A note on benign overfitting

Yen-Chi Chen
University of Washington
July 23, 2024

This section is a simplification of the following paper:

> Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). *Surprises in high-dimensional ridgeless least squares interpolation.* Annals of statistics, 50(2), 949.

In recent years, researchers have discovered an interesting phenomenon called *benign overfitting*: when the dimension increases (and sample size is fixed), the mean square error of a linear model may be decreasing! In this section, we will briefly explain how this could happen.

We will consider a special linear model called *ridgeless regression*, a combination of the usual least squared model and ridge regression. Let

$$\widehat{\beta}_{RL} = \mathsf{argmin}\left\{\|b\| : b \text{ minimizes } \|\mathbb{Y} - \mathbb{X}b\|\right\}, \tag{1}$$

where $\mathbb{Y} = (Y_1, \cdots, Y_n)^T \in \mathbb{R}^n$ is the response vector and $\mathbb{X} \in \mathbb{R}^{n\times p}$ is the feature/covariate matrix. The estimator in equation (1) is the ridgeless regression.

Here is an interesting property about $\widehat{\beta}_{RL}$:

$$\widehat{\beta}_{RL} = \begin{cases} \widehat{\beta}_{OLS} & \text{if } n > p, \\ \widehat{\beta}_{LI} & \text{if } p > n, \end{cases}$$

where $\widehat{\beta}_{OLS}$ is the ordinary least square and $\widehat{\beta}_{LI}$ is the least norm interpolator.

The least norm interpolator is defined as a limiting case of the ridge regression:

$$\widehat{\beta}_{LI} = \lim_{\lambda\to 0} \widehat{\beta}_{\lambda},$$
$$\widehat{\beta}_{\lambda} = \mathsf{argmin}_b \|\mathbb{Y} - \mathbb{X}b\| + \lambda\|b\|_2^2.$$

Here is an interesting fact: $\widehat{\beta}_{RL}$ will demonstrate the benign overfitting! See Figure 1.

1 Setup

To investigate this phenomenon, we will consider the following IID setup:

$$\mathbb{Y} = \mathbb{X}\beta^* + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2\mathbf{I}_n),$$

where $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix. Moreover, we assume that entries $\{X_{ij}\}$ are IID from $N(0,1)$. Namely, each row vectors $X_1, \cdots, X_n$ are IID from $N(0, \mathbf{I}_p)$

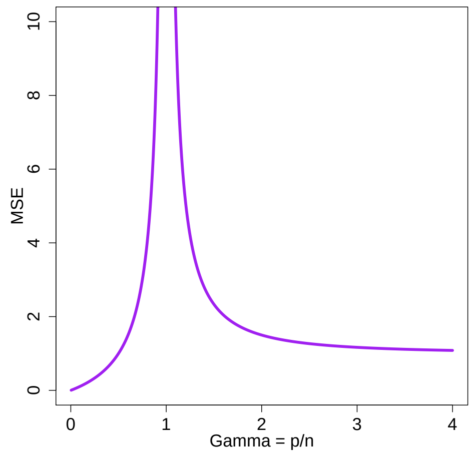

Figure 1: The MSE of the ridgeless estimator $\widehat{\beta}_{RL}$ as a function of $\gamma = \frac{p}{n}$ under $\sigma^2 = 1$ and $\|\beta^*\| = 1$.

To investigate the mean square error, we will separately analyze the bias and variance. In particular, we will consider the conditional bias and variance:

$$\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X}) \in \mathbb{R}^n, \qquad \mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) = \mathsf{Tr}[\mathsf{Cov}(\widehat{\beta}_{RL}|\mathbb{X})],$$

where $\mathsf{Cov}(\widehat{\beta}_{RL}|\mathbb{X})$ is the covariance matrix.

The conditional MSE is

$$\mathsf{MSE}(\widehat{\beta}_{RL}|\mathbb{X}) = \|\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X})\|^2 + \mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}).$$

2 Analysis on $p < n$

When $p < n$, it is clear that the bias is 0 because $\widehat{\beta}_{RL} = \widehat{\beta}$. Thus,

$$\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X}) = 0.$$

For the variance, the story is more interesting. First, let

$$\widehat{\Sigma} = \frac{\mathbb{X}^T\mathbb{X}}{n} \in \mathbb{R}^{p\times p}$$

be the (sample) covariance matrix. For the ordinary least square, we know that

$$\begin{aligned}\mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) &= \mathsf{Tr}[\mathsf{Cov}(\widehat{\beta}_{RL}|\mathbb{X})]\\ &= \mathsf{Tr}[(\mathbb{X}^T\mathbb{X})^{-1}\sigma^2]\\ &= \frac{\sigma^2}{n}\cdot \mathsf{Tr}(\widehat{\Sigma}^{-1}).\end{aligned}$$

Using the property of trace,

$$\mathsf{Tr}(\widehat{\Sigma}^{-1}) = \sum_{j=1}^{p} \mu_j^{-1}(\widehat{\Sigma}),$$

where $\mu_j(A)$ is the j-th eigenvalue of A.

To investigate the property of eigenvalues of a Gaussian covariance matrix, we will use the *Marchenko-Pastar theorem (MP theorem)*.

Theorem 1 (Marchenko-Pastar theorem) *Let $\{Z_{ij}\}$ be IID random variables with $\mathbb{E}(Z_{ij}) = 0, \mathsf{Var}(Z_{ij}) = 1$. Let $\mathbb{Z} \in \mathbb{R}^{n\times p}$ be the matrix of $\{Z_{ij}\}$. Define $\widehat{\Omega} = \frac{\mathbb{Z}^T\mathbb{Z}}{n} \in \mathbb{R}^{p\times p}$ and $S_{\widehat{\Omega}}$ be the distribution of eigenvalues of $\widehat{\Omega}$, i.e.,*

$$S_{\widehat{\Omega}}(t) = \frac{1}{p}\sum_{j=1}^{p} I(\mu_j(\widehat{\Omega}) \leq t).$$

When $n, p \to \infty, \frac{p}{n} \to \gamma < 1$, we have the following results:

1. *$S_{\widehat{\Omega}}$ converges in distribution to S_γ, where S_γ has a PDF*

$$S_{\gamma(t)} = \begin{cases} \frac{1}{2\pi\gamma}\frac{1}{t}\sqrt{(b-t)(t-a)}, & t \in [a,b] \\ 0, & \textit{Otherwise}. \end{cases}$$

   and $a = (1-\sqrt{\gamma})^2, b = (1+\sqrt{\gamma})^2$.

2. *The Stieltjes transform of $S_\gamma(t)$ is*

$$\begin{aligned} \omega_\gamma(-z) &= \int \frac{dS_\gamma(t)}{t-z} \\ &= \frac{-(1-\gamma-z)+\sqrt{(1+\gamma-z)^2-4\gamma}}{2\gamma z}. \end{aligned}$$

3. *Using L'Hospital rule, we further have*

$$\omega_\gamma(0) = \lim_{\lambda\to 0}\omega_\gamma(z) = \frac{1}{1-\gamma}.$$

The above theorem is from Chapter 3 of

Bai, Z., & Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices* (Vol. 20). New York: Springer.

The power of Theorem 1 is that the trace of inverse covariance matrix

$$\mathsf{Tr}(\widehat{\Sigma}^{-1}) = \sum_{j=1}^{p} \mu_j^{-1}(\widehat{\Sigma}),$$

can can be written as

$$
\begin{aligned}
\mathsf{Tr}(\widehat{\Sigma}^{-1}) &= p\frac{1}{p}\sum_{j=1}^{p}\mu_j^{-1}(\widehat{\Sigma}) \\
&= p\int \frac{1}{t} dS_{\widehat{\Sigma}}(t) \\
&\approx p\int \frac{1}{t} dS_{\gamma}(t) \\
&= p\cdot\omega_{\gamma}(0) = \frac{p}{1-\gamma}
\end{aligned}
$$

when $\gamma = \frac{p}{n} < 1$, which is our current setting.

To sum up,

$$
\mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) = \frac{\sigma^2}{n}\mathsf{Tr}(\widehat{\Sigma}^{-1}) \approx \sigma^2\frac{p}{n}\frac{1}{1-\gamma} = \sigma^2\frac{\gamma}{1-\gamma},
$$

so

$$
\mathsf{MSE}(\widehat{\beta}_{RL}|\mathbb{X}) \approx \sigma^2\frac{\gamma}{1-\gamma}
$$

when $\gamma = \frac{p}{n} < 1$. Thus, when γ increases, the mean square error increases as long as $\gamma < 1$.

3 Analysis on $p > n$

When $p > n$, $\widehat{\beta}_{RL} = \lim_{\lambda\to 0}\widehat{\beta}_{\lambda}$, so we will first investigate the bias and variance of the ridge regression.

A feature of the ridge regression is its closed form:

$$
\begin{aligned}
\widehat{\beta}_{\lambda} &= (\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}_p)^{-1}\mathbb{X}^T\mathbb{Y} \\
&= \mathbb{X}^T(\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}_n)^{-1}\mathbb{Y},
\end{aligned}
$$

where the last equality can be verified by multiplying $(\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}_p)$ in both sides.

Analysis of variance. We first analyze the variance.

$$
\begin{aligned}
\mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) &= \mathsf{Tr}[\mathsf{Cov}(\widehat{\beta}_{RL}|\mathbb{X})] \\
&= \lim_{\lambda\to 0}\mathsf{Tr}[\mathsf{Cov}(\widehat{\beta}_{\lambda}|\mathbb{X})], \\
\mathsf{Cov}(\widehat{\beta}_{\lambda}|\mathbb{X}) &= \mathbb{X}^T(\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}_n)^{-2}\mathbb{X}\cdot\sigma^2, \\
\mathsf{Tr}[\mathsf{Cov}(\widehat{\beta}_{\lambda}|\mathbb{X})] &= \mathsf{Tr}[\mathbb{X}^T(\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}_n)^{-2}\mathbb{X}]\cdot\sigma^2 \\
&= \mathsf{Tr}[\mathbb{X}\mathbb{X}^T(\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}_p)^{-2}]\cdot\sigma^2 \qquad \text{(trace property)} \\
&= \frac{1}{p}\mathsf{Tr}\left[\frac{\mathbb{X}\mathbb{X}^T}{p}\left(\frac{\mathbb{X}^T\mathbb{X}}{p} + \frac{n}{p}\lambda\mathbf{I}_n\right)^{-2}\right]\cdot\sigma^2 \\
&= \frac{\sigma^2}{p}\mathsf{Tr}[\widehat{Q}(\widehat{Q} + \tau\lambda\mathbf{I}_n)^{-2}],
\end{aligned}
$$

where

$$\widehat{Q} = \frac{\mathbb{X}\mathbb{X}^T}{p} \in \mathbb{R}^{n\times n}, \qquad \tau = \frac{1}{\lambda} = \frac{n}{p} < 1.$$

As $\lambda \to 0$,

$$\begin{aligned}
\mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) &= \mathsf{Tr}[\mathsf{Cov}(\widehat{\beta}_{RL}|\mathbb{X})] \\
&= \frac{\sigma^2}{p}\mathsf{Tr}[\widehat{Q}(\widehat{Q}+\tau\lambda\mathbf{I}_n)^{-2}] \\
&\approx \frac{\sigma^2}{p}\mathsf{Tr}(\widehat{Q}^{-1}) \\
&= \tau\cdot\frac{1}{n}\sum_{j=1}^{n}\mu_j^{-1}(\widehat{Q}).
\end{aligned}$$

Now we apply Theorem 1 again with swapping n, p in the setting and conclude that

$$\mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) \approx \sigma^2\frac{\tau}{1-\tau} = \frac{\sigma^2}{\gamma-1}.$$

Analysis of bias. To analyze the bias, we will use another property about $\widehat{\beta}_{RL}$ that it can be expressed by the pseudo-inverse when $p > n$:

$$\widehat{\beta}_{RL} = (\mathbb{X}^T\mathbb{X})^{\dagger}\mathbb{X}^T\mathbb{Y},$$

where for a matrix $A \in \mathbb{R}^{p\times p}$ its pseudo-inverse $A^{\dagger}$ satisfies $AA^{\dagger}A = A, A^{\dagger}AA^{\dagger} = A^{\dagger}$. Note that if A has rank $r < p$, then $\mathsf{Tr}[A^{\dagger}A] = r$.

Let $\widehat{\Omega} = \mathbb{X}^T\mathbb{X}$. A direct computation shows that

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_{RL}|\mathbb{X}) &= \widehat{\Omega}^{\dagger}\widehat{\Omega}\beta^* \\
\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X}) &= (\mathbf{I}_p - \widehat{\Omega}^{\dagger}\widehat{\Omega})\beta^* \\
\|\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X})\|^2 &= \beta^{*T}(\mathbf{I}_p - \widehat{\Omega}^{\dagger}\widehat{\Omega})\beta^*.
\end{aligned}$$

Here is an interesting property about the Gaussian vectors $X_i \sim N(0, \mathbf{I}_p)$. For any rotation matrix $U \in \mathbb{R}^{p\times p}$,

$$UX_i \overset{d}{=} X_i,$$

i.e., UX_i has identical distribution as X_i.

Thus, we can rewrite the bias as

$$\begin{aligned}
\|\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X})\|^2 &= \beta^{*T}(\mathbf{I}_p - \widehat{\Omega}^{\dagger}\widehat{\Omega})\beta^* \\
&= (U\beta^*)^T(\mathbf{I}_p - \widehat{\Omega}^{\dagger}\widehat{\Omega})(U\beta^*).
\end{aligned}$$

Now we pick $U_1, \cdots, U_p$ such that

$$U_i\beta^* = \|\beta^*\|\cdot e_i,$$

where e_i is the unit i-th coordinate vector.

Thus,

$$\|\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X})\|^2 = (U_i\beta^*)^T(\mathbf{I}_p - \widehat{\Omega}^\dagger\widehat{\Omega})(U_i\beta^*) = \|\beta^*\|^2(1 - [\widehat{\Omega}^\dagger\widehat{\Omega}]_{ii})$$

for $i = 1, \cdots, p$.

With this result, we 'average' them, which leads to

$$\|\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X})\|^2 = \frac{1}{p}\sum_{i=1}^{p}\|\beta^*\|^2(1 - [\widehat{\Omega}^\dagger\widehat{\Omega}]_{ii}) = \|\beta^*\|^2\left(1 - \frac{1}{p}\underbrace{\mathsf{Tr}(\widehat{\Omega}^\dagger\widehat{\Omega})}_{=n}\right) = \|\beta^*\|^2(1 - \frac{1}{\gamma}).$$

Putting variance and bias together, we conclude that when $p > n$,

$$\begin{aligned}
\|\mathsf{Bias}(\widehat{\beta}_{RL}|\mathbb{X})\|^2 &= \|\beta^*\|^2\left(1 - \frac{1}{\gamma}\right)\\
\mathsf{Var}(\widehat{\beta}_{RL}|\mathbb{X}) &\approx \frac{\sigma^2}{\gamma - 1}\\
\mathsf{MSE}(\widehat{\beta}_{RL}|\mathbb{X}) &\approx \|\beta^*\|^2\left(1 - \frac{1}{\gamma}\right) + \frac{\sigma^2}{\gamma - 1}.
\end{aligned}$$

When $\gamma = \frac{p}{n} \to \infty$ and $\|\beta^*\|$ remains fixed, we see that bias is converging to a fixed quantity but the variance keeps decreasing. Thus, the total mean squared error is decreasing as $\gamma \to \infty$.

4 Summary

Now we consider both regimes and conclude that

$$\mathsf{MSE}(\widehat{\beta}_{RL}|\mathbb{X}) \approx \begin{cases} \sigma^2\frac{\gamma}{1-\gamma}, & \text{when } p < n\\ \|\beta^*\|^2\left(1 - \frac{1}{\gamma}\right) + \frac{\sigma^2}{\gamma-1} & \text{when } p > n. \end{cases}$$

As $\gamma = \frac{p}{n}$ increases from 0, the MSE first increases until $\gamma = 1$, and then the MSE decreases, leading to the famous phenomenon of the benign overfitting. Figure 1 shows the asymptotic MSE under $\sigma^2 = 1$ and $\|\beta^*\|^2 = 1$.

Note that a crucial feature of the MSE decreasing is based on the assumption that $\|\beta^*\|^2$ remains fixed as $p \to \infty$. Since the total signal is fixed, the average signal on each coordinate is shrinking.