

Appendix A DiVe Learning Equations

In this section we will derive the equations that will compose the learning of DiVe. The steps will only be from DiVe Single Point, the steps for DiVe Dual Point will follow the same principles.

A.1 Activation Sigmoid

We will minimize our $\text{NLL}_{\text{NS}}(\mathcal{D})$ function using the descending gradient technique, so for each variable you have to find the Δ gradient, thus considering $\text{NLL}_{\text{NS}}(\mathcal{D})$ as:

$$\text{NLL}_{\text{NS}}(\mathcal{D}) = - \sum_{s_n \in \mathcal{D}} \sum_{i=j}^{k_{s_n}} \left[\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) - \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (1)$$

Reminder that :

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

$$f(\mathcal{X}(w_i), \mathcal{X}(c_i)) = -\frac{1}{2} \|\mathcal{X}(w_i) - \mathcal{X}(c_i)\|_2^2 + \frac{\alpha}{2} \|\mathcal{X}(w_i)\|_2^2 + \frac{\alpha}{2} \|\mathcal{X}(c_i)\|_2^2 \quad (3)$$

Thus, the gradient for the $\mathcal{X}(w_i)$ vector considering a batch size of 1:

$$\Delta_{\mathcal{X}(w_i)} \text{NLL}(\mathcal{D}) = \left[-\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (4)$$

$$= \left[-\Delta_{\mathcal{X}(w_i)} \log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(w_i)} \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (5)$$

$$= \left[-\frac{\Delta_{\mathcal{X}(w_i)} \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + 0 \right] \quad (6)$$

$$= \left[-\frac{(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))) \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) \Delta_{\mathcal{X}(w_i)} f(\mathcal{X}(w_i), \mathcal{X}(c_i))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} \right] \quad (7)$$

$$= -(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))) [-\mathcal{X}(w_i) + \mathcal{X}(c_i) + \alpha \mathcal{X}(w_i)] \quad (8)$$

Thus, the gradient for the $\mathcal{X}(c_i)$ vector considering a batch size of 1:

$$\Delta_{\mathcal{X}(c_i)} \text{NLL}(\mathcal{D}) = \left[-\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (9)$$

$$= \left[-\Delta_{\mathcal{X}(c_i)} \log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(c_i)} \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (10)$$

$$= \left[-\frac{\Delta_{\mathcal{X}(c_i)} \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + \sum_{v \in \mathcal{V}'} \frac{\Delta_{\mathcal{X}(c_i)} \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}{\sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))} \right] \quad (11)$$

$$= -\frac{(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))) \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) \Delta_{\mathcal{X}(c_i)} f(\mathcal{X}(w_i), \mathcal{X}(c_i))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + \quad (12)$$

$$\sum_{v \in \mathcal{V}'} \frac{-\sigma(f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) - \Delta_{\mathcal{X}(c_i)} f(\mathcal{X}(w_v), \mathcal{X}(c_i))}{\sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}$$

$$= -(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))) [\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha \mathcal{X}(c_i)] \quad (13)$$

$$+ \sum_{v \in \mathcal{V}'} -\sigma(f(\mathcal{X}(w_v), \mathcal{X}(c_i))) [-(\mathcal{X}(w_v) - \mathcal{X}(c_i) + \alpha \mathcal{X}(c_i))]$$

A.2 Activation Tanh

$$\text{NLL}_{\text{NS}}(\mathcal{D}) = - \sum_{s_n \in D} \sum_{i=j}^{k_{s_n}} \left[\log \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) - \sum_{v \in \mathcal{V}'} \log \tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (14)$$

$$f(\mathcal{X}(w_i), \mathcal{X}(c_i)) = -\frac{1}{2} \|\mathcal{X}(w_i) - \mathcal{X}(c_i)\|_2^2 + \frac{\alpha}{2} \|\mathcal{X}(w_i)\|_2^2 + \frac{\alpha}{2} \|\mathcal{X}(c_i)\|_2^2 \quad (15)$$

Thus, the gradient for the $\mathcal{X}(w_i)$ vector considering a batch size of 1:

$$\Delta_{\mathcal{X}(w_i)} \text{NLL}(\mathcal{D}) = \left[-\log \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (16)$$

$$= \left[-\Delta_{\mathcal{X}(w_i)} \log \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(w_i)} \sum_{v \in \mathcal{V}'} \log \tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (17)$$

$$= \left[-\frac{\Delta_{\mathcal{X}(w_i)} \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + 0 \right] \quad (18)$$

$$= \left[-\frac{(1 - \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))^2 \Delta_{\mathcal{X}(w_i)} f(\mathcal{X}(w_i), \mathcal{X}(c_i))}{\tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} \right] \quad (19)$$

$$= -\frac{(1 - \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))^2}{\tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} [-\mathcal{X}(w_i) + \mathcal{X}(c_i) + \alpha \mathcal{X}(w_i)] \quad (20)$$

Thus, the gradient for the $\mathcal{X}(c_i)$ vector considering a batch size of 1:

$$\Delta_{\mathcal{X}(c_i)} \text{NLL}(\mathcal{D}) = \left[-\log \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (21)$$

$$= \left[-\Delta_{\mathcal{X}(c_i)} \log \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(c_i)} \sum_{v \in \mathcal{V}'} \log \tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (22)$$

$$= \left[-\frac{\Delta_{\mathcal{X}(c_i)} \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + \sum_{v \in \mathcal{V}'} \frac{\Delta_{\mathcal{X}(c_i)} \tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}{\tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))} \right] \quad (23)$$

$$= -\frac{(1 - \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))^2 \Delta_{\mathcal{X}(c_i)} f(\mathcal{X}(w_i), \mathcal{X}(c_i))}{\tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + \quad (24)$$

$$\sum_{v \in \mathcal{V}'} \frac{-\frac{4 \exp(2*f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}{(\exp(2*f(\mathcal{X}(w_v), \mathcal{X}(c_i))) + 1)^2} - \Delta_{\mathcal{X}(c_i)} f(\mathcal{X}(w_v), \mathcal{X}(c_i))}{\tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}$$

$$= -\frac{(1 - \tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))^2}{\tanh(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} [\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha \mathcal{X}(c_i)] + \quad (25)$$

$$\sum_{v \in \mathcal{V}'} \frac{-\frac{4 \exp(2*f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}{(\exp(2*f(\mathcal{X}(w_v), \mathcal{X}(c_i))) + 1)^2} [-(\mathcal{X}(w_v) - \mathcal{X}(c_i) + \alpha \mathcal{X}(c_i))]}{\tanh(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}$$

A.3 Activation Exp

$$\text{NLL}_{\text{NS}}(\mathcal{D}) = - \sum_{s_n \in D} \sum_{i=j}^{k_{s_n}} \left[\log \exp(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) - \sum_{v \in \mathcal{V}'} \log \exp(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (26)$$

Reminder that:

$$f(\mathcal{X}(w_i), \mathcal{X}(c_i)) = -\frac{1}{2} \|\mathcal{X}(w_i) - \mathcal{X}(c_i)\|_2^2 + \frac{\alpha}{2} \|\mathcal{X}(w_i)\|_2^2 + \frac{\alpha}{2} \|\mathcal{X}(c_i)\|_2^2 \quad (27)$$

Thus, the gradient for the $\mathcal{X}(w_i)$ vector considering a batch size of 1:

$$\Delta_{\mathcal{X}(w_i)} \text{NLL}(\mathcal{D}) = \left[-f(\mathcal{X}(w_i), \mathcal{X}(c_i)) + \sum_{v \in \mathcal{V}'} -f(\mathcal{X}(w_v), \mathcal{X}(c_i)) \right] \quad (28)$$

$$= \left[-\Delta_{\mathcal{X}(w_i)} (f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(w_i)} \sum_{v \in \mathcal{V}'} (-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \quad (29)$$

$$= -[-\mathcal{X}(w_i) + \mathcal{X}(c_i) + \alpha\mathcal{X}(w_i)] \quad (30)$$

Thus, the gradient for the $\mathcal{X}(c_i)$ vector considering a batch size of 1:

$$\begin{aligned} \Delta_{\mathcal{X}(c_i)} \text{NLL}(\mathcal{D}) &= -(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} (-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \\ &= [-\Delta_{\mathcal{X}(c_i)}(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(c_i)} \sum_{v \in \mathcal{V}'} (-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))] \\ &= -[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)] + \sum_{v \in \mathcal{V}'} -[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)] \end{aligned} \quad (31)$$