Benign Overfitting in Linear Prediction

Peter Bartlett
CS and Statistics
UC Berkeley

June 17, 2019
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)
A new statistical phenomenon

An aside:

1. There is nothing mysterious about $p > n$ (‘overparameterization’).
   \[ \text{overparameterization} = \text{nonparametric} \]

2. There is nothing new about good prediction with zero training error for classification loss.
   \[ \text{margins analysis: regression loss vs complexity} \]

An unexplored statistical phenomenon:

- good prediction with zero regression loss on noisy training data.
- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks can be trained to fit noisy data perfectly, and they predict well.
Progress on interpolating prediction

- Interpolating nearest neighbor rules in high dimensions
  (Belkin, Hsu, Mitra, 2018)
- Kernel regression with polynomial kernels
  (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)
- Linear regression with random features
  (Belkin, Hsu and Xu, 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:

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

  \[ \theta^* := \text{arg min}_\theta \mathbb{E} \left( y - x^\top \theta \right)^2, \]

  \[ \sigma^2 := \mathbb{E}(y - x^\top \theta^*)^2. \]
### 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

\[
\min_{\theta \in \mathbb{H}} \quad \|\theta\|^2 \\
\text{s.t.} \quad \|X\theta - y\|^2 = \min_{\beta} \|X\beta - y\|^2.
\]

### Excess prediction error:

(\( \Sigma \) and \( \lambda_i \) determine importance of parameter directions)

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

Overfitting regime

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

Theorem

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

With high probability,

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

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

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

#### Lemma

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

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

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 \( 1 \ (\lambda_i \ \text{all equal}). \)
**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,

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

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: A Characterization
Benign Overfitting: A Characterization

**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.
- Overparameterization is essential for benign overfitting:
  - Number of non-zero eigenvalues: large compared to $n$,
  - Their sum: small compared to $n$,
  - Number of ‘small’ eigenvalues: large compared to $n$,
  - Small eigenvalues: roughly equal (but they can be more asymmetric if there are many more than $n$ of them).
Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: $(x^\top \theta^* \text{ and } y - x^\top \theta^*)$  
  1. $\hat{\theta}$ is a distorted version of $\theta^*$, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of $x$.  
     
     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.  
     
     Problematic.

- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?
Define the noise vector $\epsilon$ by $y = X\theta^* + \epsilon$.

**Estimator:**

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

**Excess risk:**

$$R(\hat{\theta}) = \left(\hat{\theta} - \theta^*\right)^\top \Sigma \left(\hat{\theta} - \theta^*\right)$$

$$= \theta^*\top \left( I - \hat{\Sigma}\hat{\Sigma}^\dagger \right) \left( \Sigma - \hat{\Sigma} \right) \left( I - \hat{\Sigma}^\dagger \hat{\Sigma} \right) \theta^*$$

$$+ \sigma^2 \text{tr} \left( \left( X^\top X \right)^\dagger \Sigma \right).$$
Benign Overfitting: Proof Ideas

Standard normals

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

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, $\text{tr} \left( (X^\top X)^\dagger \Sigma \right)$.

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: A Characterization

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

We say $\Sigma$ is *asymptotically benign* if

$$\lim_{n \to \infty} \left( \|\Sigma\| \sqrt{\frac{r_0(\Sigma)}{n}} + \frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma)} \right) = 0,$$

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

Example

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

The $\lambda_i$ must be almost diverging!!?!?
What kinds of eigenvalues?

Example: Finite dimension, plus isotropic noise

If

\[ \lambda_{k,n} = \begin{cases} 
  e^{-k} + \epsilon_n & \text{if } k \leq p_n, \\
  0 & \text{otherwise},
\end{cases} \]

then \( \Sigma_n \) is benign iff

- \( p_n = \omega(n) \),
- \( \epsilon_n p_n = o(n) \) and \( \epsilon_n p_n = \omega(ne^{-n}) \).

Furthermore, for \( p_n = \Omega(n) \) and \( \epsilon_n p_n = \omega(ne^{-n}) \),

\[ R(\hat{\theta}) = O\left( \frac{\epsilon_n p_n}{n} + \max\left\{ \frac{1}{n}, \frac{n}{p_n} \right\} \right). \]

Universal phenomenon: fast converging \( \lambda_i \), \( p_n \gg n \), noise in all directions.
Implications for deep learning

Neural networks versus linear prediction

Neural networks with

- width large compared to sample size,
- suitable random initialization,
- gradient descent with small step-size,

can be accurately approximated by linear functions in a certain randomly chosen Hilbert space.


- But what can we say about realistic deep network architectures?
- It seems unlikely that random features is the whole story.
Implications for adversarial examples

Label noise appears in $\hat{\theta}$

We can find a unit norm $\Delta$

such that perturbing an input $x$ by $\Delta$ changes the output enormously: even if $\Delta^T \theta^* = 0$,

$$
\left\| (x + \Delta)^T \hat{\theta} - x^T \hat{\theta} \right\|^2 \geq \frac{\sigma}{\sqrt{\lambda_{k^*} + 1}} \geq \sqrt{\frac{n}{\text{tr}(\Sigma)}} \sigma.
$$

Benign overfitting leads to huge sensitivity.
Can we extend these results to interpolating deep networks?
- Beyond 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?
- Implications for regularization methods?
- Implications for robustness?
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

But it leads to huge sensitivity to (adversarial) perturbations.