CSE 543: Deep Learning	Spring 2022
Lecture 12	
Prof. Simon Du	*Scribe: Yuhao Wan*

1 Separation between NN and kernel

Definition (Kernel method). *A linear method with an embedding $\phi : \mathbb{R}^d \mapsto \mathcal{H}$ (Hilbert space), which turns an element $f \in \mathcal{H}$ into a prediction function $y = \langle f, \phi(x) \rangle$. The method uses n samples $\{x_i\}_i^n$ where $x_i \in \mathbb{R}^d$, observes $\{y_i\}_i^n$, and requires $f \in span(\phi(x_i)_{i=1}^n), i \in [n]$.*

Theorem (Allen-Zhu and Li'20). *There exists a class of functions $\mathcal{C} \subseteq \{c : \mathbb{R}^d \mapsto \mathbb{R}\}$ and a distribution μ over $\mathbb{R}^d$ such that:*

1) For all kernel method satisfying the definition above, there exists a $c \in \mathcal{C}$ such that given $y_i = c(x_i)$, if $\mathbb{E}_{x \sim \mu}[(c(x) - \langle f, \phi(x) \rangle)^2] \leq \frac{1}{9}$, then $n \geq 2^{d-1}$.

2) There exists a simple procedure such that it can output the true c as long as $n \geq d$. This procedure can be simulated/approximated by a neural network with gradient descent.

Theorem idea: the separation between NN and kernel is that there exists a function class such that kernel method requires exponential number of samples whereas neural network requires only linear number of samples.

Proof. Define distribution μ uniform on $\{0, 1\}^d$. We consider

$$\mathcal{C} = \{c_S(x) = \prod_{s \in S} x_s\}, s \subset \{1, \cdots, d\}$$

We first prove part 2) of the theorem. Choose a basis $(e_1, \cdots, e_d)$ for $\mathcal{C}$. We observe $y_i = c(e_i)$. Note that if $i \in S$, then $y_i = -1$ and if $i \notin S$, then $y_i = 1$. We know that whether i is in S or not, so we can identify the set S. Thus we can learn the function c_S by querying only d samples.

To prove part 1) of the theorem, note that $\mathcal{C}$ is a basis for a general function class $\{f : \{-1, 1\}^d \mapsto \mathbb{R}\}$ with distribution μ where $\mathbb{E}_{x \sim \mu}[c_S(x) \cdot c_{S'}(x)] = \begin{cases} 0 & \text{if } S \neq S' \\ 1 & \text{if } S = S' \end{cases}$

Our goal is to compute a small test error

$$\mathbb{E}_{x \sim \mu}[(c_{S^*}(x) - \langle f, \phi(x) \rangle)^2]$$

By definition, $f \in$ span $(\phi(x_i)_{i=1}^n)$, so we can write $f = \sum_{i=1}^n a_i \phi(x_i)$.

Consider $x \mapsto \langle \phi(x_i), \phi(x) \rangle$. We can also write $x = \sum_{S \in [d]} \lambda_{i,S} c_S(x)$

Thus, we can write the test error in quadratic form:

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mu}[(c_{S^*}(x)-\langle f,\phi(x)\rangle)^2] &= \mathop{\mathbb{E}}_{x\sim\mu}[(c_{S^*}(x)-\sum_{S\in[d]}\sum_{i=1}^{n}a_i\lambda_{i,S}c_S(x))^2] \\ &= (1-\sum_{i}^{n}a_i\lambda_{i,S^*})^2+\sum_{S\neq S^*}(\sum_i a_i\lambda_{i,S})^2\end{aligned}$$

By assumption, if this error is less or equal to $\frac{1}{9}$, then

$$(1-\sum_{i}^{n}a_i\lambda_{i,S^*})^2\leq\frac{1}{9} \text{ and } \sum_{S\neq S^*}(\sum_i a_i\lambda_{i,S})^2\leq\frac{1}{9}$$

We will show that these two properties imply that $n\geq 2^{d-1}$ by some linear algebra.

We use the following notations (assuming $n\leq 2^d$).

$\Lambda : 2^d\times n$ matrix

$\Lambda_{S,i}=\lambda_{i,S}$

$A : n\times 2^d$ matrix

$A_{i,S^*}=a_{i,S^*}$

$\Omega=\Lambda A : 2^d\times 2^d$ matrix of rank n

We rewrite the two properties in terms of the new notations.

Property 1 is equivalent to

$$(1-\Omega_{S^*,S^*})^2\leq\frac{1}{9}$$

This implies that $\Omega_{S^*,S^*}\geq\frac{2}{3}$, and thus $\sum_{S\neq S^*}\Omega^2_{S^*,S^*}\leq\frac{1}{9}$.

In other words, the diagonal entries of Ω are at least $\frac{2}{3}$, and the sum of the off-diagonal entries (row-wise) squared is no more than $\frac{1}{9}$. The idea is to use the property that a diagonal dominant matrix has near full rank.

Formally, we consider $\Omega=\text{diag}(\Omega)+\Omega'$, where Ω' is the off-diagonal matrix.

We know that the Frobenius norm $||\Omega'||_F^2\leq\frac{2^d}{9}$ by definition, and it is equivalent to the sum of the eigenvalues of Ω'. This implies that Ω' has at most $\frac{2^d}{4}$ eigenvalues that are at least $\frac{2}{3}$.

We consider the subspace with eigenvalue strictly smaller than $\frac{2}{3}$, which has dimension at least $\frac{3}{4}\cdot 2^d$. For any x in this subspace, note that

$$||\Omega x||_2=||\text{diag}(\Omega)x+\Omega' x||_2\geq||\text{diag}(\Omega)x||_2-||\Omega' x||_2>\frac{2}{3}x-\frac{2}{3}x=0$$

This shows that $\mathrm{rank}(\Omega) \geq \frac{3}{4} \cdot 2^d$, since we have a subspace of dimension at least $\frac{3}{4} \cdot 2^d$ such that for every entry x in this subspace, the product with our matrix is strictly positive. Then this matrix has rank at least of the subspace dimension.

Thus, we have $n \geq \frac{3}{4} \cdot 2^d$. $\square$