Gradient of Loss w.r.t Output Layer Weights

Let $\mathbf{y}, \hat{\mathbf{y}} \in \mathbb{R}^{m \times 1}$, be true and predicted values for m examples.
Further

$$\hat{\mathbf{y}} = \boldsymbol{a} * W_{out} + \boldsymbol{b}^T \tag{1}$$

where $W_{out} \in \mathbb{R}^{u \times 1}$, $\boldsymbol{b} \in \mathbb{R}^{1 \times 1}$ are weights and biases of the output layer. $\boldsymbol{a} \in \mathbb{R}^{m \times u}$ be the activation at the last hidden layer.

MSE loss can be written as

$$\boldsymbol{L} = \frac{1}{m}(\mathbf{y} - \hat{\mathbf{y}})^2 \tag{2}$$

$$\boldsymbol{L} = \frac{1}{m}(\mathbf{y} - (\boldsymbol{a}W_{out} + \boldsymbol{b}^T))^2 \tag{3}$$

Note that $\boldsymbol{L} \in \mathbb{R}^{m \times 1}$. To write gradient of $\boldsymbol{L}$ with respect to W_{out}, let us consider a simple case in which we compute gradient of i-th entry of $\boldsymbol{L}$ (i.e. i-th example) with respect to W_{out}.

Now, for i-th example, let y be true value and $\hat{y}$ be predicted value. Note that both y and $\hat{y}$ are scalars. The loss is (ignore $\frac{1}{m}$ for the moment)

$$l = (y - \hat{y})^2 \tag{4}$$

$$l = (y - (\boldsymbol{a}W_{out} + \boldsymbol{b}^T))^2 \tag{5}$$

And gradient can be written as

$$\frac{\partial l}{\partial W_{out}} = 2(y - \boldsymbol{a}W_{out} - \boldsymbol{b}^T)\frac{\partial(y - \boldsymbol{a}W_{out} - \boldsymbol{b}^T)}{\partial W_{out}} \tag{6}$$

$$= 2(y - \boldsymbol{a}W_{out} - \boldsymbol{b}^T) * (-\boldsymbol{a}) \tag{7}$$

$$= 2(\hat{y} - y) * \boldsymbol{a} \tag{8}$$

To compute the gradient of all examples at the same time, suppose $m = 5$ and $u = 4$. Then we can write the vector W_{out} as

$$W_{out} = [w_1, w_2, w_3, w_4]^T \tag{9}$$

And we can write the expression $\frac{\partial l}{\partial W_{out}}$ as

$$\frac{\partial l}{\partial W_{out}} = \begin{bmatrix} \frac{\partial l}{\partial w_1} \\ \frac{\partial l}{\partial w_2} \\ \frac{\partial l}{\partial w_3} \\ \frac{\partial l}{\partial w_4} \end{bmatrix} = 2(\hat{y} - y) \begin{bmatrix} a_{i,1} \\ a_{i,2} \\ a_{i,3} \\ a_{i,4} \end{bmatrix} = 2(\hat{y} - y)\boldsymbol{a} \tag{10}$$

Hence the gradient for all m examples can be obtained by stacking $\frac{\partial l}{\partial W_{out}}$ for all examples:

$$\frac{\partial \boldsymbol{L}}{\partial W_{out}} = \begin{bmatrix} \frac{\partial \boldsymbol{L}_1}{\partial W_{out}} \\ \frac{\partial \boldsymbol{L}_2}{\partial W_{out}} \\ \frac{\partial \boldsymbol{L}_3}{\partial W_{out}} \\ \frac{\partial \boldsymbol{L}_4}{\partial W_{out}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \boldsymbol{L}_1}{\partial w_1} & \frac{\partial \boldsymbol{L}_1}{\partial w_2} & \frac{\partial \boldsymbol{L}_1}{\partial w_3} & \frac{\partial \boldsymbol{L}_1}{\partial w_4} \\ \frac{\partial \boldsymbol{L}_2}{\partial w_1} & \frac{\partial \boldsymbol{L}_2}{\partial w_2} & \frac{\partial \boldsymbol{L}_2}{\partial w_3} & \frac{\partial \boldsymbol{L}_2}{\partial w_4} \\ \frac{\partial \boldsymbol{L}_3}{\partial w_1} & \frac{\partial \boldsymbol{L}_3}{\partial w_2} & \frac{\partial \boldsymbol{L}_3}{\partial w_3} & \frac{\partial \boldsymbol{L}_3}{\partial w_4} \\ \frac{\partial \boldsymbol{L}_4}{\partial w_1} & \frac{\partial \boldsymbol{L}_4}{\partial w_2} & \frac{\partial \boldsymbol{L}_4}{\partial w_3} & \frac{\partial \boldsymbol{L}_4}{\partial w_4} \\ \frac{\partial \boldsymbol{L}_5}{\partial w_1} & \frac{\partial \boldsymbol{L}_5}{\partial w_2} & \frac{\partial \boldsymbol{L}_5}{\partial w_3} & \frac{\partial \boldsymbol{L}_5}{\partial w_4} \end{bmatrix}_{5\times 4} = 2(\hat{\mathbf{y}} - \mathbf{y}) * \boldsymbol{a} \tag{11}$$

Note that we have compactly written the expression as $2(\hat{\mathbf{y}} - \mathbf{y}) * \boldsymbol{a}$. Also, the multiplication $(\hat{\mathbf{y}} - \mathbf{y}) * \boldsymbol{a}$ is supported by numpy-broadcast.

Finally, take average of gradient vectors of all examples to get the gradient that you can use to update the weights of the output layer as follows:

$$W_{out} = W_{out} - \frac{1}{m} * \eta * \mathrm{Sum}\left(\frac{\partial \boldsymbol{L}}{\partial W_{out}}, \text{axis=0}\right) \tag{12}$$

where axis=0 denotes sum along rows.