Optimal criterion for feature learning of two-layer linear neural network in high dimensional interpolation regime

Keita Suzuki[1], Taiji Suzuki[2,3]

[1]Preferred Networks.Inc, [2]The University of Tokyo, [3]RIKEN AIP

Introduction

Full potential of feature learning in regression problems

- Most papers on feature learning only focus on classification problems .
- Many works on feature learning of regression problems aren't enough to fully understand feature learning.
  - ➢ They focus on implicit regularization, which isn't necessary optimal.

Questions:
Can we design an optimal regularization of feature mapping to fully exploit the benefit of feature learning and demonstrate its improvement over simpler models like ridge regression?

Problem Settings

Multi-output Linear Regression:

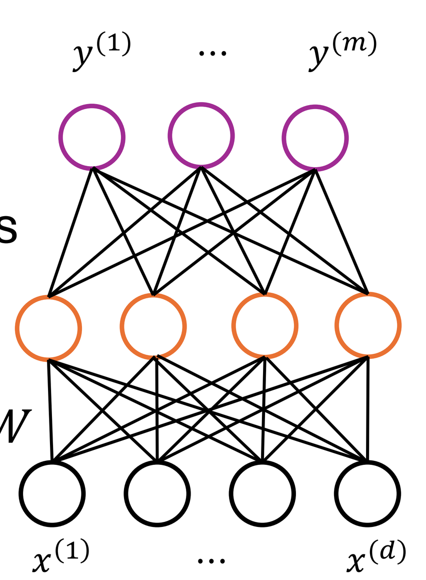
$$y_i^{(j)} = \beta_{*j}^\top x_i + \epsilon_i^{(j)} \quad (i = 1, \cdots, n, j = 1, \cdots, m)$$

Training data: $\left(x_i, \left(y_i^{(1)}, \dots, y_i^{(m)}\right)\right)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^m$:

Estimator and Model:

We estimate $y^{(i)}$ by two-layer linear neural network as

$$W^\top (W X^\top X W^\top + n\lambda I_d)^{-1} W X^\top y^{(i)}$$



Remark: If $W = I$, this estimator is equal to ridge regression.

Proposed Criterion for W:

We established direct estimator of its predictive risk:

$$R(W) := \frac{1}{m}\sum_{i=1}^m \min_\beta \frac{1}{n}\left\|y^{(i)} - XW^\top\beta\right\|^2 + \lambda\|\beta\|^2 + \underbrace{\frac{\sigma'^2}{n}\mathrm{Tr}(WX^\top XW^\top(WX^\top XW^\top + n\lambda I_d)^{-1})}_{\text{Degrees of Freedom}}$$

Remark
This can be seen as an extension of Mallows' Cp and WAIC for Ridge regression.

Main Result1: Selecting W with R(W)

Objective: $$\underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_x\left[\left(x^\top\beta_{*i} - x^\top W^\top \hat{\beta}_i(W)\right)^2\right]}_{\text{Average of predictive risk for each output}} \lesssim \text{Bias} + \text{Variance}$$

$$\hat{\beta}_i(W) = (WX^\top XW^\top + n\lambda I_d)^{-1} W X^\top y^{(i)}$$

Theorem1

For some $t > 1$ and $\delta = o(1)$, under some conditions, it holds that with high probability,

$$\text{Bias} + \text{Variance} \lesssim \max\{R(W) - \sigma^2, \delta\}.$$

Insight from Theorem1

- $R(W)$ plays a role of estimator of predictive risk.
- Minimizing $R(W)$ can lead to generalization.

➡ Q. How small $R(W) - \sigma^2$ can be? (Theorem2)

Remark: The evaluation in Theorem1 isn't uniform about W.

Theorem2

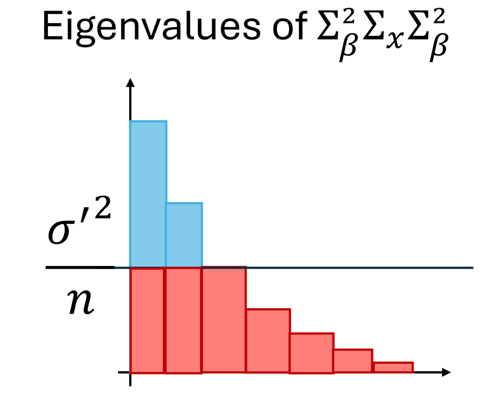
Suppose there exist $k \le n$ such that , for W such that $W^\top W = \frac{n\lambda}{\sigma'^2}\Sigma_\beta$, it holds that with high probability,

$$R(W) - \sigma^2 \lesssim \sum_{i=1}^d \min\left\{\frac{\sigma'^2}{n}, \mu_i\left(\Sigma_\beta^{1/2}\Sigma_x\Sigma_\beta^{1/2}\right)\right\}.$$

$\Sigma_x = \mathbb{E}[xx^\top]$
$\Sigma_\beta = \frac{1}{m}\sum_{i=1}^m \beta_{*i}\beta_{*i}^\top$

- Feature learning with $R(W)$ can find informative directions of Σ_β .
- Coordinate transformation with such W *can change the problem into like a kernel regime.*

Eigenvalues of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}$



Main Result2: Bayes risk optimality

Theorem 3

Suppose Σ_β is positive definite. Then under some conditions, it holds that with high probability

$$\min_W \text{Bias} + \text{Variance} \gtrsim \underbrace{\sum_{i=1}^d \min\left\{\frac{\sigma^2}{n}, \mu_i\left(\Sigma_\beta^{1/2}\Sigma_x\Sigma_\beta^{1/2}\right)\right\}}_{\text{Lower bound of optimal Bayes risk}}$$

W selected by $R(W)$ can achieve lower bound!

Proof Outline

$$\text{Bias} + \text{Variance} = \mathbb{E}_{\beta_*\sim\mathcal{N}(0,\Sigma_\beta), Y\sim\mathcal{N}(X\beta_*,\sigma^2 I)}\left[|\beta_* - \hat{\beta}(W)|_{\Sigma_x}^{\ 2}\right] := \underbrace{R(X, \sigma, \hat{\beta}(W))}_{\text{Bayes Risk}}$$

We can **obtain Bayes Estimator** as

$$\hat{\beta}_B := \mathrm{argmin}_\beta R(X, \sigma, \beta) = \left(X^\top X + \sigma^2 \underbrace{\Sigma_\beta^{-1}}_{\text{Can't access directory}}\right)^{-1} X^\top y.$$

Evaluating $R(X, \sigma, \beta_B)$ yields the lower bound.

Comparison with Ridge regression

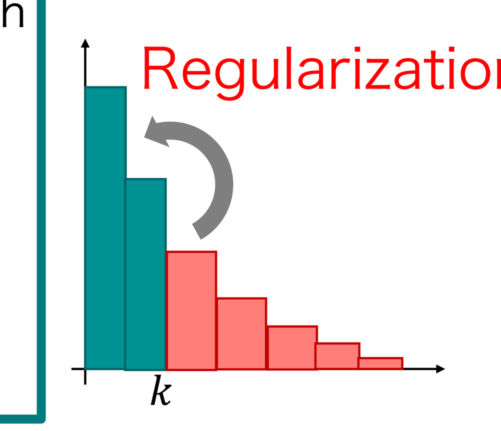
Theorem4

For $k \le n$, under some conditions, it holds with high probability,

$$\text{Bias} \approx \sum_{i=1}^k \lambda_i\sigma_i^2 \frac{(n\lambda + \sum_{i>k}\lambda_i)^2}{\lambda_i^2} + \sum_{i=k+1}^d \lambda_i\sigma_i^2,$$

$$\text{Variance} \approx \frac{k\sigma^2}{n} + \frac{\sigma^2}{n}\sum_{i=k+1}^d \frac{\lambda_i^2}{(n\lambda + \sum_{i>k}\lambda_i)^2}.$$

Eigenvalues of Σ_x



Feature learning with $R(W)$ outperforms ridge regression in the 4 cases:

1. When eigenvalues of Σ_x decreases slowly
2. When x and β_* are misaligned
3. When tail eigenvalues of Σ_x are large
4. When $y^{(i)}$ is large

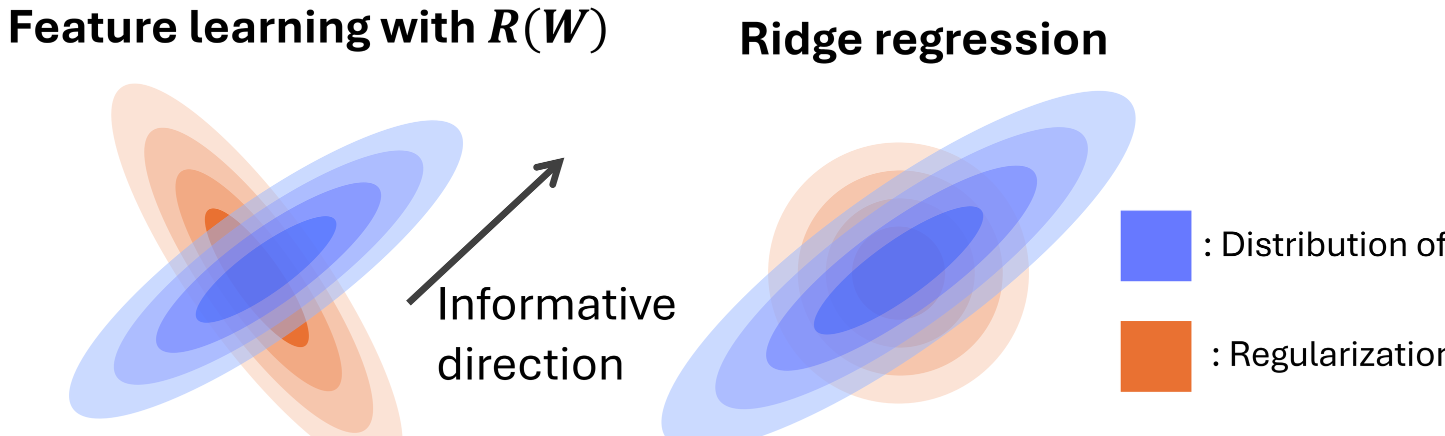
Ex. (When eigenvalues of Σ_x decreases slowly)

$$\lambda_i = \begin{cases} 1 \ (i \le n) \\ \dfrac{1}{i + 1 - n} \quad (i \ge n) \end{cases}, \quad \sigma_i^2 = \frac{n}{\log n}e^{-i}, \quad \sigma^2 = 1, \quad \sigma'^2 = 1, d \ge 2n$$

$$\sum_{i=1}^d \min\left\{\frac{\sigma'^2}{n}, \mu_i\left(\Sigma_\beta^{1/2}\Sigma_x\Sigma_\beta^{1/2}\right)\right\} = O\left(\frac{\log n}{n}\right)$$ **Feature learning with $R(W)$**

$$\text{Bias} \approx \sum_{i=1}^k \lambda_i\sigma_i^2 \frac{(n\lambda + \sum_{i>k}\lambda_i)^2}{\lambda_i^2} + \sum_{i=k+1}^d \lambda_i\sigma_i^2 = \Omega(1)$$ **Ridge regression**

Qualitative analysis



l_2 regularization with feature learning decreases variance without deteriorating bias.

Normal l_2 regularization regularize too strongly in the direction where Σ_β has large contribution.

Numerical experiments

Predictive risk of Two-layer NN, ridge regression and Bayes-optimal estimator.