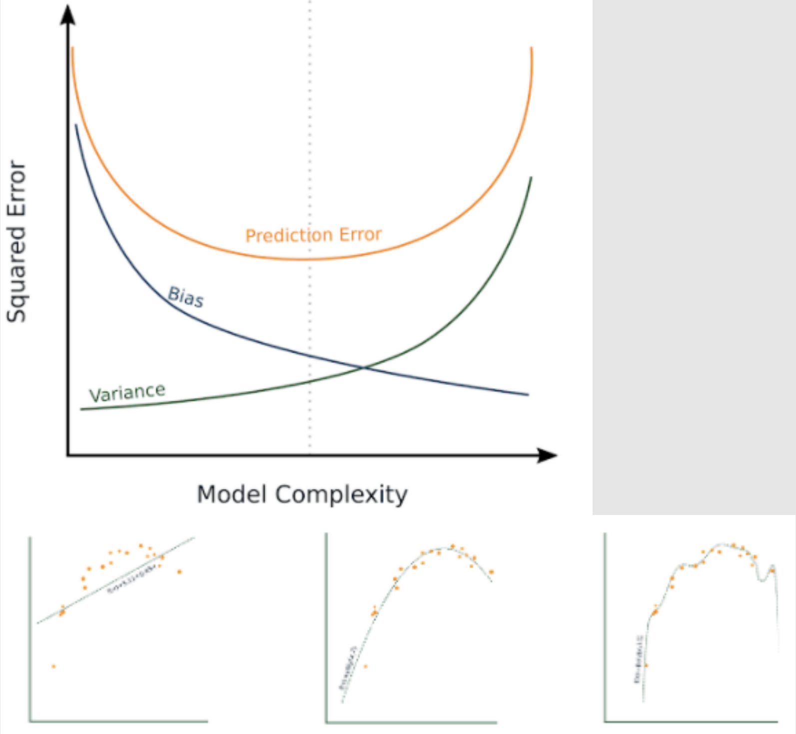
Lesson learned

Math the model complexity to

Data resources not the response complexity

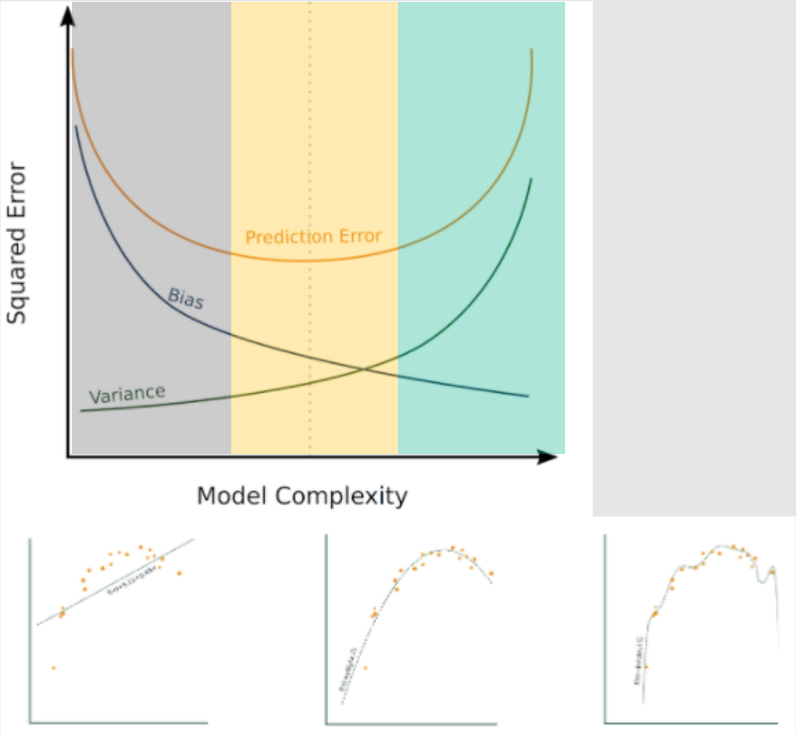
Debugging variance bias

Bias – Variance Trade-off



As we increase model complexity, bias decrease and the variance increase

Bias – Variance Trade-off



As we increase model complexity, bias decrease and the variance increase

Regime 1: high bias
Low but consistent performance
Trian MSE ≈ Test MSE

Regime 2: good trade-off
Acceptable MSE
Consistent MSE

Regime 2: high variance
MSE all over the place
Trian MSE << Test MSE

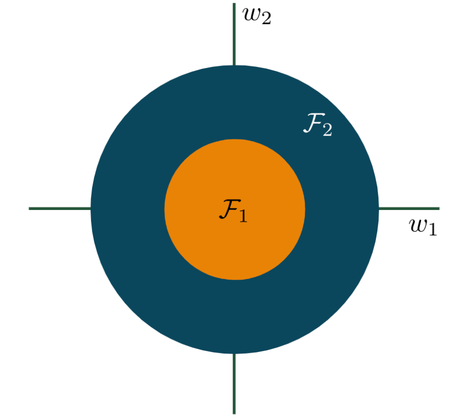
Model complexity

The representational capacity of a set of functions is an indicator of the representational richness within this set of functions.

- It's usually quite easy to compare different models of the same "type". For instance, consider the following two hypothesis sets of functions:

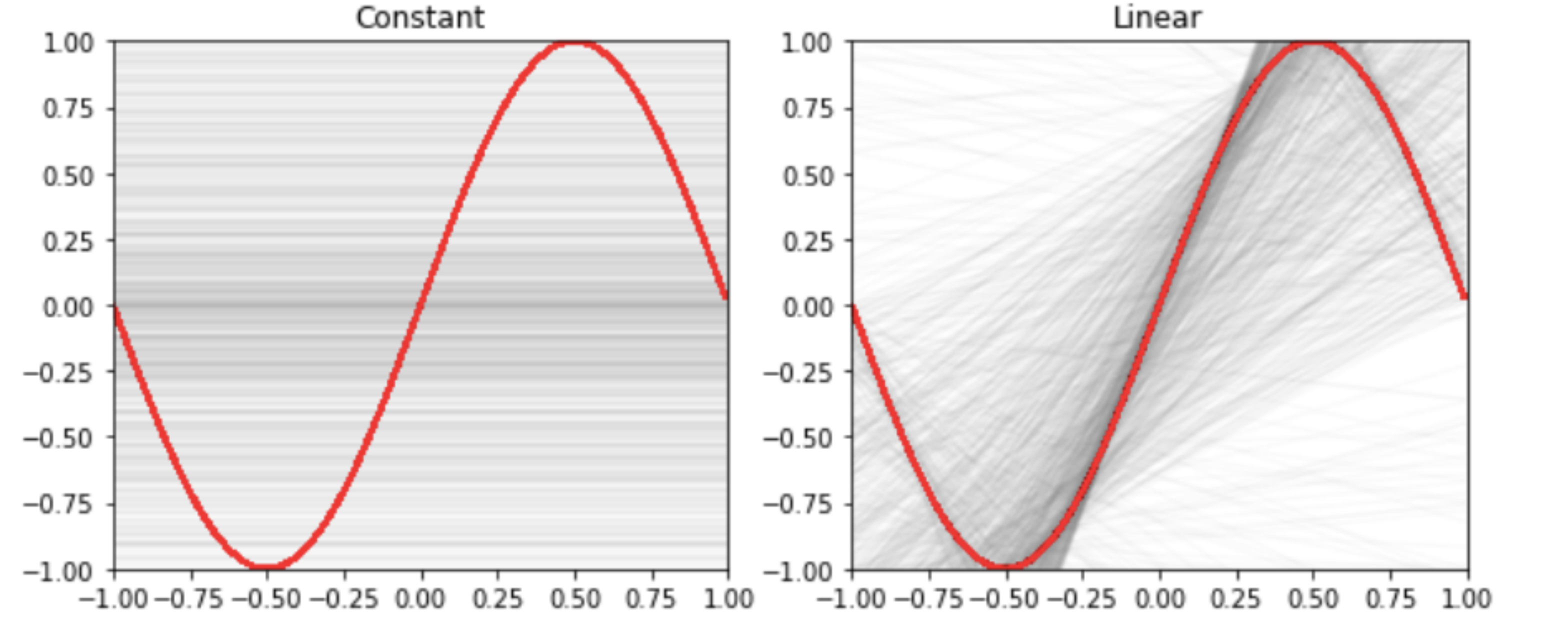
$$\mathcal{F}_1 = \{w \to w \cdot x \,|\, \|w\|_2 \leq W\}$$

$$\mathcal{F}_2 = \{w \to w \cdot x \,|\, \|w\|_2 \leq 2W\}$$

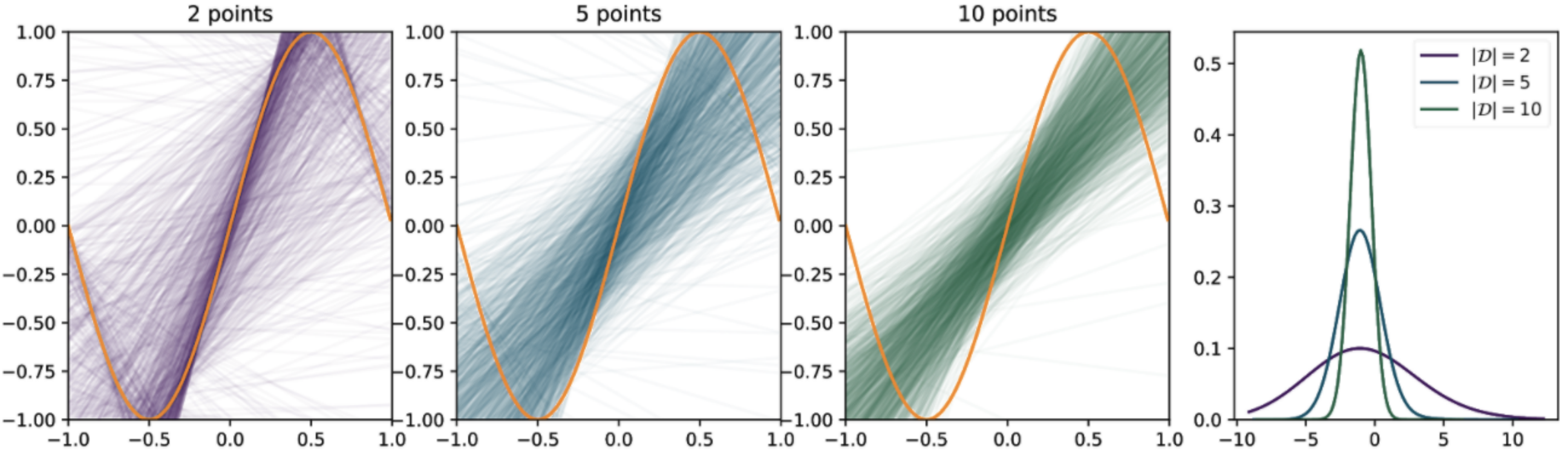


A function family with low capacity, is more likely to underfit.
A function family with high capacity, is more likely to overfit.

Experiment with two samples



Adding more data



As we increase the sample size, the variance decreases!

Theoretical results for linear regression

In linear regression

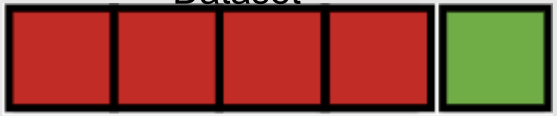
$y = x^{\mathrm{T}}\beta + \varepsilon$

$\hat{\beta} = (X^T X)^{-1} X^T$

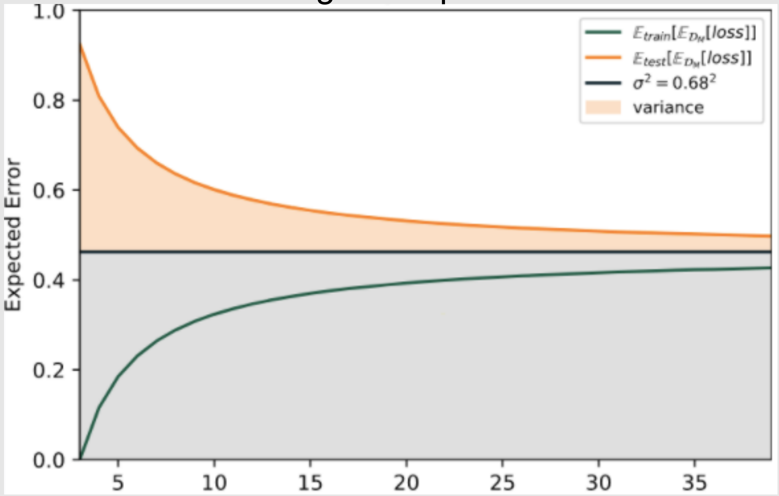
In-sample error: $\sigma^2 \left(1 - \frac{p+1}{N}\right)$

Out-of-sample error: $\sigma^2 \left(1 + \frac{p+1}{N}\right)$

Dataset



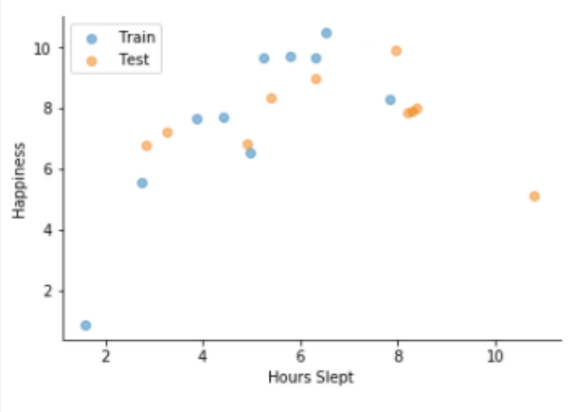
Learning curve p=1



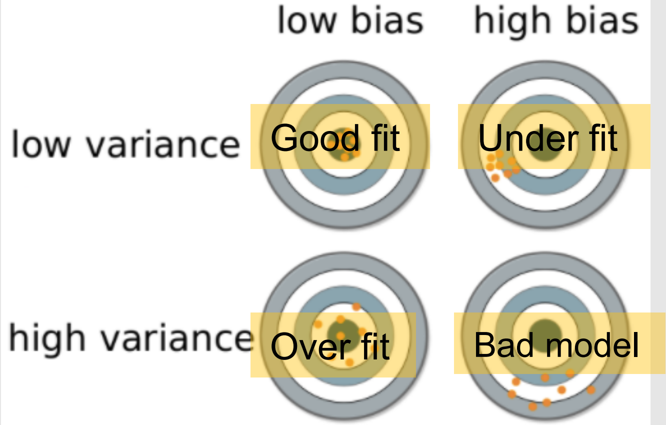
Sample size N

p=1 (with one predictor and the intercept)

Test on a hold-out set



	Underfit	Good fit	Overfit
Training MSE	Bad	**Good**	Perfect
Validation MSE	Bad	**Good**	Bad



Solution for high variance

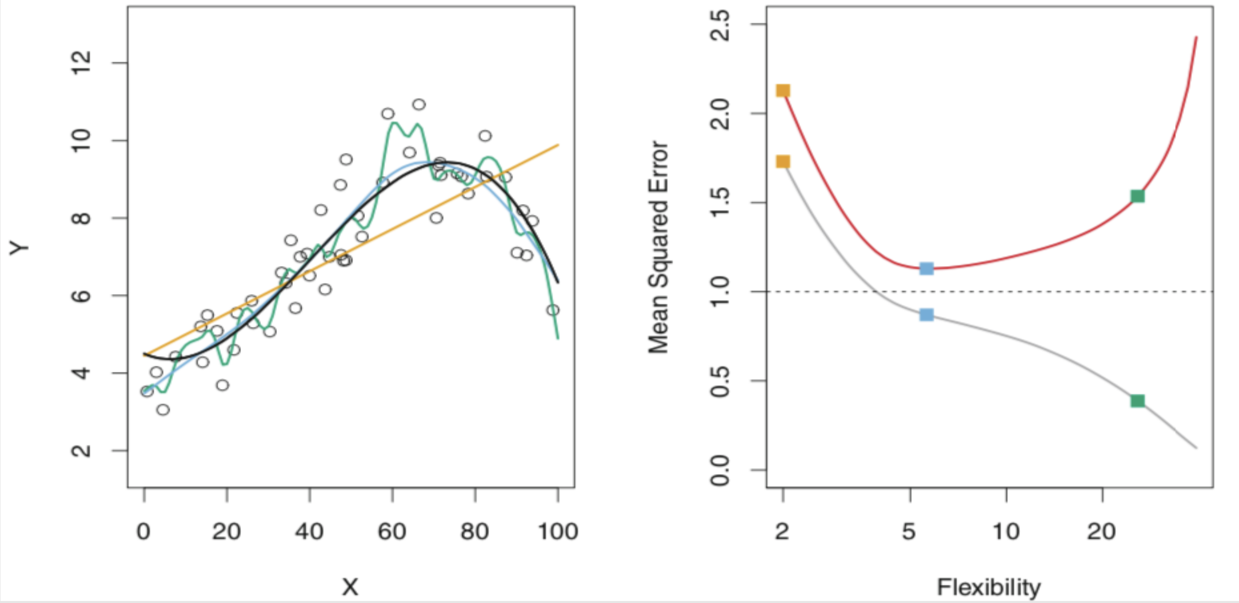
- Add more training data
- Reduce model complexity
- Bagging

Solution for high bias

- Use a more complex model
- Add extra features
- Boosting

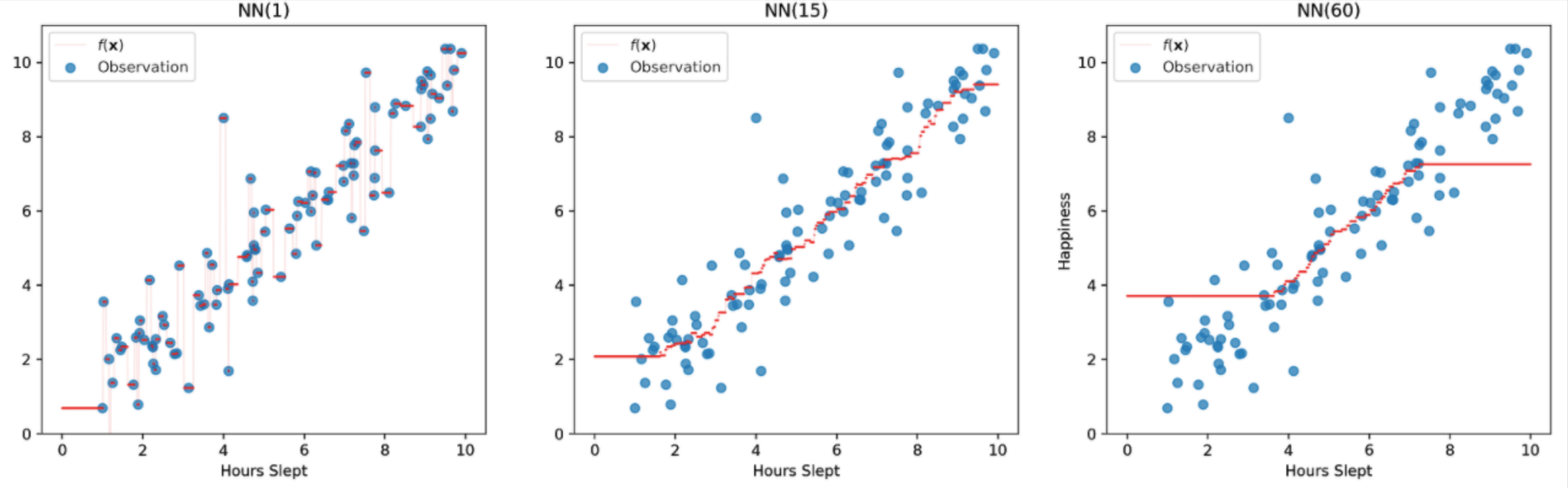
Functional View

Bias–Variance Trade-off



- Black curve is truth. Grey curve on right is training MSE, red curve is testing MSE.
- Orange, blue and green curves/squares correspond to fits different flexibility

Bias–Variance Trade-off



Which is more complex?
Which includes most bias/variance?