114-1 Machine Learning

Week 3 Assignment

Ming Hsun Wu

September 23, 2025

Problem 1.

Ryck et al., *On the approximation of functions by tanh neural networks*

Lemma 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 $\frac{s+1}{2}$ such that*

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

Moreover, the weights of $\Psi_{s,\varepsilon}$ scale as $O\Big(\varepsilon^{-s/2} \left(2(s+2)\sqrt{2M}\,\right)^{s(s+3)}\Big)$ for small ε and large s.

Lemma 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 $\psi_{s,\varepsilon} : [-M, M] \to \mathbb{R}^s$ of width $\frac{3(s+1)}{2}$ such that*

$$\max_{p \leq s} \left\| f_p - (\psi_{s,\varepsilon})_p \right\|_{W^{k,\infty}} \leq \varepsilon.$$

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

NOTE OF PROBLEM 1.

1 Lemma 1 :

Lemma statement:

Lemma 1 tells us that for any odd-degree polynomial, it is possible to find a shallow tanh neural network that approximates it. Moreover, not only can the function itself be approximated closely, but derivatives of any order can also be approximated.

Why we can do it?:

The key lies in the tanh tanh function. From calculus, we know through the Taylor expansion that many functions can be expressed as a series in powers of x. The function $tanh(x)$ is special because its expansion contains only odd-degree terms:

$$tanh(x) = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + ...$$

This makes $tanh(x)$ particularly suitable for approximating odd-power functions. By choosing appropriate scalings and combining several $tanh$ functions, one can effectively approximate various odd-degree terms.

2 Lemma 2 :

Lemma statement:

Lemma 2 further tells us that for any polynomial, there exists a neural network composed of *tanh* functions that can approximate it. The odd-degree terms are approximated in the same manner as described in Lemma 1, while the even-degree terms can also be approximated by leveraging the method from Lemma 1. Similarly, not only can the function itself be approximated with sufficient accuracy, but derivatives of any order can also be approximated.

Why we can do it?:

As stated in Lemma 1, we can approximate odd-degree polynomials. Lemma 2 further extends this result, showing that for polynomials of any degree, there exists a neural network composed of *tanh*functions that can approximate them. The key idea is to approximate even-degree terms using odd-degree terms through finite difference methods (see the formula below). Once the odd-degree terms are obtained, Lemma 1 can then be applied to establish Lemma 2.

$$(\psi)_{2n}(y) = \frac{1}{2\alpha(2n+1)}\Big(\hat{f}_{2n+1,h}(y+\alpha) - \hat{f}_{2n+1,h}(y-\alpha) - 2\sum_{k=0}^{n-1}\binom{2n+1}{2k}\alpha^{2(n-k)+1}(\psi)(y)_{2k}\Big)$$

This explains why the network in Lemma 2 needs to be larger, since additional units are required to handle the even-degree terms.

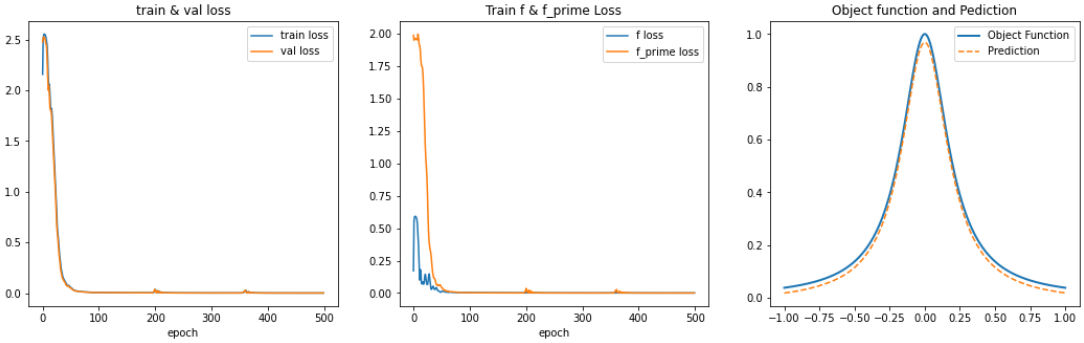
Problem 2.

If possible, I hope the teacher can write a little slower, because sometimes it is hard to tell what the teacher is writing...

PROBLEM 3.

Use the same code from Assignment 2 - programming assignment 1 to calculate the error in approximating the derivative of the given function.

SOLUTION.



NOTE OF PROBLEM 3.

About my hypothesis	
training data	300 equally spaced points in the interval $[1, 1]$.
testing data	500 equally spaced points in the interval $[1, 1]$.
$f_\theta : \mathbb{R} \to \mathbb{R}$	$f_\theta = t_3(L_3(t_2(L_2(t_1(L_1(x))))))$
	where $t_i = tanh(x)$ $L_i = W_i x + b_i$
Hidden layer and activation function	3 *linear layer* with *tanch*
Trainable params	8,577
Loss function	$mse(f_\theta(x), y) + mse(f'_\theta(x), y')$, y is true data
epoch	500

Following the HW2 code, I modified the training and validation processes by adding a step that computes the derivatives of my model's outputs using torch.autograd.grad. I then combined these results with the Runge function to calculate both the function loss and the derivative loss. The sum of these two losses was used as my defined loss function to update the model parameters.

For the trained model, the MSE on the test data is 0.000840.