Pull back / vJp rule for softmax

$$f(\underline{x}) = \underline{y}$$

$\underline{x} \in \mathbb{R}^N$ ... logit values $(x_i \in (-\infty, \infty))$

$\underline{y} \in \mathbb{R}^N$ ... probabilities $(y_i \in [0, 1])$ with $\sum_i y_i = 1$

$$y_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

task: backward propagate $\bar{\underline{y}} \in \mathbb{R}^N$ to $\bar{\underline{x}} \in \mathbb{R}^N$

$$\bar{\underline{x}}^T = \bar{\underline{y}}^T \frac{\partial \underline{y}}{\partial \underline{x}}$$

$$\bar{x}_k = \frac{\partial y_i}{\partial x_k} \bar{y}_i$$

$$\left(= y_i \delta_{ik} - y_i y_k\right)$$

see pushforward video

$$\bar{x}_k = (y_i \delta_{ik} - y_i y_k)\, \bar{y}_i$$

$$= y_i \bar{y}_i \delta_{ik} - y_i \bar{y}_i y_k$$

$$= y_k \bar{y}_k - y_i \bar{y}_i y_k$$

in symbolic notation

$$\bar{\underline{x}} = \underline{y} \odot \bar{\underline{y}} - (\underline{y}^T \bar{\underline{y}})\, \underline{y}$$

Full pullback rule

$$\boxed{\mathcal{B}\left(\text{softmax}, (\underline{x},), (\bar{\underline{y}},)\right) = \left(\left(\underbrace{\text{softmax}(\underline{x})}_{\underline{y}},\right), \left(\underbrace{\underline{y} \odot \bar{\underline{y}} - (\underline{y}^T \bar{\underline{y}})\, \underline{y}}_{\bar{\underline{x}}},\right)\right)}$$