FraseR autoencoder gradients

Beta binomial density:

$$
\begin{aligned}
BetaBin(k|n,\alpha,\beta) &= \binom{n}{k} \cdot \frac{Be(\alpha+k,\beta+n-k)}{Be(\alpha,\beta)} \\
&= \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot \frac{\Gamma(\alpha+k)\Gamma(\beta+n-k)}{\Gamma(\alpha+\beta+n)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}
\end{aligned}
$$

Parametrization with μ and ρ:

$$
\begin{aligned}
\mu &= \frac{\alpha}{\alpha+\beta} & \rho &= \frac{1}{1+\alpha+\beta} \\
\Rightarrow \alpha &= \mu\left(\frac{1-\rho}{\rho}\right) & \beta &= (\mu-1)\left(\frac{\rho-1}{\rho}\right)
\end{aligned}
$$

$\alpha+\beta$ is only dependent on ρ:

$$
\begin{aligned}
\alpha+\beta &= \mu(\frac{1}{\rho}-1) + \frac{(\mu-1)(\rho-1)}{\rho} = \frac{\mu(1-\rho)}{\rho} + \frac{-(1-\mu)-(1-\rho)}{\rho} \\
&= \frac{\mu(1-\rho)+(1-\mu)(1-\rho)}{\rho} = \frac{(1-\rho)\left(\mu+(1-\mu)\right)}{\rho} \\
&= \frac{1-\rho}{\rho} = \frac{1}{\rho} - 1
\end{aligned}
$$

Definition of μ_{ij} and x_{ij}:

$$
\begin{aligned}
Y &= XW_eW_d^T + b \\
logit(\mu_{ij}) = y_{ij} &\Rightarrow \mu_{ij} = \frac{e^{y_{ij}}}{1+e^{y_{ij}}} \\
x_{ij} &= logit\left(\frac{k_{ij}+\frac{1}{2}}{n_{ij}+1}\right)
\end{aligned}
$$

Negative log-likelihood:

$$
\begin{aligned}
nll &= -\log\left(\frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot \frac{\Gamma(\alpha+k)\Gamma(\beta+n-k)}{\Gamma(\alpha+\beta+n)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) \\
&= -\log\left(\frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}\right) - \log\left(\frac{\Gamma(\alpha+k)\Gamma(\beta+n-k)}{\Gamma(\alpha+\beta+n)}\right) - \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) \\
&= -\log(\Gamma(n+1)) + log(\Gamma(k+1)) + \log(\Gamma(n-k+1)) - \log(\Gamma(\alpha+k)) - \log(\Gamma(\beta+n-k)) \\
&\quad + \log(\Gamma(\alpha+\beta+n)) + \log(\Gamma(\alpha)) + \log(\Gamma(\beta)) - \log(\Gamma(\alpha+\beta))
\end{aligned}
$$

$$
nll_{\mathbf{W}} = \log(\Gamma(\alpha)) + \log(\Gamma(\beta)) - \log(\Gamma(\alpha+k)) - \log(\Gamma(\beta+n-k))
$$

NLL with pseudocounts ($k+1$, $n+2$):

$$
nll_{\mathbf{W}} = \log(\Gamma(\alpha)) + \log(\Gamma(\beta)) - \log(\Gamma(\alpha+k+1)) - \log(\Gamma(\beta+n-k+1))
$$

Full form of $\log(\Gamma(\alpha))$:

$$\log(\Gamma(\alpha)) = \log(\Gamma(\mu\frac{1-\rho}{\rho})) = \log\left(\Gamma\left(\frac{e^Y}{1+e^Y}\frac{1-\rho}{\rho}\right)\right)$$
$$= \log\left(\Gamma\left(\left(\frac{e^{XW_eW_d^T+b}}{1+e^{XW_eW_d^T+b}}\right)\left(\frac{1-\rho}{\rho}\right)\right)\right)$$

Derivatives:

$$\frac{d}{d\alpha}\left[\log(\Gamma(\alpha))\right] = \psi^{(0)}(\alpha)$$
$$\frac{d}{d\mu}\left[\mu\left(\frac{1-\rho}{\rho}\right)\right] = \frac{1-\rho}{\rho}$$
$$\frac{d}{dY}\left[\frac{e^Y}{1+e^Y}\right] = \frac{e^Y}{(1+e^Y)^2}$$
$$\frac{d}{dW_d}\left[XW_eW_d^T+b\right] = XW_e$$
$$\frac{d}{dW_e}\left[XW_eW_d^T+b\right] = X^TW_d$$
$$\frac{d}{db}\left[XW_eW_d^T+b\right] = 1$$

Derivatives of the first term of the negative log-likelihood:

$$\frac{d}{dW_d}\log(\Gamma(\alpha)) = \frac{d}{dW_d}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)\right)$$
$$= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)\cdot\frac{1-\rho}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T XW_e$$

$$\frac{d}{dW_e}\log(\Gamma(\alpha)) = \frac{d}{dW_e}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)\right)$$
$$= X^T\left(\psi\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)\cdot\frac{1-\rho}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)W_d$$

$$\frac{d}{db}\log(\Gamma(\alpha)) = \frac{d}{db}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)\right)$$
$$= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)\cdot\frac{1-\rho}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T$$

Derivatives of the second term of the negative log-likelihood:

$$\begin{aligned}\frac{d}{dW_d}\log(\Gamma(\beta)) &= \frac{d}{dW_d}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)\right)\right)\\ &= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)\right)\cdot\frac{\rho-1}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T XW_e\end{aligned}$$

$$\begin{aligned}\frac{d}{dW_e}\log(\Gamma(\beta)) &= \frac{d}{dW_e}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)\right)\right)\\ &= X^T\left(\psi\left(\left(\frac{e^Y}{1+e^{Y)}}-1\right)\left(\frac{\rho-1}{\rho}\right)\right)\cdot\frac{\rho-1}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)W_d\end{aligned}$$

$$\begin{aligned}\frac{d}{db}\log(\Gamma(\beta)) &= \frac{d}{db}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)\right)\right)\\ &= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)\right)\cdot\frac{\rho-1}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T\end{aligned}$$

Derivatives of the third term of the negative log-likelihood:

$$\begin{aligned}\frac{d}{dW_d}\log(\Gamma(\alpha+k+1)) &= \frac{d}{dW_d}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)+k+1\right)\\ &= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)+k+1\right)\cdot\frac{1-\rho}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T XW_e\end{aligned}$$

$$\begin{aligned}\frac{d}{dW_e}\log(\Gamma(\alpha+k+1)) &= \frac{d}{dW_e}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)+k+1\right)\\ &= X^T\left(\psi\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)+k+1\right)\cdot\frac{1-\rho}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)W_d\end{aligned}$$

$$\begin{aligned}\frac{d}{db}\log(\Gamma(\alpha+k+1)) &= \frac{d}{db}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)\right)+k+1\right)\\ &= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}\right)\left(\frac{1-\rho}{\rho}\right)+k+1\right)\cdot\frac{1-\rho}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T\end{aligned}$$

Derivatives of the fourth term of the negative log-likelihood:

$$\begin{aligned}\frac{d}{dW_d}\log(\Gamma(\beta+n-k+1)) &= \frac{d}{dW_d}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)+n-k+1\right)\right)\\ &= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)+n-k+1\right)\cdot\frac{\rho-1}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T XW_e\end{aligned}$$

$$\begin{aligned}\frac{d}{dW_e}\log(\Gamma(\beta+n-k+1)) &= \frac{d}{dW_e}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)+n-k+1\right)\right)\\ &= X^T\left(\psi\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)+n-k+1\right)\cdot\frac{\rho-1}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)W_d\end{aligned}$$

$$\begin{aligned}\frac{d}{db}\log(\Gamma(\beta+n-k+1)) &= \frac{d}{db}\log\left(\Gamma\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)+n-k+1\right)\right)\\ &= \left(\psi\left(\left(\frac{e^Y}{1+e^Y}-1\right)\left(\frac{\rho-1}{\rho}\right)+n-k+1\right)\cdot\frac{\rho-1}{\rho}\cdot\frac{e^Y}{(1+e^Y)^2}\right)^T\end{aligned}$$