WEEKLY PARTICIPATION 9

Recall that batch normalization layers are used as follows¹:

$$egin{aligned} oldsymbol{a}^\ell &= \mathrm{BN}_{oldsymbol{\gamma}^\ell,oldsymbol{eta}^\ell}(oldsymbol{W}^\elloldsymbol{o}^{\ell-1}) = oldsymbol{\gamma}^\ell\odot\left(rac{oldsymbol{W}^\elloldsymbol{o}^{\ell-1} - oldsymbol{\mu}_B}{\sqrt{oldsymbol{\sigma}_B^2 + \epsilon}}
ight) + oldsymbol{eta}^\ell \ &= oldsymbol{\gamma}^\ell\odotoldsymbol{\overline{o}}^{\ell-1} + oldsymbol{eta}^\ell, \ oldsymbol{o}^\ell &= \sigma(oldsymbol{a}^\ell). \end{aligned}$$

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

$$\overline{m{o}^{\ell-1}} = rac{m{W}^{\ell}m{o}^{\ell-1} - m{\mu}_B}{\sqrt{m{\sigma}_B^2 + \epsilon}}.$$

In the above expressions, ϵ is a small positive number to guard against division by zero, \odot 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 $\ell-1$:

$$\mu_B = \frac{1}{m} \sum_{i=1}^m (o^{\ell-1})_i, \qquad \sigma_B = \frac{1}{m} \sum_{i=1}^m ((o^{\ell-1})_i - \mu_B)^2.$$

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 $J_{o^{\ell}}(a^{\ell}) = \operatorname{diag}(\sigma'(a^{\ell}))$, and use this fact to give an expression for $\nabla_{a^{\ell}} f$, assuming that $\nabla_{o^{\ell}} f$ is known.
- (2) Verify that $J_{a^{\ell}}(\gamma) = \operatorname{diag}\left(\overline{o^{\ell-1}}\right)$, 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 $i \in [n_{\ell}]$,

$$egin{aligned} \left[oldsymbol{J}_{oldsymbol{a}^\ell}(oldsymbol{W}^\ell)
ight]_{i,:} &= \left[rac{\partial (oldsymbol{a}^\ell)_i}{\partial (oldsymbol{W}^\ell)_{p,q}}
ight]_{p,q=1}^{n_\ell,n_{\ell-1}} \ &= \operatorname{diag}\left(rac{oldsymbol{\gamma}^\ell}{\sqrt{oldsymbol{\sigma}_B^2 + \epsilon}}
ight) oldsymbol{e}_i(oldsymbol{o}^{\ell-1})^T, \end{aligned}$$

so for any vector $\boldsymbol{v} \in \mathbb{R}^{n_{\ell}}$,

$$oldsymbol{J_{a^\ell}}(oldsymbol{W}^\ell)^Toldsymbol{v} = \operatorname{diag}\left(rac{oldsymbol{\gamma}^\ell}{\sqrt{oldsymbol{\sigma}_B^2 + \epsilon}}
ight)oldsymbol{v}(oldsymbol{o}^{\ell-1})^T.$$

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

¹Notice that we do not include a bias b^{ℓ} in the call to the BN primitive because β^{ℓ} is our bias.