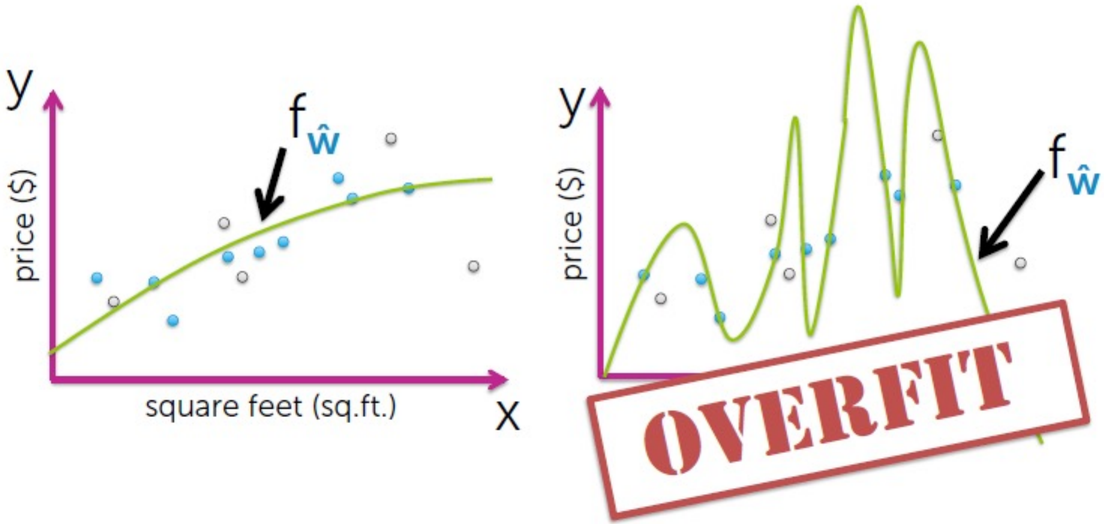
Regularization

To Address Overfitting

Flexibility of High-Order Polynomials

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p + \varepsilon_i$$



Desired Total Cost Format

We want to balance the following:

a) How well function fits the data

b) Magnitude of coefficients

Total cost

= measure of fit + measure of magnitude of coefficients

Small number indicates

= good fit to training data + reduce overfitting

Measure of Magnitude of Regression Coefficient

What summary number is indicative of size of regression coefficients?

A. Sum?
   - Positive coefficients and negative coefficients canceling effect
B. Sum of absolute value?
   - L1 norm; Lasso Regression (also known as L1 regularization)
C. Sum of squared value?
   - L2 norm; Ridge Regression (also known as L2 regularization)
D. Use both L1 norm and L2 norm?
   - Elastic Net

Ridge Regression: Part I

$$y = b + \sum w_i x_i$$

The functions with smaller w_i are better

$$L = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i \right) \right)^2 \boxed{+\lambda \sum (w_i)^2}$$

- Smaller w_i means … Smoother

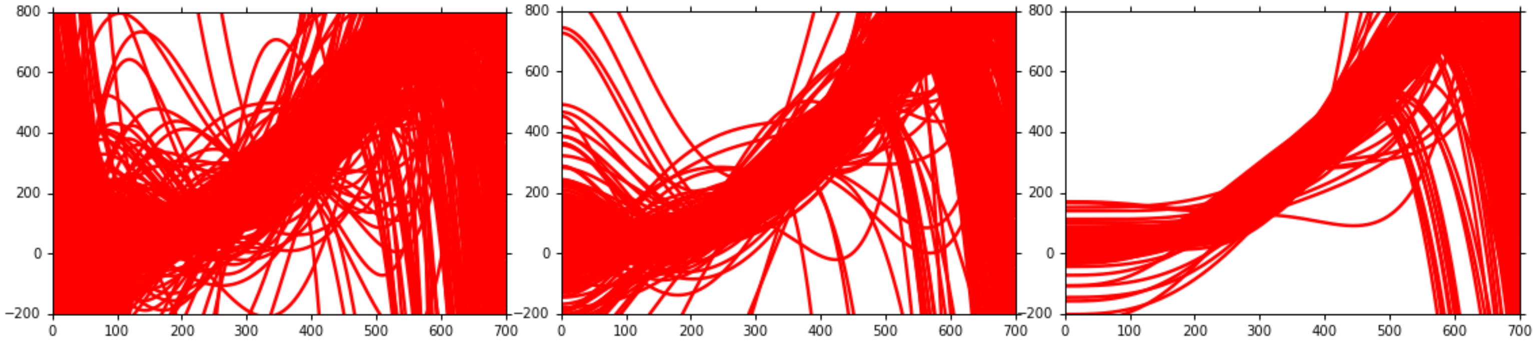
$$y = b + \sum w_i x_i$$

$$y + \sum w_i \Delta x_i = b + \sum w_i (x_i + \Delta x_i)$$

- We believe smoother function is more likely to be correct
  Do you have to apply regularization on bias?

What to do With Large Variance?

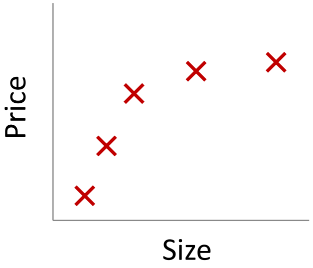
Regularization → May increase bias



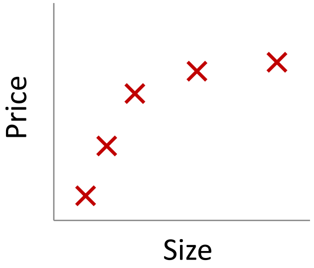
Ridge Regression: Part II

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

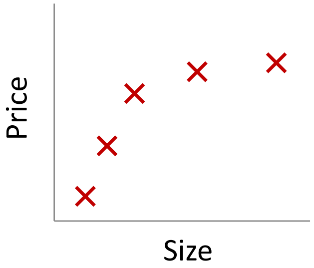
$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$



Large λ
High bias (underfit)
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$
$h_\theta(x) \approx \theta_0$



Intermediate λ
“Just right”



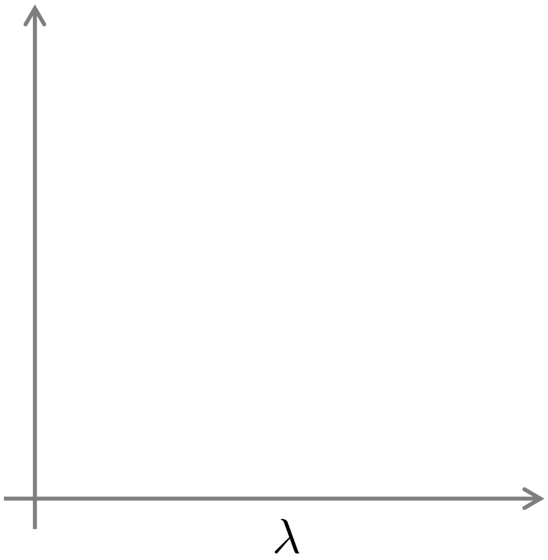
Small λ
High variance (overfit)

Bias/Variance as a Function of the Regularization Parameter

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

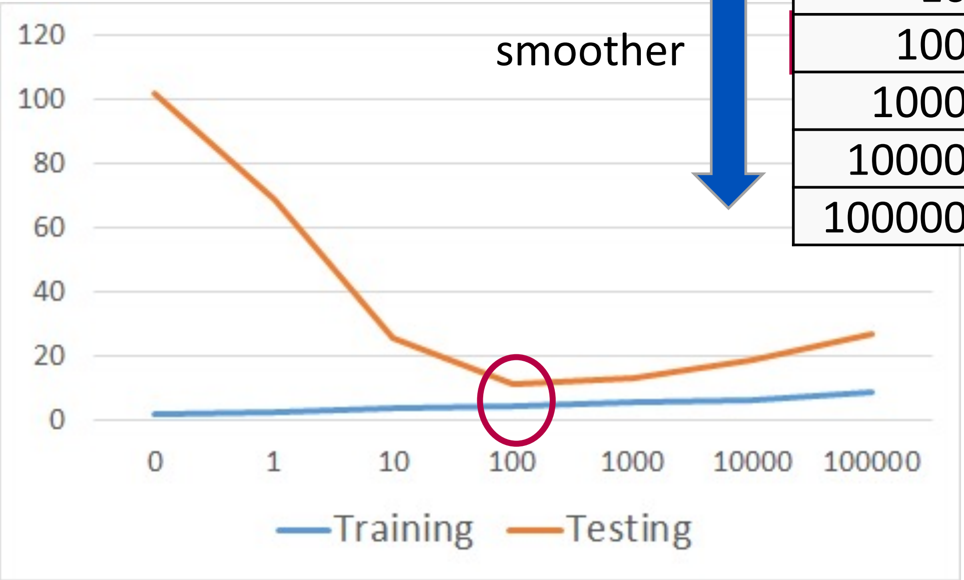
$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$



Regularization Outline

λ	Training	Testing
0	1.9	102.3
1	2.3	68.7
10	3.5	25.7
100	4.1	11.1
1000	5.6	12.8
10000	6.3	18.7
100000	8.5	26.8

smoother

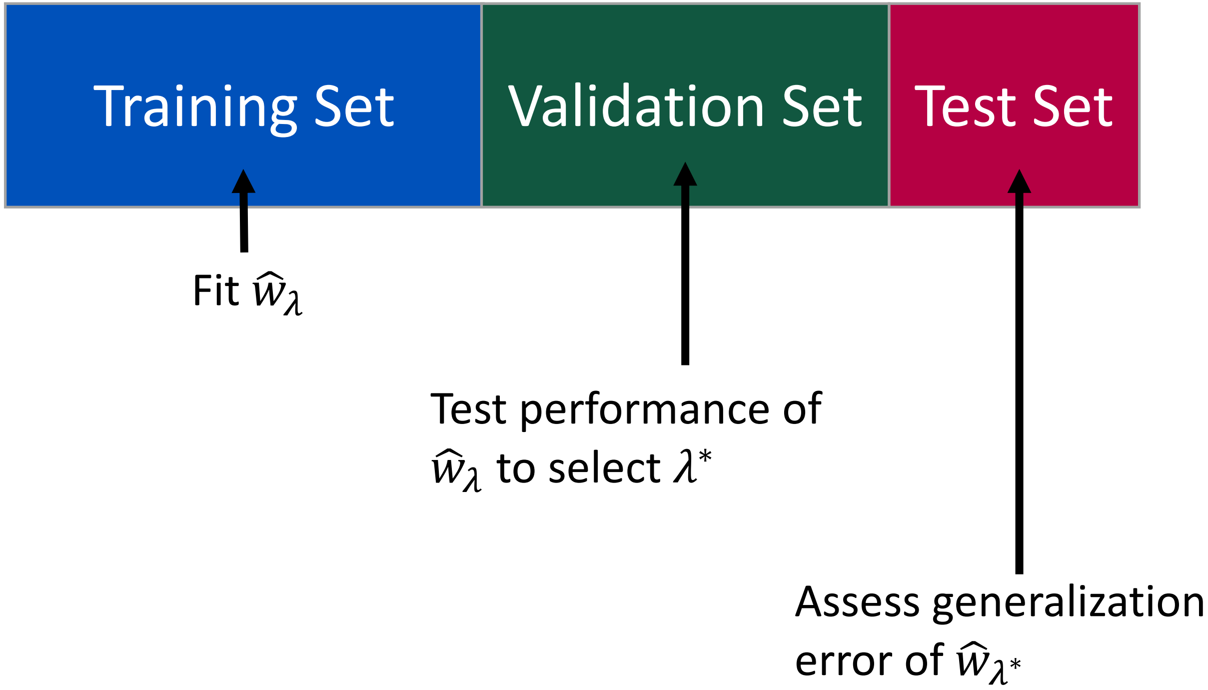


How smooth?

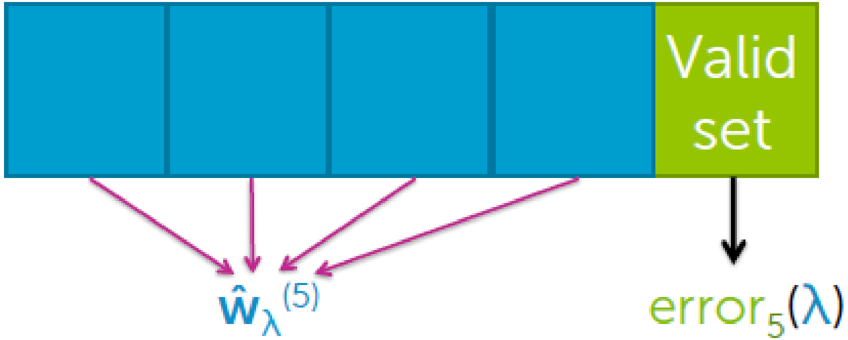
Select λ obtaining the best model

- Training error: Larger λ, considering the training error less
- We prefer smooth function, but don't be too smooth.

How to Choose Regularization Hyperparameter λ: Part I



How to Choose Regularization Hyperparameter λ: Part II



For k=1,…,K

1. Estimate $\hat{\mathbf{w}}_\lambda^{(k)}$ on the training blocks
2. Compute error on validation block: $\text{error}_k(\lambda)$

Compute average error: $CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} \text{error}_k(\lambda)$

Lasso Regression for Feature Selection: Part I

Lot size
Single Family
Year built
Last sold price
Last sale price/sqft
Finished sqft
Unfinished sqft
Finished basement sqft
floors
Flooring types
Parking type
Parking amount
Cooling
Heating
Exterior materials
Roof type
Structure style

Dishwasher
Garbage disposal
Microwave
Range / Oven
Refrigerator
Washer
Dryer
Laundry location
Heating type
Jetted Tub
Deck
Fenced Yard
Lawn
Garden
Sprinkler System
⋮

Lasso Regression for Feature Selection: Part II

$$y = b + \sum w_i x_i$$

The functions with smaller w_i are better

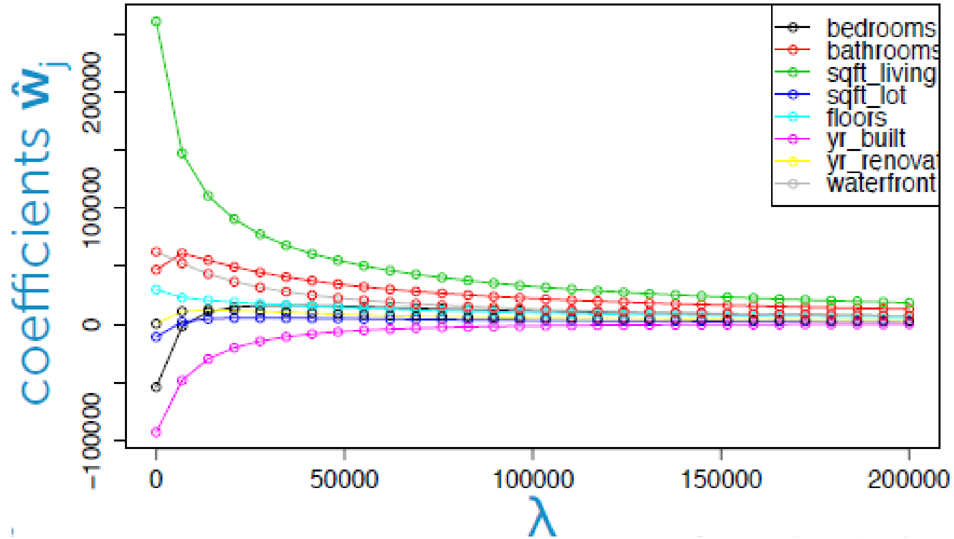
$$L = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i \right) \right)^2 \boxed{+\lambda \sum |w_i|}$$

Impact of L2 Regularization Hyperparameter

$\hat{\mathbf{w}}$ selected to minimize

tuning parameter = balance of fit and magnitude

$$\text{RSS}(\mathbf{w}) + \lambda||\mathbf{w}||_2^2$$

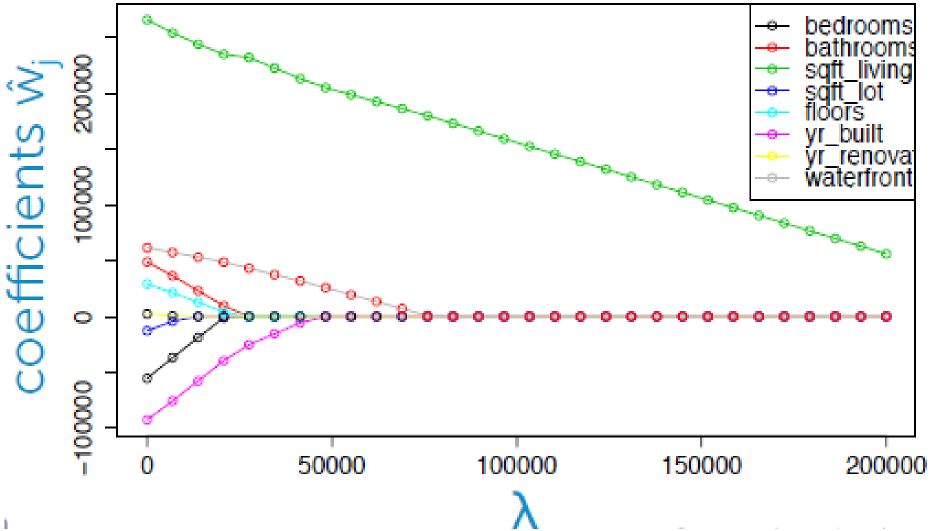


Impact of L1 Regularization Hyperparameter

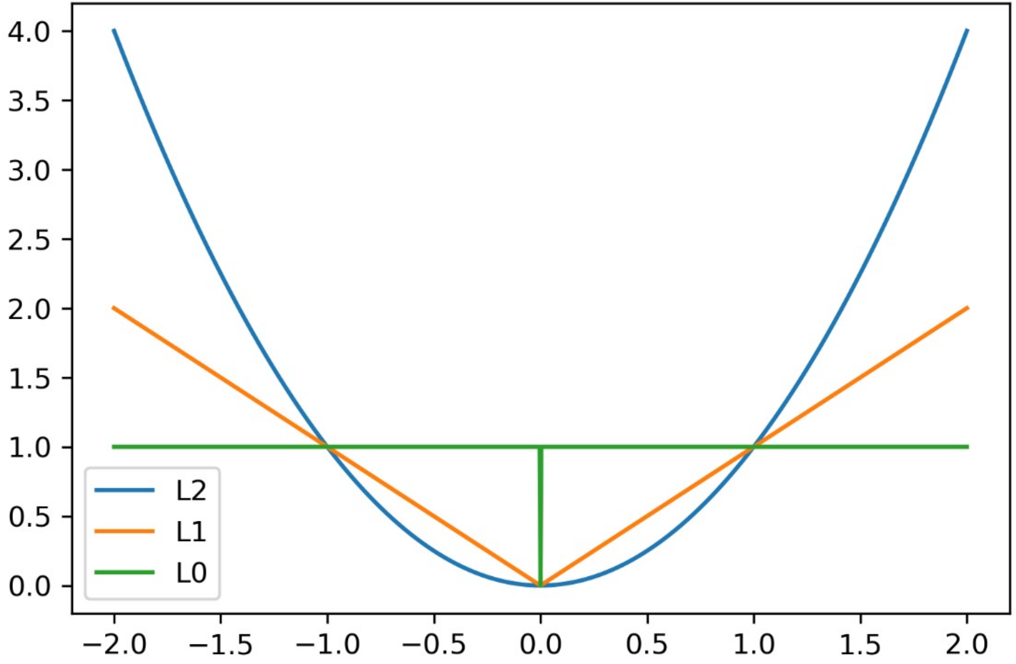
$\hat{\mathbf{w}}$ selected to minimize

tuning parameter = balance of fit and sparsity

$$RSS(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$$



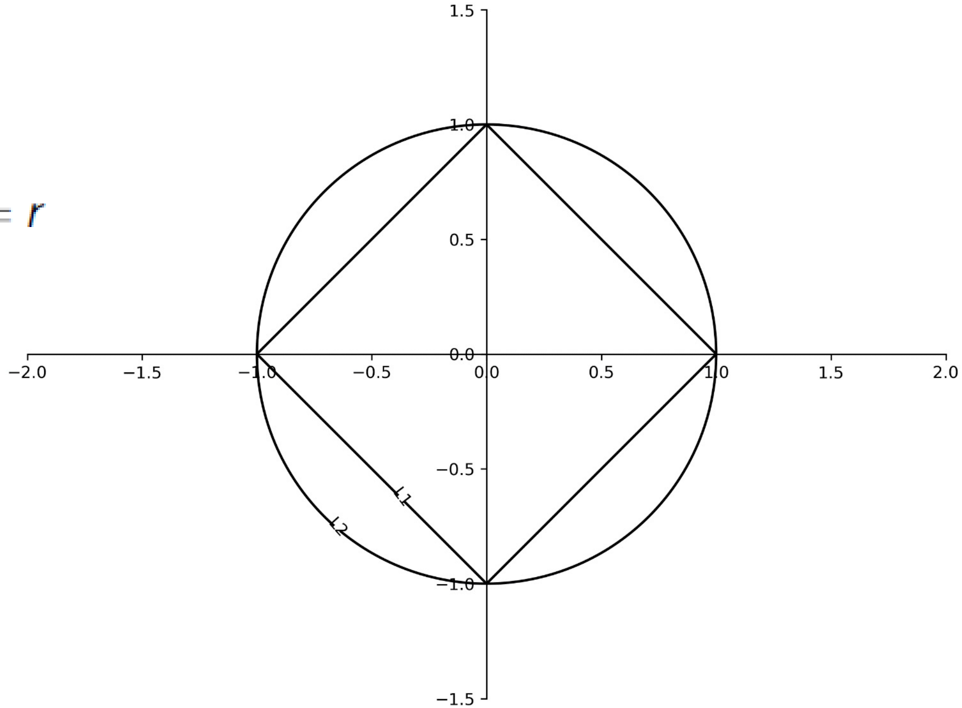
Comparison Between L1 and L2 Regularization: Part I



Comparison Between L1 and L2 Contour

L1 Contour

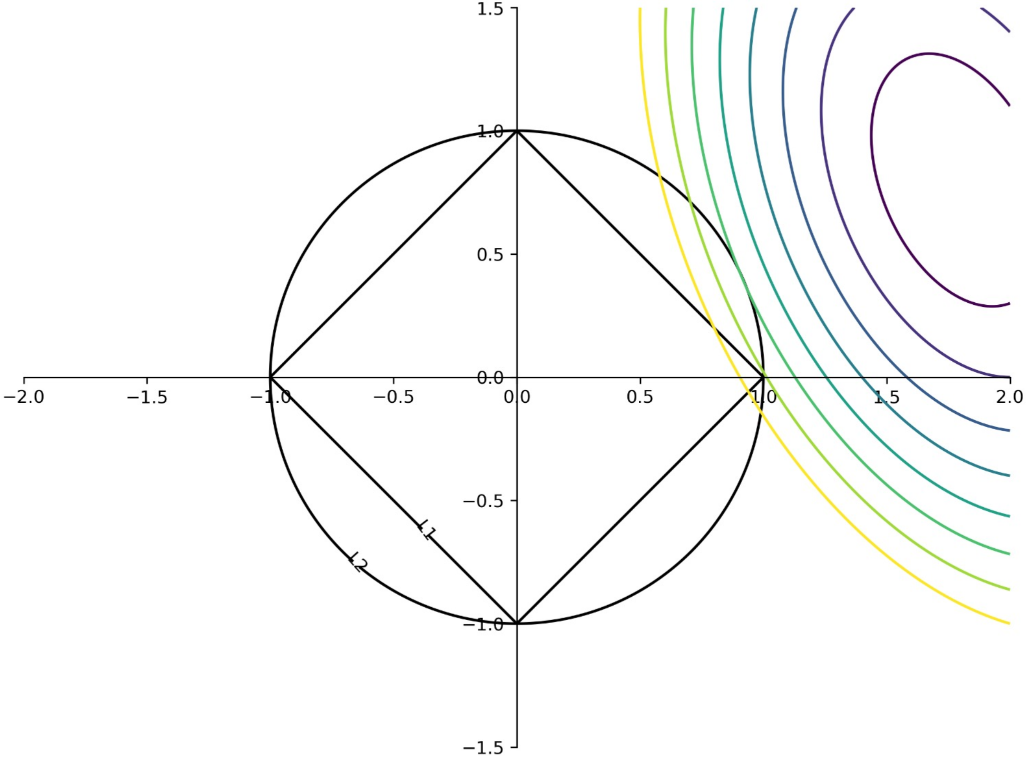
$$|w_1| + |w_2| = r$$



L2 Contour

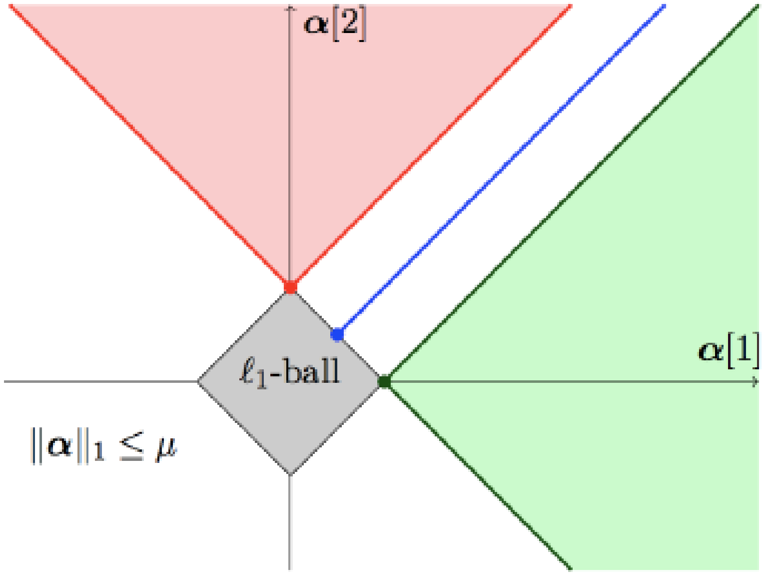
$$w_1^2 + w_2^2 = r$$

Comparison Between L1 and L2 Regularization: Part II

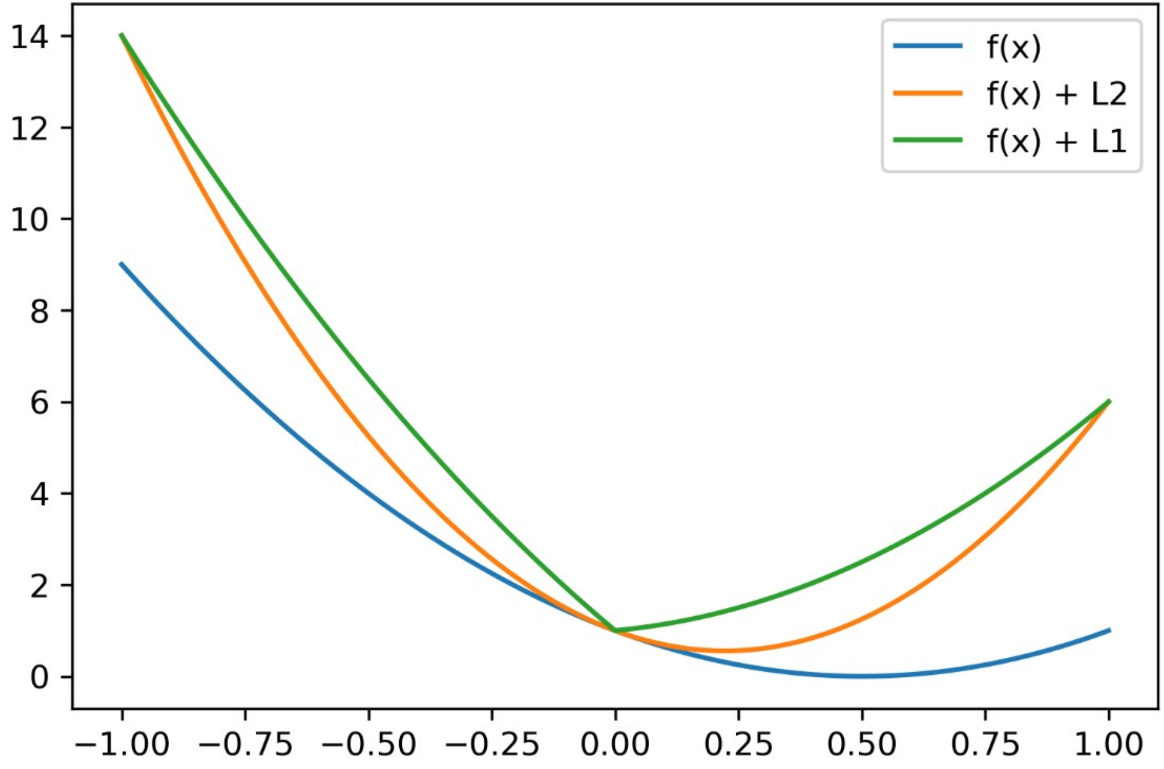


Why are Lasso Solutions Often Sparse?

- If the features are orthogonal, then the loss function contours are circles.
- The OLS solution in green or red regions implies L1 constrained solution will be at corner.



Comparison Between L1 and L2 Regularization: Part III



Elastic Net (Combination of L1 and L2)