Identities

$$(A\otimes C)(B\otimes D) = (AB)\otimes(CD) \qquad \mathrm{diag}(\boldsymbol{u})\mathbf{1} = \boldsymbol{u}$$

$$\frac{\partial}{\partial \boldsymbol{u}} f^T(\boldsymbol{u}) = \mathrm{diag}[f'(\boldsymbol{u})] \qquad \frac{\partial}{\partial \boldsymbol{v}} f^T(\boldsymbol{u}) = \frac{\partial \boldsymbol{u}^T}{\partial \boldsymbol{v}}\mathrm{diag}[f'(\boldsymbol{u})]$$

$$\frac{\partial}{\partial X^T}(AX) = I_k\otimes A, \quad X \in \mathbb{R}^{n\times k} \qquad \frac{\partial}{\partial X} f(\boldsymbol{u}) = \frac{\partial \boldsymbol{u}^T}{\partial X}\left[\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}\otimes I\right]$$

$$f(\boldsymbol{x}) = \frac{\boldsymbol{u}(\boldsymbol{x})}{v(\boldsymbol{x})} \qquad \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}) = \frac{1}{v^2}\left(\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} v - \frac{\partial v}{\partial \boldsymbol{x}}\boldsymbol{u}\right)$$

Model

procedure forward($\boldsymbol{x}; \boldsymbol{W}_{i,\dots,K}, \boldsymbol{\beta}_{i,\dots,K}$)

$\boldsymbol{z_0} \leftarrow \boldsymbol{x}$

for $i = 0, \dots, K-2$ **do**

$\boldsymbol{h}_{i+1} \leftarrow \boldsymbol{W}_i\boldsymbol{z}_i + \boldsymbol{\beta}_i$

$\boldsymbol{z}_{i+1} = \sigma(\boldsymbol{h}_{i+1})$

end for

$\boldsymbol{h}_K \leftarrow \boldsymbol{W}_{K-1}\boldsymbol{z}_{K-1} + \boldsymbol{\beta}_{K-1}$

$$\boldsymbol{z}_K \leftarrow \begin{cases} \sigma(h_K) & \text{scaler} \\ \dfrac{\exp(\boldsymbol{h}_K)}{\mathbf{1}^T\exp(\boldsymbol{h}_K)} & \text{vector} \end{cases}$$

Loss and gradients

Scaler Output

- Output

$$z_K = \sigma(h_K), \quad \frac{dz_K}{dh_K} = z_K(1 - z_K)$$

- Loss

$$J(z_K; \boldsymbol{y}) = -y\ln z_K - (1-y)\ln(1 - z_K)$$

- Loss gradients

$$\frac{d}{dz_K}J(z_K; \boldsymbol{y}) = \frac{z_K - y}{z_K(1 - z_K)}$$

$$\frac{d}{dh_K}J(z_K; \boldsymbol{y}) = \frac{dz_K}{dh_K}\frac{d}{dz_K}J(z_K; \boldsymbol{y}) = z_K - y$$

Vector Output

- Output

$$\boldsymbol{z}_K = \frac{\exp(\boldsymbol{h}_K)}{\mathbf{1}^T\exp(\boldsymbol{h}_K)} \quad \frac{\partial \boldsymbol{z}_K{}^T}{\partial \boldsymbol{h}_K} = \mathrm{diag}(\boldsymbol{z}_K) - \boldsymbol{z}_K\boldsymbol{z}_K{}^T$$

- Loss

$$J(\boldsymbol{z}_K; \boldsymbol{y}) = -\boldsymbol{y}^T\ln\boldsymbol{z}_K$$

- Loss gradients

$$\frac{\partial}{\partial \boldsymbol{z}_K}J(\boldsymbol{z}_K; \boldsymbol{y}) = \mathrm{diag}(\boldsymbol{z}_K)^{-1}\boldsymbol{y}$$

$$\frac{\partial}{\partial \boldsymbol{h}_K}J(\boldsymbol{z}_K; \boldsymbol{y}) = \frac{\partial \boldsymbol{z}_K{}^T}{\partial \boldsymbol{h}_K}\frac{\partial J}{\partial \boldsymbol{z}_K} = \boldsymbol{z}_K - \boldsymbol{y}$$