Benign Overfitting In Ridge Regression

Hwichang Jeong

2022. 01.17

Seoul National University

Introduction

- Overparametrized models empirically show good generalization performance even if trained with vanishing or negative regularization.
- Understand theoretically how this effect can occur by studying the setting of ridge regression.
- Provide non-asymptotic generalization bound for overparametrized ridge regression model depending on the covariance structure of the data.

Notation and Assumption

- $X \in \mathbb{R}^{n \times p}$: Random design matrix with i.i.d centered row.
- $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_p)$: Covariance matrix of X with $\lambda_1 \geq \cdots \geq \lambda_p$.
- $X\Sigma^{-1/2}$ are sub-gaussian vectors with sub-gaussian norm at most σ_x.
- $y = X\theta^\star + \epsilon$ is the response vector, where $\theta^\star \in \mathbb{R}^p$ is some unknown vector, and ϵ is noise that is independent of X.
- Components of ϵ are independent and have sub-gaussian norms bounded by σ_ϵ.
- $a \lesssim_{\sigma_x} b$ if there exists a constant c_x that only depends on σ_x such that $a \leq c_x b$.

Ridge Regression

- Denote the ridge estimator as

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmin}}\{\|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2\} \\ &= X^\top(\lambda I_n + XX^\top)^{-1}y\end{aligned}$$

  where assume that the matrix $(\lambda I_n + XX^\top)$ is non-degenerate.

Generalization error

- For a new independent observation x, the prediction MSE is

$$\begin{aligned}\mathbb{E}\left[\left(x\left(\hat{\theta}-\theta^*\right)\right)^2 \mid X, \varepsilon\right] &= \left\|\hat{\theta}-\theta^*\right\|_\Sigma^2 \\ &= \left\|\theta^* - X^\top\left(\lambda I_n + XX^\top\right)^{-1}(X\theta^*+\varepsilon)\right\|_\Sigma^2 \\ &\lesssim \left\|\left(I_p - X^\top\left(\lambda I_n XX^\top\right)^{-1}X\right)\theta^*\right\|_\Sigma^2 \\ &+ \left\|X^\top\left(\lambda I_n + XX^\top\right)^{-1}\varepsilon\right\|_\Sigma^2\end{aligned}$$

where $\|x\|_\Sigma := \sqrt{x^\top \Sigma x}$.

- Denote

$$B := \left\|\left(I_p - X^\top\left(\lambda I_n + XX^\top\right)^{-1}X\right)\theta^*\right\|_\Sigma^2$$

$$V := \left\|X^\top\left(\lambda I_n + XX^\top\right)^{-1}\varepsilon\right\|_\Sigma^2$$

Notation

- For any matrix $M \in \mathbb{R}^{n \times p}$ denote $M_{0:k}$ to be the matrix which is comprised of the first k columns of M, and $M_{k:\infty}$ to be the matrix comprised of the rest of the columns of M.
- For any vector $\eta \in \mathbb{R}^p$ denote $\eta_{0:k}$ to be the matrix which is comprised of the first k components of η, and $\eta_{k:\infty}$ to be the matrix comprised of the rest of the components of η.
- $\Sigma_{0:k} = \mathrm{diag}(\lambda_1, \cdots, \lambda_k)$ and $\Sigma_{k:\infty} = \mathrm{diag}(\lambda_{k+1}, \cdots)$.
- $A_k = X_{k:\infty} X_{k:\infty}^\top + \lambda I_n$
- $A_{-k} = X_{0:k-1} X_{0:k-1}^\top + X_{k:\infty} X_{k:\infty}^\top + \lambda I_n$

Theorem (1)

Suppose $\lambda \geq 0$ and it is also known that for some $\delta < 1 - 4e^{-n/c_x^2}$ with probability at least $1-\delta$ the condition number of A_k is at most L, then with probability at least $1 - \delta - 20e^{-t/c_x}$

$$\frac{B}{L^4} \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2$$

$$\frac{V}{\sigma_\varepsilon^2 t L^2} \lesssim_{\sigma_x} \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}$$

Condition number of A_k

- Choose λ to control the condition number of A_k.
- To demonstrate the applications of Theorem 1, consider three different regimes.
- If $\sum_{i>k} \lambda_i \ll n\lambda_{k+1}$ for all k, control the condition number of A_k by choosing λ.

Lemma (2)
Let $\mu_{\max}(A_k)$ and $\mu_{\min}(A_k)$ be the largest eigenvalue and lowest eigenvalue of A_k. For some absolute constant c for any $t > 0$ with probability at least $1 - 6e^{-t/c}$,

$$\mu_{\max}(A_k) \le \lambda + c\sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_i \lambda_i \right).$$

If it's additionally known that for some $\delta, L > 0$ w.p. at least $1 - \delta$

$$\|X_{1,k:\infty}\| \ge \sqrt{\mathbb{E}\|X_{1,k:\infty}\|^2 / L}$$

then w.p. at least $1 - n\delta - 4e^{-t/c}$,

$$\mu_{\min}(A_k) \ge \lambda + \frac{1}{L}\sum_{i>k}\lambda_i - c\sigma_x^2 \sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_i \lambda_i^2\right)}.$$

Theorem - Bound for particular covariance operators

Theorem (3)
Suppose there exists a large constant c_x that only depends on σ_x s.t. $n\lambda_{k+1} \gtrsim \sigma_x \sum_{i>k} \lambda_i$ for some $k < n/c_x$, then for $\lambda = n\lambda_{k+1}$ and for any $t \in (c_x, n/c_x)$, with probability at least $1 - 26e^{-t/c_x}$,

$$B \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \lambda_k^2 \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}}, \quad \frac{V}{\sigma_\varepsilon^2 t} \lesssim_{\sigma_x} \frac{k}{n} + \frac{\sum_{i>k} \lambda_i^2}{n\lambda_k^2}$$

Other regimes

- If $\sum_{i>k} \lambda_i \geq c_x n \lambda_{k+1}$ for some large constant c_x, one can control all the eigenvalues of A_k up to constant factor even for vanishing λ. (Adding small positive regularization has no effect)
- If $\sum_{i>k} \lambda_i \geq c_x n \lambda_{k+1}$ for extremely large constant c_x, one can change the bound by choosing negative λ by decreasing bias without significantly increasing the variance.

Theorem (4)

Suppose the components of the data vectors are independent and there exists a large constant c_x that only depends on σ_x s.t. $\sum_{i>k} \lambda_i \geq c_x n \lambda_{k+1}$ for some $k < n/c_x$.

(a) For any non-negative $\lambda < \sum_{i>k} \lambda_i$, for any $t \in (c_x, n/c_x)$, with probability at least $1 - 22e^{-t/c_x}$,

$$B \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{\mathbf{0}:k}\|^2_{\Sigma^{-\mathbf{1}}_{\mathbf{0}:k}} \left(\frac{\sum_{i>k} \lambda_i}{n}\right)^2, \quad \frac{V}{\sigma_\varepsilon^2 t} \lesssim_{\sigma_x} \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\left(\sum_{i>k} \lambda_i\right)^2}.$$

(b) For $\xi > c_x$ and $\lambda = -\sum_{i>k} \lambda_i + \xi \left(n\lambda_{\mathbf{1}} + \sqrt{n \sum_{i>k} \lambda_i^2}\right)$ for any $t \in (c_x, n/c_x)$ with probability at least $1 - 20e^{-t/c_x}$

$$B \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{\mathbf{0}:k}\|^2_{\Sigma^{-\mathbf{1}}_{\mathbf{0}:k}} \frac{\xi^2}{n} \left(n\lambda^2_{k+\mathbf{1}} + \sum_{i>k} \lambda_i^2\right)$$

$$\frac{V}{\sigma_\varepsilon^2 t} \lesssim_{\sigma_x} \frac{k}{n} + \frac{\sum_{i>k} \lambda_i^2}{\xi^2 \left(n\lambda^2_{k+\mathbf{1}} + \sum_{i>k} \lambda_i^2\right)}$$

Effective rank

- Define effective rank ρ_k and R_k.

$$\rho_k = \frac{\lambda + \sum_{i>k} \lambda_i}{n\lambda_{k+1}}, \quad R_k = \frac{(\lambda + \sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

- Then the bounds for bias and variance become

$$\frac{B}{L^4} \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \lambda^2_{k+1}\rho^2_k, \quad \frac{V}{\sigma^2_\varepsilon t L^2} \lesssim_{\sigma_x} \frac{k}{n} + \frac{n}{R_k}$$

Lemma

Lemma (5)
Suppose that $\lambda \geq 0$*, components of the rows of* X *are independent, and the components of the noise vector* ε *have unit variance. Then for some absolute constant* c *for any* t, k *s.t.* $t > c$ *and* $k + 2\sigma_x^2 t + \sqrt{kt}\sigma_x^2 < n/2$ $w \cdot p$*. at least* $1 - 20e^{-t/c}$,

$$V \geq \frac{1}{cn} \sum_{i=1} \min \left\{ 1, \frac{\lambda_i^2}{\sigma_x^4 \lambda_{k+1}^2 (\rho_k + 2)^2} \right\}$$

Lemma (6)
For arbitrary $\bar{\theta} \in \mathbb{R}^p$ *consider the following prior distribution on* $\theta^* : \theta^*$ *is obtained from* $\bar{\theta}$ *randomly flipping signs of all its coordinates. Suppose also that* $\lambda \geq 0$ *and it is known for some* k, δ, L *that for any* $j > k$ *w. p. at least* $1 - \delta \mu_n (A_{-j}) \geq \frac{1}{L} \left(\lambda + \sum_{i>k} \lambda_i\right)$. *Then for some absolute constant* c *for any non-negative* $t < \frac{n}{2\sigma_x^2}$ *w.p. at least* $1 - 2\delta - 4e^{-t/c}$

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{2} \sum_i \frac{\lambda_i \bar{\theta}_i^2}{\left(1 + \frac{\lambda_i}{2L\lambda_{k+1}\rho_k}\right)^2}$$

Theorem (7)

Denote

$$\underline{B} := \sum_i \frac{\lambda_i \left|\theta_i^*\right|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2}, \qquad \bar{B} := \|\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}^2 + \|\theta_{0:k}^*\|_{\Sigma_{0:k}^{-1}}^2 \left(\frac{\lambda+\sum_{i>k}\lambda_i}{n}\right)^2,$$

$$\underline{V} := \frac{1}{n}\sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+2)^2}\right\}, \quad \bar{V} := \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda+\sum_{i>k}\lambda_i\right)^2}.$$

Suppose $\rho_k \in (a, b)$ *for some* $b > a > 0$*. Then*

$$1 \le \frac{\bar{B}}{\underline{B}} \le \max\left\{(1+b)^2, \left(1+a^{-1}\right)^2\right\}, \quad 1 \le \frac{\bar{V}}{\underline{V}} \le \max\left\{(2+b)^2, \left(1+2a^{-1}\right)^2\right\}.$$

Alternatively, if $k = \min\{l : \rho_l > b\}$ *and* $b > 1/n$ *then*

$$1 \le \frac{\bar{B}}{\underline{B}} \le \max\left\{(1+b)^2, \left(1+b^{-1}\right)^2\right\}, \quad 1 \le \frac{\bar{V}}{\underline{V}} \le \max\left\{(2+b)^2, \left(1+2b^{-1}\right)^2\right\}.$$