On the approximation of functions by tanh neural networks

- **Lemma 3.1**

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

$$\max_{\substack{p \le s, \\ p \text{ odd}}} \left\| f_p - (\Psi_{s,\epsilon})_{\frac{p+1}{2}} \right\|_{W^{k,\infty}} \le \epsilon, \tag{17}$$

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

對於任何奇數次方的多項式，都可以建構一個淺層的 tanh 神經網路，使其輸出與該多項式的結果極為接近。
只要增加網路隱藏層中的神經元數量，這個近似的誤差就可以達到想要的任何微小程度。

證明過程主要透過$tanh(y)$和導數（例如函數 f(x)=x^p，p 次導數會是一個常數 p!），因為在神經網路中直接計算高階導數很複雜，因此會使用離散的導數：central finite difference 記作δ_h^p

首先對於一個函數 f(x)，一階 central finite difference 似於其一階導數：

$$\frac{f(x+h) - f(x-h)}{2h} \approx f`(x)$$

而對於$tanh(y)$，在 y=0 附近的泰勒展開式是：

$$tanh(y) = y - \frac{1}{3} * y^3 + \frac{2}{15} * y^5 - \cdots = c_1 * y + c_3 * y^3 + c_5 * y^5$$

因為展開並沒有偶次方項，y^p也在$tanh(y)$內，所以才會使用 tanh。

對於$\delta_h^p[\sigma](y)$：$\delta_h^p[\sigma](y) = \sum_{i=0}^{p}(-1)^{p-i}\binom{p}{i}\sigma\left(y + \left(i - \frac{p}{2}\right)h\right)$

將y^k代入後$\delta_h^3[tanh](y)$

1. 如果 k＜p ，那麼 y^k 的 p 次導數是 0。結果是 0。
2. 如果 k＝p ，那麼 y^p 的 p 次導數是 p！。h^p·p! （一個非零常數）。
3. 如果 k＞p ，結果會是一個新的多項式。

所以如果假如要近似y^3，$\sigma_h^3[tanh](y)$就會等於(一個常數)+(含y^2的項)+(含y^4的項)+⋯

這前幾項就會是y^3的泰勒展開式

則$\delta_h^3[\sigma](y)$展開則會等於$1 \cdot tanh(y-1.5h) - 3 \cdot tanh(y-0.5h) + 3 \cdot tanh(y+0.5h) - 1 \cdot tanh(y+1.5h)$

對應到神經元就會是

輸入層：輸入節點接收 y

隱藏層：有 4 個神經元。

神經元 1：計算 tanh(1 · y – 1.5h)輸入權重是 1，bias 是 − 1.5 h

神經元 2：計算 tanh(1 · y – 1.5h)輸入權重是 1，偏置是 − 0.5 h

神經元 3：計算 tanh(1 · y ＋ 1.5h)

神經元 4：計算 tanh(1 · y ＋ 1.5h)

輸出層：一個輸出節點。將隱藏層的 4 個輸出，分別乘以權重 +1, -3, +3, -1，然後相加

δ_h^p作用在泰勒展開式中那些高於 p 次的項則是誤差，會正比於h^2

- **Lemma 3.2**

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

$$\max_{p \le s} \left\| f_p - (\psi_{s,\epsilon})_p \right\|_{W^{k,\infty}} \le \epsilon. \tag{26}$$

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

在 Lemma 3.1 中，可以透過 Tanh 函數的奇函數特性，提出了任意奇數次方的多項式。但那是因為 Tanh 的泰勒展開式中，本身就含有這些奇數項。 但同樣使用方法針對偶數項則會失敗，因此透過 Recursive Construction 來建構。

首先先觀察$(y+\alpha)^3$和$(y-\alpha)^3$的展開式

$$(y+\alpha)^3 = y^3 + 3\alpha y^2 + 3\alpha^2 y + \alpha^3$$
$$(y-\alpha)^3 = y^3 - 3\alpha y^2 + 3\alpha^2 y - \alpha^3$$

將兩式相減：$(y+\alpha)^3 - (y-\alpha)^3 = (y^3 - y^3) + \left(3\alpha y^2 - (-3\alpha y^2)\right) + (3\alpha^2 y - 3\alpha^2 y) + \left(\alpha^3 - (-\alpha^3)\right)$

$$= 6\alpha y^2 + 2\alpha^3$$

則$y^2 = \frac{1}{6\alpha}[(y+\alpha)^3 - (y-\alpha)^3] - \frac{\alpha^2}{3}$

$-\frac{\alpha^2}{3}$可以看做y^0，$(y+\alpha)^3$和$(y-\alpha)^3$則可以根據 Lemma3.1 來近似

同理y^4的問題可以轉換成近似 $(y+\alpha)^5$+近似$(y-\alpha)^5$+近似y^2+近似y^0

因此更高次的偶次方項都可以轉換成這些形式。

- **Unanswered Questions**

在網絡架構中運用這些奇函數或是週期函數（層連結等），不知道對學習效率與準確性是否會有幫助？

用泰勒展開與差分法和梯度下降優化模擬函數的方法各有啥優缺點？

對於 p 值很大的近似x^p，神經元數量和網路深度是否會有訓練困難的問題？