Recall that batch normalization layers are used as follows:\(^1\):

\[
\begin{align*}
a^\ell &= \text{BN}_{\gamma^\ell, \beta^\ell}(W^\ell o_{\ell-1}^\ell) = \gamma^\ell \circ \left( \frac{W^\ell o_{\ell-1}^\ell - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \beta^\ell \\
&= \gamma^\ell \circ o_{\ell-1}^\ell + \beta^\ell, \\
o^\ell &= \sigma(a^\ell).
\end{align*}
\]

For convenience, we have denoted the affinely transformed output of layer \(\ell - 1\) by

\[
o_{\ell-1}^\ell = W^\ell o_{\ell-1}^\ell - \mu_B \sqrt{\sigma_B^2 + \epsilon} + \beta^\ell.
\]

In the above expressions, \(\epsilon\) is a small positive number to guard against division by zero, \(\circ\) denotes element-wise multiplication, and \(\mu_B, \sigma_B^2 \in \mathbb{R}^{n_{\ell-1}}\) are the vectors of minibatch means and minibatch variances for each neuron in layer \(\ell - 1\):

\[
\begin{align*}
\mu_B &= \frac{1}{m} \sum_{i=1}^{m} (o_{\ell-1}^\ell)_i, \\
\sigma_B &= \frac{1}{m} \sum_{i=1}^{m} ((o_{\ell-1}^\ell)_i - \mu_B)^2.
\end{align*}
\]

The scale and shift vectors \(\gamma^\ell, \beta^\ell \in \mathbb{R}^{n_{\ell-1}}\) are parameters that must be learned during training, using backpropagation. Let \(f\) be the training objective.

\(1\) Verify that \(J_{a^\ell}(a^\ell) = \text{diag}(\sigma'(a^\ell))\), and use this fact to give an expression for \(\nabla_{a^\ell} f\), assuming that \(\nabla_{a^\ell} f\) is known.

\(2\) Verify that \(J_{a^\ell}(\gamma) = \text{diag}(o_{\ell-1}^\ell)\), and use this fact to give an expression for \(\nabla_{\gamma^\ell} f\) in terms of \(\nabla_{a^\ell} f\).

\(3\) Verify that \(J_{a^\ell}(\beta^\ell) = I\), and use this fact to give an expression for \(\nabla_{\beta^\ell} f\) in terms of \(\nabla_{a^\ell} f\).

\(4\) Observe (no need to verify this) that for any \(i \in [n_{\ell}]\),

\[
\begin{align*}
[J_{a^\ell}(W^\ell)]_{i,:} &= \left[ \frac{\partial(a^\ell)_i}{\partial(W^\ell)_{p,q}} \right]_{p,q=1}^{n_{\ell}, n_{\ell-1}} \\
&= \text{diag} \left( \frac{\gamma^\ell}{\sqrt{\sigma_B^2 + \epsilon}} \right) e_i (o_{\ell-1}^\ell)^T, \\
\text{so for any vector } v \in \mathbb{R}^{n_{\ell}}, \\
J_{a^\ell}(W^\ell)^T v &= \text{diag} \left( \frac{\gamma^\ell}{\sqrt{\sigma_B^2 + \epsilon}} \right) v (o_{\ell-1}^\ell)^T.
\end{align*}
\]

Use this fact to give an expression for \(\nabla_{W^\ell} f\) in terms of \(\nabla_{a^\ell} f\).

\(^1\) Notice that we do not include a bias \(b^\ell\) in the call to the BN primitive because \(\beta^\ell\) is our bias.