# Generalization Bounds for Neural Networks

Maria-Florina (Nina) Balcan

Carnegie Mellon University

# Sample Complexity: Infinite Hypothesis Spaces
## Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

**Theorem**

$$m = O\left(\frac{1}{\varepsilon^2}\left[VCdim(H) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $|err_D(h) - err_S(h)| \leq \epsilon$.

2) Statistical Learning Theory style:

With prob. at least $1-\delta$, for all $h \in H$:

$$\mathrm{err}_D(h) \leq \mathrm{err}_S(h) + O\left(\sqrt{\frac{1}{2m}\left(\mathrm{VCdim}(H) + \ln\left(\frac{1}{\delta}\right)\right)}\right).$$

Tight bounds in the worst case.

# VC-Dimension of Neural Networks

**Theorem:** $H$ class of neural networks with $L$ layers, $W$ weights.

- Piecewise constant (linear threshold units): $\mathrm{VCdim}(H) = \tilde{O}(W)$.
  [Baum-Haussler, 1989]
- Piecewise linear (ReLUs): $\mathrm{VCdim}(H) = \tilde{O}(WL)$.
  [Bartlett-Harvey-Liaw-Mehrabian, 2017]
- Piecewise polynomial: $\mathrm{VCdim}(H) = \tilde{O}(WL^2)$.
  [Bartlett-Maiorov-Meir, 1998]

(Note: all final output values thresholded to $\{-1,1\}$)

Nearly tight bounds.

Classic VCdim bounds have a strong explicit dependence on # of parameters in the network.

Trivial if # of parameters exceeds the number of examples.

# Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Stronger Data-Dependent Bounds
- Algorithm Does Implicit Regularization (finds local optima with special properties)
- Transfer Learning

# Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Stronger Data-Dependent Bounds
- Algorithm Does Implicit Regularization (finds local optima with special properties)
- Transfer Learning

# Data Dependent Generalization Bounds

- Distribution/data dependent. Tighter for nice distributions.

## Covering Numbers Generalization Bounds

See Anthony-Bartlett, “Neural Network Learning: Theoretical Foundations”, 1999.

## Rademacher Complexity Generalization Bounds

See Bousquet-Boucheron-Lugosi, “Introduction to Statistical Learning Theory”, 2014.

# Data-Dependent Bounds for Deep Networks

E.g., very recent papers:

- Via covering numbers: "Spectrally-normalized margin bounds for neural networks". [Bartlett-Foster-Telgarsky, NIPS 2017]

- Via Rademacher complexity: "Size-independent sample complexity of neural networks". [Golowich-Rakhlin-Shamir, COLT 2018]

# Data-Dependent Bounds for Deep Networks

- Spectrally-normalized margin bounds for neural networks. [Bartlett-Foster-Telgarsky, NIPS 2017]

**Theorem:** With high probability, every $f_W$ with $R_W \le R$ satisfies

$$\Pr(M(f_W(X), Y) \le 0) \le \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[M(f_W(X_i), Y_i) \le \gamma] + \tilde{O}\left(\frac{RL}{\gamma\sqrt{n}}\right)$$

- Network with $L$ layers, parameters $W_1, \dots, W_L$:

$$f_W(x) := \sigma(W_L \sigma_{L-1}(W_{L-1} \dots \sigma_1(W_1 x) \dots))$$

$$R_W := \prod_{i=1}^{L} \|W_i\|_* \left(\sum_{i=1}^{L} \frac{\|W_i\|_{2,1}^{2/3}}{\|W\|_*^{2/3}}\right)^{3/2}$$

[$\sigma$ is 1-Lipschitz]

spectral norm

[Golowich-Rakhlin-Shamir, COLT 2018] provide a related bound via a Rademacher complexity argument

# Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Stronger Data-Dependent Bounds

- Algorithm Does Implicit Regularization (finds local optima with special properties)

  "Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations". [Li-Ma-Zhang. COLT 2018]

- Transfer Learning

  "Risk Bounds for Transferring Representations With and Without Fine-Tuning". [McNamara-Balcan. ICML 2017]

# Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Str
- Alg
  with

  "A
  N

- Tr

  "F
  T

Lots of open questions.