ECE595 / STAT598: Machine Learning I

Lecture 2.1: Regularization - Ridge Regression

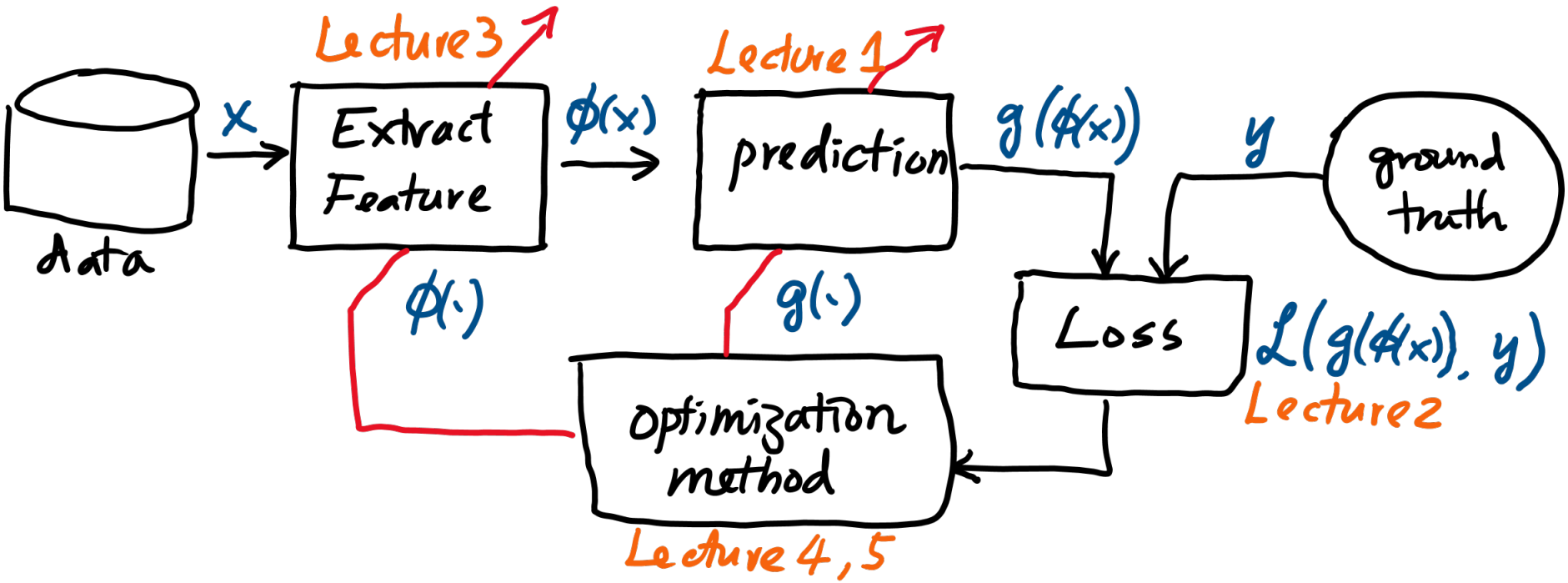
Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

Outline



Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

Lecture 2: Regularization

- Ridge Regression
  - Regularization
  - Parameter
- LASSO Regression
  - Sparsity
  - Algorithm
  - Application

Ridge Regression

- Applies to both over and under determined systems.
- The loss function of the ridge regression is defined as

$$J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

- $\|\boldsymbol{\theta}\|^2$ Regularization function
- λ: Regularization parameter
- The solution of the ridge regression is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}}\Big\{\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2\Big\} \\ &= 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}) + 2\lambda\boldsymbol{\theta} = \boldsymbol{0},\end{aligned}$$

  which gives us $\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$.
- Probabilistic interpretation: See Appendix.

Change in Eigen-values

Ridge regression improves the eigen-values:

- Eigen-decomposition of $\boldsymbol{A}^T\boldsymbol{A}$:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T \succeq 0,$$

  where $\boldsymbol{U}$ = eigen-vector matrix, $\boldsymbol{S}$ = eigen-value matrix.

- $\boldsymbol{S}$ is a diagonal matrix with non-negative entries:

$$\boldsymbol{S} = \begin{bmatrix} \clubsuit & & & \\ & \clubsuit & & \\ & & \clubsuit & \\ & & & 0 \end{bmatrix}$$

  See Tutorial on "Linear Algebra".

- Therefore, $\boldsymbol{S} + \lambda\boldsymbol{I}$ is always positive for any $\lambda > 0$, implying that

$$\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I} = \boldsymbol{U}(\boldsymbol{S} + \lambda\boldsymbol{I})\boldsymbol{U}^T \succ 0.$$

Regularization Parameter λ

- The solution of the ridge regression is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

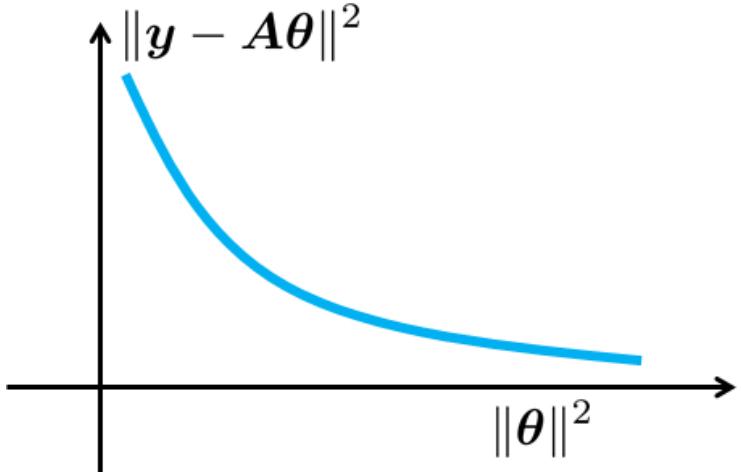
- If $\lambda \to 0$, then $\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$:

$$J(\boldsymbol{\theta}) = \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \cancel{\lambda\|\boldsymbol{\theta}\|^2}.$$

- If $\lambda \to \infty$, then $\widehat{\boldsymbol{\theta}} = \mathbf{0}$:

$$J(\boldsymbol{\theta}) = \cancel{\|\boldsymbol{A\theta} - \boldsymbol{y}\|^2} + \lambda\|\boldsymbol{\theta}\|^2.$$

- There is a trade-off curve between the two terms by varying λ.



Comparing Vanilla and Ridge

Suppose $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\theta}^* + \boldsymbol{e}$ for some ground truth $\boldsymbol{\theta}^*$ and noise vector $\boldsymbol{e}$. Then, the **vanilla linear regression** will give us

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y} \\ &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^* + \boldsymbol{e}) \\ &= \boldsymbol{\theta}^* + (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{e}\end{aligned}$$

If $\boldsymbol{e}$ has zero mean and variance σ^2, we can show that

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\theta}}] &= \boldsymbol{\theta}^*, \\ \text{Cov}[\widehat{\boldsymbol{\theta}}] &= \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}.\end{aligned}$$

Therefore, the regression coefficients are unbiased but have large variance. We can further show that the mean-squared error (MSE) is

$$\text{MSE}(\widehat{\boldsymbol{\theta}}) = \sigma^2\text{Tr}\big\{(\boldsymbol{A}^T\boldsymbol{A})^{-1}\big\}.$$

Comparing Vanilla and Ridge

On the other hand, if we use ridge regression, then

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(\lambda) &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^* + \boldsymbol{e}) \\ &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^* + (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{e}.\end{aligned}$$

Again, if $\boldsymbol{e}$ is zero mean and has a variance σ^2, then (See Reading List)

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\theta}}(\lambda)] &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^* \\ \text{Cov}[\widehat{\boldsymbol{\theta}}(\lambda)] &= \sigma^2(\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-T} \\ \text{MSE}[\widehat{\boldsymbol{\theta}}(\lambda)] &= \sigma^2\text{Tr}\{\boldsymbol{W}_\lambda(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{W}_\lambda^T\} + \boldsymbol{\theta}^{*T}(\boldsymbol{W}_\lambda - \boldsymbol{I})^T(\boldsymbol{W}_\lambda - \boldsymbol{I})\boldsymbol{\theta}^*,\end{aligned}$$

where $\boldsymbol{W}_\lambda \overset{\text{def}}{=} (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}$. In particular, we can show that

Theorem (Theobald 1974)

For $\lambda < 2\sigma^2\|\boldsymbol{\theta}^*\|^{-2}$, it holds that $MSE(\widehat{\boldsymbol{\theta}}(\lambda)) < MSE(\widehat{\boldsymbol{\theta}})$.

Geometric Interpretation

The following three problems are equivalent

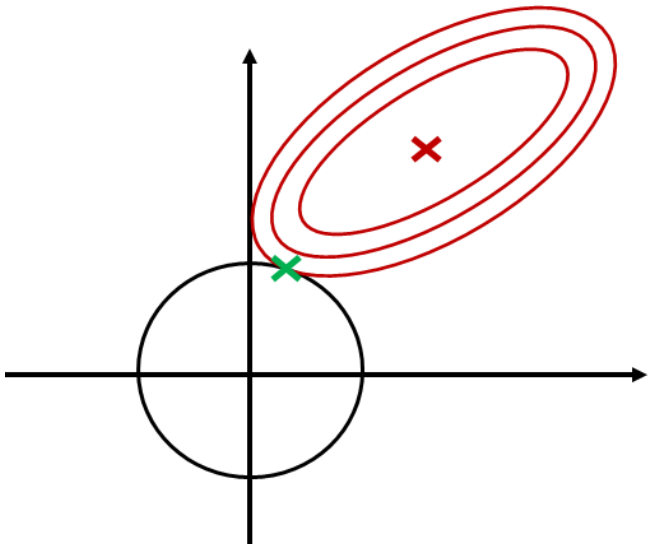
$$\boldsymbol{\theta}_\lambda^* = \underset{\boldsymbol{\theta}}{\text{argmin}} \ \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

$$\boldsymbol{\theta}_\alpha^* = \underset{\boldsymbol{\theta}}{\text{argmin}} \ \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\theta}\|^2 \leq \alpha$$

$$\boldsymbol{\theta}_\epsilon^* = \underset{\boldsymbol{\theta}}{\text{argmin}} \ \|\boldsymbol{\theta}\|^2 \quad \text{subject to} \quad \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 \leq \epsilon$$

under an appropriately chosen tuple $(\lambda, \alpha, \epsilon)$.

- Larger λ = Smaller α
- $\boldsymbol{\theta}^*$'s magnitude is tighter bounded



Choosing λ

Because the following three problems are equivalent

$$\begin{aligned}
\boldsymbol{\theta}_\lambda^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2 \\
\boldsymbol{\theta}_\alpha^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \qquad \text{subject to} \quad \|\boldsymbol{\theta}\|^2 \leq \alpha \\
\boldsymbol{\theta}_\epsilon^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{\theta}\|^2 \qquad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \leq \epsilon
\end{aligned}$$

- We can seek λ that satisfies $\|\boldsymbol{\theta}\|^2 \leq \alpha$:
  - You know how much $\|\boldsymbol{\theta}\|^2$ would be appropriate.
- We can seek λ that satisfies $\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \leq \epsilon$
  - You know how much $\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2$ would be tolerable.
- Other approaches:
  - Akaike's information criterion: Balance model fit with complexity
  - Cross validation: Leave one out
  - Generalized cross-validation: Cross-validation + weight