Lecture 3: Linear Regression

Shuai Li

John Hopcroft Center, Shanghai Jiao Tong University

shuaili8.github.io

https://shuaili8.github.io/Teaching/VE445/index.html

Last Lecture

- The classification of machine learning
  - Supervised/unsupervised/reinforcement
- Supervised learning
  - Evaluation metrics for classification
    - Accuracy/Precision/Recall/F1 score
  - Evaluation metrics for regression
    - Pearson coefficient/coefficient of determination
  - Model selection: bias/variance/generalization
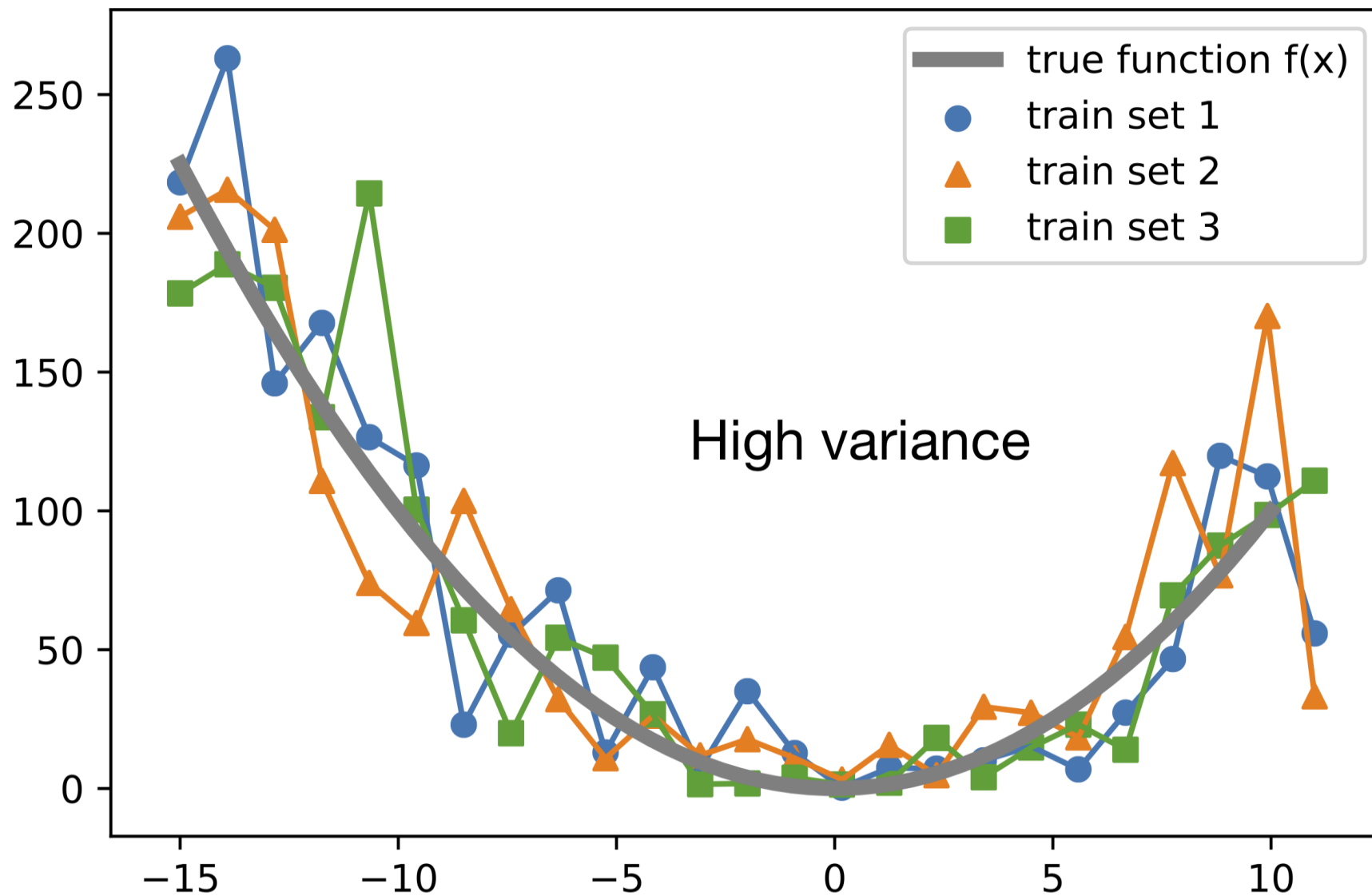 - Machine learning process
  - Generalization error bound

Bias

$$\text{Bias} = E[\hat{\theta}] - \theta.$$

Variance

$$\mathrm{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2].$$

High variance
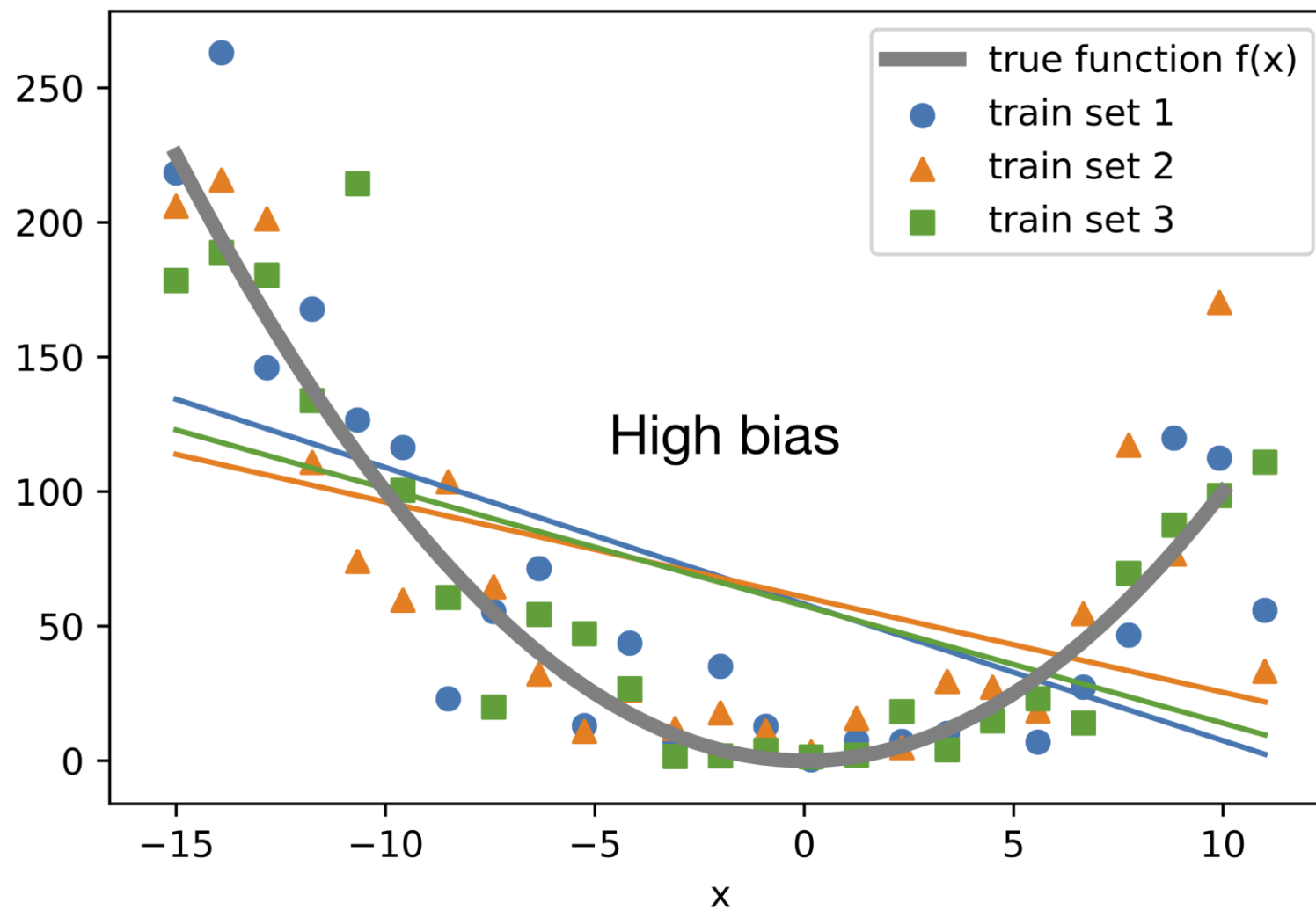
Bias-variance decomposition

True value

Estimated value

$$S = (y - \hat{y})^2$$

$$(y - \hat{y})^2 = (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2$$
$$= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y}).$$

$$E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] = 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})]$$
$$= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})]$$
$$= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}])$$
$$= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}])$$
$$= 0.$$

$$E[S] = E[(y - \hat{y})^2]$$
$$E[(y - \hat{y})^2] = (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2]$$
$$= [\text{Bias}]^2 + \text{Variance}.$$

Can be understood by replacing y and $\hat{y}$ with model parameters θ and $\hat{\theta}$

Generalization Error Bound

A Simple Case Study on Generalization Error

- Finite hypothesis set $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$
- Theorem of generalization error bound:

  For any function $f \in \mathcal{F}$, with probability no less than $1 - \delta$, it satisfies

  $$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

  where

  $$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log \frac{1}{\delta}\right)}$$

  - N: number of training instances
  - d: number of functions in the hypothesis set

Section 1.7 in Dr. Hang Li's text book.

Lemma: Hoeffding Inequality

Let $X_1, X_2, \ldots, X_n$ be bounded independent random variables $X_i \in [a, b]$, the average variable Z is

$$Z = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then the following inequalities satisfy:

$$P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

$$P(\mathbb{E}[Z] - Z \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

http://cs229.stanford.edu/extra-notes/hoeffding.pdf

Proof of Generalized Error Bound

- Assume the bounded loss function $L(y, f(x)) \in [0, 1]$

- Based on Hoeffding Inequality, for $\epsilon > 0$, we have

$$P(R(f) - \hat{R}(f) \geq \epsilon) \leq \exp(-2N\epsilon^2)$$

- As $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ is a finite set, it satisfies

$$P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geq \epsilon) = P(\bigcup_{f \in \mathcal{F}} \{R(f) - \hat{R}(f) \geq \epsilon\})$$
$$\leq \sum_{f \in \mathcal{F}} P(R(f) - \hat{R}(f) \geq \epsilon)$$
$$\leq d \exp(-2N\epsilon^2)$$

Proof of Generalized Error Bound

- Equivalence statements

$$P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geq \epsilon) \leq d \exp(-2N\epsilon^2)$$

$$\updownarrow$$

$$P(\forall f \in \mathcal{F} : R(f) - \hat{R}(f) < \epsilon) \geq 1 - d \exp(-2N\epsilon^2)$$

- Then setting

$$\delta = d \exp(-2N\epsilon^2) \quad \Leftrightarrow \quad \epsilon = \sqrt{\frac{1}{2N} \log \frac{d}{\delta}}$$

The generalized error is bounded with the probability

$$P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$$

$\square$

Generalization error bound revisit

- Finite hypothesis set $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$
- Theorem of generalization error bound:

  For any function $f \in \mathcal{F}$, with probability no less than $1 - \delta$, it satisfies

  $$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

  where

  $$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

  - *N*: number of training instances
  - *d*: number of functions in the hypothesis set

Today's lecture – Linear regression

- Normal equation

- Gradient methods

- Examples

- Probabilistic view

- Applications

- Regularization

Linear Regression

Regression example

- Given the following values of X and Y :
  $$(1,1), (2,2), (4,4), (100,100), (20, 20)$$
  what is the value of Y when X = 5?

- The answer is 5, not difficult

- What if the given values are
  $$(1,1), (2,4), (4,16), (100,10000), (20,400)$$

- Y is 25 when X =5, right?

- Rationale:
  - Look at some examples and then tries to identify the most suitable relationship between the sets X and Y
  - Using this identified relationship, try to predict the values for new examples

Regression example (cont.)

| Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

housing prices
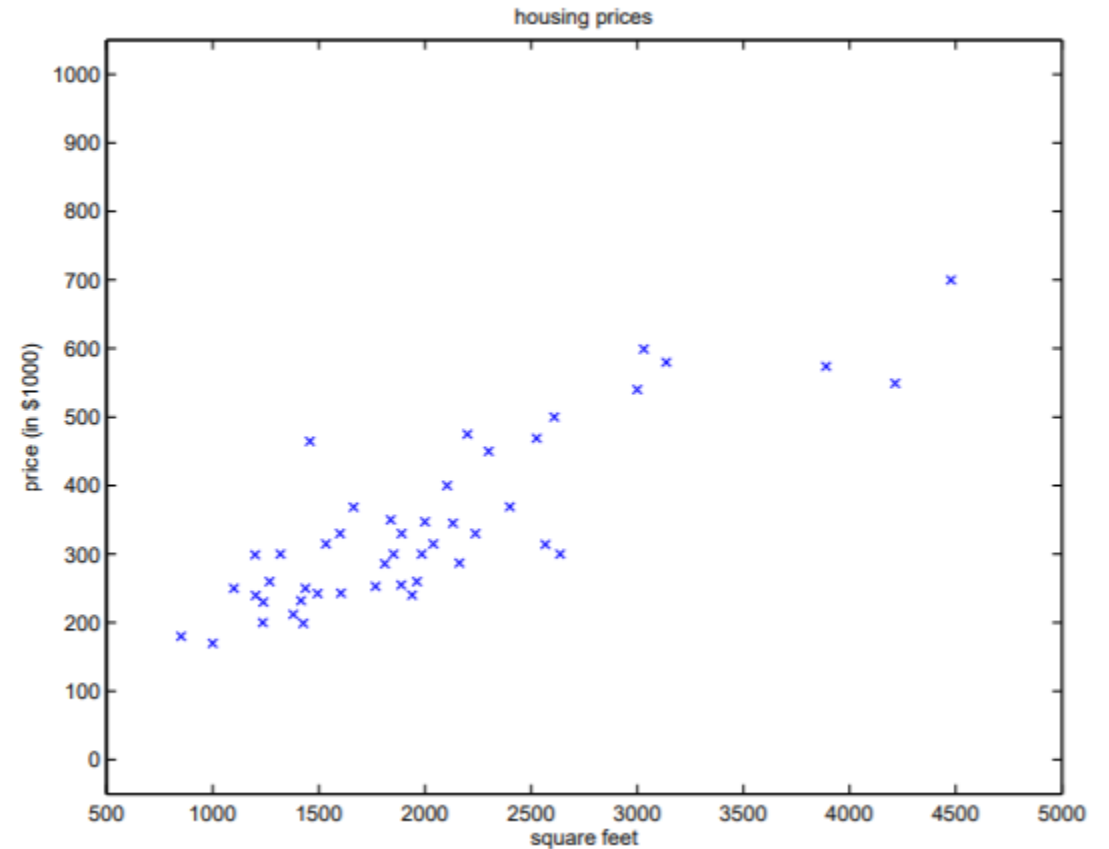
Linear regression

- Use linear relationship to approximate the function of Y on X

- How to select the most appropriate linear model?

- Error: Mean squared error (MSE)

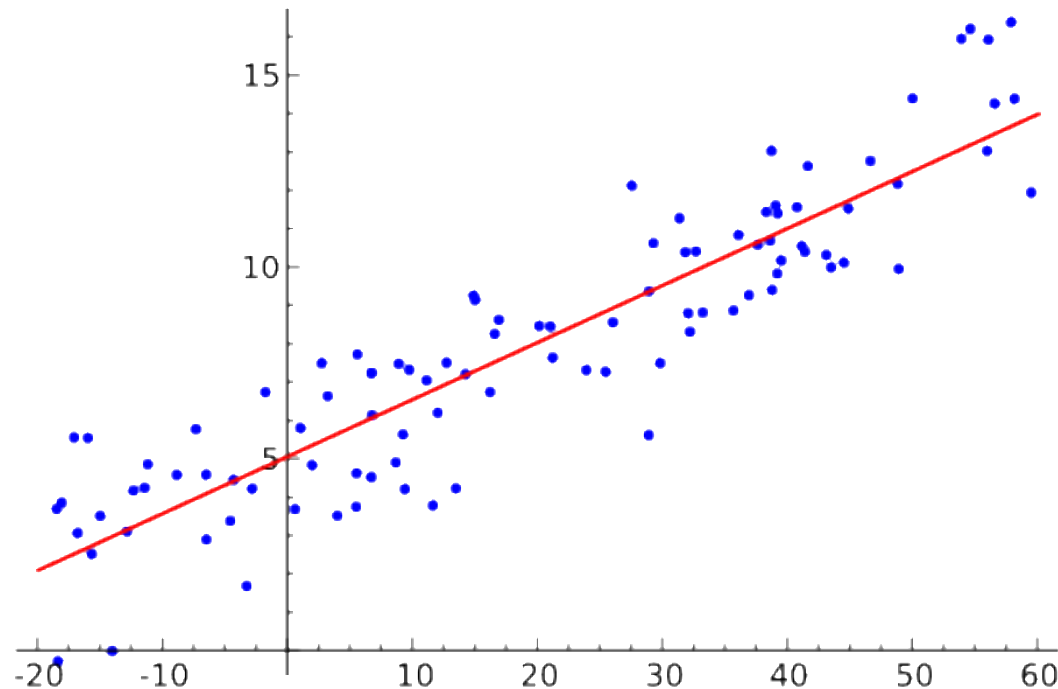$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  - Where Y and $\hat{Y}$ are the true values and predicted values

- Find the linear model with the smallest MSE

Use linear model to fit the given data
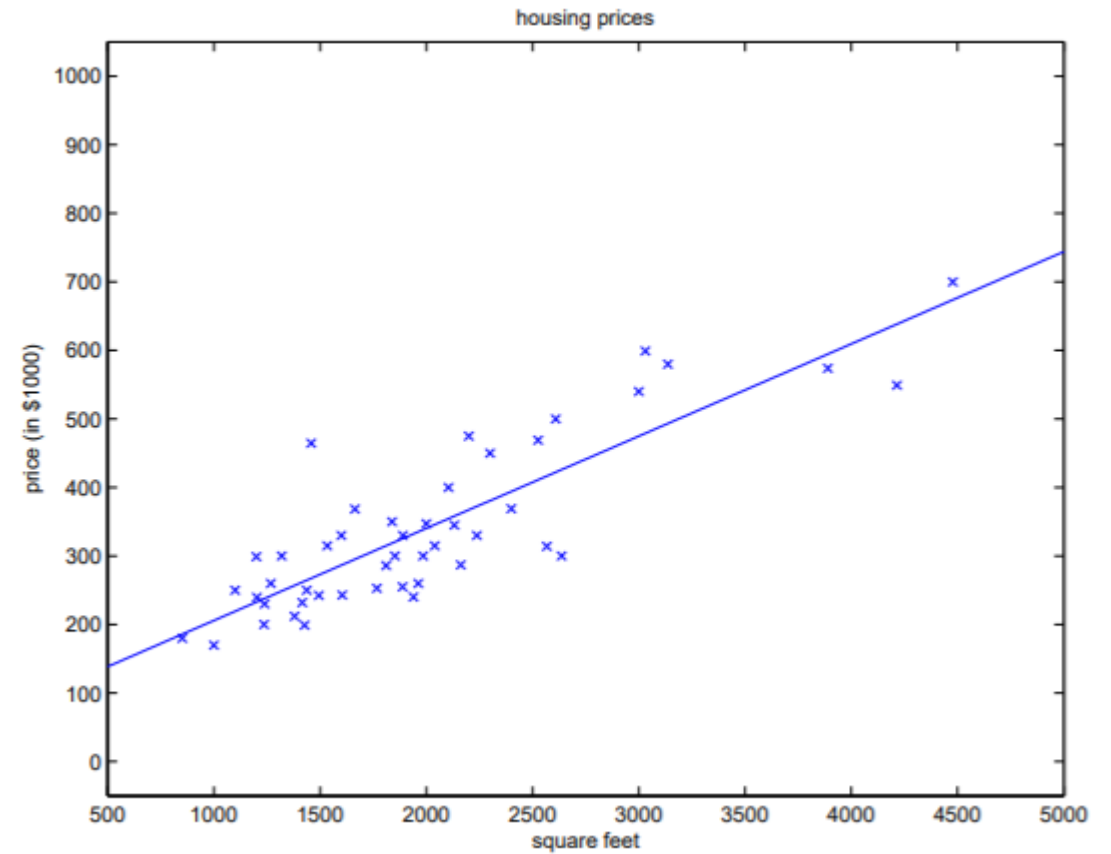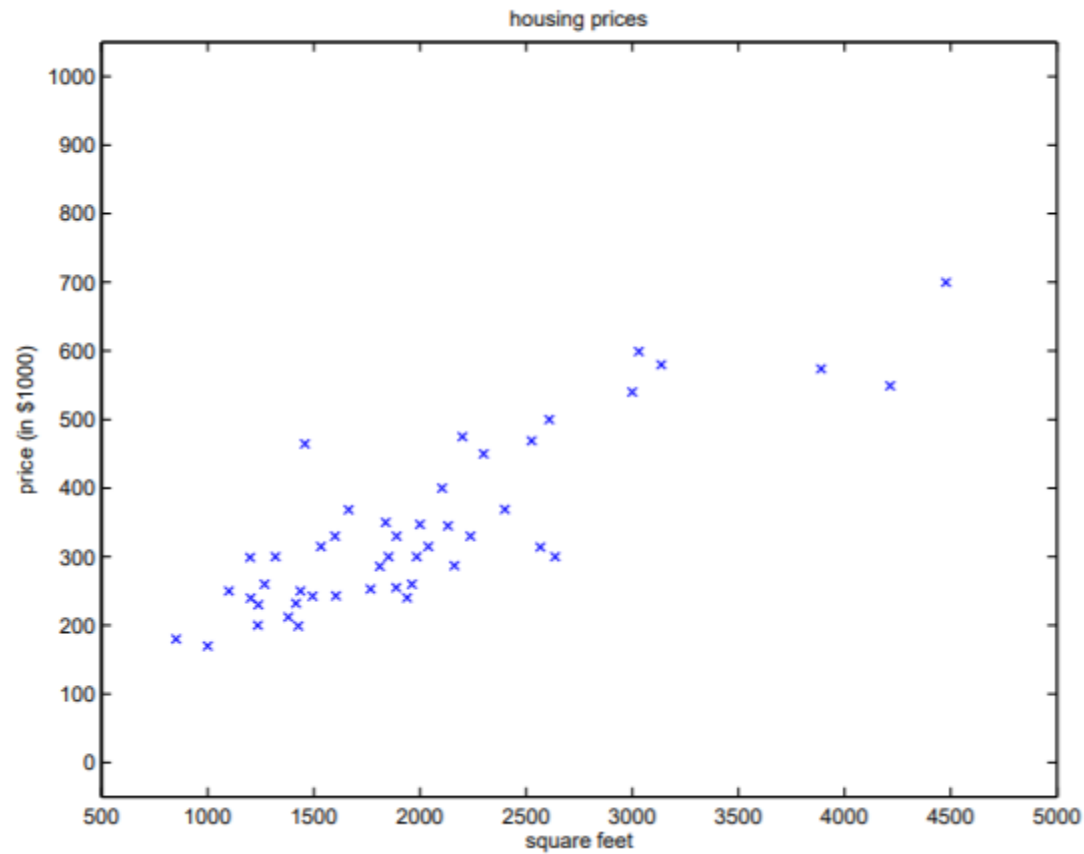
Linear regression 2D example

- In the 2D example, you are looking for a linear equation
  $y = x_1 * \theta_1 + \theta_0$ to fit the data with smallest MSE

Question

- Given the dataset
$$(1,1), (2,4), (3,5)$$
  and the linear model
$$Y = 2X + 1$$
- What is the mean squared error?

- The predicted points are
$$(1,3), (2,5), (3,7)$$
- So the mean squared error (MSE) is
$$\frac{1}{3}(2^2 + 1^2 + 2^2) = 3$$

Question 2

- Given the <real value, predicted value> pairs as:
  $< 1.2, 1.7 >, < 0.9, 0.2 >, < -0.3, 0.1 >, < 1.3, 0.3 >, < 1.1, 1.2 >$
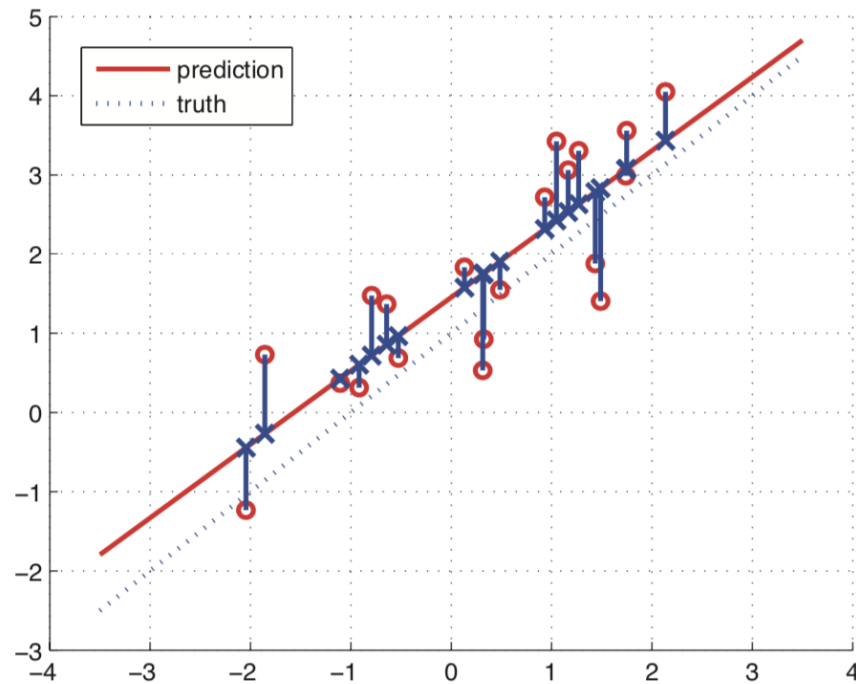 compute the means squared error

- The answer is
  $$\frac{1}{5}(0.5^2 + 0.7^2 + 0.4^2 + 1^2 + 0.1^2) = 0.382$$

How to get linear model with minimal MSE
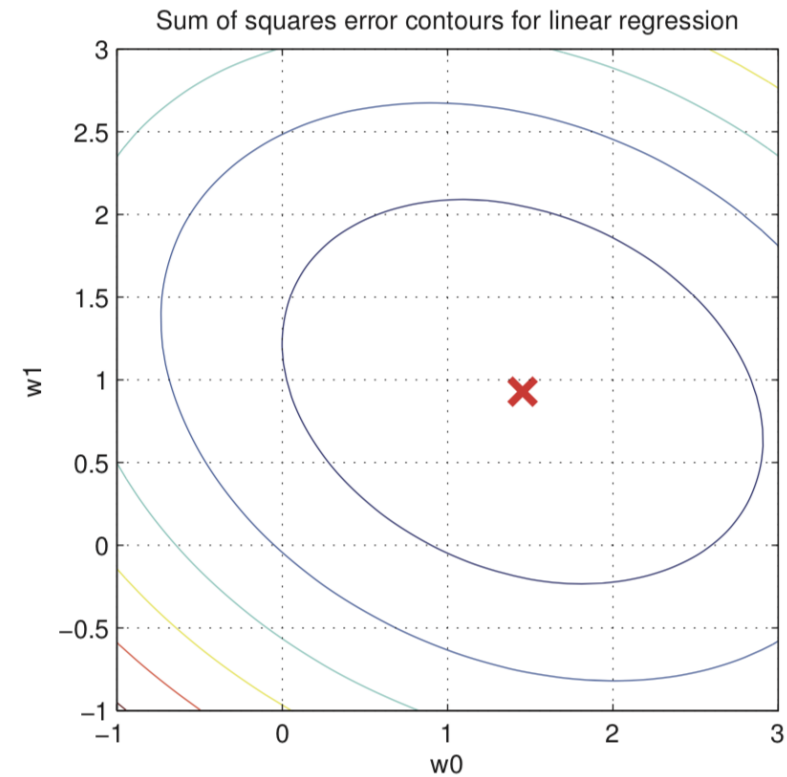
- MSE for model parameter θ:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \theta^\top x_i)^2$$

- Find an estimator $\hat{\theta}$ to minimize $J(\theta)$

- $y = \theta^\top x + b + \varepsilon$. Then we can write $x' = \left(1, x^1, \ldots, x^d\right), \theta = (b, \theta_1, \ldots, \theta_d)$, then $y = \theta^\top x' + \varepsilon$

- Note that $J(\theta)$ is a convex function in θ, so it has a unique minimal point

Interpretation

(a)

Sum of squares error contours for linear regression
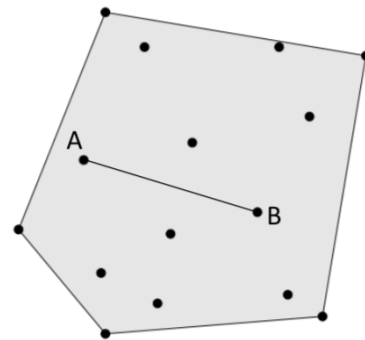
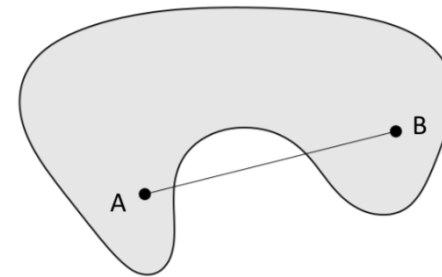(b)

Figure credit: Kevin Murphy

Convex set

- A convex set *S* is a set of points such that, given any two points A, B in that set, the line AB joining them lies entirely within *S*.

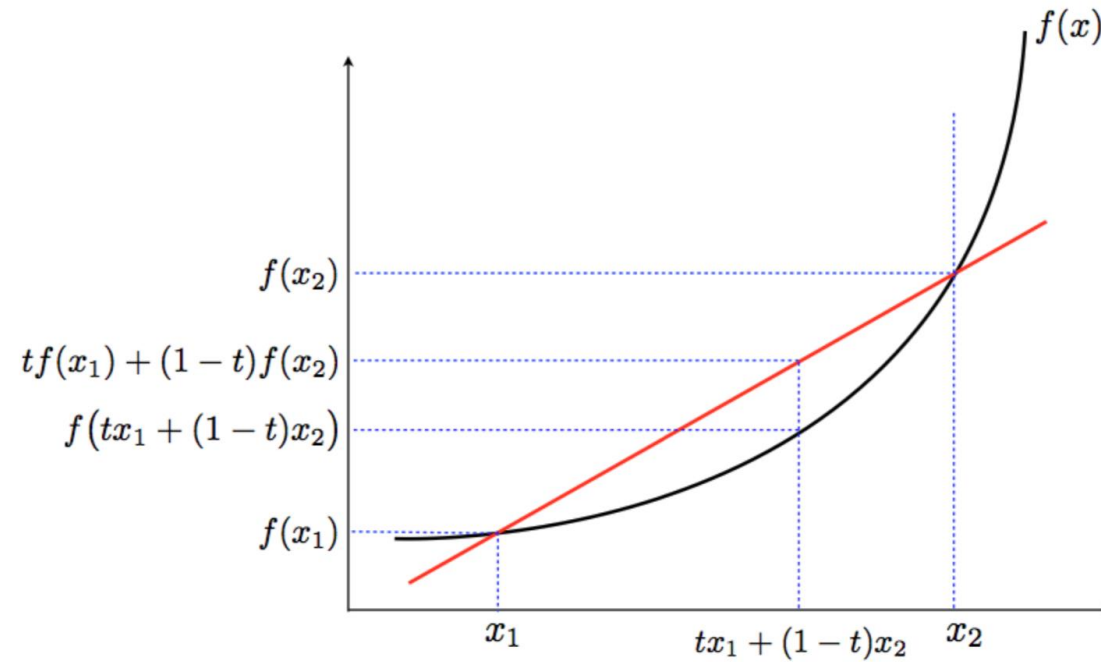$$tx_1 + (1-t)x_2 \in S$$
for all $x_1, x_2 \in S, 0 \le t \le 1$

Convex set

Non-convex set

Convex function

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if **dom** f is a convex set and

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

for all $x_1, x_2 \in \textbf{dom } f, 0 \leq t \leq 1$
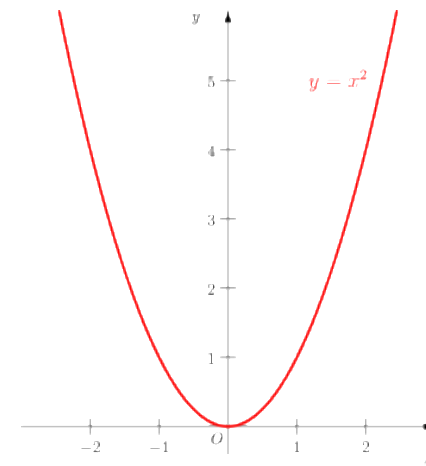
$J(\theta)$ is convex

- $f(x) = (y - x)^2 = (x - y)^2$ is convex in x

- $g(\theta) = f(\theta^\top x)$

$$g\big((1 - t)\theta_1 + t\theta_2\big)$$
$$= f\left((1 - t)\theta_1^\top x + t\theta_2^\top x\right)$$
$$\leq (1 - t)f\left(\theta_1^\top x\right) + tf\left(\theta_2^\top x\right)$$
$$= (1 - t)g(\theta_1) + tg(\theta_2)$$

- The sum of convex functions is convex
- Thus $J(\theta)$ is convex

Minimal point (Normal equation)

- $\frac{\partial J(\theta)}{\partial \theta} = \frac{2}{N} \sum_{i=1}^{N} (\theta^\top x_i - y_i) x_i = \frac{2}{N} \sum_{i=1}^{N} (x_i x_i^\top \theta - x_i y_i)$

- Letting the derivative be zero

$$\left(\sum_{i=1}^{N} x_i x_i^\top \right) \theta = \sum_{i=1}^{N} x_i y_i$$

- If we write $X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} = \begin{bmatrix} x_1^1 & \cdots & x_1^d \\ & \vdots & \\ x_N^1 & \cdots & x_N^d \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$, then

$$X^\top X \theta = X^\top y$$

Minimal point (Normal equation) (cont.)

- $X^\top X \theta = X^\top y$
- When $X^\top X$ is invertible
$$\hat\theta = (X^\top X)^{-1} X^\top y$$

- When $X^\top X$ is not invertible
$$\hat\theta = (X^\top X)^{\dagger} X^\top y$$

pseudo-inverse

- E.g. The pseudo-inverse of $\begin{bmatrix} 1 & & \\ & 2 & \\ & & 0 \end{bmatrix}$ is $\begin{bmatrix} 1 & & \\ & \frac{1}{2} & \\ & & 0 \end{bmatrix}$

Interpretation of Least square error

- A two-dim example: $x_1 * \theta_1 + x_2 * \theta_2 = y$, where x_i and y are vectors.

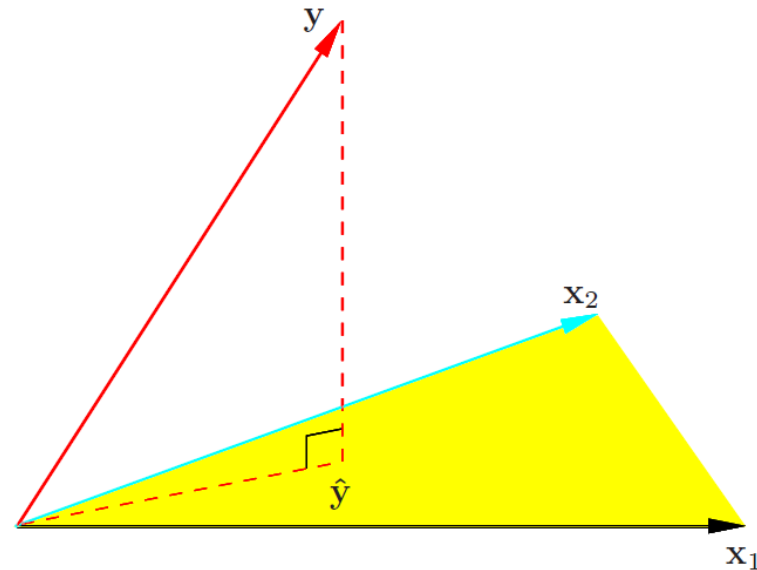- We are using the combination of x_i to approximate the projection of y at their plane.

Figure credit: Trevor Hastie

Geometric interpretation

- N=3,d=2

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

$$\underset{\hat{\mathbf{y}} \in \mathrm{span}(\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_D\})}{\mathrm{argmin}} \|\mathbf{y} - \hat{\mathbf{y}}\|_2.$$

column vectors in X

- $\hat{\theta} = (X^\top X)^{-1} X^\top y$
- $\hat{y} = X\hat{\theta} = X(X^\top X)^{-1} X^\top y$

hat matrix
(put a "hat" on y)
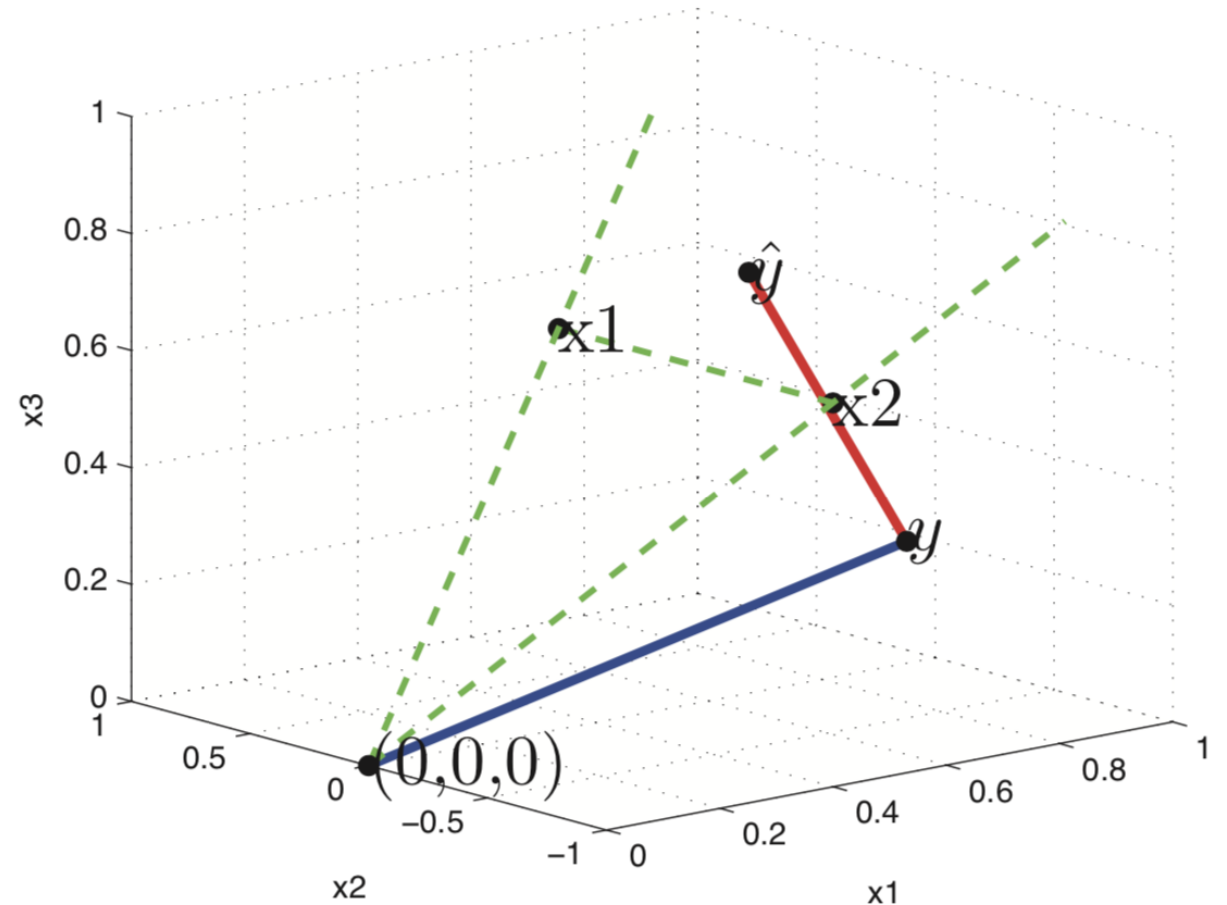

Figure credit: Kevin Murphy

29
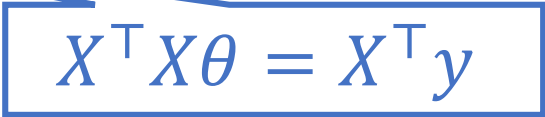
Examples

Question 1

- Given the dataset

$$(1,1), (2,4), (3,5)$$

compute the normal equation for θ, solve θ and compute the MSE

$$X^\top X \theta = X^\top y$$

- $X = \begin{bmatrix} x_1^\top \\ x_2^\top \\ x_3^\top \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 4 \\ 5 \end{bmatrix}$

$$X^\top X = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, X^\top y = \begin{bmatrix} 10 \\ 24 \end{bmatrix}$$

$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 24 \end{bmatrix}$$

- $\theta = \left[-\frac{2}{3}, 2\right], y = -\frac{2}{3} + 2x.$ MSE=$\frac{2}{9}$

Question 2

- Some economist say that the impact of GDP in 'current year' will have effect on vehicle sales 'next year'. So whichever year GDP was less, the coming year sales was lower and when GDP increased the next year vehicle sales also increased.

- Let's have the equation as $y = \theta_0 + \theta_1 x$, where
- y = number of vehicles sold in the year
- x = GDP of prior year

We need to find θ_0 and θ_1

Question 2 (cont.)

- Here is the data between 2011 and 2016.

Year	GDP	Sales of vehicle
2011	6.2	
2012	6.5	26.3
2013	5.48	26.65
2014	6.54	25.03
2015	7.18	26.01
2016	7.93	27.9
2017		30.47
2018		

Homework

- Question 1: what is the normal equation?

- Question 2: suppose the GDP increasement in 2017 is 7%, how many vehicles will be sold in 2018?

Gradient methods
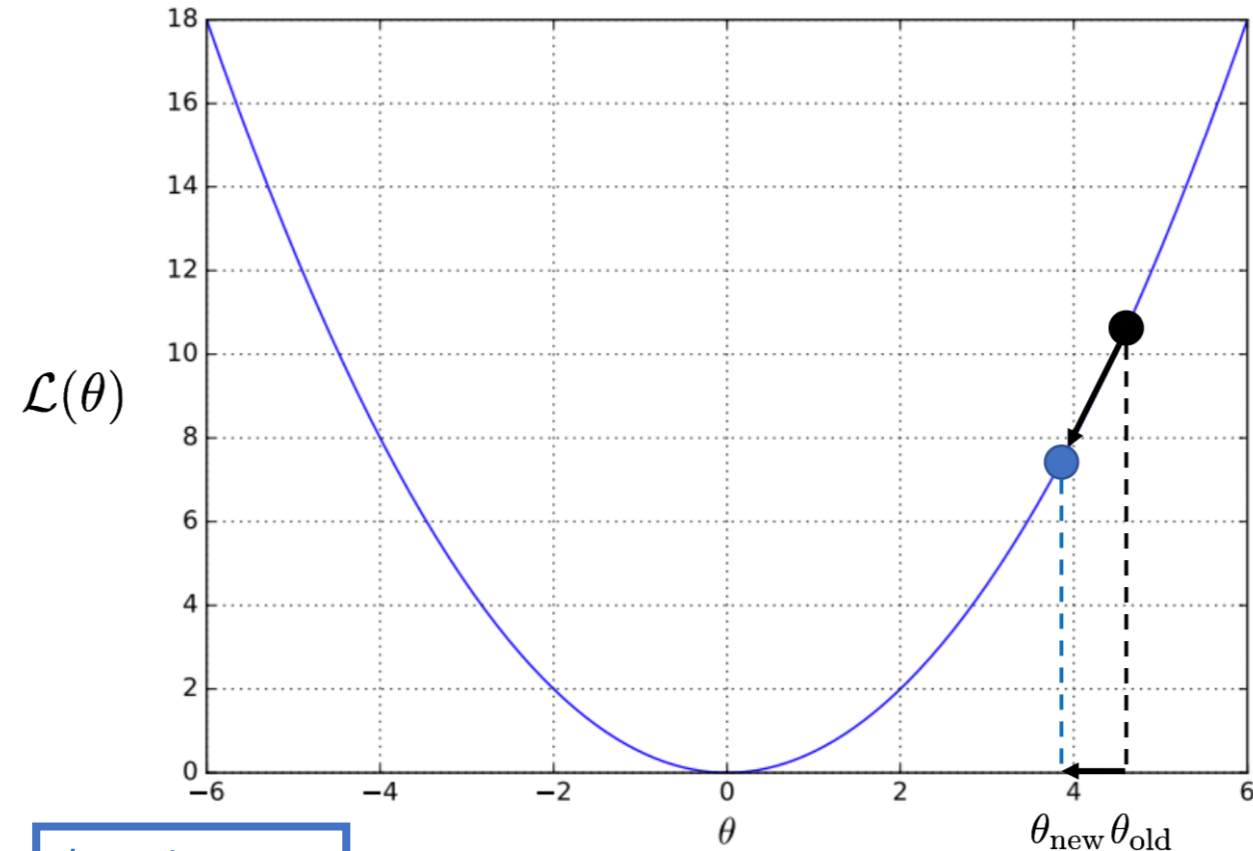
Motivation – large dataset

- Too big to compute directly
$$\hat{\theta} = (X^\top X)^{-1} X^\top y$$

- Recall the objective is to minimize the loss function

$$L(\theta) = J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta^\top x_i)^2$$

- Gradient descent method

learning rate

$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$$

Batch gradient descent

- $f_\theta(x) = \theta^\top x$

$$J(\theta) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2 \qquad \min_\theta J(\theta)$$

- Update $\theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta \dfrac{\partial J(\theta)}{\partial \theta}$ for the whole batch

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{N} \sum_{i=1}^{N} (y_i - f_\theta(x_i)) \frac{\partial f_\theta(x_i)}{\partial \theta}$$

$$= -\frac{1}{N} \sum_{i=1}^{N} (y_i - f_\theta(x_i)) x_i$$

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \frac{1}{N} \sum_{i=1}^{N} (y_i - f_\theta(x_i)) x_i$$

Stochastic gradient descent

$$J^{(i)}(\theta) = \frac{1}{2}(y_i - f_\theta(x_i))^2 \qquad \min_\theta \frac{1}{N} \sum_i J^{(i)}(\theta)$$

- Update $\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial J^{(i)}(\theta)}{\partial \theta}$ for every single instance

$$\frac{\partial J^{(i)}(\theta)}{\partial \theta} = -(y_i - f_\theta(x_i))\frac{\partial f_\theta(x_i)}{\partial \theta}$$

$$= -(y_i - f_\theta(x_i))x_i$$

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta(y_i - f_\theta(x_i))x_i$$

- Compare with BGD
  - Faster learning
  - Uncertainty or fluctuation in learning

Mini-Batch Gradient Descent

- A combination of batch GD and stochastic GD

- Split the whole dataset into *K* mini-batches

$$\{1, 2, 3, \ldots, K\}$$

- For each mini-batch *k,* perform one-step BGD towards minimizing

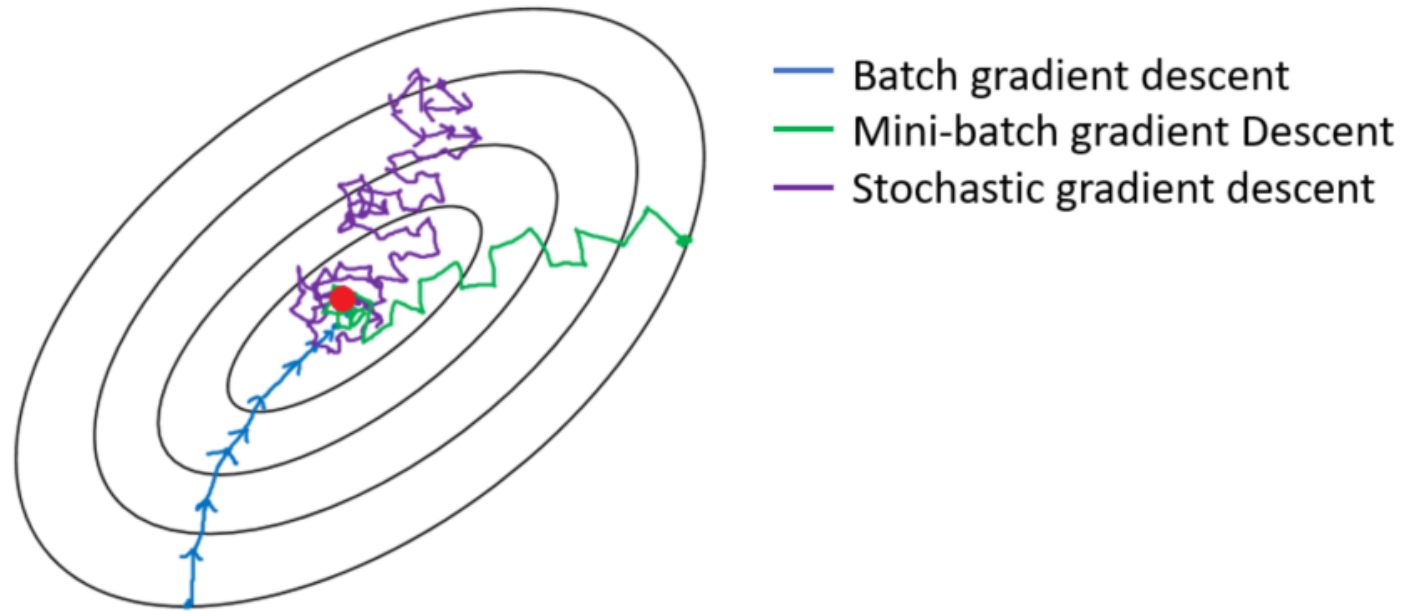$$J^{(k)}(\theta) = \frac{1}{2N_k} \sum_{i=1}^{N_k} (y_i - f_\theta(x_i))^2$$

- Update $\theta_{\text{new}} = \theta_{\text{old}} - \eta \dfrac{\partial J^{(k)}(\theta)}{\partial \theta}$ for each mini-batch

Mini-Batch Gradient Descent (cont.)

- Good learning stability (BGD)
- Good convergence rate (SGD)

- Easy to be parallelized
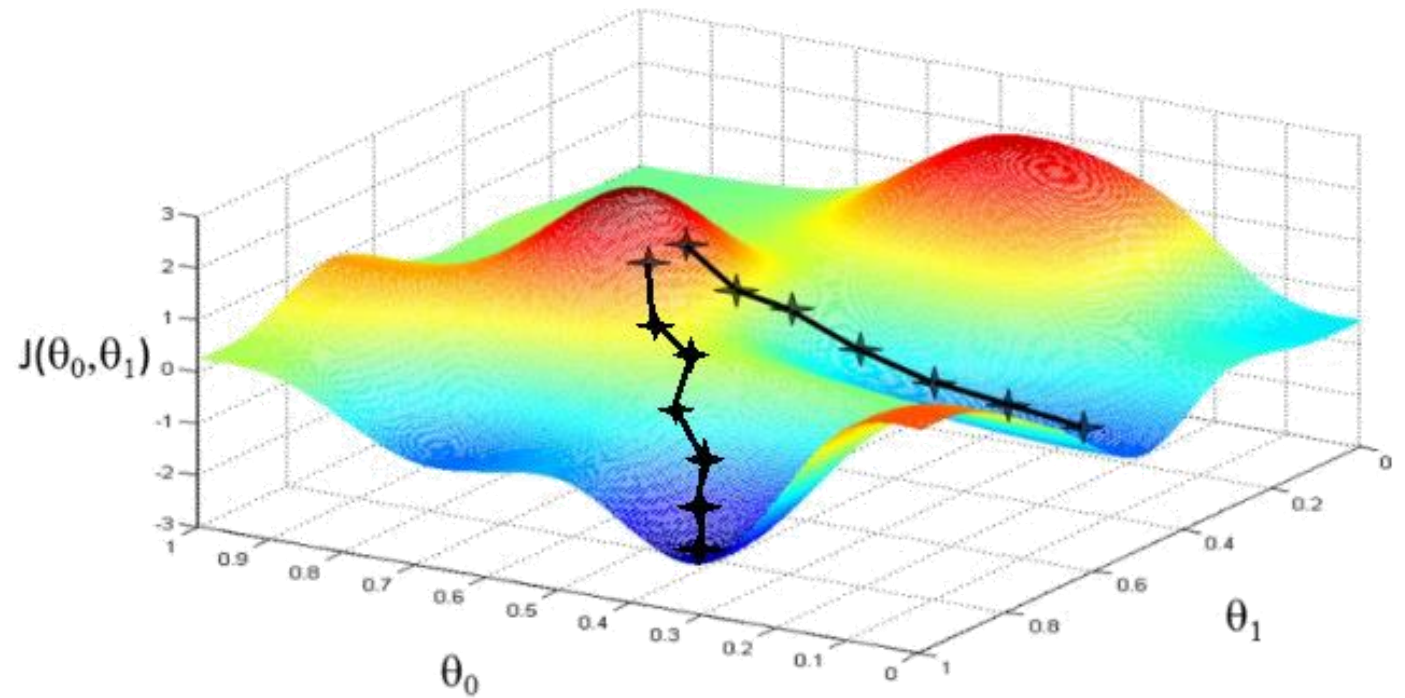 - Parallelization within a mini-batch
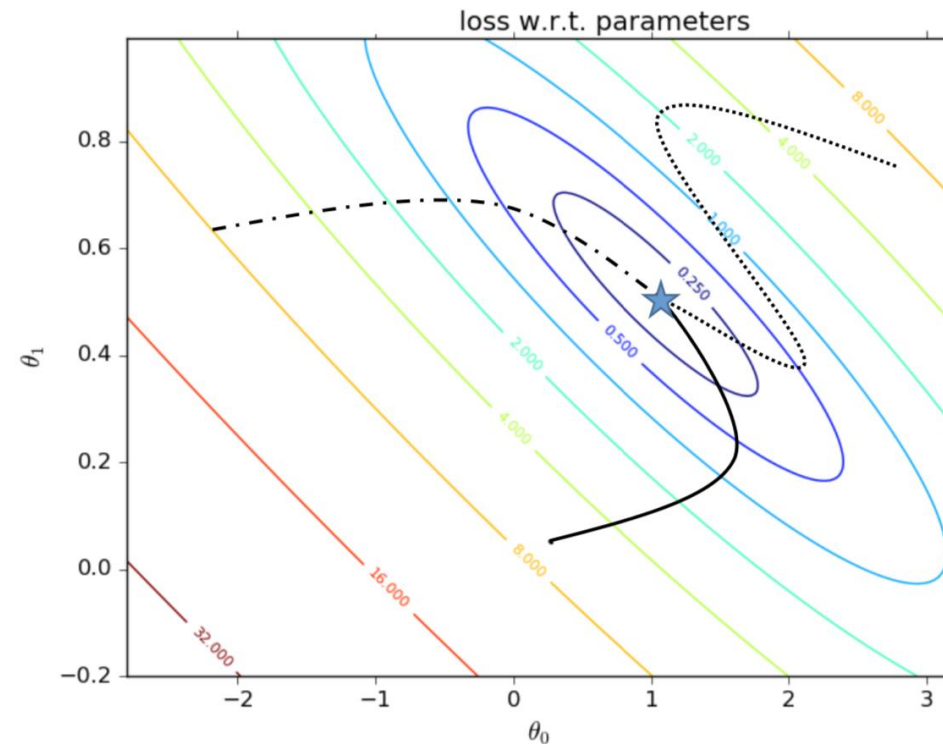
Comparisons

Legend:
- Batch gradient descent
- Mini-batch gradient Descent
- Stochastic gradient descent
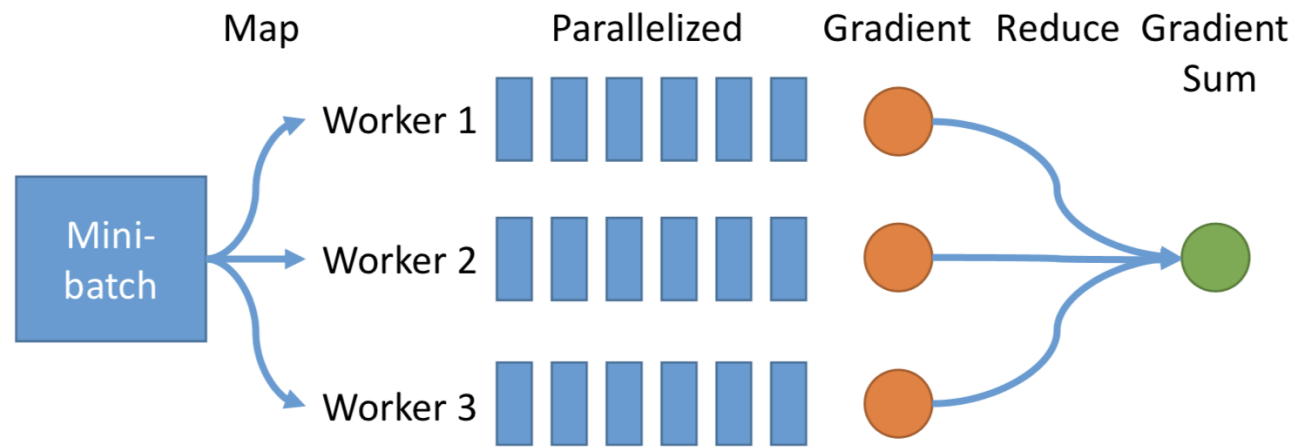
Searching

- Start with a new initial value θ
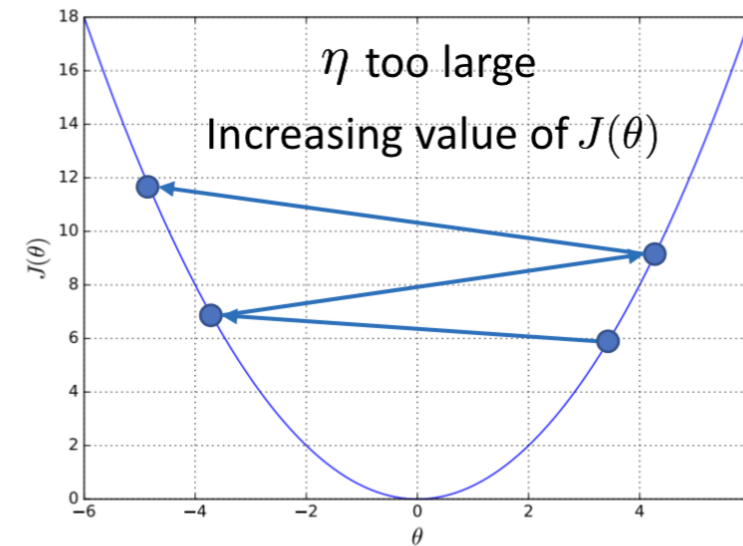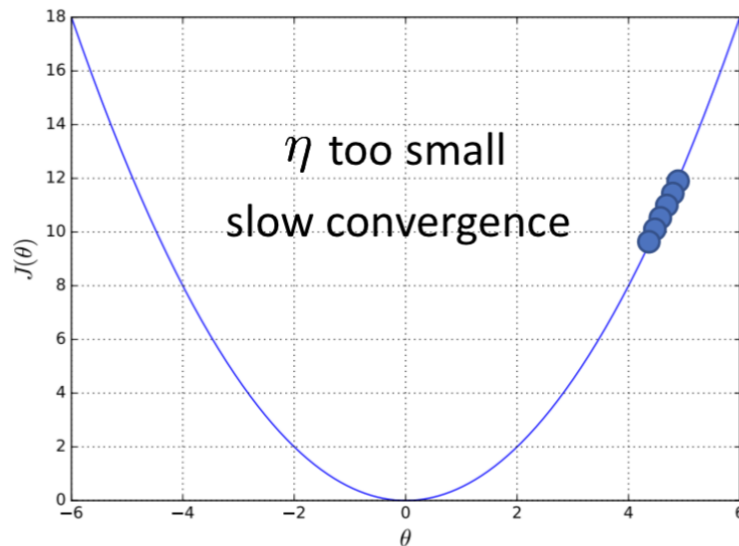- Update θ iteratively (gradient descent)
- Ends at a minimum

Uniqueness of minimum for convex objectives

- Different initial parameters and different learning algorithm lead to the same optimum

Learning rate

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial J(\theta)}{\partial \theta}$$

η too small
slow convergence

η too large
Increasing value of $J(\theta)$

- The initial point may be too far away from the optimal solution, which takes much time to converge

- May overshoot the minimum
- May fail to converge
- May even diverge

- To see if gradient descent is working, print out $J(\theta)$ for each or every several iterations. If $J(\theta)$ does not drop properly, adjust η

Probabilistic view

Probabilistic view

- Assume for each sampled $(x, y) \sim D$,
$$y = \theta^\top x + \varepsilon$$
where ε is Gaussian noise and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, or equivalently,
$y \sim \mathcal{N}(\theta^\top x, \sigma^2)$

- The linear regression estimator $\hat{\theta}$ is the maximal likelihood estimator (MLE) of the data

Maximum likelihood estimation (MLE)

- Frequentists' view

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MLE}} &= \arg\max P(X; \theta) \\
&= \arg\max P(x_1; \theta) P(x_2; \theta) \cdot \cdots \cdot P(x_n; \theta) \\
&= \arg\max \log \prod_{i=1}^{n} P(x_i; \theta) \\
&= \arg\max \sum_{i=1}^{n} \log P(x_i; \theta) \\
&= \arg\min -\sum_{i=1}^{n} \log P(x_i; \theta)
\end{aligned}
$$

MLE view

- Given dataset S, find θ to maximize the likelihood of S, which is

$$\mathbb{P}[S|\theta] = \prod_{i=1}^{N} \mathbb{P}[y_i|x_i, \theta]$$

- $\hat{\theta} = \text{argmax}_\theta \, \mathbb{P}[S|\theta]$

$$= \text{argmax}_\theta \, \log \mathbb{P}[S|\theta]$$

$$= \text{argmax}_\theta \sum_{i=1}^{N} \log \mathbb{P}[y_i|x_i, \theta]$$

- $\mathbb{P}[y_i|x_i, \theta] = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\dfrac{(y_i - \theta^{\top}x_i)^2}{2\sigma^2}}$ Gaussian distribution

MLE view (cont.)

- $\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{N} \log \mathbb{P}[y_i | x_i, \theta]$

$$\mathbb{P}[y_i | x_i, \theta] = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta^\top x_i)^2}{2\sigma^2}}$$

$$= \text{argmax}_\theta \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \theta^\top x_i)^2}{2\sigma^2}$$

$$= \text{argmax}_\theta - \sum_{i=1}^{N} (y_i - \theta^\top x_i)^2$$

$$= \text{argmin}_\theta \sum_{i=1}^{N} (y_i - \theta^\top x_i)^2$$

residual sum of squares (RSS)

$$= \text{argmin}_\theta \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta^\top x_i)^2$$

Application Examples

Trend line

- A trend line represents a trend, the long-term movement in time series data after other components have been accounted for

- E.g. Stock price, heart rate, sales volume, temperature

- Given a set of points in time t and data values y_t, find the linear relationship of y_t with respect to t

- Find a and b to minimize

$$\sum_t [y_t - (\hat{a}t + \hat{b})]^2$$

Finance: Capital asset pricing model (CAPM)

- Describes the relationship between systematic risk and expected return for assets, particularly stocks

- Is widely used throughout finance for pricing risky securities and generating expected returns for assets given the risk of those assets and cost of capital

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f)$$

where:

- $E(R_i)$ is the expected return on the capital asset
- R_f is the risk-free rate of interest such as interest arising from government bonds

- β_i (the *beta*) is the sensitivity of the expected excess asset returns to the expected excess market returns,
- $E(R_m)$ is the expected return of the market
- $E(R_m) - R_f$ is sometimes known as the *market premium*

Example of CAPM

- The risk-free rate of return (which is usually the return rate of government bonds) is 3%, and average market rate is 5%. Suppose the beta in car industry is 1.4, what is the average return rate for the car industry?

- In another way, if we have the risk-free return rate, the market return rate and the return rate of 10 car companies, how to compute the beta for car industry?

Regularization

Problems of ordinary least squares (OLS)

- Best model is to minimize both the bias and the variance

- Ordinary least squares (OLS)
  - Previous linear regression
  - Unbiased
  - Can have huge variance
    - Multi-collinearity among data
      - When predictor variables are correlated to each other and to the response variable
      - E.g. To predict patient weight by the height, sex, and diet. But height and sex are correlated
    - Many predictor variables
      - Feature dimension close to number of data points

- Solution
  - Reduce variance at the cost of introducing some bias
 - Add a penalty term to the OLS equation
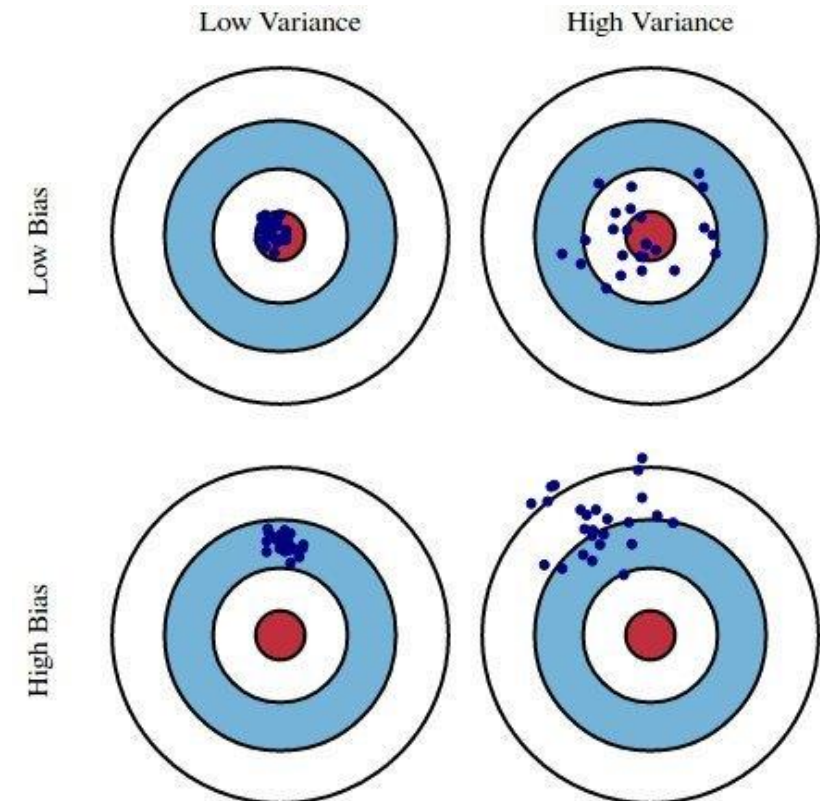
Ridge regression

- Regularization with L2 norm

$$L_{Ridge} = \frac{1}{N}(y - X\theta)^2 + \lambda||\theta||_2$$

- $\lambda \to 0, \hat{\theta}_{Ridge} \to \hat{\theta}_{OLS}$
- $\lambda \to \infty, \hat{\theta} \to 0$
- As λ becomes larger, the variance decreases and the bias increases
- λ: Trade-off between bias and variance
  - Choose by cross-validation
- Ridge regression decreases the complexity of a model but does not reduce the number of variables (compared to Lasso later)
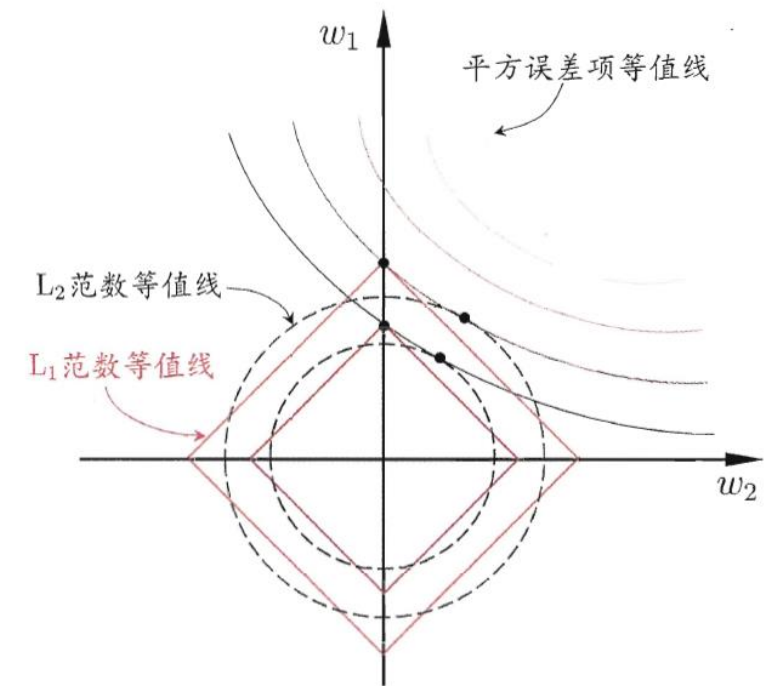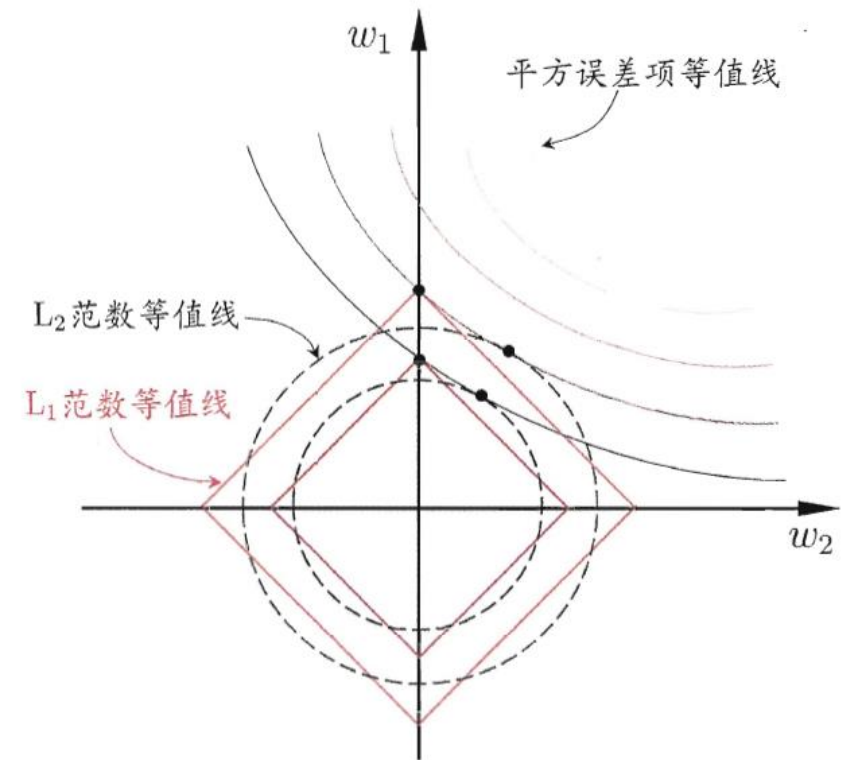
Lasso regression

- Lasso regression: linear regression with L1 norm

$$L_{Lasso} = \frac{1}{N}\,(y - X\theta)^2 + \lambda||\theta||_1$$

- Lasso eliminates some features entirely and gives a subset of predictors that helps mitigate multi-collinearity and model complexity

- Predictors do not shrink towards zero significantly that they are important and thus L1 regularization allows for feature selection (sparse selection)

Ridge vs Lasso

- Often neither one is overall better.
- Lasso can set some coefficients to zero, thus performing variable selection, while ridge regression cannot.
- Both methods allow to use correlated predictors, but they solve multicollinearity issue differently:
  - In ridge regression, the coefficients of correlated predictors are similar;
  - In lasso, one of the correlated predictors has a larger coefficient, while the rest are (nearly) zeroed.
- Lasso tends to do well if there are a small number of significant parameters and the others are close to zero
  - when only a few predictors actually influence the response
- Ridge works well if there are many large parameters of about the same value
  - when most predictors impact the response
- However, in practice, we don't know the true parameter values, so the previous two points are somewhat theoretical. Just run cross-validation to select the more suited model for a specific case.
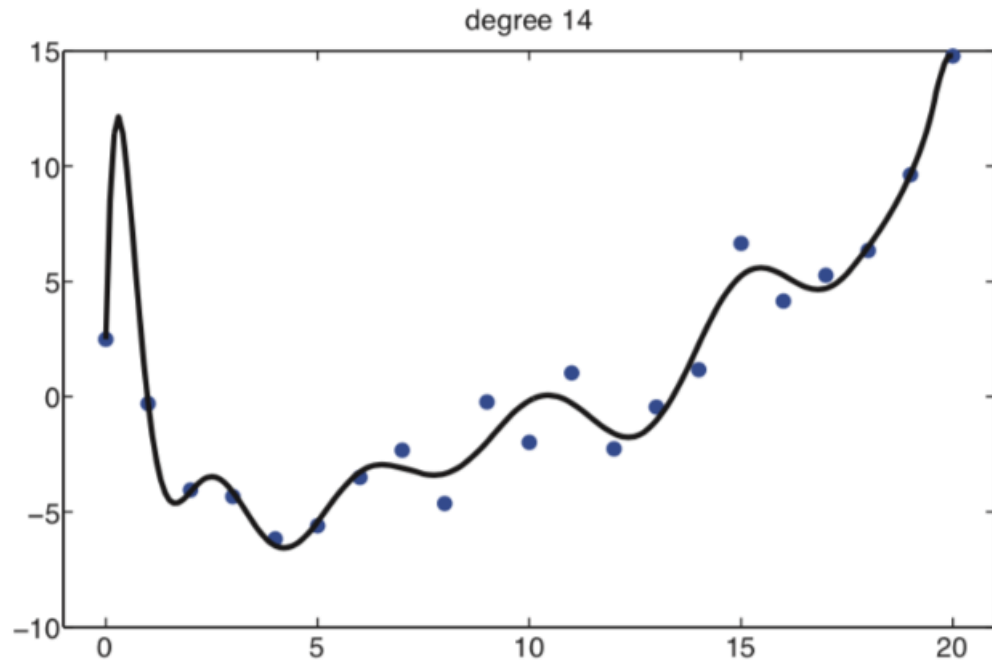- Or… combine the two!

Elastic regression

- A combination of Ridge and Lasso regression

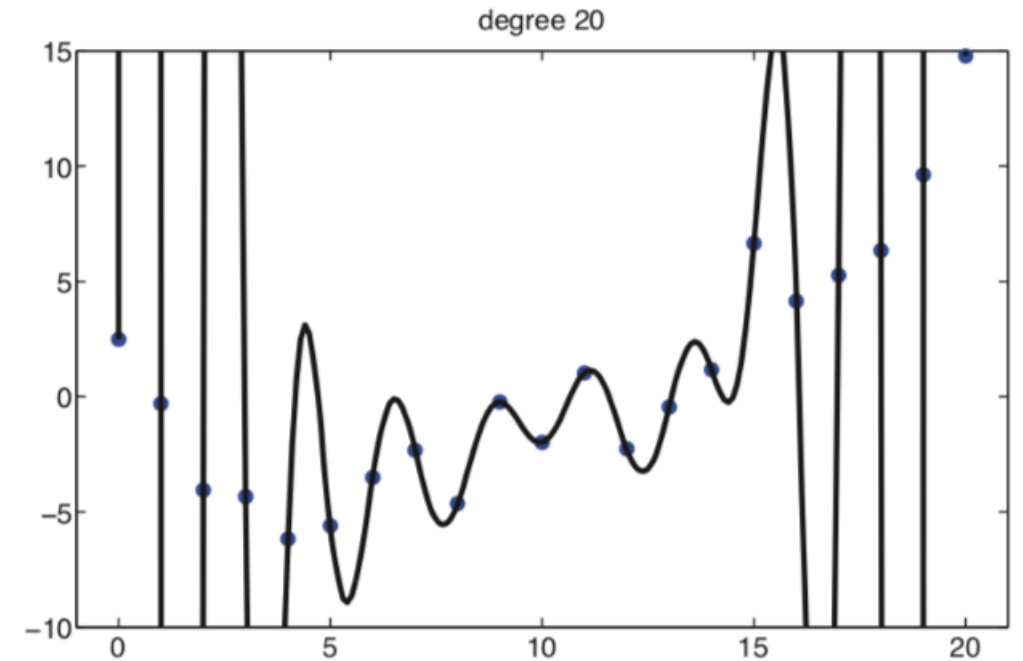$$L_{Elastic} = \frac{1}{N}(y - X\theta)^2 + \lambda_1||\theta||_1 + \lambda_2||\theta||_2$$
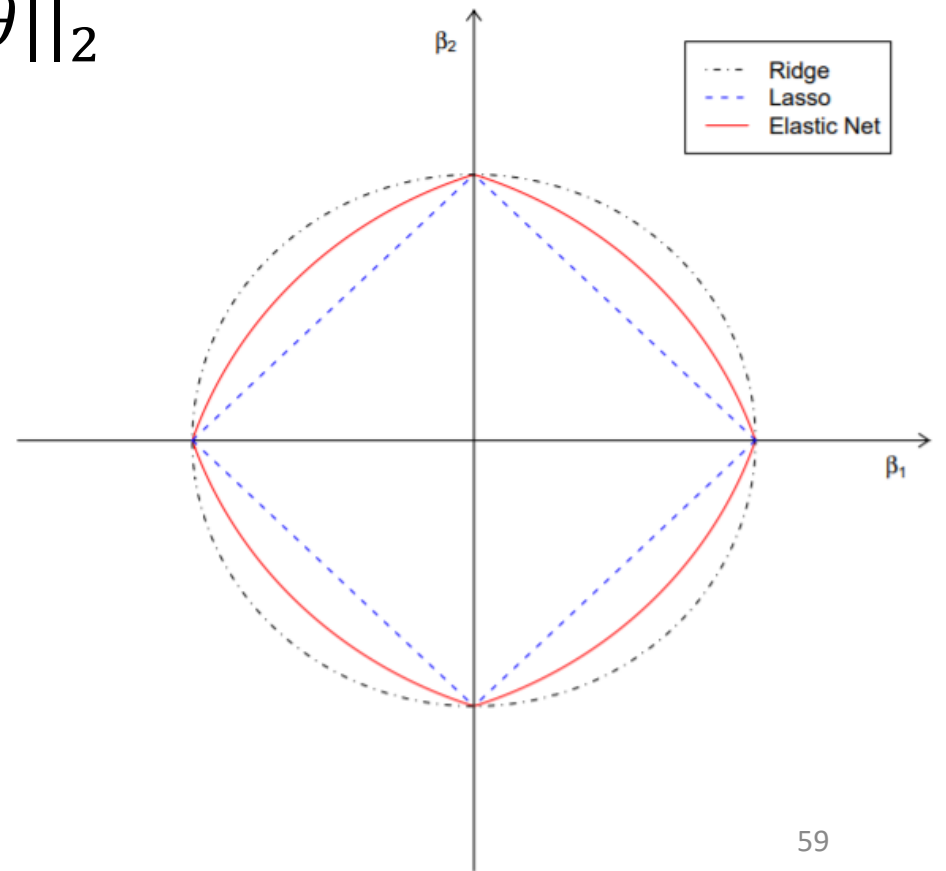
Linear regression with non-linear relationships

- E.g. $\phi(x) = (1, x, x^2, \ldots, x^d)$ and $y \sim \mathcal{N}(\theta^\top \phi(x), \sigma^2)$

(a)   (b)

Figure credit: Kevin Murphy

Summary – Linear regression

- Normal equation
- Gradient methods
- Examples
- Probabilistic view
- Applications
- Regularization

Shuai Li

https://shuaili8.github.io

Questions?

https://shuaili8.github.io/Teaching/VE445/index.html