Data X

About Me:

Polynomial Regression & Regularization

Alexander Fred Ojala

Alexander Fred Ojala
Visiting Scholar, IEOR UC Berkeley
afo@berkeley.edu

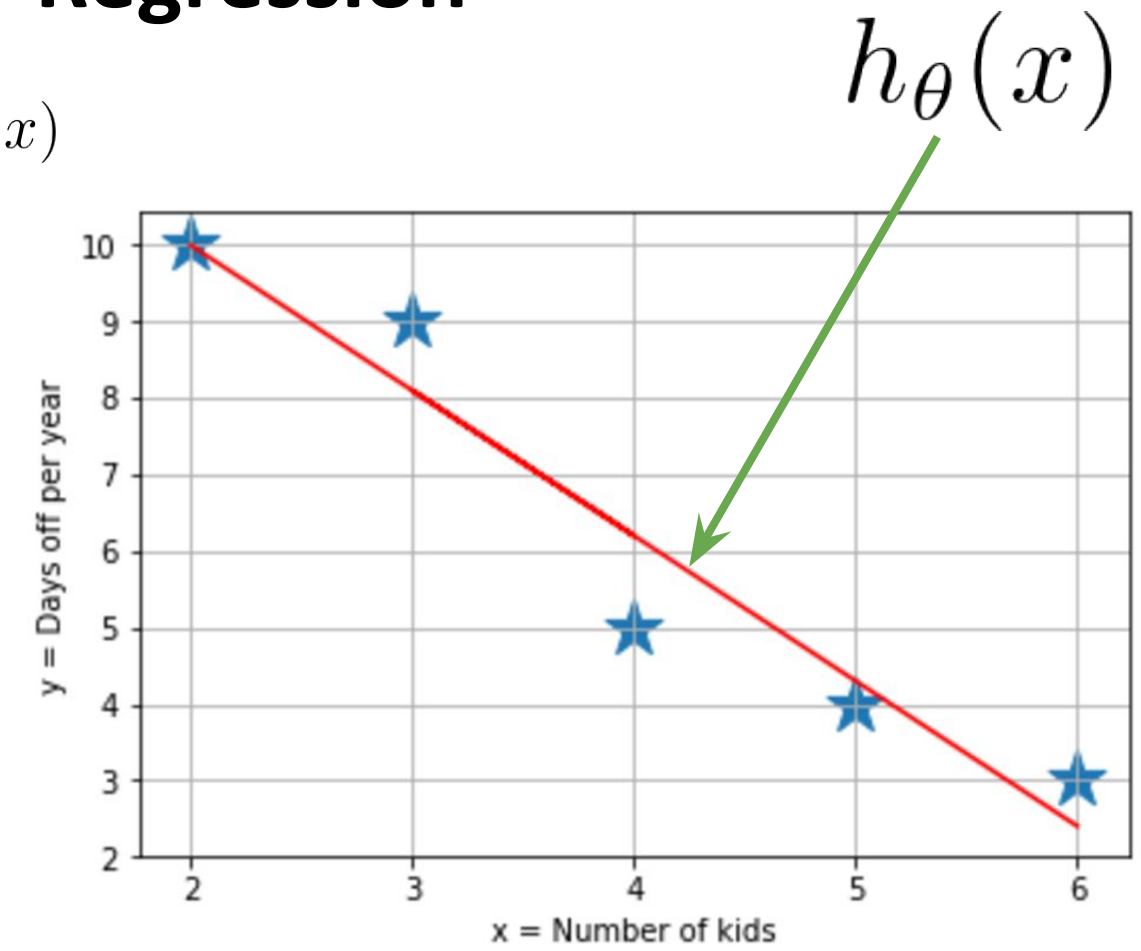
Polynomial Regression

(Simple) Linear Regression

$h_\theta(x)$

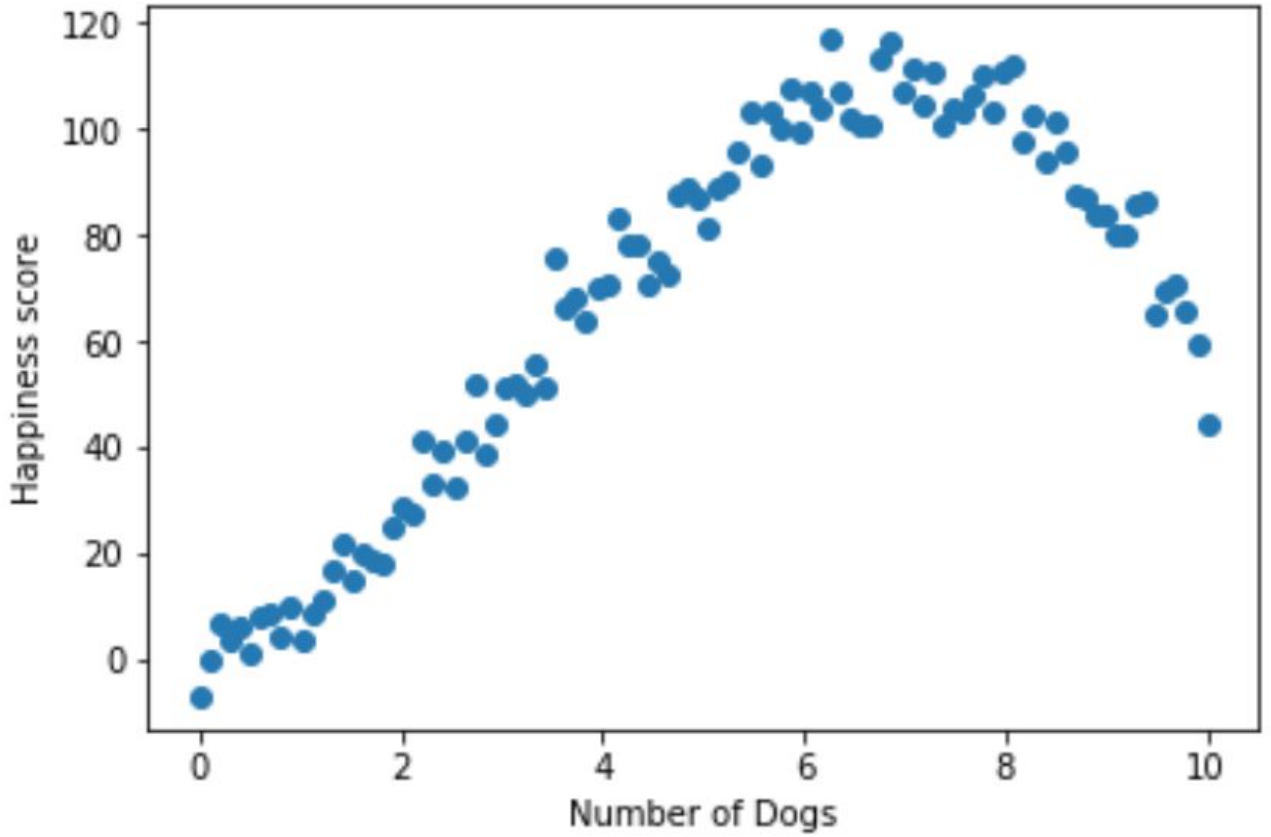
Works when have a linear relation between dependent and independent variables

$$\hat{y} = f(x, \theta) = h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x_1$$



Modeling Non-linear relationships

What if we want to model this relation?



Modeling Non-linear relationships

The best Linear Estimator

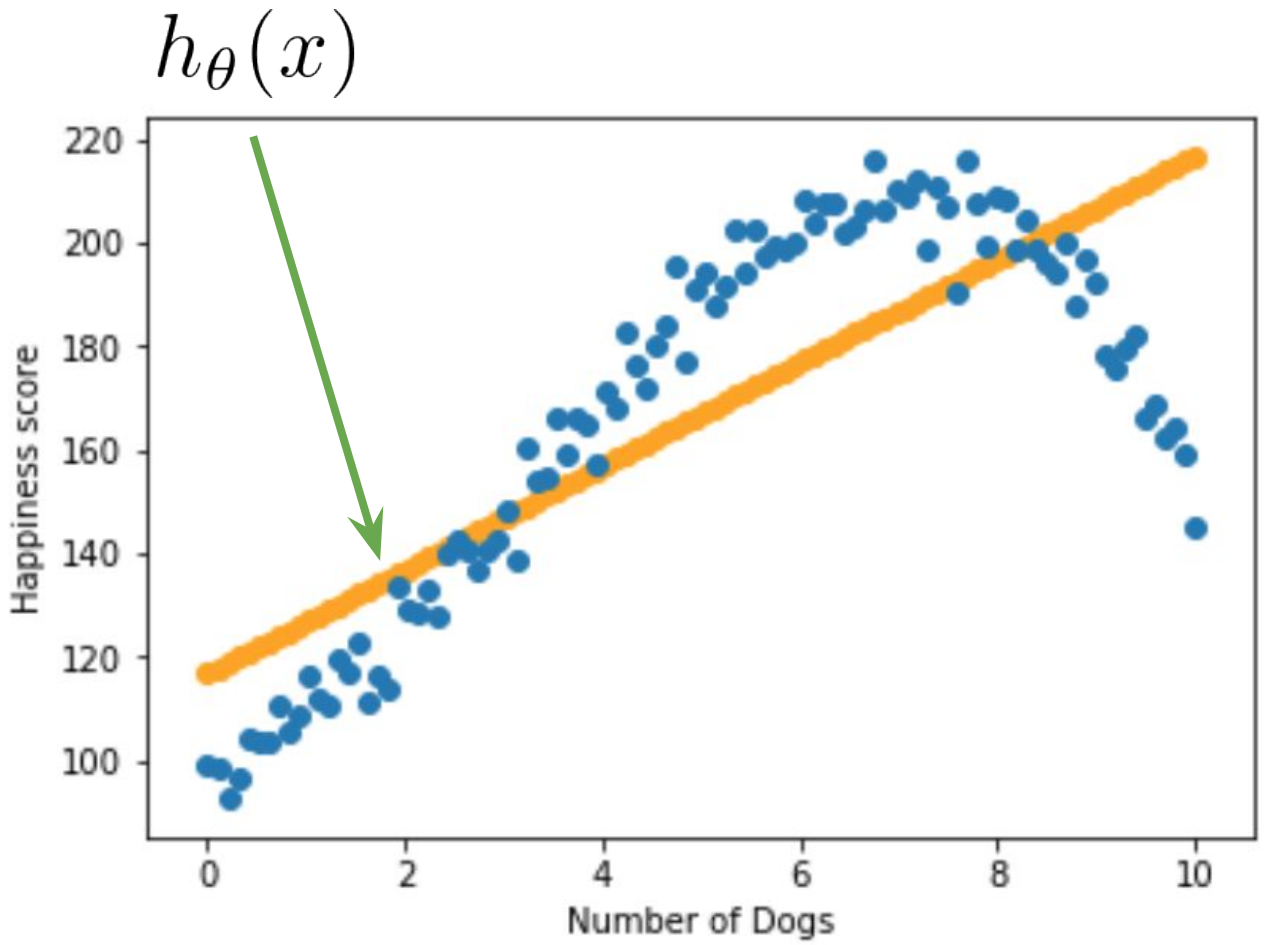
Obtained by solving the Normal Equation

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

gives us:

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x_1$$

$$\approx 117 + 10x_1$$



Polynomial Regression

Introducing (Simple) Polynomial Regression!

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_n x^n$$

- It is still a Linear Regression function we can rewrite the predictors:

x = x_1
x^2 = x_2

Find the optimal parameters by using the Normal Equations or Gradient Descent (as shown in last lecture)!

Polynomial Regression

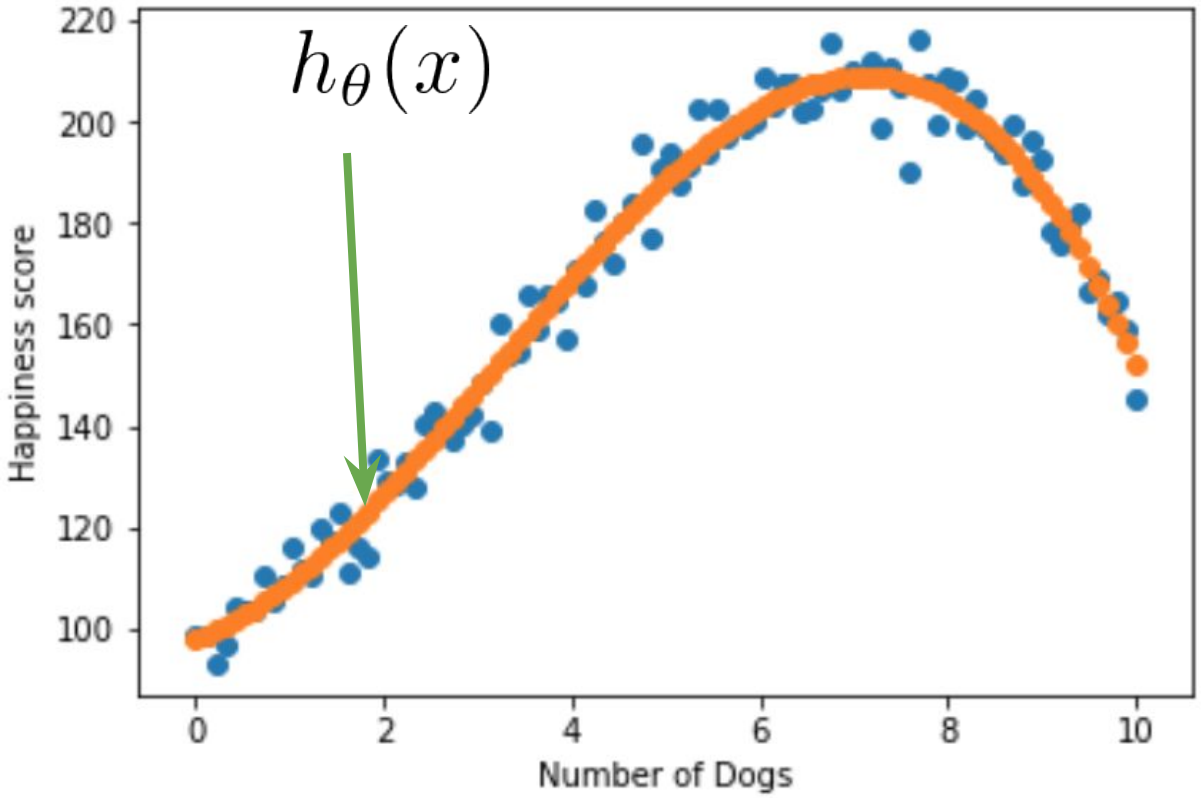
The best polynomial function for prediction (of degree 3)

Obtained by solving the Normal Equation

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

is given by:

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$
$$\approx 98 + 7x + 4.6x^2 - 0.5x^3$$



which is a much better fit to our training data!

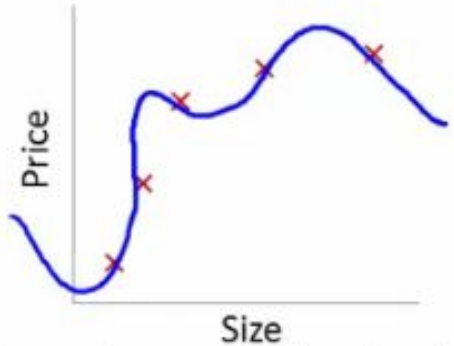
Polynomial Regression

Why don't we fit polynomial functions of very high degrees that always fit our data perfectly so that we get an error that approaches zero?

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_\infty x^\infty$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \rightarrow \mathbf{0}$$

It leads to **overfitting** (we won't predict well on new data that our model hasn't seen)

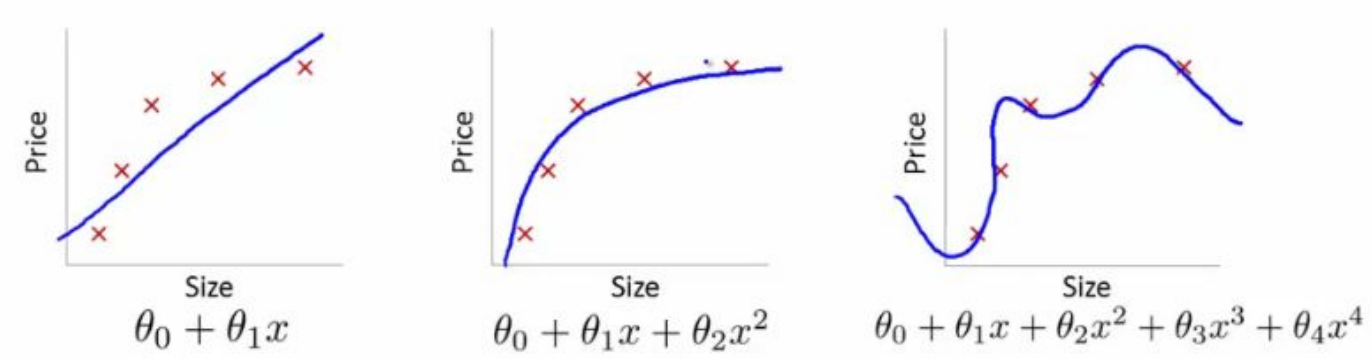


$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

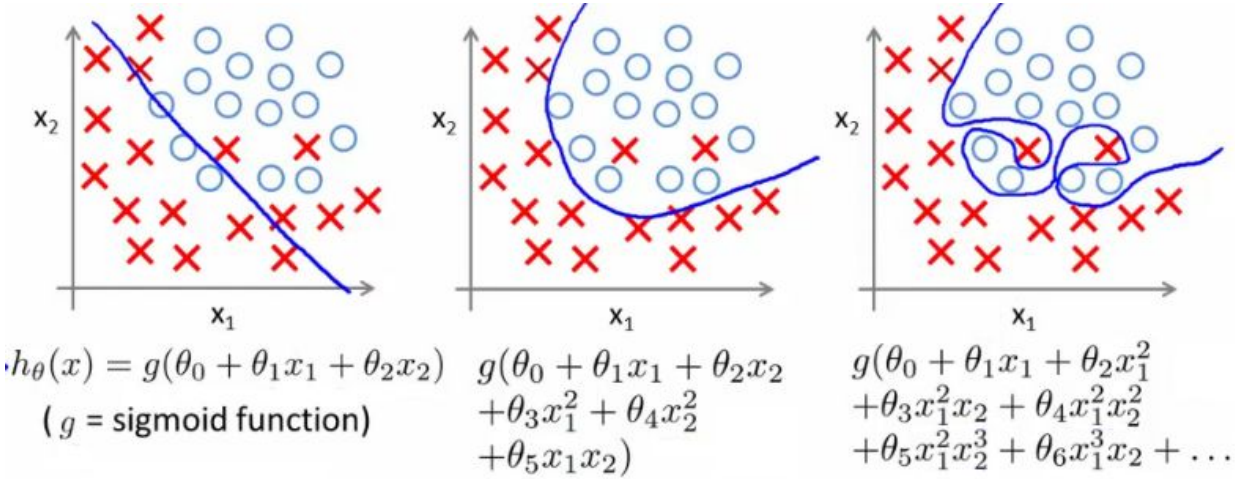
High variance
(overfit)

Bias-Variance Tradeoff:

REGRESSION CASE:



CLASSIFICATION CASE:



Bias-Variance Tradeoff:

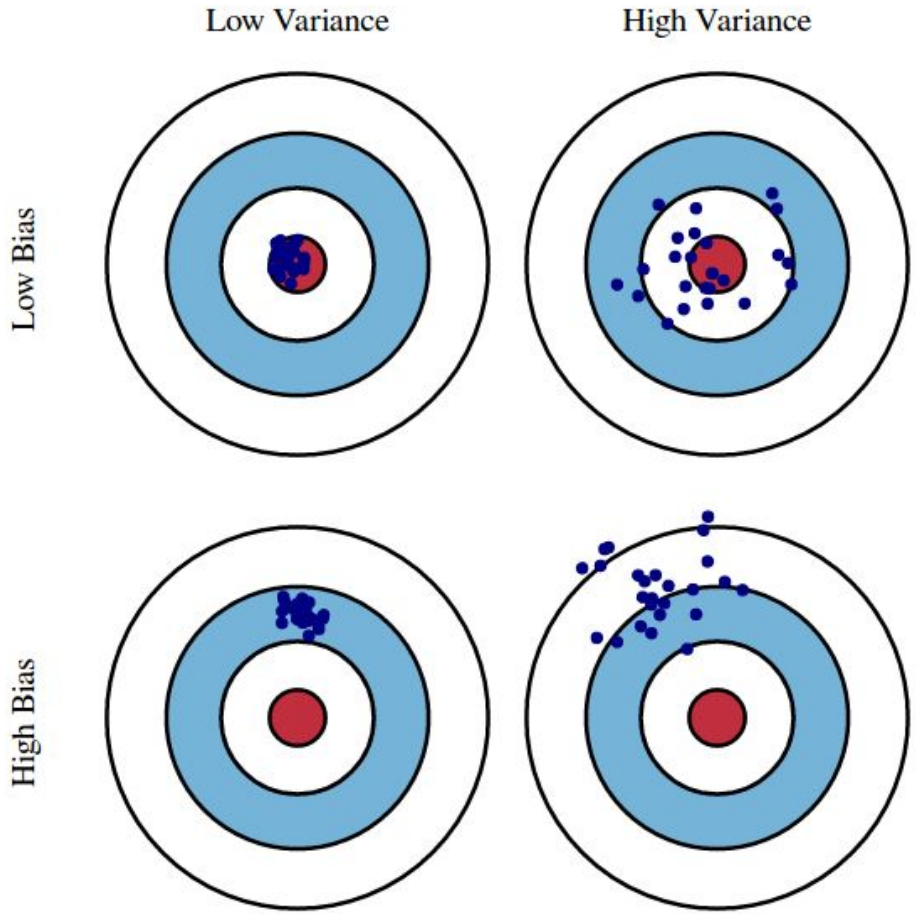


Fig. 1 Graphical illustration of bias and variance.

Regularization

Regularization

Why:

To avoid over-fitting (or performing feature selection, LASSO only)

How:

You penalize your loss function by adding a multiple of an L1 (LASSO) or an L2 (Ridge) norm of the model parameters **θ**

New loss function:

J_new(θ) = J_old(θ) + λ L_n(θ)

- λ *is a tuning parameter,*
- L_n(θ) *is the regularization norm on the parameters*

Regularization (increase error if we have too many parameters)

Non-regularized ERROR TERM: $$MSE(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

RIDGE REGRESSION (L2 NORM): $$J(\theta) = MSE(\theta) + \lambda\sum_{j}^{n}\theta_j^2$$

LASSO (L1 NORM): $$J(\theta) = MSE(\theta) + \lambda\sum_{j}^{n}|\theta_j|$$

Find optimal regularization term λ by tuning it and using Cross-validation:

- Divide your training data,
- Train your model for a fixed value of λ and test it on the remaining subsets
- Repeat this procedure while varying λ.
  Then you select the best λ that minimizes your loss function.

Regularization (increase error if we have too many parameters)

The optimal estimates of the model parameters, β, could be denoted as shown below.

This shows us the difference between Ridge and Lasso Regression

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 + \lambda \|\beta\|_1$$

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

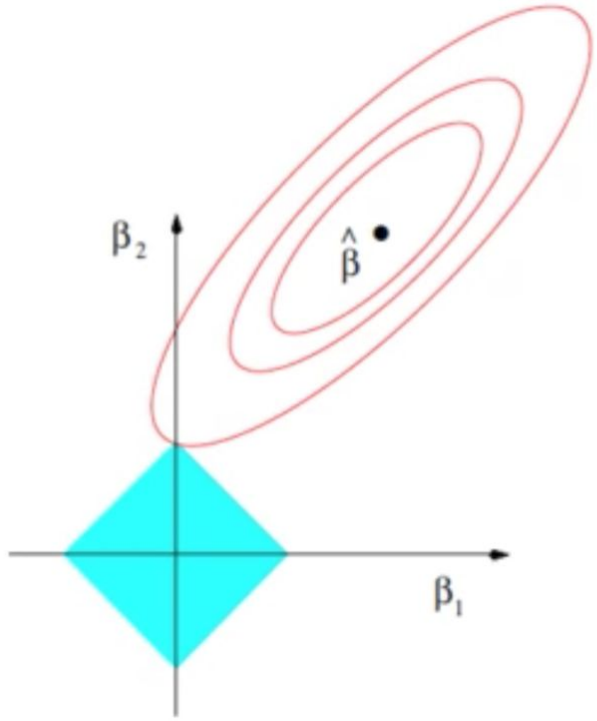
$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 \quad \text{subject to } \|\beta\|_1 \leq t$$

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_2^2 \leq t$$

Regularization (increase error if we have too many parameters)

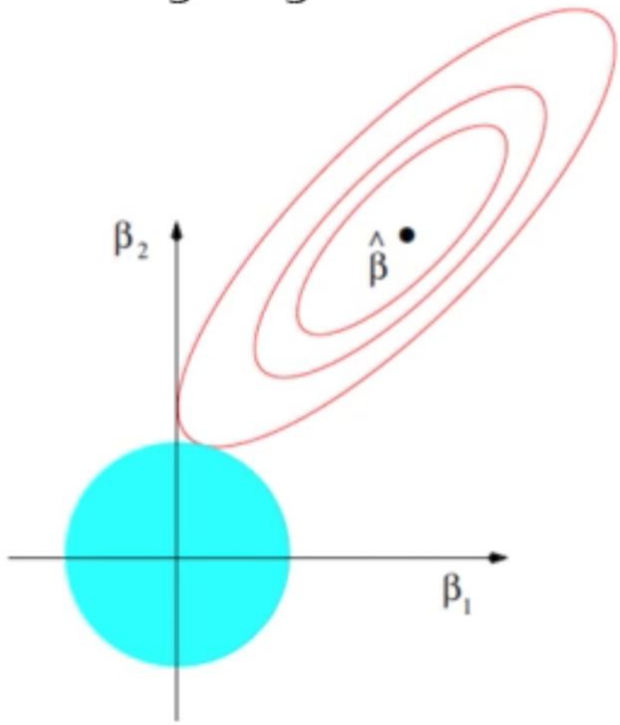
We can visualize the difference between Ridge and Lasso Regression for two parameters. Note, there is a trade-off between the optimal parameters the size of the parameters (which are constrained, to the blue areas).

Lasso Regression



$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta|_2^2 \quad \text{subject to } \|\beta\|_1 \leq t$$

Ridge Regression



$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_2^2 \leq t$$

Example Code: Regularization

End

References

- The material presented in this lecture references lecture material draws on the materials the following courses:
- Derek Kane's Data Science Tutorials: https://www.youtube.com/channel/UC33qFpcu7eHFtpZ6dp3FFXw
- Stanford – CS229 (Machine Learning) & Andrew Ng's Machine Learning at Coursera: http://cs229.stanford.edu/ & https://www.coursera.org/learn/machine-learning