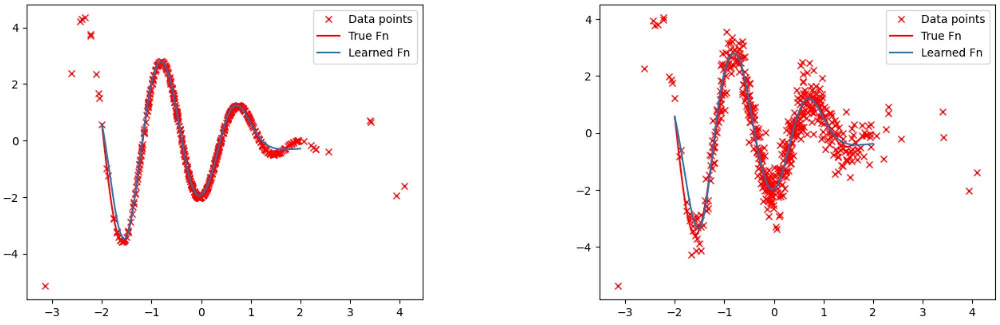
MFDNN HW2

MinGyu Shin

Problem 1 : *3-layer MLP to fit a univariate function.* & **Problem 2** : *Deep learning operates under $p \gg N$.*

# of trainable parameter = 4353 ($\gg N = 512$) (from layer1 = 64+64, from layer2 = 64*64 + 64, from layer 3 =64 + 1). The results are almost same regardless with nor without noise. It could mean that even if trainable parameter is much more than data, the large neural network do not overfit to the data.



(a) P1 result

(b) P2 result

Problem 3 : *Basic properties of CE loss.*

Assume $k \geq 2$.

$$(a) \quad 0 < \exp(f_y) < \sum_{j=1}^{k} \exp(f_j) \quad (\because \exp(f_j) > 0,\ \forall j)$$

$$\Rightarrow \qquad 0 < \frac{exp(f_y)}{\sum_{j=1}^{k} \exp(f_j)} < 1 \qquad \Rightarrow \qquad 0 < -\log(\frac{exp(f_y)}{\sum_{j=1}^{k} \exp(f_j)}) < \infty$$

(b) Note that $(\lambda e_y)_j = \begin{cases} \lambda & j = y \\ 0 & o.w. \end{cases}$. Since

$$\frac{exp(f_y)}{\sum_{j=1}^{k} \exp(f_j)} = \frac{\exp(\lambda)}{\exp(\lambda) + (k-1)\exp(0)} = \frac{\exp(\lambda)}{\exp(\lambda) + (k-1)} \to 1- \ as\ \lambda \to \infty,$$

$$l^{CE}(\lambda e_y, y) \to 0\ as\ \lambda \to \infty$$

Problem 4 : *Derivative of max.*

If given $x, I = \underset{i}{argmax} f_i(x)$ is unique, then claim that

$$\exists\ \epsilon > 0, \quad s.t. \quad f(y) = f_I(y) > f_i(y), y \in (x-\epsilon, x+\epsilon) \qquad \forall i. \tag{1}$$

Since $A = [x - 1/2, x + 1/2]$ is compact and f_i is differentiable on A, there exist M_i such that $|f_i'| < M_i$ on A. Let $\delta_i = f_I(x) - f_i(x) > 0$. Set $0 < \epsilon < \min(1/2, \{\frac{\delta_i}{M_i}\}_{i=1,\dots,N})$. Then for $y \in (x-\epsilon, x+\epsilon)$, $f_i(y) < f_i(x) + \epsilon M_i < f_i(x) + \delta_i = f_I(x)$, that is (1) is satisfied with the ϵ. Thus f is differentiable at x and

$$\frac{d}{dx} f(x) = \frac{d}{dx} f_I(x)$$

Problem 5 : *Basic Properties of activation functions.*

(a) $\sigma(\sigma(z)) = \max\{0, \delta(z)\} = \max\{0, \max\{0, z\}\} = \max\{0, z\} = \sigma(z)$

(b) $0 < \sigma'(z) = \frac{e^z}{1+e^z} < 1$. Since if $|f'(x)| \le L$, for all x,y with $x \neq y, |\frac{f(y)-f(x)}{y-x}| = |f'(c)| \le L$ with some c between x and y by Mean Value Theorem. Thus softplus has Lipschitz continuous derivatives with constant $L = 1$.

ReLU is not differentiable at 0, but just set derivative of ReLU any scaler value c at 0. That is $\rho'(z) = \begin{cases} 1 & z > 0 \\ c & z = 0 \\ 0 & z < 0 \end{cases}$. For any given L, with $x, y \in (-1/2L, 0), (0, 1/2L)$ respectively,

$$1 = |\rho'(y) - \rho'(x)| > L|y - x|.$$

Thus derivatives of RELU is not continuous and also not a Lipschitz function.

(c) Note that $\rho(z) = \frac{1-e^{-2z}}{1+e^{-2z}} = \frac{2}{1+e^{-2z}} - 1 = 2\sigma(2z) - 1$. Given $L > 1, A_1, \dots, A_L$, and $b_1, \dots, b_L$,

$$\begin{cases} C_1 = A_1/2 \quad d_1 = b_1/2 \\ C_i = A_1/4 \quad d_i = b_i/2 + A_i \cdot \mathbf{1}_{n_{i-1}}/4 \qquad i = 2, \dots, L-1 \\ C_L = A_L/2 \quad d_L = b_L + A_L \cdot \mathbf{1}_{n_{L-1}}/2 \end{cases}$$

represent identical $x \mapsto y_L$ mappings. The reverse way is similar.

Problem 6 : *Vanishing gradients.* Since $\frac{\partial}{\partial a_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \frac{\partial f_\theta}{\partial a_j}$, and $\frac{\partial f_\theta}{\partial a_j} = u_j \sigma'(a_j x + b_j) x = 0, a_j$ is never updated during training. Similarly b_j too. Thus $a_j X_i + b_j < 0$ for all $i = 1, \dots, N$.

Problem 7 : *Leaky ReLU.* $\frac{\partial f_\theta}{\partial a_j} = u_j \sigma'(a_j x + b_j) x = u_j \alpha x$ could not be zero, that is a_j would be updated during traning. Similarly b_j too.