Supplementary material

This supplementary material presents the key equations in the manuscripts. The proofs of all these equations are included in this material.

Appendix A. Joint distribution

Similar to other Boltzmann Machine models, the joint probability is the exponential function of negative energy function over the normalizing constant:

$$\begin{aligned} p\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right) &= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[-E\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right)\right] \\ &= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[\sum_{i=1}^{M} a_i \mathrm{v}_i + \sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)} \mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right) h_j^{(l)}\right. \\ &\quad \left. + \sum_{i=1}^{M}\sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} \mathrm{v}_i w_{ij}^{(1)} h_j^{(1)} + \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \bar{\gamma}_i^{(1)} \bar{\gamma}_j^{(2)} h_i^{(1)} w_{ij}^{(2)} h_j^{(2)}\right] \end{aligned}$$

where $E\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right)$ is described in Sec. 3.2 of the manuscript.

Appendix B. Marginal distribution

We marginalize out each layer to obtain the corresponding marginal distribution as follows:

$$
\begin{aligned}
\sum_{\boldsymbol{v}} p\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right) &= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \sum_{\boldsymbol{v}} \exp\left[-E\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right)\right] \\
&= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[\sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)}\mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right)h_j^{(l)} + \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\bar{\gamma}_i^{(1)}\bar{\gamma}_j^{(2)}h_i^{(1)}w_{ij}^{(2)}h_j^{(2)}\right] \\
&\times \sum_{\boldsymbol{v}} \exp\left[\sum_{i=1}^{M}\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}w_{ij}^{(1)}h_j^{(1)}\right)\right] \\
&= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[\sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)}\mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right)h_j^{(l)} + \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\bar{\gamma}_i^{(1)}\bar{\gamma}_j^{(2)}h_i^{(1)}w_{ij}^{(2)}h_j^{(2)}\right] \\
&\times \sum_{\boldsymbol{v}} \prod_{i=1}^{M} \exp\left[\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}w_{ij}^{(1)}h_j^{(1)}\right)\right] \\
&= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[\sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)}\mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right)h_j^{(l)} + \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\bar{\gamma}_i^{(1)}\bar{\gamma}_j^{(2)}h_i^{(1)}w_{ij}^{(2)}h_j^{(2)}\right] \\
&\times \prod_{i=1}^{M}\sum_{\mathrm{v}_i\in\{0,1\}} \exp\left[\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}w_{ij}^{(1)}h_j^{(1)}\right)\right]
\end{aligned}
$$

By applying the same mathematical steps, we arrive the marginal distribution of hidden layers as expected:

$$
\begin{aligned}
\sum_{\boldsymbol{h}^{(1)}} p\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right) &= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[\sum_{i=1}^{M}a_i\mathrm{v}_i + \sum_{j=1}^{K_2}\left(\bar{\gamma}_j^{(2)}\mathrm{b}_j^{(2)} + \bar{\beta}_j^{(2)}\right)h_j^{(2)}\right] \\
&\times \prod_{i=1}^{K_1}\sum_{h_i^{(1)}\in\{0,1\}} \exp\left[h_i^{(1)}\left(\bar{\gamma}_i^{(1)}\left(\mathrm{b}_i^{(1)} + \sum_{j=1}^{M}\mathrm{v}_j w_{ji}^{(1)} + \sum_{j=1}^{K_2}\bar{\gamma}_j^{(2)}w_{ij}^{(2)}h_j^{(2)}\right) + \bar{\beta}_i^{(1)}\right)\right] \\
\sum_{\boldsymbol{h}^{(2)}} p\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\Gamma}\right) &= \frac{1}{\mathcal{Z}\left(\boldsymbol{\Gamma}\right)} \exp\left[\sum_{i=1}^{M}a_i\mathrm{v}_i + \sum_{j=1}^{K_1}\left(\bar{\gamma}_j^{(1)}\mathrm{b}_j^{(1)} + \bar{\beta}_j^{(1)}\right)h_j^{(1)} + \sum_{i=1}^{M}\sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}\mathrm{v}_i w_{ij}^{(1)}h_j^{(1)}\right] \\
&\times \prod_{i=1}^{K_2}\sum_{h_i^{(2)}\in\{0,1\}} \exp\left[h_i^{(2)}\left(\bar{\gamma}_i^{(2)}\left(\mathrm{b}_i^{(2)} + \sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}h_j^{(1)}w_{ji}^{(2)}\right) + \bar{\beta}_i^{(2)}\right)\right]
\end{aligned}
$$

Appendix C. Conditional distribution

The joint distribution and marginal distributions above result in the conditional distribution of each layer given its adjacent layers.

$$
\begin{aligned}
p\left(\boldsymbol{v}|\boldsymbol{h}^{(1)};\boldsymbol{\Gamma}\right) &= \frac{p\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)}{\sum_{\boldsymbol{v}} p\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)} \\
&= \frac{\exp\left[\sum_{i=1}^{M} a_i \mathrm{v}_i + \sum_{i=1}^{M}\sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} \mathrm{v}_i w_{ij}^{(1)} h_j^{(1)}\right]}{\prod_{i=1}^{M}\sum_{\mathrm{v}_i\in\{0,1\}} \exp\left[\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} w_{ij}^{(1)} h_j^{(1)}\right)\right]} \\
&= \prod_{i=1}^{M} \frac{\exp\left[\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} w_{ij}^{(1)} h_j^{(1)}\right)\right]}{1+\exp\left[\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} w_{ij}^{(1)} h_j^{(1)}\right)\right]} \\
&= \prod_{i=1}^{M} \sigma\left(\mathrm{v}_i\left(a_i + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} w_{ij}^{(1)} h_j^{(1)}\right)\right)
\end{aligned}
$$

Since the units in the same layer are independent conditioning on the other adjacent layers, we gain the conditional probability of one visible neuron to be on:

$$
p\left(\mathrm{v}_m = 1|\boldsymbol{h}^{(1)};\boldsymbol{\Gamma}\right) \quad = \quad \sigma\left(a_m + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} w_{mj}^{(1)} h_j^{(1)}\right)
$$

Similarly, the conditional probability equations for hidden units should be:

$$
\begin{aligned}
p\left(h_n^{(1)} = 1|\boldsymbol{v},\boldsymbol{h}^{(2)};\boldsymbol{\Gamma}\right) &= \sigma\left(\mathcal{B}_{\gamma_n^{(1)},\beta_n^{(1)}}\left(\mathrm{b}_n^{(1)} + \sum_{j=1}^{M} \mathrm{v}_j w_{jn}^{(1)} + \sum_{j=1}^{K_2} \bar{\gamma}_j^{(2)} w_{nj}^{(2)} h_j^{(2)}\right)\right) \\
p\left(h_n^{(2)} = 1|\boldsymbol{h}^{(1)};\boldsymbol{\Gamma}\right) &= \sigma\left(\mathcal{B}_{\gamma_n^{(2)},\beta_n^{(2)}}\left(\mathrm{b}_n^{(2)} + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} h_j^{(1)} w_{jn}^{(2)}\right)\right)
\end{aligned}
$$

Appendix D. Data log likelihood lower bound

The data log-likelihood $\log \mathcal{L}\left(\boldsymbol{v};\boldsymbol{\Gamma}\right)$ over the data points is intractable because it requires the sum of all possible hidden states. Consequently, this function is estimated by computing its lowerbound $B(\boldsymbol{v};\boldsymbol{\Gamma},\tilde{\boldsymbol{\Gamma}})$ with respect to BNDBM parameters $\boldsymbol{\Gamma}$ and variational parameters $\tilde{\boldsymbol{\Gamma}} = \{\tilde{\mu}_i^{(l)}\}$:

$$
\log \mathcal{L}\left(\boldsymbol{v};\boldsymbol{\Gamma}\right) \quad \geq \quad \mathbb{E}_{q\left(\boldsymbol{h}|\boldsymbol{v};\tilde{\boldsymbol{\Gamma}}\right)}\left[\log p\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] + H\left[q\left(\boldsymbol{h}|\boldsymbol{v};\tilde{\boldsymbol{\Gamma}}\right)\right] = B(\boldsymbol{v};\boldsymbol{\Gamma},\tilde{\boldsymbol{\Gamma}})
$$

where $H\left[\bullet\right]$ is the entropy term while $q\left(\boldsymbol{h}|\boldsymbol{v};\tilde{\boldsymbol{\Gamma}}\right) = \prod_{l=1}^{2}\prod_{i=1}^{K_l}\left(\tilde{\mu}_i^{(l)}\right)^{h_i^{(l)}}\left(1-\tilde{\mu}_i^{(l)}\right)^{1-h_i^{(l)}}$ is the variational approximate distribution.

The explicit expression of the bound is obtained by expanding the expectation and entropy terms. To simplify the material, we ignore the parameter notations in the bound and distribution expression.

$$
\begin{aligned}
\mathbb{E}_q\left[\log p\left(\boldsymbol{v},\boldsymbol{h}\right)\right] &= \sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\log\frac{e^{-E\left(\boldsymbol{v},\boldsymbol{h}\right)}}{\mathcal{Z}} = \sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\left[-E\left(\boldsymbol{v},\boldsymbol{h}\right)\right] - \sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\log\mathcal{Z} \\
&= \sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\left[\sum_{i=1}^{M} a_i \mathrm{v}_i + \sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)}\mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right)h_j^{(l)}\right] \\
&+ \sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\left[\sum_{i=1}^{M}\sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}\mathrm{v}_i w_{ij}^{(1)}h_j^{(1)} + \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\bar{\gamma}_i^{(1)}\bar{\gamma}_j^{(2)}h_i^{(1)}w_{ij}^{(2)}h_j^{(2)}\right] - \log\mathcal{Z} \\
&= \sum_{i=1}^{M} a_i \mathrm{v}_i + \sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)}\mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right)\mathbb{E}_q\left[h_j^{(l)}\right] + \sum_{i=1}^{M}\sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}\mathrm{v}_i w_{ij}^{(1)}\mathbb{E}_q\left[h_j^{(1)}\right] \\
&+ \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\bar{\gamma}_i^{(1)}\bar{\gamma}_j^{(2)}w_{ij}^{(2)}\mathbb{E}_q\left[h_i^{(1)}\right]\mathbb{E}_q\left[h_j^{(2)}\right] - \log\mathcal{Z} \\
&= \sum_{i=1}^{M} a_i \mathrm{v}_i + \sum_{l=1}^{2}\sum_{j=1}^{K_l}\left(\bar{\gamma}_j^{(l)}\mathrm{b}_j^{(l)} + \bar{\beta}_j^{(l)}\right)\tilde{\mu}_j^{(l)} + \sum_{i=1}^{M}\sum_{j=1}^{K_1}\bar{\gamma}_j^{(1)}\mathrm{v}_i w_{ij}^{(1)}\tilde{\mu}_j^{(1)} \\
&+ \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\bar{\gamma}_i^{(1)}\bar{\gamma}_j^{(2)}w_{ij}^{(2)}\tilde{\mu}_i^{(1)}\tilde{\mu}_j^{(2)} - \log\mathcal{Z}
\end{aligned}
$$

$$
\begin{aligned}
H\left[q\left(\boldsymbol{h}|\boldsymbol{v}\right)\right] &= -\sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\log q\left(\boldsymbol{h}|\boldsymbol{v}\right) \\
&= -\sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\log\prod_{l=1}^{2}\prod_{i=1}^{K_l}\left(\tilde{\mu}_i^{(l)}\right)^{h_i^{(l)}}\left(1-\tilde{\mu}_i^{(l)}\right)^{1-h_i^{(l)}} \\
&= -\sum_{l=1}^{2}\sum_{i=1}^{K_l}\sum_{\boldsymbol{h}} q\left(\boldsymbol{h}|\boldsymbol{v}\right)\left[h_i^{(l)}\log\left(\tilde{\mu}_i^{(l)}\right) + \left(1-h_i^{(l)}\right)\log\left(1-\tilde{\mu}_i^{(l)}\right)\right] \\
&= -\sum_{l=1}^{2}\sum_{i=1}^{K_l}\left[\mathbb{E}_q\left[h_i^{(l)}\right]\log\left(\tilde{\mu}_i^{(l)}\right) + \mathbb{E}_q\left[1-h_i^{(l)}\right]\log\left(1-\tilde{\mu}_i^{(l)}\right)\right] \\
&= -\sum_{l=1}^{2}\sum_{i=1}^{K_l}\left[\tilde{\mu}_i^{(l)}\log\left(\tilde{\mu}_i^{(l)}\right) + \left(1-\tilde{\mu}_i^{(l)}\right)\log\left(1-\tilde{\mu}_i^{(l)}\right)\right]
\end{aligned}
$$

We aim to raise the lower bound and consequently maximize the log likelihood. The derivatives of the bound equations with respect to the first layer variational parameters are given as:

$$
\begin{aligned}
\frac{\partial B}{\partial \tilde{\mu}_n^{(1)}} &= \frac{\partial \mathbb{E}_q\left[\log p\left(\boldsymbol{v},\boldsymbol{h}\right)\right]}{\partial \tilde{\mu}_n^{(1)}} + \frac{\partial H\left[q\left(\boldsymbol{h}|\boldsymbol{v}\right)\right]}{\partial \tilde{\mu}_n^{(1)}} \\
&= \bar{\gamma}_n^{(1)} \mathrm{b}_n^{(1)} + \bar{\beta}_n^{(1)} + \sum_{j=1}^{M} \bar{\gamma}_n^{(1)} \mathrm{v}_j w_{jn}^{(1)} + \sum_{j=1}^{K_2} \bar{\gamma}_n^{(1)} \bar{\gamma}_j^{(2)} w_{nj}^{(2)} \tilde{\mu}_j^{(2)} \\
&- \left[\log\left(\tilde{\mu}_n^{(1)}\right) + 1 - \log\left(1-\tilde{\mu}_n^{(1)}\right) - 1\right] \\
&= \mathcal{B}_{\gamma_i^{(1)},\beta_i^{(1)}}\left(\mathrm{b}_n^{(1)} + \sum_{j=1}^{M} \mathrm{v}_j w_{jn}^{(1)} + \sum_{j=1}^{K_2} \bar{\gamma}_j^{(2)} w_{nj}^{(2)} \tilde{\mu}_j^{(2)}\right) - \log\left(\tilde{\mu}_n^{(1)}\right) + \log\left(1-\tilde{\mu}_n^{(1)}\right)
\end{aligned}
$$

By solving for $\tilde{\mu}_n^{(1)}$ where $\partial B/\tilde{\mu}_n^{(l)} = 0$, we end up with its fixed point update:

$$
\tilde{\mu}_n^{(1)} = \sigma\left(\mathcal{B}_{\gamma_n^{(1)},\beta_n^{(1)}}\left(\mathrm{b}_n^{(1)} + \sum_{j=1}^{M} \mathrm{v}_j w_{jn}^{(1)} + \sum_{j=1}^{K_2} \bar{\gamma}_j^{(2)} w_{nj}^{(2)} \tilde{\mu}_j^{(2)}\right)\right)
$$

Similarly, we arrive the update equation for $\tilde{\mu}_n^{(2)}$ depending on the current variational values of the first layer.

$$
\tilde{\mu}_n^{(2)} = \sigma\left(\mathcal{B}_{\gamma_n^{(2)},\beta_n^{(2)}}\left(\mathrm{b}_n^{(2)} + \sum_{j=1}^{K_1} \bar{\gamma}_j^{(1)} \tilde{\mu}_j^{(1)} w_{jn}^{(2)}\right)\right)
$$

Appendix E. Parameter update equations

The derivatives of the negative energy function with respect to biases and weights are given as:

$$
\begin{aligned}
\frac{\partial}{\partial a_m}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] &= \mathrm{v}_m \\
\frac{\partial}{\partial \mathrm{b}_n^{(l)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] &= \bar{\gamma}_n^{(l)} h_n^{(l)} \\
\frac{\partial}{\partial w_{mn}^{(1)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] &= \bar{\gamma}_n^{(1)} \mathrm{v}_m h_n^{(1)} \\
\frac{\partial}{\partial w_{mn}^{(2)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] &= \bar{\gamma}_m^{(1)} \bar{\gamma}_n^{(2)} h_m^{(1)} h_n^{(2)}
\end{aligned}
$$

For the normalization parameters, the shift parameters have the derivatives as:

$$
\frac{\partial}{\partial \beta_n^{(l)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] = \frac{\partial}{\partial \bar{\beta}_n^{(l)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] \frac{\partial \bar{\beta}_n^{(l)}}{\partial \beta_n^{(l)}} = h_n^{(l)}
$$

In order to compute the derivatives with respect to scale parameters, we firstly observe the fact that

$$\bar{\beta}_n^{(1)} = \beta_n^{(1)} - \frac{\gamma_n^{(1)}\mathbb{E}\left[t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right]}{\sqrt{\mathrm{Var}\left(t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right)+\epsilon}} = \beta_n^{(1)} - \bar{\gamma}_n^{(1)}\mathbb{E}\left[t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right]$$

$$\bar{\beta}_n^{(2)} = \beta_n^{(2)} - \frac{\gamma_n^{(2)}\mathbb{E}\left[t_n^{(2)}\left(\boldsymbol{h}^{(1)}\right)\right]}{\sqrt{\mathrm{Var}\left(t_n^{(2)}\left(\boldsymbol{h}^{(1)}\right)\right)+\epsilon}} = \beta_n^{(2)} - \bar{\gamma}_n^{(2)}\mathbb{E}\left[t_n^{(2)}\left(\boldsymbol{h}^{(1)}\right)\right]$$

We next replace all $\bar{\beta}_n^{(l)}$ in the energy equation by the right-hand side above and apply the chain rule with respect to $\bar{\gamma}_n^{(1)}$ and $\bar{\gamma}_n^{(2)}$:

$$\begin{aligned}
\frac{\partial}{\partial\gamma_n^{(1)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] &= \frac{\partial}{\partial\bar{\gamma}_n^{(1)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right]\frac{\partial\bar{\gamma}_n^{(1)}}{\partial\gamma_n^{(1)}} \\
&= \frac{\partial}{\partial\bar{\gamma}_n^{(1)}}\left[\left(\bar{\gamma}_n^{(1)}\mathrm{b}_n^{(1)}+\bar{\beta}_n^{(1)}\right)h_n^{(1)} + \sum_{i=1}^{M}\bar{\gamma}_n^{(1)}\mathrm{v}_i w_{in}^{(1)}h_n^{(1)} + \sum_{j=1}^{K_2}\bar{\gamma}_n^{(1)}\bar{\gamma}_j^{(2)}h_n^{(1)}w_{nj}^{(2)}h_j^{(2)}\right]\frac{\partial\bar{\gamma}_n^{(1)}}{\partial\gamma_n^{(1)}} \\
&= \left[\left(\mathrm{b}_n^{(1)}-\mathbb{E}\left[t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right]\right)h_n^{(1)} + \sum_{i=1}^{M}\mathrm{v}_i w_{in}^{(1)}h_n^{(1)} + \sum_{j=1}^{K_2}\bar{\gamma}_j^{(2)}h_n^{(1)}w_{nj}^{(2)}h_j^{(2)}\right]\frac{\partial\bar{\gamma}_n^{(1)}}{\partial\gamma_n^{(1)}} \\
&= h_n^{(1)}\frac{\left[\mathrm{b}_n^{(1)}-\mathbb{E}\left[t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right]+\sum_{i=1}^{M}\mathrm{v}_i w_{in}^{(1)}+\sum_{j=1}^{K_2}\bar{\gamma}_j^{(2)}w_{nj}^{(2)}h_j^{(2)}\right]}{\sqrt{\mathrm{Var}\left(t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right)+\epsilon}} \\
&= h_n^{(1)}\frac{\left[\mathrm{b}_n^{(1)}+\sum_{i=1}^{M}\mathrm{v}_i w_{in}^{(1)}+\sum_{j=1}^{K_2}\bar{\gamma}_j^{(2)}w_{nj}^{(2)}h_j^{(2)}-\mathbb{E}\left[t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right]\right]}{\sqrt{\mathrm{Var}\left(t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right)+\epsilon}} \\
&= h_n^{(1)}\frac{t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)-\mathbb{E}\left[t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right]}{\sqrt{\mathrm{Var}\left(t_n^{(1)}\left(\boldsymbol{v},\boldsymbol{h}^{(2)}\right)\right)+\epsilon}} \\
\frac{\partial}{\partial\gamma_n^{(2)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right] &= \frac{\partial}{\partial\bar{\gamma}_n^{(2)}}\left[-E\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\Gamma}\right)\right]\frac{\partial\bar{\gamma}_n^{(2)}}{\partial\gamma_n^{(2)}} \\
&= \frac{\partial}{\partial\bar{\gamma}_n^{(2)}}\left[\left(\bar{\gamma}_n^{(2)}\mathrm{b}_n^{(2)}+\bar{\beta}_n^{(2)}\right)h_n^{(2)} + \sum_{i=1}^{K_1}\bar{\gamma}_i^{(1)}\bar{\gamma}_n^{(2)}h_i^{(1)}w_{in}^{(2)}h_n^{(2)}\right]\frac{\partial\bar{\gamma}_n^{(2)}}{\partial\gamma_n^{(2)}} \\
&= h_n^{(2)}\frac{t_n^{(2)}\left(\boldsymbol{h}^{(1)}\right)-\mathbb{E}\left[t_n^{(2)}\left(\boldsymbol{h}^{(1)}\right)\right]}{\sqrt{\mathrm{Var}\left(t_n^{(2)}\left(\boldsymbol{h}^{(1)}\right)\right)+\epsilon}}
\end{aligned}$$

By substituting these derivatives into the log-likelihood gradient equation $\nabla_{\boldsymbol{\Gamma}}\mathcal{L} = \mathbb{E}_{\mathrm{data}}\left[-\frac{\partial E}{\partial\boldsymbol{\Gamma}}\right] - \mathbb{E}_{\mathrm{model}}\left[-\frac{\partial E}{\partial\boldsymbol{\Gamma}}\right]$, we achieve the gradient expression with respect to each parameter as follows:

$$\Delta a_m = \eta \left(\sum_{i=1}^{N_s} \frac{\mathrm{v}_m^{[i]}}{N_s} - \sum_{i=1}^{N_c} \frac{\hat{\mathrm{v}}_m^{\langle i \rangle}}{N_c} \right)$$

$$\Delta b_n^{(l)} = \eta \bar{\gamma}_n^{(l)} \left(\sum_{i=1}^{N_s} \frac{\tilde{\mu}_n^{(l)[i]}}{N_s} - \sum_{i=1}^{N_c} \frac{\hat{h}_n^{(l)\langle i \rangle}}{N_c} \right)$$

$$\Delta w_{mn}^{(1)} = \eta \bar{\gamma}_n^{(1)} \left(\sum_{i=1}^{N_s} \frac{\mathrm{v}_m^{[i]} \tilde{\mu}_n^{(1)[i]}}{N_s} - \sum_{i=1}^{N_c} \frac{\hat{\mathrm{v}}_m^{\langle i \rangle} \hat{h}_n^{(1)\langle i \rangle}}{N_c} \right)$$

$$\Delta w_{mn}^{(2)} = \eta \bar{\gamma}_m^{(1)} \bar{\gamma}_n^{(2)} \left(\sum_{i=1}^{N_s} \frac{\tilde{\mu}_m^{(1)[i]} \tilde{\mu}_n^{(2)[i]}}{N_s} - \sum_{i=1}^{N_c} \frac{\hat{h}_m^{(1)\langle i \rangle} \hat{h}_n^{(2)\langle i \rangle}}{N_c} \right)$$

$$\Delta \gamma_n^{(1)} = \eta \left(\sum_{i=1}^{N_s} \tilde{\mu}_n^{(1)[i]} \frac{t_n^{(1)} \left(\boldsymbol{v}^{[i]}, \tilde{\boldsymbol{\mu}}^{(2)[i]} \right) - \mathbb{E}[t_n^{(1)}]}{N_s \sqrt{\mathrm{Var}\left[t_n^{(1)} \right] + \epsilon}} - \sum_{i=1}^{N_c} \hat{h}_n^{(1)\langle i \rangle} \frac{t_n^{(1)} \left(\hat{\boldsymbol{v}}^{\langle i \rangle}, \hat{\boldsymbol{h}}^{(2)\langle i \rangle} \right) - \mathbb{E}[t_n^{(1)}]}{N_c \sqrt{\mathrm{Var}\left[t_n^{(1)} \right] + \epsilon}} \right)$$

$$\Delta \gamma_n^{(2)} = \eta \left(\sum_{i=1}^{N_s} \tilde{\mu}_n^{(2)[i]} \frac{t_n^{(2)} \left(\tilde{\boldsymbol{\mu}}^{(1)[i]} \right) - \mathbb{E}[t_n^{(2)}]}{N_s \sqrt{\mathrm{Var}\left[t_n^{(2)} \right] + \epsilon}} - \sum_{i=1}^{N_c} \hat{h}_n^{(2)\langle i \rangle} \frac{t_n^{(2)} \left(\hat{\boldsymbol{h}}^{(1)\langle i \rangle} \right) - \mathbb{E}[t_n^{(2)}]}{N_c \sqrt{\mathrm{Var}\left[t_n^{(2)} \right] + \epsilon}} \right)$$

$$\Delta \beta_n^{(l)} = \eta \left(\sum_{i=1}^{N_s} \frac{\tilde{\mu}_n^{(l)[i]}}{N_s} - \sum_{i=1}^{N_c} \frac{\hat{h}_n^{(l)\langle i \rangle}}{N_c} \right)$$

wherein, $\boldsymbol{v}^{[i]}$ and $\mathrm{v}_m^{[i]}$ are the i^{th} data point and and its m^{th} element. Similarly, the variational vector $\tilde{\boldsymbol{\mu}}^{(l)[i]}$ and the visible layer samples $\hat{\boldsymbol{v}}^{\langle i \rangle}$ and the hidden layer samples $\hat{\boldsymbol{h}}^{(l)\langle i \rangle}$ have their corresponding elements $\tilde{\mu}_n^{(l)[i]}$, $\hat{\mathrm{v}}_m^{\langle i \rangle}$ and $\hat{h}_n^{(l)\langle i \rangle}$ respectively.