Cross-entropy loss with regularization and its gradients

Notation:
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);

$$L = \frac{1}{n}\sum_{i}^{n} L_i + \lambda * \sum_{j}^{d}\sum_{k}^{c} W_{j,k}^2$$

$$L_i = -\ln p_{y_i} [cross - entropy]$$

$$p_k = \frac{e^{z_{y_k}}}{\sum_j^c e^{z_j}} [softmax]$$

$$z_j = X * W_j [score - for - example - X - to - belong - to - class - j]$$

$$\frac{\partial L_i}{\partial p_i} = -\frac{1}{p_i}$$

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

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

$$\frac{\partial z_j}{\partial w_j} = X$$

$$\frac{\partial L_i}{\partial w_{y_i}} = \frac{\partial L_i}{\partial p_i} \frac{\partial p_i}{\partial z_{y_i}} \frac{\partial z_{y_i}}{\partial w_{y_i}} = -\frac{1}{p_i} p_i(1 - p_i) X = (p_i - 1) X [y_i, chain - rule]$$

$$\frac{\partial L_i}{\partial w_j} = \frac{\partial L_i}{\partial p_i} \frac{\partial p_i}{\partial z_j} \frac{\partial z_j}{\partial w_j} = -\frac{1}{p_i}(-p_i p_j) X = p_j X [j \neq y_i, chain - rule]$$