Benign Overfitting

Peter Bartlett
UC Berkeley

Posner Lecture
NeurIPS 2021

Benign Overfitting

Generalization theory for neural networks

- VC theory — Advances in Neural Information Processing Systems Volume 1
- Overparameterization and large-margin classification — Advances in Neural Information Processing Systems Volume 9
- Benign overfitting in deep learning

Advances in Neural Information Processing Systems 1 (NIPS 1988)

WHAT SIZE NET GIVES VALID GENERALIZATION?*

Eric B. Baum
Department of Physics
Princeton University
Princeton NJ 08540

David Haussler
Computer and Information Science
University of California
Santa Cruz, CA 95064

ABSTRACT

We address the question of when a network can be expected to generalize from m random training examples chosen from some arbitrary probability distribution, assuming that future test examples are drawn from the same distribution. Among our results are the following bounds on appropriate sample vs. network size. Assume $0 < \epsilon \leq 1/8$. We show that if $m \geq O(\frac{W}{\epsilon} log \frac{N}{\epsilon})$ random examples can be loaded on a feedforward network of linear threshold

Probabilistic Formulations of Prediction Problems

Given data $(x_1, y_1), \ldots, (x_n, y_n)$ (observation $x_i \in \mathcal{X}$, outcome $y_i \in \mathcal{Y}$)
Assume: Independent $(x_1, y_1), \ldots, (x_n, y_n), (x, y) \sim P$
(P is a probability distribution on $\mathcal{X} \times \mathcal{Y}$).
Choose $f : \mathcal{X} \to \mathcal{Y}$
so that $f(x)$ is a good prediction of y, in the sense that $\ell(f(x), y)$ small.**Aim:** Small *risk*: $\mathbb{E}\ell_f \ell_f := \mathbb{E}\ell(f(x), y)$.

Example: Pattern classification

$$\ell_{01}(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

Example: Empirical risk minimization

Choose $f \in \mathcal{F}$ to minimize

Classification in a Probabilistic Setting

Theorem (Vapnik and Chervonenkis, 1971)

Consider $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$, $\mathcal{Y} = \{\pm 1\}$, $\ell = \ell_{01}$.
For every prob distribution P on $\mathcal{X} \times \{-1,1\}$,
with high probability over n iid examples $(x_1, y_1), \ldots, (x_n, y_n)$,
every f in $\mathcal{F}$ satisfies

$$\mathbb{E}\ell_f \leq \hat{\mathbb{E}}\ell_f + O\left(\sqrt{\frac{d_{VC}(\mathcal{F}) d_{VC}(\mathcal{F})}{n}}\right).$$

- For neural networks, VC-dimension:
  - increases with number of parameters
  - depends on nonlinearity and depth

Advances in Neural Information Processing Systems 1 (NIPS 1988)

WHAT SIZE NET GIVES VALID GENERALIZATION?*

Eric B. Baum
Department of Physics
Princeton University
Princeton NJ 08540

David Haussler
Computer and Information Science
University of California
Santa Cruz, CA 95064

ABSTRACT

We address the question of when a network can be expected to generalize from m random training examples chosen from some arbitrary probability distribution, assuming that future test examples are drawn from the same distribution. Among our results are the following bounds on appropriate sample vs. network size. Assume $0 < \epsilon \leq 1/8$. We show that if $m \geq O(\frac{W}{\epsilon} log \frac{N}{\epsilon})$ random examples can be loaded on a feedforward network of linear threshold

VC-Dimension of Neural Networks

Theorem

Consider the class $\mathcal{F}$ of $\{-1, 1\}$-valued functions computed by a network with L layers, p parameters, and k computation units with the following nonlinearities:

1. Piecewise constant (linear threshold units): $d_{VC}(\mathcal{F}) = \tilde{\Theta}(p)$.
   (Baum and Haussler, 1989)
2. Piecewise polynomial: $d_{VC}(\mathcal{F}) = \tilde{O}\left(pL^2\right)$.
   (B., Maiorov, Meir, 1998)
3. Piecewise linear (ReLUs): $d_{VC}(\mathcal{F}) = \tilde{\Theta}(pL)$.
   (B., Harvey, Liaw, Mehrabian, 2017)
4. Sigmoid: $d_{VC}(\mathcal{F}) = \tilde{O}\left(p^2k^2\right)$.
   (Karpinsky and MacIntyre, 1994)

In all cases, d_{VC} is (at least) linear in number of parameters p.

Classification in a Probabilistic Setting

Theorem (Vapnik and Chervonenkis, 1971)

Consider $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$, $\mathcal{Y} = \{\pm 1\}$, $\ell = \ell_{01}$.
For every prob distribution P on $\mathcal{X} \times \{-1,1\}$,
with probability $1-\delta$ over n iid examples $(x_1, y_1), \ldots, (x_n, y_n)$,
every f in $\mathcal{F}$ satisfies

$$\mathbb{E}\ell_f \leq \hat{\mathbb{E}}\ell_f + O\left(\sqrt{\frac{d_{VC}(\mathcal{F})}{n}}\right).$$

- For neural networks, VC-dimension:
  - increases linearly with number of parameters **need $n \gg p$?**
  - depends on nonlinearity and depth
- For uniform bounds (that is, for all distributions and all $f \in \mathcal{F}$, proportions are close to probabilities; also for near-optimal prediction with $f \in \mathcal{F}$), this inequality is tight within a constant factor.

Some Experimental Observations

Neural networks

(Steve Lawrence, Lee Giles, Ah Chung Tsoi, 1997)

monly believed to be accurate. However, the stipulation that the number of parameters must be less than the number of examples is typically believed to be true for common datasets. The results here indicate that this is not always the case.

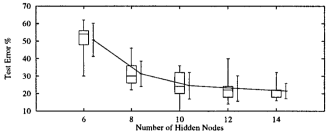


Figure 3. Face recognition example: the best generalizing network has 364 times more parameters than training points (18210 parameters).

Prediction accuracy improving with overparameterization.

Advances in Neural Information Processing Systems 9 (NIPS 1996)

For valid generalization, the size of the weights is more important than the size of the network

Peter L. Bartlett
Department of Systems Engineering
Research School of Information Sciences and Engineering
Australian National University
Canberra, 0200 Australia
Peter.Bartlett@anu.edu.au

Abstract

This paper shows that if a large neural network is used for a pattern classification problem, and the learning algorithm finds a network

Large-Margin Classification: Some Intuition

- Consider a real-valued function $f : \mathcal{X} \to \mathbb{R}$ used for classification.
- The prediction on $x \in \mathcal{X}$ is $\operatorname{sign}(f(x)) \in \{-1, 1\}$.
- Minimizing a continuous loss, such as $(f(x) - y)^2$.
- For a pattern-label pair $(x, y) \in \mathcal{X} \times \{-1, 1\}$, if $yf(x) > 0$ then f classifies x correctly.
- We call $yf(x)$ the *margin* of f on x.
- For large-margin classifiers, we might expect the fine-grained details of f (such as $d_{VC}(\mathcal{F})$) to be less important.

c.f. Support vector machines

(Boser, Guyon, Vapnik, 1992)

Large-Margin Classification with Two-Layer Networks

Theorem

(B, 1996)

Consider the following class $\mathcal{F}_B$ of two-layer neural networks defined on $\mathcal{X} = [-1, 1]^d$:

$$\mathcal{F}_B = \left\{ x \mapsto \sum_{i=1}^{k} w_i \sigma\left(v_i^T x\right) : \|w\|_1 \leq B,\ \|v_i\|_1 \leq B,\ k \geq 1 \right\},$$

where σ is 1-Lipschitz and bounded.
Then with high probability, for all $f \in \mathcal{F}_B$,

$$\mathbb{E}\ell_{01,f} \leq \hat{\mathbb{E}}\ell_{\gamma,f} \hat{\mathbb{E}}\ell_{\gamma,f} + \tilde{O}\left(\frac{B^3}{\gamma^2} \frac{B^3}{\gamma^2} \sqrt{\frac{\log d}{n}} \right).$$

Here, $\ell_{\gamma,f}(x, y) := 1[yf(x) \leq \gamma]$ penalizes margins that are less than γ.

Generalization: Margins and Size of Parameters

- A *classification* problem becomes a *regression* problem.
- For regression, the complexity of a neural network can be controlled by the *size* of the parameters, and can be independent of the number of parameters.
- We have a tradeoff between the fit to the training data (margins) and the complexity (size of parameters):

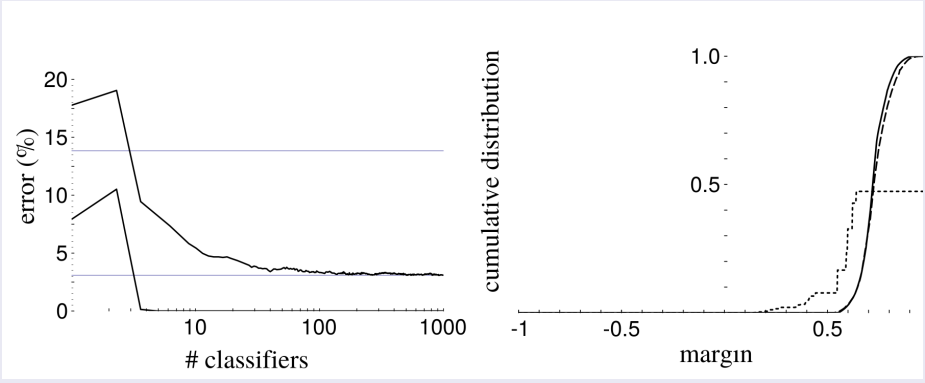
$$\mathbb{E}\ell_{01,f} \leq \hat{\mathbb{E}}\ell_{\gamma,f} + p_n(f)$$

- Even if $\hat{\mathbb{E}}\ell_{01,f} = 0$, it might be worthwhile to suffer an increased complexity penalty, $p_n(f)$, to improve $\hat{\mathbb{E}}\ell_{\gamma,f}$.

Some Experimental Observations

AdaBoost

(Rob Schapire, Yoav Freund, B, Wee Sun Lee, 1997)

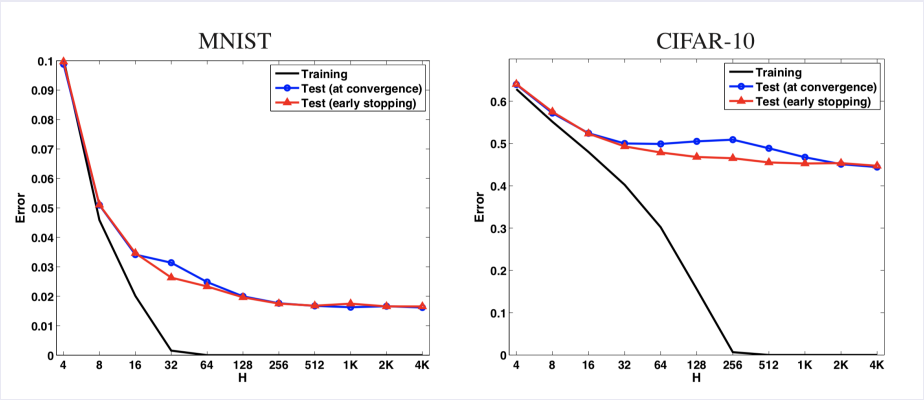


Prediction accuracy improving with overparameterization.

Some Experimental Observations

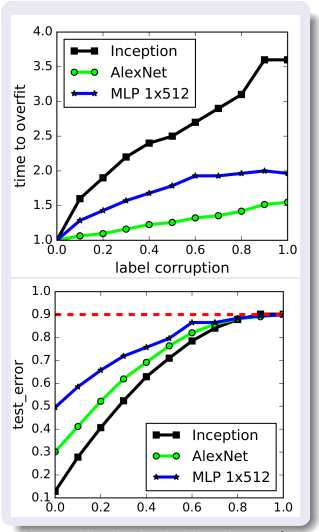
Neural networks

(Neyshabur, Tomioka, Srebro, 2015)



Prediction accuracy improving with overparameterization.

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, ICML 2018)

Anomalies in Deep Learning Practice

Deep learning methods predict accurately, but...

- They are overparameterized ($p \gg n$).
  Can control complexity some other way, e.g., scale of parameters.
- They minimize empirical risk with little or no explicit regularization.
  Can regularize implicitly, e.g., early stopping in gradient methods.
  - e.g., Yuan Yao, Lorenzo Rosasco, Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. 2007.
  - e.g., B, Mikhail Traskin, AdaBoost is Consistent, Advances in Neural Information Processing Systems 19, 2007.
- They find a perfect fit to the data.
  If f is not too complex and fits the data, i.e., $\hat{\mathbb{E}}\ell_f = 0$, then $\mathbb{E}\ell_f$ can be near zero too.
  - e.g., Leslie Valiant. A Theory of the Learnable. 1984.
- Deep learning methods find a perfect fit to *noisy* data.

Statistical Wisdom and Overfitting

"... interpolating fits... [are] unlikely to predict future data well at all."

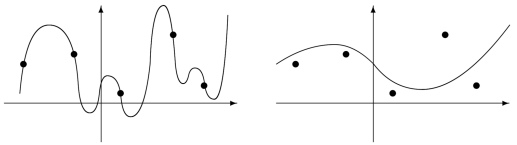
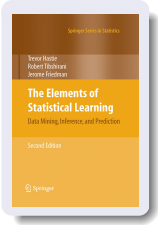
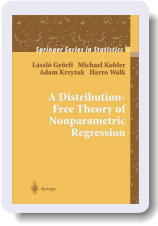


Figure 2.3. The estimate on the right seems to be more reasonable than the estimate on the left, which interpolates the data.

over $\mathcal{F}_n$. Least squares estimates are defined by minimizing the empirical L_2 risk over a general set of functions $\mathcal{F}_n$ (instead of (2.7)). Observe that it doesn't make sense to minimize (2.9) over all (measurable) functions f, because this may lead to a function which interpolates the data and hence is not a reasonable estimate. Thus one has to restrict the set of functions over



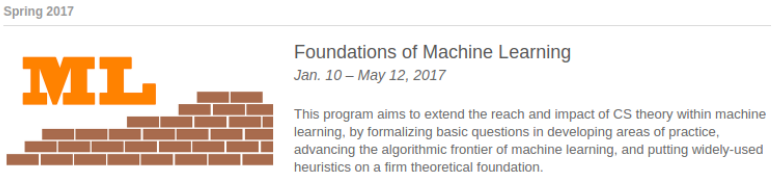


see also (B. and Rakhlin, Simons Institute, May 2019)

Benign Overfitting

A new statistical phenomenon:
good prediction with very small training error for regression loss

- 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.



Belkin, Hsu and Mitra, 2018; Belkin, Rakhlin and Tsybakov, 2018; Liang and Rakhlin, 2018;

Belkin, Hsu, Ma and Mandal, 2019; Belkin, Hsu and Xu, 2019; Hastie, Montanari, Rosset and Tibshirani, 2019; Dereziński, Liang and Mahoney, 2019; Liang, Rakhlin and Zhai, 2019; Mei and Montanari, 2019; Mitra, 2019; Muthukumar, Vodrahalli and Sahai, 2019; Nakkiran, 2019; Bunea, Strimas-Mackey, Wegkamp, 2020; Kobak, Lomond and Sanchez, 2020; Nakkiran, Venkat, Kakade and Ma, 2020; Hastie, Montanari, Rosset and Tibshirani, 2020; Mei, Misiakiewicz, Montanari, 2021; Celentano, Misiakiewicz, Montanari, 2021; Zou, Wu, Braverman, Gu and Kakade, 2021; Li, Zhou, Gretton, 2021;

Benign Overfitting

Intuition

- Benign overfitting prediction rule $\hat{f}$ decomposes as

$$\hat{f} = \hat{f}_0 + \Delta.$$

- $\hat{f}_0$ = simple component useful for *prediction*.
- Δ = spiky component useful for *benign overfitting*.
- Classical statistical learning theory applies to $\hat{f}_0$.
- Δ is not useful for prediction, but it is benign.

(Deep learning: a statistical viewpoint. B., Montanari, Rakhlin. *Acta Numerica*. 2021)

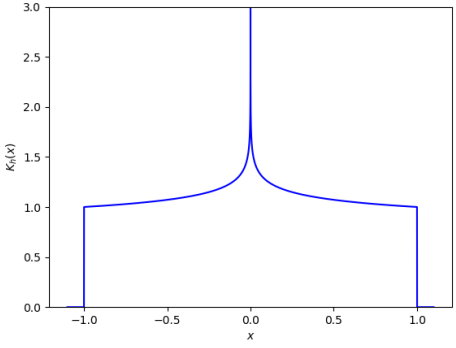
Benign Overfitting

Example: kernel smoothing with singular, compact kernels

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{y_i K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} \qquad \text{e.g., with } K_h(x) = \frac{1\left[h\|x\| \leq 1\right]}{h\|x\|^{\alpha}}.$$

Minimax rates (with suitable h).

(Belkin, Rakhlin, Tsybakov, 2018), (Belkin, Hsu, Mitra, 2018)



Outline

- *Linear regression*
- Characterizing benign overfitting
- Adversarial examples
- Ridge regression
- Model-dependent bounds

Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- Assume: (x, y) subgaussian, mean zero, well-specified
  x satisfies a small ball condition
- Define:

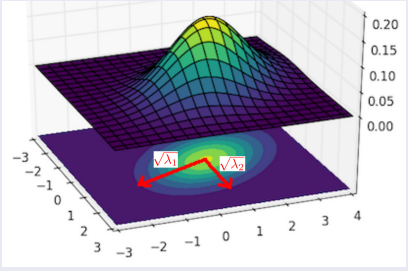
$$\mathbb{E}[y|x] = x^\top \theta^*$$

$$\exists c > 0,\ \Pr\left(\|x\|^2 < c\mathbb{E}\|x\|^2\right) \leq \delta.$$

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

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

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



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

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

$$X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Notice that gradient flow, initialized at 0:

$$\theta_0 = 0, \qquad \dot{\theta}_t = -\nabla_\theta \|X\theta - y\|^2$$

converges to the minimum norm solution.

Definitions

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)} \left(y - x^\top \hat{\theta} \right)^2 - \underbrace{\min_\theta \mathbb{E} \left(y - x^\top \theta \right)^2}_{\text{optimal prediction error}}$$

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

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

So Σ determines the importance of parameter directions.

(Recall that $\Sigma = \sum_i \lambda_i v_i v_i^\top$ for orthonormal v_i, $\lambda_1 \geq \lambda_2 \geq \cdots$.)

Benign Overfitting

Outline

- Linear regression
- *Characterizing benign overfitting*
- Adversarial examples
- Ridge regression
- Model-dependent bounds

From regularization to overfitting

Regularized linear regression

$$\begin{aligned}
&\text{ridge regression:} && \min && \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2, \\
&\text{norm constrained:} && \min && \frac{1}{n}\|X\theta - y\|^2 \\
& && \text{s.t.} && \|\theta\| \le b, \\
&\text{fit constrained:} && \min && \|\theta\| \\
& && \text{s.t.} && \frac{1}{n}\|X\theta - y\|^2 \le c.
\end{aligned}$$

- The overfitting regime: $c \ll \min_\theta \mathbb{E}\left(y - x^\top\theta\right)^2$.

Interpolating Linear Regression

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.
- Estimator $\hat{\theta} = \left(X^\top X\right)^\dagger X^\top y$ solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0. \end{aligned}$$

- 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

(B., Long, Lugosi, Tsigler, 2019), (Tsigler, B., 2020)

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

1. With high probability,

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

2. If $X = \Sigma^{1/2}Z$ where Z has independent components and θ^* is symmetrized (random sign flips of components),

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

Here, $\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k^*+1:\infty}\|^2_{\Sigma_{k^*+1:\infty}} + \|\theta^*_{1:k^*}\|^2_{\Sigma^{-1}_{1:k^*}}\left(\frac{\sum_{i>k^*}\lambda_i}{n}\right)^2$.

Effective Rank

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**

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

Examples

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

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

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

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

1. With high probability,

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

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

$$\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.$$

Benign Overfitting

- Benign overfitting prediction rule $\hat{f}$ decomposes as

$$\hat{f} = \hat{f}_0 + \Delta.$$

- $\hat{f}_0$ = *prediction* component:
  k^*-dim subspace corresponding to $\lambda_1, \ldots, \lambda_{k^*}$.
- Δ = *benign overfitting* component:
  orthogonal subspace. Δ is benign only if $R_{k^*} \gg n$.

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,
  - 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: What kinds of eigenvalues?

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,

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

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

What kinds of eigenvalues?

We say $\{\Sigma_n\}$ is *asymptotically benign* if

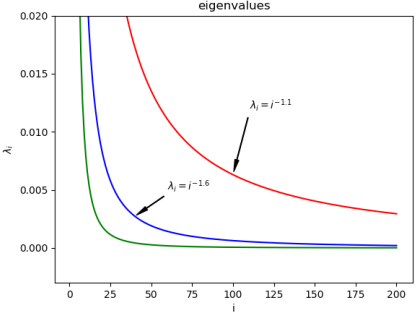
$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

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

Example

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

The $\sum_i \lambda_i$ must almost diverge!!?!



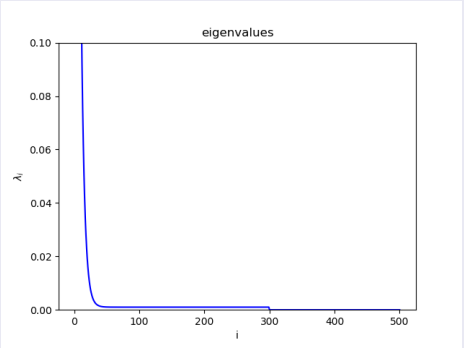
What kinds of eigenvalues?

Example: *Finite dimension*, fast λ_i decay, 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 $\epsilon_n p_n = \omega(ne^{-n})$.



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

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

Generic phenomenon: quickly converging λ_i plus noise in all directions, $p_n \gg n$.

Outline

- Linear regression
- Characterizing benign overfitting
- *Adversarial examples*
- Ridge regression
- Model-dependent bounds

Adversarial Examples in Deep Networks

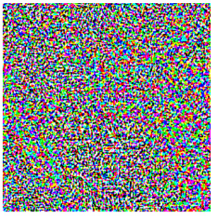
Intriguing properties of neural networks. Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow and Fergus. ICLR2014

"Panda" or "Gibbon"?

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

$=$

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

(Goodfellow, Shlens and Szegedy, ICLR2015)

Implications for adversarial examples in linear regression

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

We can find a unit norm Δ $\quad \Delta \propto X^\top (XX^\top)^{-1} \epsilon$

such that perturbing an input x by Δ changes the output enormously:
even if $\Delta^\top \theta^* = 0$, if $R(\hat{\theta}) \leq \alpha$,

$$\left((x+\Delta)^\top \hat{\theta} - x^\top \hat{\theta}\right)^2 \geq c \frac{\sigma^2 (n-k^*)}{\sum_{i>k^*} \lambda_i} \geq c \frac{\sigma^2 \|\theta^*_{1:k^*}\|}{\sqrt{\lambda_1}} \frac{1-\alpha}{\sqrt{\alpha}}.$$

Benign overfitting leads to huge sensitivity.

$\Delta =$ *benign overfitting* component is spiky:
$\|\Delta\|_{L_\infty}$ large; $\|\Delta\|_{L_2(P)}$ small.

Outline

- Linear regression
- Characterizing benign overfitting
- Adversarial examples
- *Ridge regression*
- Model-dependent bounds

Ridge Regression

Minimum norm ridge regression

$$\hat{\theta}_\lambda = X^\top \left(XX^\top + \lambda I\right)^{-1} y = \arg\min \quad \|\theta\|$$
$$\text{s.t.} \quad \theta \in \arg\min_\beta \left\{\|X\beta - y\|^2 + \lambda\|\beta\|_2^2\right\}$$

- Covers the range of solutions, from overfitting to regularized.
- Tight bounds on bias and variance for $\lambda \in \mathbb{R}$.
- Effective ranks, r_k and R_k, replaced by

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

- In some cases ($r_{k^*}(\Sigma) \gg n$), the optimal λ is *negative*: this decreases bias without significantly affecting variance.

Ridge Regression

Theorem

(Tsigler and B., 2020)

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

1. With high probability,

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

2. If $X = \Sigma^{1/2}Z$ where Z has independent components and the components of θ^* are subject to random sign flips,

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

Here, $\text{bias}(\theta^*, \Sigma, n, \lambda) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2$.

Outline

- Linear regression
- Characterizing benign overfitting
- Adversarial examples
- Ridge regression
- *Model-dependent bounds*

Model-Dependent Bounds and Benign Overfitting

Model-dependent bounds

- Suppose an algorithm returns a prediction rule f that
  - has $\hat{\mathbb{E}}\ell_f$ small, and
  - is simple (e.g., small spectral norms/approx rank of parameters/etc).
- When does this ensure that $\mathbb{E}\ell_f \leq \epsilon(f, n)$?
- e.g., via uniform deviation bounds: $\sup_{f \in F} \left| \hat{\mathbb{E}}\ell_f - \mathbb{E}\ell_f \right| \leq \epsilon(F, n)$.

Limitations of uniform convergence

(Nagarajan and Kolter, 2019)

- Sometimes uniform deviation bounds can't help.
- For a certain classification problem:
  - One gradient descent step gives an f with $\hat{\mathbb{E}}\ell_f = 0$ ... and f predicts accurately,
  - And symmetry of the data distribution implies that any uniform deviation bound for an F containing f must be large.

Model-Dependent Bounds and Benign Overfitting

Definitions

Consider the minimum norm estimator $\hat{\theta}$ in a linear regression problem.

1. A *uniform model-dependent bound* ϵ satisfies, for all sample sizes n and all mean-zero 1-subgaussian distributions P, with high probability
$$R(\hat{\theta}) \leq \epsilon(\hat{\theta}, n)$$
2. It is *bounded-antimonotonic* if, for $n_1 \leq n_2 \leq 2n_1$, $\epsilon(h, n_2) \leq c\epsilon(h, n_1)$.
3. A set $\mathcal{B} \subset \mathbb{N}$ includes "most n" if it is *strongly* $(1-\delta)$*-dense beyond* n_0, that is, $s^2 \geq n_0$ implies
$$\frac{|\mathcal{B} \cap \{s^2, ..., (s+1)^2 - 1\}|}{2s+1} \geq 1 - \delta.$$

Model-Dependent Bounds and Benign Overfitting

Theorem (B. and Long, 2020)

There are distributions D_n on $\mathbb{R}^{d_n} \times \mathbb{R}$ (X gaussian, y subgaussian) s.t.
if ϵ is a bounded-antimonotonic, uniform model-dependent bound,
then with high probability the minimum norm estimator $\hat{\theta}$ satisfies

$$R(\hat{\theta}) \lesssim \frac{1}{\sqrt{n}}$$

but nonetheless, for most n,

$$\Pr_{S \sim D_n^n} \left(\epsilon(\hat{\theta}, n) > c \right) \geq \frac{1}{2}.$$

- Natural joint distributions on training examples,
- Any ϵ that does not increase too quickly with n must sometimes be very loose: *it needs more information about the distribution.*

Model-Dependent Bounds and Benign Overfitting

Proof idea

Interpolating prediction rules are bad estimates on the training sample:

$$x_i^\top \hat{\theta} = y_i, \qquad \text{so} \qquad \mathbb{E}[(x_i^\top \hat{\theta} - x_i^\top \theta^*)^2] = \sigma^2,$$

A Poissonization approach shows that two situations are essentially indistinguishable:

1. The training sample forms a significant fraction of the support of the distribution.
2. A benign overfitting situation, where the training sample has measure zero.

A bound that is valid in both cases must be loose in the second case.

Benign Overfitting

- 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.
  - Linear prediction splits into a simple (k^*-dim) component and a benign overfitting component
  - Relies on many (roughly equally) unimportant parameter directions
  - Finite dimensional data is important:
    infinite dimension requires specific eigenvalue decay;
    it is a generic phenomenon for truncated slow decay.
- But it leads to huge sensitivity to (adversarial) perturbations.
- From interpolation to ridge regression
- Limitations of model-dependent bounds

Benign Overfitting

Next steps → beyond linear

- Linear regression: with two-layer linear networks

  (Niladri Chatterji, Phil Long, B, 2021)

  building on implicit bias results of (Azulay et al, 2021)

- Linear regression: beyond minimum Euclidean norm

  (Freddie Koehler, Lijia Zhou, Danica Sutherland and Nati Srebro, NeurIPS 2021)

- Neural networks as linear function classes:
  neural tangent kernels, random feature models

  (Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai, 2020)

  (Song Mei, Theodor Misiakiewicz, and Andrea Montanari, 2021)

- Benign overfitting in deep networks. $\hat{f} = \hat{f}_0 + \Delta$?

Benign Overfitting

SIMONS FOUNDATION

Niladri Chatterji

Phil Long

Gábor Lugosi

Andrea Montanari

Alexander Rakhlin

Alexander Tsigler

- Benign overfitting in linear regression. B., Long, Lugosi, Tsigler. PNAS 117(48):30063–30070, 2020. arXiv:1906.11300
- Benign overfitting in ridge regression. Tsigler, B. arXiv:2009.14286
- Failures of model-dependent generalization bounds for least-norm interpolation. B., Long. JMLR 22(204):1–15, 2021. arXiv:2010.08479
- Deep learning: a statistical viewpoint. B., Montanari, Rakhlin. Acta Numerica 30:87-201, 2021. arXiv:2103.09177
- The interplay between implicit bias and benign overfitting in two-layer linear networks. Chatterji, Long, B. arXiv:2108.11489