Benign Overfitting in Linear Regression

Mike Bao, Franklyn Wang

May 10, 2021

Table of Contents

Work Covered

This work covers [BLLT20]. Essentially, this paper gives a tight bound for the excess risk of a linear predictor in the overparametrized regime.

They identify nearly exactly a subregime called the “benign overfitting” subregime, where overfitting occurs; yet the excess risk does not suffer too much. We will first describe the motivation, then a setup, and then a description of the regime and how it arises.

Classical Statistical Learning Theory

- Many of the models we have seen in class have very few parameters in them.
- However, most deep models have a much larger number of parameters, frequently far more than the number of data points.
- In statistical learning theory, we are taught that models which fit every data point exactly cannot possibly generalize.

Classical Statistical Learning Theory

- Somehow models in the real world both interpolate and have low test loss.
- How is this possible?
- Informally, all the memorization has to go into dimensions that are somewhat inessential for the prediction.

Table of Contents

Benign Overfitting: Practice

- Many times when we use deep neural networks, we can add lots of noise to the training set, and the models (which are trained using standard cross-entropy losses) will continue to perform well. See, for example, [CWK20, HZZ20].
- This is quite strange – it seems that the overfitting doesn't actually hurt the network. This begs the question: maybe the overfitting doesn't matter?
- To make the problem more tractable, we look at the linear regression case.

Definitions

- Because we consider regimes where $n < p$, it is often the case that several estimators will achieve minimum least squared loss. Thus, we define the *min-norm estimator*.

Definition (Min-Norm Estimator)

Define $\hat{\theta}$ to be min-norm estimator if and only if it solves the following optimization problem:

$$\min_{\theta \in \mathbb{H}} \|\theta\|^2$$
$$\text{such that } \|X\theta - \boldsymbol{y}\|^2 = \min_{\beta} \|X\beta - \boldsymbol{y}\|^2$$

Some prior work

- Generally speaking, most generalization bounds show that $\hat{\theta} \approx \theta^*$.
- However, the notion of approximation is crucial – we often don't have $\left\|\hat{\theta} - \theta^*\right\|_2 \to 0$!
- Instead, we look at the excess risk

$$\mathbb{E}_{x,y}[\underbrace{(y - x^\top \hat{\theta})^2}_{\text{model risk}} - \underbrace{(y - x^\top \theta^*)^2}_{\text{optimal risk}}] = (\hat{\theta} - \theta^*)^\top \Sigma (\hat{\theta} - \theta^*) = \left\|\hat{\theta} - \theta^*\right\|_\Sigma^2$$

Thinking about the problem

- One way to think about the problem is that in a linear regression problem, Σ is basically enough to fully specify a data generating process.
- Thus, any insight we can obtain as to why overfitting happens can only come from thinking about the spectrum of Σ, $\lambda_1 \geq \lambda_2, \ldots \geq \lambda_d$. **These d numbers uniquely determine the risk of the algorithm!**

A teaser

Theorem ([BLLT20], Theorem 6)

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

Our Implementation

We implemented the algorithm described Google's JAX library, which is an extension to the NumPy library. Our code can be found at this GitHub repository.

Table of Contents

Benign Overfitting: Theory

- How can we decide whether overfitting is a concern in a regression problem?
- Intuitively, there are two constraints to the complexity of the problem we must consider:
  - The scale of the problem should be small compared to the sample size - the eigenvalues must decay relatively quickly
  - The problem must not be dominated by its largest eigenvalues.

Further Definitions

- Let $\mu_k(\Sigma)$ denote the k-th largest eigenvalue of Σ.
- To constrain the complexity of the regression problem, we construct the notions of *effective ranks*.

Definition (Effective Ranks)

If Σ is a covariance matrix, and $\lambda_i = \mu_i(\Sigma)$ for $i = 1, 2, \ldots,$ then define:

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

- By bounding the excess risk using effective ranks, we will be able to classify Σ as *benign* based on its eigenvalues.

Main Result

Theorem (Theorem 4)

Let $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$*. Let* $\delta < 1$ *such that* $\log(1/\delta) < n/c$*. Then, there exist constants* $b, c, c_1 > 1$ *such that the following holds. 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 $1-\delta$*. Additionally,* $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$.

Main Result

Theorem (Theorem 4)

Let $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$*. Let* $\delta < 1$ *such that* $\log(1/\delta) < n/c$*.*
Then, there exist constants $b, c, c_1 > 1$ *such that the following holds.*
If $k^* \geq n/c_1$*, then* $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$
Otherwise,

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

with probability at least $1-\delta$*. Additionally,* $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$*.*

Main Result

Theorem (Theorem 4)

Let $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. *Let* $\delta < 1$ *such that* $\log(1/\delta) < n/c$. *Then, there exist constants* $b, c, c_1 > 1$ *such that the following holds. If* $k^* \geq n/c_1$, *then* $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$
Otherwise,

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

with probability at least $1-\delta$. *Additionally,* $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$.

Main Result

Theorem (Theorem 4)

Let $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. *Let* $\delta < 1$ *such that* $\log(1/\delta) < n/c$. *Then, there exist constants* $b, c, c_1 > 1$ *such that the following holds. If* $k^* \geq n/c_1$, *then* $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$
Otherwise,

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

with probability at least $1 - \delta$. *Additionally,* $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$.

Examples of Benign Problems

Definition (Corollary to Theorem 4)

A covariance operator Σ is *benign* if

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

Theorem (Theorem 6)

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

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

and $\gamma_k = \Theta(\exp(-k/\tau))$, *then* Σ_n *is benign iff* $p_n = \omega(n)$ *and* $ne^{-o(n)} = \epsilon_n p_n = o(n)$.

Excess Risk Bound in terms of the Trace

Theorem (Lemma 7)

$$R(\hat{\theta}) \leq 2{\theta^*}^\top B\theta^* + c\sigma^2 \log\frac{1}{\delta}\,\mathrm{tr}(C)$$

with probability at least $1-\delta$*. Additionally,*

$$\mathbb{E}_\varepsilon R(\hat{\theta}) \geq {\theta^*}^\top B\theta^* + \sigma^2\,\mathrm{tr}(C)$$

where

$$B = \left(I - X^\top \left(XX^\top\right)^{-1} X\right)\Sigma\left(I - X^\top \left(XX^\top\right)^{-1} X\right)$$

$$C = \left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}$$

- We can come up with bounds on ${\theta^*}^\top B\theta^*$ based on [KL17]
- Thus, the core of this proof is to bound $\mathrm{tr}(C)$.

Proof of Lemma 7

Using the fact that $y - x^\top \theta^*$ has zero mean:

$$
\begin{aligned}
R(\hat{\theta}) &= \mathbb{E}_{x,y}\left(y - x^\top \hat{\theta}\right)^2 - \mathbb{E}\left(y - x^\top \theta^*\right)^2 \\
&= \mathbb{E}_{x,y}\left(y - x^\top \theta^* + x^\top\left(\theta^* - \hat{\theta}\right)\right)^2 - \mathbb{E}\left(y - x^\top \theta^*\right)^2 \\
&= \mathbb{E}_{x}\left(x^\top\left(\theta^* - \hat{\theta}\right)\right)^2
\end{aligned}
$$

Proof of Lemma 7 (cont.)

$$
\begin{aligned}
R(\hat{\theta}) &= \mathbb{E}_x \left(x^\top \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* - x^\top X^\top \left(XX^\top \right)^{-1} \varepsilon \right)^2 \\
&\leq 2\mathbb{E}_x \left(x^\top \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* \right)^2 \\
&\quad + 2\mathbb{E}_x \left(x^\top X^\top \left(XX^\top \right)^{-1} \varepsilon \right)^2 \\
&= 2\theta^* B \theta^* + 2\varepsilon^\top C \varepsilon
\end{aligned}
$$

Lemma 36 from [PG19] finishes the proof by showing

$$
\varepsilon^\top C \varepsilon \leq \sigma^2 \operatorname{tr}(C)(2t+1) + 2\sigma^2 \sqrt{\operatorname{tr}(C)^2 \left(t^2 + t\right)} \leq (4t+2)\sigma^2 \operatorname{tr}(C)
$$

Trace Decomposition

Theorem (Lemma 8)

Consider a covariance operator Σ and $\lambda_n > 0$. Write its spectral decomposition $\Sigma = \sum_j \lambda_j v_j v_j^\top$, where the orthonormal $v_j \in \mathbb{H}$ are the eigenvectors corresponding to the λ_j. For i with $\lambda_i > 0$, define $z_i = Xv_i/\sqrt{\lambda_i}$. Then,

$$\operatorname{tr}(C) = \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i \right]$$

Furthermore, if $\lambda_i > 0$, letting $A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top$, we have

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2}$$

Key Step

One of the most important steps is an understanding of the eigenvalues of A_{-i}.

Lemma

There is a constant c such that for any $k \geq 0$ with probability at least $1 - 2e^{-n/c}$,

$$\frac{1}{c}\sum_{j>k}\lambda_j - c\lambda_{k+1}n \leq \mu_{k+1}(A_{-i}) \leq c\left(\sum_{j>k}\lambda_j + \lambda_{k+1}n\right)$$

and if $r_k(\Sigma) \geq bn$,

$$\frac{1}{c}\lambda_{k+1}r_k(\Sigma) \leq \mu_n(A_{-i}) \leq \mu_{k+1}(A_{-i}) \leq c\lambda_{k+1}r_k(\Sigma).$$

Upper Bound on the Trace

Lemma

There are constants $b, c \geq 1$ *such that if* $0 \leq k \leq n/c$, $r_k(\Sigma) \geq bn$, *and* $l \leq k$ *then with probability at least* $1 - 7e^{-n/c}$,

$$\operatorname{tr}(C) \leq c\left(\frac{l}{n} + n\frac{\sum_{i>l}\lambda_i^2}{(\sum_{i>k}\lambda_i)^2}\right)$$

Sketch of Upper Bound on Trace

The idea is to write

$$\mathrm{tr}(C) = \sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i$$

so it suffices to bound each set of terms independently.
For the first set of terms, we can obtain that

$$z_i^\top A_{-i}^{-2} z_i \leq \mu_n(A_{-i})^{-2} \|z_i\|^2 \leq \frac{c_1^2 \|z\|^2}{(\lambda_{k+1} r_k(\Sigma))^2}$$

and (where $\Pi_{\mathcal{L}_i}$ is a projection onto the lowest $n-k$ eigenvectors of A_{-i}).

$$z_i^\top A_{-i}^{-1} z \geq (\Pi_{\mathcal{L}_i} z)^\top A_{-i}^{-1} (\Pi_{\mathcal{L}_i} z) \geq \mu_{k+1}(A_{-i})^{-1} \|(\Pi_{\mathcal{L}_i} z)\|^2 \geq \frac{\|\Pi_{\mathcal{L}_i} z\|^2}{c_1 \lambda_{k+1} r_k(\Sigma)}$$

Upper Bound on the Trace, contd.

For the second sum we get

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \leq \sum_{i>l} \mu_n(A_{-i})^{-2} \lambda_i^2 \left\|z_i\right\|^2 \leq \frac{c_1^2 \sum_{i>l} \lambda_i^2 \left\|z_i\right\|^2}{(\lambda_{k+1} r_k(\Sigma))^2}$$

This can be bounded with standard concentration inequalities.

Lemma

Suppose $\{\lambda_i\}_{i=1}^\infty$ is a non-increasing sequence of non-negative numbers such that $\sum_{i=1}^\infty \lambda_i < \infty$, and $\{\xi_i\}_{i=1}^\infty$ are independent centered σ-subexponential random variables. Then for some universal constant a for any $t > 0$ with probability at least $1 - 2e^{-t}$,

$$\left|\sum_i \lambda_i \xi_i\right| \leq a\sigma \max\left(t\lambda_1, \sqrt{t \sum_i \lambda_i^2}\right)$$

Lower Bound on the Trace

Lemma

There is a constant c such that for any $i \geq 1$ with $\lambda_i > 0$, and any $0 \leq k \leq n/c$, with probability at least $1 - 5e^{-n/c}$,

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \frac{1}{cn}\left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}$$

Proof Sketch of Lower Bound on the Trace

- The proof relies on several previous results [Lemma 10, Corollary 13] but we demonstrate the style of proofs used in this paper to establish the main result, as well as the other lemmas presented.
- Lemma 10 establishes that with probability $1 - 2e^{-n/c_1}$

$$z_i^\top A_{-i}^{-1} z_i \geq \frac{\left\|\Pi_{\mathcal{L}_i} z_i\right\|^2}{c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right)}$$

Proof Sketch of Lower Bound on the Trace (cont.)

- Corollary 13 establishes that with probability at least $1 - 3e^{-n/c_1}$

$$\|\Pi_{\mathcal{L}_i} z_i\|^2 \geq n - a\sigma_x^2(k + t + \sqrt{tn}) \geq n/c_2$$

- Combining the two previous results with a union bound, with probability at least $1 - 5e^{-n/c_1}$:

$$z_i^\top A_{-i}^{-1} z_i \geq \frac{n}{c_3\left(\sum_{j>k}\lambda_j + \lambda_{k+1} n\right)}$$

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{\left(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i\right)^2} \geq \left(\frac{c_3\left(\sum_{j>k}\lambda_j + \lambda_{k+1} n\right)}{\lambda_i n} + 1\right)^{-2} \frac{z_i^\top A_{-i}^{-2} z_i}{\left(z_i^\top A_{-i}^{-1} z_i\right)^2}$$

Proof Sketch of Lower Bound on the Trace (cont.)

Using Corollary 13 again and the Cauchy-Schwarz Inequality, we obtain our desired result with proper choice of c_4:

$$\begin{aligned}\frac{z_i^\top A_{-i}^{-2} z_i}{\left(z_i^\top A_{-i}^{-1} z_i\right)^2} &\geq \frac{z_i^\top A_{-i}^{-2} z_i}{\left\|A_{-i}^{-1} z_i\right\|^2 \left\|z_i\right\|^2} \\ &= \frac{1}{\left\|z_i\right\|^2} \geq \frac{1}{n + a\sigma_x^2(t + \sqrt{nt})} \geq \frac{1}{c_4 n}\end{aligned}$$

Applications and Future Work

- Benign overfitting was first observed in deep neural networks
- Theorem 4 is connected to *neural tangent kernels (NTKs)*, which are regimes where neural networks can be well-approximated by linear functions
- NTKs generally have high dimension and slowly decaying eigenvalues of the covariates, which are required for benign overfitting
- However, it is an open question of whether a version of Theorem 4 can be applied more generally to neural networks

References I

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler.
Benign overfitting in linear regression.
Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.

Daniel Chiu, Franklyn Wang, and Scott Duke Kominers.
Generalization by recognizing confusion.
arXiv preprint arXiv:2006.07737, 2020.

Lang Huang, Chao Zhang, and Hongyang Zhang.
Self-adaptive training: beyond empirical risk minimization.
Advances in Neural Information Processing Systems, 33, 2020.

References II

Vladimir Koltchinskii and Karim Lounici.
Concentration inequalities and moment bounds for sample covariance operators.
Bernoulli, 23(1), Feb 2017.

Stephen Page and Steffen Grunewalder.
Ivanov-regularised least-squares estimators over large rkhss and their interpolation spaces.
arXiv:1706.03678 [math, stat], Jun 2019.