Benign Overfitting in Linear Prediction

Peter Bartlett
CS and Statistics
UC Berkeley

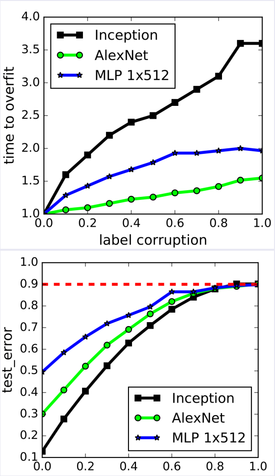
June 17, 2019

Phil Long

Gábor Lugosi

Alexander Tsigler

Overfitting in Deep Networks



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

- 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.*

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

Benign Overfitting

A new statistical phenomenon

- An aside:
  1. There is nothing mysterious about $p > n$ ('overparameterization').
     overparameterization = nonparametric
  2. There is nothing new about good prediction with zero training error for classification loss. 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.

Interpolating Prediction Rules

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:

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

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

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

Definitions

Minimum norm estimator

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

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

Excess prediction error:

(Σ and λ_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, $y_1 = x_1^\top \hat{\theta}, \ldots, y_n = x_n^\top \hat{\theta}$.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

Benign Overfitting: 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,

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

2.

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

Notions of Effective Rank

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

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

Lemma

$$1 \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$

Notions of Effective Rank

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

Examples

1. $r_0(I_p) = R_0(I_p) = p$.
2. If $\mathrm{rank}(\Sigma) = p$, we can write

$$r_0(\Sigma) = \mathrm{rank}(\Sigma) s(\Sigma), \qquad R_0(\Sigma) = \mathrm{rank}(\Sigma) S(\Sigma),$$

$$\text{with} \qquad s(\Sigma) = \frac{1/p \sum_{i=1}^{p} \lambda_i}{\lambda_1}, \qquad S(\Sigma) = \frac{\left(1/p \sum_{i=1}^{p} \lambda_i\right)^2}{1/p \sum_{i=1}^{p} \lambda_i^2}.$$

Both s and S lie between $1/p$ ($\lambda_2 \approx 0$) and 1 (λ_i all equal).

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,

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

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

Intuition

- The mix of eigenvalues of Σ 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 assymmetric if there are many more than n of them).

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, 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?

Benign Overfitting: Proof Ideas

Bias-variance decomposition

Define the noise vector ϵ by $y = X\theta^* + \epsilon$.

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

$$\begin{aligned}\text{Excess risk:} \qquad R(\hat{\theta}) &= \left(\hat{\theta} - \theta^*\right)^\top \Sigma \left(\hat{\theta} - \theta^*\right) \\ &= {\theta^*}^\top \left(I - \hat{\Sigma}\hat{\Sigma}^\dagger\right)\left(\Sigma - \hat{\Sigma}\right)\left(I - \hat{\Sigma}^\dagger\hat{\Sigma}\right)\theta^* \\ &\quad + \sigma^2 \mathrm{tr}\left(\left(X^\top X\right)^\dagger \Sigma\right).\end{aligned}$$

Benign Overfitting: Proof Ideas

Standard normals

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

where $z_i = X v_i / \sqrt{\lambda_i}$ for $\Sigma = \sum_j \lambda_j v_j v_j^\top$, and

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

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 $n/R_{k^*}(\Sigma)$.

Benign Overfitting: Proof Ideas

Lower bound

- The excess expected loss is at least as big as the same trace term, $\mathrm{tr}\left(\left(X^\top X\right)^\dagger \Sigma\right)$.
- 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,

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

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

What kinds of eigenvalues?

We say Σ 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 Σ is benign iff $\alpha = 1$ and $\beta > 1$.

The λ_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 Σ_n is benign iff

- $p_n = \omega(n)$,
- $\epsilon_n p_n = o(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$. $(n \geq 40 \implies ne^{-n} < 2^{-52})$

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 λ_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.

(Li and Liang, 2018), (Du, Poczós, Zhai, Singh, 2018), (Du, Lee, Li, Wang, Zhai, 2018), (Arora, Du, Hu, Li, Wang, 2019).

- 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 Δ $\qquad {\scriptstyle \Delta \propto X^\top (XX^\top)^{-1}\epsilon}$

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

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

Benign overfitting leads to huge sensitivity.

Interpolating prediction

- 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?

Benign Overfitting in Linear Regression

- 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.