A. Supplementary Materials for *Bidirectional Learning for Time-series Models with Hidden Units*

Here, we derive specific learning rules suggested by (27)-(28) as well as those with approximation with (29). These learning rule can be derived in a way similar to the learning rules (18)-(22) are derived from (17). We also provide some of the details, which are omitted in the derivation of (18)-(22).

The learning rules for $\mathbf{U}$ and $\mathbf{Z}$ are derived from (27)-(28) as follows:

$$\mathbf{U}^{[d]} \leftarrow \mathbf{U}^{[d]} + \eta \, \log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \boldsymbol{\alpha}^{[s-1]} \, (\mathbf{h}^{[s]} - \langle \mathbf{H}^{[\mathbf{s}]} \rangle_\phi)^\top \tag{32}$$

$$\mathbf{Z}^{[d]} \leftarrow \mathbf{Z}^{[d]} + \eta \, \log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \boldsymbol{\beta}^{[s-1]} \, (\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_\phi)^\top \tag{33}$$

$$\mathbf{U}^{[\delta]} \leftarrow \mathbf{U}^{[\delta]} + \eta \, \log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \mathbf{x}^{[s-\delta]} \, (\mathbf{h}^{[s]} - \langle \mathbf{H}^{[\mathbf{s}]} \rangle_\phi)^\top \tag{34}$$

$$\mathbf{Z}^{[\delta]} \leftarrow \mathbf{Z}^{[\delta]} + \eta \, \log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \mathbf{h}^{[s-\delta]} \, (\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_\phi)^\top \tag{35}$$

for $1 \le \delta < d$, where $\langle \mathbf{H}^{[s]} \rangle_\phi$ denotes the expected values of $\mathbf{h}^{[s]}$ with respect to the conditional distribution given by the following p_ϕ:

$$p_\phi(\mathbf{h}^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = \frac{1}{Z'} \exp(-E_\phi(\mathbf{h}^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]})) \tag{36}$$

for any binary vectors $\mathbf{h}^{[s]}$, where Z' is a normalization factor for the probabilities to sum up to one, and

$$E_\phi(\mathbf{h}^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = -\sum_{\delta=1}^{d-1} (\mathbf{x}^{[s-\delta]})^\top \mathbf{U}^{[\delta]} \, \mathbf{h}^{[s]} - \sum_{\delta=1}^{d-1} (\mathbf{h}^{[s-\delta]})^\top \mathbf{Z}^{[\delta]} \, \mathbf{h}^{[s]} - (\boldsymbol{\alpha}^{[s-1]})^\top \, \mathbf{U}^{[d]} \mathbf{h}^{[s]} - (\boldsymbol{\beta}^{[s-1]})^\top \, \mathbf{Z}^{[d]} \mathbf{h}^{[s]}. \tag{37}$$

The energy in (37) can be decomposed into the energy associated with each hidden unit j as follows:

$$E_\phi(\mathbf{h}^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = \sum_{j \in \mathcal{H}} E_{\phi,j}(h_j^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) \tag{38}$$

where $\mathcal{H}$ denotes the set of hidden units, and

$$E_{\phi,j}(h_j^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = -\sum_{\delta=1}^{d-1} (\mathbf{x}^{[s-\delta]})^\top \mathbf{U}_{:,j}^{[\delta]} \, h_j^{[s]} - \sum_{\delta=1}^{d-1} (\mathbf{h}^{[s-\delta]})^\top \mathbf{Z}_{:,j}^{[\delta]} \, h_j^{[s]} - (\boldsymbol{\alpha}^{[s-1]})^\top \mathbf{U}_{:,j}^{[d]} \, h_j^{[s]} - (\boldsymbol{\beta}^{[s-1]})^\top \mathbf{Z}_{:,j}^{[d]} \, h_j^{[s]}, \tag{39}$$

where $\mathbf{U}_{:,j}$ denotes a column vector corresponding to the j-th column of $\mathbf{U}$, and $\mathbf{Z}_{:,j}$ is defined analogously.

Then (36) can be expressed as

$$p_\phi(\mathbf{h}^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = \prod_{j \in \mathcal{H}} p_{\phi,j}(h_j^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}), \tag{40}$$

where

$$p_{\phi,j}(h_j^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = \frac{\exp(-E_{\phi,j}(h_j^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}))}{\exp(-E_{\phi,j}(h_j^{[s]} = 0 | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]})) + \exp(-E_{\phi,j}(h_j^{[s]} = 1 | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}))} \tag{41}$$

$$= \frac{\exp(-E_{\phi,j}(h_j^{[s]} | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}))}{1 + \exp(-E_{\phi,j}(h_j^{[s]} = 1 | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}))}. \tag{42}$$

The j-th element of $\langle \mathbf{H}^{[s]} \rangle_\phi$ is then given by

$$\langle H_j^{[s]} \rangle_\phi = p_{\phi,j}(h_j^{[s]} = 1 | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) \tag{43}$$

In (32)-(35), the value of $\langle \mathbf{H}^{[s]} \rangle_\phi$ is computed with the latest values of ϕ. Let $\phi^{[t-1]}$ be the value of ϕ immediately before step t. With the recursive computation of (29), the learning rules of (32)-(35) are approximated with the following learning rules:

$$\mathbf{U}^{[d]} \leftarrow \mathbf{U}^{[d]} + \eta\,(1-\gamma)\,\log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \gamma^{t-1-s}\,\boldsymbol{\alpha}^{[s-1]}\,(\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_{\phi^{[s-1]}})^\top \tag{44}$$

$$\mathbf{Z}^{[d]} \leftarrow \mathbf{Z}^{[d]} + \eta\,(1-\gamma)\,\log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \gamma^{t-1-s}\,\boldsymbol{\beta}^{[s-1]}\,(\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_{\phi^{[s-1]}})^\top \tag{45}$$

$$\mathbf{U}^{[\delta]} \leftarrow \mathbf{U}^{[\delta]} + \eta\,(1-\gamma)\,\log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \gamma^{t-1-s}\,\mathbf{x}^{[s-\delta]}\,(\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_{\phi^{[s-1]}})^\top \tag{46}$$

$$\mathbf{Z}^{[\delta]} \leftarrow \mathbf{Z}^{[\delta]} + \eta\,(1-\gamma)\,\log p_\theta(\mathbf{x}^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) \sum_{s=\ell}^{t-1} \gamma^{t-1-s}\,\mathbf{h}^{[s-\delta]}\,(\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_{\phi^{[s-1]}})^\top \tag{47}$$

for $1 \le \delta < d$, where the quantity such as

$$G'_{t-1} \equiv \sum_{s=\ell}^{t-1} \gamma^{t-1-s}\,\boldsymbol{\alpha}^{[s-1]}\,(\mathbf{h}^{[s]} - \langle \mathbf{H}^{[s]} \rangle_{\phi^{[s-1]}})^\top \tag{48}$$

can be computed recursively as

$$G'_t \leftarrow \gamma\, G'_{t-1} + (1-\gamma)\,\boldsymbol{\alpha}^{[t-1]}\,(\mathbf{h}^{[t]} - \langle \mathbf{H}^{[t]} \rangle_{\phi^{[t-1]}})^\top. \tag{49}$$

One may consider real-valued units as well (Dasgupta & Osogami, 2017; Osogami, 2016). For example, each of $x_i^{[t]}$ and $h_j^{[t]}$ may have a Gaussian distribution with the following density:

$$p_{\theta,i}(x_i^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) = \frac{1}{\sqrt{2\,\pi\,\sigma_i^2}} \exp\left(- \frac{\left(x_i^{[t]} - E_{\theta,i}(x_i^{[t]} = 1 | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]})\right)^2}{2\,\sigma_i^2} \right) \tag{50}$$

$$p_{\phi,i}(h_j^{[t]} | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]}) = \frac{1}{\sqrt{2\,\pi\,\sigma_j^2}} \exp\left(- \frac{\left(h_j^{[t]} - E_{\phi,j}(h_j^{[t]} = 1 | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]})\right)^2}{2\,\sigma_j^2} \right), \tag{51}$$

where σ_i^2 and σ_j^2 are variance parameters, $E_{\phi,j}$ is given by (39), and $E_{\theta,i}(x_i^{[t]} = 1 | \mathbf{x}^{[<t]}, \mathbf{h}^{[<t]})$ is given by

$$E_{\theta,i}(x_i^{[s]} = 1 | \mathbf{x}^{[<s]}, \mathbf{h}^{[<s]}) = -b_i - \sum_{\delta=1}^{d-1} (\mathbf{x}^{[s-\delta]})^\top \mathbf{W}_{:,i}^{[\delta]} - \sum_{\delta=1}^{d-1} (\mathbf{h}^{[s-\delta]})^\top \mathbf{V}_{:,i}^{[\delta]} - (\boldsymbol{\alpha}^{[s-1]})^\top \mathbf{W}_{:,i}^{[d]} - (\boldsymbol{\beta}^{[s-1]})^\top \mathbf{V}_{:,i}^{[d]}. \tag{52}$$