Problem 1

$\nabla_{h^{(K)}} L_y(\hat{y}) = \nabla_{\hat{y}} L_y(\hat{y})$ because the last layer of the network is $h^{(K)}$ which serves as the final hidden layer and the output layer. So the final layer can be referred to as $h^{(K)}$ or $\hat{y}$, which means the expressions on either side of the equation are equivalent.

Problem 2

$$\nabla_{a^{(k)}} L_y(\hat{y}) = \frac{\partial L}{\partial g} \frac{\partial g}{\partial a^{(k)}} = \nabla_g L_y(\hat{y}) \times g'(a^{(k)}) \times 1$$

by the multivariable chain rule. Also, $g(a^{(k)}) = \hat{y}$ from slide 6 of the notes. Further, $a^{(k)}$ is a vector so we can use the gradient when differentiating it. So we can write the previous equation as

$$\nabla_{\hat{y}} L_y(\hat{y}) \times g'(a^{(k)})$$

Since g is applied element wise, the multiplication is not regular vector multiplication so we have

$$\nabla_{\hat{y}} L_y(\hat{y}) \odot g'(a^{(k)})$$

Problem 3

$$\nabla_{W^{(k)T}} L_y(\hat{y}) = \frac{\partial L}{\partial a^{(k)}} \frac{a^{(k)}}{\partial W^{(k)}}$$

so

$$\nabla_{W^{(k)}} L_y(\hat{y}) = (\frac{\partial L}{\partial a^{(k)}} \frac{a^{(k)}}{\partial W^{(k)}})^T = \frac{a^{(k)T}}{\partial W^{(k)}} \frac{\partial L^T}{\partial a^{(k)}}$$

The transpose of $a^{(k)T} = b^{(k)} + h^{(k-1)T} W^{(k)}$. And the partial with respect to W there is $h^{(k-1)}$. Now taking the gradient of L^T with respect to $a^{(k)}$ and multiplying by the partial of a with respect to W we get

$$h^{(k-1)} (\nabla_{a^{(k)}} L_y(\hat{y}))^T$$

Problem 4

$$\nabla_{h^{(k-1)T}} L_y(\hat{y}) = \frac{\partial L}{\partial a^{(k)}} \frac{a^{(k)}}{\partial h^{(k-1)}}$$

so

$$\nabla_{h^{(k-1)T}} L_y(\hat{y}) = (\frac{\partial L}{\partial a^{(k)}} \frac{a^{(k)}}{\partial h^{(k-1)}})^T = \frac{a^{(k)T}}{\partial h^{(k-1)T}} \frac{\partial L^T}{\partial a^{(k)}}$$

The transpose of $a^{(k)T} = b^{(k)} + h^{(k-1)T} W^{(k)}$. And the partial with respect to h^T there is $W^{(k)}$. Now taking the gradient of L^T with respect to $a^{(k)}$ and multiplying by the partial of a with respect to h^T we get

$$W^{(k)} (\nabla_{a^{(k)}} L_y(\hat{y}))^T$$