Supplementary material for the paper structured recurrent temporal restricted Boltzmann machines

Roni Mittelman, Benjamin Kuipers, Silvio Savarese, Honglak Lee

June, 2014

1 Problem title

The RTRBM describes the joint probability distribution of the visible units vector $\mathbf{v}_t \in \mathbb{R}^{N_v}$, and hidden units vector $\mathbf{h}_t \in \mathbb{R}^{N_h}$ at time step t, using a conditional RBM which depends on the hidden input $\mathbf{r}_{t-1}$. The RTRBM network is illustrated in Figure **??**. The joint probability distribution of $\mathbf{v}_t, \mathbf{h}_t$ for any $t > 1$ (given $\mathbf{r}_{t-1}$) takes the following form:

$$P(\mathbf{v}_t, \mathbf{h}_t; \mathbf{r}_{t-1}) = \frac{1}{Z_{r_{t-1}}} \exp\{-E(\mathbf{v}_t, \mathbf{h}_t; \mathbf{r}_{t-1})\} \tag{1.1}$$
$$E(\mathbf{v}_t, \mathbf{h}_t; \mathbf{r}_{t-1}) = -(\mathbf{h}_t^\top \mathbf{W}\mathbf{v}_t + \mathbf{c}^\top \mathbf{v}_t + \mathbf{b}^\top \mathbf{h}_t + \mathbf{h}_t^\top \mathbf{U}\mathbf{r}_{t-1}),$$

where $\mathbf{W} \in \mathbb{R}^{N_h \times N_v}$, $\mathbf{U} \in \mathbb{R}^{N_h \times N_h}$, $\mathbf{c} \in \mathbb{R}^{N_v}$, $\mathbf{b} \in \mathbb{R}^{N_h}$ are model parameters, and $Z_{\mathbf{r}_{t-1}}$ denotes a normalization factor which depends on $\mathbf{r}_{t-1}$ and the other model parameters (we used the subscript $\mathbf{r}_{t-1}$ since the dependency on the input $\mathbf{r}_{t-1}$ is the major difference compared to the RBM). For $t = 1$, the joint distribution of $\mathbf{v}_1, \mathbf{h}_1$ takes the form of a standard RBM with the hidden units biases $\mathbf{b}_{init} \in \mathbb{R}^{N_h}$.
The inputs $\mathbf{r}_t$ (where $t \in \{1, \ldots, T-1\}$) are obtained from a RNN given $\mathbf{v}_1, ..., \mathbf{v}_t$:

$$\mathbf{r}_t = \begin{cases} \sigma(\mathbf{W}\mathbf{v}_t + \mathbf{b} + \mathbf{U}\mathbf{r}_{t-1}), & \text{if } t > 1 \\ \sigma(\mathbf{W}\mathbf{v}_t + \mathbf{b}_{init}), & \text{if } t = 1 \end{cases} \tag{1.2}$$

where the logistic function $\sigma(x) = (1 + \exp(-x))^{-1}$ is applied to each element of its argument vector. The motivation for the choice of $\mathbf{r}_t$ is that using the RBM associated with time instant t, we have that $\mathbb{E}[\mathbf{h}_t | \mathbf{v}_t] = \mathbf{r}_t$; i.e., it is the expected value of the hidden units vector.

The joint probability distribution of the visible and hidden units of the RTRBM with length T takes the form:

$$P(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1}) = P(\mathbf{v}_1, \mathbf{h}_1) \prod_{t=2}^{T} P(\mathbf{v}_t, \mathbf{h}_t; \mathbf{r}_{t-1}) = \frac{\exp\{E(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1})\}}{Z \cdot Z_{\mathbf{r}_1} \cdots Z_{\mathbf{r}_{T-1}}} \tag{1.3}$$

where Z denotes the normalization factor for the first RBM at $t = 1$, and where

$$E(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1}) = -\Big(\mathbf{h}_1^\top \mathbf{W}\mathbf{v}_1 + \mathbf{c}^\top \mathbf{v}_1 + \mathbf{b}_{init}^\top \mathbf{h}_1 + \sum_{t=2}^{T} \Big(\mathbf{h}_t^\top \mathbf{W}\mathbf{v}_t + \mathbf{c}^\top \mathbf{v}_t + \mathbf{b}^\top \mathbf{h}_t + \mathbf{h}_t^\top \mathbf{U}\mathbf{r}_{t-1}\Big)\Big) \tag{1.4}$$

Since $Z, Z_{\mathbf{r}_1} \ldots Z_{\mathbf{r}_{T-1}}$ are independent of $\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1}$, we can rewrite (1.3) using

$$P(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1}) = \frac{\exp\{E(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1})\}}{\bar{Z}} \tag{1.5}$$

where $\bar{Z} = \sum_{\{\mathbf{v}'_t, \mathbf{h}_t\}_{t=1}^T} E(\{\mathbf{v}'_t, \mathbf{h}_t\}_{t=1}^T; \{\mathbf{r}_t\}_{t=1}^{T-1})$. Note that $\{\mathbf{r}_t\}_{t=1}^{T-1}$ are not a function of $\{\mathbf{v}'_t\}_{t=1}^T$.

1.1 Inference in the RTRBM

Given the hidden inputs $\mathbf{r}_{t-1}$ $(t > 1)$, the conditional distributions are factorized and take the form:

$$\begin{aligned} P(h_{t,j} = 1 | \mathbf{v}_t, \mathbf{r}_{t-1}) &= \sigma(\sum_i w_{j,i} v_{t,i} + b_j + \sum_l u_{j,m} r_{t-1,m}), \\ P(v_{t,i} = 1 | \mathbf{h}_t, \mathbf{r}_{t-1}) &= \sigma(\sum_j w_{j,i} h_{t,j} + c_i), \end{aligned} \tag{1.6}$$

For $t = 1$, the posterior of $\mathbf{h}_1$ given $\mathbf{v}_1$ has the same for as in RBM, where $\mathbf{b}_{init}$ replaces $\mathbf{b}$. The above conditional probabilities can also be used to generate samples $\mathbf{v}_1, ..., \mathbf{v}_T$ (i.e., by repeatedly running Gibbs sampling for $t = 1, ..., T$).
Further, note that, given the hidden inputs $\mathbf{r}_1, ..., \mathbf{r}_{T-1}$, all the RBMs (corresponding to different time frames) are decoupled; thus, sampling can be performed using block Gibbs sampling for each RBM independently. This fact is useful in deriving the CD approximation for the RTRBM.

1.2 Learning in the RTRBM

In order to learn the parameters, we need to obtain the partial derivatives of the log-likelihood, $\log P(\mathbf{v}_1, ..., \mathbf{v}_T)$, with respect to the model parameters. Using the CD approximation to compute these derivatives requires the gradients of energy function (1.4) with respect to all the model parameters. We separate the energy function into the following two terms: $E = -\mathcal{H} - \mathcal{Q}_2$, where

$$\mathcal{H} = \mathbf{h}_1^\top \mathbf{W}\mathbf{v}_1 + \mathbf{c}^\top \mathbf{v}_1 + \mathbf{b}_{init}^\top \mathbf{h}_1 + \sum_{t=2}^{T} \mathbf{h}_t^\top \mathbf{W}\mathbf{v}_t + \mathbf{c}\mathbf{v}_t + \mathbf{b}^\top \mathbf{h}_t, \tag{1.7}$$

$$\mathcal{Q}_2 = \sum_{t=2}^{T} \mathbf{h}_t^T \mathbf{U}\mathbf{r}_{t-1}. \tag{1.8}$$

Taking the gradients of $\mathcal{H}$ with respect to the model parameters is straightforward, and therefore we focus on $\mathcal{Q}_2$. To compute the partial derivative of $\mathcal{Q}_2$ with respect to a model parameter θ, we first compute the gradient of $\mathcal{Q}_2$ with respect to $\mathbf{r}_t$, for $t = 1, \ldots, T-1$, which can be computed recursively using the backpropagation-through-time (BPTT) algorithm. These gradients are then used with the chain rule to compute the derivatives with respect to all the model parameters. In the next subsection, we use the BPTT to derive the gradient with respect to $\mathbf{U}$. The other gradients can be computed similarly.

1.2.1 Calculating $\nabla_{\mathbf{r}_t} \mathcal{Q}_2$ using BPTT

We observe that $\mathcal{Q}_2$ can be computed recursively using:

$$\mathcal{Q}_t = \sum_{\tau=t}^{T} \mathbf{h}_\tau^\top \mathbf{U}\mathbf{r}_{\tau-1} = \mathcal{Q}_{t+1} + \mathbf{h}_t^\top \mathbf{U}\mathbf{r}_{t-1} \tag{1.9}$$

where $\mathcal{Q}_{T+1} = 0$. Using the chain rule and (1.12), we have

$$\begin{aligned} \frac{\partial}{\partial r_{t,m}} \mathcal{Q}_{t+1} &= \sum_{m'} \frac{\partial \mathcal{Q}_{t+2}}{\partial r_{t+1,m'}} \frac{\partial r_{t+1,m'}}{\partial r_{t,m}} + \frac{\partial}{\partial r_{t,m}} (\mathbf{h}_{t+1}^\top \mathbf{U}\mathbf{r}_t) \\ &= \sum_{m'} \frac{\partial \mathcal{Q}_{t+2}}{\partial r_{t+1,m'}} r_{t+1,m'}(1 - r_{t+1,m'}) u_{m',m} + \sum_{m'} h_{t+1,m'} u_{m',m} \end{aligned} \tag{1.10}$$

where $r_{t+1,m'}(1 - r_{t+1,m'})$ is obtained from the partial derivative of the sigmoid function. Equation (1.13) can also be expressed in vector form using:

$$\nabla_{\mathbf{r}_t} \mathcal{Q}_{t+1} = \mathbf{U}^\top (\nabla_{\mathbf{r}_{t+1}} \mathcal{Q}_{t+2} \odot \mathbf{r}_{t+1} \odot (\mathbf{1} - \mathbf{r}_{t+1}) + \mathbf{h}_{t+1}), \tag{1.11}$$

where $\odot$ denotes element-wise product. Since $\mathcal{Q}_{t+1}$ is not a function of $\mathbf{r}_1, \ldots, \mathbf{r}_{t-1}$, we have that $\nabla_{\mathbf{r}_t} \mathcal{Q}_2 = \nabla_{\mathbf{r}_t} \mathcal{Q}_{t+1}$, and therefore the necessary partial derivatives can be computed recursively using (1.14).

1.2.2 Calculating $\nabla_{\mathbf{r}_t} \mathcal{Q}_2$ using BPTT

We observe that $\mathcal{Q}_2$ can be computed recursively using:

$$\mathcal{Q}_t = \sum_{\tau=t}^{T} \mathbf{h}_\tau^\top \mathbf{U}\mathbf{r}_{\tau-1} = \mathcal{Q}_{t+1} + \mathbf{h}_t^\top \mathbf{U}\mathbf{r}_{t-1} \tag{1.12}$$

where $\mathcal{Q}_{T+1} = 0$. Using the chain rule and (1.12), we have

$$\begin{aligned}\frac{\partial}{\partial r_{t,m}}\mathcal{Q}_{t+1} &= \sum_{m'}\frac{\partial \mathcal{Q}_{t+2}}{\partial r_{t+1,m'}}\frac{\partial r_{t+1,m'}}{\partial r_{t,m}} + \frac{\partial}{\partial r_{t,m}}(\mathbf{h}_{t+1}^\top \mathbf{U}\mathbf{r}_t) \\ &= \sum_{m'}\frac{\partial \mathcal{Q}_{t+2}}{\partial r_{t+1,m'}} r_{t+1,m'}(1-r_{t+1,m'})u_{m',m} \\ &+ \sum_{m'} h_{t+1,m'}u_{m',m} \end{aligned} \tag{1.13}$$

where $r_{t+1,m'}(1-r_{t+1,m'})$ is obtained from the partial derivative of the sigmoid function. Equation (1.13) can also be expressed in vector form using:

$$\nabla_{\mathbf{r}_t}\mathcal{Q}_{t+1} = \mathbf{U}^\top(\nabla_{\mathbf{r}_{t+1}}\mathcal{Q}_{t+2} \odot \mathbf{r}_{t+1} \odot (\mathbf{1}-\mathbf{r}_{t+1}) + \mathbf{h}_{t+1}), \tag{1.14}$$

where $\odot$ denotes element-wise product. Since $\mathcal{Q}_{t+1}$ is not a function of $\mathbf{r}_1, \ldots, \mathbf{r}_{t-1}$, we have that $\nabla_{\mathbf{r}_t}\mathcal{Q}_2 = \nabla_{\mathbf{r}_t}\mathcal{Q}_{t+1}$, and therefore the necessary partial derivatives can be computed recursively using (1.14).

1.2.3 Calculating the partial derivatives with respect to the model parameters

In order to compute the derivatives with respect to $\mathbf{U}$, we use the chain rule and (1.12):

$$\begin{aligned}\frac{\partial \mathcal{Q}_2}{\partial u_{m,m'}} &= \sum_{t=2}^{T}\left(\frac{\partial \mathcal{Q}_{t+1}}{\partial r_{t,m}}\frac{\partial r_{t,m}}{\partial u_{m,m'}} + \frac{\partial}{\partial u_{m,m'}}(\mathbf{h}_t^\top \mathbf{U}\mathbf{r}_{t-1})\right) \\ &= \sum_{t=2}^{T}\left(\frac{\partial \mathcal{Q}_{t+1}}{\partial r_{t,m}} r_{t,m}(1-r_{t,m}) + h_{t,m}\right) r_{t-1,m'} \end{aligned} \tag{1.15}$$

where when taking the derivative of $\mathbf{h}_t^\top \mathbf{U}\mathbf{r}_{t-1}$ with respect to $u_{m,m'}$ we regard $\mathbf{r}_{t-1}$ as a constant, since the contribution of its derivative is factored in through $\nabla_{\mathbf{r}_{t-1}}\mathcal{Q}_t$.
Next, we use the CD approximation with (1.13), to show that the update rule for $\mathbf{U}$, that is related to $\mathcal{Q}_2$ (not including $\mathcal{H}$), takes the form:

$$\Delta_{\mathbf{U}}^{\mathcal{Q}_2} = \sum_{t=2}^{T}\left(\mathcal{D}_{t+1} \odot \mathbf{r}_t \odot (\mathbf{1}-\mathbf{r}_t) + \mathbb{E}_{\mathbf{h}_t|\mathbf{v}_t,\mathbf{r}_{t-1}}[\mathbf{h}_t] - \mathbb{E}_{\mathbf{v}'_t,\mathbf{h}_t|\mathbf{r}_{t-1}}[\mathbf{h}_t]\right)\mathbf{r}_{t-1}^\top \tag{1.16}$$

where

$$\mathcal{D}_t = \mathbb{E}_{\mathbf{h}_t,\ldots,\mathbf{h}_T|\mathbf{v}_t,\ldots\mathbf{v}_T,\mathbf{r}_1\ldots,\mathbf{r}_{T-1}}[\nabla_{\mathbf{r}_{t-1}}\mathcal{Q}_t] - \mathbb{E}_{\mathbf{h}_t,\ldots,\mathbf{h}_T,\mathbf{v}'_t,\ldots\mathbf{v}'_T|\mathbf{r}_1\ldots,\mathbf{r}_{T-1}}[\nabla_{\mathbf{r}_{t-1}}\mathcal{Q}_t]. \tag{1.17}$$

Proof. The partial derivative of $\mathcal{Q}_2$ with respect to $u_{m,m'}$ takes the form:

$$\frac{\partial \mathcal{Q}_2}{\partial u_{m,m'}} = \sum_{t=2}^{T} h_{t,m}r_{t-1,m'} + \sum_{t=2}^{T}\frac{\partial \mathcal{Q}_2}{\partial r_{t,m}}\frac{\partial r_{t,m}}{\partial u_{m,m'}} = \sum_{t=2}^{T} h_{t,m}r_{t-1,m'} + \sum_{t=2}^{T}\frac{\partial \mathcal{Q}_2}{\partial r_{t,m}} r_{t,m}(1-r_{t,m})r_{t-1,m'} \tag{1.18}$$

The derivative of the log-likelihood with respect to $u_{m,m'}$ takes the form

$$\Delta^{\mathcal{H}+\mathcal{Q}_2}_{u_{m,m'}} = \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T|\mathbf{v}_1,...,\mathbf{v}_T,\mathbf{r}_1,...,\mathbf{r}_{T-1}}\left[\frac{\partial}{\partial u_{m,m'}}(\mathcal{H}+\mathcal{Q}_2)\right] - \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T,\mathbf{v}'_1,...,\mathbf{v}'_T;\mathbf{r}_1,...,\mathbf{r}_{T-1}}\left[\frac{\partial}{\partial u_{m,m'}}(\mathcal{H}+\mathcal{Q}_2)\right]$$
$$= \Delta^{\mathcal{H}}_{u_{m,m'}} + \Delta^{\mathcal{Q}_2}_{u_{m,m'}}, \tag{1.19}$$

where we defined

$$\Delta^{\mathcal{H}}_{u_{m,m'}} = \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T|\mathbf{v}_1,...,\mathbf{v}_T,\mathbf{r}_1,...,\mathbf{r}_{T-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{H}] - \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T,\mathbf{v}'_1,...,\mathbf{v}'_T;\mathbf{r}_1,...,\mathbf{r}_{T-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{H}]$$
$$\Delta^{\mathcal{Q}_2}_{u_{m,m'}} = \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T|\mathbf{v}_1,...,\mathbf{v}_T,\mathbf{r}_1,...,\mathbf{r}_{T-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{Q}_2] - \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T,\mathbf{v}'_1,...,\mathbf{v}'_T;\mathbf{r}_1,...,\mathbf{r}_{T-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{Q}_2] \tag{1.20}$$

Using (1.18) in $\Delta^{\mathcal{Q}_2}_{u_{m,m'}}$ we have that

$$\Delta^{\mathcal{Q}_2}_{u_{m,m'}} = \sum_{t=2}^{T}\left(\mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T|\mathbf{v}_1,...,\mathbf{v}_T,\mathbf{r}_1,...,\mathbf{r}_{T-1}}[h_{t,m}] - \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T,\mathbf{v}'_1,...,\mathbf{v}'_T;\mathbf{r}_1,...,\mathbf{r}_{T-1}}[h_{t,m}]\right) r_{t-1,m'}$$
$$+ \sum_{t=2}^{T}\left(\mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T|\mathbf{v}_1,...,\mathbf{v}_T,\mathbf{r}_1,...,\mathbf{r}_{T-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{Q}_2] - \mathbb{E}_{\mathbf{h}_1,...,\mathbf{h}_T,\mathbf{v}'_1,...,\mathbf{v}'_T;\mathbf{r}_1,...,\mathbf{r}_{T-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{Q}_2]\right)$$
$$\times r_{t,m}(1-r_{t,m})r_{t-1,m'}$$
$$= \sum_{t=2}^{T}\left(\mathbb{E}_{\mathbf{h}_t|\mathbf{v}_t,\mathbf{r}_{t-1}}[h_{t,m}] - \mathbb{E}_{\mathbf{h}_t,\mathbf{v}'_t;\mathbf{r}_{t-1}}[h_{t,m}]\right) r_{t-1,m'}$$
$$+ \sum_{t=2}^{T}\left(\mathbb{E}_{\mathbf{h}_t|\mathbf{v}_t,\mathbf{r}_{t-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{Q}_2] - \mathbb{E}_{\mathbf{h}_t,\mathbf{v}'_t;\mathbf{r}_{t-1}}[\frac{\partial}{\partial u_{m,m'}}\mathcal{Q}_2]\right)$$
$$\times r_{t,m}(1-r_{t,m})r_{t-1,m'} = \sum_{t=2}^{T}\left(\mathbb{E}_{\mathbf{h}_t|\mathbf{v}_t,\mathbf{r}_{t-1}}[h_{t,m}] - \mathbb{E}_{\mathbf{h}_t,\mathbf{v}'_t;\mathbf{r}_{t-1}}[h_{t,m}] + \mathcal{D}_{t+1,m}r_{t,m}(1-r_{t,m})\right) r_{t-1,m'} \tag{1.21}$$

which in vector for is identical to (1.16). □

In the following we show that similarly to (1.14), $\mathcal{D}_t$ can be computed recursively using:

$$\mathcal{D}_{t+1} = \mathbf{U}^\top(\mathcal{D}_{t+2}\odot\mathbf{r}_{t+1}\odot(\mathbf{1}-\mathbf{r}_{t+1}) + \mathbb{E}_{\mathbf{h}_{t+1}|\mathbf{v}_{t+1};\mathbf{r}_t}[\mathbf{h}_{t+1}] - \mathbb{E}_{\mathbf{v}'_{t+1},\mathbf{h}_{t+1}|\mathbf{r}_t}[\mathbf{h}_{t+1}]), \tag{1.22}$$

and $\mathcal{D}_{T+1} = 0$.

Proof. Plugging (1.14) into (1.17), we have that:

$$\begin{aligned}
\mathcal{D}_{t+1} &= \mathbb{E}_{\mathbf{h}_{t+1},\ldots,\mathbf{h}_T|\mathbf{v}_{t+1},\ldots,\mathbf{v}_T,\mathbf{r}_1,\ldots,\mathbf{r}_T}[\mathbf{U}^\top(\nabla_{\mathbf{r}_{t+1}}\mathcal{Q}_{t+2}\odot\mathbf{r}_{t+1}\odot(\mathbf{1}-\mathbf{r}_{t+1})+\mathbf{h}_{t+1})] \\
&- \mathbb{E}_{\mathbf{h}_{t+1},\ldots,\mathbf{h}_T,\mathbf{v}'_{t+1},\ldots,\mathbf{v}'_T;\mathbf{r}_1,\ldots,\mathbf{r}_T}[\mathbf{U}^\top(\nabla_{\mathbf{r}_{t+1}}\mathcal{Q}_{t+2}\odot\mathbf{r}_{t+1}\odot(\mathbf{1}-\mathbf{r}_{t+1})+\mathbf{h}_{t+1})] \\
&= \mathbf{U}^\top[(\mathbb{E}_{\mathbf{h}_{t+1},\ldots,\mathbf{h}_T|\mathbf{v}_{t+1},\ldots,\mathbf{v}_T,\mathbf{r}_1,\ldots,\mathbf{r}_T}[\nabla_{\mathbf{r}_{t+1}}\mathcal{Q}_{t+2}]-\mathbb{E}_{\mathbf{h}_{t+1},\ldots,\mathbf{h}_T,\mathbf{v}'_{t+1},\ldots,\mathbf{v}'_T;\mathbf{r}_1,\ldots,\mathbf{r}_T}[\nabla_{\mathbf{r}_{t+1}}\mathcal{Q}_{t+2}]) \\
&\odot\mathbf{r}_{t+1}\odot(1-\mathbf{r}_{t+1})+\mathbb{E}_{\mathbf{h}_{t+1}|\mathbf{v}_{t+1},\mathbf{r}_t}[\mathbf{h}_{t+1}]-\mathbb{E}_{\mathbf{h}_{t+1},\mathbf{v}'_{t+1};\mathbf{r}_t}[\mathbf{h}_{t+1}]] \\
&= \mathbf{U}^\top\left(\mathcal{D}_{t+2}\odot\mathbf{r}_{t+1}\odot(1-\mathbf{r}_{t+1})+\mathbb{E}_{\mathbf{h}_{t+1}|\mathbf{v}_{t+1},\mathbf{r}_t}[\mathbf{h}_{t+1}]-\mathbb{E}_{\mathbf{h}_{t+1},\mathbf{v}'_{t+1};\mathbf{r}_t}[\mathbf{h}_{t+1}]\right) \quad (1.23)
\end{aligned}$$

□

The model parameters $\theta \in \{\mathbf{W},\mathbf{U},\mathbf{b},\mathbf{b}_{init},\mathbf{c}\}$ are updated via gradient ascent (e.g., $\theta := \theta + \eta\Delta_\theta^{\mathcal{H}+\mathcal{Q}_2}$) where

$$\begin{aligned}
\Delta_\theta^{\mathcal{H}+\mathcal{Q}_2} &= \mathbb{E}_{\{\mathbf{h}_t\}_{t=1}^T|\{\mathbf{v}_t,\mathbf{r}_t\}_{t=1}^T}[\nabla_\theta\mathcal{H}] \quad (1.24)\\
&- \mathbb{E}_{\{\mathbf{h}_t,\mathbf{v}'_t\}_{t=1}^T|\{\mathbf{r}_t\}_{t=1}^T}[\nabla_\theta\mathcal{H}]+\Delta_\theta^{\mathcal{Q}_2}
\end{aligned}$$

For the other model parameters, we have $\Delta_{\mathbf{c}}^{\mathcal{Q}_2}=0$, and

$$\Delta_{\mathbf{W}}^{\mathcal{Q}_2} = \sum_{t=1}^{T-1}(\mathcal{D}_{t+1}\odot\mathbf{r}_t\odot(\mathbf{1}-\mathbf{r}_t))\mathbf{v}_t^\top \quad (1.25)$$

$$\Delta_{\mathbf{b}}^{\mathcal{Q}_2} = \sum_{t=2}^{T-1}(\mathcal{D}_{t+1}\odot\mathbf{r}_t\odot(\mathbf{1}-\mathbf{r}_t)) \quad (1.26)$$

$$\Delta_{\mathbf{b}_{init}}^{\mathcal{Q}_2} = \mathcal{D}_2\odot\mathbf{r}_1\odot(\mathbf{1}-\mathbf{r}_1) \quad (1.27)$$