Benign Overfitting In Linear Regression

P.L. Bartlett, P.M. Long, G. Lugosi, and A. Tsigler (2020)

Yuha Park
17, Jan, 2022

Contents

Introduction

- **the benign overfitting phenomenon**: deep neural networks seem to predict well, even with a perfect fit to noisy training data *(overfitting can perform well)*
- Motivated by this phenomenon, consider when a perfect fit to training data in linear regression is compatible with accurate prediction

- the purposes of the paper are:
  1) to consider the simplest setting where we might hope to witness this phenomenon: linear regression
  2) to give a characterization of linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy

Definitions and Notation

Definition 1 (Linear Regression). A linear regression problem in a separable Hilbert space $\mathbb{H}$ is defined by a random covariate vector $x \in \mathbb{H}$ and outcome $y \in \mathbb{R}$. We define

1) the covariate operator $\Sigma = \mathbb{E}[xx^\top]$ and
2) the optimal parameter vector $\theta^* \in \mathbb{H}$, satisfying

$$\mathbb{E}(y - x^\top \theta^*)^2 = \min_\theta \mathbb{E}(y - x^\top \theta)^2.$$

Definitions and Notation

Assumption.

1) x and y are mean zero;
2) $x = V\Lambda^{1/2}z$, where $\Sigma = V\Lambda V^\top$ is the spectral decomposition of Σ and z has components that are independent σ_x^2 sub-Gaussian with σ_x a positive constant: that is, for all $\lambda \in \mathbb{H}$,

$$\mathbb{E}[exp(\lambda^\top z)] \leq \exp(\sigma_x^2 \|\lambda\|^2/2),$$

where $\|\cdot\|$ is the norm in the Hilbert space $\mathbb{H}$;
3) the conditional noise variance is bounded below by some constant σ^2,

$$\mathbb{E}\left[(y - x^\top\theta)^2 \middle| x\right] \geq \sigma^2;$$

Definitions and Notation

Assumption (Continued).

4) $y - x^\top \theta^*$ is σ_y^2 sub-Gaussian conditionally on x: that is, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\lambda(y - x^\top \theta^*))\middle| x\right] \leq \exp(\sigma_y^2 \lambda^2/2)$$

(note that this implies $\mathbb{E}[y|x] = x^\top \theta^*$); and

5) almost surely, the projection of the data X on the space orthogonal to any eigenvector of Σ spans a space of dimension n, where X is the linear map from $\mathbb{H}$ to $\mathbb{R}^n$ corresponding to $x_1, \cdots, x_n \in \mathbb{H}$ so that $X\theta \in \mathbb{R}^n$ has ith component ${x_i}^\top \theta$ (a training sample $(x_1, y_1), \cdots, (x_n, y_n)$: n independent pairs with the same distribution as (x, y)).

Definitions and Notation

Notation.

- the excess risk of the estimator $R(\theta) := \mathbb{E}_{x,y}\left[\left(y - x^\top \theta\right)^2 - \left(y - x^\top \theta^*\right)^2\right]$, where $\mathbb{E}_{x,y}$ denotes the conditional expectation given all random quantities other than x, y;

- $\mu_1(\Sigma) \geq \mu_2(\Sigma) \geq \cdots$: the eigenvalues of Σ;
- $\|\Sigma\|$: the operator norm of Σ;
- I: the identity operator on $\mathbb{H}$;
- I_n: the $n \times n$ identity matrix.

Definitions and Notation

Definition 2 (Minimum Norm Estimator). Given data $X \in \mathbb{H}^n$ and $\mathbf{y} \in \mathbb{R}^n$, the minimum norm estimator $\hat{\theta}$ solves the optimization problem

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{such that} \quad & \|X\theta - \mathbf{y}\|^2 = \min_{\beta} \|X\beta - \mathbf{y}\|^2. \end{aligned}$$

$\Rightarrow$ The minimum norm solution is given by

$$\hat{\theta} = X^\top (XX^\top)^{-1} \mathbf{y}.$$

Definition 3 (Effective Ranks). For the covariance operator Σ, define $\lambda_i = \mu_i(\Sigma)$ for $i = 1, 2, \cdots$. If $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$ for $k \geq 0$, define

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}$$

Main Results

Theorem 1. *For any σ_x, there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem from Definition 1. Define*

$$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\},$$

where the minimum of the empty set is defined as ∞. Suppose that $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* \geq n/c_1$, then $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$. Otherwise,

$$R(\hat{\theta}) \leq c\left(\|\theta^*\|^2\|\Sigma\|\max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right)$$
$$+ c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$$

with probability at least $1 - \delta$, and

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

Main Results

Theorem 1(Continued). *Moreover, there are universal constants* a_1, a_2, n_0 *such that, for all* $n \geq n_0$*, for all* Σ*, and for all* $t \geq 0$*, there is a* θ^* *with* $\|\theta^*\| = t$ *such that, for* $x \sim \mathcal{N}(0, \Sigma)$ *and* $y|x \sim \mathcal{N}(x^\top \theta^*, \|\theta^*\|^2 \|\Sigma\|)$ *with probability at least* $1/4$*,*

$$R(\hat{\theta}) \geq \frac{1}{a_1} \|\theta^*\|^2 \|\Sigma\| \, \mathbb{1}\left[\frac{r_0(\Sigma)}{n \log(1 + r_0(\Sigma))} \geq a_2\right].$$

Definition 4. A sequence of covariance operators Σ_n with $\|\Sigma_n\| = 1$ is benign if

$$\lim_{n\to\infty} \frac{r_0(\Sigma_n)}{n} = \lim_{n\to\infty} \frac{k_n^*}{n} = \lim_{n\to\infty} \frac{n}{R_{k_n^*}(\Sigma_n)} = 0.$$

Main Results

Theorem 2.

1) *If* $\mu_k(\Sigma) = k^{-\alpha} ln^{-\beta}((k+1)e/2)$, *then* Σ *is benign if and only if* $\alpha = 1$ *and* $\beta > 1$.
2) *If*

$$\mu_k(\Sigma_n) = \begin{cases} \gamma_k + \epsilon_n & \text{if } k \leq p_n, \\ 0 & \text{otherwise} \end{cases}$$

and $\gamma_k = \Theta(exp(-k/\tau))$, *then* Σ_n *with* $\|\Sigma_n\| = 1$ *is benign if and only if* $p_n = \omega(n)$ *and* $ne^{-o(n)} = \epsilon_n p_n = o(n)$. *Furthermore, for* $p_n = \Omega(n)$ *and* $\epsilon_n p_n = ne^{-o(n)}$,

$$R(\hat{\theta}) = O\Big(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\Big\{\frac{1}{n}, \frac{n}{p_n}\Big\}\Big).$$

Relevance to Deep Neural Networks

- the connection appears by considering regimes where deep neural networks are well approximated by linear functions of their parameters
- very wide neural networks can be accurately approximated by linear functions in an appropriate Hilbert space
- covariance eigenvalues that are constant or slowly decaying in a high (but finite)-dimensional space might be important in the deep network setting also

Conclusions

1) characterizes when the phenomenon of benign overfitting occurs in high-dimensional linear regression with Gaussian data and more generally
2) gives finite sample excess risk bounds that reveal the covariance structure that ensures that the minimum norm interpolating prediction rule has near-optimal prediction accuracy
3) the characterization depends on two notions of the effective rank of the data covariance operator
4) overparameterization is essential for benign overfitting: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size
5) data that lie in a large but finite-dimensional space exhibit the benign overfitting phenomenon with a much wider range of covariance properties than data that lie in an infinite-dimensional space