Lecture Notes: Theoretical Foundations of Deep Learning

Part 1: Decomposition of Risk (Week 6)

1.1 Problem Setup

- **Data:** $S = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ i.i.d., where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- **Goal:** Find a hypothesis $f : \mathcal{X} \to \mathcal{Y}$ within a hypothesis class $\mathcal{F}$ to minimize loss.
- **Risk Definitions:**
    - **Population Risk:** $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$.
    - **Empirical Risk:** $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$.
- **Hypothesis Space Constraints:** We often consider a constrained class $\mathcal{F}_\delta = \{f \in \mathcal{F} : c(f) \leq \delta\}$ (where $c(f)$ is a complexity measure, e.g., norm).
- **Estimator:** $\hat{f} = \arg\min_{f\in\mathcal{F}_\delta} \hat{R}(f)$.
- **Optimal:** $f^* = \arg\min_{f\in\mathcal{F}} R(f)$ (Global optimum), $f^*_{\mathcal{F}_\delta} = \arg\min_{f\in\mathcal{F}_\delta} R(f)$ (Restricted optimum).

1.2 The Decomposition

The excess risk can be decomposed into three components:

$$R(\hat{f}) - R(f^*) = \underbrace{[R(\hat{f}) - R(f^*_{\mathcal{F}_\delta})]}_{\text{Estimation Error}} + \underbrace{[R(f^*_{\mathcal{F}_\delta}) - R(f^*)]}_{\text{Approximation Error}}$$

More strictly, taking the supremum over the class for the estimation part:

$$R(\hat{f}) - \inf_{f\in\mathcal{F}_\delta} R(f) \leq 2\sup_{f\in\mathcal{F}_\delta} |R(f) - \hat{R}(f)| + \text{Optimization Error}$$

1. **Approximation Error:** Distance between the full target space and our restricted hypothesis class $\mathcal{F}_\delta$.
2. **Estimation Error (Generalization Gap):** The difference between empirical performance and population performance, bounded by the uniform convergence of the class: $\sup |R(f) - \hat{R}(f)|$.
3. **Optimization Error:** Failure to find the global minimum of $\hat{R}(f)$.

Part 2: The Curse of Dimensionality (Week 6)

Standard non-parametric estimators suffer from the curse of dimensionality.

- **Scenario:** Target function f^* is L-Lipschitz.
- **Error Rate:** To achieve an error of ϵ, the number of samples n required scales exponentially with dimension d.

$$\mathbb{E}[|\hat{f}(x) - f^*(x)|^2] \approx O(n^{-1/d})$$

- **Implication:** For high-dimensional data (large d), standard local averaging methods fail. Deep learning attempts to overcome this by exploiting compositional structures (like Barron spaces) rather than just local smoothness.

Part 3: Universal Approximation Theorem (UAT) (Week 6)

3.1 Statement

A 2-layer neural network with a "sigmoidal" activation function is dense in the space of continuous functions $C(I_n)$ on a compact set I_n.

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

3.2 Key Definitions

1. **Sigmoidal Function:** $\sigma : \mathbb{R} \to [0, 1]$ such that $\lim_{z\to\infty} \sigma(z) = 1$ and $\lim_{z\to-\infty} \sigma(z) = 0$.
2. **Discriminatory Function:** A function σ is discriminatory if for a signed measure μ,

$$\int_{I_n} \sigma(w^T x + b) d\mu(x) = 0 \quad \forall w, b \implies \mu \equiv 0.$$

3.3 Proof Sketch (Hahn-Banach & Riesz Representation)

This proof relies on **Functional Analysis** (proof by contradiction).

1. Let S be the subspace of neural networks. Assume S is **not** dense in $C(I_n)$.

2. **Hahn-Banach Theorem:** There exists a bounded linear functional L on $C(I_n)$ such that $L(g) = 0$ for all $g \in S$, but $L \neq 0$.
3. **Riesz Representation Theorem (RRT):** Any bounded linear functional L on $C(I_n)$ can be represented uniquely by a signed regular Borel measure μ:

$$L(f) = \int_{I_n} f(x) d\mu(x)$$

4. Since $\sigma(w^T x + b) \in S$, we have:

$$\int_{I_n} \sigma(w^T x + b) d\mu(x) = 0 \quad \forall w, b$$

5. **Discriminatory Property:** It is proven (Lemma) that continuous sigmoidal functions are discriminatory. Therefore, the condition above implies $\mu = 0$.
6. **Contradiction:** If $\mu = 0$, then $L = 0$, which contradicts step 2. Thus, S must be dense.

Part 4: Approximation Error & Maurey's Theorem (Week 7)

While UAT guarantees existence (density), it does not quantify the rate (efficiency) or the number of neurons needed. We use **Barron Spaces** and **Maurey's Theorem** for this.

4.1 Maurey's Theorem (Jones-Barron)

Theorem: Let H be a Hilbert space. Let $G \subset H$ be a subset such that $||g|| \leq B$ for all $g \in G$. Let f be in the closure of the convex hull of G ($f \in \overline{\text{conv}(G)}$).
Then, for any $N \geq 1$, there exists a function f_N which is a convex combination of N elements from G such that:

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

4.2 Proof Idea: Probabilistic Method

This proof is crucial for understanding why N neurons approximate well.

1. Since $f \in \text{conv}(G)$, we can write $f = \sum \alpha_j h_j$ where $\sum \alpha_j = 1, \alpha_j \geq 0$.
2. Define a random variable Z taking values in $\{h_j\}$ with probability $P(Z = h_j) = \alpha_j$.

- $\mathbb{E}[Z] = \sum \alpha_j h_j = f$.
- $\|Z\| \leq B$ almost surely.

3. Let $Z_1, \dots, Z_N$ be i.i.d. copies of Z. Define the approximation $f_N = \frac{1}{N} \sum_{i=1}^{N} Z_i$.
4. Analyze the expected squared error:

$$\mathbb{E}[\|f - f_N\|^2] = \mathbb{E}\left[\left\|\mathbb{E}[Z] - \frac{1}{N}\sum_{i=1}^{N} Z_i\right\|^2\right] = \text{Var}(f_N)$$

5. By independence:

$$\text{Var}\left(\frac{1}{N}\sum Z_i\right) = \frac{1}{N^2}\sum \text{Var}(Z_i) = \frac{1}{N}\text{Var}(Z)$$

6. Since $\text{Var}(Z) = \mathbb{E}[\|Z\|^2] - \|f\|^2 \leq B^2$, we get:

$$\mathbb{E}[\|f - f_N\|^2] \leq \frac{B^2}{N}$$

7. Since the expectation is bounded by B^2/N, there must exist at least one specific realization f_N satisfying the bound.

4.3 Implication for Neural Networks

If the target function f^* lies in a "Barron Space" (defined by spectral properties of its Fourier transform), it can be approximated by a 2-layer network with N neurons with error $O(1/N)$ (squared error) or $O(1/\sqrt{N})$ (RMSE).

- Notably, this rate is **independent of input dimension d**, avoiding the curse of dimensionality for this specific function class.

Part 5: Estimation Error & Rademacher Complexity (Week 7 & 8)

To bound the generalization error (Estimation Error), we measure the capacity of the hypothesis class using Rademacher Complexity.

5.1 Definition

Let $\mathcal{F}$ be a hypothesis class and $S = \{z_1, \dots, z_n\}$ be a fixed sample.
The **Empirical Rademacher Complexity** is:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right]$$

where σ_i are i.i.d. Rademacher variables ($P(\sigma_i = 1) = P(\sigma_i = -1) = 0.5$).

- **Intuition:** It measures how well the class $\mathcal{F}$ can correlate with random noise. A rich class can fit noise perfectly (high complexity).

5.2 Generalization Bound

With probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \leq 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

5.3 Complexity of 2-Layer ReLU Networks (Week 8)

We want to bound $\mathfrak{R}_n(\mathcal{F})$ for the class of 2-layer ReLU networks with bounded path-norm.

Class Definition:

$$\mathcal{F}_{m,\sigma,B} = \left\{ f(x) = \sum_{j=1}^{m} \beta_j \sigma(w_j^T x) \;\middle|\; \sum_{j=1}^{m} |\beta_j| \|w_j\|_2 \leq B \right\}$$

Assumption: Data is bounded $\|x\|_2 \leq C$.

Derivation Steps (Key Proof):

1. **Homogeneity:** Since ReLU is positive homogeneous ($\alpha\sigma(x) = \sigma(\alpha x)$ for $\alpha > 0$), we can re-parameterize weights such that $\|w_j\|_2 = 1$ and absorb the magnitude into β_j. The constraint becomes $\sum |\tilde{\beta}_j| \leq B$.

$$f(x) = \sum_{j=1}^{m} \tilde{\beta}_j \sigma(\tilde{w}_j^T x), \quad \|\tilde{w}_j\|_2 = 1$$

2. **Supremum Bound:**

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

   Since the inner sum is linear in β, the supremum occurs at an extreme point (one active neuron).

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

3. **Talagrand's Contraction Lemma:** Since $\sigma(\cdot)$ (ReLU) is 1-Lipschitz and $\sigma(0) = 0$, we can remove it from the Rademacher average (costing at most a factor of 2, or 1 depending on the version).

$$\mathbb{E}_\sigma \left[\sup_{\|w\| \le 1} \sum \sigma_i \sigma(w^T x_i) \right] \le \mathbb{E}_\sigma \left[\sup_{\|w\| \le 1} \sum \sigma_i (w^T x_i) \right]$$

4. **Cauchy-Schwarz:**

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

5. **Final Calculation:**

$$\mathbb{E}_\sigma \left[\left\| \sum \sigma_i x_i \right\|_2 \right] \le \sqrt{\mathbb{E} \left\| \sum \sigma_i x_i \right\|^2} = \sqrt{\sum \|x_i\|^2} = \sqrt{nC^2}$$

   (Using Jensen's inequality and independence of σ_i).

Result:

$$\mathrm{Rad}(\mathcal{F}_{m,\sigma,B}) \le \frac{2BC}{\sqrt{n}}$$

5.4 Conclusion on Total Error

Combining Maurey's Theorem (Approximation) and Rademacher Complexity (Estimation):

$$\text{Total Error} \le O\left(\frac{1}{\sqrt{m}}\right) + O\left(\frac{BC}{\sqrt{n}}\right)$$

- First term: Approximation error (decreases with network width m).
- Second term: Estimation error (decreases with sample size n).
- **Significance:** The bound depends on the *norms* (B, C) and sample size, not the dimension d. This suggests that with proper regularization (controlling B), deep learning can generalize well even in high dimensions.