Understanding of High-dimensional Ridgeless Least-Square Interpolation

Libin Liang

Advisor: Professor Zhiqiang Tan
Department of Statistics

April 13, 2022

Background

Interpolating Estimators: Estimators that have zero error in training set.

Ridgeless Least-Square: The min$-l^2$ norm least square estimator
$(\hat{\beta} = (X^TX)^+X^TY$ or $\hat{\beta} = \lim_{\lambda\to 0}\hat{\beta}_\lambda = (X^TX+\lambda I)^+X^TY)$

Why the High-dimensional Ridgeless Least-Square Interpolation is of interest?

- Interpolating Estimators such as Neural Network can have good generalization results in practical application.
- High-dimensional Least-Square Interpolator is one of the simplest interpolating estimators we can study.
- Ridgeless Least-Square Interpolator will be selected by **Gradient Descent** given zero initial and proper learning rate.

Overview

- Models Discussion:
  1. Linear Regression with Isotropic Features
  2. Latent Space Model
- Characteristics of Covariance Matrix
- Promising Direction

Linear Regression with Isotropic Features

$y_i = x_i^T \beta + \epsilon_i$

where $x_i \in \mathbb{R}^p$, the components of x_i are independent, zero mean, unit variance and with bounded moments of all order.

Numerical Study Setting:
$x_i \sim N(0, I_p)$ $\epsilon_i \sim N(0, 1)$
$\beta = (\frac{1}{\sqrt{p}}, ..., \frac{1}{\sqrt{p}})$
number of sample $n = 200$
$p = 100 - 1200$

Repeat 100 times for each pair (n, p) and take the average of in-sample error and out-sample error.

Linear Regression with Isotropic Features

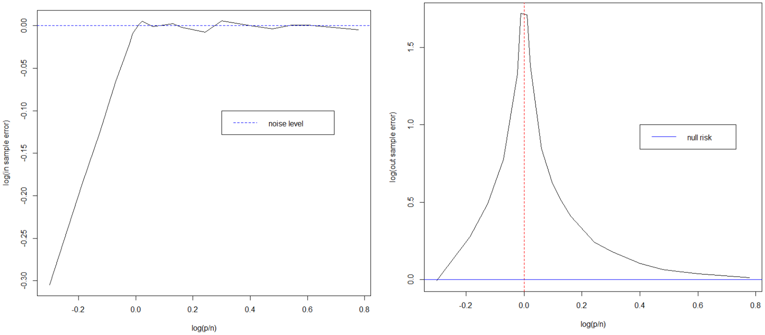

Figure: In-sample error(Left) VS Out-sample error(Right)

Linear Regression with Isotropic Features

$$Bias_X = E[(x_0^T((X^TX)^+X^TX\beta - \beta))^2|X]$$

$$Variance_X = E[(x_0^T(X^TX)^+X^T\epsilon)^2|X]$$

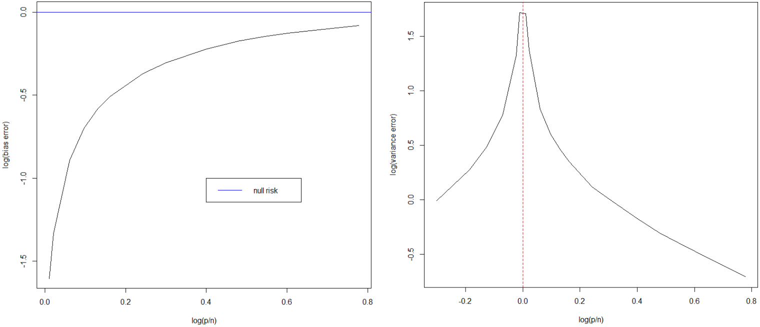

Figure: Bias error(Left) VS Variance (Right)

Linear Regression with Isotropic Features

- Overparameterization helps reducing the variance but it will increase the bias

$\tilde{\beta} = (X^T X)^+ X^T X \beta$ is the projection of β onto the eigenvectors space of $X^T X$.

If x_i is isotropic, the direction of eigenvectors space of $X^T X$ are symmetric. And the number of eigenvectors of $X^T X$ is n.

$||\tilde{\beta}||_2^2 = \tilde{\beta}^T \beta \approx ||\beta||_2^2 * \frac{n}{p} \Rightarrow ||\tilde{\beta} - \beta||_2^2 = ||\beta||_2^2 * \frac{p-n}{p}$

What happen if the eigenvectors space of $X^T X$ is more aligned with β when p is increasing?

Latent Space Model

Latent covariates: $z_i \in \mathbb{R}^d, \quad i = 1, ..., n$ with components are independent.

True Model: $y_i = \theta_i^T z_i + \xi_i, \ \ \xi_i \sim N(0, \sigma_\xi^2)$

We only observe: $x_i = (x_{i1}, ..., x_{ip}) \in \mathbb{R}^p$

$x_{ij} = w_j^T z_i + u_{ij}$, where $w_j \in \mathbb{R}^d$ and $u_{ij} \sim N(0, 1)$

Notice that $Var(w_j^T z_i)/var(u_{ij}) = w_j^T w_j =$ SNR for x_j

Latent Space Model

Let $W = \begin{pmatrix} w_1^T \\ \vdots \\ w_p^T \end{pmatrix}$

The linear model wrt to y and x is

$$y_i = x_i^T \beta + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

with

$\Sigma_x = I_p + WW^T$

$\beta = E[x_0 x_0^T]^{-1} E[x_0 y] = W(I_d + W^T W)^{-1} \theta$

$\sigma^2 = \sigma_\xi^2 + \theta^T (I_d + W^T W) \theta$

Latent Space Model

Experiment Setting:

- $z_i \sim_{i.i.d} N(0, I_d)$, $\quad \theta = (1/\sqrt{d}, ..., 1/\sqrt{d})^T$, $\quad \xi_i \sim N(0, 1)$.
- Averge SNR for $(x_{i1}, ..., x_{ip})$ is 1, $\frac{1}{p}\sum_{i=1}^{p} w_j^T w_j = \frac{1}{p} tr(WW^T) = 1$
- The singular values of the W are the same, WLOG, $W = \begin{pmatrix} \sqrt{\frac{p}{d}} I_d \\ 0 \end{pmatrix}$.
- d = 20, n = 200, p = 100 - 1200

Repeat 100 times for each pair (n, p) and take the average of in-sample error and out-sample error.

Errors are wrt to model $y_i = x_i\beta + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

Latent Space Model

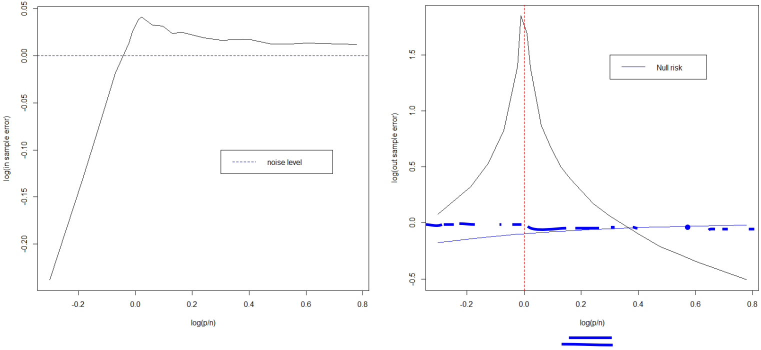

Figure: In-sample error(Left) VS Out-sample error(Right)

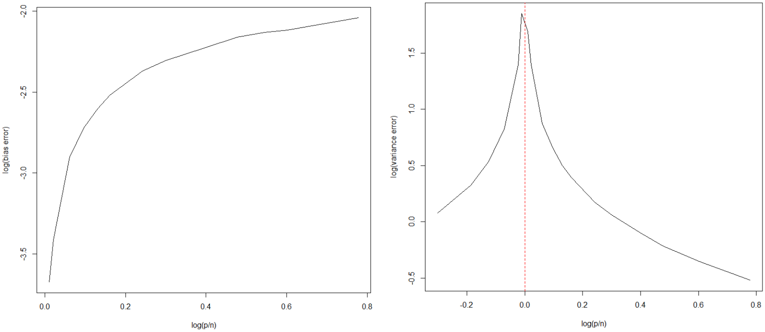

Figure: Bias error(Left) VS Variance (Right)

Latent Space Model

- Overparameterization still helps reducing the variance but increases the bias
- The scale of the bias error is much smaller than the null risk and the variance error.
- The total risk do not converge to null risk and keep decreasing with higher overparameterization and then converge to a risk that is lower than the noise level.

Characteristics of Covariance Matrix

Isotropic Features Linear model: $\Sigma_x = I_p, \quad \beta = (1/\sqrt{p}, ..., 1/\sqrt{p})^T$

Latent Space model:

$$\Sigma_x = \begin{pmatrix} (\frac{p}{d}+1)I_d & 0_{(dp-d)} \\ 0_{p-d\times d} & I_{p-d} \end{pmatrix} \quad \beta = (\sqrt{p}/(p+d), ..., \sqrt{p}/(p+d), 0, ..., 0)^T$$

For latent space model, in overparameterization scheme:

- A gap between leading eigenvalues and tailed eigenvalues
- The gap gets larger as p increases.
- The true β lines in the space of leading eigenvectors
- The number of tailed eigenvalues are large and decay slowly.

Characteristics of Covariance Matrix

$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ be the eigenvalues of Σ_x
$v_1, ..., v_p$ are the corresponding eigenvector

$\hat{H}_n(s) = \frac{1}{p}\sum_{i=1}^{p} 1_{\{s \geq \lambda_i\}}$ $\quad$ $\hat{G}_n(s) = \frac{1}{||\beta||_2^2}\sum_{i=1}^{p} < \beta, v_i >^2 1_{\{s \geq \lambda_i\}}$

Theorem 2(Hastie , Montanari, etc 2019)
Let $\gamma = \frac{p}{n}$ and c_0 is the solution of $1 - \frac{1}{\gamma} = \int \frac{s}{1+c_0\gamma s} d\hat{H}_n(s)$

Define $B(\hat{H}_n, \hat{G}_n, \gamma) = ||\beta||^2\{1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}\} \int \frac{s}{(1+c_)\gamma s)^2} d\hat{G}_n(s)$

$V(\hat{H}_n, \gamma) = \sigma^2\gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}$

Given $\hat{H}_n(s) \to H(s)$ and $\hat{G}_n(s) \to G(s)$ and certain assumptions, we have

$Bias_X \to B(H, G, \gamma)$ and $Variance_X \to V(H, \gamma)$

Characteristics of Covariance Matrix

- The effect of the magnitude of the gap

$$\frac{1}{p}\sum_{i=1}^{p} w_j^T w_j = \frac{1}{p} tr(WW^T) = \mu = 0.001$$

$$\Sigma_x = \begin{pmatrix} (\frac{p}{d} * \mu + 1) I_d & 0_{d \times p-d} \\ 0_{p-d \times d} & I_{p-d} \end{pmatrix} \qquad \beta = (\sqrt{\mu p}/(\mu p + d), ..., \sqrt{\mu p}/(\mu p + d), 0, ..., 0)^T$$

$$\frac{p}{n} \to \gamma \text{ and } \frac{d}{p} \to \psi \text{ and } \frac{d}{n} = \gamma * \psi = 0.1$$

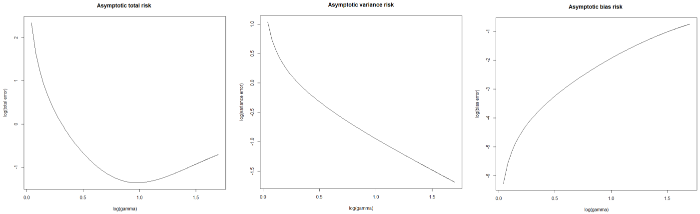

Figure: Asymptotic Total risk(Left) VS Asymptotic Variance (Middle) VS Asymptotic Bias(Right) for $\mu = 0.001$

Characteristics of Covariance Matrix

- The effect of the tailed eigenvalues decays slowly

$$\Sigma_x = \begin{pmatrix} (\frac{p}{d}+1)I_d & 0_{d\times p-d} \\ 0_{p-d\times d} & \Lambda_{p-d} \end{pmatrix} \text{ and } \Lambda = Diag(\lambda_{p-d+1}, ..., \lambda_p)$$

$\frac{1}{p-d}\sum_i \mathbf{1}_{\{s\geq\lambda_i\}} \to s^\alpha (s \in [0,1], \alpha > 0)$, small α relates to fast decay rate.

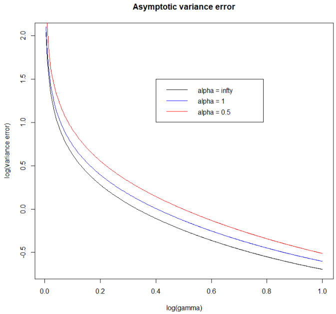

Figure: Asymptotic Variance for different decay rates

Promising direction

- Give a set of conditions directly w.r.t the covariance structure of the covariates such that the overparameterization is benefit
    1. The magnitude of the gap w.r.t eigenvalues.(Constant? How Fast it should grow?)
    2. How close should the true β be aligned with the leading eigenvectors.
    3. How slow should the tailed eigenvalues decay?
- The results from Haste & Montanari is based on the assumption that the covariates is linear transformatiom from random vector with independent components, what can we say if the covariates are sub-gaussian random vectors but not independent.
- With the results of ridgeless interpolation, can we have a better understaning of the generalization of other interpolating models such as random features and neural network?

Thank you!

Reference

- Trevor Hastie, Andrea Montanari, Saharon Rosset , Ryan J. Tibshirani (2019) *SURPRISES IN HIGH-DIMENSIONAL RIDGELESS LEAST SQUARES INTERPOLATION*
- Alexander Tsigler, Peter L. Bartlett (2020) *BENIGN OVERFITTING IN RIDGE REGRESSION*
- Peter L. Bartletta, Philip M. Longc , Gabor Lugosid, and Alexander Tsiglera (2020) *BENIGN OVERFITTING IN LINEAR REGRESSION*