Bias-Variance Tradeoff

Overview of the second half the semester

1. A little bit Learning Theory

2. Make our linear models nonlinear (Kernel)

3. How to combine multiple classifiers into a stronger one (Bagging & Boosting)?

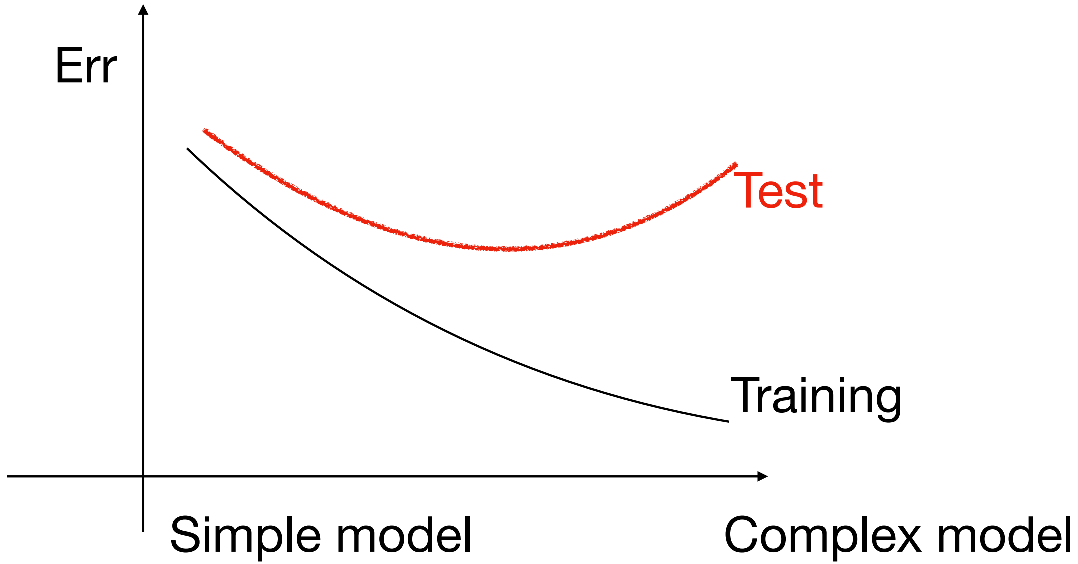
4. Intro of Neural Networks (old and new)

Objective

Understand Bias-Variance tradeoff — When and why your ML models work (or don't work)

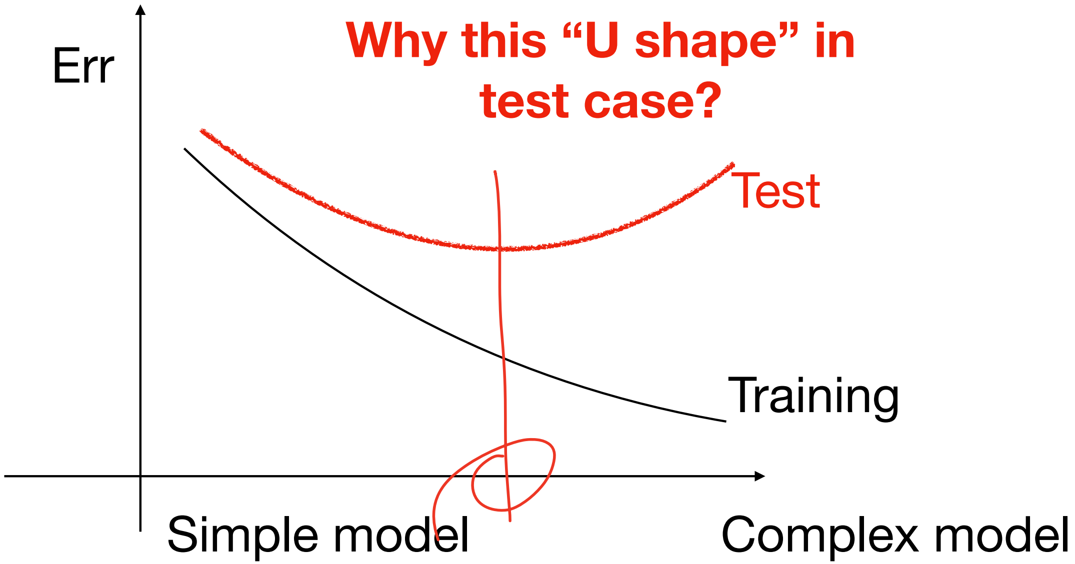
Objective

Understand Bias-Variance tradeoff — When and why your ML models work (or don't work)



Objective

Understand Bias-Variance tradeoff — When and why your ML models work (or don't work)



Outline of Today

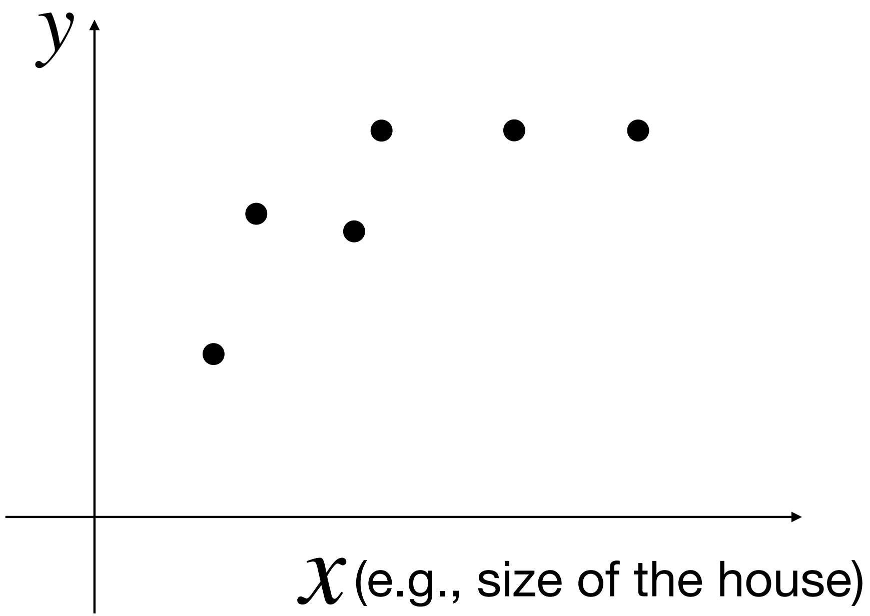
1. Intro on Underfitting/Overfitting and Bias/Variance

2. Derivation of the Bias-Variance Decomposition

Bayes optimal predictor

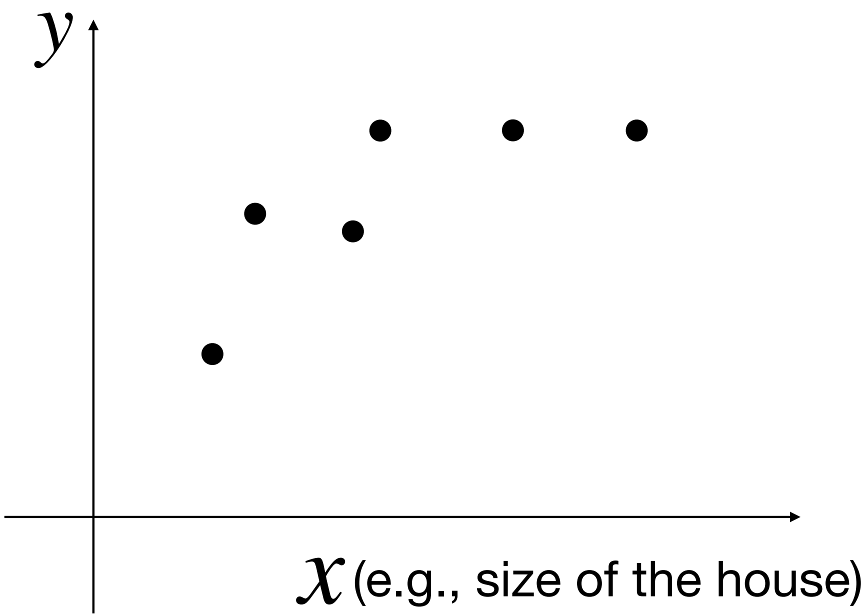
Consider regression problem w/ dataset $\mathscr{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$

$\ell(h) = (h(x) - y)^2$



Bayes optimal predictor

Consider regression problem w/ dataset $\mathscr{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$

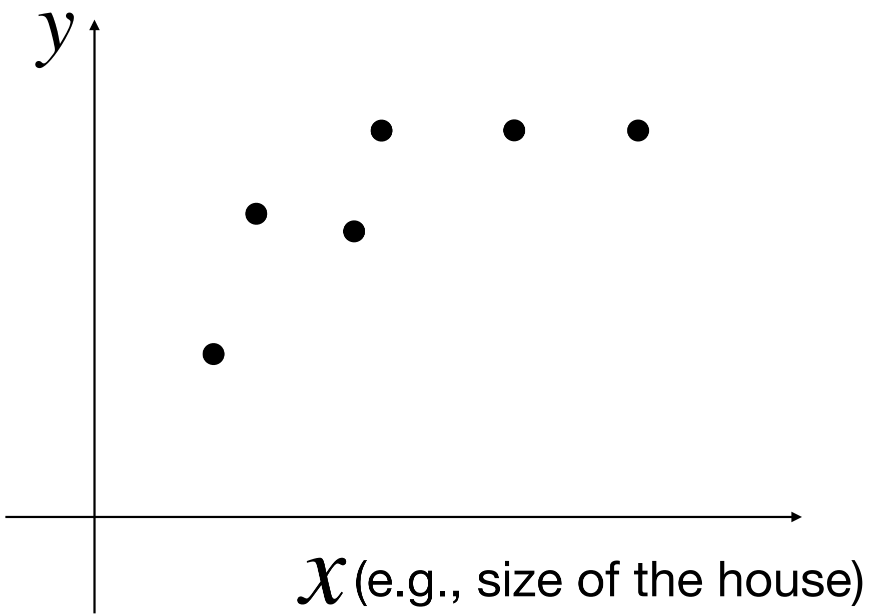


The Bayes optimal regressor:

$$\bar{y}(x) := \mathbb{E}[y \mid x]$$

Bayes optimal predictor

Consider regression problem w/ dataset $\mathscr{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$



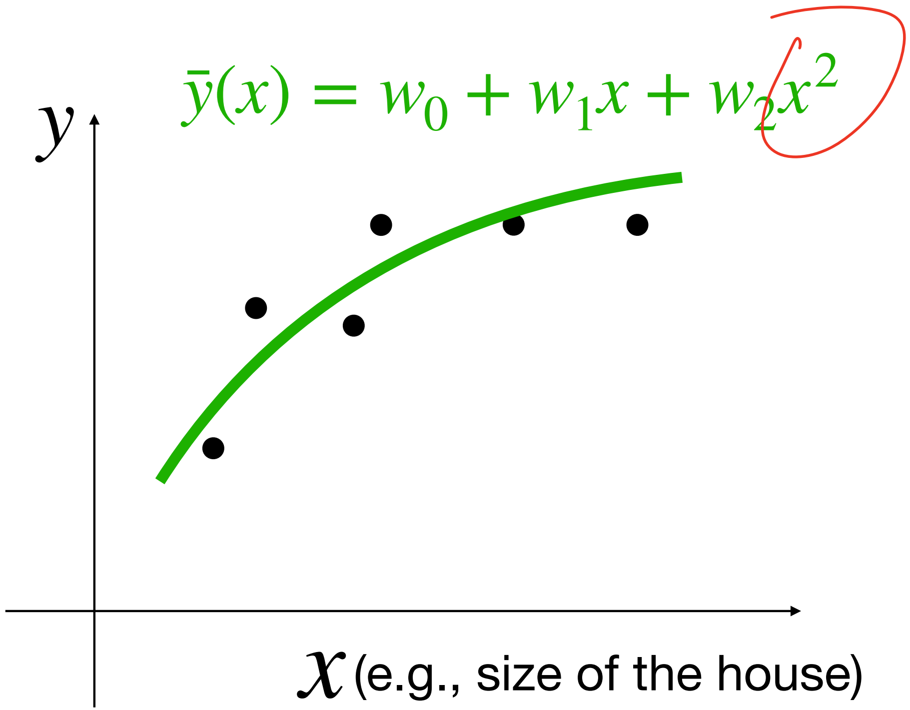
The Bayes optimal regressor:

$$\bar{y}(x) := \mathbb{E}[y \mid x]$$

The best we could do, cannot beat this one

Bayes optimal predictor

Consider regression problem w/ dataset $\mathscr{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$

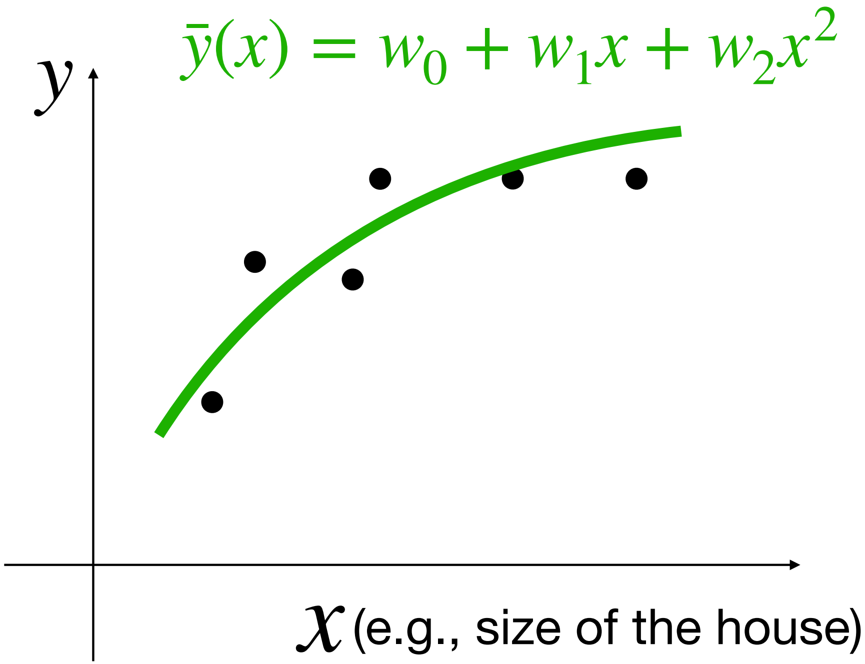


The Bayes optimal regressor:

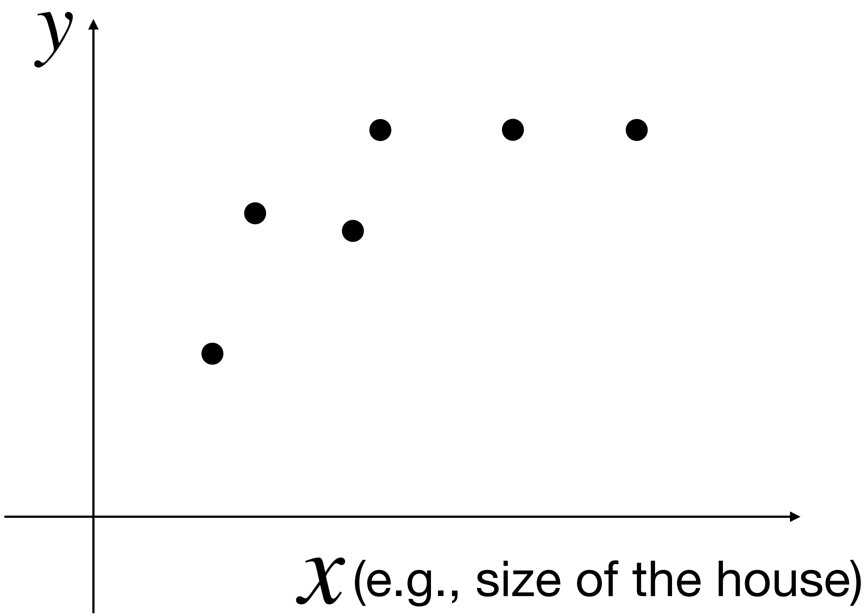
$$\bar{y}(x) := \mathbb{E}[y \mid x]$$

The best we could do, cannot beat this one

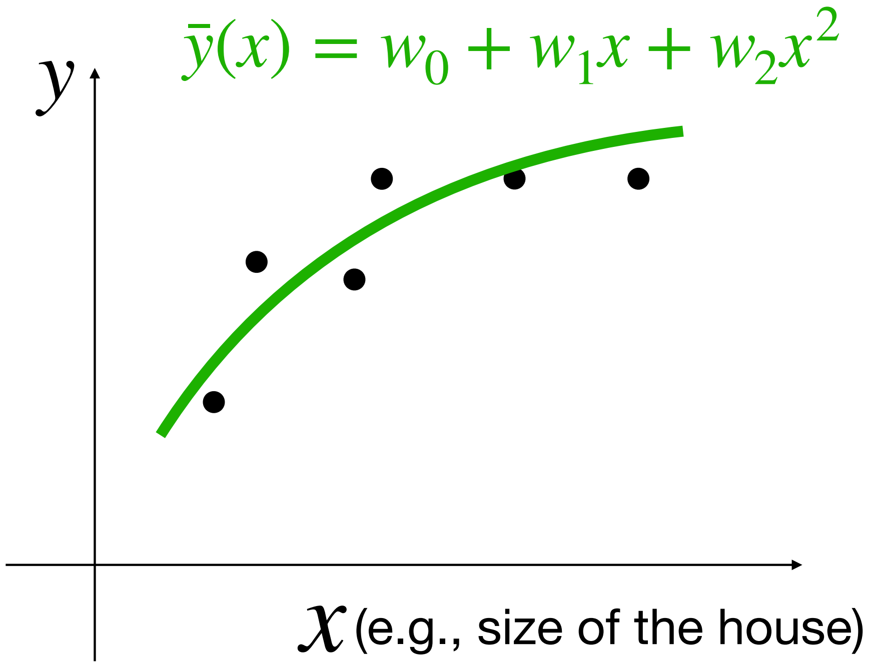
Underfitting



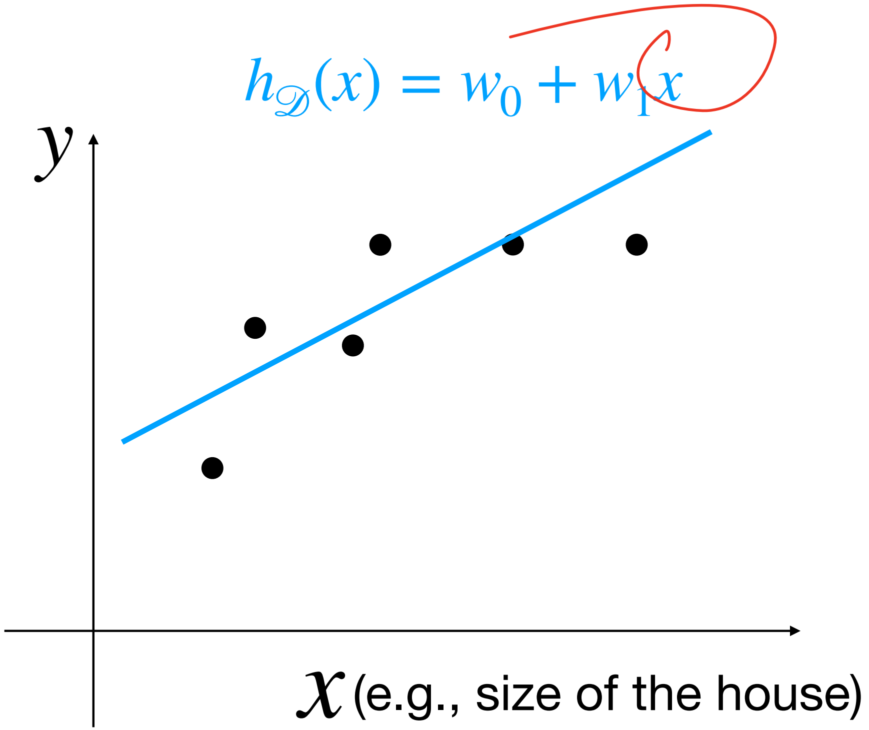
(Just right)



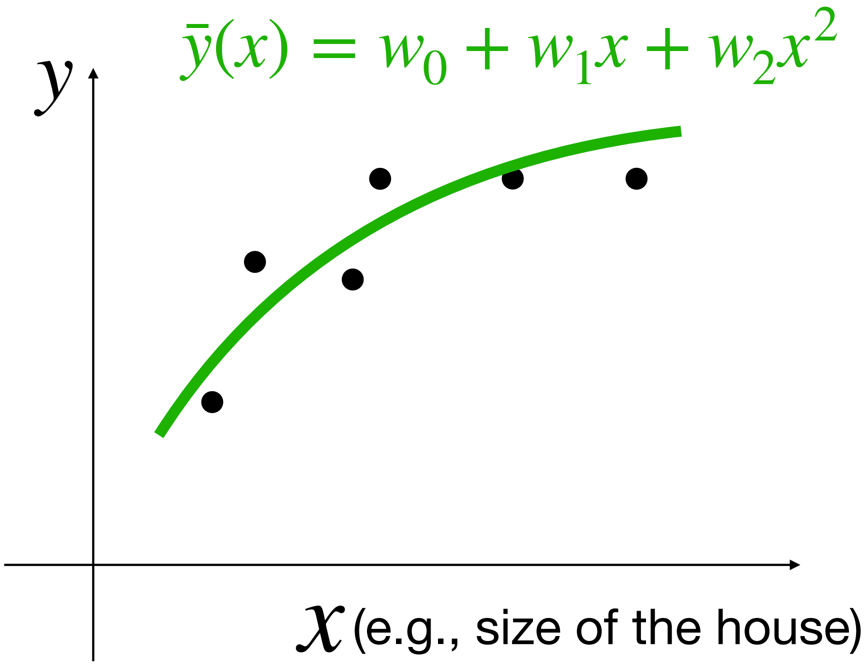
Underfitting



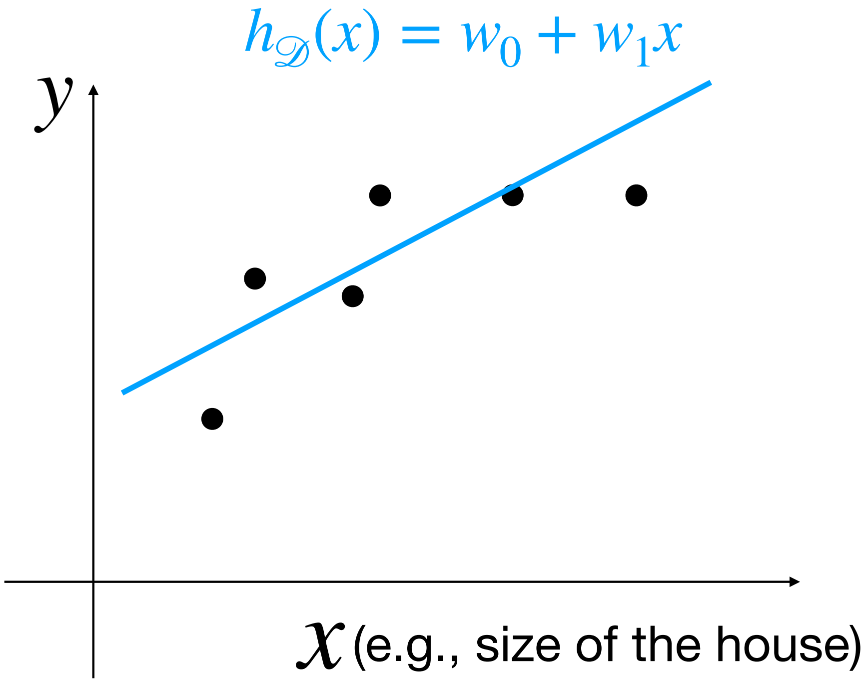
(Just right)



Underfitting



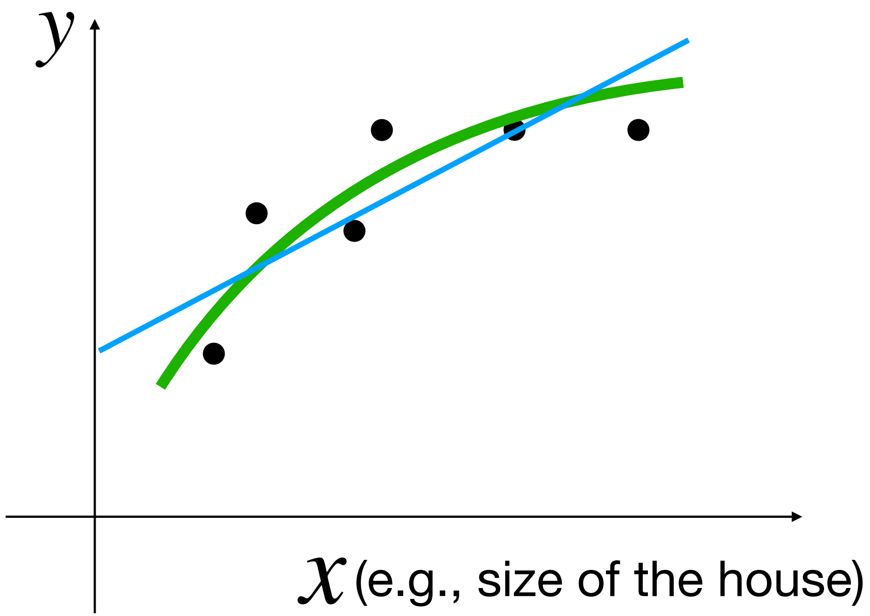
(Just right)



Underfitting

Underfitting

Just right versus Underfitting

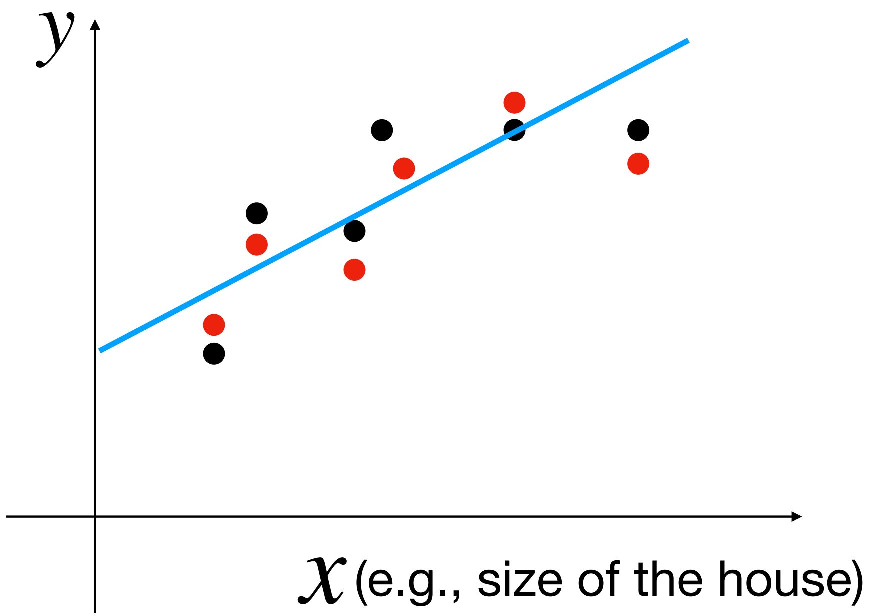


Bias:

Bias towards to linear models

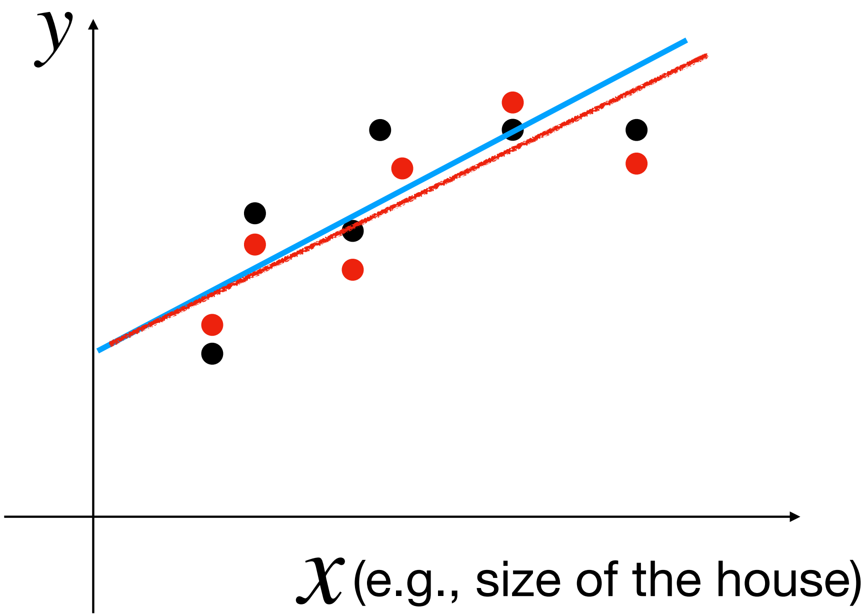
Underfitting

Now let's redo linear regression on a different dataset $\mathscr{D}'$ (but from the same distribution)



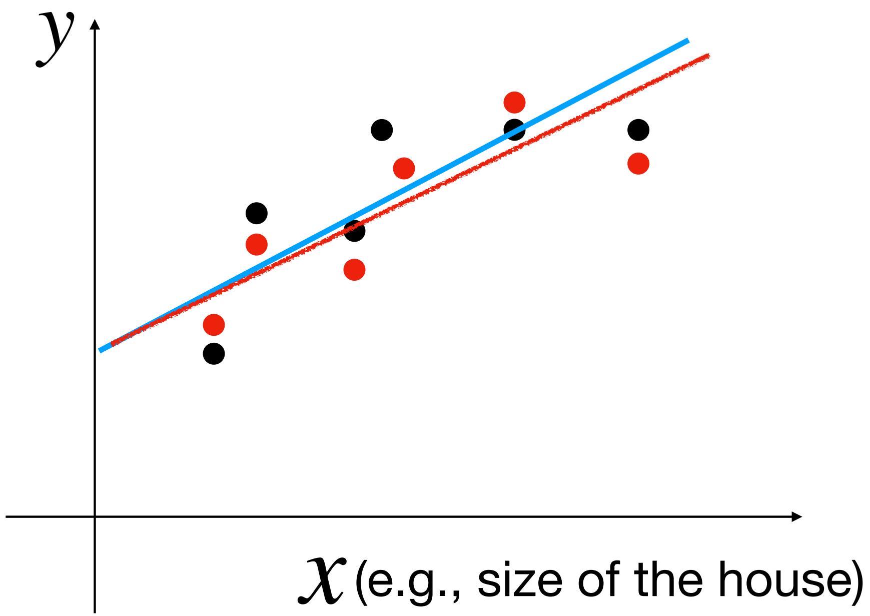
Underfitting

Now let's redo linear regression on a different dataset $\mathscr{D}'$ (but from the same distribution)



Underfitting

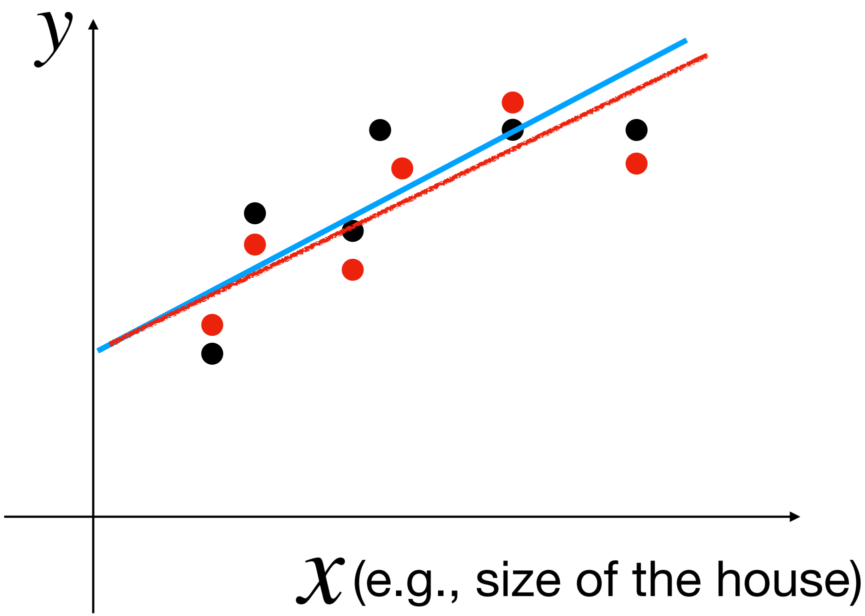
Now let's redo linear regression on a different dataset $\mathscr{D}'$ (but from the same distribution)



The new linear function does not differ too much from the old one

Underfitting

Now let's redo linear regression on a different dataset $\mathscr{D}'$ (but from the same distribution)

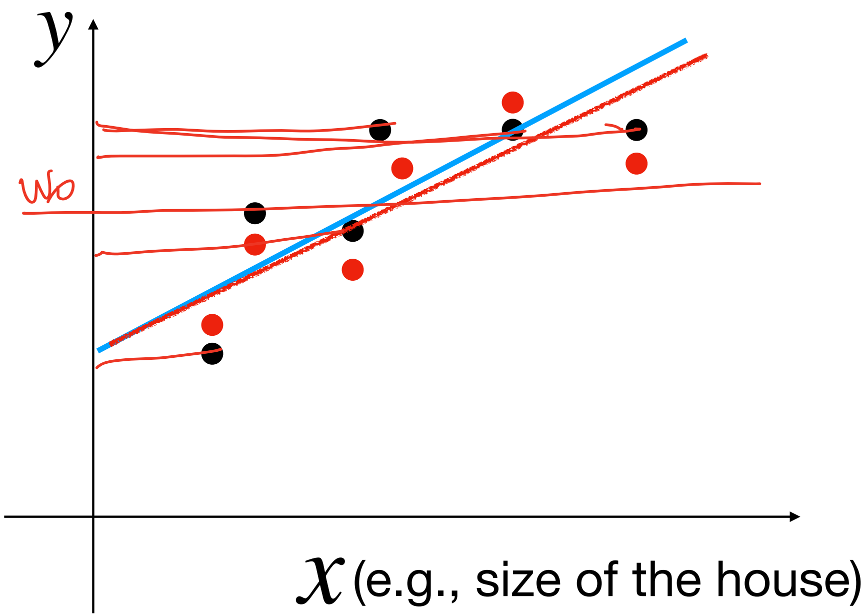


The new linear function does not differ too much from the old one

This is called low variance

Underfitting

Now let's redo linear regression on a different dataset $\mathcal{D}'$ (but from the same distribution)



The new linear function does not differ too much from the old one

This is called low variance

Q: what happens when our linear predictor is $h(x) = w_0$?

$\hat{w_0} = \frac{1}{n} \sum_{i=1}^{n} y_i = \mathbb{E}_y[y]$

Summary on underfitting

1. Often our model is too simple, i.e.., we bias towards too simple models

Summary on underfitting

1. Often our model is too simple, i.e.., we bias towards too simple models

2. This causes underfitting, i.e., we cannot capture the trend in the data

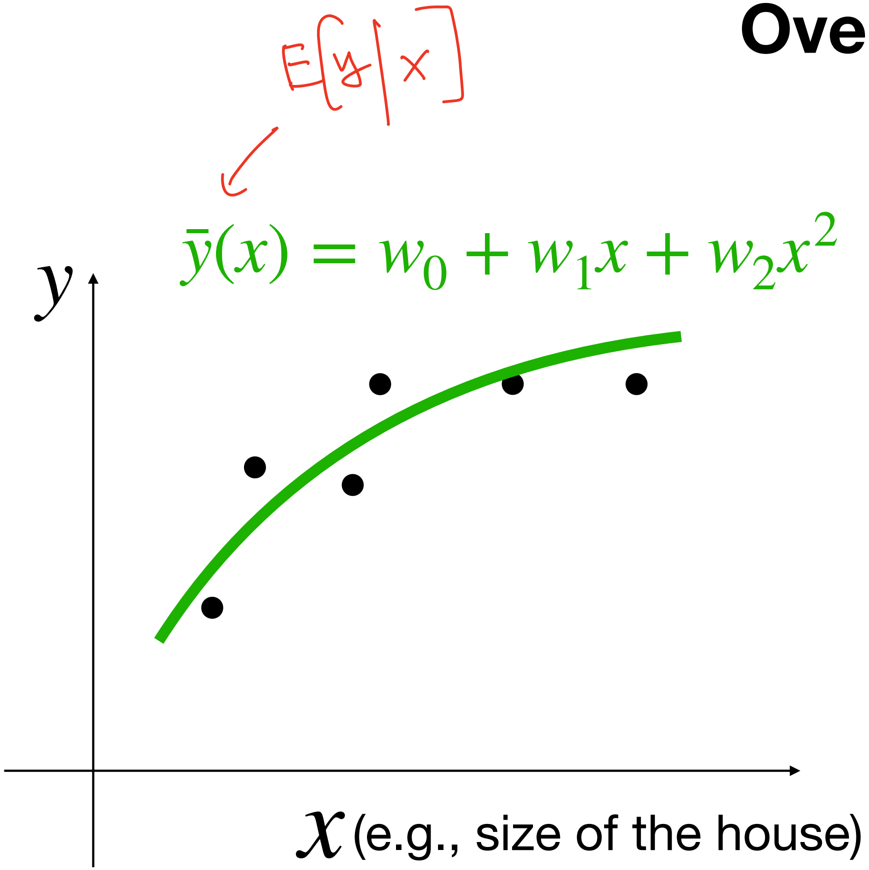
Summary on underfitting

1. Often our model is too simple, i.e.., we bias towards too simple models

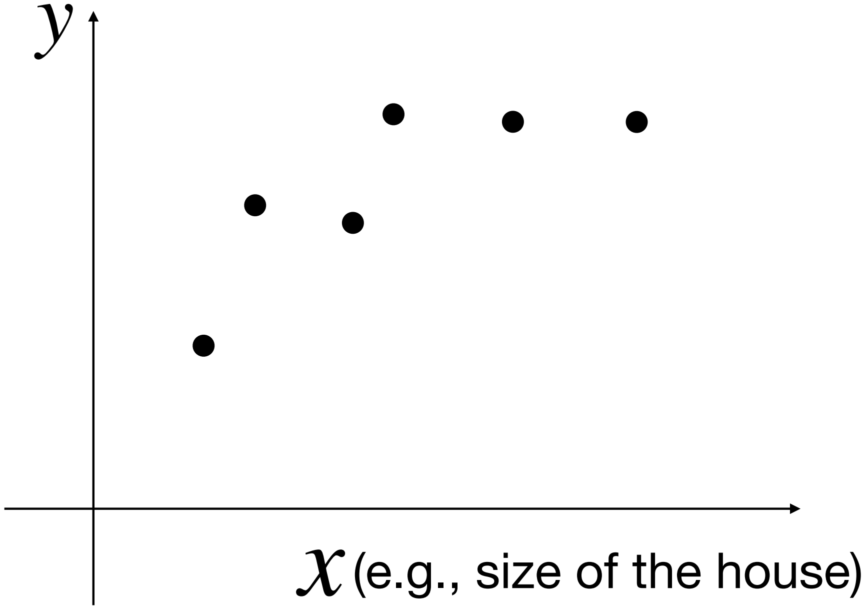
2. This causes underfitting, i.e., we cannot capture the trend in the data

3. In this case, we have large bias, but low variance (think about the $h(x) = w_0$ case)

Overfitting

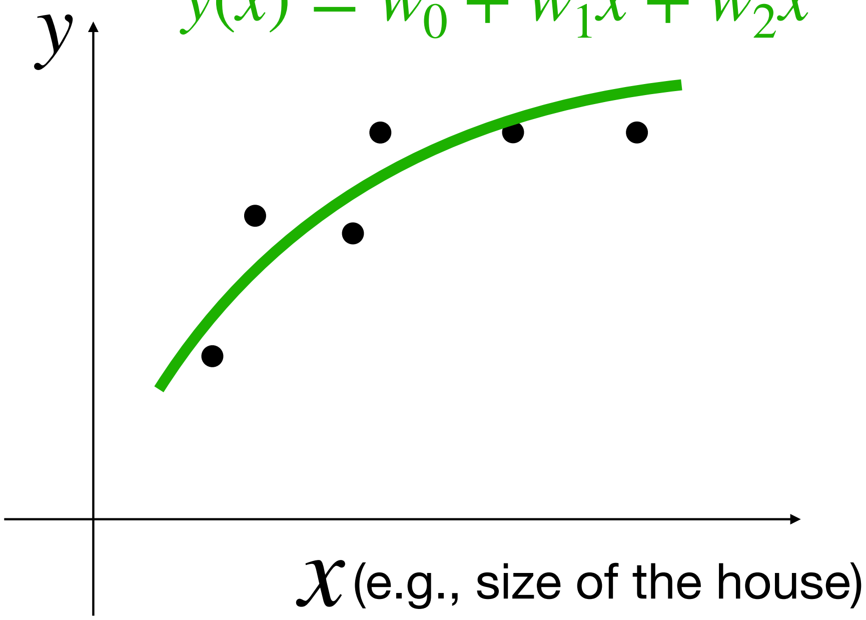


(Just right)



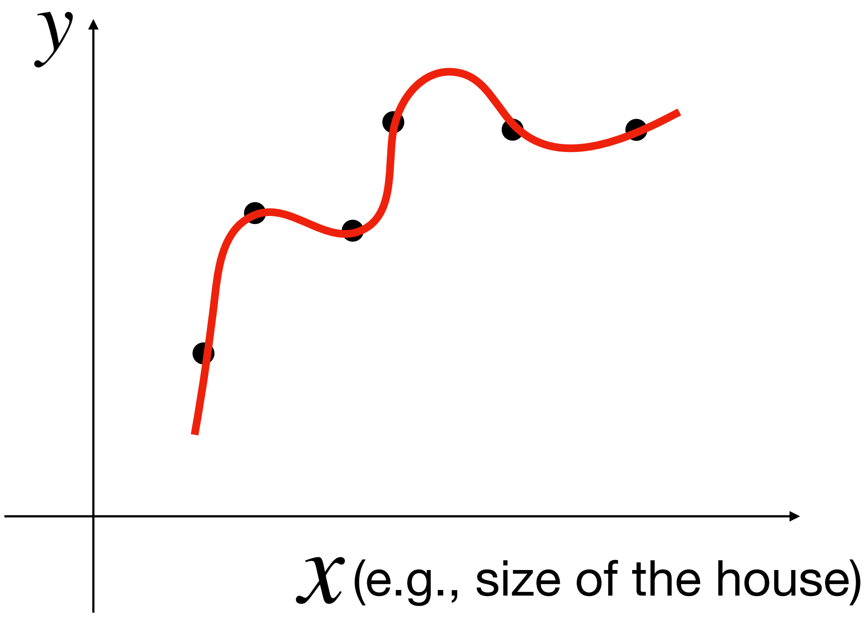
Overfitting

$$\bar{y}(x) = w_0 + w_1 x + w_2 x^2$$

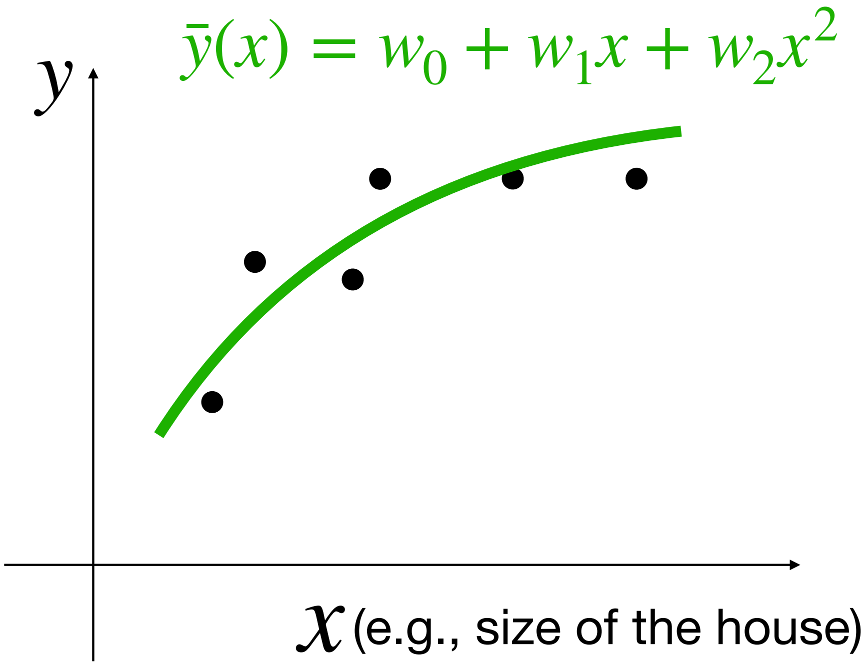


(Just right)

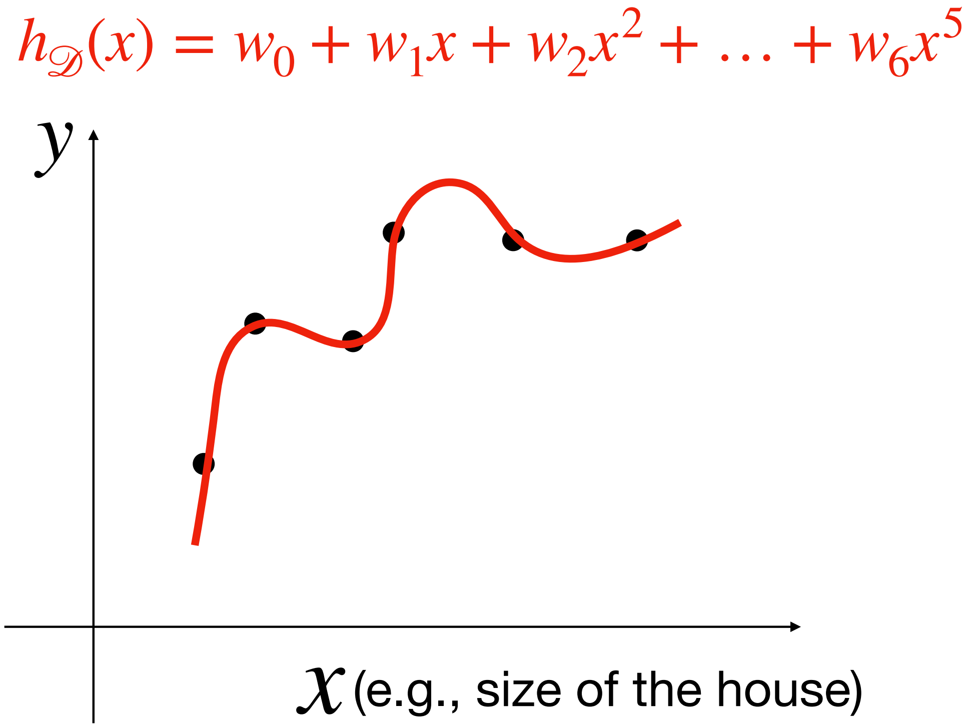
$$h_{\mathcal{D}}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_6 x^5$$



Overfitting



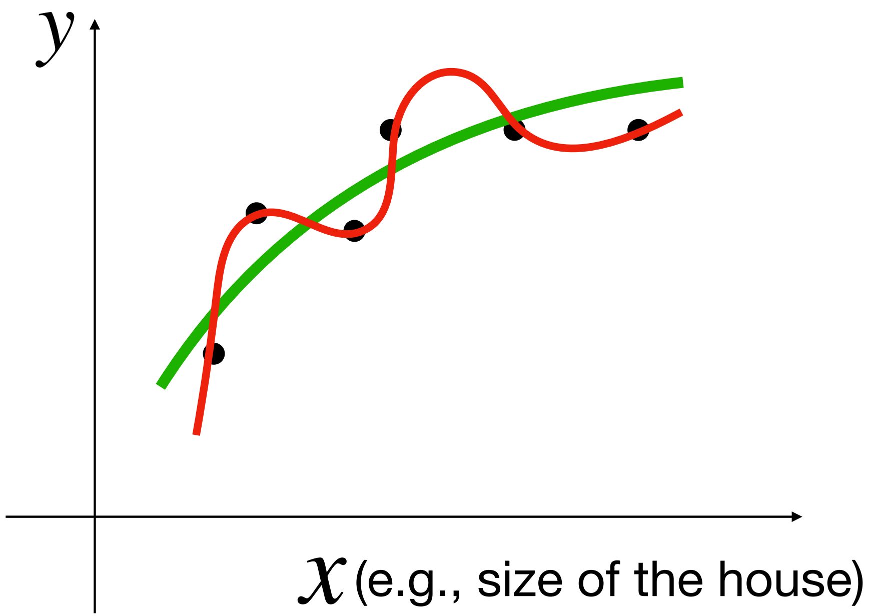
(Just right)



Overfitting

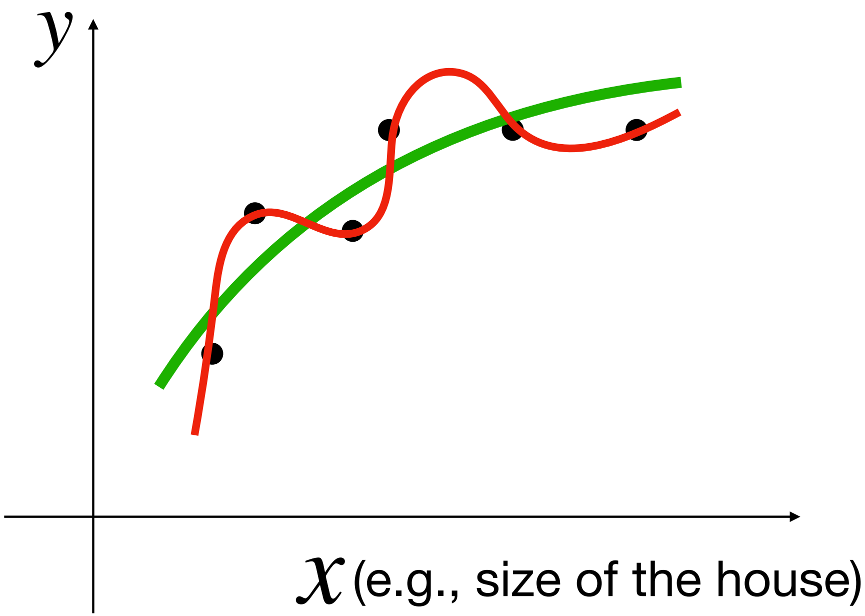
Overfitting

Just right versus Overfitting



Overfitting

Just right versus Overfitting

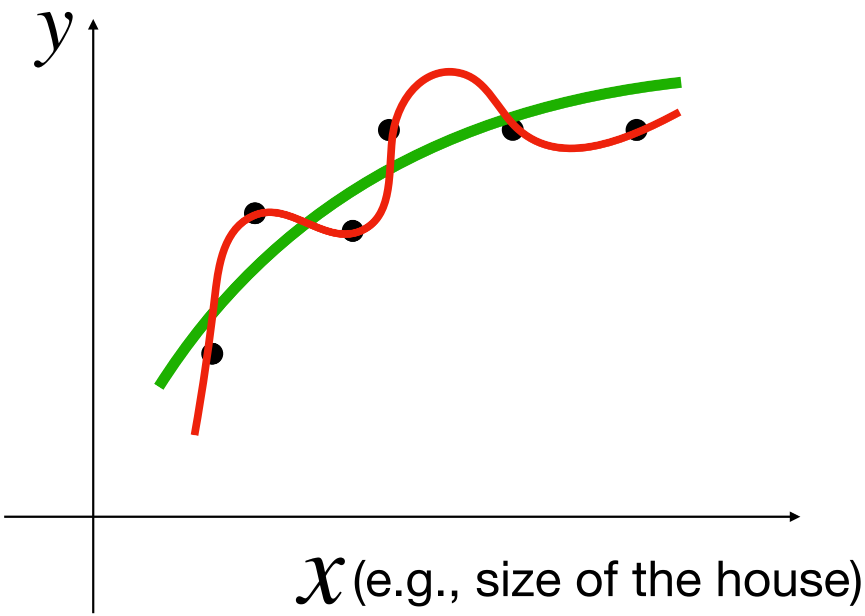


No strong bias:

Our hypothesis class is all polynomials up to 5-th order

Overfitting

Just right versus Overfitting



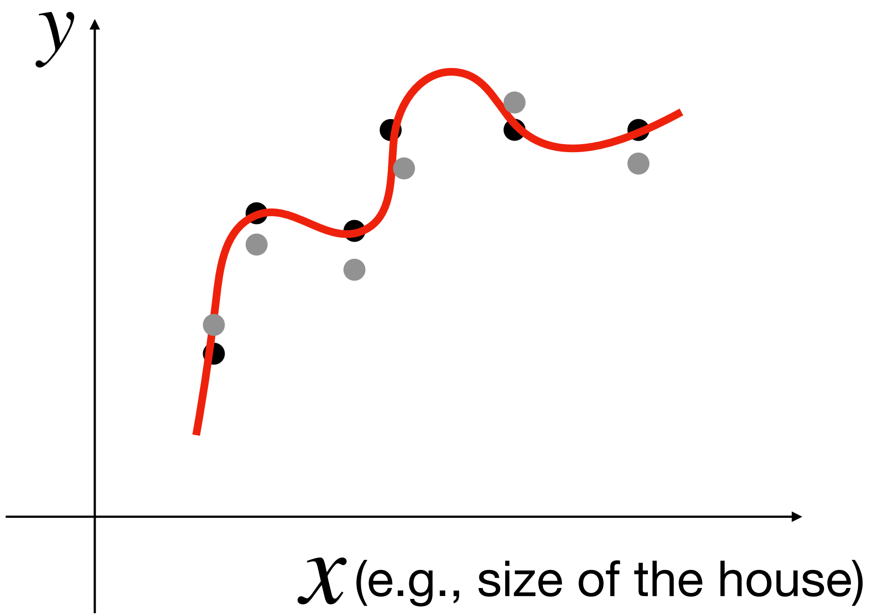
No strong bias:

Our hypothesis class is all polynomials up to 5-th order

i.e., a priori, no strong bias towards linear or quadratic, or cubic, etc

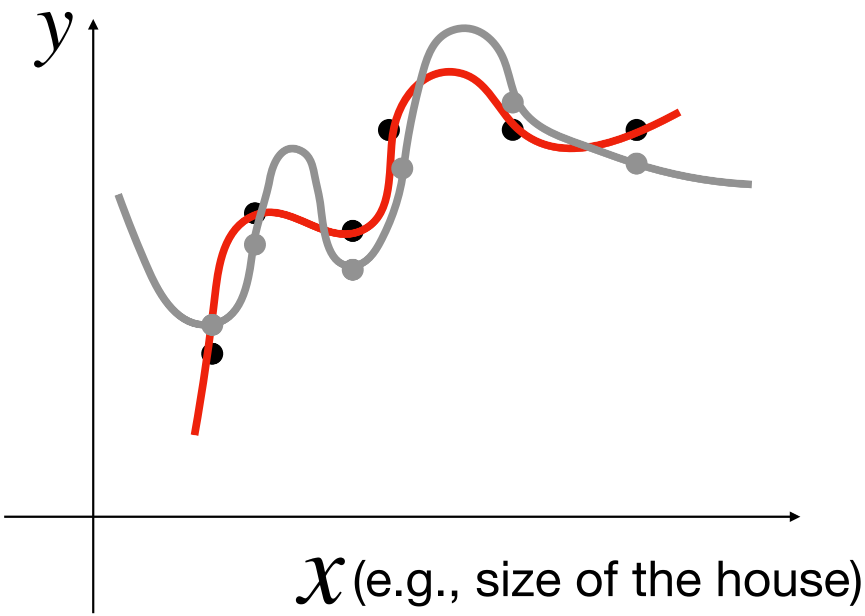
Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathscr{D}'$



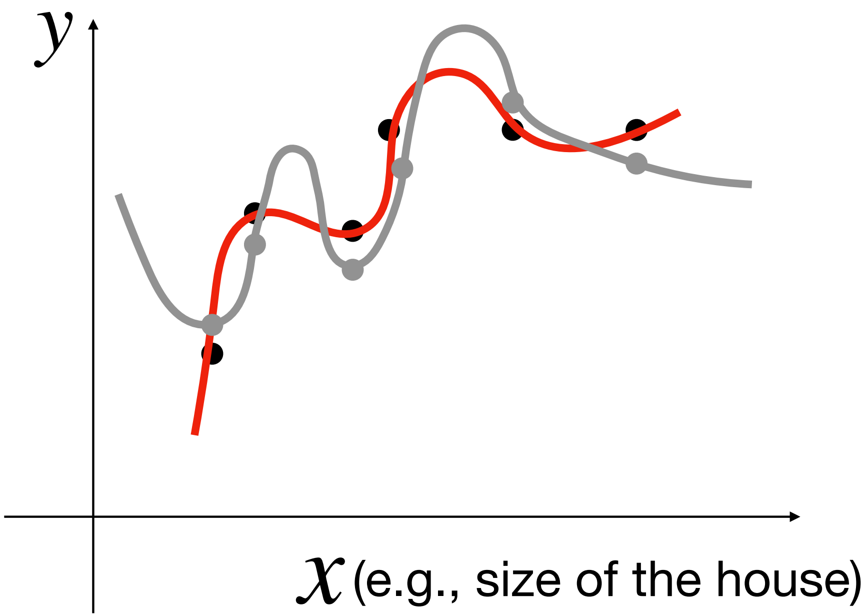
Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathscr{D}'$



Overfitting

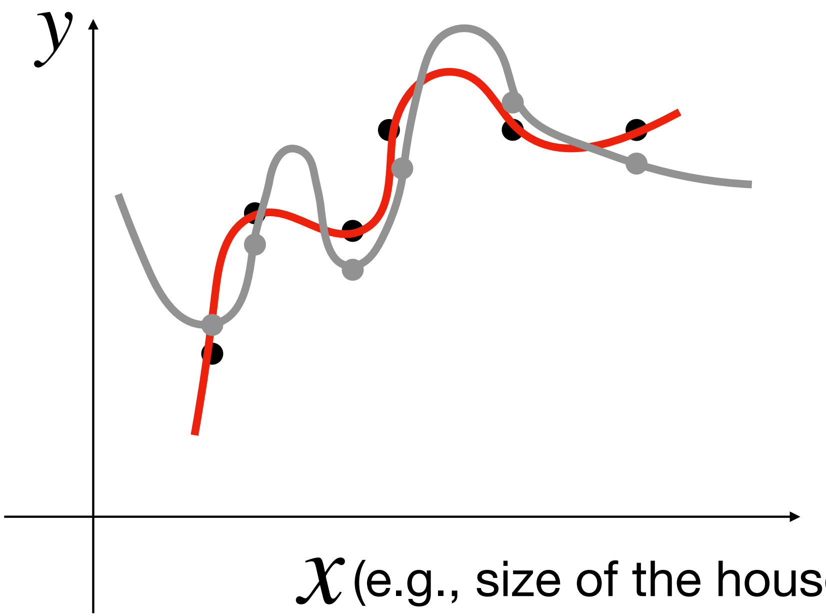
Redo the higher-order polynomial fitting on different dataset $\mathscr{D}'$



The new function can differ a lot from the old one

Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathscr{D}'$



x (e.g., size of the house)

The new function can differ a lot from the old one

This is called high variance

Summary on Overfitting

1. Often our model is too complex (e.g., can fit noise perfectly to achieve zero training error)

Summary on Overfitting

1. Often our model is too complex (e.g., can fit noise perfectly to achieve zero training error)

2. This causes overfitting, i.e., cannot generalize well on unseen test example

Summary on Overfitting

1. Often our model is too complex (e.g., can fit noise perfectly to achieve zero training error)

2. This causes overfitting, i.e., cannot generalize well on unseen test example

3. In this case, we have small bias, but large variance
(tiny change on the dataset cause large change in the fitted functions)

Outline of Today

1. Intro on Underfitting/Overfitting and Bias/Variance

2. Derivation of the Bias-Variance Decomposition

Generalization error

Given dataset $\mathscr{D}$, a hypothesis class $\mathscr{H}$, squared loss $\ell(h, x, y) = (h(x) - y)^2$, denote $h_{\mathscr{D}}$ as the ERM solution

$$h_D = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \left(h(x_i) - y_i\right)^2$$

Generalization error

Given dataset $\mathscr{D}$, a hypothesis class $\mathscr{H}$, squared loss $\ell(h, x, y) = (h(x) - y)^2$, denote $h_{\mathscr{D}}$ as the ERM solution

We are interested in the generalization error of $h_{\mathscr{D}}$:

$$\mathbb{E}_{\mathscr{D}} \mathbb{E}_{x,y \sim P} (h_{\mathscr{D}}(x) - y)^2$$

Generalization error

Given dataset $\mathscr{D}$, a hypothesis class $\mathscr{H}$, squared loss $\ell(h, x, y) = (h(x) - y)^2$, denote $h_{\mathscr{D}}$ as the ERM solution

We are interested in the generalization error of $h_{\mathscr{D}}$:

$$\mathbb{E}_{\mathscr{D}}\mathbb{E}_{x,y\sim P}(h_{\mathscr{D}}(x) - y)^2$$

Q: how to estimate this in practice?

The expectation of our model $h_{\mathscr{D}}$

Since $h_{\mathscr{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right]$$

$D_1 \sim P, \quad D_2 \sim P, \quad \cdots \quad D_{10} \sim P$

LR: $\downarrow$ LR $\downarrow$ LR $\downarrow$

$\hat{w}_1 \quad \hat{w}_2 \quad \hat{w}_{10}$

$$\bar{w} = \sum_{i=1}^{10} \hat{w}_i / 10$$

The expectation of our model $h_{\mathscr{D}}$

Since $h_{\mathscr{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right]$$

In other words, we have:

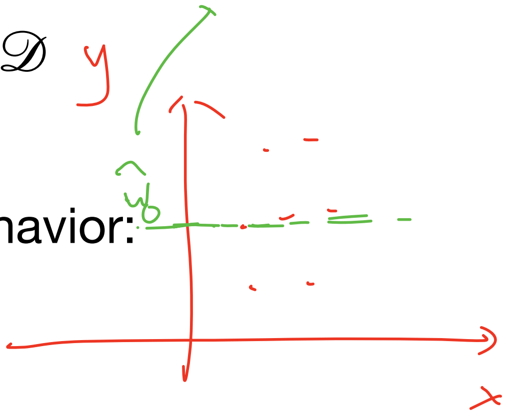
$$\bar{h}(x) = \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}(x)\right], \forall x$$

$$\hat{w}_1 \ldots \hat{w}_{10}$$

$$\bar{w}^T x = \frac{1}{10} \sum_{i=1}^{10} \hat{w}_i^T x$$

$$\min_{w_0} \sum_{i=1}^{n} (w_0 - y_i)^2 \qquad \hat{w}_0 \xrightarrow{LLN} \mathbb{E}_y[y]$$

The expectation of our model $h_{\mathscr{D}}$



Since $h_{\mathscr{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right]$$

Q: what is $\bar{h}$ is the case where hypothesis is $h(x) = w_0$?

In other words, we have:

$$\bar{h}(x) = \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}(x)\right], \forall x$$

$$\bar{h} = \mathbb{E}_y[y]$$

The expectation of our model $h_{\mathscr{D}}$

Since $h_{\mathscr{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right]$$

In other words, we have:

$$\bar{h}(x) = \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}(x)\right], \forall x$$

Q: what is $\bar{h}$ is the case where hypothesis is $h(x) = w_0$?

A: $\bar{h}(x) = \mathbb{E}_y[y]$

Formal definition of Bias and Variance

$$\bar{h} := \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

Bayes opt

Bias^2: (squared)difference between $\bar{h}$ and the best $\bar{y}(x)$, i.e., $\mathbb{E}_x\left(\bar{y}(x) - \bar{h}(x)\right)^2$

Formal definition of Bias and Variance

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

Bias^2: (squared)difference between $\bar{h}$ and the best $\bar{y}(x)$, i.e., $\mathbb{E}_x\left(\bar{y}(x) - \bar{h}(x)\right)^2$

Difference between our mean and the best

Formal definition of Bias and Variance

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

Bias^2: (squared)difference between $\bar{h}$ and the best $\bar{y}(x)$, i.e., $\mathbb{E}_x\left(\bar{y}(x) - \bar{h}(x)\right)^2$

Difference between our mean and the best

Variance: difference from $\bar{h}$ and $h_{\mathscr{D}}$, i.e., $\mathbb{E}_{\mathscr{D}}\mathbb{E}_x\left(h_{\mathscr{D}}(x) - \bar{h}(x)\right)^2$

Formal definition of Bias and Variance

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

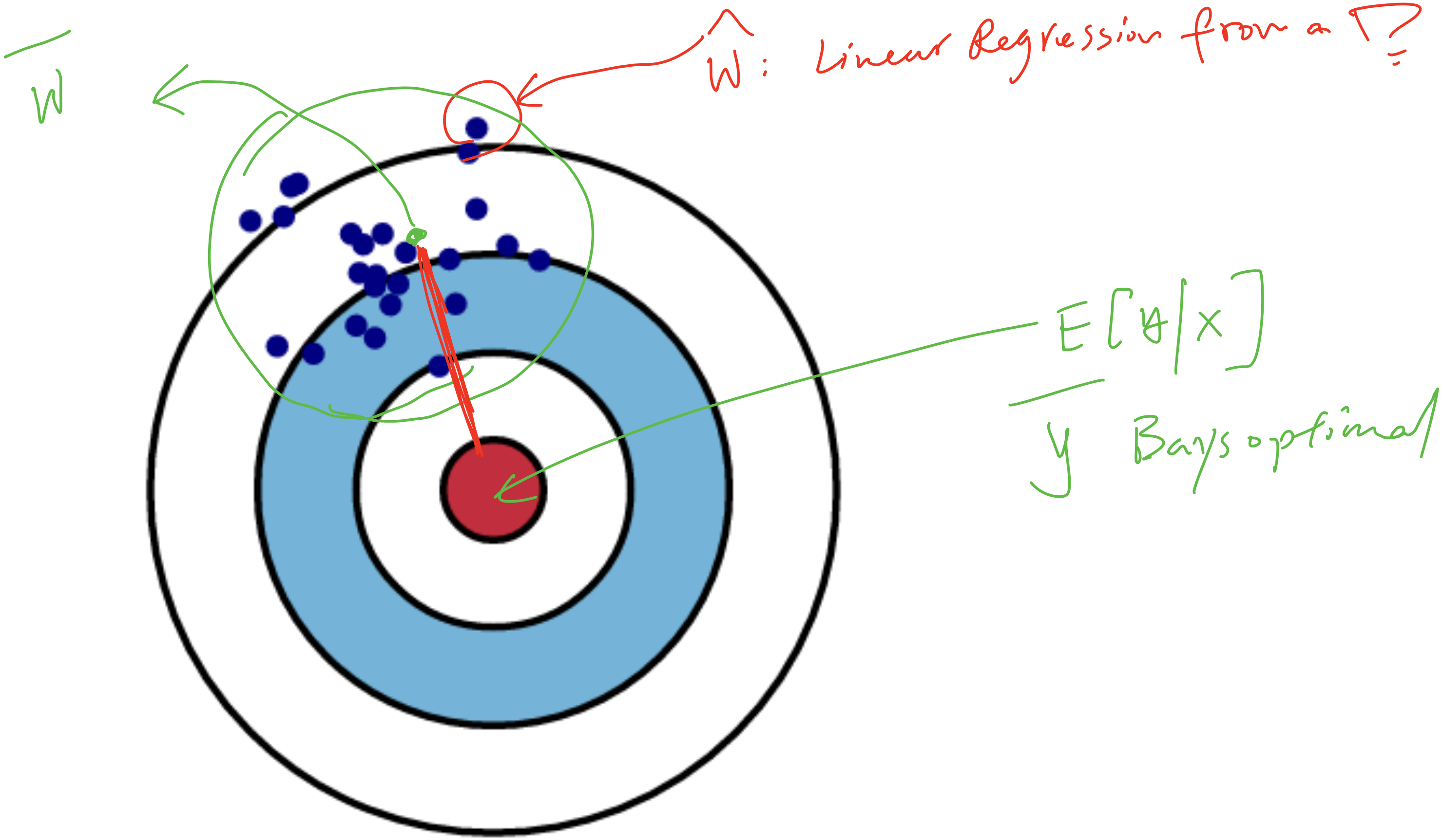
Bias^2: (squared)difference between $\bar{h}$ and the best $\bar{y}(x)$, i.e., $\mathbb{E}_x \left(\bar{y}(x) - \bar{h}(x)\right)^2$

Difference between our mean and the best

Variance: difference from $\bar{h}$ and $h_{\mathscr{D}}$, i.e., $\mathbb{E}_{\mathscr{D}}\mathbb{E}_x \left(h_{\mathscr{D}}(x) - \bar{h}(x)\right)^2$

Fluctuation of our random model around its mean

Bias-Variance illustration



Generalization error decomposition

$$\bar{h} := \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

What we gonna show now:

Generalization error decomposition

$$\bar{h} := \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

What we gonna show now:

$$\mathbb{E}_{\mathcal{D}}\mathbb{E}_{x,y\sim P}(h_{\mathcal{D}}(x) - y)^2$$

= **Bias**^{2} + **Variance** + Noise (unavoidable, independent of Algs/models)

Generalization error decomposition

$$\bar{h} := \mathbb{E}_{\mathscr{D}}\left[h_{\mathscr{D}}\right] \qquad \bar{y}(x) := \mathbb{E}[y \mid x]$$

What we gonna show now:

$$\mathbb{E}_{\mathscr{D}}\mathbb{E}_{x,y\sim P}(h_{\mathscr{D}}(x) - y)^2$$

= **Bias**^{2} + **Variance** + Noise (unavoidable, independent of Algs/models)

We will use the following trick twice: $(x-y)^2 = (x-z)^2 + (z-y)^2 + 2(x-z)(z-y)$

$(x - z + z - y)^2$

$\to \mathbb{E}_{D} \mathbb{E}_{xy}$

$\bar{h} = \mathbb{E}_{D}[h_D]$

$$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2$$

$$= \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x) + \bar{h}(x) - y)^2$$

$\bar{h} = \mathbb{E}_{D}[h_D]$

$$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2$$

$$= \mathbb{E}(\underbrace{h_{\mathcal{D}}(x) - \bar{h}(x)}_{①} + \underbrace{\bar{h}(x) - y}_{②})^2$$

$$= \underbrace{\mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2}_{①^2} + \underbrace{\mathbb{E}(\bar{h}(x) - y)^2}_{②^2} + 2\mathbb{E}_{\mathcal{D},x,y}\left[\underbrace{(h_{\mathcal{D}}(x) - \bar{h}(x))(\bar{h}(x) - y)}_{① \times ②}\right]$$

Var

$$\mathbb{E}(h_{\mathscr{D}}(x) - y)^2$$

$$= \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x) + \bar{h}(x) - y)^2$$

$$= \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 + 2\mathbb{E}_{\mathscr{D},x,y}\left[(h_{\mathscr{D}}(x) - \bar{h}(x))(\bar{h}(x) - y)\right]$$

Def of $\bar{h}$:

$$\bar{h} = \mathbb{E}_{D}[h_D]$$

This term is zero since:

$$\mathbb{E}_{x,y,\mathscr{D}}\left[(h_{\mathscr{D}}(x) - \bar{h}(x))(\bar{h}(x) - y)\right]$$

$$= \mathbb{E}_{x,y}\left[\mathbb{E}_{\mathscr{D}}(h_{\mathscr{D}}(x) - \bar{h}(x)) \cdot (\bar{h}(x) - y)\right]$$

$$= \mathbb{E}_{x,y}\left[(\bar{h}(x) - \bar{h}(x)) \cdot (\bar{h}(x) - y)\right]$$

$= 0$

Bias: $\bar{h} - \underbrace{E[y|x]}_{:=\bar{y}}$

$$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2$$

$$-\bar{y}(x) + \bar{y}(x)$$

$$= \underbrace{\mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2}_{\text{Variance}} + \mathbb{E}(\bar{h}(x) - y)^2$$

$$\mathbb{E}(h_{\mathscr{D}}(x) - y)^2$$

$$= \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2$$

Variance

$$\mathbb{E}(h_{\mathscr{D}}(x) - y)^2$$

$$= \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2$$

Variance

$\bar{y}(x) = \mathbb{E}[y|x]$

$$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2$$

$$= \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2$$

Complete square

$$= \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2$$

Variance

$$\mathbb{E}(h_{\mathscr{D}}(x) - y)^2$$

$$= \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2$$

Variance

$$= \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2$$

$$= \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2$$

$$+ 2\mathbb{E}(\bar{h}(x) - \bar{y}(x))(\bar{y}(x) - y)$$

$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2$

$= \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2$

Variance

Bias^2

$= \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2$

$= \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2$

$+ 2\mathbb{E}(\bar{h}(x) - \bar{y}(x))(\bar{y}(x) - y)$

$\mathbb{E}_{x,y} := \mathbb{E}_{x} \mathbb{E}_{y|x}$

This term is zero since:

$= \mathbb{E}_x \left[(\bar{h}(x) - \bar{y}(x)) \cdot \mathbb{E}_{y|x}(\bar{y}(x) - y) \right]$

$= \mathbb{E}_x \left[(\bar{h}(x) - \bar{y}(x)) \cdot (\bar{y}(x) - \mathbb{E}_{y|x}[y]) \right] = 0$

$\mathbb{E}_{y|x}[y] := \bar{y}(x)$

Putting the derivations together, we arrive at:

$$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2 = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2$$

Variance

Bias^2

Putting the derivations together, we arrive at:

$$\mathbb{E}(h_{\mathscr{D}}(x) - y)^2 = \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2$$

Variance

Bias^2

Noise

Putting the derivations together, we arrive at:

$$\mathbb{E}(h_{\mathscr{D}}(x) - y)^2 = \mathbb{E}(h_{\mathscr{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2$$

Variance

Bias^2

Noise

Note that the noise term is independent of training algorithms / models