Overparameterization and Benign Overfitting in Linear Regression

Peter Bartlett
CS and Statistics
UC Berkeley

May 11, 2019

Phil Long  Gábor Lugosi  Alexander Tsigler
Overfitting in Deep Networks

- Deep networks can be trained to zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for noisy problems.
- No tradeoff between fit to training data and complexity!
- *Benign overfitting.*

(Zhang, Bengio, Hardt, Recht, Vinyals, 2017)

also (Belkin, Hsu, Ma, Mandal, 2018)
Benign Overfitting

A new statistical phenomenon

- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.
- An unexplored statistical phenomenon: good prediction with zero training error for regression loss.
- A similar phenomenon was observed in the mid-90s for classification problems. It led to the development of the margins analysis: zero classification loss still leaves room for tradeoffs between complexity and training error for regression loss.
- An aside: there is nothing mysterious about $p > n$ (‘overparameterization’). Overparameterization $=$ nonparametric
Progress on interpolating prediction

- Interpolating nearest neighbor rules in high dimensions
  (Belkin, Hsu, Mitra, 2018)
- Kernel regression with kernels defined in terms of the Euclidean inner product
  (Liang and Rakhlin, 2018)
- Kernel smoothing with singular kernels
  (Belkin, Rakhlin, Tsybakov, 2018)
- Linear regression with $p, n \to \infty, p/n \to \gamma$
  (Hastie, Montanari, Rosset, Tibshirani, 2019)
Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- $(x, y)$ Gaussian, mean zero.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \quad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$

$$\theta^* := \arg\min_{\theta} \mathbb{E} \left( y - x^\top \theta \right)^2,$$

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

Minimum norm estimator

- Data: $X \in \mathbb{H}^n$, $y \in \mathbb{R}^n$.
- Estimator $\hat{\theta} = (X^\top X)^\dagger X^\top y$, which solves

$$\min_{\theta \in \mathbb{H}} \|\theta\|^2$$

$$\text{s.t.} \quad \|X\theta - y\|^2 = \min_{\beta} \|X\beta - y\|^2.$$ 

Excess prediction error:

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)} \left[ (y - x^\top \hat{\theta})^2 - (y - x^\top \theta^*)^2 \right] = (\hat{\theta} - \theta^*)^\top \Sigma (\hat{\theta} - \theta^*).$$

($\Sigma$ determines importance of parameter directions)
Overfitting regime

- We consider situations where $\min_\beta \| X\beta - y \|^2 = 0$.
- Hence, $y_1 = x_1^\top \hat{\theta}, \ldots, y_n = x_n^\top \hat{\theta}$.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?
Benign Overfitting: Main Result

Theorem

For universal constants $b$, $c$, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min \left\{ k \geq 0 : r_k(\Sigma) \geq bn \right\}$,

1. With high probability,

$$ R(\hat{\theta}) \leq c \left( \|\theta^*\|^2 \sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sigma^2 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right), $$

2. $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c} \min \left\{ \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1 \right\}$. 

Notions of Effective Rank

Definition (Effective Ranks)
Recall that \( \lambda_1 \geq \lambda_2 \geq \cdots \) are the eigenvalues of \( \Sigma \).
For \( k \geq 0 \), if \( \lambda_{k+1} > 0 \), define the effective ranks

\[
\begin{align*}
  r_k(\Sigma) &= \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, & R_k(\Sigma) &= \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}.
\end{align*}
\]

Lemma

\[1 \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).\]
Notions of Effective Rank

\[ r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \quad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}. \]

Examples

1. \( r_0(I_p) = R_0(I_p) = p. \)

2. If \( \text{rank}(\Sigma) = p \), we can write

\[ r_0(\Sigma) = \text{rank}(\Sigma) s(\Sigma), \quad R_0(\Sigma) = \text{rank}(\Sigma) S(\Sigma), \]

with

\[ s(\Sigma) = \frac{1/p \sum_{i=1}^p \lambda_i}{\lambda_1}, \quad S(\Sigma) = \frac{(1/p \sum_{i=1}^p \lambda_i)^2}{1/p \sum_{i=1}^p \lambda_i^2}. \]

Both \( s \) and \( S \) lie between \( 1/p \) (\( \lambda_2 \approx 0 \)) and \( 1 \) (\( \lambda_i \) all equal).
Benign Overfitting: Main Result

**Theorem**

For universal constants $b$, $c$, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min \{ k \geq 0 : r_k(\Sigma) \geq bn \}$,

1. With high probability,

$$R(\hat{\theta}) \leq c \left( \|\theta^*\|^2 \sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sigma^2 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right),$$

2. $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c} \min \left\{ \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1 \right\}$. 
Benign Overfitting: Main Result

Intuition

- The mix of eigenvalues of $\Sigma$ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and
  2. how errors in $\hat{\theta}$ affect prediction accuracy.

- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions.

- Need $\text{tr}(\Sigma)/n$ small; need $r_{k^*}(\Sigma)/n \geq b$ and $R_{k^*}(\Sigma)/n$ large.

- Overparameterization is essential for benign overfitting: many eigenvalues not too far below $\lambda_{k^*+1}$. 
Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: \((\text{corresponding to } x^\top \theta^* \text{ and } y - x^\top \theta^*)\)
  1. \(\hat{\theta}\) is a distorted version of \(\theta^*\), because the sample \(x_1, \ldots, x_n\) distorts our view of the covariance of \(x\).
  
  \textit{Not a problem, even in high dimensions (}p > n\text{).}

  2. \(\hat{\theta}\) is corrupted by the noise in \(y_1, \ldots, y_n\).

\textit{Problematic.}

- When can the label noise be hidden in \(\hat{\theta}\) without hurting predictive accuracy?
Bias-variance decomposition

Define the noise vector $\epsilon = y - X\theta^*$.

Estimator:
$$\hat{\theta} = (X^\top X)^\dagger X^\top y = (X^\top X)^\dagger X^\top (X\theta^* + \epsilon),$$

Excess risk:
$$R(\hat{\theta}) = (\hat{\theta} - \theta^*)^\top \Sigma (\hat{\theta} - \theta^*)$$
$$= \theta^*^\top (I - \hat{\Sigma}\hat{\Sigma}^\dagger) (\Sigma - \hat{\Sigma}) (I - \hat{\Sigma}^\dagger\hat{\Sigma}) \theta^*$$
$$+ \sigma^2 \text{tr} \left( (X^\top X)^\dagger \Sigma \right).$$
Benign Overfitting: Proof Ideas

Standard normals

\[
\text{tr} \left( \left( X^T X \right)^\dagger \Sigma \right) = \text{tr} \left( \Sigma^{1/2} X^T \left( XX^T \right)^{-2} X \Sigma^{1/2} \right) \\
= \sum_{i=1}^{\infty} \lambda_i^2 z_i^\top A^{-2} z_i \\
= \sum_{i=1}^{\infty} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},
\]

where \( z_i = Xv_i / \sqrt{\lambda_i} \) for \( \Sigma = \sum_j \lambda_j v_j v_j^\top \), and

\[
A = \sum_{i=1}^{\infty} \lambda_i z_i z_i^\top, \quad A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top.
\]

Now \( z_i \sim \mathcal{N}(0, I_n) \) and \( z_i \) and \( A_{-i} \) are independent.
Benign Overfitting: Proof Ideas

Concentration

If \( r_k(\Sigma) \geq bn \), then

\[
\frac{1}{c} \lambda_{k+1} r_k(\Sigma) \leq \mu_n(A) \leq \mu_{k+1}(A) \leq c \lambda_{k+1} r_k(\Sigma),
\]

where \( \mu_1(A) \geq \cdots \geq \mu_n(A) \) are the eigenvalues of \( A = \sum_i \lambda_i z_i z_i^\top \).

- Split the trace into “heavy” directions, which cost \( 1/n \) each, and “light” directions, which cost \( n/R_k(\Sigma) \).
- The excess expected loss is at least as big as the same trace term.
- When \( A \) and \( A_{-i} \) are concentrated, the same split gives a lower bound within a constant factor of the upper bound.
  (And otherwise, the excess expected loss is at least a constant.)
Benign Overfitting: Main Result

**Theorem**

For universal constants $b$, $c$, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min \{ k \geq 0 : r_k(\Sigma) \geq bn \}$, then

1. With high probability,
   
   $$ R(\hat{\theta}) \leq c \left( \| \theta^* \|^2 \sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sigma^2 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right), $$

2. 
   
   $$ \mathbb{E} R(\hat{\theta}) \geq \frac{\sigma^2}{c} \min \left\{ \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1 \right\}. $$
Interpolation: far from the regime of a tradeoff between fit to training data and complexity.

In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well:

The noise is hidden in many unimportant directions.
  - Relies on overparameterization
  - ... and lots of unimportant parameters
Can we extend these results to interpolating deep networks?

- There are recent results showing that in extremely wide networks, a gradient flow stays near a linear approximation. But these conditions seem unnatural; representation learning with linear combinations of random features?
- Benign overfitting with these nonlinear functions?
- What is the analog of the minimum norm linear prediction rule?
- What role does the optimization method play?
Interpolation: far from the regime of a tradeoff between fit to training data and complexity.

In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well:

The noise is hidden in many unimportant directions.

- Relies on overparameterization
- ... and lots of unimportant parameters