# Machine Learning HW3

Student ID: 111652015

## (1) Reading and Explaining Lemmas

[Lemma 3.1] Let $k \in \mathbb{N}_0$ and $s \in 2\mathbb{N}-1$. Then it holds that for all $\varepsilon > 0$ there exists a shallow tanh neural network $\Psi_{s,\varepsilon} : [-M, M] \to \mathbb{R}^{\frac{s+1}{2}}$ of width $\frac{s+1}{2}$ s.t.

$$\max_{\substack{p \le s \\ p \text{ odd}}} \left\| f_p - (\Psi_{s,\varepsilon})_{\frac{p+1}{2}} \right\|_{W^{k,\infty}} \;\le\; \varepsilon.$$

Moreover, the weights of $\Psi_{s,\varepsilon}$ scale as

$$O\Big(\varepsilon^{\frac{s}{2}} \big(2(s+1)\sqrt{2M}\big)^{s(s+3)}\Big) \quad \text{for small } \varepsilon \text{ and large } s.$$

---

**Lemma 3.1** explains that using a single-hidden-layer neural network $\Psi_{s,\varepsilon}$ (activation function tanh) can approximate all odd monomials ($x^p,\ p$ is odd).
The function $\Psi_{s,\varepsilon} : [-M, M] \to \mathbb{R}^{\frac{s+1}{2}}$ means on the interval from $-M$ to $M$. If the maximal degree is $s$, then there are $s$ monomials $(x^1, x^3, x^5, \dots, x^s)$.
Since the neural network is single-layer and uses tanh as the activation function, it can be written as

$$g(x) = \sum_{i=1}^{N} a_i \tanh(b_i x + c_i) + d.$$

And $\max_{\substack{p \le s \\ p \text{ odd}}} \|f_p - (\Psi_{s,\varepsilon})_{\frac{p+1}{2}}\|_{W^{k,\infty}} \le \varepsilon$ means the odd monomials $f_p$ with degree less than $s$, and the prediction error of the neural network can all be controlled within $\varepsilon$, and $k$ denotes the order. For each $m$-th derivative of $f_p$, $m \le k$, there exists $g_p^{(m)}(x)$ so that their difference is less than $\varepsilon$.
The $O$ in the last line says that the magnitudes of the weights in the neural network are proportional to

$$\varepsilon^{-\,s/2} \left(2(s{+}1)\sqrt{\mu}\right)^{s(s+3)}.$$

$\Rightarrow$ the smaller $\varepsilon$ is, or the more odd monomials are required, the larger the weights.

---

[Lemma 3.2] Let $k \in \mathbb{N}_0$, $s \in 2\mathbb{N}-1$ and $M > 0$. For every $\varepsilon > 0$, there exists a shallow tanh neural network $\Psi_{s,\varepsilon} : [-M, M] \to \mathbb{R}^{\frac{3(s+1)}{2}}$ of width $\frac{s+1}{2}$ s.t.

$$\max_{p \le s} \left\| f_p - (\Psi_{s,\varepsilon}) \right\|_{W^{k,\infty}} \;\le\; \varepsilon.$$

Furthermore, the weights scale as $O\Big(\varepsilon^{-\,s/2} \,((s+2)\sqrt{M})^{\frac{3s(s+3)}{2}}\Big)$ for small $\varepsilon$ and large $s$.

---

**Difference from 3.1** Slightly different from 3.1 is that, Lemma 3.2 assumes that to simultaneously fit $x, x^2, \ldots, x^s$ the number of hidden-layer neurons required is $\frac{3(s+1)}{2}$; the rest is the same as Lemma 3.1.

## (2) Unanswered Question

1. How do the size of the training dataset and the number of training epochs affect the effectiveness of training?
2. Does there exist a formula that can describe data size, number of repetitions, and other parameters, and the influence on the error of the target of the trained neural network?