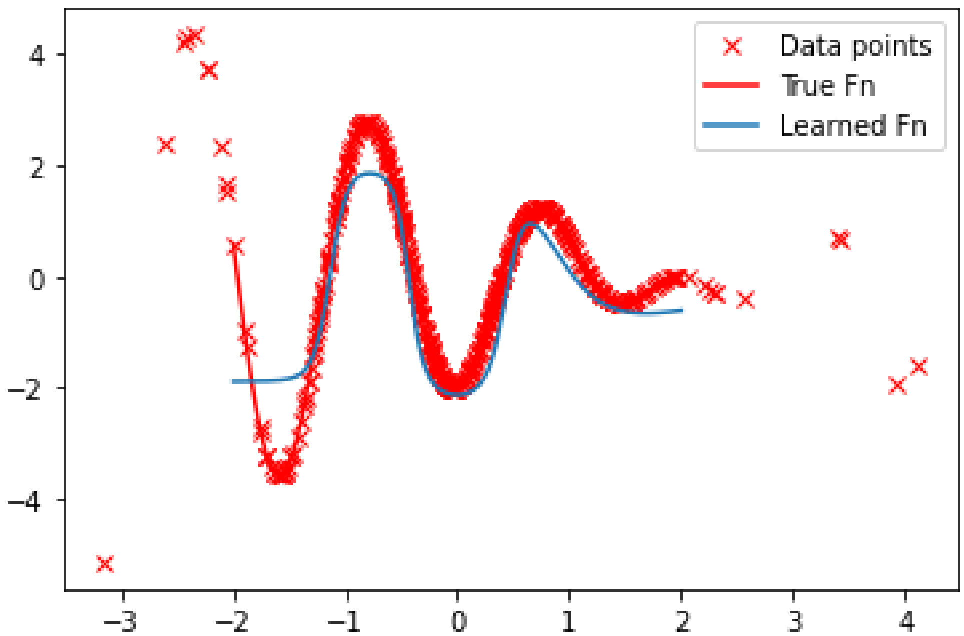
#1.

threelayerSGD.py

```python

import torch
import numpy as np
from torch import nn, optim
from torch.nn import functional as F
from torch.utils.data import TensorDataset, DataLoader
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

alpha = 0.1
K = 1000
B = 128
N = 512

def f_true(x) :
    return (x-2) * np.cos(x*4)

torch.manual_seed(0)
X_train = torch.normal(0.0, 1.0, (N,))
y_train = f_true(X_train)
X_val = torch.normal(0.0, 1.0, (N//5,))
y_val = f_true(X_val)

train_dataloader = DataLoader(TensorDataset(X_train.unsqueeze(1), y_train.unsqueeze(1)), batch_si
test_dataloader = DataLoader(TensorDataset(X_val.unsqueeze(1), y_val.unsqueeze(1)), batch_size=B)

'''
unsqueeze(1) reshapes the data into dimension [N,1],
where is 1 the dimension of an data point.

The batchsize of the test dataloader should not affect the test result
so setting batch_size=N may simplify your code.
In practice, however, the batchsize for the training dataloader
is usually chosen to be as large as possible while not exceeding
the memory size of the GPU. In such cases, it is not possible to
use a larger batchsize for the test dataloader.
'''

class MLP(nn.Module):
    def __init__(self):

        super().__init__()
        self.linear1 = nn.Linear(1, 64, bias=True)
        self.linear2 = nn.Linear(64, 64, bias=True)
        self.linear3 = nn.Linear(64, 1, bias=True)


    def forward(self, x):
        x = x.float().view(-1, 1)
        x = nn.functional.sigmoid(self.linear1(x))
        x = nn.functional.sigmoid(self.linear2(x))
        x = (self.linear3(x))

        return x


model = MLP()
loss_function = nn.MSELoss()
model . linear1 . weight . data = torch . normal (0 , 1 , model . linear1 . weight . shape )
model . linear1 . bias . data = torch . full ( model . linear1 . bias . shape , 0.03)
model . linear2 . weight . data = torch . normal (0 , 1 , model . linear2 . weight . shape )
model . linear2 . bias . data = torch . full ( model . linear2 . bias . shape , 0.03)
model . linear3 . weight . data = torch . normal (0 , 1 , model . linear3 . weight . shape )
model . linear3 . bias . data = torch . full ( model . linear3 . bias . shape , 0.03)

optimizer = torch.optim.SGD(model.parameters(), lr = alpha)

for epoch in range(K):
    for x, y in train_dataloader:
        optimizer.zero_grad()
        train_loss = loss_function(model(x) , y)
        train_loss.backward()

        optimizer.step()


with torch.no_grad():
    xx = torch.linspace(-2,2,1024).unsqueeze(1)
    plt.plot(X_train,y_train,'rx',label='Data points')
    plt.plot(xx,f_true(xx),'r',label='True Fn')
    plt.plot(xx, model(xx),label='Learned Fn')
plt.legend()
plt.show()


'''
When plotting torch tensors, you want to work with the
torch.no_grad() context manager.

When you call plt.plot(...) the torch tensors are first converted into
numpy arrays and then the plotting proceeds.
However, our trainable model has requires_grad=True to allow automatic
gradient computation via backprop, and this option prevents
converting the torch tensor output by the model to a numpy array.
Using the torch.no_grad() context manager resolves this problem
as all tensors are set to requires_grad=False within the context manager.

An alternative to using the context manager is to do
plt.plot(xx, model(xx).detach().clone())
The .detach().clone() operation create a copied pytorch tensor that
has requires_grad=False.

To be more precise, .detach() creates another tensor with requires_grad=False
(it is detached from the computation graph) but this tensor shares the same
underlying data with the original tensor. Therefore, this is not a genuine
copy (not a deep copy) and modifying the detached tensor will affect the
original tensor is weird ways. The .clone() further proceeds to create a
genuine copy of the detached tensor, and one can freely manipulate and change it.
(For the purposes of plotting, it is fine to just call .detach() without
.clone() since plotting does not change the tensor.)

This discussion will likely not make sense to most students at this point of the course.
We will revisit this issue after we cover backpropagation.
```



Pytorch로 training을 해보면 위와 같은 결과를 얻을 수 있다.

#2.

먼저 parameter의 개수를 계산해보자. 일단,

$64 \times 1 + 64 \times 64 + 64 \times 1$ 개의 parameter가 존재하며, bias를 추가로 계산해보면

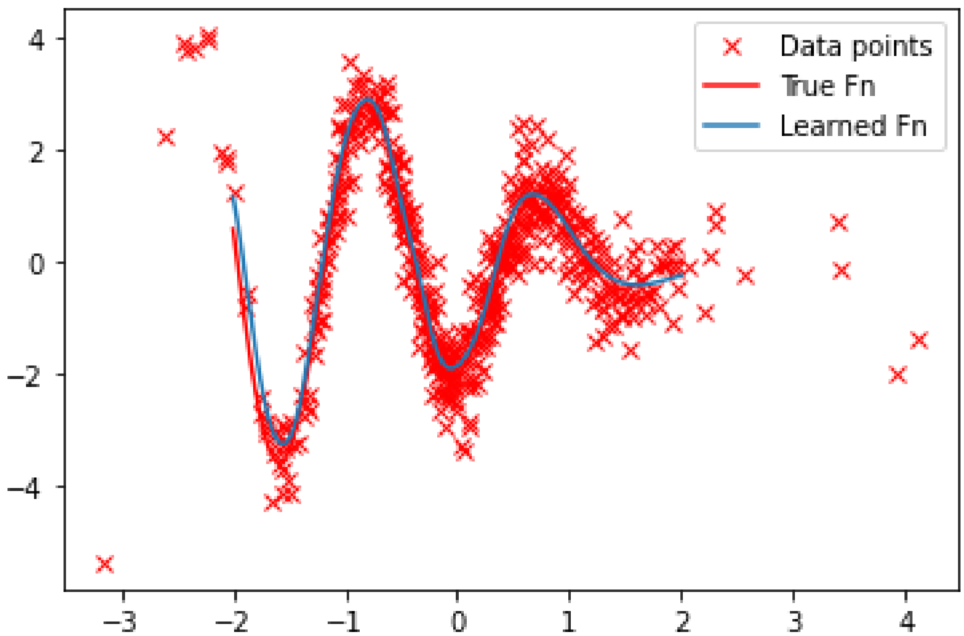
$64 + 64 + 1$ 개가 추가로 존재함을 알 수 있다.

이는 총 4353개이다. (p>N!)

Y -train 에 noise 를 추가하고 동일한 실험을 반복하였다. 코드와 결과는 다음과 같았다. 실험 결과, 그래프의 개형이 비슷하게 나타났고, outlier 점에 의한 overfitting 현상도 오히려 감소한 것과 같은 모습이 보여졌다.

threelayerSGD.py ×

```python
import numpy as np
from torch import nn, optim
from torch.nn import functional as F
from torch.utils.data import TensorDataset, DataLoader
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

alpha = 0.1
K = 1000
B = 128
N = 512

def f_true(x) :
    return (x-2) * np.cos(x*4)

torch.manual_seed(0)
X_train = torch.normal(0.0, 1.0, (N,))
#y_train = f_true(X_train)
y_train = f_true ( X_train ) + torch . normal (0 , 0.5 , X_train . shape )
X_val = torch.normal(0.0, 1.0, (N//5,))
y_val = f_true(X_val)

train_dataloader = DataLoader(TensorDataset(X_train.unsqueeze(1), y_train.unsqueeze(1)), batch_si
test_dataloader = DataLoader(TensorDataset(X_val.unsqueeze(1), y_val.unsqueeze(1)), batch_size=B)

'''
unsqueeze(1) reshapes the data into dimension [N,1],
where is 1 the dimension of an data point.

The batchsize of the test dataloader should not affect the test result
so setting batch_size=N may simplify your code.
In practice, however, the batchsize for the training dataloader
is usually chosen to be as large as possible while not exceeding
the memory size of the GPU. In such cases, it is not possible to
use a larger batchsize for the test dataloader.
'''

class MLP(nn.Module):
    def __init__(self):

        super().__init__()
        self.linear1 = nn.Linear(1, 64, bias=True)
        self.linear2 = nn.Linear(64, 64, bias=True)
        self.linear3 = nn.Linear(64, 1, bias=True)


    def forward(self, x):
        x = x.float().view(-1, 1)
        x = nn.functional.sigmoid(self.linear1(x))
        x = nn.functional.sigmoid(self.linear2(x))
        x = (self.linear3(x))

        return x


model = MLP()
loss_function = nn.MSELoss()
model . linear1 . weight . data = torch . normal (0 , 1 , model . linear1 . weight . shape )
model . linear1 . bias . data = torch . full ( model . linear1 . bias . shape , 0.03)
model . linear2 . weight . data = torch . normal (0 , 1 , model . linear2 . weight . shape )
model . linear2 . bias . data = torch . full ( model . linear2 . bias . shape , 0.03)
model . linear3 . weight . data = torch . normal (0 , 1 , model . linear3 . weight . shape )
model . linear3 . bias . data = torch . full ( model . linear3 . bias . shape , 0.03)

optimizer = torch.optim.SGD(model.parameters(), lr = alpha)

for epoch in range(K):
    for x, y in train_dataloader:
        optimizer.zero_grad()
        train_loss = loss_function(model(x) , y)
        train_loss.backward()

        optimizer.step()


with torch.no_grad():
    xx = torch.linspace(-2,2,1024).unsqueeze(1)
    plt.plot(X_train,y_train,'rx',label='Data points')
    plt.plot(xx,f_true(xx),'r',label='True Fn')
    plt.plot(xx, model(xx),label='Learned Fn')
plt.legend()
plt.show()


'''
When plotting torch tensors, you want to work with the
torch.no_grad() context manager.

When you call plt.plot(...) the torch tensors are first converted into
numpy arrays and then the plotting proceeds.
However, our trainable model has requires_grad=True to allow automatic
gradient computation via backprop, and this option prevents
converting the torch tensor output by the model to a numpy array.
Using the torch.no_grad() context manager resolves this problem
as all tensors are set to requires_grad=False within the context manager.

An alternative to using the context manager is to do
plt.plot(xx, model(xx).detach().clone())
The .detach().clone() operation create a copied pytorch tensor that
has requires_grad=False.

To be more precise, .detach() creates another tensor with requires_grad=False
(it is detached from the computation graph) but this tensor shares the same
underlying data with the original tensor. Therefore, this is not a genuine
copy (not a deep copy) and modifying the detached tensor will affect the
original tensor is weird ways. The .clone() further proceeds to create a
genuine copy of the detached tensor, and one can freely manipulate and change it.
(For the purposes of plotting, it is fine to just call .detach() without
.clone() since plotting does not change the tensor.)

This discussion will likely not make sense to most students at this point of the course.
We will revisit this issue after we cover backpropagation.

'''
```



Data points
True Fn
Learned Fn
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4

#3 $\ell^{CE}(f, y) = -\log\left(\frac{\exp(f_y)}{\sum_{j=1}^{k}\exp(f_j)}\right)$, $f \in \mathbb{R}^k$, $y \in \{1, \cdots, k\}$

(a) $\frac{\exp(f_y)}{\sum_{j=1}^{k}\exp(f_j)} = \frac{\exp(f_y)}{\exp(f_1)+\cdots+\exp(f_y)+\cdots+\exp(f_k)}$ Since $\exp(f_i) > 0$ for all $i \in \{1, \dots, k\}$,

$0 < \frac{\exp(f_y)}{\sum \exp(f_j)} < 1$

Also, we know that $-\infty < \log(x) < 0$ when $x \in (0, 1)$

Therefore, $-\infty < \log\left(\frac{\exp(f_y)}{\sum\exp(f_j)}\right) < 0$ and thus, $0 < \ell^{CE}(f, y) < \infty$

(b) $\ell^{CE}(\lambda e_y, y) = -\log\left(\frac{\exp(\lambda)}{(k-1)+\exp(\lambda)}\right) = \log\left(\frac{e^\lambda + k-1}{e^\lambda}\right) = \log\left(1 + \frac{k-1}{e^\lambda}\right) \xrightarrow{\lambda\to\infty} \log(1) = 0\,!$

#4 Let $f(x) = \max\{f_1(x), \cdots, f_k(x)\}$. $f_i(x)$: diff. univariate fn.

pf) for a given x, maximizing index $I = \arg\max_i \{f_i(x)\}$ is unique.

Therefore, $f_I(x) > f_i(x)$ holds for all $i \neq I$

Since $\forall i$, $f_i(x)$ is differentiable, $\forall i$ $f_i(x)$ is continuous. $\therefore$ $\forall i \neq I$ $f_I(x) - f_i(x)$ is continuous.

By the definition of continuity, $f_I(x) - f_i(x) > 0$, $\exists \delta > 0$. s.t if $x-\delta < t < x+\delta$, $f_I(t) - f_i(t) > 0$ holds.

This means that $\forall t \in (x-\delta, x+\delta)$, $f_I(t) = f(t)$.

Therefore for $|h| < |\delta|$, $\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{f_I(x+h)-f_I(x)}{h}$ holds.

i.e. $f'(x) = f_I'(x)\,!$ ▢

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

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

(b) Softplus $\sigma(z) = \log(1+e^z)$, $\sigma'(z) = \dfrac{e^z}{1+e^z}$

Let's show that $\sigma(z)$ has Lipschitz continuous derivative

$$|\sigma'(x) - \sigma'(y)| = \left| \frac{e^x}{1+e^x} - \frac{e^y}{1+e^y} \right| = \left| \frac{e^x + e^x e^y - e^y - e^x e^y}{(1+e^x)(1+e^y)} \right| = \left| \frac{e^x - e^y}{(1+e^x)(1+e^y)} \right|$$

$$= \left| \frac{e^c}{(1+e^x)(1+e^y)} \right| |x-y| \quad \exists c,\ c \in (\min(x,y), \max(x,y)) \text{ by Mean Value Theorem}$$

Since $c < \max(x,y)$, $e^c < 1+e^x$ or $e^c < 1+e^y$ holds,

$$\therefore |\sigma'(x) - \sigma'(y)| = \left| \frac{e^c}{(1+e^x)(1+e^y)} \right| |x-y| < |x-y|$$ $\therefore$ Softplus has Lipschitz continuous derivative.

On the other hand, Consider ReLU, $\sigma(z) = \max(0, z)$

$$\frac{|\sigma'(z) - \sigma'(-z)|}{|z-(-z)|} = \frac{+1}{2|z|} \to \infty \text{ as } z \to 0$$

$\therefore \forall L, \exists \delta > 0$ s.t if $0 < |z| < \delta$, $|\sigma'(z) - \sigma'(-z)| > L|z-(-z)|$. $\therefore$ ReLU does not have Lipschitz continuous derivative!

(c)

$$\rho(z) = \frac{1-e^{-2z}}{1+e^{-2z}} = \frac{e^{2z}-1}{e^{2z}+1} = \frac{2e^{2z} - (1+e^{2z})}{1+e^{2z}} = 2\frac{1}{1+e^{-2z}} - 1 = 2\sigma(2z) - 1.$$

$L > 1$, $A_1 \dots, A_L$, $b_1 \dots b_L$ 이 주어져서, $y_L = A_L y_{L-1} + b_L, \dots, y_1 = \sigma(A_1 x + b_1)$ 이라 하자.

① $$y_L = A_L \sigma(A_{L-1} y_{L-2} + b_{L-1}) + b_L = C_L \rho(C_{L-1} y_{L-2} + d_{L-1}) + d_L$$

$\Rightarrow \boxed{C_L = \frac{1}{2}A_L,\ d_L = b_L + \frac{1}{2}A_L\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}}$ $$= C_L\left(2\sigma\left(\boxed{2C_{L-1}y_{L-2} + 2d_{L-1}}\right)\right) + d_L - C_L\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

$$A_{L-1}\sigma(A_{L-2}y_{L-3} + b_{L-2}) + b_{L-1} = 2C_{L-1}\rho(C_{L-2}y_{L-3} + d_{L-2}) + 2d_{L-1}$$

$$= 2C_{L-1}\left(2\sigma(2C_{L-2}y_{L-3} + 2d_{L-2}) - \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right) + 2d_{L-1}$$

$$= 4C_{L-1}\sigma\left(\boxed{2C_{L-2}y_{L-3} + 2d_{L-2}}\right) + 2d_{L-1} - 2C_{L-1}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

$\Rightarrow C_{L-1} = \frac{1}{4}A_{L-1}$, $d_{L-1} = \frac{1}{2}\left(b_{L-1} + \frac{1}{2}A_{L-1}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right)$

$= \frac{1}{2}b_{L-1} + \frac{1}{4}A_{L-1}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

같은 꼴로 완성!

② $\boxed{C_i = \frac{1}{4}A_i,\ d_i = \frac{1}{2}b_i + \frac{1}{4}A_i\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \quad (1 < i < L)}$ (1로 구성된 $\mathbb{R}^i$ vector)

③ $\boxed{C_1 = \frac{1}{2}A_1,\ d_1 = \frac{1}{2}b_1}$

임의의 A_i, b_i에 대하여, 위와 같이 C_i, d_i를 잡으면 두 MLP는 equivalent 하다!

pf) <i> L=2 인 경우

$y_2 = A_2 y_1 + b_2$ $\qquad$ $\bar{y}_2 = C_2 \bar{y}_1 + d_2 = \frac{1}{2} A_2 \bar{y}_1 + b_2 + \frac{1}{2} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$y_1 = \sigma(A_1 x + b_1)$ $\qquad$ $\bar{y}_1 = \rho(C_1 x + d_1) = \rho\left(\frac{1}{2} A_1 + \frac{1}{2} b_1\right)$

$$\bar{y}_2 = \frac{1}{2} A_2 \left(2\sigma(A_1 x + b_1) \cancel{- \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} \right) + b_2 \cancel{+ \frac{1}{2} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}}$$

$$= A_2 \sigma(A_1 x + b_1) + b_2 = y_2.$$

<ii> L=3 인 경우

$y_3 = A_3 y_2 + b_3$ $\qquad$ $\bar{y}_3 = \frac{1}{2} A_3 \bar{y}_2 + b_3 + \frac{1}{2} A_3 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$y_2 = \sigma(A_2 y_1 + b_2)$ $\qquad$ $\bar{y}_2 = \rho\left(\frac{1}{4} A_2 \bar{y}_1 + \frac{1}{2} b_2 + \frac{1}{4} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right)$

$y_1 = \sigma(A_1 x + b_1)$ $\qquad$ $\bar{y}_1 = \rho\left(\frac{1}{2} A_1 x + \frac{1}{2} b_1\right) = 2\sigma(A_1 x + b_1) - \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$$\bar{y}_2 = \rho\left(\frac{2}{4} A_2 \sigma(A_1 x + b_1) \cancel{- \frac{1}{4} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} + \frac{1}{2} b_2 \cancel{+ \frac{1}{4} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} \right)$$

$$= 2\sigma\left(A_2 \sigma(A_1 x + b_1) + b_2 \right) - \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

$$\bar{y}_3 = A_3 \sigma\left(A_2 \sigma(A_1 x + b_1) + b_2 \right) + b_3 = y_3 \ !$$

<iii> L 일때 성립 가정.

즉 $y_L = A_L y_{L-1} + b_L = \bar{y}_L = \frac{1}{2} A_L \bar{y}_{L-1} + b_L + \frac{1}{2} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$ 인 상황이라 하자. ⋯⋯(*)

C_L은 $\frac{1}{4} A_L$로, d_L은 $\frac{1}{2} b_L + \frac{1}{4} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$로 바꿀때 $y_{L+1} = \bar{y}_{L+1}$ 임을 보임으로서 L+1 일때 성립함을 보이자.

$$\cancel{A_{L+1}} \sigma(A_L y_{L-1} + b_L) \cancel{+ b_{L+1}} = \frac{1}{2} A_{L+1} \rho\left(\frac{1}{4} A_L \bar{y}_{L-1} + \frac{1}{2} b_L + \frac{1}{4} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \right) + b_{L+1} + \frac{1}{2} A_{L+1} \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

$$= \cancel{A_{L+1}} \sigma\left(\frac{1}{2} A_L \bar{y}_{L-1} + b_L + \frac{1}{2} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \right) \cancel{+ b_{L+1}}$$

$\iff A_L y_{L-1} + b_L = \frac{1}{2} A_L \bar{y}_{L-1} + b_L + \frac{1}{2} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$ ⋯ (*)와 동일한 식! 즉 성립한다!

즉 <i>, <ii>, <iii>, M.I 에 의하여 전 page 에 주는대로 C_i, d_i 를 잡으면 동일한 MLP를 얻게 된다는 사실을 엄밀히 증명하였다.

#6. $a_j^0 X_i + b_j^0 < 0 \quad \forall i \in \{1, \dots, N\}$. $\quad \sigma$: ReLU

We want to show that $a_j^k X_i + b_j^k < 0$ for all i and k (to show that jth ReLU output remains dead)

$$\frac{\partial}{\partial a_j} l(f_\theta(x_i), Y_i) = \frac{\partial l}{\partial f}\frac{\partial f}{\partial a_j} = \frac{\partial l}{\partial f}\frac{\partial u^T \sigma(a x_i + b)}{\partial a_j} = \frac{\partial l}{\partial f} u_j \cdot \sigma'(a_j x_i + b_j) x_i$$

$$= \begin{cases} 0 & \text{if } a_j X_i + b_j < 0 \\ \frac{\partial l}{\partial f} u_j X_i & \text{if } a_j X_i + b_j > 0 \end{cases}$$

$$\frac{\partial}{\partial b_j} l(f_\theta(x_i), Y_i) = \frac{\partial l}{\partial f}\frac{\partial f}{\partial b_j} = \frac{\partial l}{\partial f}\frac{\partial u^T \sigma(a x_i + b)}{\partial b_j} = \frac{\partial l}{\partial f} u_j \sigma'(a_j X_i + b_j)$$

$$= \begin{cases} 0 & \text{if } a_j X_i + b_j < 0 \\ \frac{\partial l}{\partial f} u_j & \text{if } a_j X_i + b_j > 0 \end{cases}$$

If we think about SGD, $\theta_{k+1} = \theta_k - lr \times \frac{\partial l}{\partial \theta_k}$.

Since <i> $a_j^0 X_i + b_j^0 < 0$ ($\forall i$),

<ii> $a_j^k X_i + b_j^k < 0$. Then, $\frac{\partial}{\partial a_j} l = 0$ and $\frac{\partial l}{\partial b_j} = 0$ for all i.

Thus, a_j^{k+1}, b_j^{k+1} remains unchanged and thus $a_j^{k+1} X_i + b_j^{k+1} < 0$ for all i.

By Mathematical induction, <i>, <ii>, we proved that $\forall i, k$, $a_j^k X_i + b_j^k < 0$ holds.

The jth ReLU output is dead throughout training.

#7 If we use Leaky ReLU, $\sigma(z) = \begin{cases} z & z>0 \\ \alpha z & z\le 0 \end{cases}$

$$\frac{\partial \ell(f_\theta(X_i), Y_i)}{\partial a_j} = \begin{cases} \left(\frac{\partial \ell}{\partial f} u_j X_i\right)\alpha & (a_j X_i + b_j < 0) \\ \left(\frac{\partial \ell}{\partial f} u_j X_i\right) & (a_j X_i + b_j > 0) \end{cases}$$

$$\frac{\partial \ell(f_\theta(X_i), Y_i)}{\partial b_j} = \begin{cases} \left(\frac{\partial \ell}{\partial f} u_j\right)\alpha & (a_j X_i + b_j < 0) \\ \frac{\partial \ell}{\partial f} u_j & (a_j X_i + b_j > 0) \end{cases}$$

$\alpha \neq 0$ 이라면, 비록 $a_j^0 X_i + b_j^0 < 0$ 더라도

$\frac{\partial \ell}{\partial a_j} = \left(\frac{\partial \ell}{\partial f} u_j X_i\right)\alpha$: not identically zero

$\frac{\partial \ell}{\partial b_j} = \left(\frac{\partial \ell}{\partial f} u_j\right)\alpha$: not identically zero 이므로,

SGD 에 의해 $a_j^{k+1} = a_j^k - \frac{\partial \ell}{\partial a_j} \times$ learning_rate

$b_j^{k+1} = b_j^k - \frac{\partial \ell}{\partial b_j} \times$ learning_rate 와 같이 update 되고,

#6 번 문제와 같이 gradient 가 "exactly vanish" 하지는 않는다!