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, this steps for DiVe Dual Point will follow the same principles.

A.1 Activation Sigmoid

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 $\boldsymbol{\Delta}$ gradient, thus considering $\mathrm{NLL}_{\mathrm{NS}}(\mathcal{D})$ as:

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

Reminder that :

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

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

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

$$\boldsymbol{\Delta}_{\mathcal{X}(w_i)}\mathrm{NLL}(\mathcal{D}) = \left[-\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{4}$$

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

$$= \left[-\frac{\boldsymbol{\Delta}_{\mathcal{X}(w_i)} \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))} + 0 \right] \tag{6}$$

$$= \left[-\frac{(1 - \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))))\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i)))\boldsymbol{\Delta}_{\mathcal{X}(w_i)} f(\mathcal{X}(w_i), \mathcal{X}(c_i))}{\sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))} \right] \tag{7}$$

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

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

$$\boldsymbol{\Delta}_{\mathcal{X}(c_i)}\mathrm{NLL}(\mathcal{D}) = \left[-\log \sigma(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v \in \mathcal{V}'} \log \sigma(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \right] \tag{9}$$

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

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

$$\begin{aligned} &= -\frac{(1-\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i))))\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))\boldsymbol{\Delta}_{\mathcal{X}(c_i)} f(\mathcal{X}(w_i),\mathcal{X}(c_i))}{\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))} + \\ &\sum_{v\in\mathcal{V}'} \frac{-\sigma(f(\mathcal{X}(w_v),\mathcal{X}(c_i)))\sigma(-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) - \boldsymbol{\Delta}_{\mathcal{X}(c_i)} f(\mathcal{X}(w_v),\mathcal{X}(c_i))}{\sigma(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))} \end{aligned} \tag{12}$$

$$\begin{aligned} &= -(1-\sigma(f(\mathcal{X}(w_i),\mathcal{X}(c_i))))[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)] \\ &+ \sum_{v\in\mathcal{V}'} -\sigma(f(\mathcal{X}(w_v),\mathcal{X}(c_i)))[-(\mathcal{X}(w_v) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i))] \end{aligned} \tag{13}$$

A.2 Activation Tanh

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

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

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

$$\boldsymbol{\Delta}_{\mathcal{X}(w_i)}\mathrm{NLL}(\mathcal{D}) = \left[-\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \sum_{v\in\mathcal{V}'} \log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i))) \right] \tag{16}$$

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

$$= \left[-\frac{\boldsymbol{\Delta}_{\mathcal{X}(w_i)} \tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))} + 0 \right] \tag{18}$$

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

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

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

$$\mathbf{\Delta}_{\mathcal{X}(c_i)}\mathrm{NLL}(\mathcal{D}) = \left[-\log\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))) + \sum_{v\in\mathcal{V}'}\log\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))\right] \tag{21}$$

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

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

$$\begin{aligned}= &-\frac{(1-\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))^2)\mathbf{\Delta}_{\mathcal{X}(c_i)}f(\mathcal{X}(w_i),\mathcal{X}(c_i))}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i))} + \\ &\sum_{v\in\mathcal{V}'}\frac{-\frac{4\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}{(\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))+1)^2}-\mathbf{\Delta}_{\mathcal{X}(c_i)}f(\mathcal{X}(w_v),\mathcal{X}(c_i))}{\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}\end{aligned} \tag{24}$$

$$\begin{aligned}= &-\frac{(1-\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))^2)}{\tanh(f(\mathcal{X}(w_i),\mathcal{X}(c_i)))}[\mathcal{X}(w_i)-\mathcal{X}(c_i)+\alpha\mathcal{X}(c_i)] + \\ &\sum_{v\in\mathcal{V}'}\frac{-\frac{4\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}{(\exp(2*f(\mathcal{X}(w_v),\mathcal{X}(c_i)))+1)^2}[-(\mathcal{X}(w_v)-\mathcal{X}(c_i)+\alpha\mathcal{X}(c_i))]}{\tanh(-f(\mathcal{X}(w_v),\mathcal{X}(c_i)))}\end{aligned} \tag{25}$$

A.3 Activation Exp

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

Reminder that:

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

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

$$\mathbf{\Delta}_{\mathcal{X}(w_i)}\mathrm{NLL}(\mathcal{D}) = \left[-f(\mathcal{X}(w_i),\mathcal{X}(c_i)) + \sum_{v\in\mathcal{V}'}-f(\mathcal{X}(w_v),\mathcal{X}(c_i))\right] \tag{28}$$

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

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

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

$$\begin{aligned}
\boldsymbol{\Delta}_{\mathcal{X}(c_i)}\mathrm{NLL}(\mathcal{D}) &= -(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \sum_{v\in\mathcal{V}'}(-f(\mathcal{X}(w_v), \mathcal{X}(c_i))) \\
&= [-\boldsymbol{\Delta}_{\mathcal{X}(c_i)}(f(\mathcal{X}(w_i), \mathcal{X}(c_i))) + \boldsymbol{\Delta}_{\mathcal{X}(c_i)}\sum_{v\in\mathcal{V}'}(-f(\mathcal{X}(w_v), \mathcal{X}(c_i)))] \\
&= -[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)] + \sum_{v\in\mathcal{V}'} -[\mathcal{X}(w_i) - \mathcal{X}(c_i) + \alpha\mathcal{X}(c_i)]
\end{aligned} \tag{31}$$