Supplementary Material for: Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style[*]

Graham W. Taylor and Geoffrey E. Hinton

May 6, 2009

1 Introduction

In this document, we provide additional details for variants of Conditional Restricted Boltzmann Machines (CRBMs). Specifically we focus on each of the four models compared in the Quantitative Evaluation (Sec. 4.4). We collect the formulae required for contrastive divergence learning of parameters, synthesis from a trained model by alternating Gibbs samping, and forward prediction from a trained model by following the gradient of the free energy.

1.1 Notation

As in the original paper, we employ the notation $\boldsymbol{v}_{<t}$ to denote a "history" vector at time t which is a concatenation of the last N configurations of the visible units: $\boldsymbol{v}_{<t} = \boldsymbol{v}_{t-N}, \boldsymbol{v}_{t-N+1}, \ldots, \boldsymbol{v}_{t-1}$. We use i to index the D current visible units, $\boldsymbol{v}_t$, j to index the H current hidden units, $\boldsymbol{h}_t$, and k to index the $N \cdot D$-element history vector, $\boldsymbol{v}_{<t}$. In our discussion, we assume real-valued Gaussian visible units with unit standard deviation and stochastic binary hidden units. When we want to assign both a dimensional and time index to a vector, we use a comma-separated subscript (e.g. $v_{i,t}$). Since we commonly apply the sigmoid function, we use the shorthand notation:

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{1}$$

2 Conditional Restricted Boltzmann Machines

The Conditional Restricted Boltzmann Machine (CRBM) is parameterized by $\theta = \{\boldsymbol{a}, \boldsymbol{b}, A, B, W\}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are static biases on the visible and hidden units, respectively, A and B represent directed connections from the past visible units to the current visible and hidden units, respectively, and W represent undirected connections between visible and hidden units. Note that by setting $A = [0]$ and $B = [0]$ the CRBM reduces to an RBM.

[*] To appear in the Proceedings of the 26th International Conference on Machine Learning

2.1 Energy function

The energy function for a CRBM is:

$$E(\boldsymbol{v}_t, \boldsymbol{h}_t | \boldsymbol{v}_{<t}, \theta) = \frac{1}{2} \sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_{ij} W_{ij} v_{i,t} h_{j,t} - \sum_j \hat{b}_{j,t} h_{j,t} \tag{2}$$

where $\hat{a}_{i,t}$ is the "dynamic" bias on visible unit i and $\hat{b}_{j,t}$ is the "dynamic" bias on hidden unit j. The dynamic biases have both a static component and a contribution from the recent past history of the visible units:

$$\hat{a}_{i,t} = a_i + \sum_k A_{ki} v_{k,<t}, \tag{3}$$

$$\hat{b}_{j,t} = b_j + \sum_k B_{kj} v_{k,<t}. \tag{4}$$

2.2 Inference

Given a setting of the visible units at time t, $\boldsymbol{v}_t$, and the recent history, $\boldsymbol{v}_{<t}$, the posterior distribution over the hidden units is factorial:

$$p(\boldsymbol{h}_t | \boldsymbol{v}_t, \boldsymbol{v}_{<t}, \theta) = \prod_j p(h_{j,t} = 1 | \boldsymbol{v}_t, \boldsymbol{v}_{<t}, \theta)^{[h_{j,t}=1]} \left(1 - p(h_{j,t} = 1 | \boldsymbol{v}_t, \boldsymbol{v}_{<t}, \theta)\right)^{[h_{j,t}=0]} \tag{5}$$

where $[\cdot]$ are indicator variables and

$$p(h_{j,t} = 1 | \boldsymbol{v}_t, \boldsymbol{v}_{<t}, \theta) = \sigma \left(\hat{b}_{j,t} + \sum_i W_{ij} v_{i,t} \right). \tag{6}$$

This means that given $\boldsymbol{v}_t$ and $\boldsymbol{v}_{<t}$ the hidden units can be sampled in parallel.

2.3 Reconstruction

Similarly, given a setting of the hidden units at time t, $\boldsymbol{h}_t$, and the recent history, $\boldsymbol{v}_{<t}$, the "reconstruction" distribution over the visible units is also factorial:

$$p(\boldsymbol{v}_t | \boldsymbol{h}_t, \boldsymbol{v}_{<t}, \theta) = \prod_i p(v_{i,t} | \boldsymbol{h}_t, \boldsymbol{v}_{<t}, \theta) \tag{7}$$

where $p(v_{i,t} | \boldsymbol{h}_t, \boldsymbol{v}_{<t}, \theta)$ is a univariate Gaussian whose mean is the total input to visible unit i:

$$p(v_{i,t} | \boldsymbol{h}_{\boldsymbol{t}}, \boldsymbol{v}_{<t}, \theta) = \mathcal{N} \left(\hat{a}_{i,t} + \sum_j W_{ij} h_{j,t}, 1 \right) \tag{8}$$

This means that given $\boldsymbol{h}_t$ and $\boldsymbol{v}_{<t}$ the visible units can be sampled in parallel.

2.4 Contrastive divergence learning

The contrastive divergence (CD) updates for the parameters, θ, follow from the energy function (Eq. 2). For a sequence of length T, we sum the gradients over all windows[1] of size $N+1$:

$$\Delta W_{ij} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} h_{j,t} \rangle_{\text{data}} - \langle v_{i,t} h_{j,t} \rangle_{\text{recon}}\right) \tag{9}$$

$$\Delta A_{ki} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} v_{k,<t} \rangle_{\text{data}} - \langle v_{i,t} v_{k,<t} \rangle_{\text{recon}}\right) \tag{10}$$

$$\Delta B_{kj} \propto \sum_{t=N+1}^{T} \left(\langle h_{j,t} v_{k,<t} \rangle_{\text{data}} - \langle h_{j,t} v_{k,<t} \rangle_{\text{recon}}\right) \tag{11}$$

$$\Delta a_{i} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} \rangle_{\text{data}} - \langle v_{i,t} \rangle_{\text{recon}}\right) \tag{12}$$

$$\Delta b_{j} \propto \sum_{t=N+1}^{T} \left(\langle h_{j,t} \rangle_{\text{data}} - \langle h_{j,t} \rangle_{\text{recon}}\right) \tag{13}$$

where $\langle \cdot \rangle_{\text{data}}$ is an expectation with respect to the data distribution, and $\langle \cdot \rangle_{\text{recon}}$ is the K-step reconstruction distribution as obtained by alternating Gibbs sampling (i.e. iterating between updating all hiddens using Eq. 6 and all visibles using Eq. 47), starting with the visible units clamped to the training data. Note that we estimate these expectations empirically.

2.5 Free energy

Under the CRBM, the probability of observing a configuration of visible units, $\boldsymbol{v}_t$, given the recent past, $\boldsymbol{v}_{<t}$ can be obtained by marginalizing out the hidden units:

$$\begin{aligned} p(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta) &= \frac{1}{Z(\boldsymbol{v}_{<t},\theta)} \sum_{h_{j,t}\in\{0,1\}^H} \exp\left(-E(\boldsymbol{v}_t, \boldsymbol{h}_t|\boldsymbol{v}_{<t},\theta)\right) && (14) \\ &= \frac{1}{Z(\boldsymbol{v}_{<t},\theta)} \sum_{h_{j,t}\in\{0,1\}^H} \exp\left(-\frac{1}{2}\sum_i (v_{i,t}-\hat{a}_{i,t})^2 + \sum_{ij} W_{ij} v_{i,t} h_{j,t} + \sum_j \hat{b}_{j,t} h_{j,t}\right) && (15) \\ &= \frac{1}{Z(\boldsymbol{v}_{<t},\theta)} \exp\left(-\frac{1}{2}\sum_i (v_{i,t}-\hat{a}_{i,t})^2\right) \sum_{h_{j,t}\in\{0,1\}^H} \exp\left(\sum_{ij} W_{ij} v_{i,t} h_{j,t} + \sum_j \hat{b}_{j,t} h_{j,t}\right) && (16) \end{aligned}$$

where Eq. 16 follows from the properties of exponent rules and the fact that the quadratic term does not depend on the hidden units. Unfortunately $p(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta)$ is intractable to compute since $Z(\boldsymbol{v}_{<t},\theta)$ involves an integration/sum over all possible settings of the visible/hidden units:

$$Z(\boldsymbol{v}_{<t},\theta) = \sum_{h'_{j,t}\in\{0,1\}^H} \int_{\boldsymbol{v}'_t} \exp\big(\big(-E(\boldsymbol{v}'_t, \boldsymbol{h}'_t|\boldsymbol{v}_{<t})\big)\, d\boldsymbol{v}'_t. \tag{17}$$

However, $\log p(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta)$ can be computed up to a constant, which is useful for scoring observations under a fixed model. Even if we ignore $Z(\boldsymbol{v}_{<t},\theta)$ it is not immediately obvious how

[1] We consider a window to be a group of successive frames of data.

to proceed from Eq. 16 because it involves a sum over all possible configurations of the hidden units: $\sum_{h_{j,t}\in\{0,1\}^H} \exp\left(\sum_{ij} W_{ij}v_{i,t}h_{j,t} + \sum_j \hat{b}_{j,t}h_{j,t}\right)$. But the sums within the exponential decouple the hiddens, such that each unit, j, can be considered in turn. Furthermore, since the hidden units are binary, we only need to consider two possible states for each unit. We start by "pulling out" the term corresponding to unit H from the sum, explicitly considering its two states, and leave the terms corresponding to the other units in the sum. We use the notation $\backslash H$ to mean the set of units not including H.

$$\begin{aligned}
&\sum_{h_{j,t}\in\{0,1\}^H} \exp\left(\sum_{ij} W_{ij}v_{i,t}h_{j,t} + \sum_j \hat{b}_{j,t}h_{j,t}\right) \\
&= \sum_{h_{H,t}\in\{0,1\}} \exp\left(\sum_{ij} W_{ij}v_{i,t}h_{j,t} + \sum_j \hat{b}_{j,t}h_{j,t}\right) \sum_{h_{j,t}\in\{0,1\}^{H-1}, j\in\backslash H} \exp\left(\sum_{ij} W_{ij}v_{i,t}h_{j,t} + \sum_j \hat{b}_{j,t}h_{j,t}\right) \qquad (18) \\
&= \left(1+\exp\left(\sum_i W_{iH}v_{i,t} + \hat{b}_{H,t}\right)\right) \sum_{h_{j,t}\in\{0,1\}^{H-1}, j\in\backslash H} \exp\left(\sum_{ij} W_{ij}v_{i,t}h_{j,t} + \sum_j \hat{b}_{j,t}h_{j,t}\right). \qquad (19)
\end{aligned}$$

By applying this same process to the remaining hidden units $H-1, H-2, \ldots, 1$, we get:

$$\sum_{h_{j,t}\in\{0,1\}^H} \exp\left(\sum_{ij} W_{ij}v_{i,t}h_{j,t} + \sum_j \hat{b}_{j,t}h_{j,t}\right) = \prod_j \left(1+\exp\left(\sum_i W_{ij}v_{i,t} + \hat{b}_{j,t}\right)\right). \qquad (20)$$

The "free energy" which is the negative log probability up to a constant can therefore easily be computed:

$$\begin{aligned}
F(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta) &= -\log(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta) + \log Z(\boldsymbol{v}_{<t},\theta) \qquad (21) \\
&= \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_j \log\left(1+\exp\left(\sum_i W_{ij}v_{i,t} + \hat{b}_{j,t}\right)\right) \qquad (22)
\end{aligned}$$

In addition to scoring observations under a model, the free energy can be used to fill in complete or partial observations of $\boldsymbol{v}_t$, given $\boldsymbol{v}_{<t}$ by following its gradient with respect to the visible units:

$$\frac{\partial}{\partial v_{i,t}} F(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta) = (v_{i,t} - \hat{a}_{i,t}) - \sum_j \frac{\partial}{\partial v_{i,t}} \log\left(1+\exp\left(\sum_i W_{ij} v_{i,t} + \hat{b}_{j,t}\right)\right) \tag{23}$$

$$= (v_{i,t} - \hat{a}_{i,t}) - \sum_j \frac{1}{1+\exp\left(\sum_i W_{ij} v_{i,t} + \hat{b}_{j,t}\right)} \frac{\partial}{\partial v_{i,t}} \left(1+\exp\left(\sum_i W_{ij} v_{i,t} + \hat{b}_{j,t}\right)\right) \tag{24}$$

$$= (v_{i,t} - \hat{a}_{i,t}) - \sum_j \frac{1}{1+\exp\left(\sum_i W_{ij} v_{i,t} + \hat{b}_{j,t}\right)} \exp\left(\sum_i W_{ij} v_{i,t} + \hat{b}_{j,t}\right) W_{ij} \tag{25}$$

$$= (v_{i,t} - \hat{a}_{i,t}) - \sum_j \frac{1}{1+\exp\left(-\sum_i W_{ij} v_{i,t} - \hat{b}_{j,t}\right)} W_{ij} \tag{26}$$

$$= v_{i,t} - \left(\hat{a}_{i,t} + \sum_j W_{ij} \sigma(\sum_i W_{ij} v_{i,t} + \hat{b}_{j,t})\right). \tag{27}$$

The gradient with respect to particular visible unit (Eq. 27) has an intuitive form: it is the difference between its current value and the value that would be obtained by mean-field reconstruction.

3 Factored Conditional Restricted Boltzmann Machines

A Factored Conditional Restricted Boltzmann Machine (FCRBM) is a low-rank version of the CRBM where the weight matrix A is factored into a pair of weight matrices $A^{\boldsymbol{v}_{<t}}$ and $A^{\boldsymbol{v}}$, the weight matrix B is factored into a pair of weight matrices $B^{\boldsymbol{v}_{<t}}$ and $B^{\boldsymbol{h}}$, and the weight matrix W is factored into a pair of weight matrices $W^{\boldsymbol{v}}$ and $W^{\boldsymbol{h}}$. We use m, n and f to index the deterministic "factors" assigned to each type of pairwise interaction.

3.1 Energy function

The energy function for a FCRBM is:

$$E(\boldsymbol{v}_t, \boldsymbol{h}_t|\boldsymbol{v}_{<t},\theta) = \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_f \sum_{ij} W^{\boldsymbol{v}}_{if} W^{\boldsymbol{h}}_{jf} v_{i,t} h_{j,t} - \sum_i \hat{b}_{j,t} h_{j,t} \tag{28}$$

$$= \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_f \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} - \sum_i \hat{b}_{j,t} h_{j,t} \tag{29}$$

where the dynamic component of the biases is also factored:

$$\hat{a}_{i,t} = a_i + \sum_m \sum_k A^{\boldsymbol{v}}_{im} A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} = a_i + \sum_m A^{\boldsymbol{v}}_{im} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t}, \tag{30}$$

$$\hat{b}_{j,t} = b_j + \sum_n \sum_k B^{\boldsymbol{h}}_{jn} B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} = b_j + \sum_n B^{\boldsymbol{h}}_{jn} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t}. \tag{31}$$

3.2 Inference

The posterior distribution for a FCRBM is factorial. Given the current and recent past visible states, the probability that a hidden unit, j, is active is given by:

$$p(h_{j,t} = 1|\boldsymbol{v}_t, \boldsymbol{v}_{<t}, \theta) = \sigma\left(\hat{b}_{j,t} + \sum_f W^{\boldsymbol{h}}_{jf} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t}\right). \tag{32}$$

3.3 Reconstruction

Given $\boldsymbol{h}_t$ and $\boldsymbol{v}_{<t}$, the reconstruction distribution at each visible unit is:

$$p(v_{i,t}|\boldsymbol{h_t}, \boldsymbol{v}_{<t}, \theta) = \mathcal{N}\left(\hat{a}_{i,t} + \sum_f W^{\boldsymbol{v}}_{if} \sum_j W^{\boldsymbol{h}}_{if} h_{j,t}, 1\right). \tag{33}$$

3.4 Contrastive divergence learning

The CD updates for the parameters of a FCRBM are given by:

$$\Delta W^{\boldsymbol{v}}_{if} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t}\rangle_{\text{data}} - \langle v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t}\rangle_{\text{recon}}\right) \tag{34}$$

$$\Delta W^{\boldsymbol{h}}_{jf} \propto \sum_{t=N+1}^{T} \left(\langle h_{j,t} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t}\rangle_{\text{data}} - \langle h_{j,t} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t}\rangle_{\text{recon}}\right) \tag{35}$$

$$\Delta A^{\boldsymbol{v}_{<t}}_{km} \propto \sum_{t=N+1}^{T} \left(\langle v_{k,<t} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t}\rangle_{\text{data}} - \langle v_{k,<t} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t}\rangle_{\text{recon}}\right) \tag{36}$$

$$\Delta A^{\boldsymbol{v}}_{im} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t}\rangle_{\text{data}} - \langle v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t}\rangle_{\text{recon}}\right) \tag{37}$$

$$\Delta B^{\boldsymbol{v}_{<t}}_{kn} \propto \sum_{t=N+1}^{T} \left(\langle v_{k,<t} \sum_j B^{\boldsymbol{h}}_{hn} h_{j,t}\rangle_{\text{data}} - \langle v_{k,<t} \sum_j B^{\boldsymbol{h}}_{hn} h_{j,t}\rangle_{\text{recon}}\right) \tag{38}$$

$$\Delta B^{\boldsymbol{h}}_{jn} \propto \sum_{t=N+1}^{T} \left(\langle h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t}\rangle_{\text{data}} - \langle h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t}\rangle_{\text{recon}}\right) \tag{39}$$

The updates for the static biases are given by Eq. 12 and 13.

3.5 Free energy

The free energy for a FCRBM is derived similarly to that of a CRBM:

$$F(\boldsymbol{v}_t|\boldsymbol{v}_{<t}, \theta) = \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_j \log\left(1 + \exp\left(\sum_f W^{\boldsymbol{h}}_{jf} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} + \hat{b}_{j,t}\right)\right). \tag{40}$$

The gradient of the free energy with respect to a visible unit, i is:

$$\frac{\partial}{\partial v_{i,t}} F(\boldsymbol{v}_t|\boldsymbol{v}_{<t},\theta) = v_{i,t} - \left(\hat{a}_{i,t} + \sum_j W^{\boldsymbol{v}}_{if} \sum_f W^{\boldsymbol{h}}_{jf} \sigma(\sum_f W^{\boldsymbol{h}}_{jf} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} + \hat{b}_{j,t}) \right). \tag{41}$$

4 Conditional Restricted Boltzmann Machines with contextual multiplicative interactions

To also capture context (e.g. motion style) we introduce one or more "one-hot"-encoded contextual label units, $\boldsymbol{y}$, which are connected to real-valued, deterministic features, $\boldsymbol{z}$. The features "gate" each existing pairwise interaction in the CRBM. The weight matrices, A, B, and W become tensors, where the third dimension in each corresponds to the feature dimensions, indexed by l.

The relation between labels, $\boldsymbol{y}$, and features, $\boldsymbol{z}$, is linear:

$$z_{l,t} = \sum_p R_{pl} y_{p,t}. \tag{42}$$

Note that we can also provide the continuous features $\boldsymbol{z}_t$ directly, instead of learning the mapping R. In that case, we condition on $\boldsymbol{z}_t$ instead of $\boldsymbol{y}_t$ and do not use Eq. 42.

4.1 Energy function

The energy function of the multiplicative CRBM is:

$$E(\boldsymbol{v}_t, \boldsymbol{h}_t|\boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta) = \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_{ijl} W_{ijl} v_{i,t} h_{j,t} z_{l,t} - \sum_j \hat{b}_{j,t} h_{j,t} \tag{43}$$

where the dynamic biases are:

$$\hat{a}_{i,t} = a_i + \sum_{kl} A_{kil} v_{k,<t} z_{l,t}, \tag{44}$$

$$\hat{b}_{j,t} = b_j + \sum_{kl} B_{kjl} v_{k,<t} z_{l,t}. \tag{45}$$

4.2 Inference

Given $\boldsymbol{v}_t$, $\boldsymbol{v}_{<t}$, and $\boldsymbol{y}_t$, the probability that a hidden unit, j, is active is given by:

$$p(h_{j,t} = 1|\boldsymbol{v}_t, \boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta) = \sigma\left(\hat{b}_{j,t} + \sum_{il} W_{ijl} v_{i,t} z_{l,t} \right). \tag{46}$$

4.3 Reconstruction

Given $\boldsymbol{h}_t$, $\boldsymbol{v}_{<t}$, and $\boldsymbol{y}_t$, the reconstruction distribution at each visible unit, i, is:

$$p(v_{i,t}|\boldsymbol{h_t}, \boldsymbol{v}_{<t}, \theta) = \mathcal{N}\left(\hat{a}_{i,t} + \sum_{jl} W_{ijl} h_{j,t} z_{l,t}, 1 \right) \tag{47}$$

4.4 Contrastive divergence learning

The CD updates for the parameters of the multiplicative CRBM are given by:

$$\Delta W_{ijl} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} h_{j,t} z_{l,t}\rangle_{\text{data}} - \langle v_{i,t} h_{j,t} z_{l,t}\rangle_{\text{recon}}\right), \tag{48}$$

$$\Delta A_{kil} \propto \sum_{t=N+1}^{T} \left(\langle v_{i,t} v_{k,<t} z_{l,t}\rangle_{\text{data}} - \langle v_{i,t} v_{k,<t} z_{l,t}\rangle_{\text{recon}}\right), \tag{49}$$

$$\Delta B_{kjl} \propto \sum_{t=N+1}^{T} \left(\langle h_{j,t} v_{k,<t} z_{l,t}\rangle_{\text{data}} - \langle h_{j,t} v_{k,<t} z_{l,t}\rangle_{\text{recon}}\right), \tag{50}$$

$$\Delta R_{pl} \propto \sum_{t=N+1}^{T} \left(\langle C_{l,t} y_{p,t}\rangle_{\text{data}} - \langle C_{l,t} y_{p,t}\rangle_{\text{recon}}\right) \tag{51}$$

where

$$C_{l,t} = \sum_{ij} W_{ijl} v_{i,t} h_{j,t} + \sum_{ik} A_{kil} v_{i,t} v_{k,<t} + \sum_{jk} B_{kjl} h_{j,t} v_{k,<t}. \tag{52}$$

The updates for the static biases are given by Eq. 12 and 13.

4.5 Free energy

The free energy for the multiplicative CRBM is given by:

$$F(\boldsymbol{v}_t|\boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta) = \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_j \log\left(1 + \exp\left(\sum_{il} W_{ijl} v_{i,t} z_{l,t} + \hat{b}_{j,t}\right)\right) \tag{53}$$

The gradient of the free energy with respect to a visible unit, i is:

$$\frac{\partial}{\partial v_{i,t}} F(\boldsymbol{v}_t|\boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta) = v_{i,t} - \left(\hat{a}_{i,t} + \sum_j \sum_l W_{ijl} z_{l,t} \sigma(\sum_{il} W_{ijl} v_{i,t} z_{l,t} + \hat{b}_{j,t})\right). \tag{54}$$

5 Factored Conditional Restricted Boltzmann Machines with contextual multiplicative interactions

Similar to how we have factored the CRBM, we can also factor the CRBM with contextual, multiplicative interactions. The weight tensor A is factored into a three pairwise interaction matrices: $A^{\boldsymbol{v}_{<t}}$, $A^{\boldsymbol{v}}$, and A^{z}. The weight tensor B is factored into three pairwise interaction matrices: $B^{\boldsymbol{v}_{<t}}$, $B^{\boldsymbol{h}}$, and B^{z}. The weight tensor W is factored into three pairwise interaction matrices: $W^{\boldsymbol{v}}$, $W^{\boldsymbol{h}}$, and W^{z}. Again we use m, n and f to index the deterministic "factors" assigned to each type of pairwise interaction. The superscripts on the weights indicate to which type of unit the factors connect. Note that we do not factor the feature weights, R, (although we could) so Eq. 42 does not change.

5.1 Energy function

The energy function of the factored, multiplicative CRBM is:

$$E\left(\boldsymbol{v}_t, \boldsymbol{h}_t | \boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta\right) = \frac{1}{2}\sum_i \left(v_{i,t} - \hat{a}_{i,t}\right)^2 - \sum_f \sum_{ijl} W^{\boldsymbol{v}}_{if} W^{\boldsymbol{h}}_{jf} W^{\boldsymbol{z}}_{lf} v_{i,t} h_{j,t} z_{l,t} - \sum_j \hat{b}_{j,t} h_{j,t} \tag{55}$$

$$= \frac{1}{2}\sum_i \left(v_{i,t} - \hat{a}_{i,t}\right)^2 - \sum_f \left(\sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \right) - \sum_j \hat{b}_{j,t} h_{j,t}. \tag{56}$$

where the dynamic biases are:

$$\hat{a}_{i,t} = a_i + \sum_m \sum_{kl} A^{\boldsymbol{v}}_{im} A^{\boldsymbol{v}_{<t}}_{km} A^{\boldsymbol{z}}_{lm} v_{k,<t} z_{l,t} = a_i + \sum_m \left(A^{\boldsymbol{v}}_{im} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} \sum_l A^{\boldsymbol{z}}_{lm} z_{l,t} \right), \tag{57}$$

$$\hat{b}_{j,t} = b_j + \sum_n \sum_{kl} B^{\boldsymbol{h}}_{jn} B^{\boldsymbol{v}_{<t}}_{kn} B^{\boldsymbol{z}}_{ln} v_{k,<t} z_{l,t} = b_j + \sum_n \left(B^{\boldsymbol{h}}_{jn} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} \sum_l B^{\boldsymbol{z}}_{ln} z_{l,t} \right). \tag{58}$$

Note that the dynamic component of Eq. 57 and Eq. 58 is simply the total input to the visible/hidden unit via the factors. The total input is a three-way product between the input to the factors (coming from the past and from the style features) and the weight from the factors to the visible/hidden unit.

5.2 Inference

Given $\boldsymbol{v}_t$, $\boldsymbol{v}_{<t}$, and $\boldsymbol{y}_t$, the probability that a hidden unit, j, is active is given by:

$$p(h_{j,t} = 1 | \boldsymbol{v}_t, \boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta) = \sigma\left(\hat{b}_{j,t} + \sum_f \left(W^{\boldsymbol{h}}_{jf} \sum_i W^{\boldsymbol{v}}_{ij} v_{i,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t}) \right) \right) \tag{59}$$

5.3 Reconstruction

Given $\boldsymbol{h}_t$, $\boldsymbol{v}_{<t}$, and $\boldsymbol{y}_t$, the reconstruction distribution at each visible unit, i, is:

$$p(v_{i,t} | \boldsymbol{h}_t, \boldsymbol{v}_{<t}, \boldsymbol{y}_t, \theta) = \mathcal{N}\left(\hat{a}_{i,t} + \sum_f \left(W^{\boldsymbol{v}}_{if} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \right), 1 \right) \tag{60}$$

5.4 Contrastive divergence learning

The CD updates for the parameters of the factored, multiplicative CRBM have an intuitive form. The gradient with respect to a weight that connects a unit to a factor is the difference of two expectations of products. Each product involves three terms: the activity of the respective unit, and the total input to the factor from each of the two other sets of units involved in the three-way relationship:

$$\Delta W^{\boldsymbol{v}}_{if} \propto \sum_t \left(\langle v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \rangle_{\text{data}} - \langle v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \rangle_{\text{recon}} \right), \tag{61}$$

$$\Delta W^{\boldsymbol{h}}_{jf} \propto \sum_t \left(\langle h_{j,t} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \rangle_{\text{data}} - \langle h_{j,t} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \rangle_{\text{recon}} \right), \tag{62}$$

$$\Delta W^{\boldsymbol{z}}_{lf} \propto \sum_t \left(\langle z_{l,t} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \rangle_{\text{data}} - \langle z_{l,t} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \rangle_{\text{recon}} \right), \tag{63}$$

$$\Delta A^{\boldsymbol{v}}_{im} \propto \sum_t \left(\langle v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} \sum_l A^{\boldsymbol{z}}_{lm} z_{l,t} \rangle_{\text{data}} - \langle v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} \sum_l A^{\boldsymbol{z}}_{lm} z_{l,t} \rangle_{\text{recon}} \right), \tag{64}$$

$$\Delta A^{\boldsymbol{v}_{<t}}_{km} \propto \sum_t \left(\langle v_{k,<t} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t} \sum_l A^{\boldsymbol{z}}_{lm} z_{l,t} \rangle_{\text{data}} - \langle v_{k,<t} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t} \sum_l A^{\boldsymbol{z}}_{lm} z_{l,t} \rangle_{\text{recon}} \right), \tag{65}$$

$$\Delta A^{\boldsymbol{z}}_{lm} \propto \sum_t \left(\langle z_{l,t} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} \rangle_{\text{data}} - \langle z_{l,t} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} \rangle_{\text{recon}} \right), \tag{66}$$

$$\Delta B^{\boldsymbol{h}}_{jn} \propto \sum_t \left(\langle h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} \sum_l B^{\boldsymbol{z}}_{ln} z_{l,t} \rangle_{\text{data}} - \langle h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} \sum_l B^{\boldsymbol{z}}_{ln} z_{l,t} \rangle_{\text{recon}} \right), \tag{67}$$

$$\Delta B^{\boldsymbol{v}_{<t}}_{kn} \propto \sum_t \left(\langle v_{k,<t} \sum_j B^{\boldsymbol{h}}_{jn} h_{j,t} \sum_l B^{\boldsymbol{z}}_{ln} z_{l,t} \rangle_{\text{data}} - \langle v_{k,<t} \sum_j B^{\boldsymbol{h}}_{jn} h_{j,t} \sum_l B^{\boldsymbol{z}}_{ln} z_{l,t} \rangle_{\text{recon}} \right), \tag{68}$$

$$\Delta B^{\boldsymbol{z}}_{ln} \propto \sum_t \left(\langle z_{l,t} \sum_j B^{\boldsymbol{h}}_{jn} h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} \rangle_{\text{data}} - \langle z_{l,t} \sum_j B^{\boldsymbol{h}}_{jn} h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} \rangle_{\text{recon}} \right), \tag{69}$$

$$\Delta R_{pl} \propto \sum_{t=N+1}^{T} \left(\langle C_{l,t} y_{p,t} \rangle_{\text{data}} - \langle C_{l,t} y_{p,t} \rangle_{\text{recon}} \right) \tag{70}$$

where

$$\begin{aligned} C_{l,t} = {} & \sum_f \left(W^{\boldsymbol{z}}_{lf} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_j W^{\boldsymbol{h}}_{jf} h_{j,t} \right) + \sum_m \left(A^{\boldsymbol{z}}_{lm} \sum_i A^{\boldsymbol{v}}_{im} v_{i,t} \sum_k A^{\boldsymbol{v}_{<t}}_{km} v_{k,<t} \right) \\ & + \sum_n \left(B^{\boldsymbol{z}}_{ln} \sum_j B^{\boldsymbol{h}}_{jn} h_{j,t} \sum_k B^{\boldsymbol{v}_{<t}}_{kn} v_{k,<t} \right). \end{aligned} \tag{71}$$

The updates for the static biases are given by Eq. 12 and 13.

5.5 Free energy

The free energy for the factored, multiplicative CRBM is given by:

$$F(\boldsymbol{v}_t | \boldsymbol{v}_{<t}, \theta) = \frac{1}{2} \sum_i \left(v_{i,t} - \hat{a}_{i,t} \right)^2 - \sum_j \log \left(1 + \exp D_{j,t} \right). \tag{72}$$

where $D_{j,t}$ is the total input to a hidden unit, j, at time t:

$$D_{j,t} = \sum_f \left(W^{\boldsymbol{h}}_{jf} \sum_i W^{\boldsymbol{v}}_{if} v_{i,t} \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \right) + \hat{b}_{j,t} \tag{73}$$

The gradient of the free energy with respect to a visible unit, *i* is:

$$\frac{\partial}{\partial v_{i,t}} F(\boldsymbol{v}_t | \boldsymbol{v}_{<t}, \theta) = v_{i,t} - \left(\hat{a}_{i,t} + \sum_j W^{\boldsymbol{v}}_{if} \sum_f \left(W^{\boldsymbol{h}}_{jf} \sigma \left(D_{j,t} \right) \sum_l W^{\boldsymbol{z}}_{lf} z_{l,t} \right) \right).$$