# Bias-Variance Tradeoff

# Announcements

# 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)

# Outline of Today

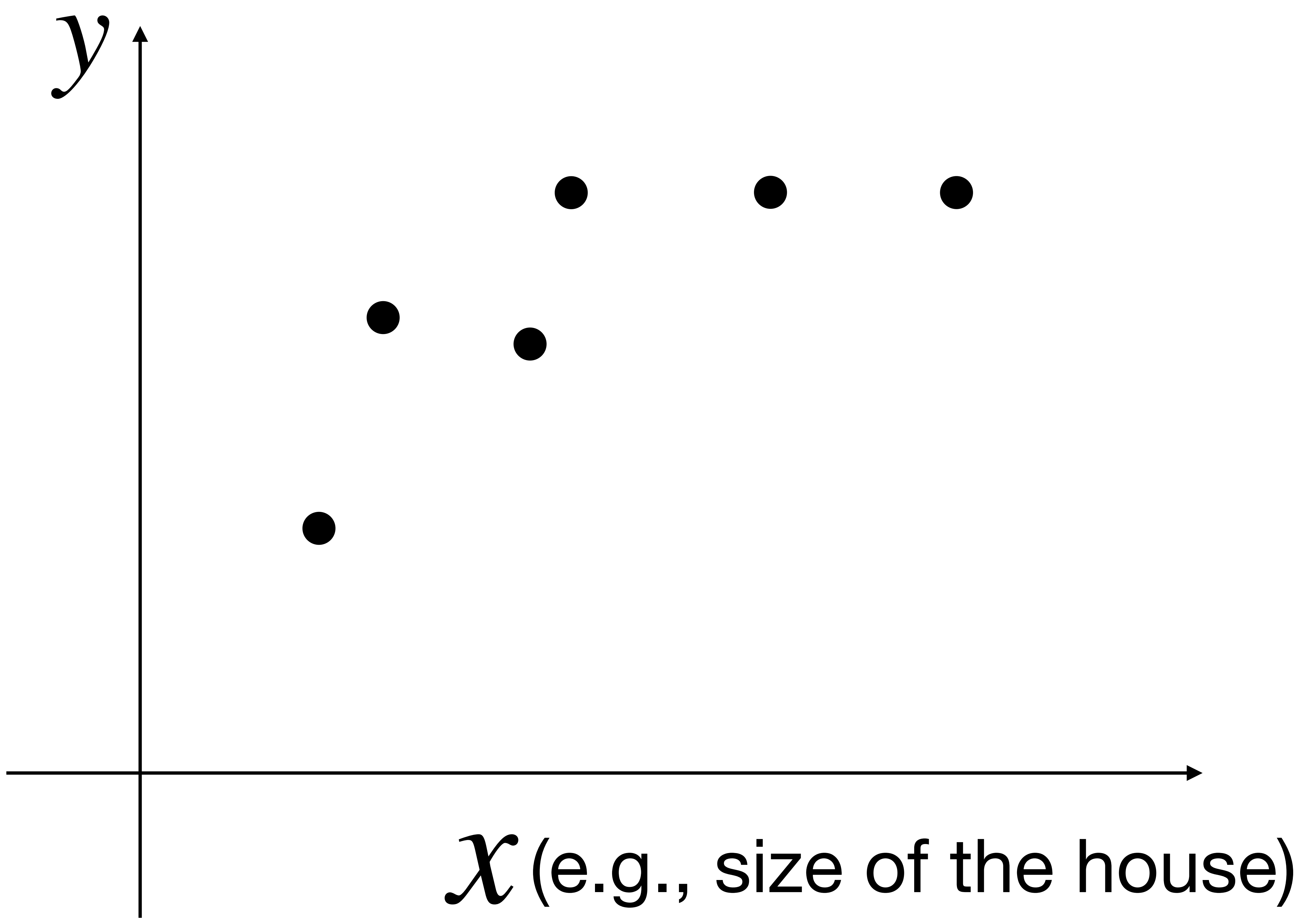
1. Intro on Underfitting/Overfitting and Bias/Variance

2. Derivation of the Bias-Variance Decomposition

3. Example on Ridge Linear Regression

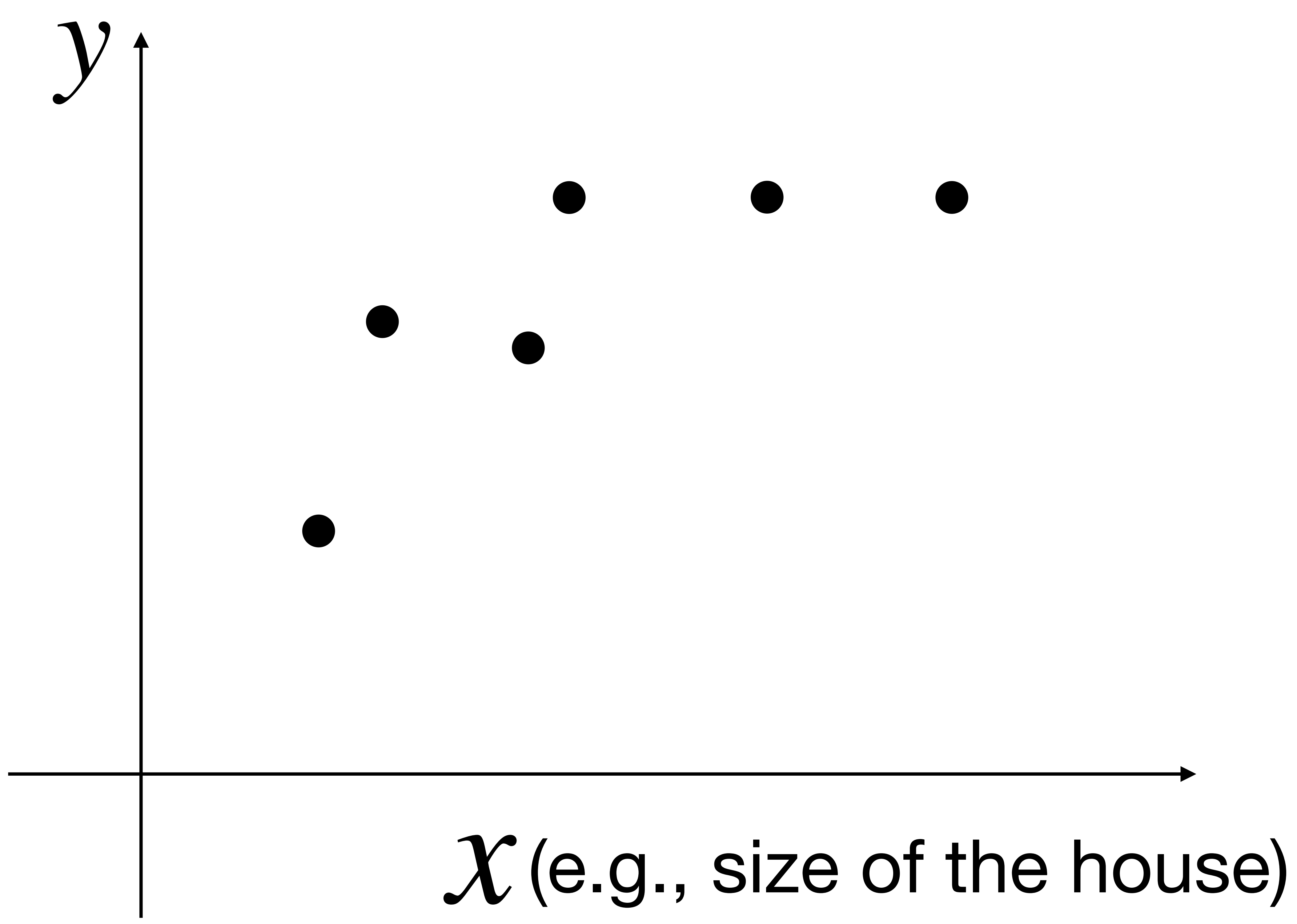
# Bayes optimal predictor

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



# Bayes optimal predictor

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

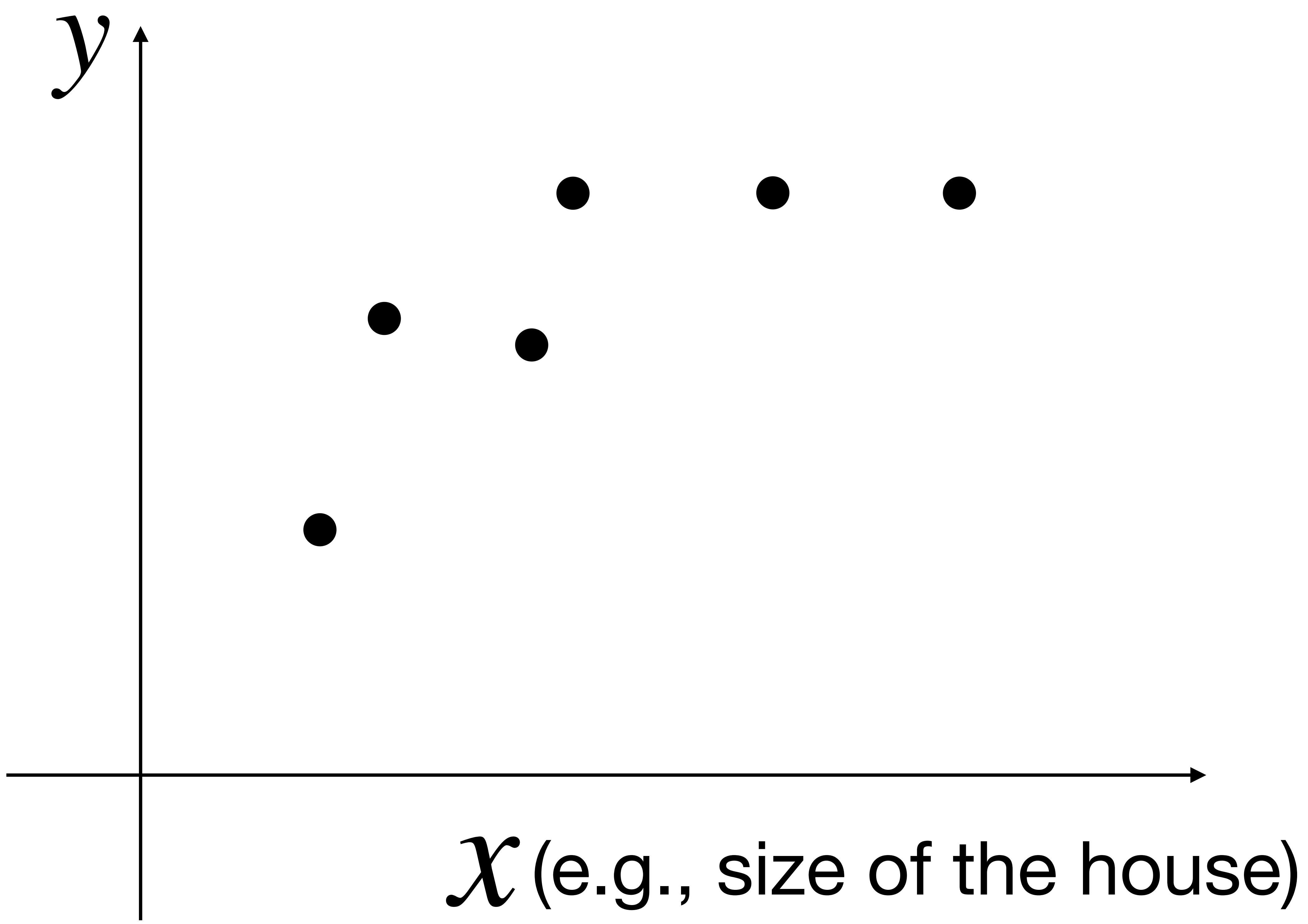



The Bayes optimal regressor:

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

# Bayes optimal predictor

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



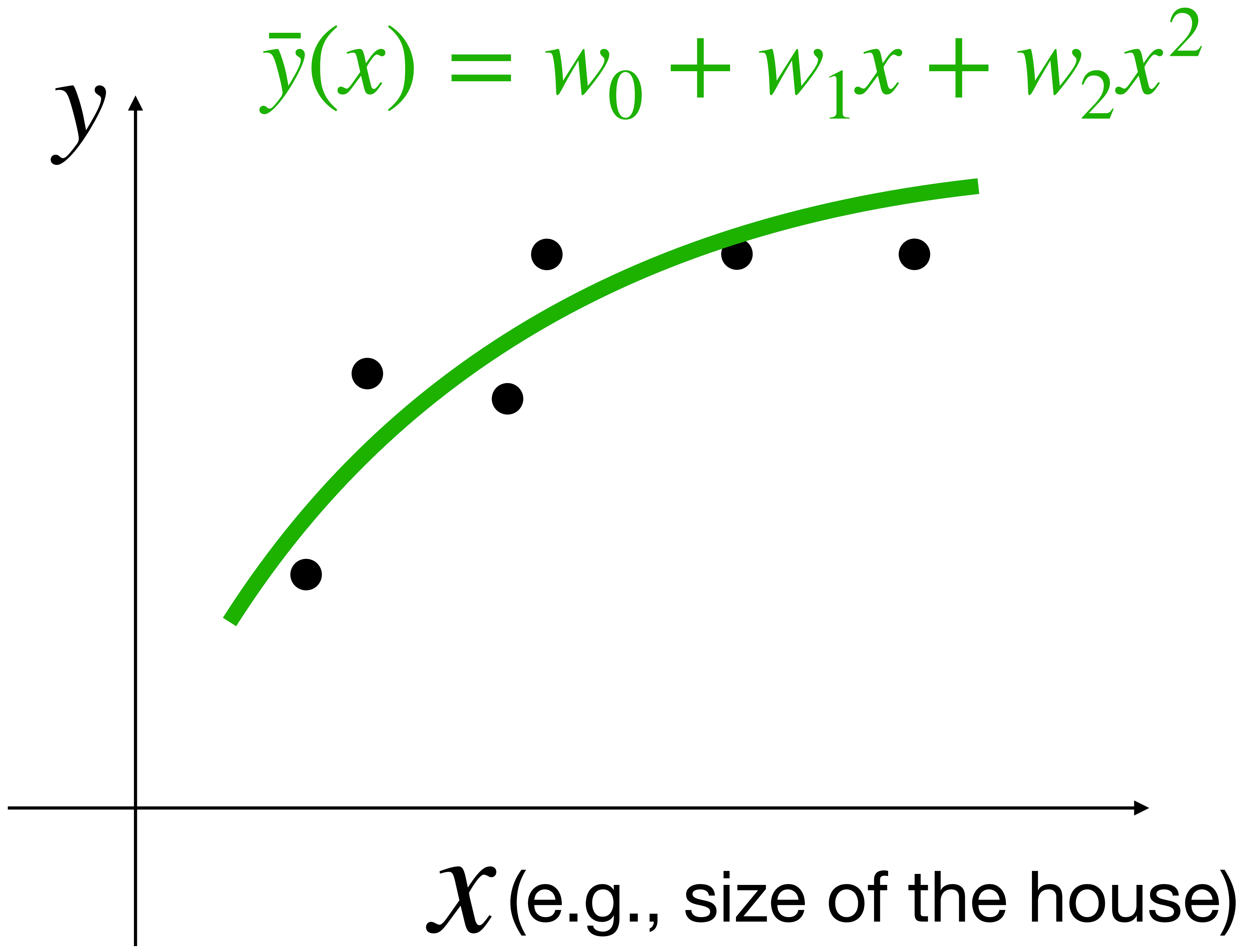

The Bayes optimal regressor:

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

The best we could do, cannot beat this one

# Bayes optimal predictor

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

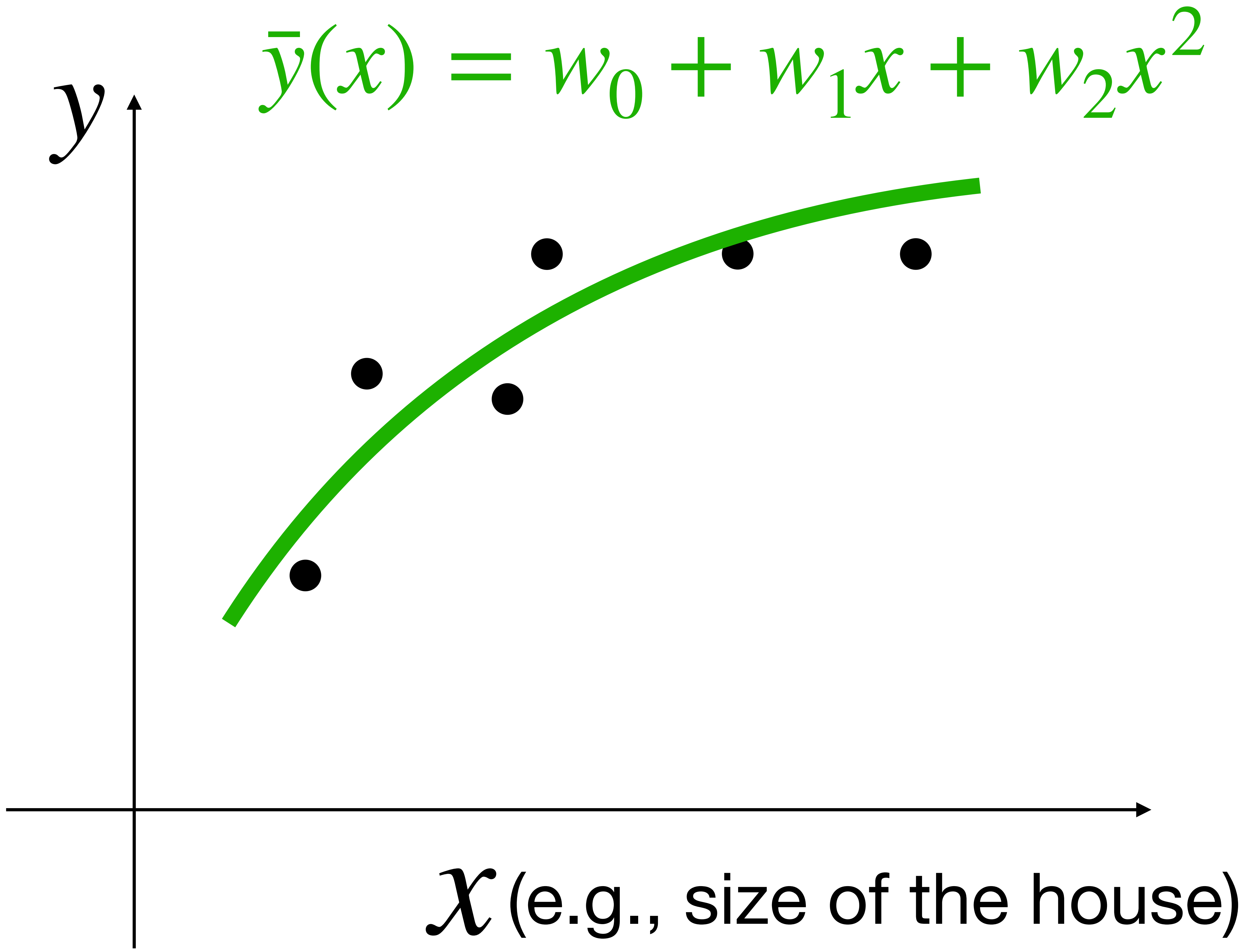



The Bayes optimal regressor:

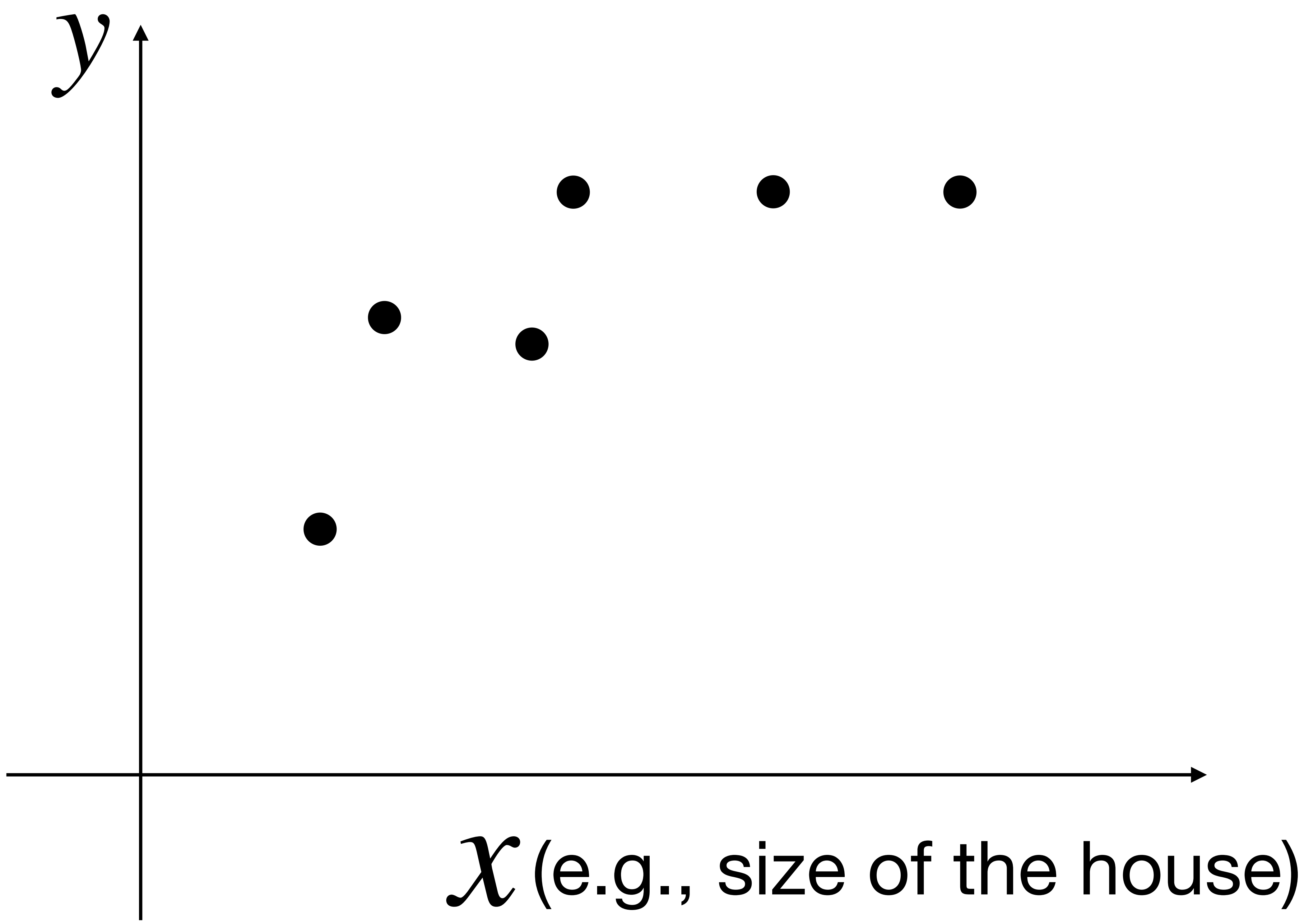
$$\bar{y}(x) := \mathbb{E}[y \mid x]$$

The best we could do, cannot beat this one

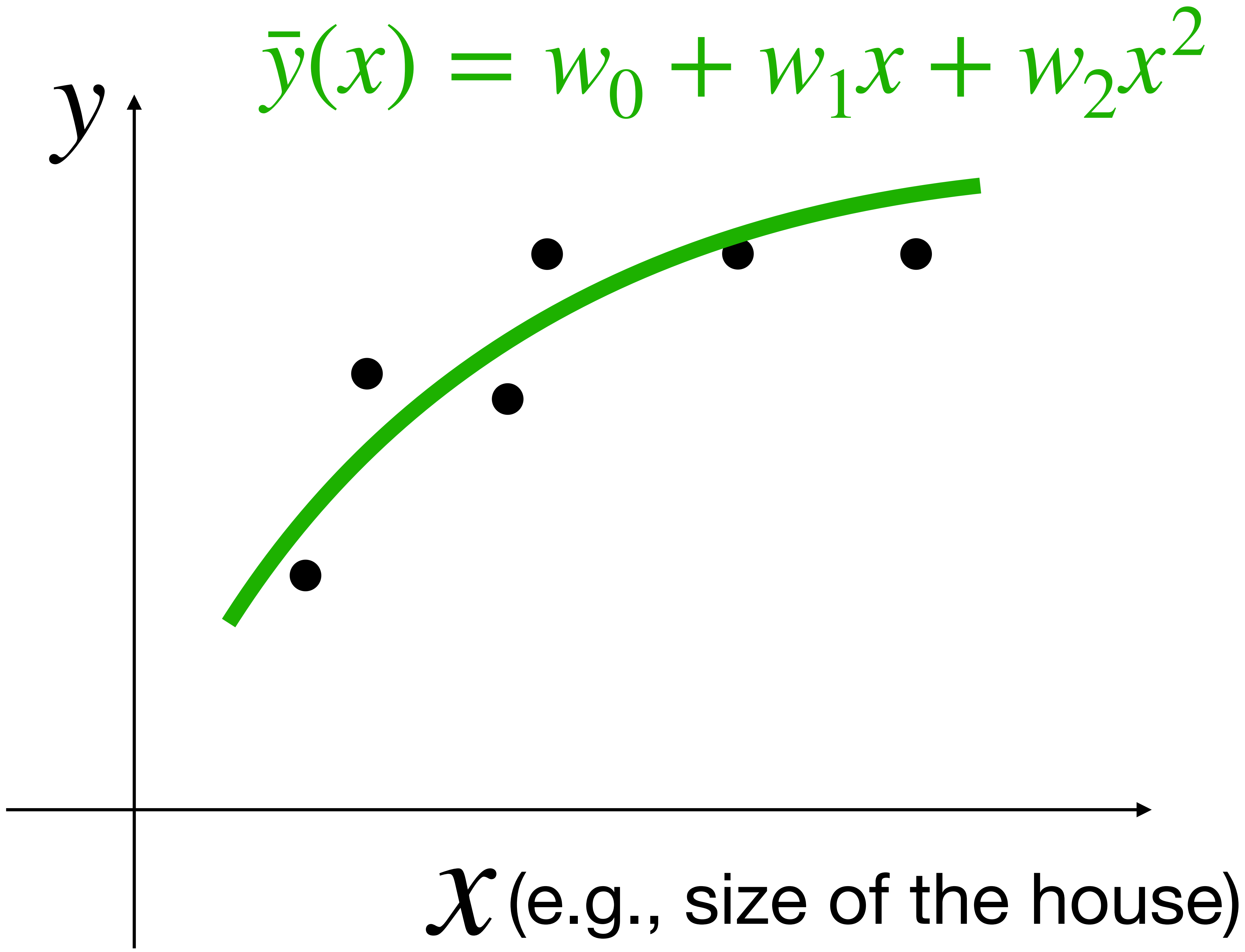
# Underfitting



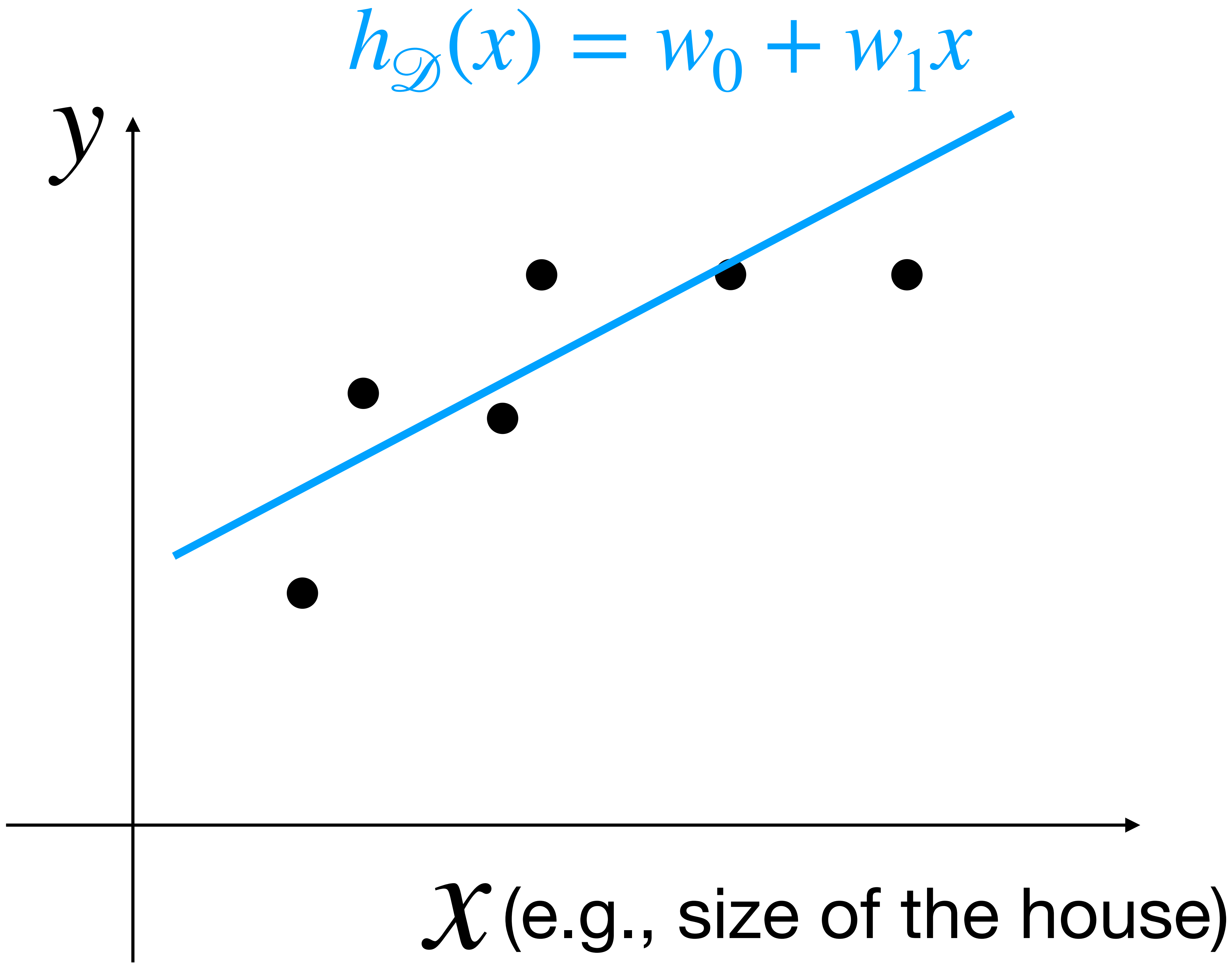

(Just right)



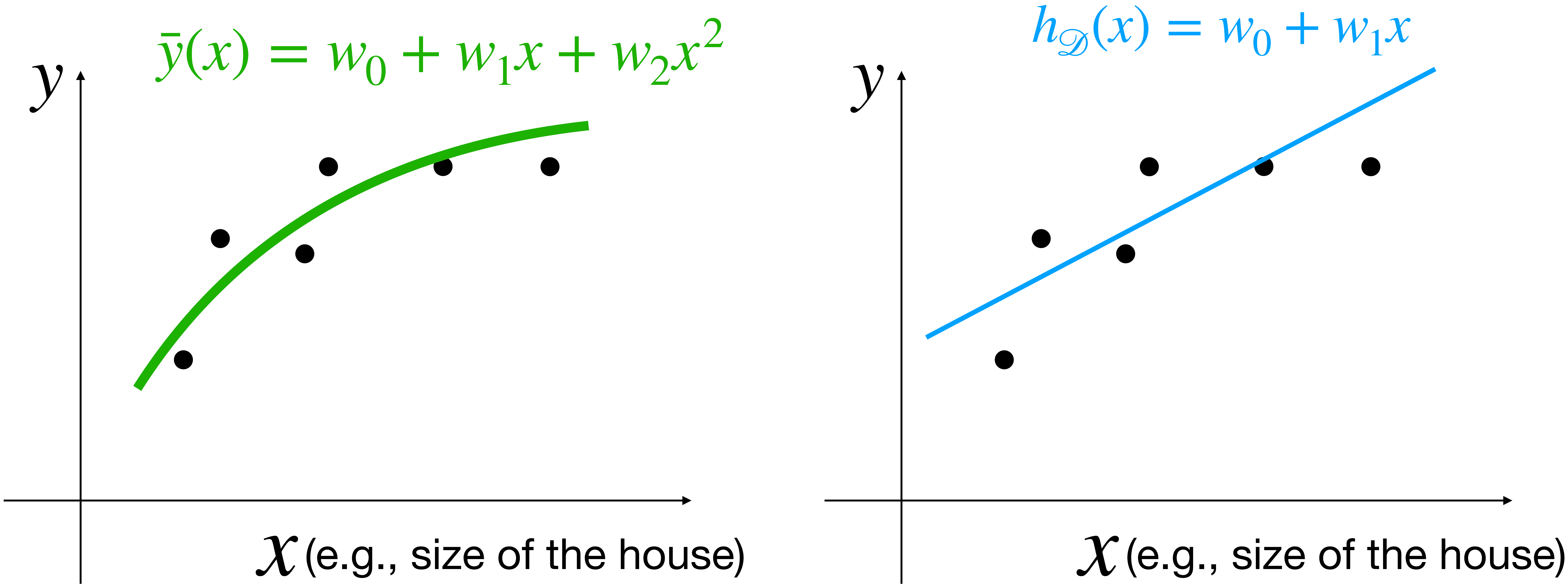
# Underfitting




(Just right)



# Underfitting

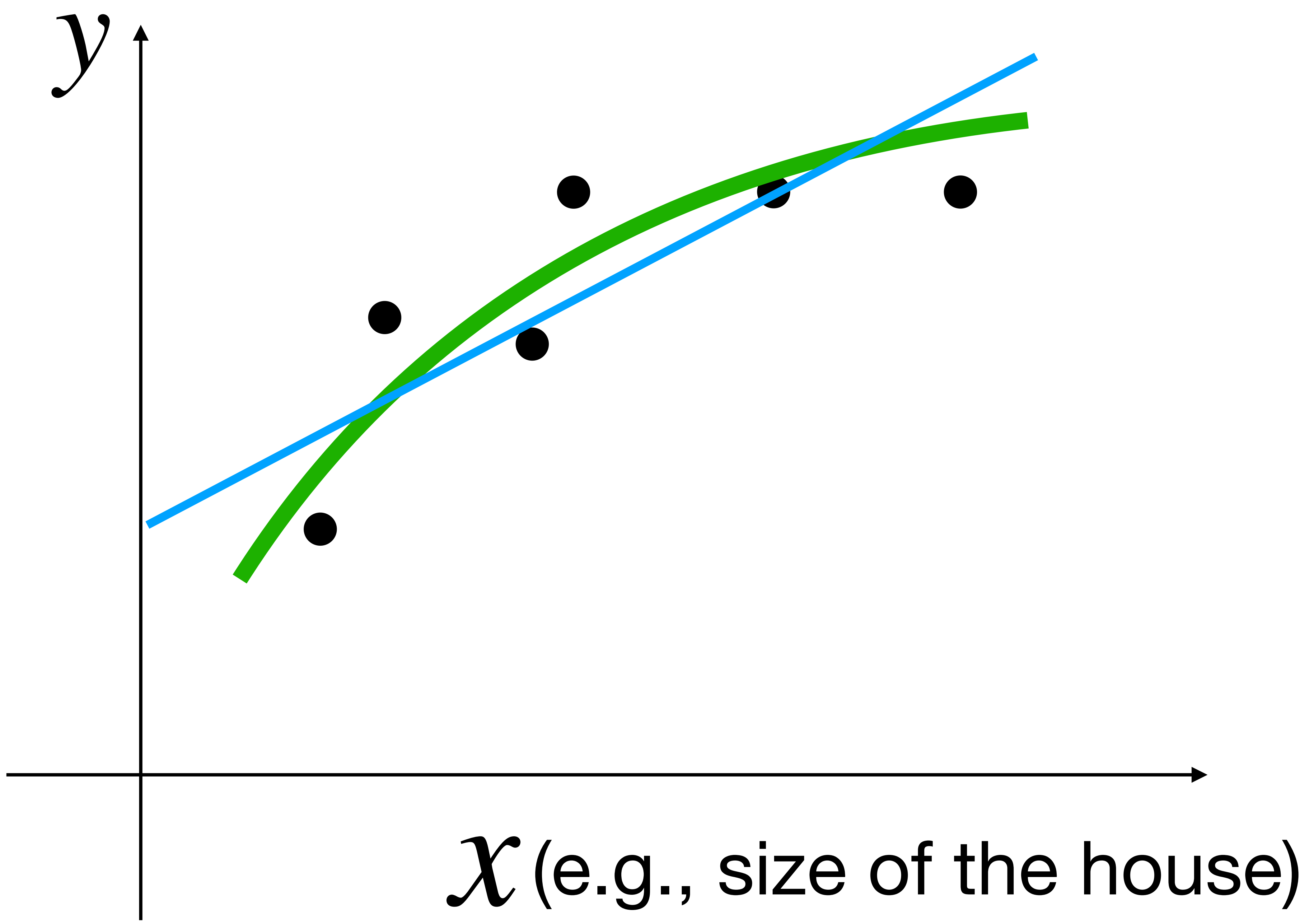



(Just right)

Underfitting

# Underfitting

Just right versus Underfitting

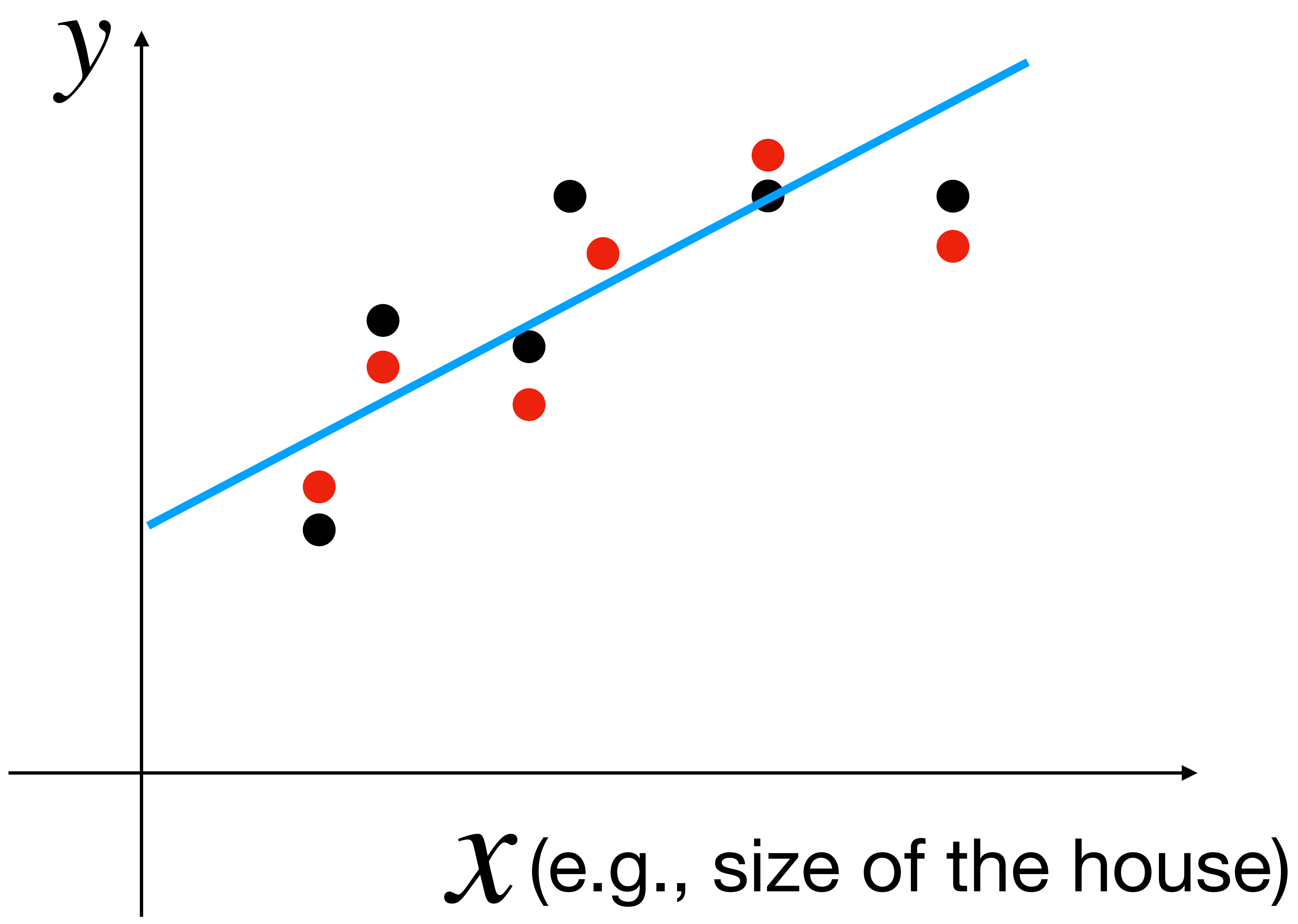



Bias:

Bias towards to linear models

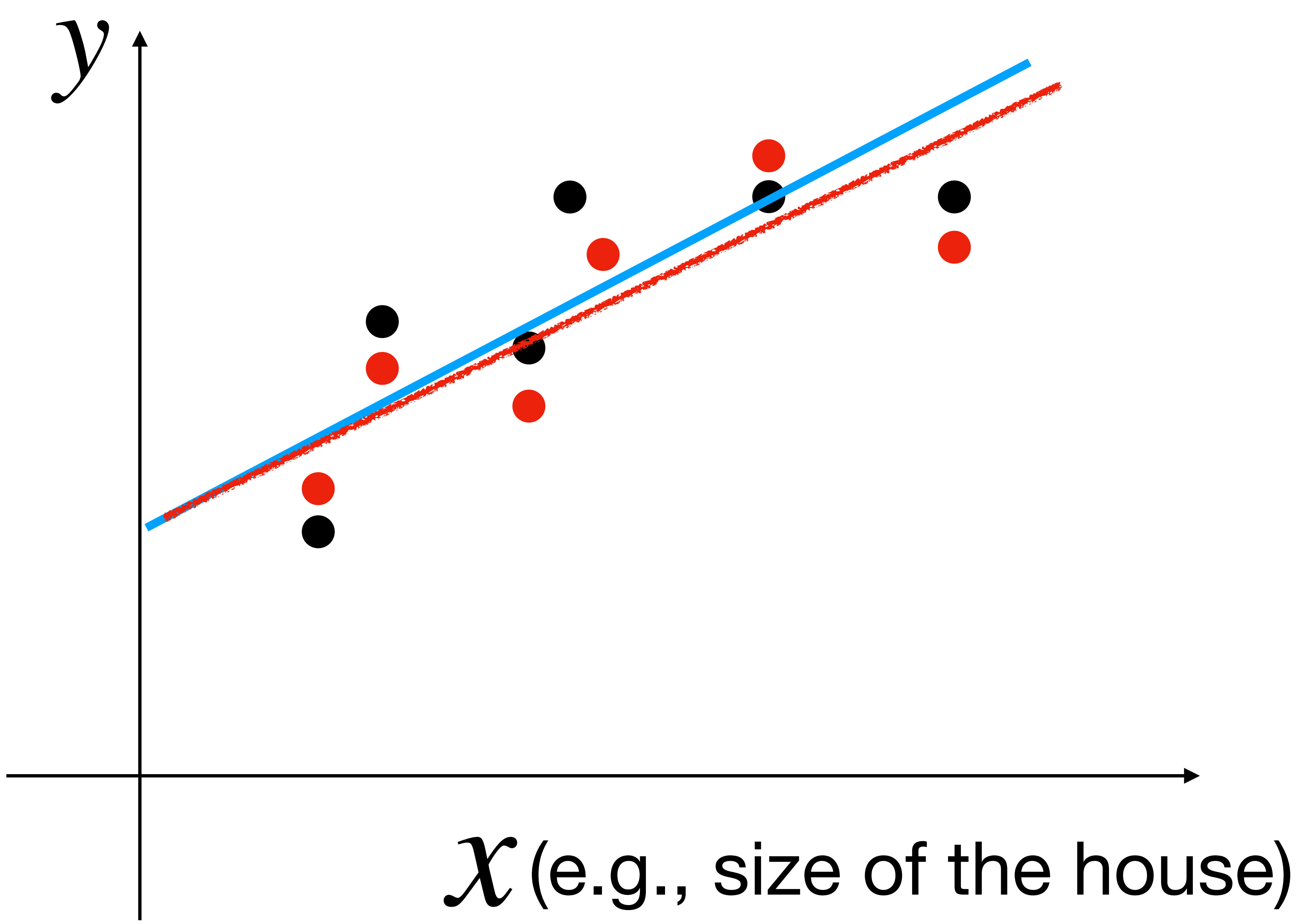
# Underfitting

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



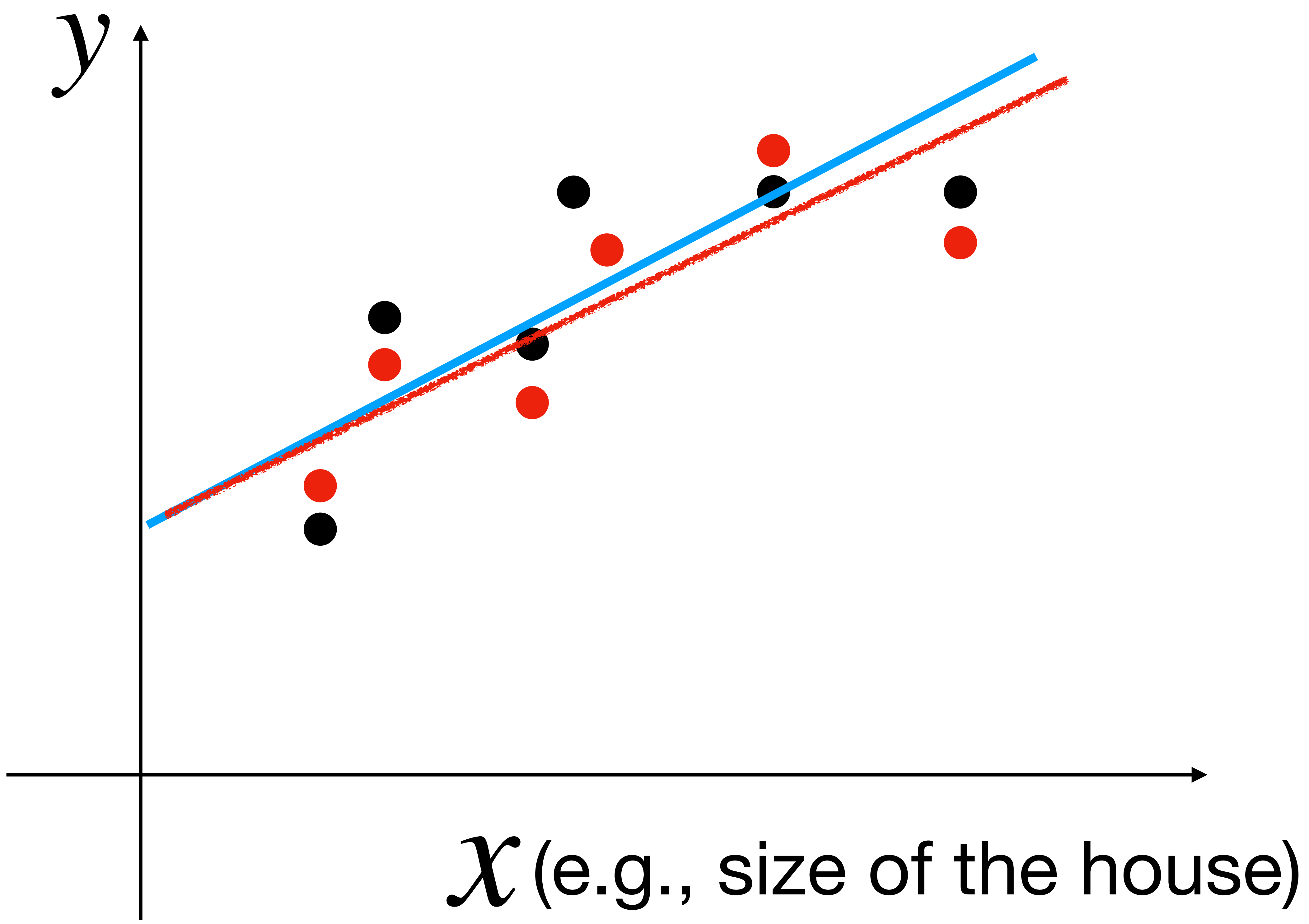
# Underfitting

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



# Underfitting

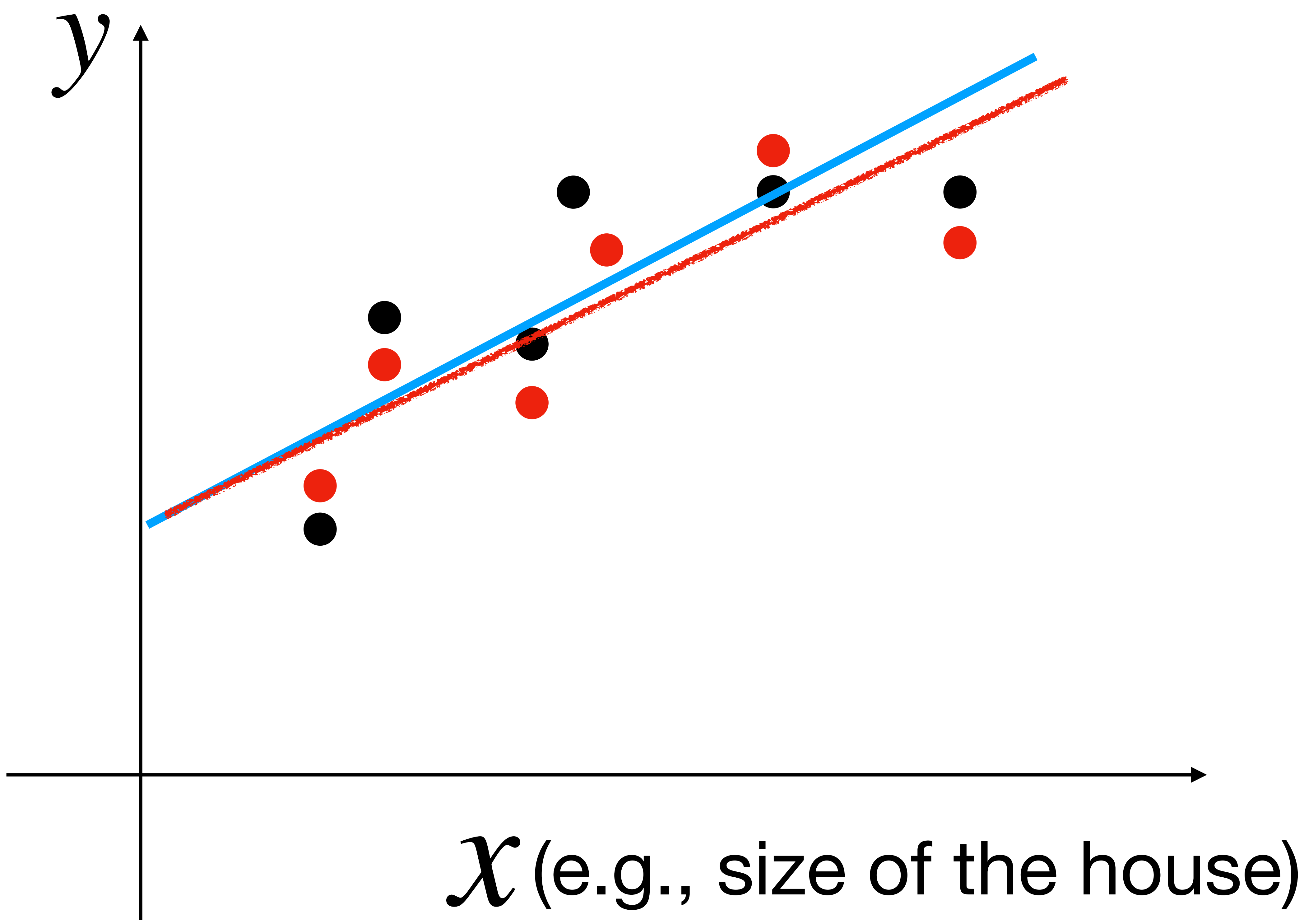
Now let's redo linear regression on a different dataset $\mathscr{D}'$, but from the same distribution




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

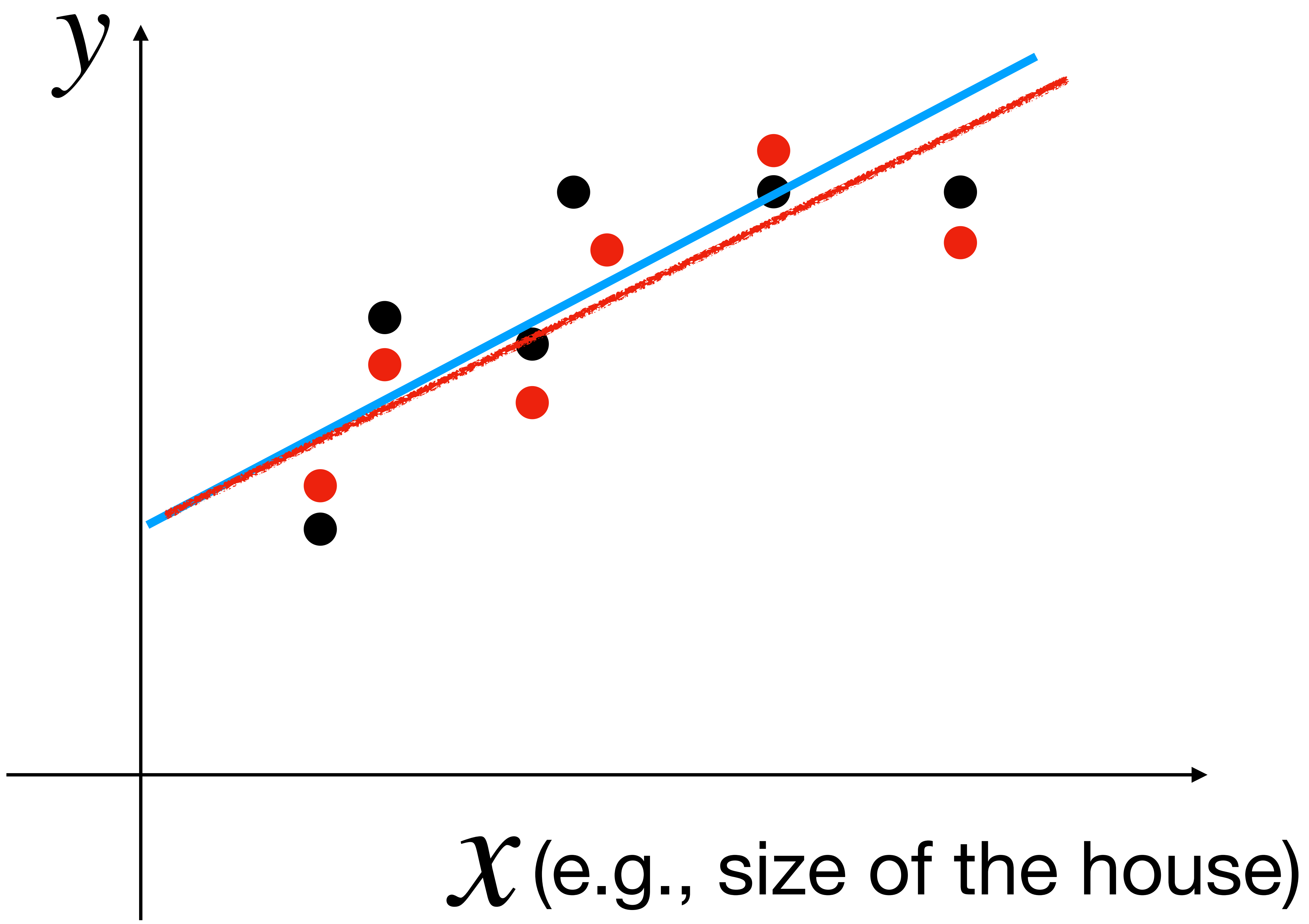



The new linear function does not differ too much from the old one

This is called low variance

# Underfitting

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



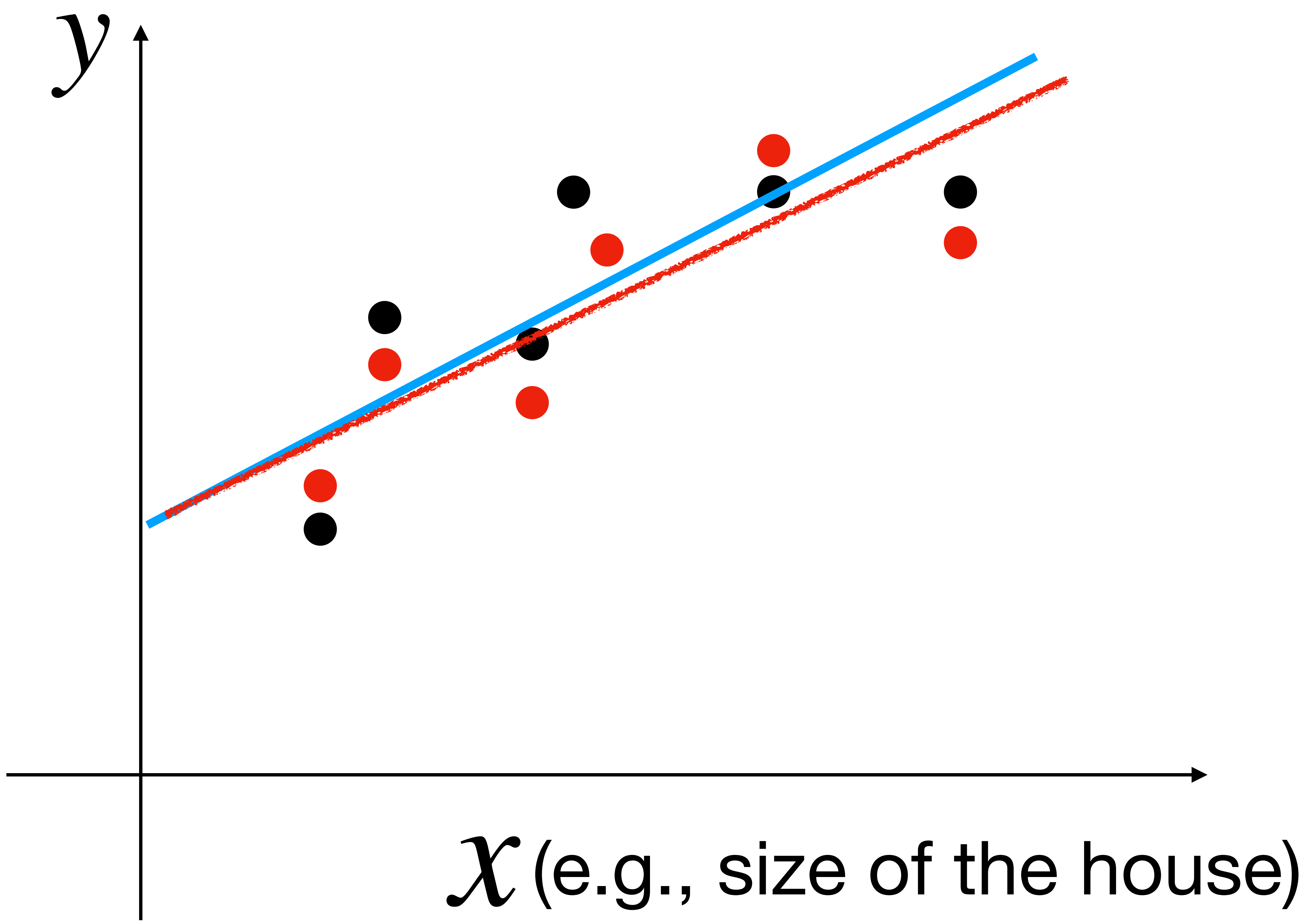

The new linear function does not differ too much from the old one

This is called low variance

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

# 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

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

A: in this case, $w_0$ models the mean of the y in data

# 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

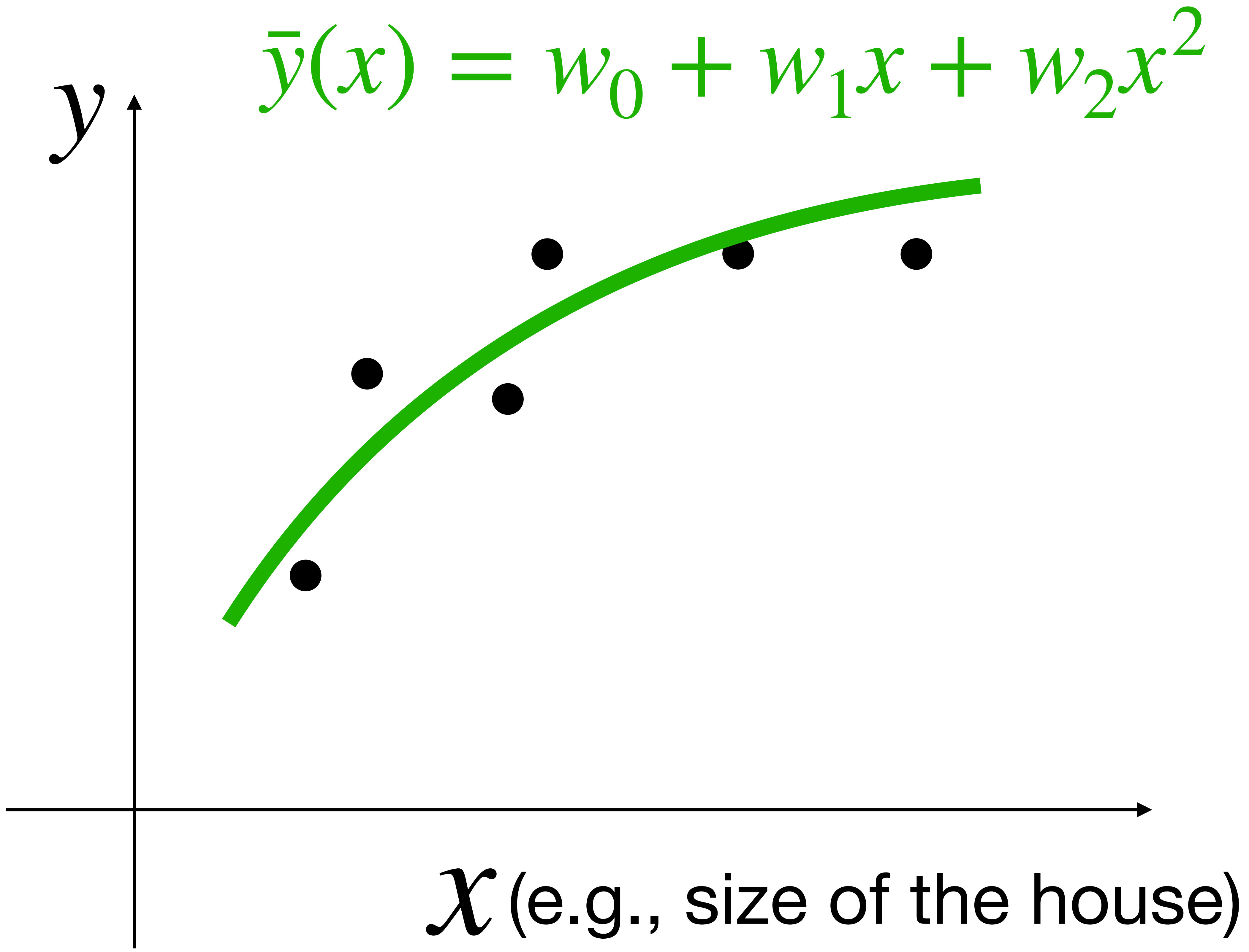
# Summary on underfitting

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

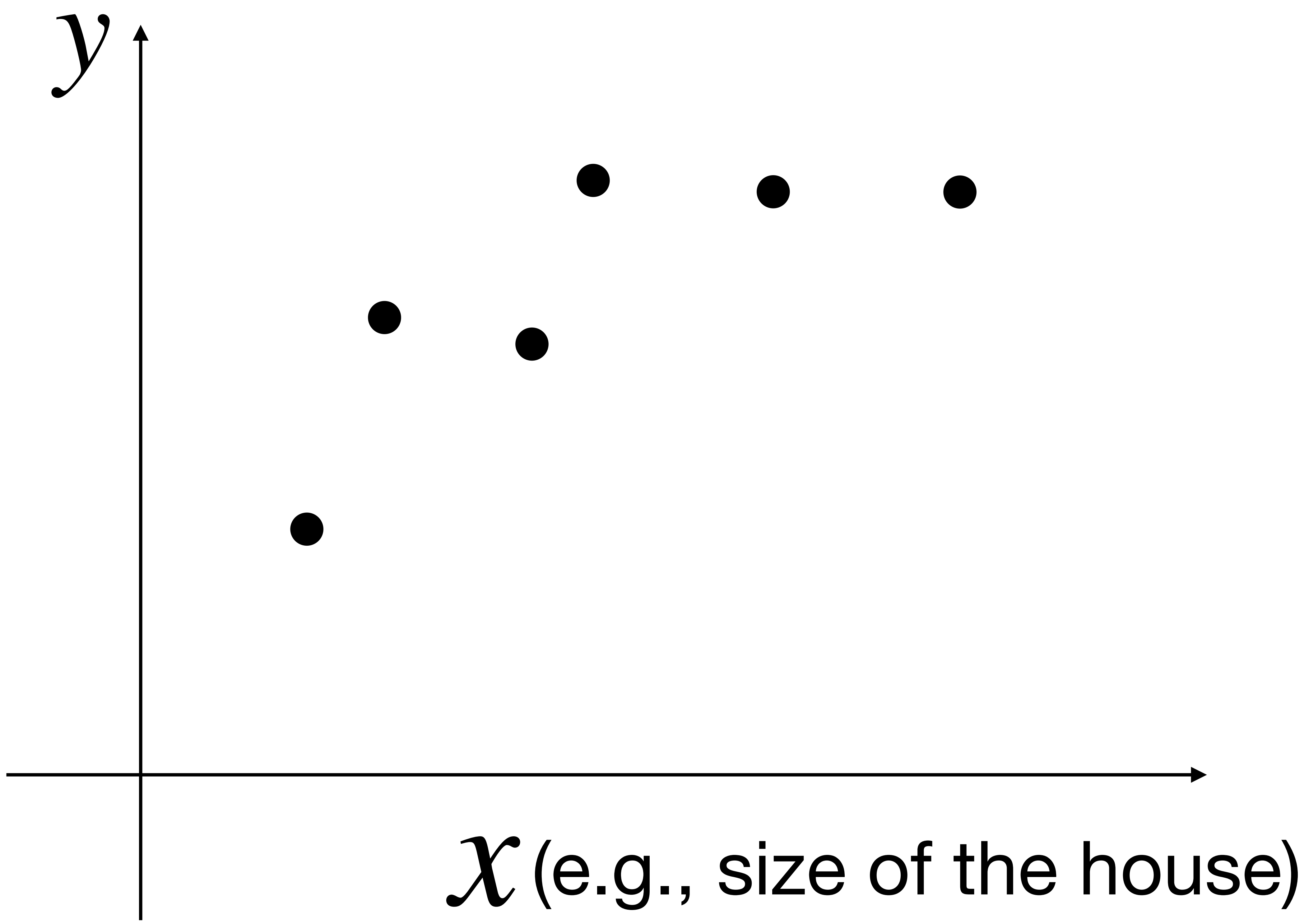
2. This causes underfitting, i.e., we cannot capture the trend in the data

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

# Overfitting

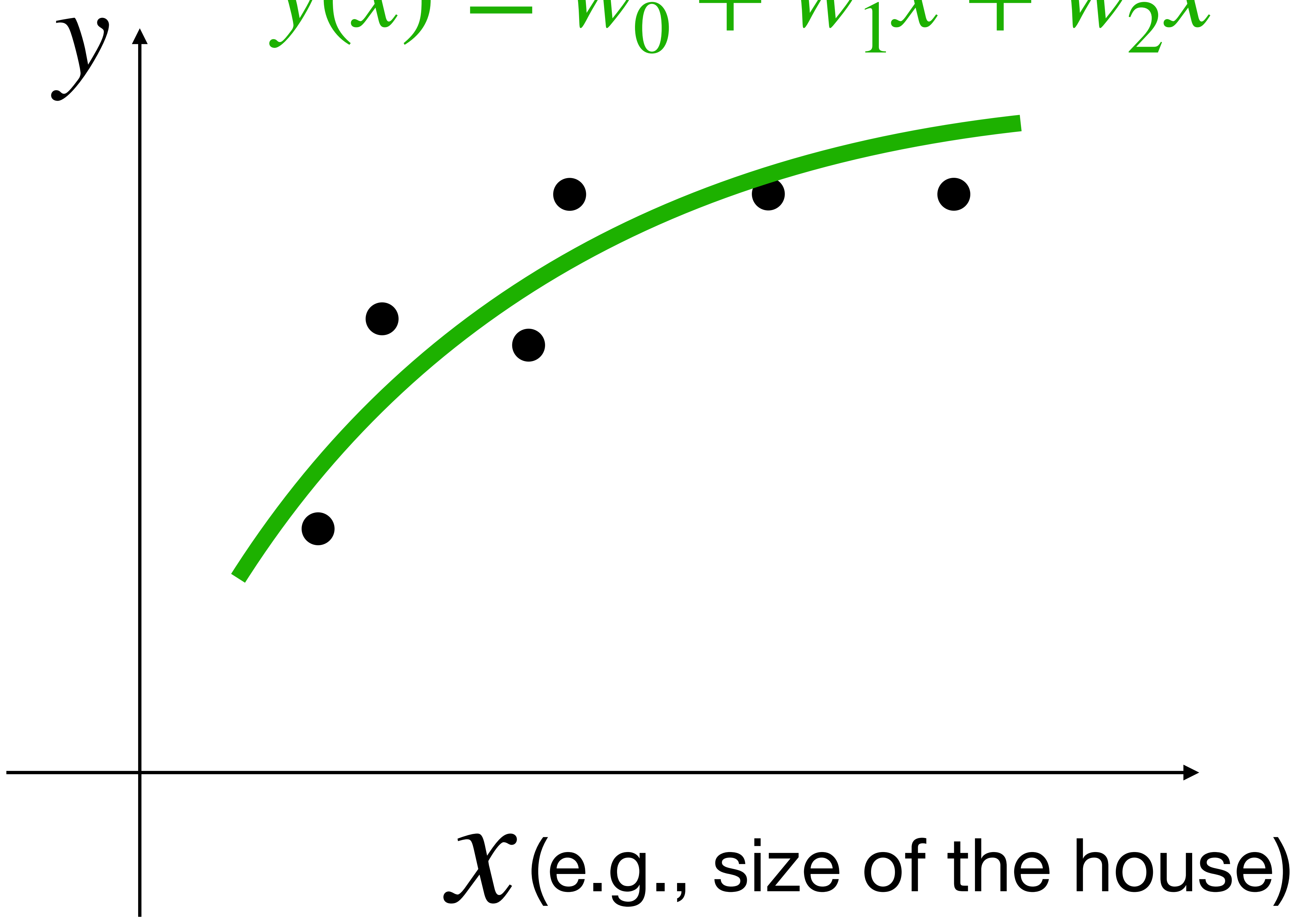



(Just right)



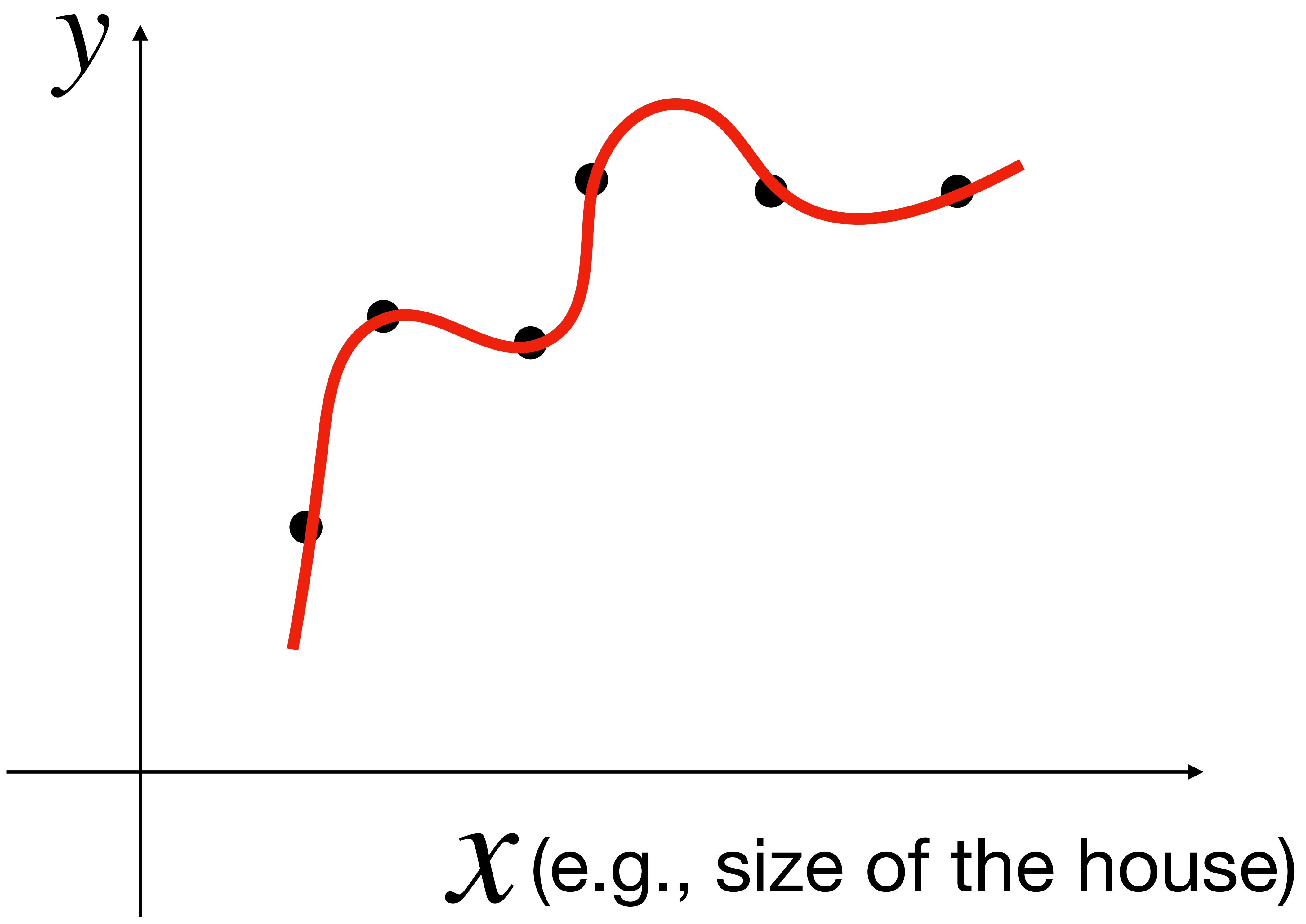
# Overfitting

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

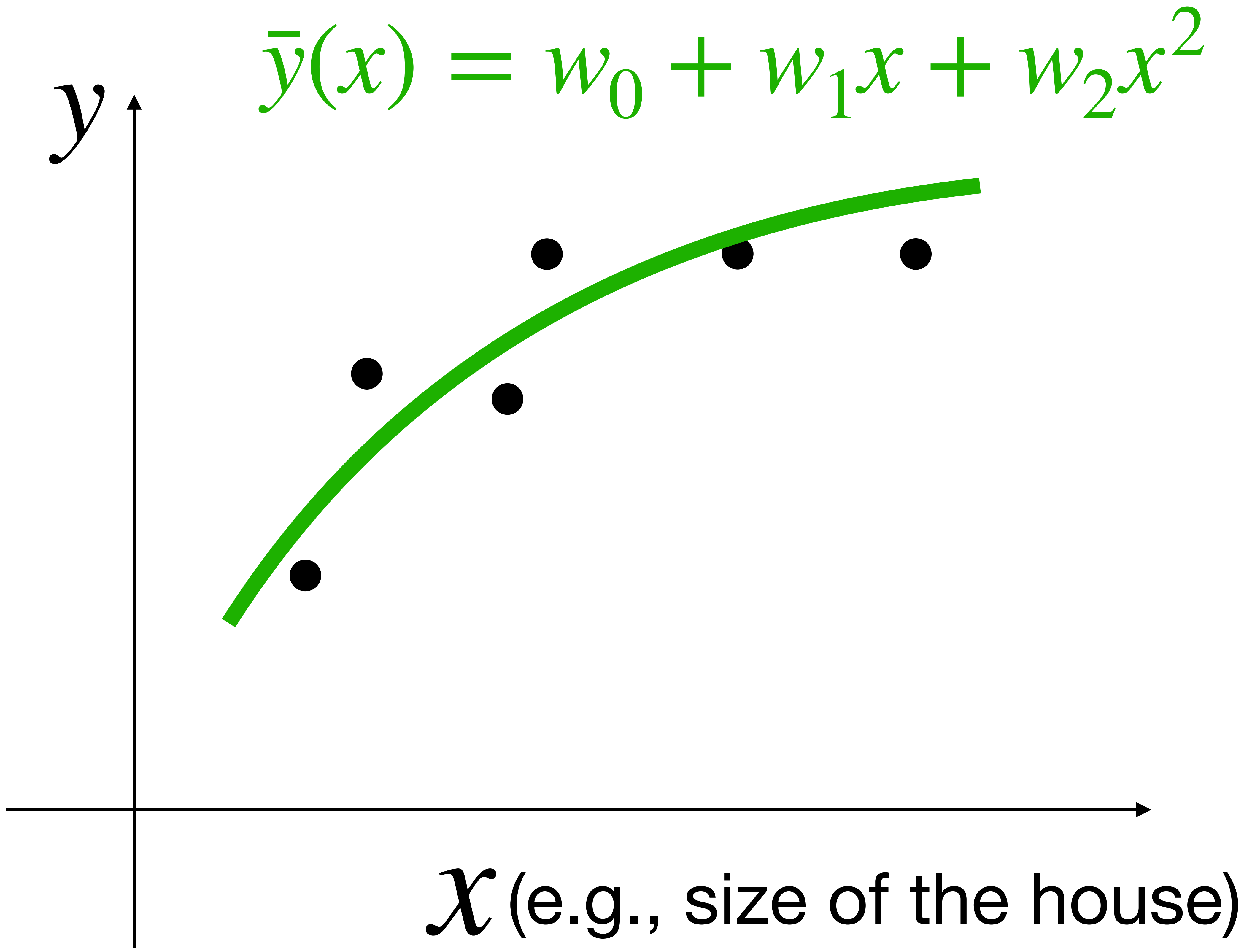



(Just right)

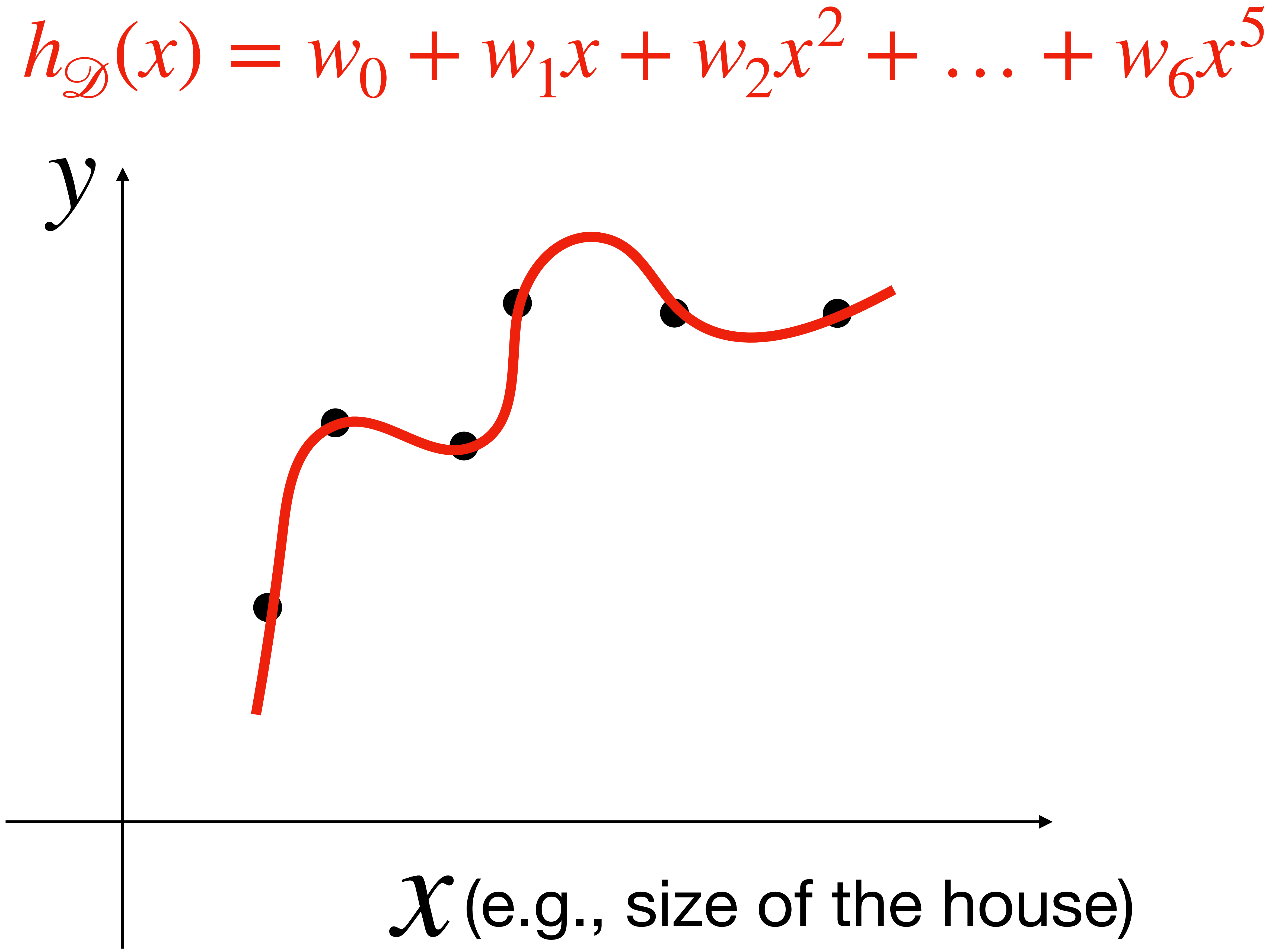
$$h_{\mathscr{D}}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_6 x^5$$



# Overfitting



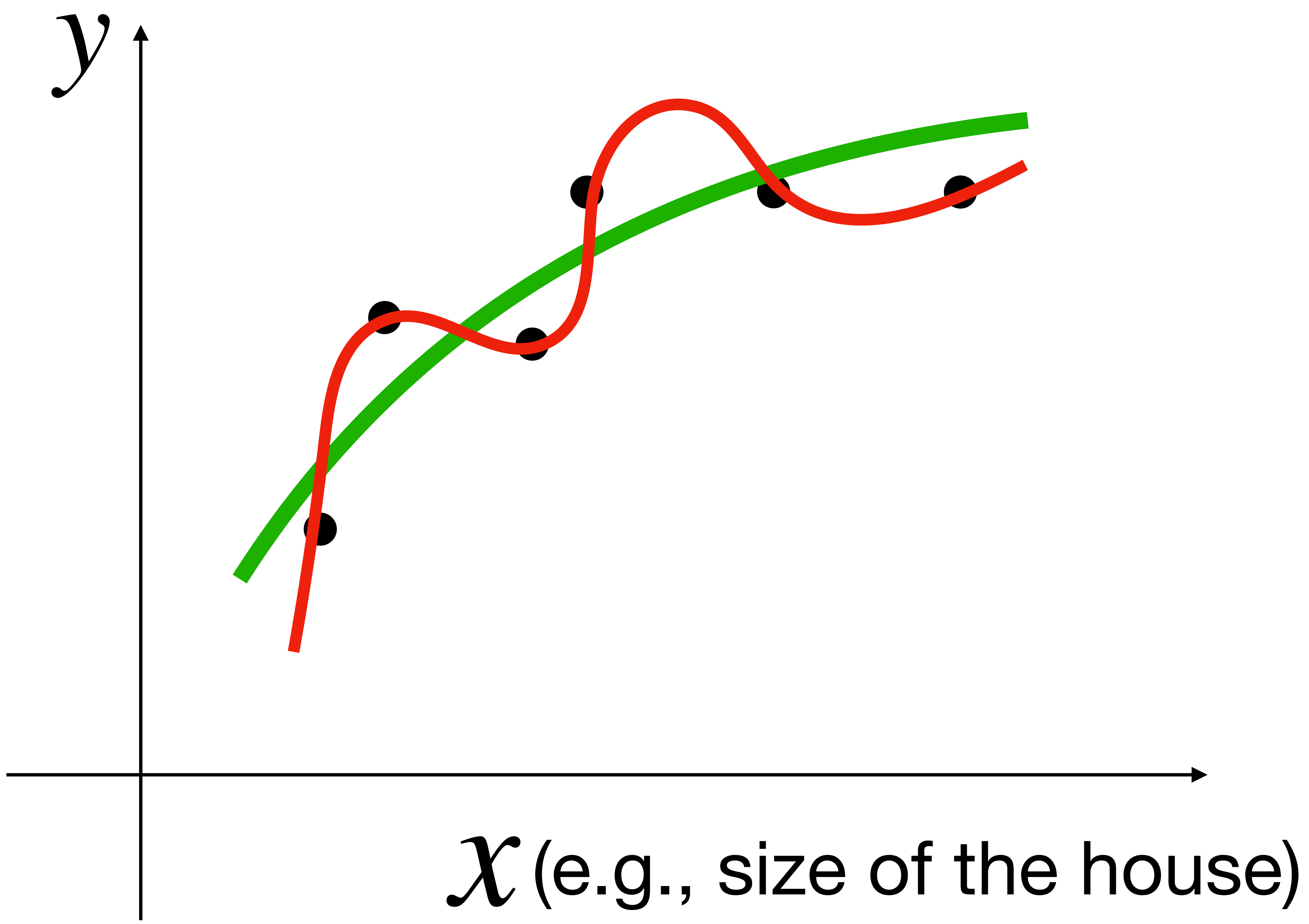

(Just right)




Overfitting

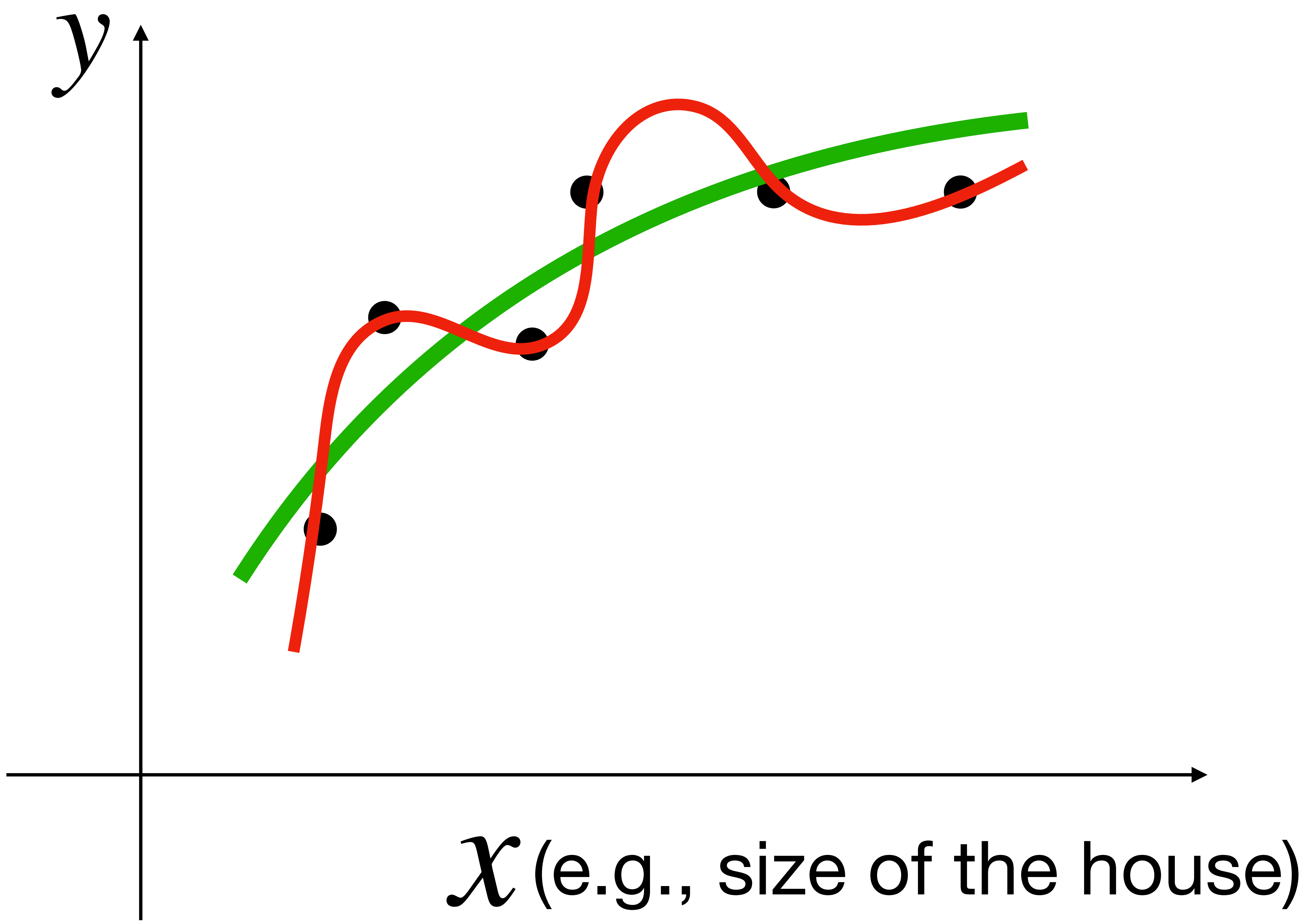
# Overfitting

Just right versus Overfitting



# Overfitting

Just right versus Overfitting

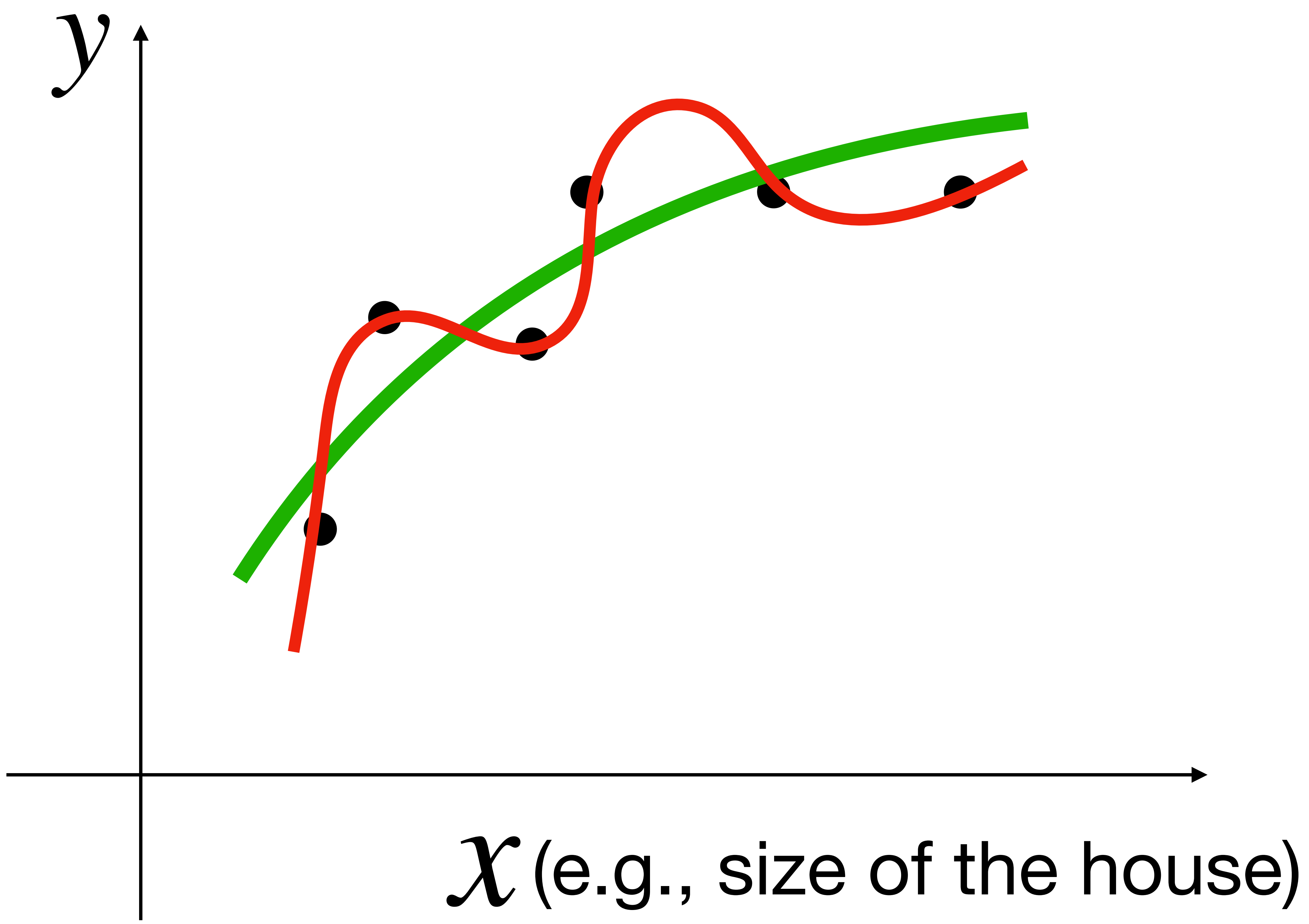



No strong bias:

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

# Overfitting

Just right versus Overfitting



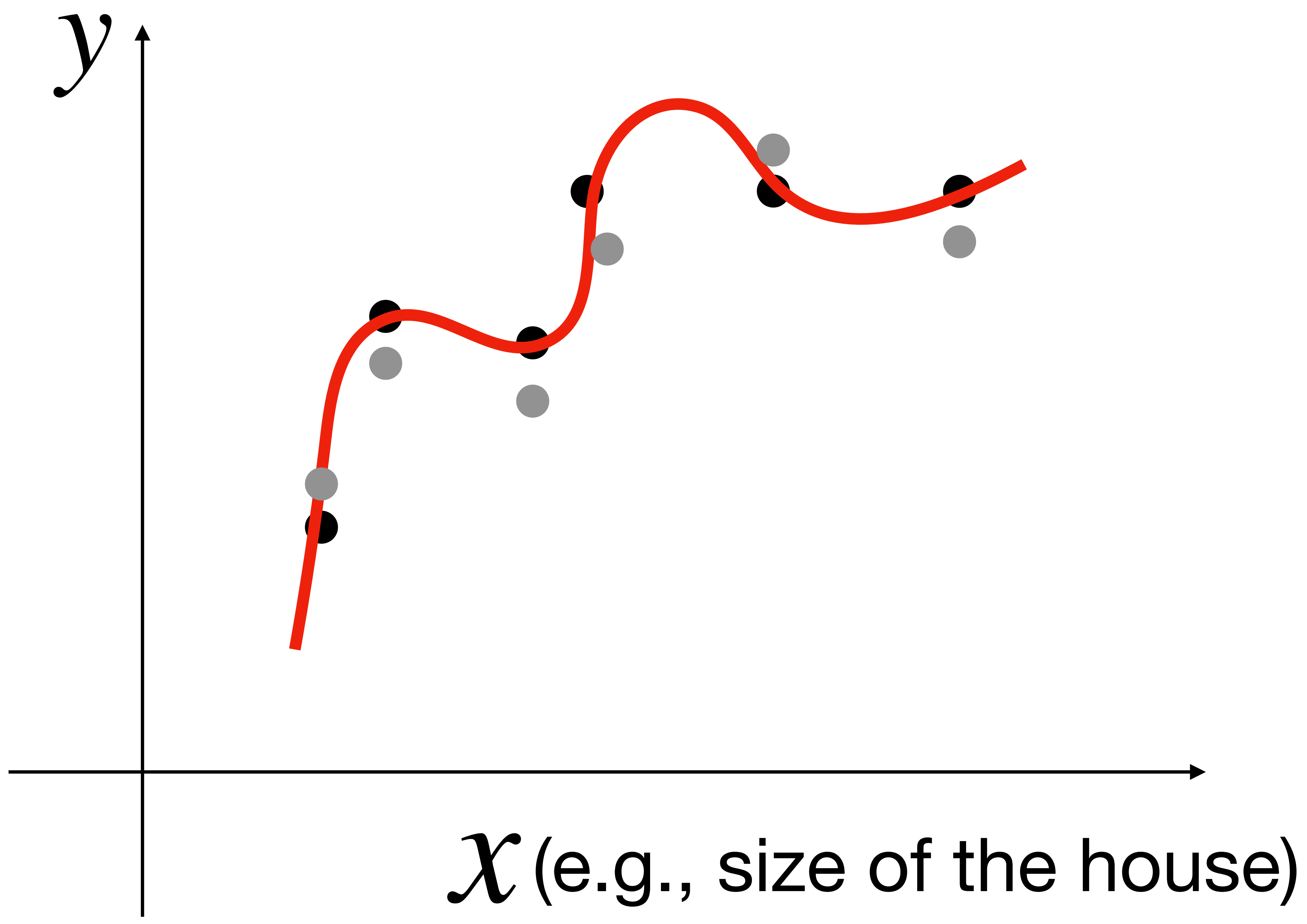

No strong bias:

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

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

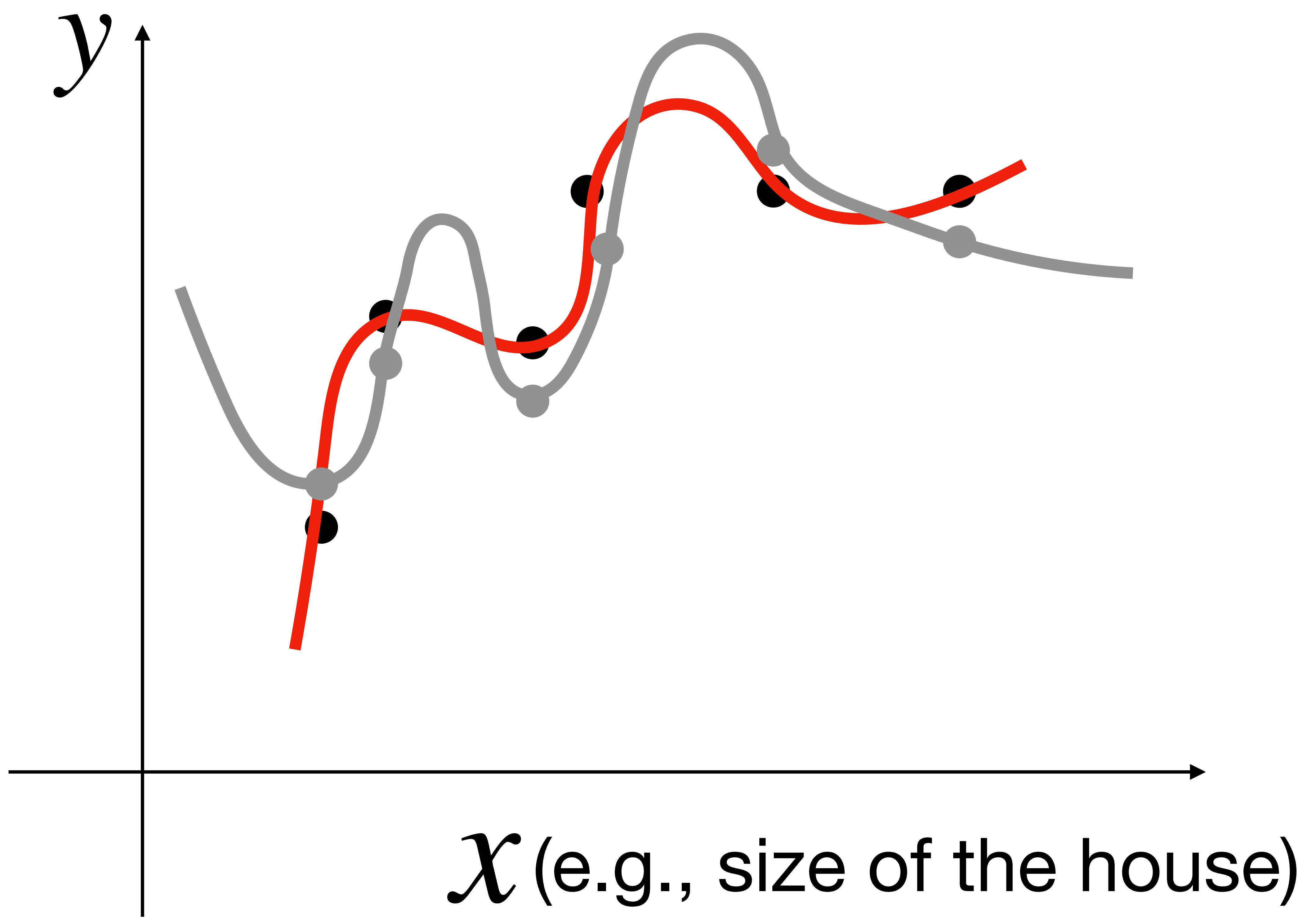
# Overfitting

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



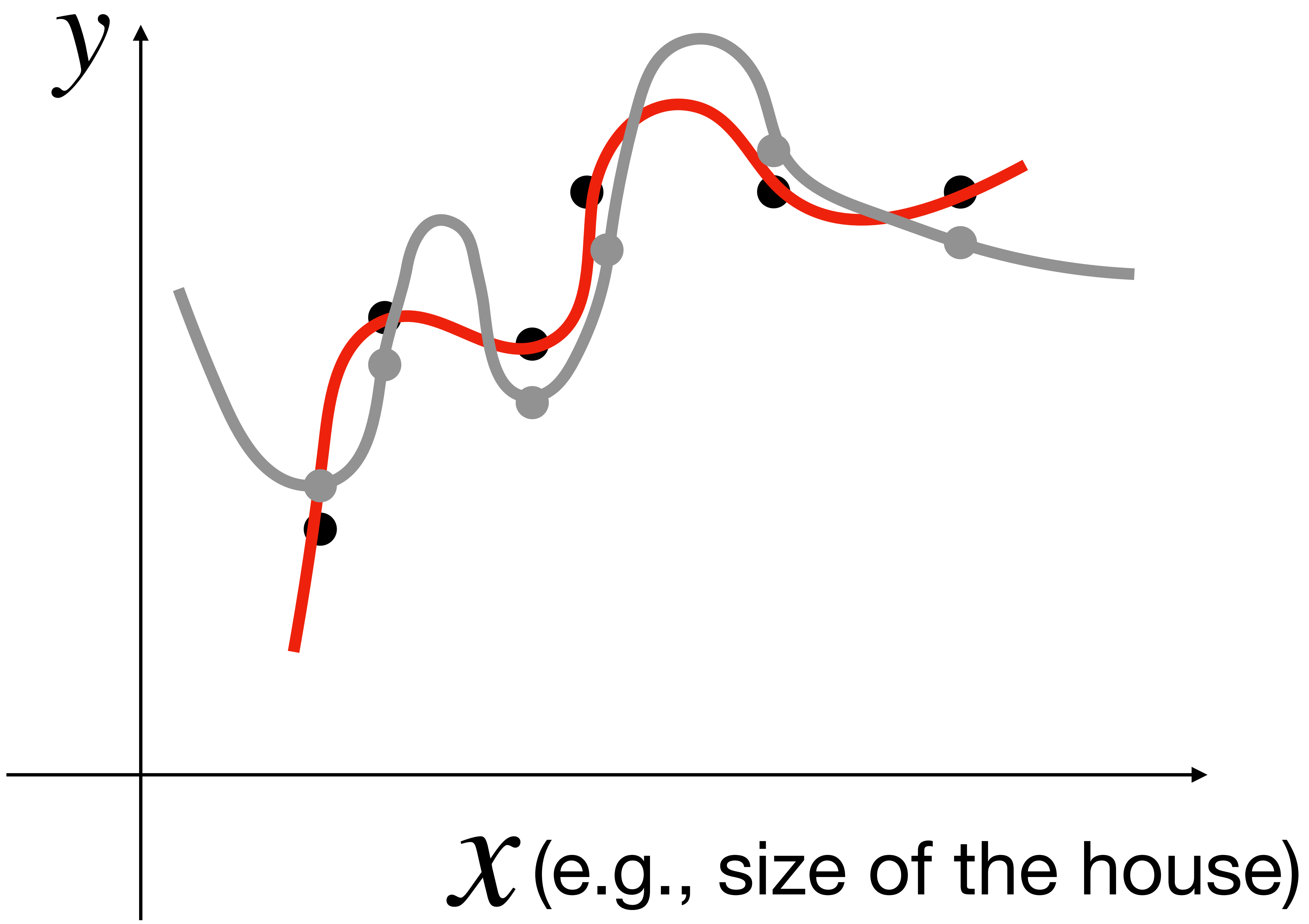
# Overfitting

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



# Overfitting

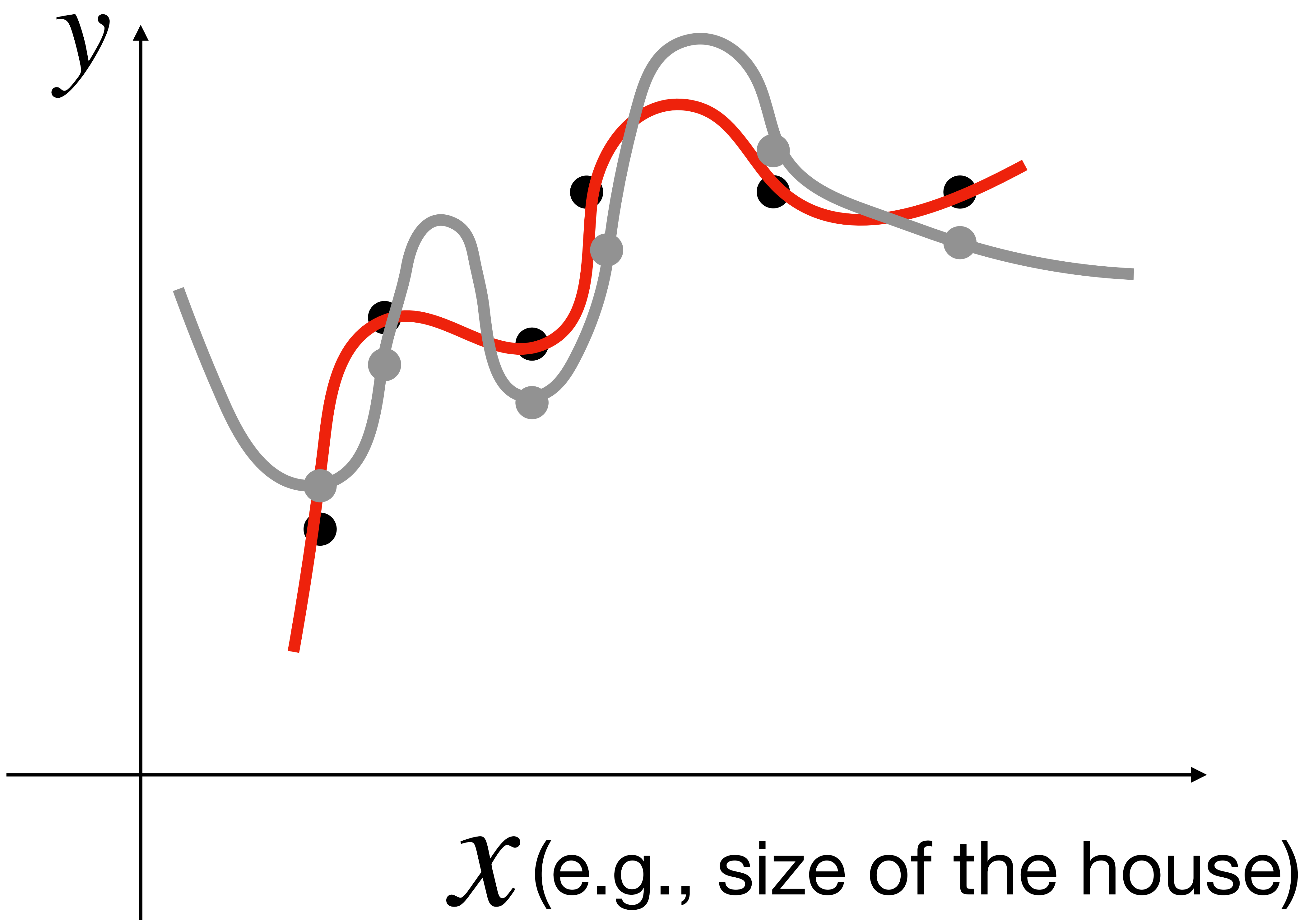
Redo the higher-order polynomial fitting on different dataset $\mathscr{D}'$




The new linear function does differ a lot from the old one

# Overfitting

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




The new linear function does 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

3. Example on Ridge Linear Regression

# 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

# 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 bound 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 bound 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:

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

# 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:

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

# 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:

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

# The expectation of our model $h_{\mathcal{D}}$

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

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

In other words, we have:

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

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

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

Bias: 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

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

Bias: 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: 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

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

Bias: 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 $h_{\mathcal{D}}$, i.e., $\mathbb{E}_{\mathcal{D}}\mathbb{E}_x\left(h_{\mathcal{D}}(x) - \bar{h}(x)\right)^2$

Fluctuation of our random model around its mean

# 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** + **Variance** + Noise (unavoidable, independent of Algs/models)

# 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** + **Variance** + Noise (unavoidable, independent of Algs/models)

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

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

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

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

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

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

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

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

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

This term is zero since:

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

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

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

This term is zero since:

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

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

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

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

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

This term is zero since:

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

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

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

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

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

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

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

Variance

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

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

Variance

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

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

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

Variance

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

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

Variance

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

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

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

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

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

Variance

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

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

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

This term is zero since:

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

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

Variance

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

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

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

This term is zero since:

$$= \mathbb{E}_{\mathscr{D}}\mathbb{E}_x(\bar{h}(x) - \bar{y}(x)) \cdot \mathbb{E}_{y|x}(\bar{y}(x) - y)$$

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

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

Variance

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

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

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

This term is zero since:

$$= \mathbb{E}_{\mathcal{D}}\mathbb{E}_x(\bar{h}(x) - \bar{y}(x)) \cdot \mathbb{E}_{y|x}(\bar{y}(x) - y)$$

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

## Putting the derivations together, we arrive at:

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

Variance

Bias

## 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 Noise

## Putting the derivations together, we arrive at:

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

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

# Outline of Today

1. Intro on Underfitting/Overfitting and Bias/Variance

2. Derivation of the Bias-Variance Decomposition

3. Example on Ridge Linear Regression

# Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)

# Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)

But $y_i \sim (w^\star)^\top x_i + \epsilon_i, \; \epsilon_i \sim \mathcal{N}(0,1)$

# Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)

But $y_i \sim (w^\star)^\top x_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0,1)$

(This is called LR w/ fixed design)

# Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)

But $y_i \sim (w^\star)^\top x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,1)$

(This is called LR w/ fixed design)

(So the only randomness of our dataset $\mathcal{D} = \{x_i, y_i\}$ is coming from the noises $\epsilon_i$)

# Ex: Ridge Linear regression

Ridge Linear Regression formulation

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

# Ex: Ridge Linear regression

Ridge Linear Regression formulation

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

**What we will show now:**

Larger $\lambda$ (model becomes "simpler") => larger bias, but smaller variance

# Ex: Ridge Linear regression

Ridge Linear Regression formulation

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

**What we will show now:**

Larger $\lambda$ (model becomes "simpler") => larger bias, but smaller variance

(Q: think about the case where $\lambda \to \infty$, what happens to $\hat{w}$?)

# Ex: Ridge Linear regression

Denote $X = [x_1, \ldots, x_n], Y = [y_1, \ldots, y_n]^\top, \epsilon = [\epsilon_1, \ldots, \epsilon_n]^\top$

# Ex: Ridge Linear regression

Denote $X = [x_1, \ldots, x_n], Y = [y_1, \ldots, y_n]^\top, \epsilon = [\epsilon_1, \ldots, \epsilon_n]^\top$

$$\hat{w} = \arg\min_w \|X^\top w - Y\|_2^2 + \lambda \|w\|_2^2$$

# Ex: Ridge Linear regression

Denote $X = [x_1, \ldots, x_n], Y = [y_1, \ldots, y_n]^\top, \epsilon = [\epsilon_1, \ldots, \epsilon_n]^\top$

$$\hat{w} = \arg\min_w \|X^\top w - Y\|_2^2 + \lambda \|w\|_2^2$$

Since $y_i = (w^\star)^\top x_i + \epsilon_i$ we have $Y = X^\top w^\star + \epsilon$

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY \; = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

Source of the randomness of $\hat{w}$

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

Source of the randomness of $\hat{w}$

**Let us compute $\mathbb{E}_\epsilon[\hat{w}]$:**

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

Source of the randomness of $\hat{w}$

**Let us compute $\mathbb{E}_\epsilon[\hat{w}]$:**

$$\mathbb{E}_\epsilon[\hat{w}] = (XX^\top + \lambda I)^{-1} X[X^\top w^\star + \mathbb{E}_\epsilon[\epsilon]]$$

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY \; = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

Source of the randomness of $\hat{w}$

**Let us compute $\mathbb{E}_\epsilon[\hat{w}]$:**

$$\mathbb{E}_\epsilon[\hat{w}] = (XX^\top + \lambda I)^{-1} X[X^\top w^\star + \mathbb{E}_\epsilon[\epsilon]]$$

$$= (XX^\top + \lambda I)^{-1} XX^\top w^\star$$

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY \; = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

Source of the randomness of $\hat{w}$

**Let us compute $\mathbb{E}_\epsilon[\hat{w}]$:**

$$\mathbb{E}_\epsilon[\hat{w}] = (XX^\top + \lambda I)^{-1} X[X^\top w^\star + \mathbb{E}_\epsilon[\epsilon]]$$

$$= (XX^\top + \lambda I)^{-1} XX^\top w^\star$$

$$= (XX^\top + \lambda I)^{-1} (XX^\top + \lambda I - \lambda I) w^\star$$

# Ex: Ridge Linear regression

Recall we have closed form solution for Ridge LR

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY = (XX^\top + \lambda I)^{-1} X(X^\top w^\star + \epsilon)$$

Source of the randomness of $\hat{w}$

**Let us compute $\mathbb{E}_\epsilon[\hat{w}]$:**

$$\mathbb{E}_\epsilon[\hat{w}] = (XX^\top + \lambda I)^{-1} X[X^\top w^\star + \mathbb{E}_\epsilon[\epsilon]]$$

$$= (XX^\top + \lambda I)^{-1} XX^\top w^\star$$

$$= (XX^\top + \lambda I)^{-1}(XX^\top + \lambda I - \lambda I) w^\star = w^\star - \lambda (XX^\top + \lambda I)^{-1} w^\star$$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - \lambda(XX^\top + \lambda)^{-1}\lambda w^\star$$

Bias term: $\displaystyle\sum_{i=1}^{n} \left((\mathbb{E}[\hat{w}] - w^\star)^\top x_i\right)^2$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - \lambda(XX^\top + \lambda)^{-1}\lambda w^\star$$

Bias term: $\sum_{i=1}^{n} \left( (\mathbb{E}[\hat{w}] - w^\star)^\top x_i \right)^2$

$$= \sum_{i=1}^{n} \left( (\lambda(XX^\top + \lambda)^{-1} w^\star)^\top x_i \right)^2$$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - \lambda(XX^\top + \lambda)^{-1}\lambda w^\star$$

Bias term: $\sum_{i=1}^{n} \left( (\mathbb{E}[\hat{w}] - w^\star)^\top x_i \right)^2$

$$= \sum_{i=1}^{n} \left( (\lambda(XX^\top + \lambda)^{-1} w^\star)^\top x_i \right)^2$$

$$= \lambda^2 (w^\star)^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} w^\star$$

# Ex: Ridge Linear regression

$$\text{Bias } = \lambda^2 (w^\star)^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} w^\star$$

# Ex: Ridge Linear regression

Bias $= \lambda^2 (w^\star)^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} w^\star$

Eigendecomposition on $XX^\top = U \Sigma U^\top$

# Ex: Ridge Linear regression

$$\text{Bias } = \lambda^2 (w^\star)^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} w^\star$$

Eigendecomposition on $XX^\top = U \Sigma U^\top$

$$= (w^\star)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1/\lambda + 1)^2} & 0 & 0\ldots \\ 0 & \frac{\sigma_2}{(\sigma_2/\lambda + 1)^2} & 0\ldots \\ \ldots & \ldots & \ldots \\ 0, & \ldots & \frac{\sigma_d}{(\sigma_d/\lambda + 1)^2} \end{bmatrix} U^\top w^\star$$

# Ex: Ridge Linear regression

$$\text{Bias } = \lambda^2 (w^\star)^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} w^\star$$

Eigendecomposition on $XX^\top = U\Sigma U^\top$

$$= (w^\star)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1/\lambda + 1)^2} & 0 & 0\ldots \\ 0 & \frac{\sigma_2}{(\sigma_2/\lambda + 1)^2} & 0\ldots \\ \ldots & \ldots & \ldots \\ 0, & \ldots & \frac{\sigma_d}{(\sigma_d/\lambda + 1)^2} \end{bmatrix} U^\top w^\star$$

Q: how does bias behave when $\lambda \to +\infty$

# Ex: Ridge Linear regression

$$\text{Bias } = \lambda^2 (w^\star)^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} w^\star$$

Eigendecomposition on $XX^\top = U \Sigma U^\top$

$$= (w^\star)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1/\lambda + 1)^2} & 0 & 0\ldots \\ 0 & \frac{\sigma_2}{(\sigma_2/\lambda + 1)^2} & 0\ldots \\ \ldots & \ldots & \ldots \\ 0, & \ldots & \frac{\sigma_d}{(\sigma_d/\lambda + 1)^2} \end{bmatrix} U^\top w^\star$$

Q: how does bias behave when $\lambda \to +\infty$

Q: how does bias behave when $\lambda \to 0$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - (XX^\top + \lambda)^{-1}\lambda w^\star$$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - (XX^\top + \lambda)^{-1}\lambda w^\star$$

Variance term: $\sum_{i=1}^{n} \mathbb{E}(\hat{w}^\top x_i - \mathbb{E}[\hat{w}]^\top x_i)^2$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - (XX^\top + \lambda)^{-1}\lambda w^\star$$

Variance term: $\sum_{i=1}^{n} \mathbb{E}(\hat{w}^\top x_i - \mathbb{E}[\hat{w}]^\top x_i)^2$

$$= \sum_{i=1}^{d} \sigma_i^2/(\sigma_i + \lambda)^2$$

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - (XX^\top + \lambda)^{-1}\lambda w^\star$$

Variance term: $\sum_{i=1}^{n} \mathbb{E}(\hat{w}^\top x_i - \mathbb{E}[\hat{w}]^\top x_i)^2$

$$= \sum_{i=1}^{d} \sigma_i^2/(\sigma_i + \lambda)^2$$

(Optional — tedious but basic computation, see note)

# Ex: Ridge Linear regression

$$\mathbb{E}[\hat{w}] = w^\star - (XX^\top + \lambda)^{-1}\lambda w^\star$$

Variance term: $\sum_{i=1}^{n} \mathbb{E}(\hat{w}^\top x_i - \mathbb{E}[\hat{w}]^\top x_i)^2$

$$= \sum_{i=1}^{d} \sigma_i^2 / (\sigma_i + \lambda)^2$$

(Optional — tedious but basic computation, see note)

Q: how does Var behave when $\lambda \to +\infty$

Q: how does Var behave when $\lambda \to 0$

# Ex: Ridge Linear regression

In summary, for Ridge LR:

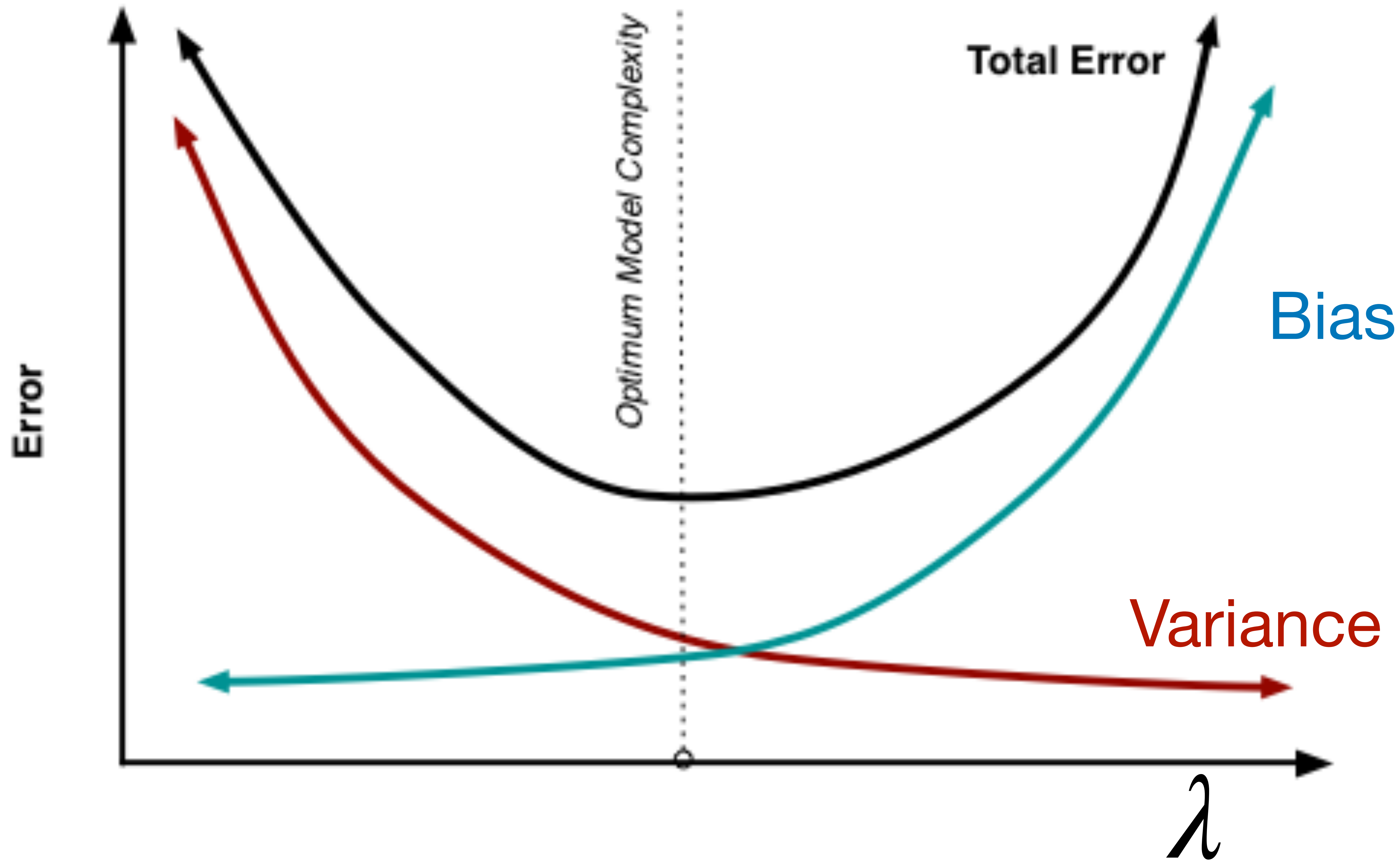
Smaller regularization penalty $\lambda$ => smaller bias, but larger variance

Larger regularization penalty $\lambda$ => larger bias, but smaller variance

# Ex: Ridge Linear regression

Tuning $\lambda$ allows us to control the generalization error of Ridge LR solution:

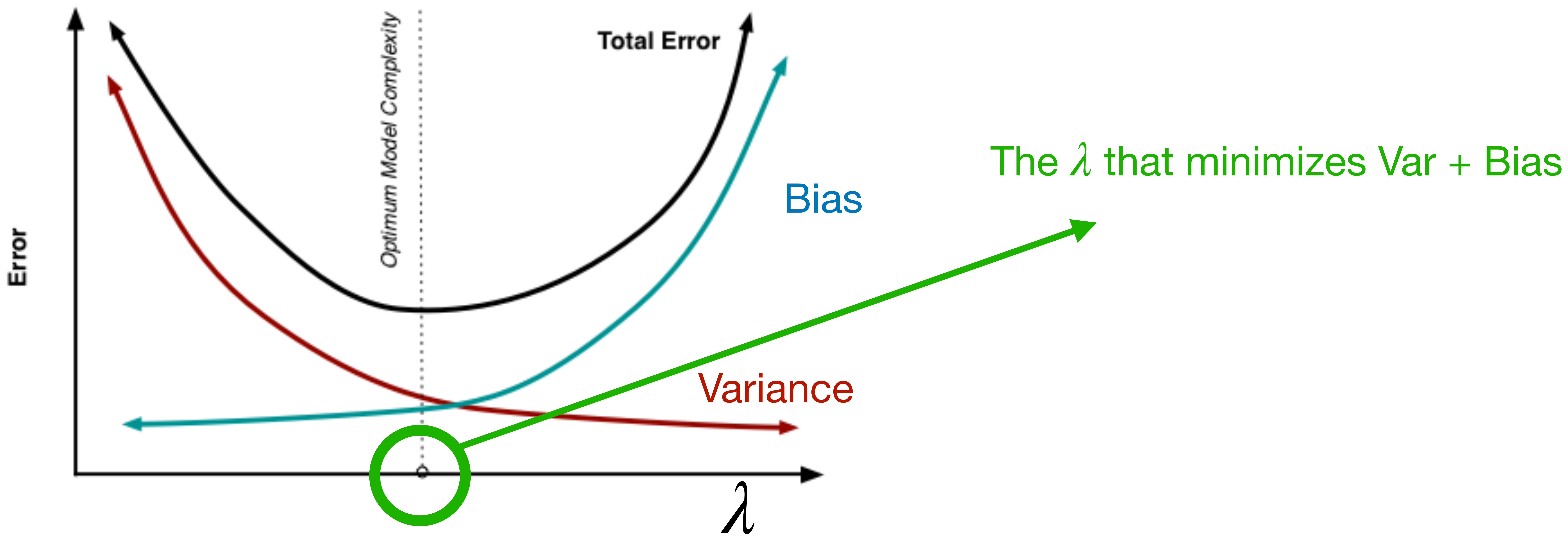
$$\mathbb{E}(\hat{w}^\top x - y)^2 = \text{Variance} + \text{Bias} + \text{Inherent noise}$$



# Ex: Ridge Linear regression

Tuning $\lambda$ allows us to control the generalization error of Ridge LR solution:

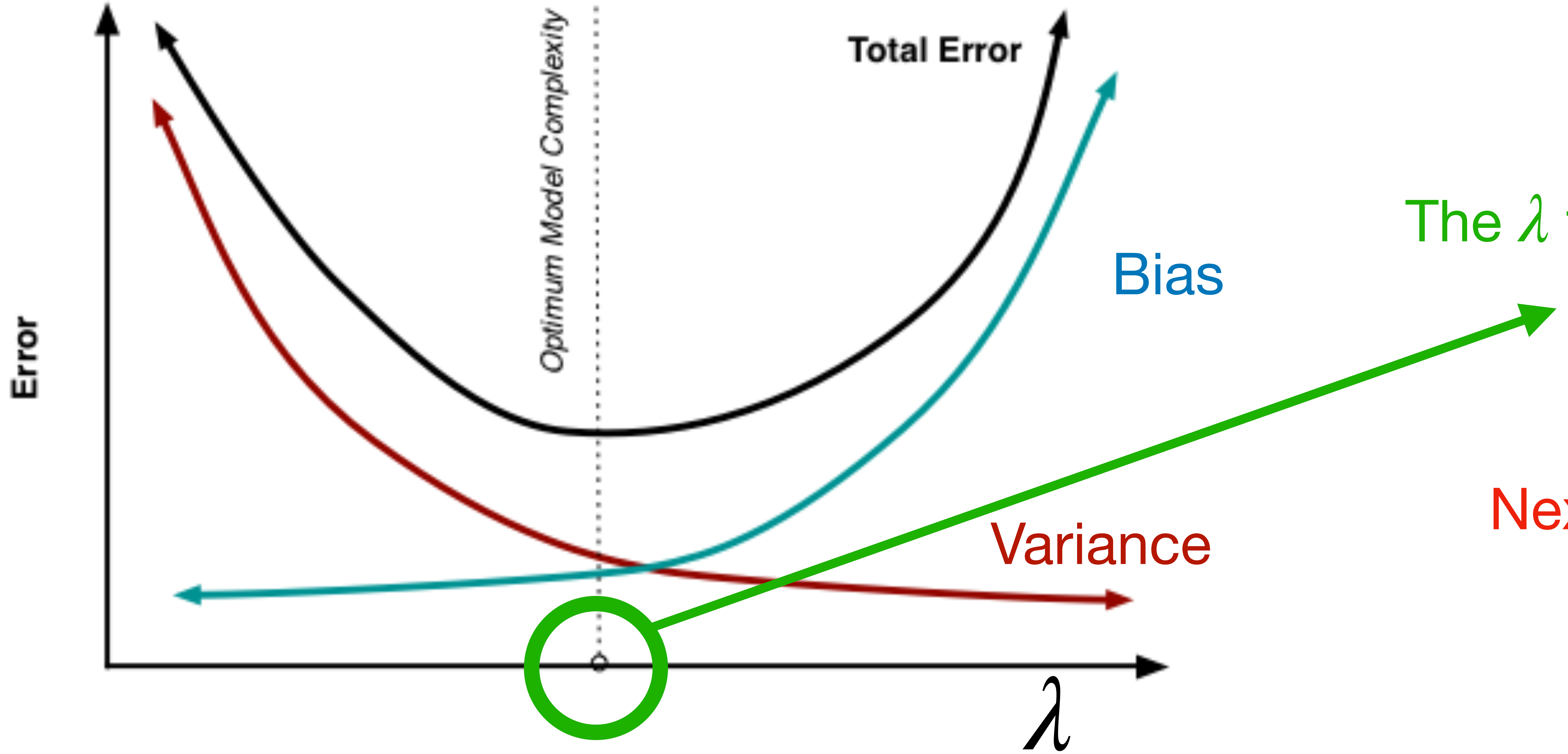
$$\mathbb{E}(\hat{w}^\top x - y)^2 = \text{Variance} + \text{Bias} + \text{Inherent noise}$$



# Ex: Ridge Linear regression

Tuning $\lambda$ allows us to control the generalization error of Ridge LR solution:

$$\mathbb{E}(\hat{w}^\top x - y)^2 = \text{Variance} + \text{Bias} + \text{Inherent noise}$$




The $\lambda$ that minimizes Var + Bias

Next lecture: how to select that in practice