Appendix A DIVE LEARNING EQUATIONS

In this section we derive the equations that compose the learning of DiVe. The steps below are only 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:

$$\text{NLL}_{\text{NS}}(\mathcal{D}) = -\sum_{s_n \in D} \sum_{i=j}^{k_{s_n}} \left[\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) - \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{10}$$

Reminder that:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

$$f(\mathcal{X}(w_i), \mathcal{X}(c_i)) = -\frac{1}{2}||\mathcal{X}(w_i) - \mathcal{X}(c_i)||_2^2 + \frac{\alpha}{2}||\mathcal{X}(w_i)||_2^2 + \frac{\alpha}{2}||\mathcal{X}(c_i)||_2^2 \tag{12}$$

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

$$\Delta_{\mathcal{X}(w_i)}\text{NLL}(\mathcal{D}) = \left[-\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{13}$$

$$= \left[-\Delta_{\mathcal{X}(w_i)} \log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(w_i)} \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{14}$$

$$= \left[-\frac{\Delta_{\mathcal{X}(w_i)} \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + 0 \right] \tag{15}$$

$$= \left[-\frac{(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))\Delta_{\mathcal{X}(w_i)} f(\mathcal{X}(w_i), \mathcal{X}(c_i))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))} \right] \tag{16}$$

$$= -(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))[-\mathcal{X}(w_i) + \mathcal{X}(c_i) + \alpha\mathcal{X}(w_i)] \tag{17}$$

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

$$\Delta_{\mathcal{X}(c_i)}\text{NLL}(\mathcal{D}) = \left[-\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{18}$$

$$= \left[-\Delta_{\mathcal{X}(c_i)} \log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \Delta_{\mathcal{X}(c_i)} \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{19}$$

$$= \left[-\frac{\Delta_{\mathcal{X}(c_i)} \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + \sum_{v \in \mathcal{V}'} \frac{\Delta_{\mathcal{X}(c_i)} \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))}{\sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))} \right] \tag{20}$$

$$= -\frac{(1-\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i))))\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))\Delta_{\mathcal{X}(c_i)}f(\mathcal{X}(w_i),\mathcal{X}(c_i))}{\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))} + \sum_{v\in\mathcal{V}'}\frac{-\sigma(f(\mathcal{X}(w_v),\mathcal{X}(c_i)))\sigma(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))-\Delta_{\mathcal{X}(c_i)}f(\mathcal{X}(w_v),\mathcal{X}(c_i))}{\sigma(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))} \tag{21}$$

$$= -(1-\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i))))[\mathcal{X}(w_i)-\mathcal{X}(c_i)+\alpha\mathcal{X}(c_i)] + \sum_{v\in\mathcal{V}'}-\sigma(f(\mathcal{X}(w_v),\mathcal{X}(c_i)))[-(\mathcal{X}(w_v)-\mathcal{X}(c_i)+\alpha\mathcal{X}(c_i))] \tag{22}$$

A.2 Activation Tanh

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

$$\mathrm{NLL}_{\mathrm{NS}}(\mathcal{D}) = -\sum_{s_n\in D}\sum_{i=j}^{k_{s_n}}\left[\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) - \sum_{v\in\mathcal{V}'}\log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))\right] \tag{23}$$

Reminder that:

$$\tanh(x) = \frac{e^x+e^{-1}}{e^x-e^{-1}} \tag{24}$$

$$f(\mathcal{X}(w_i),\mathcal{X}(c_i)) = -\frac{1}{2}||\mathcal{X}(w_i)-\mathcal{X}(c_i)||_2^2 + \frac{\alpha}{2}||\mathcal{X}(w_i)||_2^2 + \frac{\alpha}{2}||\mathcal{X}(c_i)||_2^2 \tag{25}$$

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

$$\Delta_{\mathcal{X}(w_i)}\mathrm{NLL}(\mathcal{D}) = \left[-\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \sum_{v\in\mathcal{V}'}\log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))\right] \tag{26}$$

$$= \left[-\Delta_{\mathcal{X}(w_i)}\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \Delta_{\mathcal{X}(w_i)}\sum_{v\in\mathcal{V}'}\log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))\right] \tag{27}$$

$$= \left[-\frac{\Delta_{\mathcal{X}(w_i)}\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))} + 0\right] \tag{28}$$

$$= \left[-\frac{(1-\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))^2)\Delta_{\mathcal{X}(w_i)}f(\mathcal{X}(w_i),\mathcal{X}(c_i))}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))}\right] \tag{29}$$

$$= -\frac{(1-\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))^2)}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))}[-\mathcal{X}(w_i)+\mathcal{X}(c_i)+\alpha\mathcal{X}(w_i)] \tag{30}$$

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

$$\Delta_{\mathcal{X}(c_i)}\text{NLL}(\mathcal{D}) = \left[-\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \sum_{v\in\mathcal{V}'} \log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) \right] \tag{31}$$

$$= \left[-\Delta_{\mathcal{X}(c_i)}\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \Delta_{\mathcal{X}(c_i)}\sum_{v\in\mathcal{V}'} \log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) \right] \tag{32}$$

$$= \left[-\frac{\Delta_{\mathcal{X}(c_i)}\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))} + \sum_{v\in\mathcal{V}'} \frac{\Delta_{\mathcal{X}(c_i)}\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}{\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))} \right] \tag{33}$$

$$= -\frac{(1-\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))^2)\Delta_{\mathcal{X}(c_i)}f(\mathcal{X}(w_i),\mathcal{X}(c_i))}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))} + \sum_{v\in\mathcal{V}'} \frac{-\frac{4\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}{(\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))+1)^2}-\Delta_{\mathcal{X}(c_i)}f(\mathcal{X}(w_v),\mathcal{X}(c_i))}{\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))} \tag{34}$$

$$= -\frac{(1-\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))^2)}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))}[\mathcal{X}(w_i)-\mathcal{X}(c_i)+\alpha\mathcal{X}(c_i)] + \sum_{v\in\mathcal{V}'} \frac{-\frac{4\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}{(\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))+1)^2}[-(\mathcal{X}(w_v)-\mathcal{X}(c_i)+\alpha\mathcal{X}(c_i))]}{\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))} \tag{35}$$

A.3 Activation Exp

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:

$$\text{NLL}_{\text{NS}}(\mathcal{D}) = -\sum_{s_n\in D}\sum_{i=j}^{k_{s_n}} \left[\log\exp(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) - \sum_{v\in\mathcal{V}'} \log\exp(-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) \right] \tag{36}$$

Lembrando que:

$$f(\mathcal{X}(w_i),\mathcal{X}(c_i)) = -\frac{1}{2}||\mathcal{X}(w_i)-\mathcal{X}(c_i)||_2^2 + \frac{\alpha}{2}||\mathcal{X}(w_i)||_2^2 + \frac{\alpha}{2}||\mathcal{X}(c_i)||_2^2 \tag{37}$$

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

$$\Delta_{\mathcal{X}(w_i)}\text{NLL}(\mathcal{D}) = \left[-f(\mathcal{X}(w_i),\mathcal{X}(c_i)) + \sum_{v\in\mathcal{V}'} -f(\mathcal{X}(w_v),\mathcal{X}(c_i)) \right] \tag{38}$$

$$= \left[-\Delta_{\mathcal{X}(w_i)}(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \Delta_{\mathcal{X}(w_i)}\sum_{v\in\mathcal{V}'} (-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) \right] \tag{39}$$

$$= -[-\mathcal{X}(w_i)+\mathcal{X}(c_i)+\alpha\mathcal{X}(w_i)] \tag{40}$$

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

$$\begin{aligned}\Delta_{\mathcal{X}(c_i)}\mathrm{NLL}(\mathcal{D}) &= -(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \sum_{v\in\mathcal{V}'}(-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) \\ &= [-\Delta_{\mathcal{X}(c_i)}(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \Delta_{\mathcal{X}(c_i)}\sum_{v\in\mathcal{V}'}(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))] \\ &= -[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)] + \sum_{v\in\mathcal{V}'} -[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)]\end{aligned} \tag{41}$$