Chapter 23
The universal approximation theorem for sigmoidal functions

As we have discussed in Chapters 18 and 19, we are limited to hypothesis spaces which are compact subset of $\mathbb{C}(X)$. Specifically we have considered hypothesis spaces defined as bounded subsets of (Mercer) kernel-based RKHS. Provided that the eigenfunctions φ_i of the corresponding integral operator are available, we can express the empirical target function $\hat{h}_N$ as a linear combination of φ_i:

$$\hat{h}_N = \sum_{i=1}^{\mathscr{N}} \alpha_i \varphi_i,$$

and the empirical risk minimization problem (23.1) is equivalent to

$$\boxed{\min_{\alpha:=[\alpha_1,\ldots,\alpha_{\mathscr{N}}]} \mathscr{R}_N(\alpha) = \frac{1}{N}\sum_{i}^{N}\left(\sum_{i=1}^{\mathscr{N}} \alpha_i \varphi_i(\mathbf{x}_i) - y_i\right)^2.} \tag{23.1}$$

In general, constructing (evaluating) φ_i is not a trivial task and we have to resort to universal bases that are available at no cost. One of the most popular universal bases in machine learning consists of sigmoidal functions. The purpose of this chapter is to show that, similar to the eigenfunctions φ, the sigmoidal functions is dense in $\mathbb{C}(X)$. We follow closely the original proof of Cybenko [14].

23.1 The universal approximation theorem

Definition 23.1 (Sigmoidal functions). Any function $\sigma : \mathbb{R} \ni t \mapsto \sigma(t) \in \mathbb{R}$ with the property

$$\sigma(t) \to \begin{cases} 1 & \text{when } t \to +\infty \\ 0 & \text{when } t \to -\infty \end{cases}$$

is called a *sigmoidal* function or simply *sigmoid*.

We denote by $\mathscr{M}(X)$ the space of *finite* signed regular Borel measures on X.

Definition 23.2 (Discriminatory). σ is called *discriminatory* if for any $\mu \in \mathscr{M}(X)$

$$\int_X \sigma\left(\mathbf{y}^T\mathbf{x}+b\right)\mu(d\mathbf{x}) = 0, \quad \forall \mathbf{y} \in \mathbb{R}^k \text{ and } b \in \mathbb{R}, \tag{23.2}$$

implies $\mu \equiv 0$.

Lemma 23.1 (Discriminatory of sigmoidal functions). *If a sigmoidal function is bounded and measurable, then it is discriminatory.*

Proof. We want to show that (23.2) implies $\mu \equiv 0$, and it is sufficient to consider the following class of sigmoidal functions

$$\sigma_\lambda\left(\lambda\left(\mathbf{y}^T\mathbf{x}+b\right)+\varphi\right) \overset{\lambda\to+\infty}{\to} \gamma(\mathbf{x}) := \begin{cases} 1 & \text{if } \mathbf{y}^T\mathbf{x}+b>0 \\ 0 & \text{if } \mathbf{y}^T\mathbf{x}+b<0 \\ \sigma(\varphi) & \text{if } \mathbf{y}^T\mathbf{x}+b=0 \end{cases}.$$

Since $\sigma_\lambda\left(\lambda\left(\mathbf{y}^T\mathbf{x}+b\right)+\varphi\right)$ converges everywhere to $\gamma(\mathbf{x})$ and $\gamma(\mathbf{x})$ is integrable, i.e.,

$$\int_X \gamma(\mathbf{x})\,\mu(d\mathbf{x}) \le \int_X \max\{1,\sigma(\varphi)\}\,\mu(d\mathbf{x}) = \max\{1,\sigma(\varphi)\}\,\mu(X) < \infty,$$

we can apply the dominated convergence theorem 23.2 to obtain

$$\lim_{\lambda\to+\infty}\int_X \sigma\left(\lambda\mathbf{y}^T\mathbf{x}+b+\varphi\right)\mu(d\mathbf{x}) = \int_X \gamma(\mathbf{x})\,\mu(d\mathbf{x}) = \mu(H(\mathbf{y},b)) + \sigma(\varphi)\,\mu(B(\mathbf{y},b)),$$

where $H(\mathbf{y},b) := \left\{\mathbf{x} : \mathbf{y}^T\mathbf{x}+b>0\right\}$ and $B(\mathbf{y},b) := \left\{\mathbf{x} : \mathbf{y}^T\mathbf{x}+b=0\right\}$. From (23.2), we have

$$\mu(H(\mathbf{y},b)) + \sigma(\varphi)\,\mu(B(\mathbf{y},b)) = 0, \quad \forall \mathbf{y}, b, \text{ and } \varphi,$$

which, by passing φ to the limits $-\infty$ and then ∞, implies

$$\mu(H(\mathbf{y},b)) = 0, \text{ and } \mu(B(\mathbf{y},b)) = 0, \quad \forall \mathbf{y}, b,$$

that is, the measure of any half plane under μ vanishes. We now show that this implies $\mu \equiv 0$. To that end, we observe that

$$\int_X \mathbb{1}_{[\theta,\infty)}\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = \mu(H(\mathbf{y},-\theta)) + \mu(B(\mathbf{y},-\theta)) = 0, \text{ and}$$
$$\int_X \mathbb{1}_{(\theta,\infty)}\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = \mu(H(\mathbf{y},-\theta)) = 0,$$

thus by linearity we conclude that

$$\int_X \chi\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = 0$$

for any *simple function* χ supported on any interval in $\mathbb{R}$. Since the set of all simple functions is dense in the space of all measurable functions (see Lemma 23.2), we have

$$\hat{\mu}(\mathbf{y}) := \int_X e^{i\mathbf{y}^T\mathbf{x}} \mu(d\mathbf{x}) = \int_X \cos\left(\mathbf{y}^T\mathbf{x}\right) \mu(d\mathbf{x}) + i \int_X \sin\left(\mathbf{y}^T\mathbf{x}\right) \mu(d\mathbf{x}) = 0, \forall \mathbf{y} \in \mathbb{R}^k,$$

that is, the characteristic function $\hat{\mu}(\mathbf{y})$ of μ is identically zero. Theorem 23.3 can now be applied to end the proof. □

Now comes the main result of the chapter.

Theorem 23.1 (Approximation capability of sigmoidal functions). *Let σ be a continuous discriminatory sigmoidal functions. The set*

$$M := \left\{\sigma\left(\mathbf{y}^T\mathbf{x} + b\right)\right\}, \quad \forall \mathbf{y} \in \mathbb{R}^k, b \in \mathbb{R},$$

is dense in $\mathbb{C}(X)$. In particular, for any $f \in \mathbb{C}(X)$ and any $\varepsilon > 0$, there exists an $n \in \mathbb{N}$ such that

$$\left\| \sum_{i=1}^n \alpha_i \sigma\left(\mathbf{y}_i^T\mathbf{x} + b_i\right) - f(\mathbf{x}) \right\|_\infty < \varepsilon.$$

Proof. We proceed by contradiction using the Hahn-Banach theorem 23.4. Suppose $M \subset \mathbb{C}(X)$ is not dense in $\mathbb{C}(X)$ and thus its closure $\overline{M}$ is a proper subset of $\mathbb{C}(X)$. By the Hahn-Banach extension theorem, there exists a bounded linear functional $\mathscr{L}$ on $\mathbb{C}(X)$ such that $\mathscr{L}|_{\overline{M}} = 0$, but $\mathscr{L} \neq 0$. By a Riesz representationt theorem 23.5, there exists a unique finite signed regular measure μ on $\mathbb{C}(X)$ such that

$$\mathscr{L}(f) = \int_X f \mu(d\mathbf{x}), \quad \forall f \in \mathbb{C}(X).$$

Now taking $f(\mathbf{x}) = \sigma\left(\mathbf{y}^t\mathbf{x} + b\right)$ we have

$$\mathscr{L}(\sigma) = \int_X \sigma\left(\mathbf{y}^T\mathbf{x} + b\right) \mu(d\mathbf{x}) = 0, \quad \forall \mathbf{y} \in \mathbb{R}^k, b \in \mathbb{R},$$

which, by discrimninatory of sigmoidal functions in Lemma 23.1, implies

$$\mu \equiv 0,$$

and hence $\mathscr{L} \equiv 0$, a contradiction, and this ends the proof. □

23.2 Appendix

Theorem 23.2 (Dominated convergence theorem). *Let $f_n(\mathbf{x})$ be sequence of measurable functions with respect to μ such that*

- $\lim_{n\to\infty} f_n(\mathbf{x}) = f(\mathbf{x})$ *for any* $\mathbf{x} \in X$,
- $|f_n(\mathbf{x})| \leq g(\mathbf{x})$, *for all* $\mathbf{x}$, *n and*

- $g(\mathbf{x}) \in L^1(X, \mu)$.

Then

i) $f(\mathbf{x}) \in L^1(X, \mu)$.
ii) $\lim_{n\to\infty} \int_X f_n(\mathbf{x})\, \mu(d\mathbf{x}) = \int_X f(\mathbf{x})\, \mu(d\mathbf{x})$.

That is, (a.e) pointwise convergence implies L^1-convergence.

Definition 23.3 (Simple functions). A simple function is a finite combination of characteristic (or indicator) functions. In particular,

$$f(\mathbf{x}) := \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\mathbf{x}),$$

where $A_i \subset X$ and $n \in \mathbb{N}$ is finite, is called a simple function.

Definition 23.4 (Denseness). A set $B \subset X$ is dense in X if for any $\mathbf{x} \in X$, there exists a sequence in B converges to $\mathbf{x}$ with respect to the topology in X.

Lemma 23.2 (Denseness of simple functions). *The space of simple functions is dense in the space of measurable functions. That is, for any measurable $F(\mathbf{x})$, there exists a sequence of simple functions $f^m(\mathbf{x})$ such that*

$$F(\mathbf{x}) = \lim_{m\to\infty} f^m(\mathbf{x}), \quad \forall \mathbf{x} \in X.$$

Definition 23.5 (Characteristic function of a measure). For a probability measure μ on X, the characteristic function is defined as

$$\hat{\mu}(\mathbf{y}) := \int_X e^{i\mathbf{y}^T\mathbf{x}} \mu(d\mathbf{x}),$$

for any $\mathbf{y} \in X$.

Theorem 23.3 (Equivalence of a measure and its characteristic functions). *For any two probability measures μ_1 and μ_2 on X, we have*

$$\hat{\mu}_2(\mathbf{y}) = \hat{\mu}_2(\mathbf{y}), \forall \mathbf{y} \in X \quad \textit{implies} \quad \mu_1 = \mu_2.$$

In particular, if $\hat{\mu}(\mathbf{y}) = 0$ for all $\mathbf{y} \in X$, then $\mu = 0$.

Theorem 23.4 (Hahn-Banach). *Let W be a real vector space, $V \subset W$, p a sublinear functional on W and L a linear functional on V such that $L(f) \leq p(f)$ for any $f \in V$. Then there exists a linear functional $\mathscr{L}$ on W such that $\mathscr{L}(f) \leq p(f)$ for all $f \in W$ and $\mathscr{L}|_V = L$.*

Theorem 23.5 (Riesz representation theorem for $\mathbb{C}(X)$). *Let X be a compact metric spaces and $\mathscr{L}$ be a linear functional on $\mathbb{C}(X)$. There exists a unique finite regular signed measure $\mu \in \mathscr{M}(X)$ such that*

$$\mathscr{L}(f) = \int_X f(\mathbf{x})\, \mu(d\mathbf{x}), \quad \forall f \in \mathbb{C}(X).$$