Deep learning HW1

Shanyin Tong, st3255@nyu.edu

February 21, 2021

1 Theory

1.1 Two-Layer Neural Nets

1.2 Regression Task

(a) five steps

1) Build the model to represent the architecture of the neural network in `Pytorch`:

$$\text{Linear}_1 \to f \to \text{Linear}_2 \to g. \tag{1}$$

2) Feed forward to get outputs/predictions: for input $\boldsymbol{x}$, obtain output $\hat{\boldsymbol{y}}$ through forward pass of the neural network (like composition of functions)

$$\hat{\boldsymbol{y}} = g(\text{Linear}_2(f(\text{Linear}_1(\boldsymbol{x})))). \tag{2}$$

3) Evaluate the loss function: using data $\boldsymbol{y}$ and prediction from forward pass $\hat{\boldsymbol{y}}$:

$$l_{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2. \tag{3}$$

4) Backward pass to compute gradient: first initiate the gradient by 0, and then back propagate and accumulate the derivative information at each layer to evaluate the gradient of loss function l with respect to neural network parameter (denoted by $\frac{\partial l}{\partial w}$) using the chain rule.

5) Update the parameter w using the gradient obtained in step 4:

$$w = w - \eta \frac{\partial l}{\partial w}, \tag{4}$$

where η is the step size.

(b) forward pass in table 1

(c) downward pass in table 2

(d) For this part and the following text, I will assume $\boldsymbol{z}_2 \in \mathbb{R}^r$ and use notation $z_{j,i}$ to denote the ith component of vector $\boldsymbol{z}_j$, $j = 1, 2, 3$.

$$\left(\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}\right)_{ij} = \frac{\partial f(z_{1,i})}{\partial z_{1,j}} = \frac{\partial (z_{1,i})^+}{\partial z_{1,j}} = \begin{cases} 0, & \text{if } i \neq j \text{ or } z_{1,i} < 0, \\ 1, & \text{if } i = j \text{ and } z_{1,i} \geq 0. \end{cases}, \tag{5}$$

Table 1: forward pass

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$
f	$\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$	$\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}$
Linear_2	$\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}$	$\boldsymbol{W}^{(2)}\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}+\boldsymbol{b}^{(2)}$
g	$\boldsymbol{W}^{(2)}\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}+\boldsymbol{b}^{(2)}$	$\boldsymbol{W}^{(2)}\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}+\boldsymbol{b}^{(2)}$
Loss	$\boldsymbol{W}^{(2)}\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}+\boldsymbol{b}^{(2)}$	$\\|\boldsymbol{W}^{(2)}\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}+\boldsymbol{b}^{(2)}-y\\|^2$

Table 2: downward pass

Parameter	Gradient
$\boldsymbol{W}^{(1)}$	$\boldsymbol{x}\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}\boldsymbol{W}^{(2)}\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$
$\boldsymbol{b}^{(1)}$	$\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}\boldsymbol{W}^{(2)}\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$
$\boldsymbol{W}^{(2)}$	$\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$
$\boldsymbol{b}^{(2)}$	$\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$

where $z_{1,i}=\sum_{j=1}^{n} W_{ij}^{(1)} x_j + b_i^{(1)}$.

$$\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}=\operatorname{diag}\left\{\mathbb{1}_{[0,\infty)}\left(\boldsymbol{z}_1\right)\right\}\in\mathbb{R}^{r\times r}, \tag{6}$$

is a diagonal matrix. Here, $\boldsymbol{z}_1=\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$, $\operatorname{diag}(\cdot)$ builds a diagonal matrix with the input vector as the diagonal elements, $\mathbb{1}_{[0,\infty)}(\cdot)$ is the element-wise indicator function of $[0,\infty)$, defined as:

$$\mathbb{1}_{[0,\infty)}(x)=\begin{cases}1, & \text{if } x\geq 0,\\ 0, & \text{if } x<0.\end{cases} \tag{7}$$

$$\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}\right)_{ij}=\frac{\partial g(z_{3,i})}{\partial z_{3,j}}=\frac{\partial z_{3,i}}{\partial z_{3,j}}=\begin{cases}1, & \text{if } i=j,\\ 0, & \text{if } i\neq j.\end{cases} \tag{8}$$

Thus,

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}=I\in\mathbb{R}^{K\times K}. \tag{9}$$

$$\left(\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\right)_j=\frac{\partial l_{MSE}(\hat{\boldsymbol{y}},\boldsymbol{y})}{\partial \hat{y}_j}=\frac{\partial\\|\hat{\boldsymbol{y}}-\boldsymbol{y}\\|^2}{\partial \hat{y}_j}=2(\hat{y}_j-y_j), \tag{10}$$

where $\hat{y}_j=\sum_{p=1}^{r} W_{jp}^{(2)}\left(\sum_{q=1}^{n} W_{pq}^{(1)} x_q + b_p^{(1)}\right)^{+}+b_j^{(2)}$. Thus,

$$\frac{\partial l}{\partial \hat{\boldsymbol{y}}}=2(\hat{\boldsymbol{y}}-\boldsymbol{y})^{\top}\in\mathbb{R}^{1\times K}, \tag{11}$$

where $\hat{\boldsymbol{y}}=\boldsymbol{W}^{(2)}\left(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)^{+}+\boldsymbol{b}^{(2)}$.

1.3 Classification

(a) Need to change f, g in the forward pass (b) with logistic sigmoid function σ, shown as below (I don't use the explicit form of σ here because it makes the formula too long) (c) is the same as before, the only change is from using different f, so we

Table 3: forward pass, where $\sigma(z) = (1 + \exp(-z))^{-1}$ (applied element-wisely)

Layer	Input	Output
Linear$_1$	$\boldsymbol{x}$	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$	$\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right)$
Linear$_2$	$\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right)$	$\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}$	$\sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right)$
Loss	$\sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right)$	$\lVert\sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right) - y\rVert^2$

use $\sigma(\cdot)$ to replace $(\cdot)^+$ in table 4, because the NN architecture does not change. (d) also changes for this problem,

Table 4: downward pass

Parameter	Gradient
$\boldsymbol{W}^{(1)}$	$\boldsymbol{x}\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}\boldsymbol{W}^{(2)}\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$
$\boldsymbol{b}^{(1)}$	$\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}\boldsymbol{W}^{(2)}\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$
$\boldsymbol{W}^{(2)}$	$\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right)\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$
$\boldsymbol{b}^{(2)}$	$\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$

$$
\begin{aligned}
\left(\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}\right)_{ij} &= \frac{\partial \sigma(z_{1,i})}{\partial z_{1,j}} = \frac{\partial (1 + \exp(-z_{1,i}))^{-1}}{\partial z_{1,j}} \\
&= \begin{cases} 0, & \text{if } i \neq j, \\ \dfrac{\exp(-z_{1,i})}{(1 + \exp(-z_{1,i}))^2}, & \text{if } i = j, \end{cases}
\end{aligned} \tag{12}
$$

where $z_{1,i} = \sum_{j=1}^{n} W_{ij}^{(1)} x_j + b_i^{(1)}$. Thus,

$$
\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} = \sigma'(\boldsymbol{z}_1) = \operatorname{diag}\left\{\frac{\exp(-\boldsymbol{z}_1)}{(1 + \exp(-\boldsymbol{z}_1))^2}\right\} \in \mathbb{R}^{r \times r}, \tag{13}
$$

where $\boldsymbol{z}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$, the function exp and other operations here use element-wise evaluation. Similarly,

$$
\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}\right)_{ij} = \frac{\partial \sigma(z_{3,i})}{\partial z_{3,j}} = \begin{cases} 0, & \text{if } i \neq j, \\ \dfrac{\exp(-z_{3,i})}{(1 + \exp(-z_{3,i}))^2}, & \text{if } i = j, \end{cases} \tag{14}
$$

where $z_{3,i} = \sum_{p=1}^{r} W_{ip}^{(2)} \sigma\left(\sum_{q=1}^{n} W_{pq}^{(1)} x_q + b_p^{(1)}\right) + b_i^{(2)}$. Thus,

$$
\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} = \sigma'(\boldsymbol{z}_3) = \operatorname{diag}\left\{\frac{\exp(-\boldsymbol{z}_3)}{(1 + \exp(-\boldsymbol{z}_3))^2}\right\} \in \mathbb{R}^{K \times K}, \tag{15}
$$

where $\boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}$. Since the loss function does not change, similar to 1.2(d),

$$\left(\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\right)_j = 2(\hat{y}_j - y_j), \tag{16}$$

where $\hat{y}_j = \sigma\left(\sum_{p=1}^{r} W_{jp}^{(2)}\sigma\left(\sum_{q=1}^{n} W_{pq}^{(1)}x_q + b_p^{(1)}\right) + b_j^{(2)}\right)$. Thus,

$$\frac{\partial l}{\partial \hat{\boldsymbol{y}}} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top \in \mathbb{R}^{1\times K}, \tag{17}$$

where $\hat{\boldsymbol{y}} = \sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right)$.

(b) First of all, beside the changes we obtain from using σ, the loss evaluation step of the forward pass (b) changes, because we use different loss function, The backward

Table 5: forward pass, where $\sigma(z) = (1 + \exp(-z))^{-1}$ (applied element-wisely) and $l_{BCE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$.

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$	$\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right)$
Linear_2	$\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right)$	$\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}$	$\sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right)$
Loss	$\sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right)$	$l_{BCE}(\sigma\left(\boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right), \boldsymbol{y})$

pass (c) does not change, the elements of $\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$ and $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$ in (d) are the same as we discussed for using σ case (see sec 1.3(a) (13) and (15)). The only change is $\frac{\partial l}{\partial \hat{\boldsymbol{y}}}$ since l changes.

$$\begin{aligned}\left(\frac{\partial l}{\partial \hat{\boldsymbol{y}}}\right)_j &= \frac{\partial l_{BCE}(\hat{\boldsymbol{y}}, \boldsymbol{y})}{\partial \hat{y}_j} = \frac{\partial \frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]}{\partial \hat{y}_j} \\ &= -\frac{1}{K}\frac{\partial[y_j \log(\hat{y}_j) + (1 - y_j)\log(1 - \hat{y}_j)]}{\partial \hat{y}_j} = -\frac{1}{K}\left[\frac{y_j}{\hat{y}_j} - \frac{1 - y_j}{1 - \hat{y}_j}\right]\end{aligned} \tag{18}$$

Thus,

$$\frac{\partial l}{\partial \hat{\boldsymbol{y}}} = -\frac{1}{K}\left[\frac{\boldsymbol{y}}{\hat{\boldsymbol{y}}} - \frac{1 - \boldsymbol{y}}{1 - \hat{\boldsymbol{y}}}\right]^\top \in \mathbb{R}^{1\times K}, \tag{19}$$

where the operations are done element-wisely.

(c) For sigmoid function $\sigma(z) = (1 + \exp(-z))^{-1}$, we know its gradient $\sigma'(z) = \frac{\exp(-z)}{(1+exp(-z))^2} = \frac{1}{2+\exp(z)+\exp(-z)}$ becomes close to 0 when $|z|$ is large, which means the elements of $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$ and $\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$ is quite small. Based on table 2, this small Jacobian will leads to small gradient with respect to all parameter, especially for $\boldsymbol{W}^{(1)}$ and $\boldsymbol{b}^{(1)}$ (first few layers), because they have both $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$ and $\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$ inside. So the parameter $\boldsymbol{W}^{(1)}$ and $\boldsymbol{b}^{(1)}$ will not change/be update much through training, which is a waste of the layer. Thus, to make full use of the layer, the activation function $f(z) = (z)^+$ is better, because its gradient is 1 for all positive z, so the elements of $\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$ are $O(1)$, thus this activation function provides enough updates for the parameter in the first few layers, which helps train the neural network.