

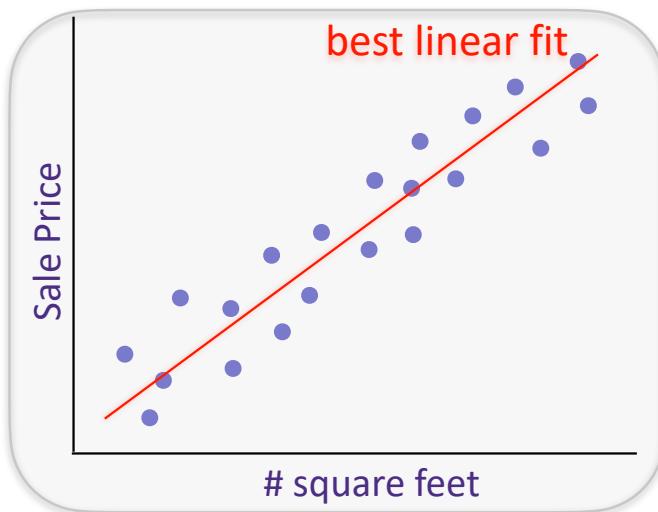
Generalized Linear Regression and Bias-Variance Tradeoffs

The regression problem

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

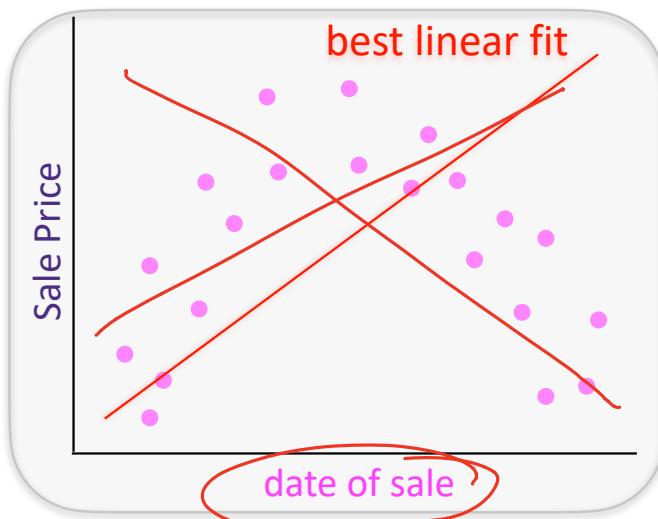
$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

The regression problem

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

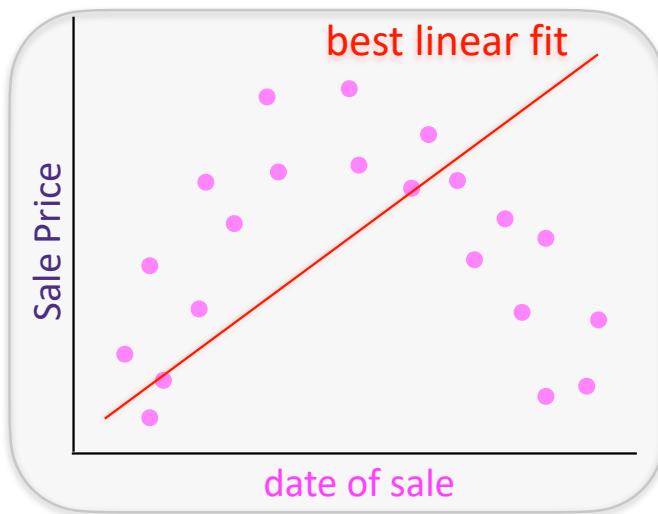


The regression problem

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Best linear model of data of sale is a very poor fit!

Either because date of sale doesn't predict price well, or...

... because the relationship isn't linear.

Process

Decide on a model → Quadratic fn, polynomial of degree p/..

Find the function which fits the data best

Choose a loss function → least squares

Pick the function which minimizes loss
on data

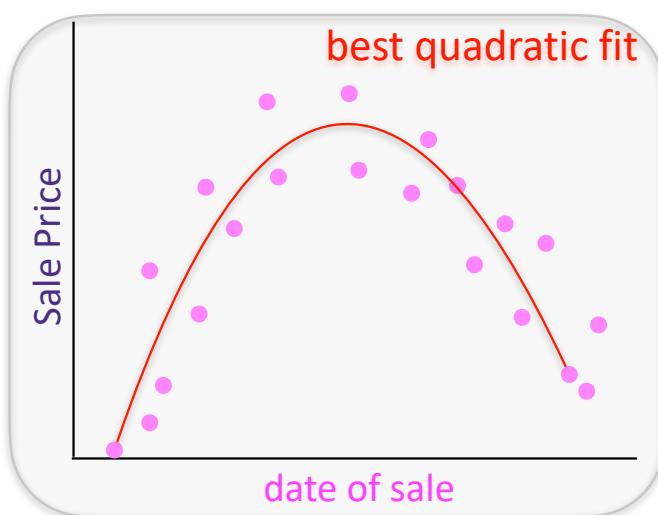
Use function to make prediction on new examples

Quadratic Regression

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

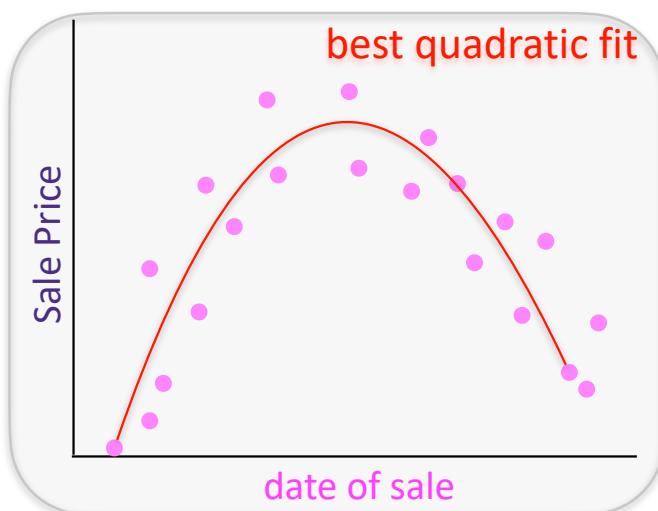
$$y_i \approx \sum_{j=1}^d x_{i,j} w_{j,1} + x_{i,j}^2 w_{j,2}$$

Quadratic Regression

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n \quad y_i \in \mathbb{R}$

Hypothesis: quadratic *(Model)*

$$y_i \approx \sum_{j=1}^d x_{i,j} w_{j,1} + x_{i,j}^2 w_{j,2}$$

i. satırın j. feature'si

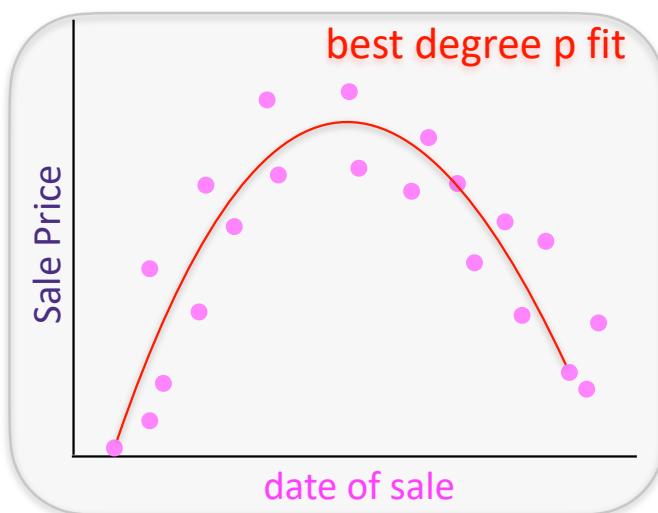
i. evin j. özelliği

Polynomial regression

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

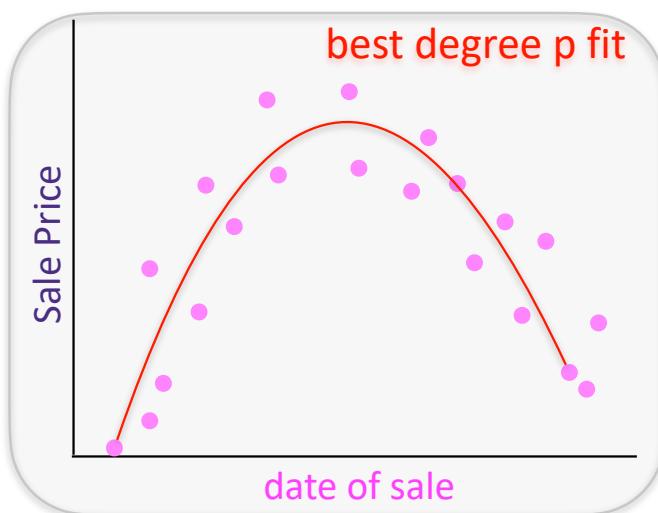
$$y_i \approx \sum_{j=1}^d \sum_{\ell=1}^p x_{i,j}^\ell w_{j,\ell}$$

Polynomial regression

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

Hypothesis: *Model*
degree p polynomial

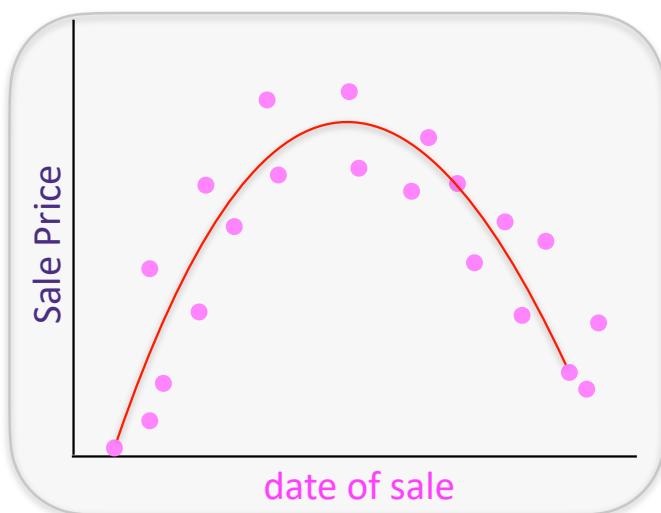
$$y_i \approx \sum_{j=1}^d \sum_{\ell=1}^p x_{i,j}^\ell w_{j,\ell}$$

Generalized linear regression

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

$$y_i \approx \sum_{\ell=1}^p h_\ell(x_i)^T w_\ell$$

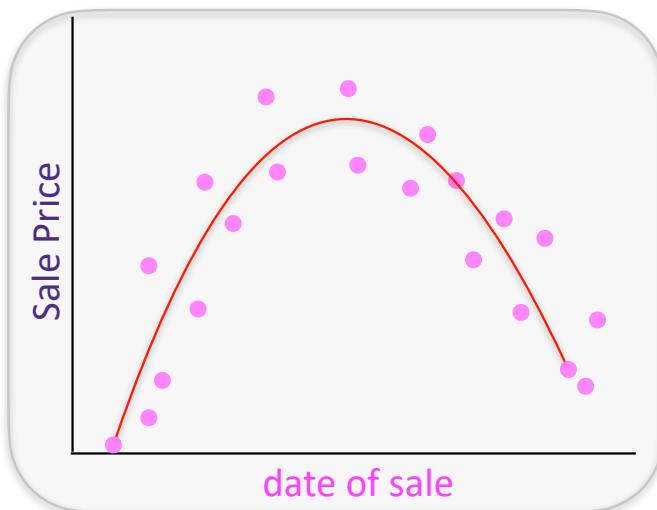
i. eve hl uygulaniyor

Generalized linear regression

Given past sales data on [zillow.com](#), predict:

y = House sale price from

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $x_i \in \mathbb{R}^d$
 $\{(x_i, y_i)\}_{i=1}^n$ $y_i \in \mathbb{R}$

Hypothesis:

generalized linear fn of x

$$y_i \approx \sum_{\ell=1}^p h_\ell(x_i)^T w_\ell$$

feature vector
for house
 i

Process

Decide on a model →

Find the function which fits the data best

Choose a loss function → Least squares

Pick the function which minimizes loss
on data

Use function to make prediction on new
examples

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

in $d=1$:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$$

for $d > 1$, generate

The regression problem

1 feature var. ve iliski linear degil.

$h: \mathbb{R}^d \rightarrow \mathbb{R}^p$

burada x 'in p ye kadar kuvvetini aliyor (mesela $p=2$ ise x, x^2)

ve bu sayede non-linear iliski yakalanmaya calisiliyor.

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n$$

$$\begin{aligned} x_i &\in \mathbb{R}^d \\ y_i &\in \mathbb{R} \end{aligned}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$h: \mathbb{R}^d \rightarrow \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

in $d=1$:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$$

for $d>1$, generate

$$\{u_j\}_{j=1}^p \subset \mathbb{R}^d$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

in d=1: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$

for d>1, generate

$$\{u_j\}_{j=1}^p \subset \mathbb{R}^d$$

$$h_j(x) = \frac{1}{1 + \exp(u_j^T x)}$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

in d=1: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$

for d>1, generate

$$\{u_j\}_{j=1}^p \subset \mathbb{R}^d$$
$$h_j(x) = (u_j^T x)^2$$
$$h_j(x) = \frac{1}{1 + \exp(u_j^T x)}$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

in d=1: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$

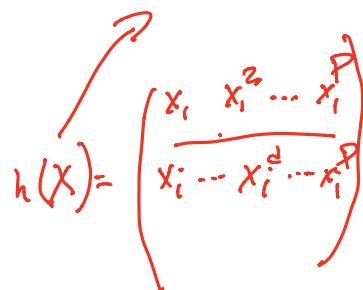
for d>1, generate

$$\{u_j\}_{j=1}^p \subset \mathbb{R}^d$$

$$h_j(x) = (u_j^T x)^2$$

$$h_j(x) = \frac{1}{1 + \exp(u_j^T x)}$$

$$h_j(x) = \cos(u_j^T x)$$


$$h(x) = \begin{pmatrix} x_1 & x_1^2 & \dots & x_1^p \\ x_2 & \dots & x_2^2 & \dots & x_2^p \\ \vdots & & \ddots & & \vdots \\ x_d & \dots & x_d^2 & \dots & x_d^p \end{pmatrix}$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Transformed data:
$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

~~Hypothesis: linear~~

$$y_i \approx x_i^T w$$

~~Loss: least squares~~

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$



Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

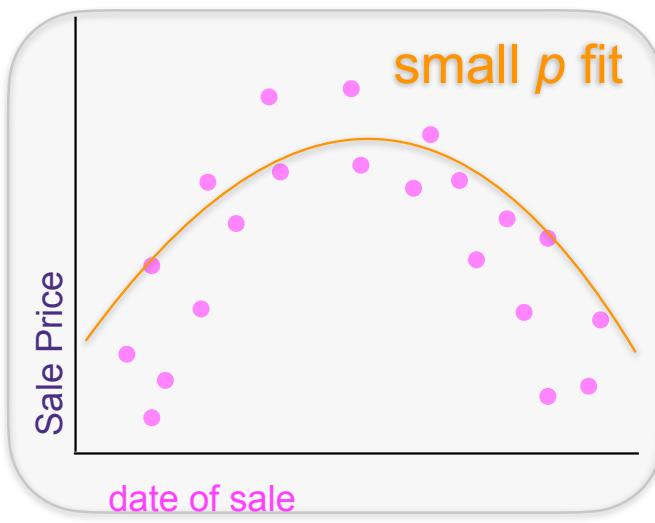
$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$



Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

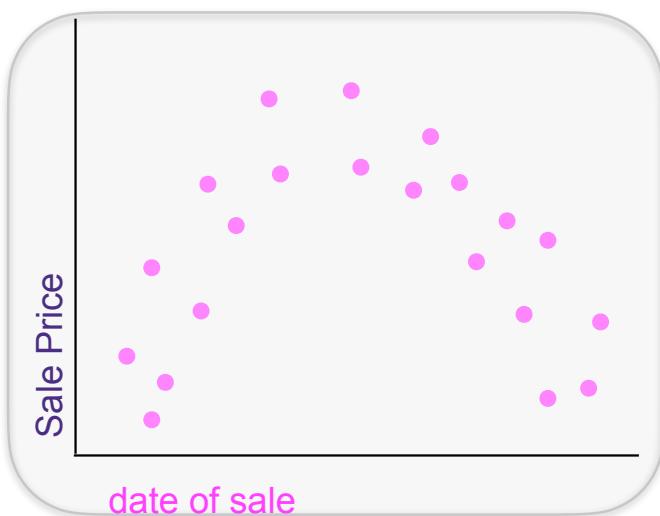
$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$



Transformed data:
$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$



Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

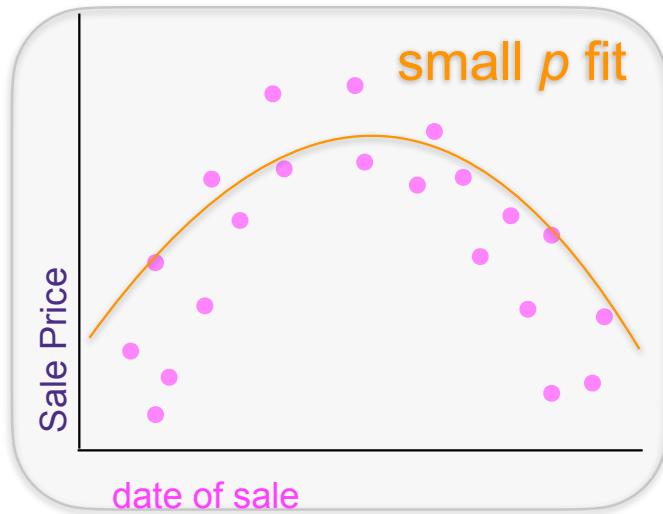
$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

Which is better?

A: large p



B: small p

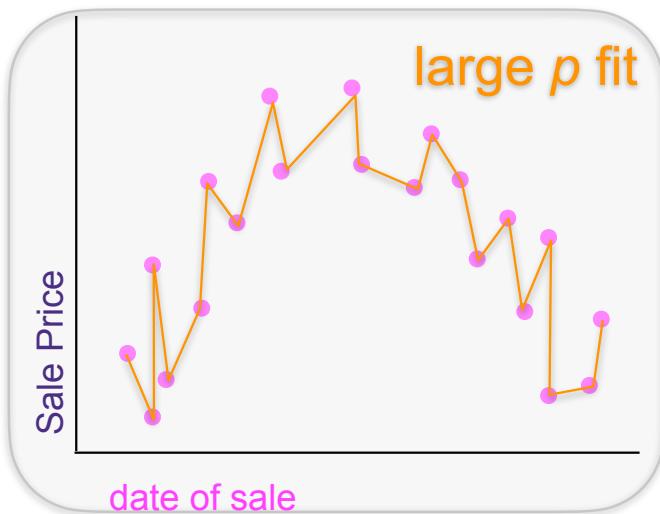


Bias-Variance Tradeoff

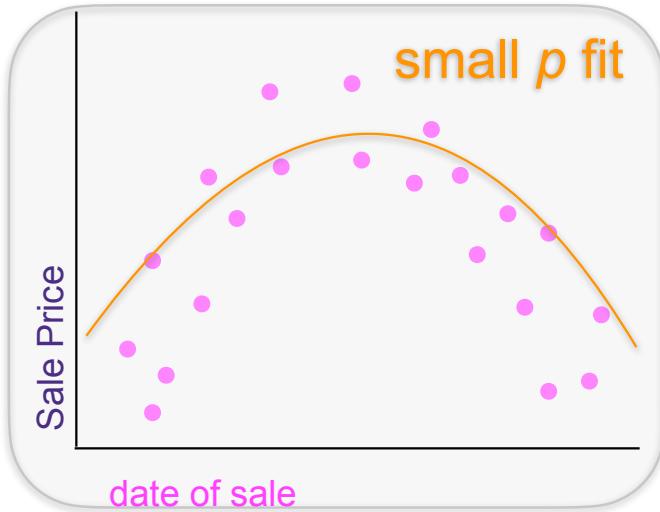
What is our goal, anyway?

What do these two graphs imply?

A: large p

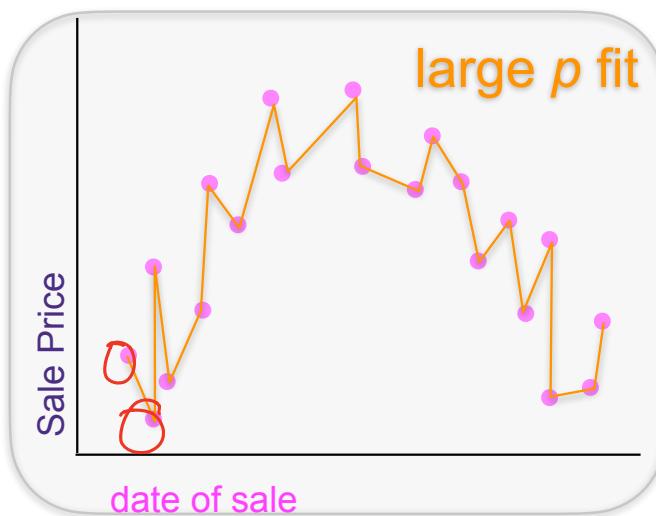


B: small p

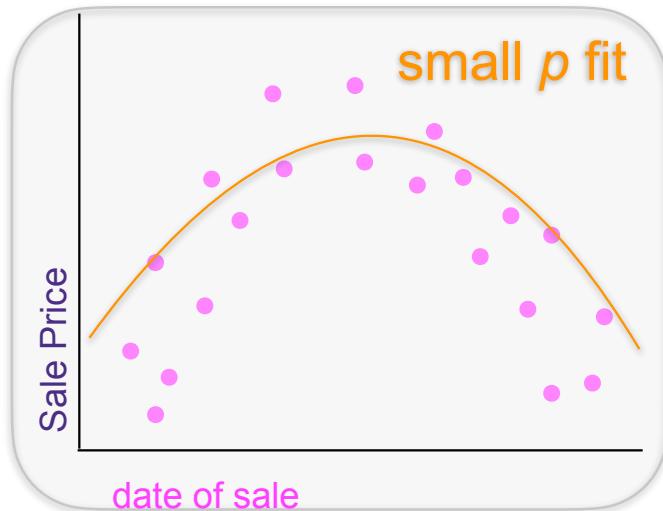


What do these two graphs imply?

A: large p



B: small p



Least squares loss is 0 on training data

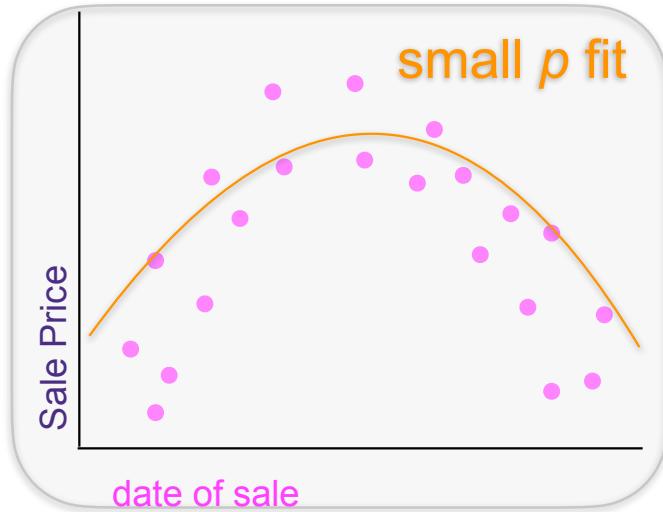
What do these two graphs imply?

A: large p



Least squares loss is 0 on training data

B: small p



Least squares loss is > 0 on training data

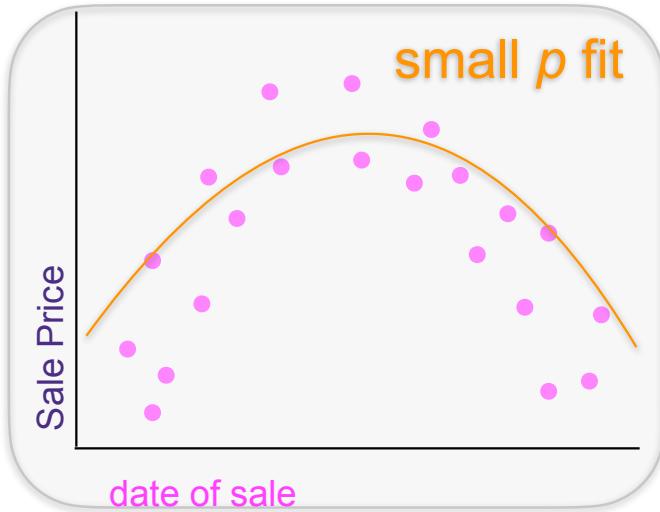


What do these two graphs imply?

A: large p



B: small p



Least squares loss is 0 on training data

Least squares loss is > 0 on training data

Least squares training error is lower on A than B

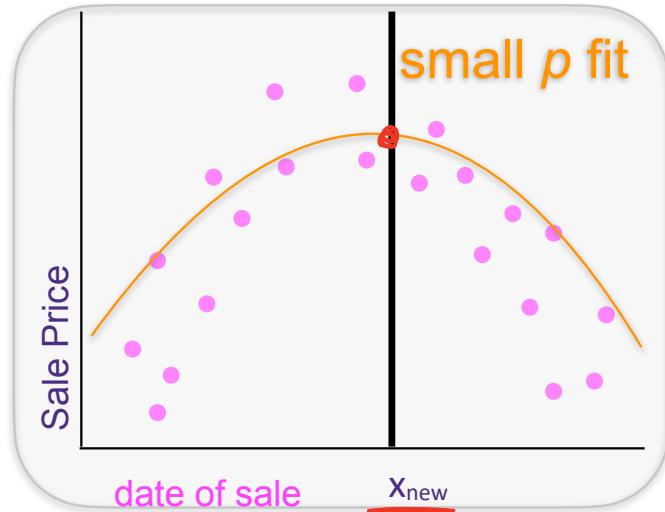
Predicting sale price for a new house: A vs B

Our goal is to predict prices for new houses

A: large p



B: small p



that “look like” the houses in our training data

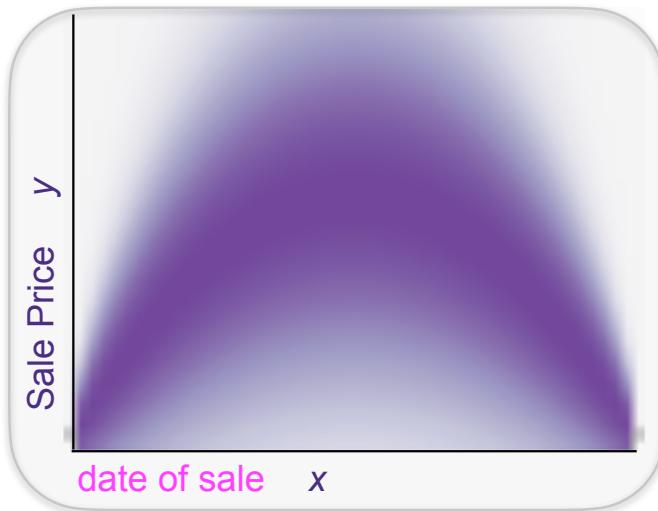
What do we mean by “look like”?

On *average* over a house drawn from this distribution, we want to make a good prediction.

This will work best if our training data came from the same distribution...

What do we mean by “look like”?

$$P_{XY}(X = x, Y = y)$$



On *average* over a house drawn from this distribution, we want to make a good prediction.

This will work best if our training data came from the same distribution...

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

our loss
our loss

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

(Least squares)

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

Thus far, we've been using η which is a:

- Linear functions of X
- Degree p polynomials of X
- Linear "generalization" of X in p dimensions

Statistical Learning

Eger bir date of sale biliyorsak, ve bu Pxy dagilimini biliyorsak, yapmak istedigimiz sey, o date of sale'a gore orijinal dagilimdan average price'i tahmin etmek.

house price

Goal: Predict Y given X

$$P_{XY}(X = x, Y = y)$$

$$Y \in [0, 1]$$

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

$$\eta(x) = \arg \min_{c \in [0, 1]} \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$

house
price
for
that
date
of
sale

particular
date of sale.

ETA burada bu date of sale'a sahip evler icin ortalama degeri tahmin ediyor

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

$$\Rightarrow \boxed{\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X \left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x] \right]}$$
$$\eta(x) = \arg \min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X \left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x] \right]$$

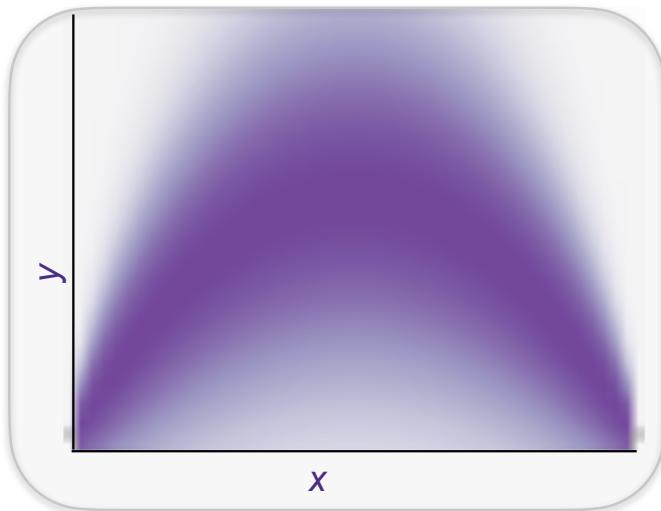
$$\eta(x) = \arg \min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$

Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

$$P_{XY}(X = x, Y = y)$$

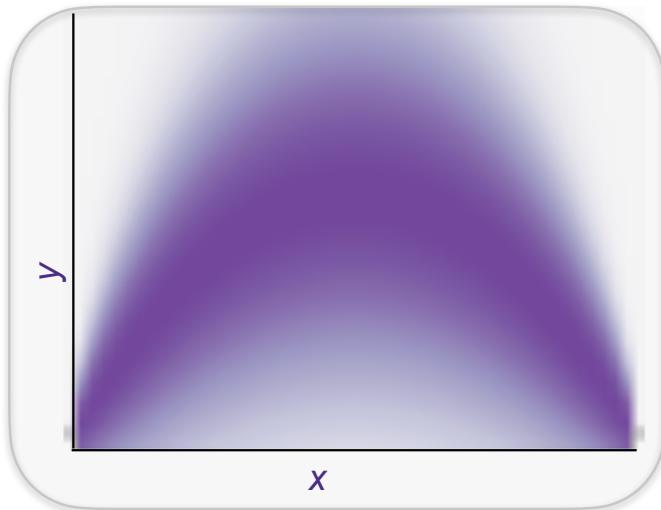


Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

~~Σ~~

$$P_{XY}(X = x, Y = y)$$



Statistical Learning

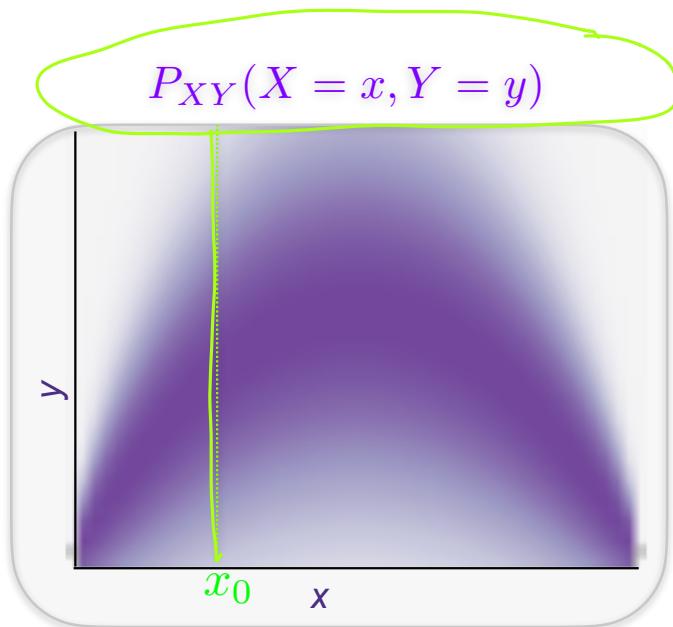
$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

$$P_{XY}(X = x, Y = y)$$

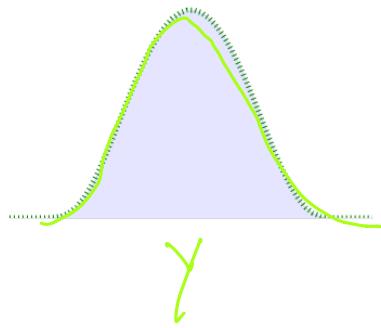


Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$



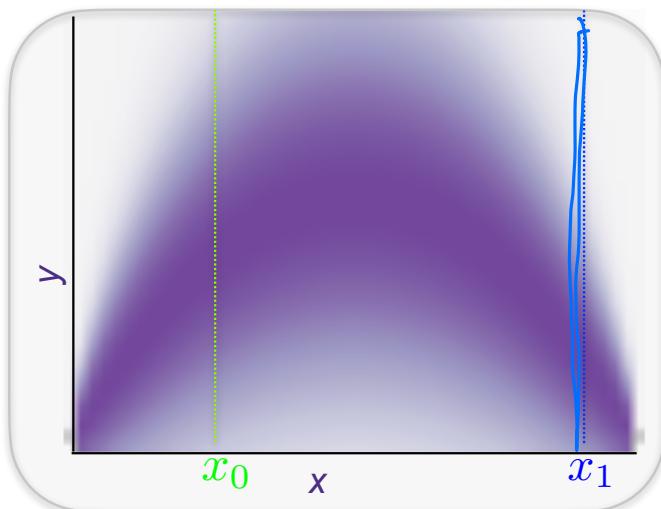
$$P_{XY}(Y = y|X = x_0)$$



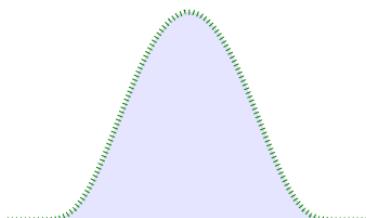
Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

$$P_{XY}(X = x, Y = y)$$

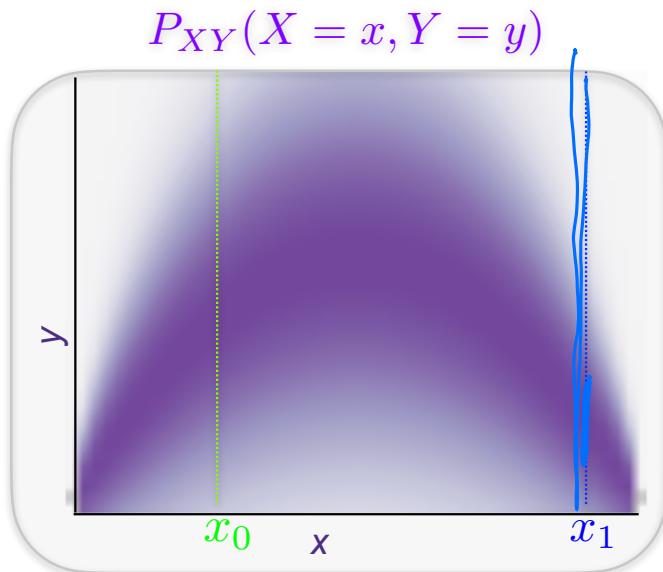


$$P_{XY}(Y = y|X = x_0)$$

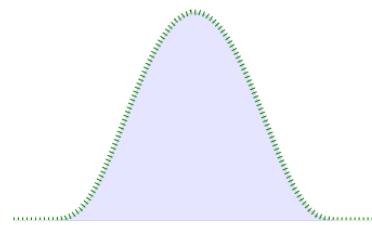


Statistical Learning

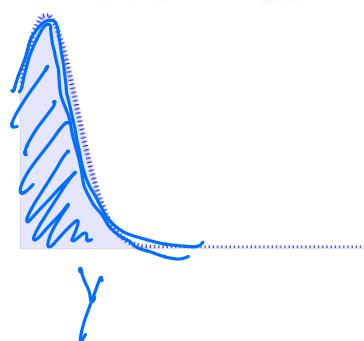
$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$



$$P_{XY}(Y = y|X = x_0)$$



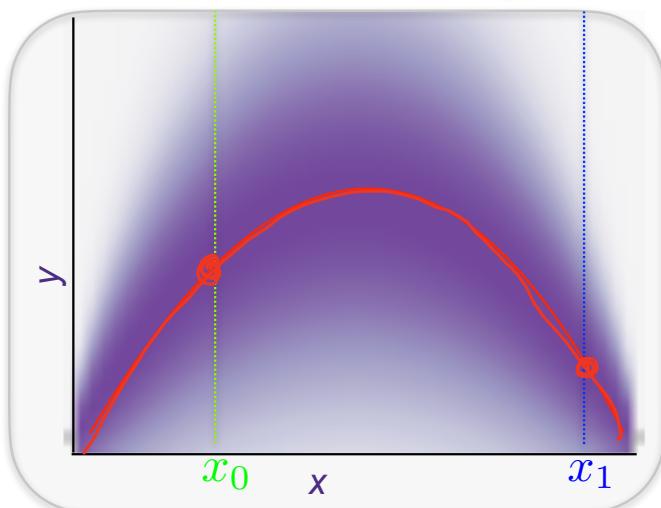
$$P_{XY}(Y = y|X = x_1)$$



Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

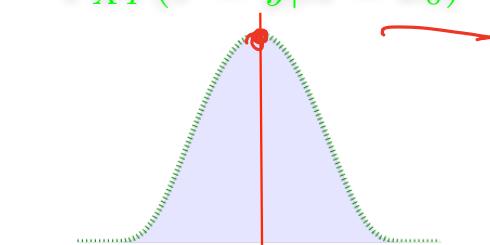
$$P_{XY}(X = x, Y = y)$$



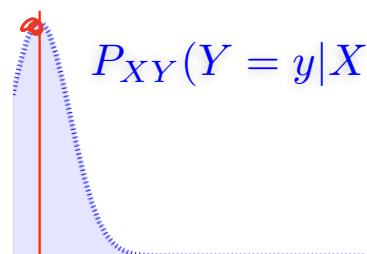
Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$P_{XY}(Y = y|X = x_0)$$

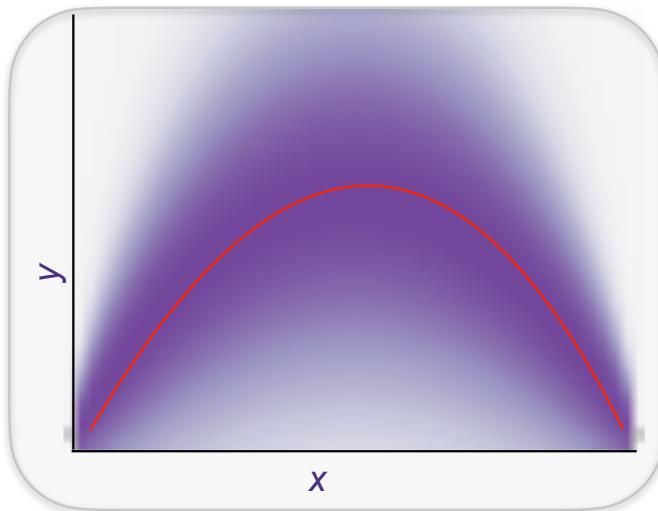


$$P_{XY}(Y = y|X = x_1)$$



Statistical Learning

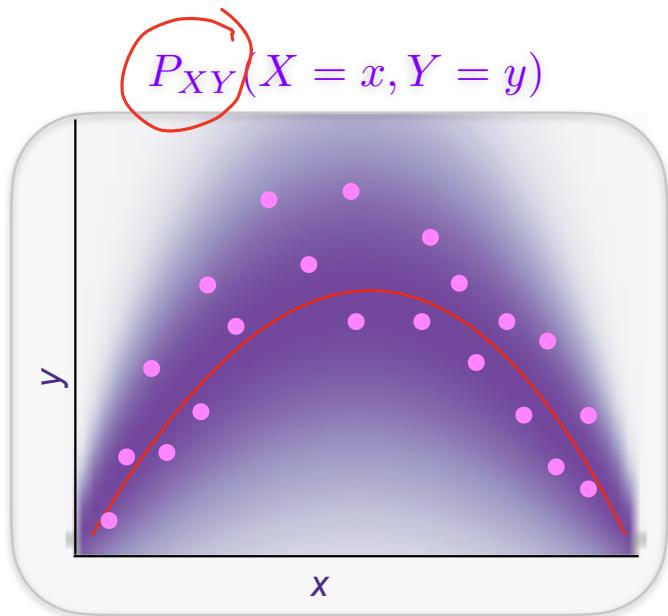
$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Statistical Learning



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

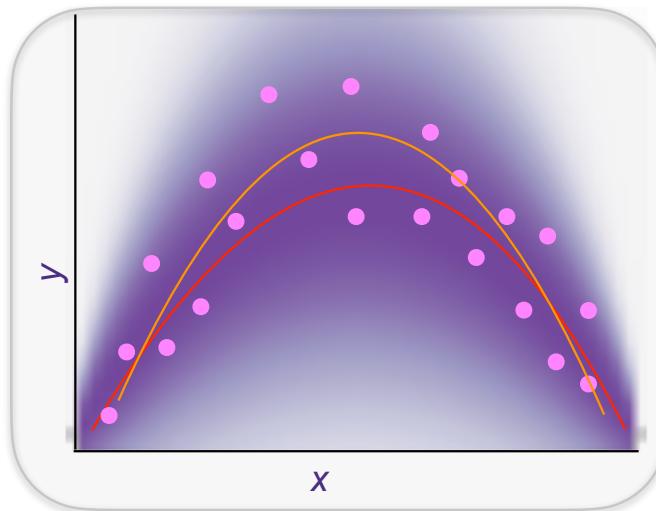
But we only have samples:

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

\

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

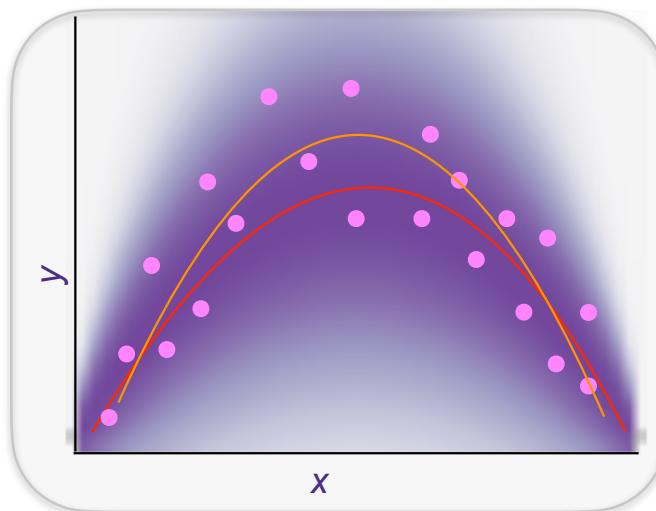
and are restricted to a function class (e.g., linear) so we compute:
quadratic

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

\mathcal{F} : linear functions
quadratic functions

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

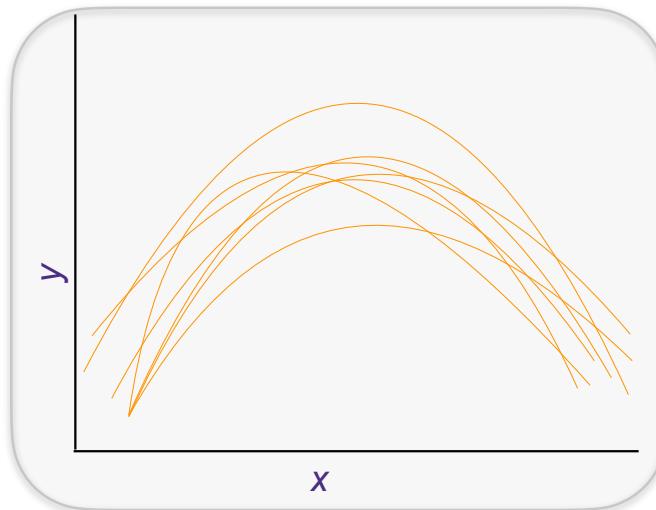
and are restricted to a function class (e.g., linear)
so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

We care about future predictions: $\mathbb{E}_{XY}[(Y - \hat{f}(X))^2]$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ results in different \hat{f}

Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

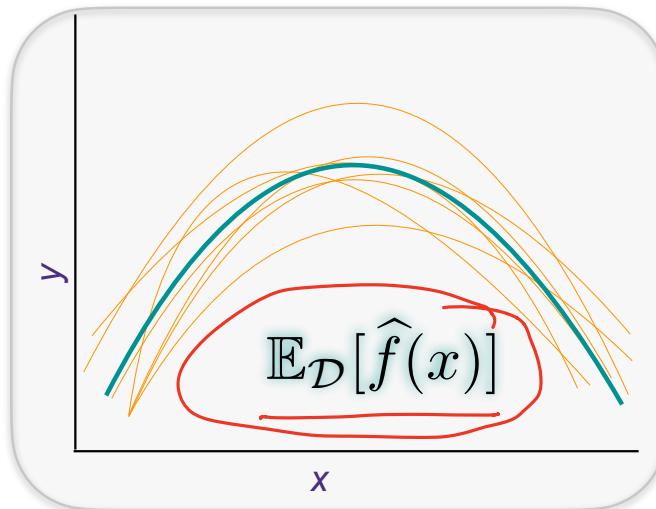
and are restricted to a function class (e.g., linear)
so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$



Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ results in different \hat{f}

Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

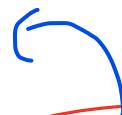
$$(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

and are restricted to a function class (e.g., linear)
so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Bias-Variance Tradeoff

ideal



$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

we learn this func f.

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(\underline{Y} - \underline{\hat{f}}_{\mathcal{D}}(x))^2] | X = x]$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \hat{f}_{\mathcal{D}}(x))^2] | X = x]$$


Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \hat{f}_{\mathcal{D}}(x))^2] | X = x]$$

irreducible error

Caused by stochastic
label noise

learning error

Caused by either using too “simple”
of a model or not enough
data to learn the model accurately

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$
$$\widehat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$
$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2] | X = x] = \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \widehat{f}_{\mathcal{D}}(x))^2] | X = x]$$
$$= \mathbb{E}_{Y|X} \left[\underbrace{\mathbb{E}_{\mathcal{D}}[(Y - \eta(x))^2]}_{A^2} + 2(Y - \eta(x))(\eta(x) - \widehat{f}_{\mathcal{D}}(x)) \right.$$
$$\left. + \underbrace{(\eta(x) - \widehat{f}_{\mathcal{D}}(x))^2}_{B^2} \right]$$

irreducible error

Caused by stochastic
label noise

learning error
Caused by either using too “simple”
of a model or not enough
data to learn the model accurately

Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[(Y - \eta(x))^2] = \mathbb{E}_{Y|X}[(Y - \hat{f}(x))^2] + \mathbb{E}_{Y|X}[(\hat{f}(x) - \eta(x))^2]$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2]|X = x] = \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \hat{f}_{\mathcal{D}}(x))^2]|X = x]$$

$$(A + B)^2$$

$$= \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x))^2 + 2(Y - \eta(x))(\eta(x) - \hat{f}_{\mathcal{D}}(x)) + (\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]|X = x]$$

$$= \mathbb{E}_{Y|X}[(Y - \eta(x))^2|X = x] + \mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]$$

irreducible error

Caused by stochastic
label noise

Buranın
expectation'i 0
oluyor.

learning error
Caused by either using too “simple”
of a model or not enough
data to learn the model accurately

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Learning Error $\left[\frac{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]}{\uparrow}$ CS

$$\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]$$

How much our estimation vary from the average value for a particular x?

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \widehat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \widehat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)]) + \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X=x]$$
$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$
$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]) + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))] \\ &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \end{aligned}$$

2A+B

A: $\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)$
B: $(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2$
B2: $(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)) \\ &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= (\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]\end{aligned}$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)) \\ &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= (\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]\end{aligned}$$

bias squared

Bias-Variance Tradeoff

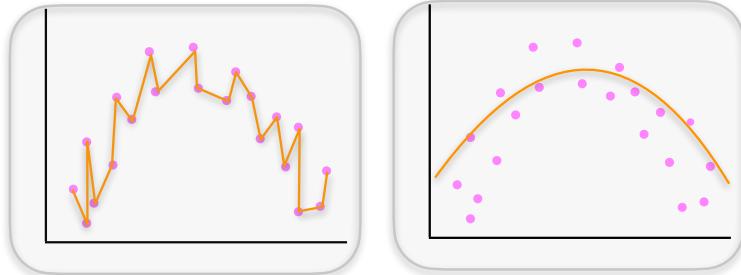
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)) \\ &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= (\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]\end{aligned}$$

bias squared **variance**

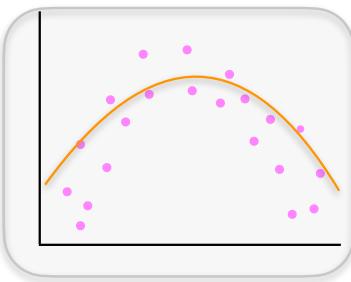
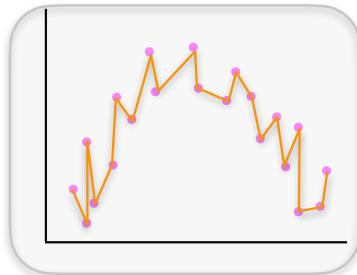
Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{bias squared}}$$



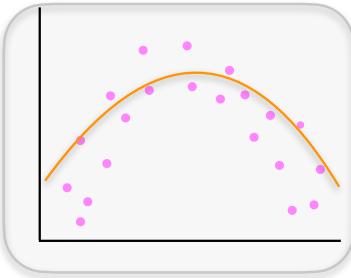
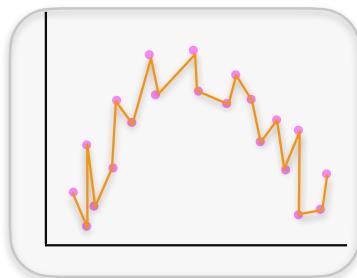
Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{bias squared variance}}$$



Bias-Variance Tradeoff

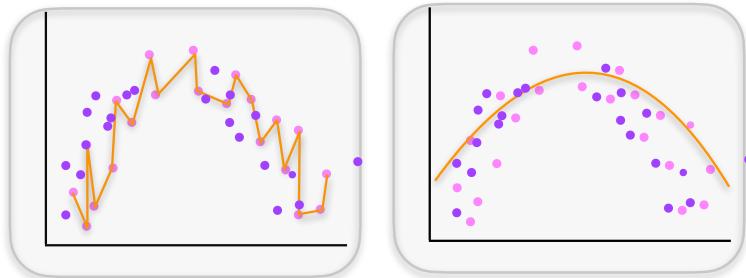
$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2}_{\text{bias squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}}$$



If we re-drew our data, what the LS training error estimator look like for generalized linear functions in small p/large p dimensions?

Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{bias squared variance}}$$



variance measures how much our estimator change when we re-drew our data and compute the ls estimator again.

If we re-drew our data, what the LS training error estimator look like for generalized linear functions in small p/large p dimensions?

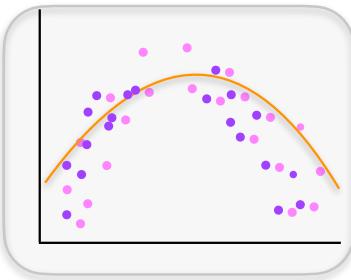
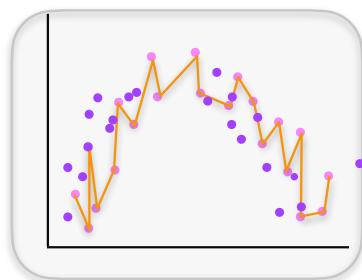
bias ne olduguna bak.

small p ----> high bias, low variance

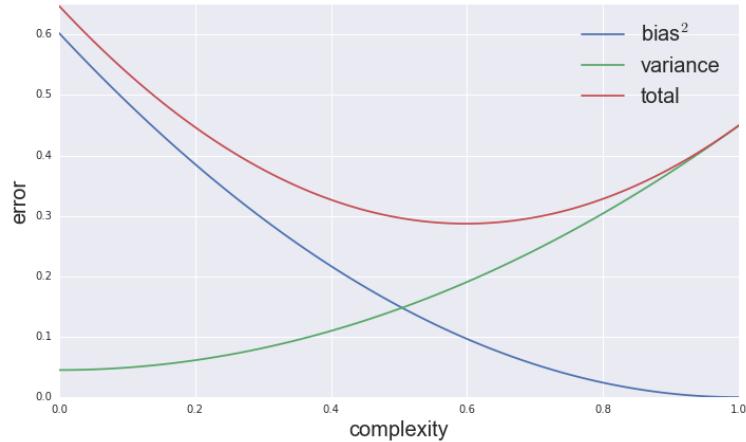
large p ----> low bias, high variance

Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{bias squared variance}}$$



If we re-drew our data, what the LS training error estimator look like for generalized linear functions in small p/large p dimensions?



model complexity arttıkça bias error azalıyor.

variance error artıyor.

Example: Linear LS

→ Unbiased

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) =$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} x$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x]}$$

irreducible error

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x] = \sigma^2}$$

irreducible error

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x] = \sigma^2 \quad \frac{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2}{}$$

irreducible error

bias squared

Example: Linear LS: compute bias

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 .$$

bias squared

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} =$$

variance

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}$$

variance

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

variance

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

variance $= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

variance

$$= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T)$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

variance

$$\begin{aligned} &= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x \\ &= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T) \end{aligned}$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

variance

$$\begin{aligned} &= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x \\ &= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T) \end{aligned}$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma \quad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\text{if } y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

$$\begin{aligned} \mathbf{variance} &= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x \\ &= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T) \end{aligned}$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma \quad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x} [$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

$$\begin{aligned} \text{variance} &= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x \\ &= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T) \end{aligned}$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma \quad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x} [\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]]$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

$$\begin{aligned} \text{variance} &= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x \\ &= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T) \end{aligned}$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma \quad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x} [\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]] = \frac{\sigma^2}{n} \mathbb{E}_X [\text{Trace}(\Sigma^{-1} XX^T)] = \frac{p\sigma^2}{n}$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

$$\begin{aligned} \text{variance} &= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x \\ &= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T) \end{aligned}$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma \quad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x} [\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]] = \frac{\sigma^2}{n} \mathbb{E}_X [\text{Trace}(\Sigma^{-1} XX^T)] = \frac{p\sigma^2}{n}$$

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\frac{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x] = \sigma^2}{\text{irreducible error}} \quad \frac{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 = 0}{\text{bias squared}}$$

$$\mathbb{E}_{X=x} \left[\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \right] = \frac{p\sigma^2}{n}$$