Mathematics of Deep Learning
Lecture 11: Generalization Bounds: Regularized "A Priori", and Unregularized "A Posteriori" Case Using VC Dimension

Prof. Predrag R. Jelenković

Time: Tuesday 4:10-6:40pm

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
Office: 812 Schapiro Research Bldg.
Phone: (212) 854-8174
Email: predrag@ee.columbia.edu
URL: http://www.ee.columbia.edu/∼predrag

Final Project

Rough Paper Outline, about 15 pages:

1. Introduction: e.g., describe the general problem area, DL and specific subtopic(s). Brief literature review, etc.
2. Detailed Problem Description: More detailed literature review for a selected problem(s), detailed description of the known results, theoretical or experimental, etc.
3. Some Reproduction: Theoretical or Experimental partial or full reproduction of the results. For example, run some simulations that illustrate main results.
4. New Results: Theoretical or Experimental Describe in detail your results. If experimental, describe the experiments and results. Explain clearly the graphs and tables from experiments, etc.
5. Discussion and conclusion: e.g., try to draw general inferences from your results. Compare to the known results from the literature, etc.

Final Project

The key difference from other courses and guiding questions:

- What did you learn about a neural network?
  - The focus should be on NN properties instead of applications.
- How do the changes in NN impact its performance?
  - The changes could be: architecture (e.g., width/depth), activation function, training method, normalization, dropout...
  - In class, we focused on plain vanilla feed-forward networks, but you could choose other types, e.g., ResNets.
- You could center your questions on one or more of the general themes we focused on in class:
  1. Approximation and interpolation theory and the impact of depth.
  2. Optimization landscape and global convergence.
  3. Generalization theory: conditions for small/bounded testing errors.
- Many of the problems we formulated in the context of wide/over-parametrized networks with two types of scaling: NTK/lazy training or mean-field/active training.

Final Project

- **Deliverables**:
  - **Paper**: about 15 pages - the most important part.
  - **Presentations**: about 10min each, 10 slides
  - **Software**: Document your code well
- **First slot for presentatios**: April 25, during the last class
  3% EC for those presenting on April 25.
- additional presentation slots during study/exam week: TBA
- **Project due**: During the exam week of May 5-12: TBA
- Academic Honesty - do not plagiarize; Turnitin will be used to check for originality

Have Fun and Good Luck!

Bounding Generalization Error

The main objective of statistical learning is to predict well on future data.

- How do we measure the error?
  - Regression: Quadratic error/loss

$$\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$$

  - Classification: $$\ell(\hat{y}, y) = 1_{\{\hat{y} \neq y\}}$$

- $\{\mathcal{H}\}$: Hypothesis class of functions, e.g., all functions, $h(x, w)$, that can be generated by a NN of a certain architecture.

- Empirical Risk/Loss - total training error: For a data sample $S = \{(x_i, y_i)\}_{i=1}^n$ and $h \in \mathcal{H}$

$$\hat{L}_n = \hat{L}_n(h) \equiv L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Shai's book [ML2014] uses $L_S(h)$ notation; we'll use these interchangeably.

Empirical Risk Minimization and True Error

During training one typically minimizes the empirical risk (ERM), and obtain $\hat{h}_n \equiv h_S$

$$\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \hat{L}(h)$$

True Risk=Population Risk

- Let $x \in \mathcal{X}, y \in \mathcal{Y}$, say $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} \subset \mathbb{R}$, $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Define a probability measure on $\mathcal{Z}$, denoted by $\mathcal{D}$ in [ML2014] book
- Let $\{z_i = (x_i, y_i)\}_{i=1}^n$ be i.i.d. random variables on a product probability space $\mathcal{Z}^n$.
- Then, for $h \in \mathcal{H}$, we define a true risk/population risk as

$$L(h) \equiv L_{\mathcal{D}}(h) := \mathbb{E}_{x,y} \ell(h(x), y)$$

Probably Approximately Correct (PAC) Learning

The ultimate goal is to find h that minimizes the true risk, i.e.,

$$h \in \arg\min_{h \in \mathcal{H}} L(h)$$

But this is often impossible, leading to the more relaxed definition

> **Definition** (PAC Learnability) A hypothesis class $\mathcal{H}$ is (agnostic) PAC learnable with respect to a set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, if there exists a learning algorithm, which returns $\hat{h} \equiv \hat{h}(S)$ for a sample S of size $|S| = m$ with the following property: for every $0 < \epsilon, \delta < 1$ there exists $m(\epsilon, \delta)$, such that for all $m \geq m(\epsilon, \delta)$,
>
> $$\mathbb{P}\left[\left\{S \in \mathcal{Z}^m \,:\, L(\hat{h}(S)) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon\right\}\right] \;\geq\; 1 - \delta$$

- PAC learning was introduced by Leslie Valiant (1984), who won for it the Turing Award in 2010.

Learning Infinite Hypothesis Classes

For most hypothesis classes $|\mathcal{H}| = \infty$: What can be done here?

Common measures of complexity of hypothesis classes

- Vapnik-Chervonenkis (VC) dimension
  Chapter 6, 28 & 20 in Shai's [ML2014] book
- Rademacher Complexity
  Chapter 26 in [ML2014] book; this was recently used in
  - A Priori Estimates For Two-layer Neural Networks, by Weinan et al., Jan 2019.
  - Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., Jan 2019.
- PAC - Bayes: Chapter 31 in [ML20014]
  used recently in several NN papers, e.g.: see Neyshabur et al.
- Compression Bounds: Chapter 30 in [ML2014]

Rademacher Complexity

- To simplify the notation, let $f(z) \equiv f(h,z) = \ell(h,z)$ and let $\mathcal{F}$ be the set of these functions
- Then, for $f \in \mathcal{F}$, the population/true risk end empirical risk are equal to

$$L(f) = \mathbb{E}_z[f(z)], \qquad \hat{L}_S(f) = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$$

Definition Let σ_i be i.i.d. with $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$. Then, the *Rademacher Complexity* of $\mathcal{F}$ with respect to sample S is defined as

$$R(\mathcal{F}, S) := \frac{1}{m}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)$$

Back to Generalization Bounds

Lemma 26.2 The expected value of the representativeness of S is bounded by

$$\mathbb{E}_S[\mathrm{Rep}(\mathcal{F}, S)] = \mathbb{E}_S\left[\sup_{f \in \mathcal{F}} |\hat{L}_S(f) - L(f)|\right] \le 2\mathbb{E}_S[R(\mathcal{F}, S)]$$

Theorem 26.6 (Shais' book) Assume that for all z and $h \in \mathcal{H}$, we have $|\ell(h, z)| \le c$, and recall $S = \{z_1, \ldots, z_m\}$ is the sample. Then, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$ (and in particular for $h = ERM_{\mathcal{H}}(S)$),

$$L(h) - \hat{L}_S(h) \le 2\mathbb{E}_S[R(\mathcal{F}, S)] + c\sqrt{\frac{2}{m}\log\left(\frac{2}{\delta}\right)}$$

Rademacher Calculus

Hence, to make the preceding theorem useful, we must find the ways to estimate

$$\mathbb{E}_S[R(\mathcal{F}, S)] = ?$$

With a small abuse of notation, from any set $A \in R^m$ let us define

$$R(A) := \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{a}\in A}\sum_{i=1}^{m} a_i\sigma_i\right]$$

This lemma is immediate for the definition. (Why?)

Lemma 26.6 For any $A \subset \mathbb{R}^m$, $c \in \mathbb{R}$, and $\boldsymbol{a}_0 \in \mathbb{R}^m$, we have

$$R(\{c\boldsymbol{a} + \boldsymbol{a}_0 : \boldsymbol{a} \in A\}) \leq |c|R(A).$$

Contraction Lemma

Lemma 26.9 (Contraction Lemma) For each $i \in \{1, \ldots, m\}$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be ρ-Lipschitz, i.e., for any $\alpha, \beta \in \mathbb{R}$, we have $\phi_i(\alpha) - \phi_i(\beta) \leq \rho|\alpha - \beta|$. Furthermore, for any $\in \mathbb{R}$, let $\boldsymbol{\phi}(\boldsymbol{a}) = (\phi_1(a_1), \ldots, \phi_m(a_m))$ and $\boldsymbol{\phi} \circ A = \{\boldsymbol{\phi}(\boldsymbol{a}) : \boldsymbol{a} \in A\}$. Then,

$$R(\boldsymbol{\phi} \circ A) \leq \rho R(A).$$

Rademacher complexity of a finite set $A = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}, \boldsymbol{a}_i \in \mathbb{R}^m$, grows logarithmically with the size of the set $|A| = n$.

Lemma 26.8 (Massarat Lemma) Let $\bar{\boldsymbol{a}} = (1/n) \sum_{i=1}^{n} \boldsymbol{a}_i$. Then,

$$R(A) \leq \max_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \bar{\boldsymbol{a}}\| \frac{\sqrt{2 \log(|A|)}}{m}.$$

Rademacher Complexity of Linear Classes

Using Massarat's lemma, we can derive the following lemma.

Let

$$\mathcal{H}_1 = \{\boldsymbol{x} \to \langle \boldsymbol{w}, \boldsymbol{x} \rangle : \|\boldsymbol{w}\|_1 \leq 1\}$$

Lemma 26.11 Let $S = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ be vectors in $\boldsymbol{x}_i \in \mathbb{R}^n$. Then,

$$R(\mathcal{H}_1 \circ S) \leq \max_i \|\boldsymbol{x}_i\|_\infty \sqrt{\frac{2\log(2n)}{m}}$$

- Combining this lemma with contraction lemma allows for use in NNs.

A Priori Estimates For Two-layer NNs

A Priori Estimates For Two-layer Neural Networks, Weinan et al., 2019.
The consider two ways of bounding the error (in some norm):

- f^*: true function that we want to estimate
- $\hat{f}_n$: numerical estimate based on a sample of size n
- A priori error estimate
$$\|\hat{f}_n - f^*\| = O(\|f^*\|)$$
- A posteriori error estimate
$$\|\hat{f}_n - f^*\| = O(\|\hat{f}_n\|)$$
- In this context, most of the recent work on generalization error of NNs can be viewed as "a posteriori".
Values of $\|\hat{f}_n\|$ are often huge, yielding often vacuous bounds.

Paper Notation

- Training set: $S = \{(x_i, y_i)\}_{i=1}^{n}$; i.i.d.samples from a distribution $\rho_{x,y}$
- True (target) function: $f^*(x) = \mathbb{E}[y|x]$ where $y = f(x) + \xi$ with ξ being the noise.
- $f^*(x) : [-1,1]^d \to [0,1]$
- Two layer neural network

$$f(x;\theta) = \sum_{k=1}^{m} a_k \sigma(w_k^\top x)$$

  where $w_k \in \mathbb{R}^d, a_k \in \mathbb{R}$ and $\theta = \{(a_k, w_k)\}_{k=1}^{m}$
- $\sigma(x) : \mathbb{R} \to \mathbb{R}$: activation function
  $\sigma(x)$ scale free: $\sigma(\alpha x) = \alpha\sigma(x), \alpha \geq 0, x \in \mathbb{R}$
  e.g., ReLU or Leaky ReLU

Training

- Loss function: $\ell(y, y') = (y - y')^2/2$
- Ultimate goal: minimize the population (true) risk

$$L(\theta) = \mathbb{E}_{x,y}[\ell(f(x;\theta), y)]$$

- In practice: minimize the empirical risk

$$\hat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

- Regularized empirical risk:

$$J_\lambda(\theta) := \hat{L}_n(\theta) + \lambda(\|\theta\|_{\mathcal{P}} + 1)$$

  where $\|\theta\|_{\mathcal{P}} := \sum_{k=1}^{m} |a_k| \|w_k\|$ is the path norm (Neyshabur et al., 2015)

- Regularized ERM: $$\hat{\theta}_{n,\lambda} = \arg\min_{\theta} J_\lambda(\theta)$$

  where $\lambda > 0$ is the tuning parameter.

Barron Space

Based on pioneering work by Barron (1993).

- Barron function: A function $f : \Omega \to \mathbb{R}$ is called a Barron if it admits the following representation

$$f(x) = \int_{S^d} a(w)\sigma(w^\top x)d\pi(w),$$

  where π is a probability distribution over $S^d = \{x : \|x\|_1 = 1\}$ and $a(\cdot)$ is a scalar function.

- Barron norm: Let f be a Barron function.
  - Denote by Θ_f all possible representations of f

$$\Theta_f = \left\{(a, \pi) : f(x) = \int_{S^d} a(w)\sigma(w^\top x)d\pi(w)\right\}$$

  - Barron norm $\gamma_p(f)$:

$$\gamma_p(f) := \inf_{(a,\pi)\in\Theta_f} \left(\int_{S^d} |a(w)|^p d\pi(w)\right)^{1/p}$$

Barron Space

- Barron space
$$\mathcal{B}_p(\Omega) = \{f(x) : \gamma_p(f) < \infty\}$$

- Since π is a probability distribution, by Hölder's inequality
$$\gamma_p(f) \leq \gamma_q(f), \qquad \text{if } q \geq p > 0.$$
and thus
$$\mathcal{B}_\infty(\Omega) \subset \cdots \subset \mathcal{B}_2(\Omega) \subset \mathcal{B}_1(\Omega)$$

Theorem 3.1 For any $f \in \mathcal{B}_2(\Omega)$, there exists a 2-layer NN $f(x; \tilde{\theta})$ of width m with $\|\tilde{\theta}\|_{\mathcal{P}} \leq 2\gamma_2(f)$, such that

$$\mathbb{E}_x(f(x) - f(x; \tilde{\theta}))^2 \leq \frac{3\gamma_2^2}{m}.$$

Intuition: Integral representation ensures a good approximation

$$f(x) \approx \frac{1}{m}\sum_{k=1}^{m} a(w_k)\sigma(w_k^\top x)$$

Main Result

Noiseless clase: $\xi = 0$; Also, assume $\ln(2d) \geq 2$ and $\hat{\gamma}_2 = \max(1, \gamma_2(f))$.

Theorem 4.1 Assume $f^* \in \mathcal{B}_2(\Omega)$ and $\lambda \geq \lambda_n := 4\sqrt{2\ln(2d)/n}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of the training set S,

$$\mathbb{E}_x[(f(x;\tilde{\theta}_{n,\lambda}) - f^*(x))^2] \lesssim \frac{\gamma_2^2(f^*)}{m} + \lambda\hat{\gamma}_2(f^*) \tag{1}$$

$$+ \frac{1}{\sqrt{n}}(\hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}) \tag{2}$$

Remark Note that is we choose $\lambda = 4\sqrt{2\ln(2d)/n}$ and $m \geq \sqrt{n}$, then

$$\mathbb{E}_x[(f(x;\tilde{\theta}_{n,\lambda}) - f^*(x))^2] = O\left(\frac{1}{\sqrt{n}}\right)$$

Is $O(1/\sqrt{n})$ the best generalizations bound?

Consider this easy example

- Let $y_1, \ldots, y_n \in \mathbb{R}$ be i.i.d. variables with distribution $\mathcal{N}(\mu, \sigma)$
- Consider a hypothesis class $\mathcal{H} = \{f(x) = \text{constant}\}$
- Empirical risk minimization with quadratic loss

$$\hat{y} = \arg\min_{y \in \mathcal{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - y)^2$$

- Then, $\hat{y}$ has normal distribution $\mathcal{N}(\mu, \sigma/\sqrt{n})$ since

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- Therefore, for any $0 < \delta < 1$, there exists c_δ, such that

$$\mathbb{P}\left[|\hat{y} - \mu| \leq \frac{c_\delta}{\sqrt{n}} \right] = 1 - \delta$$

- Hence, $O(1/\sqrt{n})$ is the best we can hope for. Most results have this, but the constant is what really determines the quality of the bound.

Comparison With Kernel Methods

- For fixed π, the following is a Reproducing Kernel Hilbert Space (RKHS)

$$\mathcal{H}_\Omega = \left\{ f = \int_{S^d} a(w)\sigma(x^\top w) d\pi(w) : \|f\|_{\mathcal{H}_\pi} < \infty \right\}$$

  where

$$\|f\|^2_{\mathcal{H}_\pi} := \mathbb{E}_\pi[|a(w)|^2]$$

  The corresponding kernel (see Rahimi&Rechet, 2008) is defined by

$$k_\pi(x, x') = \mathbb{E}_\pi[\sigma(x^\top w)\sigma(x'^\top w)].$$

- Note that Barron's space is much reacher

$$\mathcal{B}_2(\omega) = \cup_\pi \mathcal{H}_\pi(\Omega)$$

Comparison With Kernel Methods

More formally

- Consider $f^* \in \mathcal{B}_2(\omega)$ and let a^*, π^* be its best representation, i.e.

$$\gamma_2^2(f^*) = \mathbb{E}_{\pi^*}[|a^*(w)|^2]$$

- For fixed π_0, if π^* is absolutely continuous withe respect to π_0

$$\begin{aligned} f^* &= \int_{S^d} a^*(w)\sigma(x^\top w)d\pi^*(w) \\ &= \int_{S^d} a^*(w)\frac{d\pi^*}{d\pi_0}\sigma(x^\top w)d\pi_0(w) \end{aligned}$$

and

$$\|f\|^2_{\mathcal{H}_{\pi_0}} = \mathbb{E}_{\pi_0}[|a^*(w))\frac{d\pi^*}{d\pi_0}|^2] \geq \gamma_2(f^*)$$

- The best generalization bound using kernel k_{π_0} is (Caponnetto&De Vito, 2007) of the order

$$\frac{\|f\|_{\mathcal{H}_{\pi_0}}}{\sqrt{n}} \geq \frac{\gamma_2(f^*)}{\sqrt{n}}$$

where the inequality follows from Theorem 4.1

A Posterior Bound

The proof of Theorem 4.1 uses the following a posterior bound.

Theorem 5.2 (A posterior generalization bound) Assume that the loss function $\ell(\cdot, y)$ is A-Lipschitz continuous and bounded by B. Then, for any δ, with probability at least $1-\delta$, over the choice of the training set S, we have, for any 2-layer NN

$$|L(\theta) - \hat{L}_n| \leq 4A\sqrt{\frac{2\ln(2d)}{n}}(\|\theta\|_{\mathcal{P}} + 1) + B\sqrt{\frac{2\ln(2c(\|\theta\|_{\mathcal{P}} + 1)^2/\delta)}{n}},$$

where $c = \sum_{k=1}^{\infty} 1/k^2$, and $\|\theta\|_{\mathcal{P}} := \sum_{k=1}^{m} |a_k|\|w_k\|$.

Note that the generalization gap is roughly bounded by $\|\theta\|_{\mathcal{P}}/\sqrt{n}$

The **proof** is based on Theorem 26.5 in Shais' book, which we covered, and the next lemma.

Rademacher Complexity of 2-Layer NN

Lemma B.3 Let $\mathcal{F}_C = \{f_m(x;\theta) | \|\theta\|_{\mathcal{P}} \leq C\}$ be the set of 2-L NN with path norm bounded by C. Then

$$R_n(\mathcal{F}_C) \leq 2C\sqrt{\frac{2\ln(2d)}{n}}$$

We proved this lemma in the last class using Lemmas 26.9 & 26.11 from [ML] book.

Proof of Theorem 5.2

- Decompose the hypothesis class, $\mathcal{F}$, into $\mathcal{F} = \cup_{l=1}^{\infty} \mathcal{F}_l$, where

$$\mathcal{F}_l = \{f_m(x;\theta) | \|\theta\|_{\mathcal{P}} \leq l\}$$

- Set $\delta_l = \delta/(cl^2)$, where $c = \sum_{l=1}^{\infty} 1/l^2$
- Now, we can decompose the event as

$$\left\{ |L(\theta) - \hat{L}_n(\theta)| \geq 4A\sqrt{\frac{2\ln(2d)}{n}}(\|\theta\|_{\mathcal{P}} + 1) + B\sqrt{\frac{2\ln(2c(\|\theta\|_{\mathcal{P}} + 1)^2/\delta)}{n}} \right\}$$
$$\subset \bigcup_{l=1}^{\infty} \left\{ \sup_{\|\theta_{\mathcal{P}}\| \leq l} |L(\theta) - \hat{L}_n(\theta)| \geq 4Al\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{2\ln(2/\delta_l)}{n}} \right\}$$

- Finally, use union bound, apply Theorem 26.5 from the [ML] book and the preceding Lemma B.3.
- Combining all the bounds, we obtain that the desired result holds with probability at least $1 - \delta$, $\delta = \sum \delta_l$.

Proof of Main Theorem 4.1

- The main idea is to bound $\|\hat{\theta}_n\|_{\mathcal{P}}$ by a well behaved path norm $\|\tilde{\theta}\|_{\mathcal{P}}$ from Theorem 3.1 since $\|\tilde{\theta}\|_{\mathcal{P}} \leq 2\gamma_2(f^*)$.
- From Theorem 5.2 we have with probability at least $1 - \delta$

$$L(\hat{\theta}_{n,\lambda}) \leq \hat{L}(\hat{\theta}_{n,\lambda}) + \lambda_n(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1) + 3\sqrt{\frac{\ln(2c(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1)^2/\delta)}{n}}$$

$$\leq J_\lambda(\hat{\theta}_{n,\lambda}) + 3\sqrt{\frac{\ln(2c(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1)^2/\delta)}{n}}, \qquad (3)$$

  where the last inequality is due to the choice of $\lambda \geq \lambda_n = 4\sqrt{2\ln(2d)/n}$
- The first term can be bounded as

$$J_\lambda(\hat{\theta}_{n,\lambda}) \leq J_\lambda(\tilde{\theta}),$$

  which follows from the definition of $\hat{\theta}_{n,\lambda}$, where $\tilde{\theta}$ is the same as in Theorem 3.1

Proof of Main Theorem 4.1

- Then, recalling $\lambda_n = 4\sqrt{2\ln(2d)/n}$ and the claim of Theorem 5.2,

$$\begin{aligned}
J(\tilde{\theta}) &= \hat{L}_n(\tilde{\theta}) + \lambda(\|\tilde{\theta}\|_{\mathcal{P}} + 1) \\
&\leq L(\tilde{\theta}) + (\lambda_n + \lambda)(\|\tilde{\theta}\|_{\mathcal{P}} + 1) + 2\sqrt{\frac{2\ln(2c(\|\tilde{\theta}\|_{\mathcal{P}} + 1)^2/\delta}{n}} \\
&\leq L(\tilde{\theta}) + 6\lambda\hat{\gamma}_2(f^*) + 2\sqrt{\frac{2\ln(2c(1 + 2\gamma_2(f^*))^2/\delta}{n}} \\
&\leq L(\tilde{\theta}) + 8\lambda\hat{\gamma}_2(f^*) + 2\sqrt{\frac{\ln(2c/\delta)}{n}}
\end{aligned}$$

  where in second to the last inequality we used $\|\tilde{\theta}\|_{\mathcal{P}} \leq 2\gamma_2(f^*)$, and in the last $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \ln(1+a) \leq a, a \geq 0, b \geq 0$.
  (The preceding inequality is stated as Proposition 5.1.)

- Next, we bound the second term in Equation (3)

$$\sqrt{n}\sqrt{\frac{\ln(2c(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1)^2/\delta)}{n}} \leq \sqrt{\ln(2nc/\delta)} + \sqrt{2\frac{\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}}{\sqrt{n}}}$$

Proof of Main Theorem 4.1

- Next, $\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}$ can be bounded using

$$\lambda(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}+1) \leq J_\lambda(\hat{\theta}_{n,\lambda}) \leq J_\lambda(\tilde{\theta}) \leq L(\tilde{\theta})+8\lambda\hat{\gamma}_2(f^*)+2\sqrt{\frac{\ln(2c/\delta)}{n}},$$

which we showed earlier, implying (Proposition 5.2)

$$\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} \leq \frac{L(\tilde{\theta})}{\lambda} + 8\hat{\gamma}_2(f^*) + \frac{1}{2}\sqrt{\ln(2c/\delta)}.$$

- Finally, combining the preceding bounds in (3), yields

$$L(\hat{\theta}_{n,\lambda}) \lesssim L(\tilde{\theta})+8\lambda\hat{\gamma}_2(f^*)+\frac{3}{\sqrt{n}}\left(\sqrt{\frac{L(\tilde{\theta})}{n^{1/2}\lambda}} + \hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}\right),$$

which, in combination with $L(\tilde{\theta}) \leq 3\gamma_2^2(f^*)/m$ from Theorem 3.1, yields the proof.

VC Dimension

- VC dimension: common characterization of sample complexity
- Introduced by Vapnik & Chervonenkis (VC) in 1970
- Can be used to characterize the sample complexity of NNs
- In contrast to Weinan et al., this is an example of bounding the generalization by $\hat{f}$ instead of the target function f^*.
  - A priori error estimate (Weinan et al.)
  $$\|\hat{f}_n - f^*\| = O(\|f^*\|)$$
  - A posteriori error estimate (VC dimension and the book)
  $$\|\hat{f}_n - f^*\| = O(\|\hat{f}_n\|)$$

VC Dimension: Definition

- Let $\mathcal{H}$ be a class of functions from $\mathcal{X} \to \{\pm 1\}$ (or $\{0,1\}$)
- Let $C = \{x_1, \ldots, x_{|C|}\} \subset \mathcal{X}$
- Let $\mathcal{H}_C$ be the restriction of $\mathcal{H}$ to C, namely, $\mathcal{H}_C = \{h_C : h \in \mathcal{H}\}$ where $h_C : C \to \{\pm 1\}$ is s.t. $h_C(x_i) = h(x_i)$ for every $x_i \in C$
- Observe: we can represent each h_C as the vector $(h(x_1), \ldots, h(x_{|C|})) \in \{\pm 1\}^{|C|}$
- Therefore: $|\mathcal{H}_C| \leq 2^{|C|}$
- We say that $\mathcal{H}$ shatters C if $|\mathcal{H}_C| = 2^{|C|}$
- $\mathrm{VCdim}(\mathcal{H}) = \sup\{|C| \ : \ \mathcal{H} \text{ shatters } C\}$
- That is, the VC dimension is the maximal size of a set C such that $\mathcal{H}$ gives no prior knowledge w.r.t. C

VC dimension — Examples

To show that $\mathrm{VCdim}(\mathcal{H}) = d$ we need to show that:

1. There exists a set C of size d which is shattered by $\mathcal{H}$.
2. Every set C of size $d+1$ is not shattered by $\mathcal{H}$.

VC dimension — Examples

Threshold functions: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{x \mapsto 1_{\{x-\theta\geq 0\}} : \theta \in \mathbb{R}\}$

- Show that $\{0\}$ (or any other one-point set) is shattered
- Show that any two points cannot be shattered since no function from $\mathcal{H}$ can result in $\{1, 0\}$ immage

VC dimension — Examples

Intervals: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_{a,b} : a < b \in \mathbb{R}\}$, where $h_{a,b}(x) = 1$ iff $x \in [a, b]$, and zero otherwise.

- Show that $\{0, 1\}$ (or any other two point set $\{c_1, c_2\}$ is shattered
- Show that any three points cannot be shattered since no function from $\mathcal{H}$ could result in labeling $\{1, 0, 1\}$

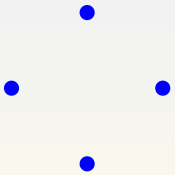
- Note that $\mathcal{H}$ is a 2-parameter class and $\mathrm{VCdim}(\mathcal{H}) = 2$

VC dimension — Examples

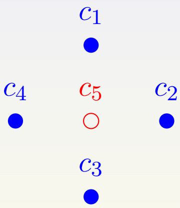
Axis aligned rectangles: $\mathcal{X} = \mathbb{R}^2$,
$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 < a_2 \text{ and } b_1 < b_2\}$, where
$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = 1$ iff $x_1 \in [a_1, a_2]$ and $x_2 \in [b_1, b_2]$

Show:

Shattered



Not Shattered



No function from $\mathcal{H}$ can map $\{c_1, \ldots, c_5\} \to \{1, 1, 1, 1, 0\}$

Note that $\mathcal{H}$ is a 4-parameter class and $\mathrm{VCdim}(\mathcal{H}) = 4$

VC dimension — Examples

Finite classes:

- Show that the VC dimension of a finite $\mathcal{H}$ is at most $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$
  - since C cannot be shattered if $|\mathcal{H}| < 2^{|C|}$

- There can be arbitrary gap between $\mathrm{VCdim}(\mathcal{H})$ and $\log_2(|\mathcal{H}|)$
  - e.g., consider $\mathcal{X} = \{1, 2, \cdots, k\}$ and consider $\mathcal{H} = \{\text{step functions on } \mathcal{X}\}$
  - Then, $|\mathcal{H}| = k$, but $\mathrm{VCdim}(\mathcal{H}) = 1$

VC dimension — Examples

Halfspaces: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$

- Show that $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ is shattered
- Show that any $d+1$ points cannot be shattered
- Hence, $\mathrm{VCdim}(\mathcal{H}) = d$

- Note again that $\mathcal{H}$ is a d-parameter class and $\mathrm{VCdim}(\mathcal{H}) = d$
- In general, one can expect that the VC dimension of a hypothesis class is equal to the number of parameters.
  - However, this is not true in general, e.g., consider $h_\theta(x) = \lceil \sin(\theta x)/2 \rceil \Rightarrow \mathrm{VCdim}(\mathcal{H}) = \infty$.

The Fundamental Theorem of Statistical Learning

Theorem (Theorem 6.8 in [ML] book)

Let $\mathcal{H}$ be a hypothesis class of binary classifiers with $\mathrm{VCdim}(\mathcal{H}) = d$. *Then, there are absolute constants C_1, C_2, s.t.*

1. *(Agnostic: $\hat{h}$ may not be in $\mathcal{H}$) $\mathcal{H}$ is PAC learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
2. *(Realizable case: $\hat{h} \in \mathcal{H}$) $\mathcal{H}$ is PAC learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Furthermore, this sample complexity is achieved by the ERM rule.

- We'll give a sketch of the proof of part 2: the realizable case.
- The agnostic case is given in Chapter 28 and is based on Rademacher complexity: more specifically Massarat Lemma 26.8, which we covered before.

Proof of the lower bound – main ideas

- Suppose $\mathrm{VCdim}(\mathcal{H}) = d$ and let $C = \{x_1, \ldots, x_d\}$ be a shattered set
- Consider the distribution $\mathcal{D}$ supported on C s.t.

$$\mathcal{D}(\{x_i\}) = \begin{cases} 1 - 4\epsilon & \text{if } i = 1 \\ 4\epsilon/(d-1) & \text{if } i > 1 \end{cases}$$

- If we see m i.i.d. examples then the expected number of examples from $C \setminus \{x_1\}$ is $4\epsilon m$
- If $m < \frac{d-1}{8\epsilon}$ then $4\epsilon m < \frac{d-1}{2}$ and therefore, we have no information on the labels of at least half the examples in $C \setminus \{x_1\}$
- Best we can do is to guess, but then our error is $\geq \frac{1}{2} \cdot 2\epsilon = \epsilon$

Proof of the upper bound – main ideas

- Recall the proof for finite class:
  - For a single hypothesis, we've shown that the probability of the event: $L_S(h) = 0$ given that $L_{(\mathcal{D},f)} > \epsilon$ is at most $e^{-\epsilon m}$
  - Then we applied the union bound over all "bad" hypotheses, to obtain the bound on ERM failure: $|\mathcal{H}|\, e^{-\epsilon m}$
- If $\mathcal{H}$ is infinite, or very large, the union bound yields a meaningless bound

Proof of the upper bound – main ideas

- The two samples trick: show that

$$\begin{aligned}&\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[\exists h\in\mathcal{H}_B : L_S(h)=0]\\ &\quad\le 2\underset{S,T\sim\mathcal{D}^m}{\mathbb{P}}[\exists h\in\mathcal{H}_B : L_S(h)=0 \text{ and } L_T(h)\ge\epsilon/2]\end{aligned}$$

- Symmetrization: Since S,T are i.i.d., we can think on first sampling $2m$ examples and then splitting them to S,T at random
- If we fix h, and $S\cup T$, the probability to have $L_S(h)=0$ while $L_T(h)\ge\epsilon/2$ is $\le e^{-\epsilon m/4}$
- Once we fixed $S\cup T$, we can take a union bound over $\mathcal{H}_{S\cup T}$

For more details, see Section 28.3 in [ML] book.

Sauer-Shelah-Perles Lemma

Let

$$\tau_{\mathcal{H}}(m) := \max_{C \in \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

In words, $\tau_{\mathcal{H}}(m)$ is the number of functions from $C \to \{0, 1\}$ that can be realized by restricting $\mathcal{H}$ to $\mathcal{H}_C$.

Lemma (Sauer-Shelah-Perles)

Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all $C \subset \mathcal{X}$ s.t. $|C| = m > d + 1$ we have

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$$

- The lemma shows that when $m > d + 1$, $\tau_{\mathcal{H}}(m)$ the grows polynomially, rather than exponentially in m.
- The proof is by induction in m, see p. 49 in the [ML] book.

VC Dimension of Neural Networks

Recall the graph notation for NNs:

- A neural network is obtained by connecting many neurons together
- We focus on feedforward networks, formally defined by a directed acyclic graph $G = (V, E)$
- Input nodes: nodes with no incoming edges
- Output nodes: nodes without out going edges
- Weights: $w : E \to \mathbb{R}$
- Each neuron (node) receives as input:

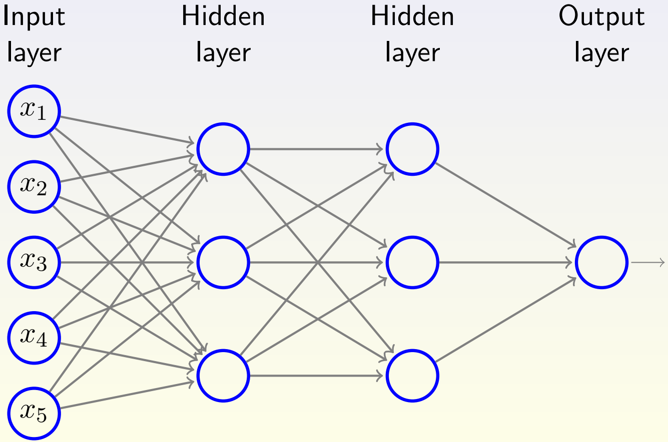
$$a[v] = \sum_{u \to v \in E} w[u \to v] o[u]$$

and output

$$o[v] = \sigma(a[v])$$

Multilayer Neural Networks

- Neurons are organized in layers: $V = \cup_{t=0}^{T} V_t$, and edges are only between adjacent layers
- Example of a multilayer neural network of depth 3 and size 6



Neural Network Hypothesis Class

- Given a neural network (V, E, σ, w), we obtain a hypothesis $h_{V,E,\sigma,w} : \mathbb{R}^{|V_0|-1} \to \mathbb{R}^{|V_T|}$
- We refer to (V, E, σ) as the architecture, and it defines a hypothesis class by

$$\mathcal{H}_{V,E,\sigma} \;=\; \{h_{V,E,\sigma,w} \;:\; w \text{ is a mapping from } E \text{ to } \mathbb{R}\} \;.$$

- The architecture is our “Prior knowledge” and the learning task is to find the weight function w

Neural Network Sample Complexity

- **Theorem 1:** (σ - step/sign function) The VC dimension of $\mathcal{H}_{V,E,\mathrm{sign}}$ is $O(|E|\log(|E|))$.
- **Theorem 2:** (σ - any sigmoidal function) The VC dimension of $\mathcal{H}_{V,E,\sigma}$, for σ being the sigmoidal function, is $\Omega(|E|^2)$.

- Representation trick: In practice, we only care about networks where each weight is represented using $O(1)$ bits, and therefore the VC dimension of such networks is $O(|E|)$, no matter what σ is.

Neural Network Sample Complexity

Proof of Theorem 1:

- Let $\tau_{\mathcal{H}}(m) = \max_{C \in \mathcal{X}:|C|=m} |\mathcal{H}_C|$, where $\mathcal{H}_C$ is the restriction to C of binary valued functions in $\mathcal{H}$
- NN has T layers: $0, 1, 2, \ldots, T$ with V_t nodes at layer t.
- Then, $\mathcal{H}$ can be written as a composition

$$\mathcal{H} = \mathcal{H}^{(T)} \circ \cdots \circ \mathcal{H}^{(1)}$$

- Furthermore, each class $\mathcal{H}^{(t)}$ can be decomposed per each neuron

$$\mathcal{H}^{(t)} = \mathcal{H}^{(t,1)} \times \cdots \times \mathcal{H}^{t,|V_t|}$$

Neural Network Sample Complexity

Proof of Theorem 1:

- Then
$$\tau_{\mathcal{H}^{(t)}}(m) \leq \prod_{i=1}^{|V_t|} \tau_{\mathcal{H}^{(t,i)}}(m)$$
- Let $d_{t,i}$ be the number of edges that are headed to the ith neuron of layer t.
- Since each neuron is a homogenous half-space hypothesis class and the VC dimension of the the half-spaces is the dimension of their input, by Sauer's lemma,
$$\tau_{\mathcal{H}^{(t,i)}}(m) \leq \left(\frac{em}{d_{t,i}}\right)^{d_{t,i}} \leq (em)^{d_{t,i}}$$
implying
$$\tau_{\mathcal{H}}(m) \leq (em)^{\sum_{t,i} d_{t,i}} = (em)^{|E|}$$

Neural Network Sample Complexity

Proof of Theorem 1:

- Now, if we assume that m points are shattered, we must have

  $$2^m \leq (em)^{|E|}$$

  implying

  $$m \leq |E| \log(em)/\log(2)$$

  resulting in

  $$m \leq O(|E| \log(|E|)),$$

  which concludes the proof.

Generalization Bound for Unregularized NNs

Theorem Let $\mathcal{H} = (V, E, \sigma)$ be a hypothesis class of binary classifiers of multilayer NN with step function activation σ. Then, there are absolute constants C_1, C_2, s.t. $\mathcal{H}$ is (agnostic) PAC learnable with sample complexity

$$C_1 \frac{|E| \log(|E|) + \log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}(\epsilon, \delta) \le C_2 \frac{|E| \log(|E|) + \log(1/\delta)}{\epsilon^2}$$

Furthermore, this sample complexity is achieved by the ERM rule.

- If σ is any sigmoid, $|E| \log(|E|)$ should be replaced by $|E|^2$
- Hence, we need either regularization/shrinkage or prior knowledge on the target function to reduce the sample complexity: e.g., Weinan et. al (2019), Neyshabur et. al. (2015-)

Recent Results on VC-dim of NNs With ReLUs

- Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks, Barttlet et al., 2019.
- Compute tight upper and lower bounds on the VC-dimension of deep neural networks with the ReLU activation function.
- Let W be the number of weights and L be the number of layers. Then, the paper
  - proves that the VC-dimension is $O(WL\log(W/L))$
  - provides examples with VC-dimension $\Omega(WL\log(W/L))$
- Roughly VCdim$(\mathcal{H}) \approx WL$

Implicit Bias of Gradient Descent

- Could it be that we are finding nice generalizable solutions due to GD optimization?
  - In lazy trining, GD does not move the parameters much, which restricts the size of the hypothesis class, i.e., GD acts as an implicit regularizer.
- For linear predictors with linearly separable data, Soudry, Hoffer, and Srebro (2017) show that GD on the cross-entropy loss is implicitly biased towards a maximum margin direction.
  - Bias of GD towards margin maximization means that gradient descent "prefers" a solution which is likely to generalize well, and not just achieve low empirical risk.
- The preceding work inspired many other results, e.g.: Ji and Telgarsky 2019+; Gunasekar et al. 2018; Lyu and Li 2019; Chizat and Bach 2020; Ji et al. 2020.
- Interesting topic for further research, or project. Maybe I'll say a bit more on this next week.

Reading

- VC-dimension: Chapters 6 and 28; VCdim of NNs: Theorem 20.6 in Chapter 20 in [ML] book.
- Recent paper on VC dimension
  - Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks, Barttlet et al., 2019.
- Generalization bounds for NNs
  - A Priori Estimates For Two-layer Neural Networks, by Weinan et al., Jan 2019.
  - Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., Jan 2019.
  - Norm-based capacity control in neural networks, by Neyshabur et. al 2015.
  - Exploring generalization in deep learning, by Neyshabur et. al 2017.
  - A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks, by Neyshabur et. al 2018.
  - Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks, by Neyshabur et. al 2018.
  - See the references in the preceding papers and follow their citations

Have Fun!