Two layer net gradients

The neural network that was used for image classification has the following structure:

0.) Input layer (X)

1.) Hidden layer ($W_{(1)}, b_{(1)}$)

2.) Output layer ($W_{(2)}, b_{(2)}$)

First layer (hidden layer):
$z_{(1)} = XW_{(1)} + b_{(1)}$
$a_{(1)} = relu(z_{(1)})$
Second layer (output layer):
$z_{(2)} = a_{(1)}W_{(2)} + b_{(2)}$
$p = softmax(z_{(2)})$

Notation:

- X - the input (X.shape=(n, d), X_i - the ith example with shape=$(1, d)$)
- $W_{(1)}, b_{(1)}$ - the weights and bias arrays for the hidden layer;
- $W_{(2)}, b_{(2)}$ - the weights and bias arrays for the output layer;
- n - number of train examples ($batch\_size$);
- d - number of dimensions of an example;
- c - number of classes (10 classes for CIFAR-10);
- y - labels of correct classes (y_i - the correct label for the ith example);

To quantify the unhappiness with predictions on the training set, cross-entropy loss with regularization was used. All the details about this loss function and its gradients are presented below.

$L = \frac{1}{n}\sum_{i}^{n} L_i + \lambda \sum_{l\_nr}^{2}\sum_{j}^{d}\sum_{k}^{c} W^2_{(l\_nr)(j,k)}$ [cross-entropy loss with regularization for a batch of size n]
$L_i = -\ln p_{y_i}$ [cross-entropy loss for example i]
$p_k = \frac{e^{z_{(2)(k)}}}{\sum_j^c e^{z_{(2)(j)}}}$ [softmax]
$z_{(2)(j)} = X_i W_{(2)(j)} + b_{(2)(j)}$ [class score $z_{(2)(j)}$ for example X_i]
$a_{(1)(i)} = relu(z_{(1)(i)}) = maximum(0, z_{(1)(i)})$

Second layer gradients (for $W_{(2)}$ and $b_{(2)}$):

$$\frac{\partial z_{(2)(j)}}{\partial W_{(2)(j)}} = a_{(1)(i)}$$

$$\frac{\partial z_{(2)(j)}}{\partial b_{(2)(j)}} = 1$$

$$\frac{\partial z_{(2)(j)}}{\partial a_{(1)(i)}} = W_{(2)(j)}\ [a_{(1)(i)} \text{ is a row vector that represents the activations of the } i\text{th example (1st layer)}]$$

$$\frac{\partial p_k}{\partial z_{(2)(k)}} = \frac{e^{z_{(2)(k)}} \sum_j^c e^{z_{(2)(j)}} - e^{z_{(2)(k)}} e^{z_{(2)(k)}}}{(\sum_j^c e^{z_{(2)(j)}})^2} = \frac{e^{z_{(2)(k)}}}{\sum_j^c e^{z_{(2)(j)}}} - \frac{e^{z_{(2)(k)}}}{\sum_j^c e^{z_{(2)(j)}}} \frac{e^{z_{(2)(k)}}}{\sum_j^c e^{z_{(2)(j)}}} = p_k - p_k^2 = p_k(1 - p_k)$$

$$\frac{\partial p_k}{\partial z_{(2)(l)}} = -\frac{e^{z_{(2)(k)}} e^{z_{(2)(l)}}}{(\sum_j^c e^{z_{(2)(j)}})^2} = -\frac{e^{z_{(2)(k)}}}{\sum_j^c e^{z_{(2)(j)}}} \frac{e^{z_{(2)(l)}}}{\sum_j^c e^{z_{(2)(j)}}} = -p_k p_l [l \neq k]$$

$$\frac{\partial L_i}{\partial p_{y_i}} = -\frac{1}{p_{y_i}}$$

$$\frac{\partial L_i}{\partial W_{(2)(y_i)}} = \frac{\partial L_i}{\partial p_{y_i}} \frac{\partial p_{y_i}}{\partial z_{(2)(y_i)}} \frac{\partial z_{(2)(y_i)}}{\partial W_{(2)(y_i)}} = -\frac{1}{p_{y_i}} p_{y_i}(1 - p_{y_i}) a_{(1)(i)} = (p_{y_i} - 1) a_{(1)(i)} [y_i]$$

$$\frac{\partial L_i}{\partial W_{(2)(j)}} = \frac{\partial L_i}{\partial p_{y_i}} \frac{\partial p_{y_i}}{\partial z_{(2)(j)}} \frac{\partial z_{(2)(j)}}{\partial W_{(2)(j)}} = -\frac{1}{p_{y_i}} (-p_{y_i} p_j) a_{(1)(i)} = p_j a_{(1)(i)} [j \neq y_i]$$

$$\frac{\partial L_i}{\partial W_{(2)(j)}} = \frac{\partial L_i}{\partial p_{y_i}} \frac{\partial p_{y_i}}{\partial z_{(2)(j)}} \frac{\partial z_{(2)(j)}}{\partial W_{(2)(j)}} = \begin{cases} (p_{y_i} - 1) a_{(1)(i)} & \text{if } j = y_i \\ p_j a_{(1)(i)} & \text{if } j \neq y_i \end{cases}$$

$$\frac{\partial L_i}{\partial b_{(2)(j)}} = \frac{\partial L_i}{\partial p_{y_i}} \frac{\partial p_{y_i}}{\partial z_{(2)(j)}} \frac{\partial z_{(2)(j)}}{\partial b_{(2)(j)}} = \begin{cases} (p_{y_i} - 1) & \text{if } j = y_i \\ p_j & \text{if } j \neq y_i \end{cases}$$

First layer gradients (for $W_{(1)}$ and $b_{(1)}$):

$$\frac{\partial z_{(1)(j)}}{\partial W_{(1)(j)}} = X_i$$

$$\frac{\partial z_{(1)(j)}}{\partial b_{(1)(j)}} = 1$$

$$\frac{\partial a_{(1)(i)}}{\partial z_{(1)(j)}} = \frac{\partial relu(z_{(1)(j)})}{\partial z_{(1)(j)}} = \frac{\partial max(0, z_{(1)(j)})}{\partial z_{(1)(j)}} = \begin{cases} 1 & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases}$$

$$\frac{\partial a_{(1)(i)}}{\partial W_{(1)(j)}} = \frac{\partial a_{(1)(i)}}{\partial z_{(1)(j)}} \frac{\partial z_{(1)(j)}}{\partial W_{(1)(j)}} = \begin{cases} X_i & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases}$$

$$\frac{\partial a_{(1)(i)}}{\partial b_{(1)(j)}} = \frac{\partial a_{(1)(i)}}{\partial z_{(1)(j)}} \frac{\partial z_{(1)(j)}}{\partial b_{(1)(j)}} = \begin{cases} 1 & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases}$$

$$\frac{\partial L_i}{\partial W_{(1)(j)}} = \frac{\partial L_i}{\partial p_{y_i}} \frac{\partial p_{y_i}}{\partial z_{(2)(j)}} \frac{\partial z_{(2)(j)}}{\partial a_{(1)(i)}} \frac{\partial a_{(1)(i)}}{\partial z_{(1)(j)}} \frac{\partial z_{(1)(j)}}{\partial W_{(1)(j)}} = \begin{cases} j = y_i : \begin{cases} X_i^\top (p_{y_i} - 1) W_{(2)(y_i)}^\top & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases} \\ j \neq y_i : \begin{cases} X_i^\top p_j W_{(2)(y_i)}^\top & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases} \end{cases}$$

$$= \begin{cases} z_{(1)(j)} > 0 : \begin{cases} X_i^\top (p_{y_i} - 1) W_{(2)(y_i)}^\top & \text{if } j = y_i \\ X_i^\top p_j W_{(2)(y_i)}^\top & \text{if } j \neq y_i \end{cases} \\ z_{(1)(j)} \leq 0 : 0 \end{cases}$$

$$\frac{\partial L_i}{\partial b_{(1)(j)}} = \frac{\partial L_i}{\partial p_{y_i}} \frac{\partial p_{y_i}}{\partial z_{(2)(j)}} \frac{\partial z_{(2)(j)}}{\partial a_{(1)(i)}} \frac{\partial a_{(1)(i)}}{\partial z_{(1)(j)}} \frac{\partial z_{(1)(j)}}{\partial b_{(1)(j)}} = \begin{cases} j = y_i : \begin{cases} (p_{y_i} - 1) W_{(2)(y_i)}^\top & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases} \\ j \neq y_i : \begin{cases} p_j W_{(2)(y_i)}^\top & \text{if } z_{(1)(j)} > 0 \\ 0 & \text{if } z_{(1)(j)} \leq 0 \end{cases} \end{cases}$$

$$= \begin{cases} z_{(1)(j)} > 0 : \begin{cases} (p_{y_i} - 1) W_{(2)(y_i)}^\top & \text{if } j = y_i \\ p_j W_{(2)(y_i)}^\top & \text{if } j \neq y_i \end{cases} \\ z_{(1)(j)} \leq 0 : 0 \end{cases}$$

Gradients from the regularization loss (for $W_{(1)}, W_{(2)}$):

$$Reg\_loss = \lambda \sum_{l\_nr}^{2} \sum_{j}^{d} \sum_{k}^{c} W^2_{(l\_nr)(j,k)}$$

$$\frac{\partial Reg\_loss}{\partial W_{(l\_nr)(j,k)}} = 2\lambda W_{(l\_nr)(j,k)}$$

Previous gradients were calculated for the example X_i that is a row vector with d dimensions. The gradients for a batch of size n can be calculated as the mean gradients from the data loss plus the gradients from the regularization loss.

$$L = \frac{1}{n} \sum_{i}^{n} L_i + \lambda \sum_{l\_nr}^{2} \sum_{j}^{d} \sum_{k}^{c} W^2_{(l\_nr)(j,k)}$$

$$\frac{\partial L}{\partial W_{(l\_nr)(j)}} = \frac{1}{n} \sum_{i}^{n} \frac{\partial L_i}{\partial W_{(l\_nr)(j)}} + \frac{\partial Reg\_loss}{\partial W_{(l\_nr)(j,k)}}$$

$$\frac{\partial L}{\partial b_{(l\_nr)(j)}} = \frac{1}{n} \sum_{i}^{n} \frac{\partial L_i}{\partial b_{(l\_nr)(j)}}$$