Generalization in Deep Networks

Peter Bartlett

UC Berkeley

December 9, 2017
What determines the statistical complexity of a deep network?

- VC theory: Number of parameters
- Margins analysis: Size of parameters
- Understanding generalization failures
Outline

What determines the statistical complexity of a deep network?

- **VC theory:** *Number of parameters*
- Margins analysis: *Size of parameters*
- Understanding generalization failures
Assume network maps to $\{-1, 1\}$. (Threshold its output)

Data generated by a probability distribution $P$ on $\mathcal{X} \times \{-1, 1\}$.

Want to choose a function $f$ such that $P(f(x) \neq y)$ is small (near optimal).
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

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

- 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
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):
   \[
   \text{VCdim}(\mathcal{F}) = \tilde{O}(p).
   \]
   (Baum and Haussler, 1989)

2. Piecewise linear (ReLUs):
   \[
   \text{VCdim}(\mathcal{F}) = \tilde{O}(pL).
   \]
   (B., Harvey, Liaw, Mehrabian, 2017)

3. Piecewise polynomial:
   \[
   \text{VCdim}(\mathcal{F}) = \tilde{O}(pL^2).
   \]
   (B., Maiorov, Meir, 1998)

4. Sigmoid:
   \[
   \text{VCdim}(\mathcal{F}) = \tilde{O}(p^2k^2).
   \]
   (Karpinsky and Maclntyre, 1994)
Neural networks with many parameters, trained on small data sets, sometimes generalize well.

Eg: Face recognition (Lawrence et al, 1996)

$m = 50$ training patterns.
What determines the statistical complexity of a deep network?
- VC theory: Number of parameters
- Margins analysis: Size of parameters
- Understanding generalization failures
Consider a real-valued function $f : \mathcal{X} \rightarrow \mathbb{R}$ used for classification.

The prediction on $x \in \mathcal{X}$ is $\text{sign}(f(x)) \in \{-1, 1\}$.

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 $x$.

We can view a larger margin as a more confident correct classification.

Minimizing a continuous loss, such as

$$
\sum_{i=1}^{n} (f(X_i) - Y_i)^2,
$$

encourages large margins.

For large-margin classifiers, we should expect the fine-grained details of $f$ to be less important.
Generalization: Margins and Size of Parameters

Theorem (B., 1996)

1. 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

$$\Pr(\text{sign}(f(X)) \neq Y) \leq \frac{1}{n} \sum_{i=1}^{n} 1[Y_i f(X_i) \leq \gamma] + \tilde{O}\left(\sqrt{\text{fat}_{\mathcal{F}}(\gamma)} \frac{n}{n}\right).$$

2. If functions in $\mathcal{F}$ are computed by two-layer sigmoid networks with each unit’s weights bounded in $1$-norm, that is, $\|w\|_1 \leq B$, then

$$\text{fat}_{\mathcal{F}}(\gamma) = \tilde{O}((B/\gamma)^2).$$

- The bound depends on the margin loss plus an error term.
- Minimizing quadratic loss or cross-entropy loss leads to large margins.
- $\text{fat}_{\mathcal{F}}(\gamma)$ is a scale-sensitive version of VC-dimension. Unlike the VC-dimension, it need not grow with the number of parameters.
Generalization: Margins and Size of Parameters

1996: Sigmoid networks

- Qualitative behavior explained by small weights theorem.

2017: Deep ReLU networks

- How to measure the complexity of a ReLU network?
What determines the statistical complexity of a deep network?

- VC theory: Number of parameters
- Margins analysis: Size of parameters
- Understanding generalization failures
Explaining Generalization Failures

CIFAR10

6: frog
9: truck
9: truck
4: deer
1: automobile

1: automobile
2: bird
7: horse
8: ship
3: cat

4: deer
7: horse
7: horse
2: bird
9: truck

9: truck
9: truck
3: cat
2: bird
6: frog

http://corochann.com/
Explaining Generalization Failures

Training margins on CIFAR10 with true and random labels

- How does this match the large margin explanation?
- Need to account for the *scale* of the neural network functions.
- What is the appropriate notion of the size of these functions?
arXiv:1706.08498
New results for generalization in deep ReLU networks

- Measuring the size of functions computed by a network of ReLUs. (c.f. sigmoid networks: the output $y$ of a layer has $\|y\|_\infty \leq 1$, so $\|w\|_1 \leq B$ keeps the scale under control.)
- Large multiclass versus binary classification.

Definitions

- Consider operator norms: For a matrix $A_i$,

$$\|A_i\|_* := \sup_{\|x\| \leq 1} \|A_i x\|.$$  

- Multiclass margin function for $f : \mathcal{X} \to \mathbb{R}^m, y \in \{1, \ldots, m\}$:

$$M(f(x), y) = f(x)_y - \max_{i \neq y} f(x)_i.$$
Generalization in Deep Networks

**Theorem**

With high probability, every \( f_A \) with \( R_A \leq r \) satisfies

\[
\Pr(M(f_A(X), Y) \leq 0) \leq \frac{1}{n} \sum_{i=1}^{n} 1[M(f_A(X_i), Y_i) \leq \gamma] + \tilde{O}\left(\frac{rL}{\gamma \sqrt{n}}\right).
\]

**Definitions**

Network with \( L \) layers, parameters \( A_1, \ldots, A_L \):

\[
f_A(x) := \sigma_L(A_L \sigma_{L-1}(A_{L-1} \cdots \sigma_1(A_1 x) \cdots)).
\]

Scale of \( f_A \): \( R_A := \prod_{i=1}^{L} \|A_i\|_* \left(\sum_{i=1}^{L} \frac{\|A_i\|_{2,1}^{2/3}}{\|A_i\|_*^{2/3}}\right)^{3/2}. \)

(Assume \( \sigma_i \) is 1-Lipschitz, inputs normalized.)
Explaining Generalization Failures

Stochastic Gradient Training Error on CIFAR10

(Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, 2017)
How does this match the large margin explanation?
If we rescale the margins by $R_A$ (the scale parameter):
Explaining Generalization Failures

If we rescale the margins by $R_A$ (the scale parameter):

Rescaled cumulative margins on MNIST
Theorem

With high probability, every \( f_A \) with \( R_A \leq r \) satisfies

\[
\Pr(M(f_A(X), Y) \leq 0) \leq \frac{1}{n} \sum_{i=1}^{n} 1[M(f_A(X_i), Y_i) \leq \gamma] + \tilde{O}\left(\frac{r^L \gamma}{\sqrt{n}}\right).
\]

Network with \( L \) layers, parameters \( A_1, \ldots, A_L \):

\[
f_A(x) := \sigma(A_L \sigma_{L-1}(A_{L-1} \cdots \sigma_1(A_1 x) \cdots)).
\]

Scale of \( f_A \): \( R_A := \prod_{i=1}^{L} \|A_i\|_* \left(\sum_{i=1}^{L} \frac{\|A_i\|_2^{2/3}}{\|A_i\|_*^{2/3}}\right)^{3/2}.
\]
Explaining Generalization Failures

epoch 10
epoch 100

- green: cifar Lipschitz
- blue: cifar [random] Lipschitz
Explaining Generalization Failures

- epoch 10
- epoch 100
- excess risk 0.3
- excess risk 0.9
- cifar excess risk
- cifar Lipschitz
- cifar [random] excess risk
- cifar [random] Lipschitz
With appropriate normalization, the margins analysis is qualitatively consistent with the generalization performance.

Margin bounds extend to residual networks.

Lower bounds?

Regularization: explicit control of operator norms?

Role of depth?

Interplay with optimization?
Outline

- What determines the statistical complexity of a deep network?
  - VC theory: Number of parameters
  - Margins analysis: Size of parameters
  - Understanding generalization failures