1.2 Regression Task

(a)

1. Forward pass (run the model and the get predicted outputs)
2. Compute the loss
3. Zero out the gradients
4. Run backward pass to compute gradient of the loss w.r.t. learnable parameters
5. Update parameters by performing a step of the optimizer

(b)

Layer	Input	Output
$Linear_1$	$\mathbf{x}$	$\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)}$
f	$\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)}$	$(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}$
$Linear_2$	$(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}$	$\mathbf{W}^{(2)}(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}+\mathbf{b}^{(2)}$
g	$\mathbf{W}^{(2)}(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}+\mathbf{b}^{(2)}$	$\mathbf{W}^{(2)}(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}+\mathbf{b}^{(2)}$
Loss	$(\mathbf{W}^{(2)}(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}+\mathbf{b}^{(2)},\mathbf{y})$	$\Vert\mathbf{W}^{(2)}(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})^{+}+\mathbf{b}^{(2)}-\mathbf{y}\Vert^{2}$

(c)

Parameter	Gradient
$\mathbf{W}^{(1)}$	$\mathbf{x}\frac{\partial l}{\partial \mathbf{z_3}}\mathbf{W}^{(2)}\frac{\partial \mathbf{z_2}}{\partial \mathbf{z_1}}$
$\mathbf{b}^{(1)}$	$\frac{\partial l}{\partial \mathbf{z_3}}\mathbf{W}^{(2)}\frac{\partial \mathbf{z_2}}{\partial \mathbf{z_1}}$
$\mathbf{W}^{(2)}$	$\mathbf{z_2}\frac{\partial l}{\partial \hat{\mathbf{y}}}\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z_3}}$
$\mathbf{b}^{(2)}$	$\frac{\partial l}{\partial \hat{\mathbf{y}}}\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z_3}}$

(d)

1. $$\left(\frac{\partial \mathbf{z_2}}{\partial \mathbf{z_1}}\right)_{ij}=\begin{cases}1, & \text{if } \mathbf{z_1}_{ij}>0\\ 0, & \text{otherwise.}\end{cases}$$
2. $$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z_3}}=\mathbf{I}.$$

   Where $\mathbf{I}$ is the $\mathrm{K}\times\mathrm{K}$ identity matrix.
3. $$\frac{\partial l}{\partial \hat{\mathbf{y}}}=2(\hat{\mathbf{y}}-\mathbf{y})^{\top}.$$

1.3 Classification Task

(a)

1. (b)

Layer	Input	Output
$Linear_1$	$\mathbf{x}$	$\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)}$
f	$\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)}$	$\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})$
$Linear_2$	$\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})$	$\mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)}$
g	$\mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)}$	$\sigma(\mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)})$
Loss	$\sigma(\mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)}),\mathbf{y})$	$\lVert\sigma(\mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)})-\mathbf{y}\rVert^2$

2. (c) - Nothing changes
3. (d)

$$\left(\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z_3}}\right)_{ij} = \begin{cases} \hat{y}_i(1-\hat{y}_i), & \text{if } i=j \\ 0, & \text{otherwise.} \end{cases}$$

$$\left(\frac{\partial \mathbf{z_2}}{\partial \mathbf{z_1}}\right)_{ij} = \begin{cases} z_{2i}(1-z_{2i}), & \text{if } i=j \\ 0, & \text{otherwise.} \end{cases}$$

(b)

1. (b) - No changes
2. (c) - No changes
3. (d)

$$\left(\frac{\partial l}{\partial \hat{\mathbf{y}}}\right)_{ij} = \frac{-1}{K}\left(\frac{y_i}{\hat{y}_i} - \frac{(1-y_i)}{(1-\hat{y}_i)}\right)\delta_{ij}.$$

where δ_{ij} is the Kronecker delta function.

(c) The sigmoid function is bounded from both above and below. If the network learns parameters that saturate the sigmoids in the network their gradient propagated back will be close to zero. This, in very deep networks, results in a vanishing gradient and the network stops learning.

The reasoning behind the vanishing gradient is due to the fact that gradients are multiplied as they are backpropagated which compounds the effect and quickly turns small values to zero.