1. Show $\nabla_{h^{(K)}} L_y(\hat{y}) = \nabla_{\hat{y}} L_y(\hat{y})$

   Since $h^{(K)}$ represents the output of the last layer K of the network, it is also the prediction ($\hat{y}$) of the model. Therefore, $h^{(K)} = \hat{y}$, meaning the gradient w.r.t. one equals the gradient w.r.t. to the other.

2. Show $\nabla_{a^{(K)}} L_y(\hat{y}) = g'(a^{(K)})^\top \odot \nabla_{\hat{y}} L_y(\hat{y})$

   Since g is an element-wise function, using the chain rule results in the following element-wise product

$$\nabla_{a^{(K)}} L_y(\hat{y}) = \nabla_{\hat{y}} L_y(\hat{y}) \odot \frac{dh^{(K)}}{da^{(K)}}$$

   where

$$\nabla_{\hat{y}} L_y(\hat{y}), \frac{dh^{(K)}}{da^{(K)}} \in \mathbb{R}^{1\times n}$$

   to keep consistent with our definition of gradient dimensions. We know that

$$\frac{dh^{(K)}}{da^{(K)}} = g'(a^{(K)})$$

   but since $h^{(K)} \in \mathbb{R}^n$, the derivative $g'(a^{(K)})$ will yield the same dimensions. The element-wise multiplication can not occur unless the quantity is in $\mathbb{R}^{1\times n}$. Therefore we transpose it, and due to the commutative property of element-wise multiplication, we can bring the transposed quantity out front to get

$$\nabla_{a^{(K)}} L_y(\hat{y}) = g'(a^{(K)})^\top \odot \nabla_{\hat{y}} L_y(\hat{y})$$

3. Show $\nabla_{W^{(K)}} L_y(\hat{y}) = h^{(K-1)}(\nabla_{a^{(K)}} L_y(\hat{y}))$

   Again with the chain rule, we know

$$\nabla_{W^{(K)}} L_y(\hat{y}) = \nabla_{a^{(K)}} L_y(\hat{y}) \frac{da^{(K)}}{dW^{(K)}}$$

   where

$$a^{(K)} = W^{(K)} h^{(K-1)} + b^{(K)}$$

   so

$$\frac{da^{(K)}}{dW^{(K)}} = \frac{d}{dW^{(K)}} \left[W^{(K)} h^{(K-1)} + b^{(K)} \right]$$

   The derivative of $a^{(K)}$ w.r.t. matrix $W^{(K)}$ is easier to compute using vectorization, where $vec : \mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$. An example definition is shown below

$$vec(ABC) = \left(C^\top \otimes A\right) vec(B)$$

   where $\otimes$ is the Kronecker product operator. In addition, the bias is dropped since its derivative w.r.t. $W^{(K)}$ is 0. So

$$W^{(K)} h^{(K-1)} = \left((h^{(K-1)})^\top \otimes I \right) vec(W^{(K)})$$

$$\frac{da^{(K)}}{dw^{(K)}} = (h^{(K-1)})^\top \otimes I$$

Therefore, the vectorized gradient equals

$$\nabla_{w^{(K)}} L_y(\hat{y}) = \nabla_{a^{(K)}} L_y(\hat{y}) \left((h^{(K-1)})^\top \otimes I \right)$$

Transpose the result to make things easier down the line

$$\nabla_{w^{(K)}} L_y(\hat{y})^\top = \left(h^{(K-1)} \otimes I \right) \nabla_{a^{(K)}} L_y(\hat{y})^\top$$

To get back to our gradient from our vectorized gradient, we use inverse vectorization. Inverse vectorization is defined as $vec^{-1} : \mathbb{R}^{nm} \to \mathbb{R}^{n \times m}$, meaning

$$\nabla_{W^{(K)}} L_y(\hat{y}) = vec^{-1} \left(\nabla_{w^{(K)}} L_y(\hat{y}) \right)$$

so

$$\nabla_{W^{(K)}} L_y(\hat{y}) = vec^{-1} \left(\left(h^{(K-1)} \otimes I \right) \nabla_{a^{(K)}} L_y(\hat{y})^\top \right)^\top$$

We know that $\nabla_{a^{(K)}} L_y(\hat{y}) \in \mathbb{R}^{1 \times n}$, so $vec\left(\nabla_{a^{(K)}} L_y(\hat{y})^\top\right) = \nabla_{a^{(K)}} L_y(\hat{y})^\top$. We can rewrite what we have above

$$= vec^{-1} \left(\left(h^{(K-1)} \otimes I \right) vec\left(\nabla_{a^{(K)}} L_y(\hat{y})^\top \right) \right)^\top$$

$$= vec^{-1} \left(vec\left(\nabla_{a^{(K)}} L_y(\hat{y})^\top (h^{(K-1)})^\top \right) \right)^\top$$

$$= \left(\nabla_{a^{(K)}} L_y(\hat{y})^\top (h^{(K-1)})^\top \right)^\top$$

So that

$$\nabla_{W^{(K)}} L_y(\hat{y}) = h^{(K-1)} \nabla_{a^{(K)}} L_y(\hat{y})$$

4. Show $\nabla_{h^{(K-1)}} L_y(\hat{y}) = (\nabla_{a^{(K)}} L_y(\hat{y})) W^{(K)}$

   Similar to 3.,

$$\nabla_{h^{(K-1)}} L_y(\hat{y}) = \nabla_{a^{(K)}} L_y(\hat{y}) \frac{da^{(K)}}{dh^{(K-1)}}$$

   where

$$a^{(K)} = W^{(K)} h^{(K-1)} + b^{(K)}$$

   so

$$\frac{da^{(K)}}{dh^{(K-1)}} = \frac{d}{dh^{(K-1)}} \left[W^{(K)} h^{(K-1)} + b^{(K)} \right]$$

   The derivative of $a^{(K)}$ w.r.t. vector $h^{(K-1)}$ is simply the matrix $W^{(K)}$, meaning

$$\frac{da^{(K)}}{dh^{(K-1)}} = W^{(K)}$$

   So that

$$\nabla_{h^{(K-1)}} L_y(\hat{y}) = \nabla_{a^{(K)}} L_y(\hat{y}) W^{(K)}$$