CS5489
Lecture 5.2: Regularized Linear Regression: Ridge and Lasso Regression

Kede Ma

City University of Hong Kong (Dongguan)

Slide template by courtesy of Benjamin M. Marlin

Outline

Regularized Linear Regression

- Why adding regularization?
  - Control the parameter space to avoid **overfitting**
  - Obtain numerically more stable solutions
    - E.g., when $\mathbf{X}^T\mathbf{X}$ is not invertible
  - Enforce the desired parameter space as a prior
    - E.g., select a subset of features that are good for prediction by enforcing the weights of irrelevant features to zero
- What is the price that we pay?
  - The regularization term will bias the least squares estimate

Ridge Regression

- Add regularization term to OLS:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \frac{\alpha}{2}\|\mathbf{w}\|_2^2$$

- The first term is the data-fit term
  - Sum-squared error of the predictions
- The second term is the regularization term
  - $\|\mathbf{w}\|_2^2 = \sum_{j=1}^{N} w_j^2$ penalizes large weights
  - α is the hyperparameter that controls the amount of "shrinkage"
    - Larger α means more shrinkage
    - $\alpha = 0$ reduces to OLS
- Has a probabilistic interpretation:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}; \sigma^2, \alpha) \propto p(\mathbf{y}|\mathbf{w}, \mathbf{X}; \sigma^2) p(\mathbf{w}; \alpha)$$

But, Why?

- Why should we penalize large weights in the first place?
- Equivalently, why smaller weights (i.e., weights that close to zero) are more preferred than larger weights?
- **Reason 1**: Smaller weights are more robust to perturbations of input features
  - Consider $f_1(x) = 2x + 0.1$ and $f_2(x) = 100x - 98$
  - If we add a small perturbation to $x \rightarrow x + 0.1$
  - Then, $\Delta f_1(x) = 0.2$ and $\Delta f_2(x) = 10$
- **Reason 2**: There are better chances to zero out some input features x that are redundant or uninformative, leading to more accurate prediction of y

Ridge Regression

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \frac{\alpha}{2}\|\mathbf{w}\|_2^2$$

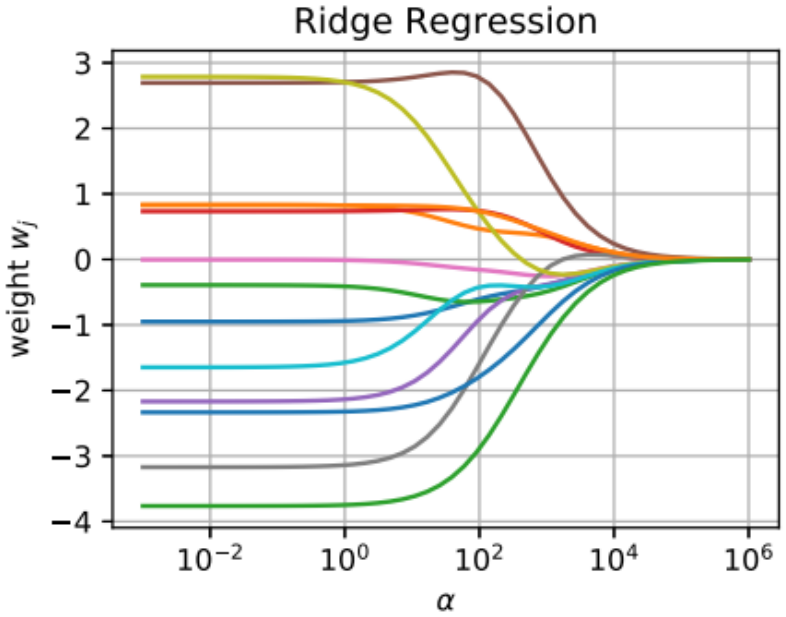
- Has a closed-form solution:

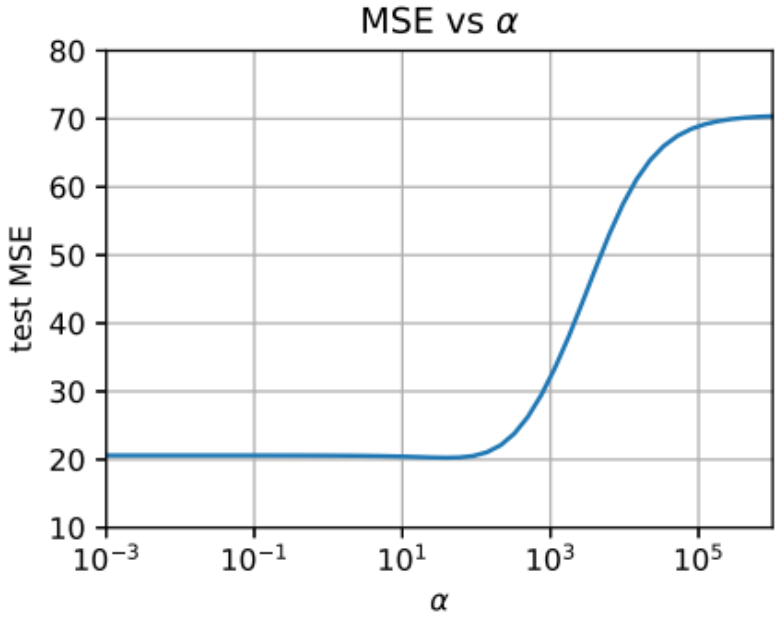
$$\mathbf{w}^\star = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- The term "ridge regression" comes from the closed-form solution, where a "ridge" is added to the diagonal of $\mathbf{X}^T\mathbf{X}$

Example on Boston Data

- The Boston housing data actually has 13 features
- Vary α from 10^{-3} (little shrinkage) to 10^{6} (lots of shrinkage)
  - For small α, all weights are non-zero
  - For large α, all weights shrink to zero
  - Somewhere in between lies the best model... (selected using cross-validation)





Interpretation

```
weight : feature description
-3.613 : LSTAT   % lower status of the population
-2.854 : DIS     weighted distances to five Boston employment centres
 2.783 : RM      average number of rooms per dwelling
-2.223 : PTRATIO pupil-teacher ratio by town
 2.105 : RAD     index of accessibility to radial highways
-1.856 : NOX     nitric oxides concentration (parts per 10 million)
-1.097 : TAX     full-value property-tax rate per $10,000
-0.854 : CRIM    per capita crime rate by town
 0.816 : B       1000(Bk - 0.63)^2 where Bk is the proportion of blacks by t
own
 0.756 : CHAS    Charles River dummy variable (= 1 if tract bounds river; 0
otherwise)
 0.714 : ZN      proportion of residential land zoned for lots over 25,000 s
q.ft.
-0.540 : INDUS   proportion of non-retail business acres per town
-0.061 : AGE     proportion of owner-occupied units built prior to 1940
```

- Which weights are most important?
  - Negative weights indicate factors that decrease the house price
    - Examples: LSTAT (having higher percentage of lower status population), DIS (distance to business areas), PTRATIO (higher student-teacher ratio)
  - Positive weights indicate factors that increase the house price
    - Examples: RM (having more rooms), RAD (proximity to highways)

Outline

Better Shrinkage

- With ridge regression, some weights are small but still non-zero
  - These are less important, but somehow still necessary
- To get better shrinkage to zero, we can change the regularization term to encourage more weights to be zero

Lasso Regression

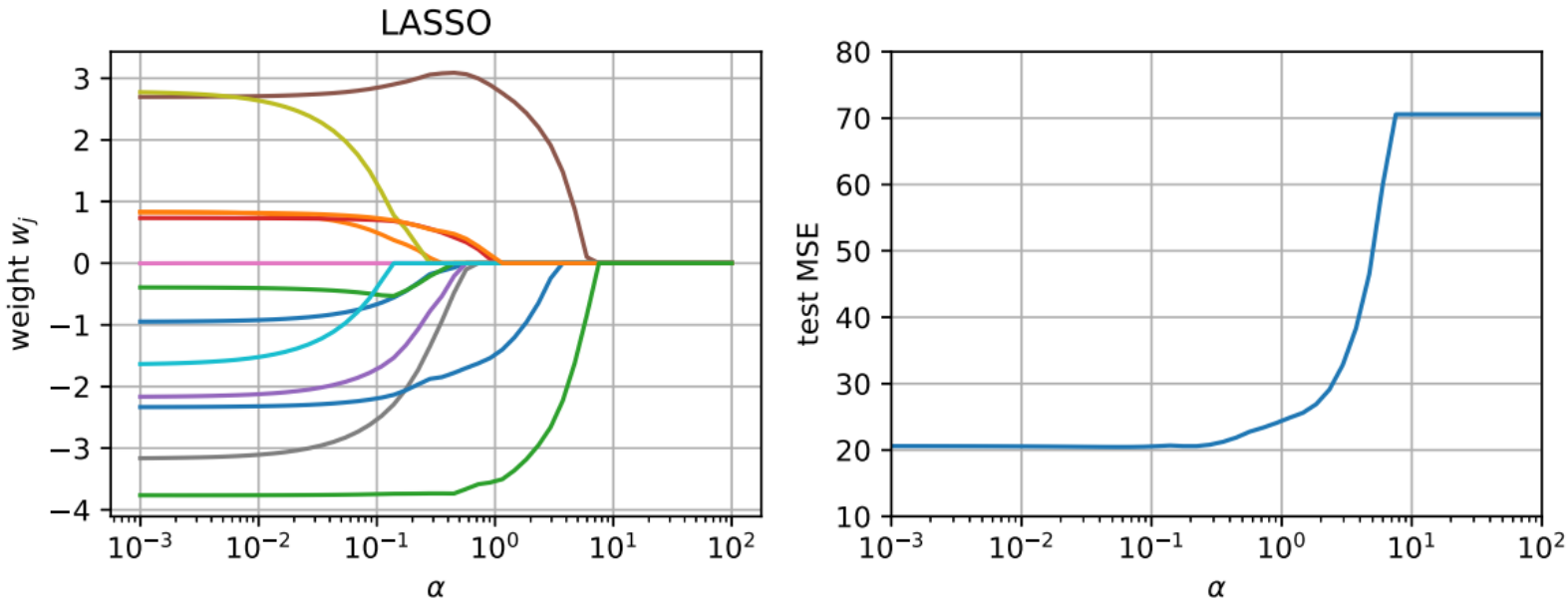
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \alpha\|\mathbf{w}\|_1$$

- LASSO = "least absolute shrinkage and selection operator"
- Keep the same data fit term, but change the regularization term:
  - Sum of absolute weight values: $\|\mathbf{w}\|_1 = \sum_{j=1}^{N} |w_j|$
  - When a weight is close to zero, the regularization term will force it to be equal to zero
- It is a convex optimization problem with no closed-form solution in general. However, it can be solved efficiently using an algorithm called **least angle regression**

`https://en.wikipedia.org/wiki/Least-angle_regression`

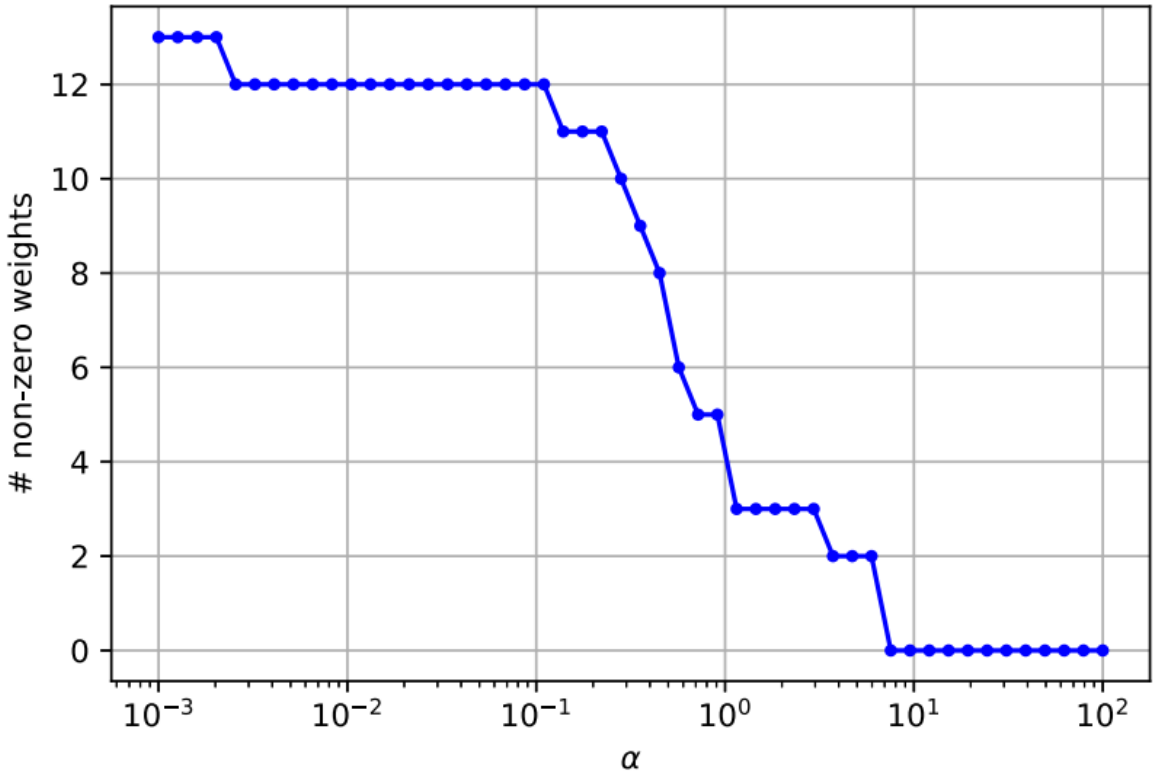
Example on Boston Data

- Vary α from 10^{-3} to 10^{2}



Feature Selection

- Select α to obtain a given number of features



Interpretation

```
weight : feature description
-3.585 : LSTAT   % lower status of the population
 2.996 : RM      average number of rooms per dwelling
-1.625 : PTRATIO pupil-teacher ratio by town
 0.275 : B       1000(Bk - 0.63)^2 where Bk is the proportion of blacks by t
own
 0.225 : CHAS    Charles River dummy variable (= 1 if tract bounds river; 0
otherwise)
-0.000 : TAX     full-value property-tax rate per $10,000
-0.000 : RAD     index of accessibility to radial highways
-0.000 : DIS     weighted distances to five Boston employment centres
-0.000 : AGE     proportion of owner-occupied units built prior to 1940
-0.000 : NOX     nitric oxides concentration (parts per 10 million)
-0.000 : INDUS   proportion of non-retail business acres per town
 0.000 : ZN      proportion of residential land zoned for lots over 25,000 s
q.ft.
-0.000 : CRIM    per capita crime rate by town
```

- Weights for unimportant features are set to zero
  - TAX, RAD, DIS, AGE, ...
- Important features have non-zero weights
  - LSTAT, RM, PTRATIO, B, CHAS

Why Shrinkage?

- Under the orthogonal design ($\mathbf{X}^T\mathbf{X} = \mathbf{I}$)
  - Linear regression: $\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$

$$\mathbf{w}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}$$

  - Ridge regression: $\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \frac{\alpha}{2}\|\mathbf{w}\|_2^2$

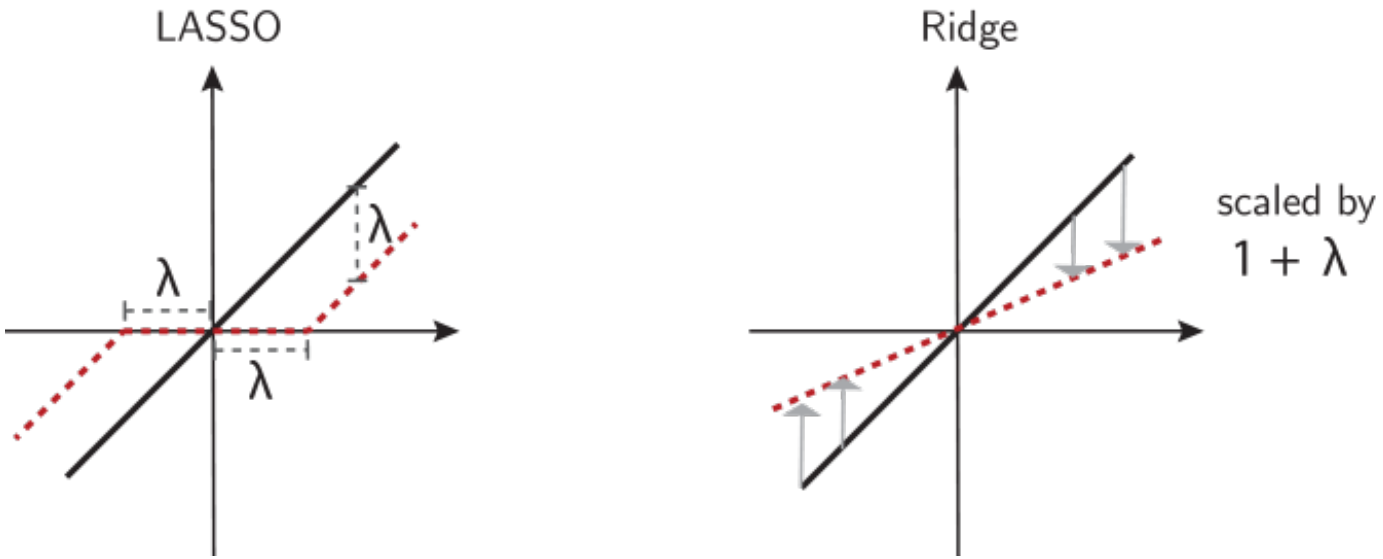
$$\mathbf{w}_{\text{RR}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}/(1+\alpha) = \mathbf{w}_{\text{LS}}/(1+\alpha)$$

  - Lasso regression: $\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \alpha\|\mathbf{w}\|_1$

$$\mathbf{w}_{\text{LR}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2+\alpha\|\mathbf{w}\|_1 = \text{sign}(\mathbf{w}_{\text{LS}})\max(0, |\mathbf{w}_{\text{LS}}|-\alpha))$$

Ridge vs Lasso

- Ridge regression: $\mathbf{w}_{\text{RR}} = \mathbf{w}_{\text{LS}}/(1+\lambda)$
- Lasso regression: $\mathbf{w}_{\text{LR}} = \text{sign}(\mathbf{w}_{\text{LS}})\max(0, |\mathbf{w}_{\text{LS}}| - \lambda))$



Linear Regression Summary

- OLS needs at least $M \geq N + 1$ data cases to learn a model with an N dimensional feature vector
- Ridge and Lasso work when $\mathbf{X}^T\mathbf{X}$ is close to singular
  - E.g., caused by co-linear features ($x_i \approx ax_j + b$)
- MSE objective function for OLS, Ridge, and Lasso is sensitive to noise and **outliers**
  - Regularization (Ridge and Lasso) can prevent very large weights, and reduce the possibility of overfitting to outliers
- All fail when output y **non-linearly** depends on input features $\mathbf{x}$