Lecture Notes: Learning Theory & Neural Network Approximation

I. Basic Decomposition of Risk (Week 6 & Week 8)

1. Problem Setup

- **Data:** Pairs (x_i, y_i) sampled i.i.d from a distribution $\mathcal{D}$. $x \in \mathcal{X}$ (high dimensional), $y \in \mathcal{Y}$ (label).
- **Map:** $f : \mathcal{X} \to \mathcal{Y}$.
- **Loss Function:** $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- **Definitions of Risk:**
  - **Population Risk:** $R(f) = \mathbb{E}_{(x,y)}[l(f(x), y)]$.
  - **Empirical Risk:** $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i)$.

2. Hypothesis Space and Complexity

- **Hypothesis Space ($\mathcal{F}$):** A normed space $\{f : \mathcal{X} \to \mathcal{Y}\}$.
- **Constrained Space ($\mathcal{F}_\delta$):** $\{f \in \mathcal{F} : r(f) \leq \delta\}$, where $r(f)$ is a complexity measure (e.g., Euclidean norm of weights, number of parameters, GD iterations).
- **Empirical Risk Minimization (ERM):**

$$\hat{f} = \text{argmin}_{f \in \mathcal{F}_\delta} \hat{R}(f)$$

3. Error Decomposition

The excess risk $R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)$ can be decomposed into three components:

$$\text{Error} = \underbrace{\text{Optimization}}_{\text{Opt}} + \underbrace{\text{Generalization}}_{\text{Gen}} + \underbrace{\text{Approximation}}_{\text{App}}$$

Derivation:

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = \underbrace{[R(\hat{f}) - \hat{R}(\hat{f})] + [\hat{R}(\hat{f}) - \hat{R}(f^*)] + [\hat{R}(f^*) - R(f^*)]}_{\text{Estimation / Generalization}} + \underbrace{[R(f^*) - \inf_{f \in \mathcal{F}} R(f)]}_{\text{Approximation}}$$

- **Approximation Error:** $\inf_{f \in \mathcal{F}_\delta} R(f) - \inf_{f \in \mathcal{F}} R(f)$. This measures the limitation of the constrained hypothesis class $\mathcal{F}_\delta$ compared to the full space.
- **Generalization Error:** bounded by $2 \cdot \sup_{f \in \mathcal{F}_\delta} |R(f) - \hat{R}(f)|$.
- **Optimization Error:** Arises if the algorithm fails to find the global minimum of the empirical risk (often assumed zero in theoretical bounds).

II. Generalization Error and Curse of Dimensionality (Week 6)

1. Generalization vs. Approximation Trade-off

- As the hypothesis class complexity increases (e.g., d increases):
  - **Approximation Error** decreases.
  - **Generalization (Estimation) Error** increases (Variance increases).
- **Variance Relation:** $\mathrm{Var}[\hat{R}(f) - R(f)] \propto \frac{1}{n}\mathrm{Var}[l]$.

2. Curse of Dimensionality

- Let the target function be $f^*(x)$. If we assume f^* is Lipschitz continuous (L-Lipschitz), meaning $|f^*(x) - f^*(z)| \leq L\|x - z\|$.
- **MSE Bound:** $\mathbb{E}|\hat{f}(x) - f^*(x)|^2 \leq 4L^2\|x - x_i\|^2$.
- To maintain a stable generalization error ϵ, the number of samples n required scales exponentially with dimension d:

$$n \sim \left(\frac{1}{\epsilon}\right)^d$$

III. Universal Approximation Theorem (UAT) (Week 6)

1. The Theorem

Neural networks with a single hidden layer are dense in $C(I_n)$ (continuous functions on a compact set).

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j)$$

- σ: Sigmoid function ($\mathbb{R} \to [0, 1]$), continuous, $\sigma(+\infty) = 1, \sigma(-\infty) = 0$.

2. Functional Analysis Tools for Proof

- **Discriminatory Function:** A function σ is discriminatory if for a measure $\mu \in M(I_n)$, $\int \sigma(w^T x + b) d\mu(x) = 0 \implies \mu = 0$.
- **Hahn-Banach / Riesz Representation Theorem (RRT):**
  - If a subspace U is **not dense** in X, there exists a bounded linear functional L such that $L|_U = 0$ but $L \neq 0$.
  - By RRT, on $C(K)$, every bounded linear functional corresponds to a unique signed measure μ. Thus, $L(f) = \int f d\mu$.

3. Proof Sketch (By Contradiction)

1. Let U be the subspace of neural networks.
2. Assume U is **not** dense. By RRT, there exists a non-zero measure μ such that $\int g(x) d\mu(x) = 0$ for all $g \in U$.
3. Since $\sigma(w^T x + b) \in U$, we have $\int \sigma(w^T x + b) d\mu(x) = 0$ for all w, b.
4. Since σ is discriminatory (Lemma 1), this implies $\mu = 0$.
5. Contradiction. Therefore, U is dense.

IV. Fourier Analysis & Barron Spaces (Week 7)

1. Barron Space

To avoid the curse of dimensionality, we define a space constrained by spectral properties.

- **Barron Norm:**

$$\|f\|_{\mathcal{B}^s(\Omega)} := \inf_{f_e|_\Omega = f} \int (1+|\xi|)^s |\hat{f}_e(\xi)| d\xi$$

  This penalizes high-frequency components in the Fourier transform $\hat{f}$.

2. Approximation Rate using Maurey's Theorem

We aim to approximate functions in the Barron space using a convex combination of neurons (2-layer NN).

Maurey's Theorem (Probabilistic Method):
Let H be a Hilbert space and $G \subset H$ with $\|g\| \leq B$ for $g \in G$. Let $f \in \overline{\text{conv}(G)}$. Then there exists $f_N = \frac{1}{N}\sum_{i=1}^N g_i$ such that:

$$\|f - f_N\|^2 \leq \frac{B^2}{N}$$

Proof Steps:

1. Since $f \in \overline{\text{conv}(G)}$, $f \approx \sum_{j=1}^m \alpha_j h_j$ where $\sum \alpha_j = 1, \alpha_j > 0$.
2. Define a random variable X taking value h_j with probability α_j.
3. $\mathbb{E}[X] = \sum \alpha_j h_j = f$.
4. Consider the estimator $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$.
5. Compute Expected Error (Variance):

$$\mathbb{E}\left[\|f - \frac{1}{N}\sum X_i\|^2\right] = \frac{1}{N}(\mathbb{E}[\|X\|^2] - \|f\|^2) \leq \frac{B^2}{N}$$

   (Since $\mathbb{E}[\|X\|^2] \leq B^2$).

Conclusion: The approximation error for this class is $O(1/\sqrt{N})$, independent of dimension d.

V. Rademacher Complexity & Generalization Bounds (Week 7 & 8)

1. Definition

Rademacher Complexity measures the ability of a function class $\mathcal{F}$ to fit random noise $\xi_i \in \{+1, -1\}$.

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i) \right]$$

2. Generalization Bound

With probability $1 - \delta$:

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \leq 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}$$

3. Complexity of 2-Layer ReLU Networks

We analyze the class $\mathcal{F}_{m,\sigma,B} = \{f_\theta(x) = \sum_{j=1}^{m} \beta_j \sigma(w_j^T x) : C(\theta) \leq B\}$, where $C(\theta) = \sum |\beta_j| \|w_j\|_2$.
Assume data is bounded $\|x\|_2 \leq C$.

Derivation:

1. **Scaling Property:** For ReLU, $\alpha\sigma(x) = \sigma(\alpha x)$ for $\alpha > 0$. We can normalize weights such that $\|w_j\|_2 = 1$ and absorb magnitude into β_j.
2. **Rademacher Definition:**

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \frac{1}{n} \mathbb{E}_\xi \left[\sup_{\|\beta\|_1 \leq B, \|w_j\| \leq 1} \sum_{i=1}^{n} \xi_i \sum_{j=1}^{m} \beta_j \sigma(w_j^T x_i) \right]$$

3. **Supremum Bound:** Move sum over j out. The sup over the convex combination is achieved at a vertex (single neuron).

$$\leq \frac{B}{n} \mathbb{E}_\xi \left[\sup_{\|w\| \leq 1} \left| \sum_{i=1}^{n} \xi_i \sigma(w^T x_i) \right| \right]$$

4. **Talagrand's Contraction Lemma:** Since ReLU is 1-Lipschitz, $\mathfrak{R}(\phi \circ \mathcal{F}) \leq 2\mathfrak{R}(\mathcal{F})$. (Note: Note uses factor 2 for absolute value removal/symmetry).

$$\leq \frac{2B}{n}\mathbb{E}_\xi \left[\sup_{\|w\|\leq 1} \sum_{i=1}^{n} \xi_i w^T x_i \right]$$

5. **Linear Class Bound:**

$$\sup_{\|w\|\leq 1} w^T \left(\sum \xi_i x_i \right) = \left\| \sum \xi_i x_i \right\|_2$$

Using Jensen's inequality and independence of ξ:

$$\mathbb{E} \left\| \sum \xi_i x_i \right\| \leq \sqrt{\mathbb{E} \left\| \sum \xi_i x_i \right\|^2} = \sqrt{\sum \mathbb{E}[\xi_i^2] \|x_i\|^2} = \sqrt{nC^2}$$

6. **Final Result:**

$$\mathfrak{R}_n(\mathcal{F}_{m,\sigma,B}) \leq \frac{2BC}{\sqrt{n}}$$

4. Summary of Total Error

$$\text{Total Error} \leq \underbrace{\frac{2BC}{\sqrt{n}}}_{\text{Estimation}} + \underbrace{\frac{C'}{\sqrt{m}}}_{\text{Approximation (Maurey)}} + \text{Optimization Error}$$

This confirms that neural networks can overcome the curse of dimensionality if the function belongs to the appropriate Barron space (error depends on n, not d).