Accurate Prediction From Interpolation: A New Challenge for Statistical Learning Theory

Peter Bartlett
CS and Statistics
UC Berkeley

March 14, 2019

SIMONS INSTITUTE
for the Theory of Computing

The statistics of deep learning: fit versus complexity

Use training data to choose parameters

The statistics of deep learning: fit versus complexity

Typical theorem

prediction error $\leq$ training error + complexity penalty

Outline

- Empirical process theory for classification**Empirical process theory for classification**
- Margins analysis: relating classification to regression
- Interpolation: Where is the tradeoff between fit and complexity?
- Interpolation in linear regression

Problem formulation

- Independent training data $(x_1, y_1), \ldots, (x_n, y_n) \sim P$
- Prediction rule $f : \mathcal{X} \to \mathcal{Y}$
- Loss $\ell(y, f(x))$
- Prediction error: $\mathbb{E}\ell(y, f(x))$

VC Theory

Theorem (Vapnik and Chervonenkis)

Suppose $\mathcal{F} \subseteq \{-1, 1\}^{\mathcal{X}}$.
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

$$P(f(x) \neq y) \leq \frac{1}{n} |\{i : f(x_i) \neq y_i\}| + \left(\frac{c}{n}(\mathrm{VCdim}(\mathcal{F}) + \log(1/\delta))\right)^{1/2}.$$

- For uniform bounds (that is, for all distributions and all $f \in \mathcal{F}$, proportions are close to probabilities), this inequality is tight—within a constant factor.
- For neural networks, VC-dimension:
  - increases with number of parameters
  - depends on nonlinearity and depth

VC-Dimension of Neural Networks

Theorem

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

1. Piecewise constant (linear threshold units): $\mathrm{VCdim}(\mathcal{F}) = \tilde{\Theta}(p)$.

   (Baum and Haussler, 1989)

2. Piecewise linear (ReLUs): $\mathrm{VCdim}(\mathcal{F}) = \tilde{\Theta}(pL)$.

   (B., Harvey, Liaw, Mehrabian, 2017)

3. Piecewise polynomial: $\mathrm{VCdim}(\mathcal{F}) = \tilde{O}\left(pL^2\right)$.

   (B., Maiorov, Meir, 1998)

Generalization in Neural Networks: Number of Parameters

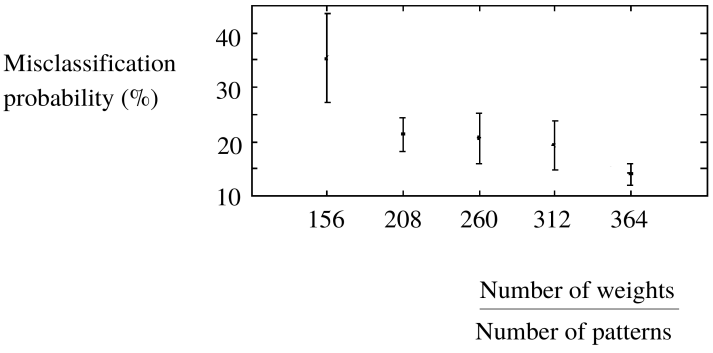
NIPS 1996

Experimental Results

Neural networks with many parameters, trained on small data sets, sometimes generalize well.

Eg: Face recognition (Lawrence *et al*, 1996)

$m = 50$ training patterns.



Generalization: Margins and Size of Parameters

With high probability over n training examples $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{\pm 1\}$, every $f \in \mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ has

$$\Pr(\mathrm{sign}(f(X)) \neq Y) \leq \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + p_n(\mathcal{F})$$

1. (B., 1996) $\ell = L$-Lipschitz approximation of step function;
$p_n =$ (scale-sensitive dimension) $= \tilde{O}((LB)^d/\sqrt{n})$.
2. (B. and Mendelson, 2000)
$p_n =$ (Rademacher averages) $= \tilde{O}(LB^d/\sqrt{n})$.
3. (B., Foster and Telgarsky, 2017)
$p_n =$ (covering numbers) $= \tilde{O}(LRd/\sqrt{n})$.

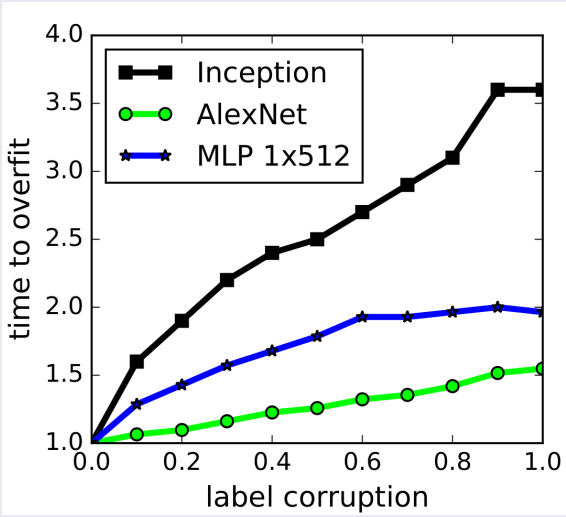
- The bound combines training error (with the L-Lipschitz surrogate loss ℓ) with a complexity penalty, $p_n(\mathcal{F})$.
- $p_n(\mathcal{F})$ need not grow with the number of parameters.
- e.g., $\mathcal{F} = d$-layer sigmoid networks with each unit's weights bounded

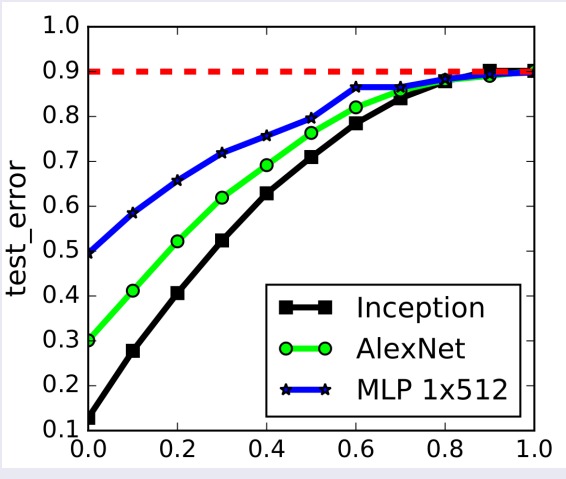
see also (Neyshabur, Tomioka and Srebro, 2015), (Golowich, Rakhlin and Shamir, 2018)

Outline

- Empirical process theory for classification
- Margins analysis: relating classification to regression
- **Interpolation: Where is the tradeoff between fit and complexity?**
- Interpolation in linear regression

Interpolation in Deep Networks: A New Challenge for Statistical Learning Theory





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

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

$$\Pr(\mathrm{sign}(f(X)) \neq Y) \leq \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) + p_n(\mathcal{F})$$

(Belkin, Hsu, Ma, Mandal, 2018)

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)

Interpolation in Linear Regression

Phil Long

Gábor Lugosi

Alexander Tsigler

Linear regression

- Training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$.
- Linear functions: $f_\theta(x) = x^\top \theta$.
- Squared error: $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.
- Least squares linear prediction: θ^* minimizes $\mathbb{E}\ell(y, f_\theta(x))$.
- Choose $\hat{\theta}$ to interpolate: $\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) = 0$.

  Hence, $y_1 = f_{\hat{\theta}}(x_1), \ldots, y_n = f_{\hat{\theta}}(x_n)$ (need $p \geq n$).
- Which interpolating f_θ? Choose $\hat{\theta}$ to minimize $\|\theta\|$.

Interpolation in Linear Regression

Interpolation for linear prediction

- Excess expected loss, $\mathbb{E}\ell(y, f_{\hat{\theta}}(x)) - \mathbb{E}\ell(y, f_{\theta^*}(x))$ has two components:
  (corresponding to $f_{\theta^*}(x)$ and $y - f_{\theta^*}(x)$)
  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 in high dimensions.
- When can we hide the label noise in $\hat{\theta}$ without hurting predictive accuracy?

Interpolation in Linear Regression

Accurate interpolating prediction as dimension p_n grows

- Suppose the covariance of x is in two pieces:
  - a constant piece (of dimension k), and
  - a 'tail' (of dimension $p_n - k$)) that gets longer and flatter with n.
- Denote the variance in the 'tail' directions $\gamma_1 \geq \cdots \geq \gamma_{p_n - k}$.

Theorem

If the 'tail' is long and flat:

- a small proportion of variance in any direction, $\dfrac{\gamma_1}{\sum_i \gamma_i} = o(1/n)$,
- total variance $\sum_i \gamma_i = o(n)$,

then for jointly gaussian (x, y), with high probability,

$$\mathbb{E}\ell(y, f_{\hat{\theta}}(x)) - \mathbb{E}\ell(y, f_{\theta^*}(x)) = O\left(\left(\frac{k}{n}\right)^{1/4} + \left(\frac{n\gamma_1}{\sum_i \gamma_i}\right)^{1/4}\right) \to 0.$$

Interpolating prediction

- Interpolation: far from the regime of a tradeoff between fit to training data and complexity.
- In high-dimensional linear regression, if the covariance has a long, flat tail, the minimum norm interpolant can hide the noise in these many unimportant directions.
  - Relies on overparameterization
  - ... and lots of unimportant parameters
- Can we extend these results to interpolating deep networks?
  - What is the analog of the minimum norm linear prediction rule?
  - What role does the optimization method play?