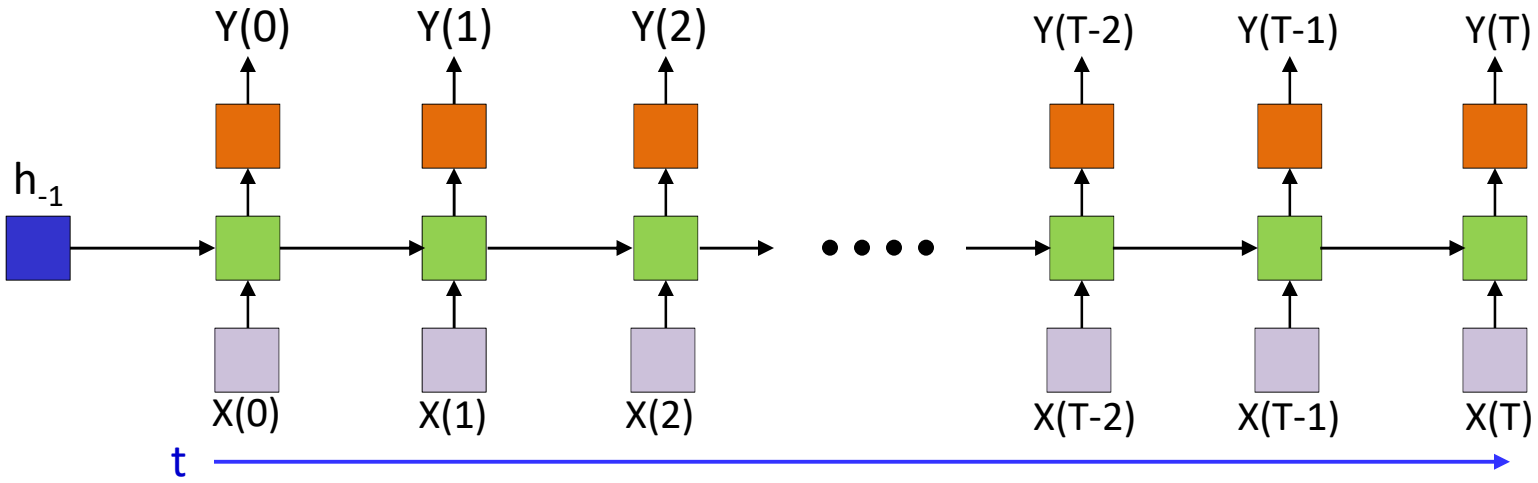
How do we *train* the network



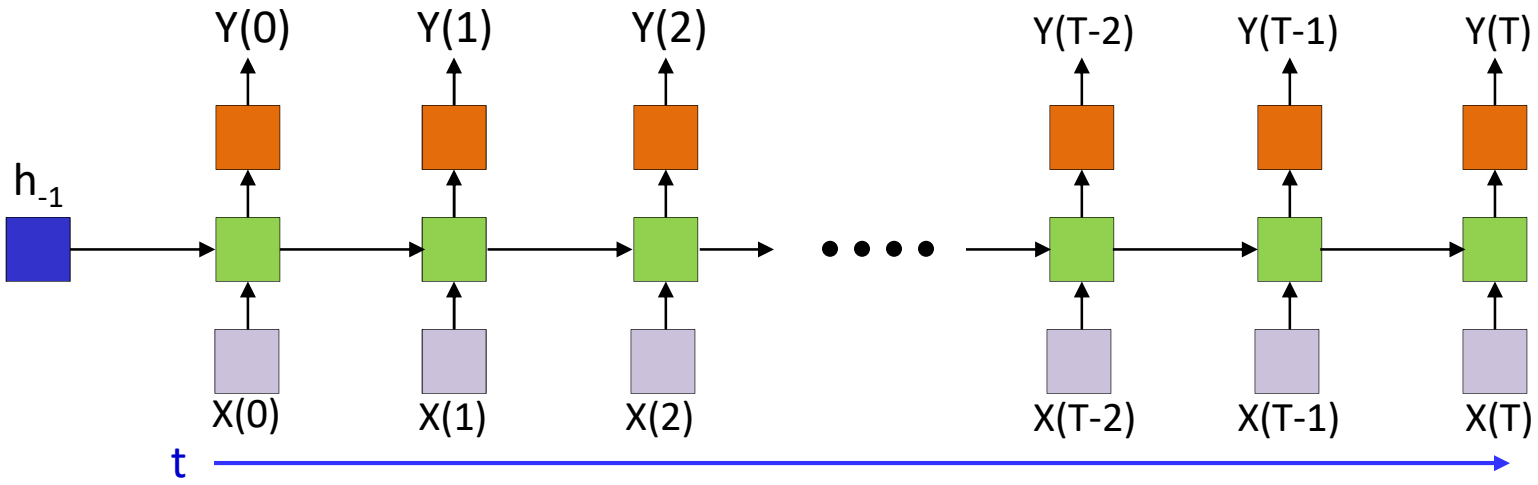
- Back propagation through time (BPTT)
- Given a collection of *sequence* inputs
  - $(\mathbf{X}_i, \mathbf{D}_i)$, where
  - $\mathbf{X}_i = X_{i,0}, \ldots, X_{i,T}$
  - $\mathbf{D}_i = D_{i,0}, \ldots, D_{i,T}$
- Train network parameters to minimize the error between the output of the network $\mathbf{Y}_i = Y_{i,0}, \ldots, Y_{i,T}$ and the desired outputs
  - This is the most generic setting. In other settings we just "remove" some of the input or output entries

Loss Function

1. Note that $D(0), D(1), \ldots, D(T)$ are the desired output sequence.
2. $Y(0), Y(1), \ldots, Y(T)$ is the output sequence generated by the RNN.
3. Thus, the loss will capture distance between two sequences and is denoted as $DIV\left(\{Y(0), Y(1), \ldots, Y(T)\}, \{D(0), D(1), \ldots, D(T)\}\right)$.
4. An example of the loss is where $D(t)$ is compared to only $Y(t)$ and the overall loss is a sum of local loss values as follows.

$$DIV\left(\{Y(0), Y(1), \ldots, Y(T)\}, \{D(0), D(1), \ldots, D(T)\}\right)$$
$$= \sum_{t=0}^{T} Div\left(Y(t), D(t)\right)$$
$$\Rightarrow \quad \frac{\partial DIV\left(\{Y(0), Y(1), \ldots, Y(T)\}, \{D(0), D(1), \ldots, D(T)\}\right)}{\partial Y(t)}$$
$$= \frac{\partial Div\left(Y(t), D(t)\right)}{\partial Y(t)}$$

Training: Forward pass



- For each training input:
- Forward pass: pass the entire data sequence through the network, generate outputs

BPTT: Forward Pass

1. Let the input $X(t)$ has dimension N for all $t = 0, \ldots, T$.
2. Let the hidden layer has N_1 nodes.
3. Then the pre-activation value $Z_i^{(1)}(t)$ of the i^{th} node in the hidden layer at time t is computed as follow. Note that it takes $X(t)$ and $\mathbf{h}(t-1)$ as input.

$$Z_i^{(1)}(t) = \sum_{j=1}^{N} w_{ji}^{(1)} X_j(t) + \sum_{j=1}^{N_1} w_{ji}^{(11)} h_j^{(1)}(t-1) + b_i^{(1)}, \; i = 1 \ldots N_1$$

4. Output of i^{th} node in the hidden layer is found by applying activation function $f^{(1)}$ on $Z_i^{(1)}(t)$.

$$h_i^{(1)}(t) = f^{(1)}\left(Z_i^{(1)}(t)\right), \; i = 1 \ldots N_1$$

BPTT: Forward Pass

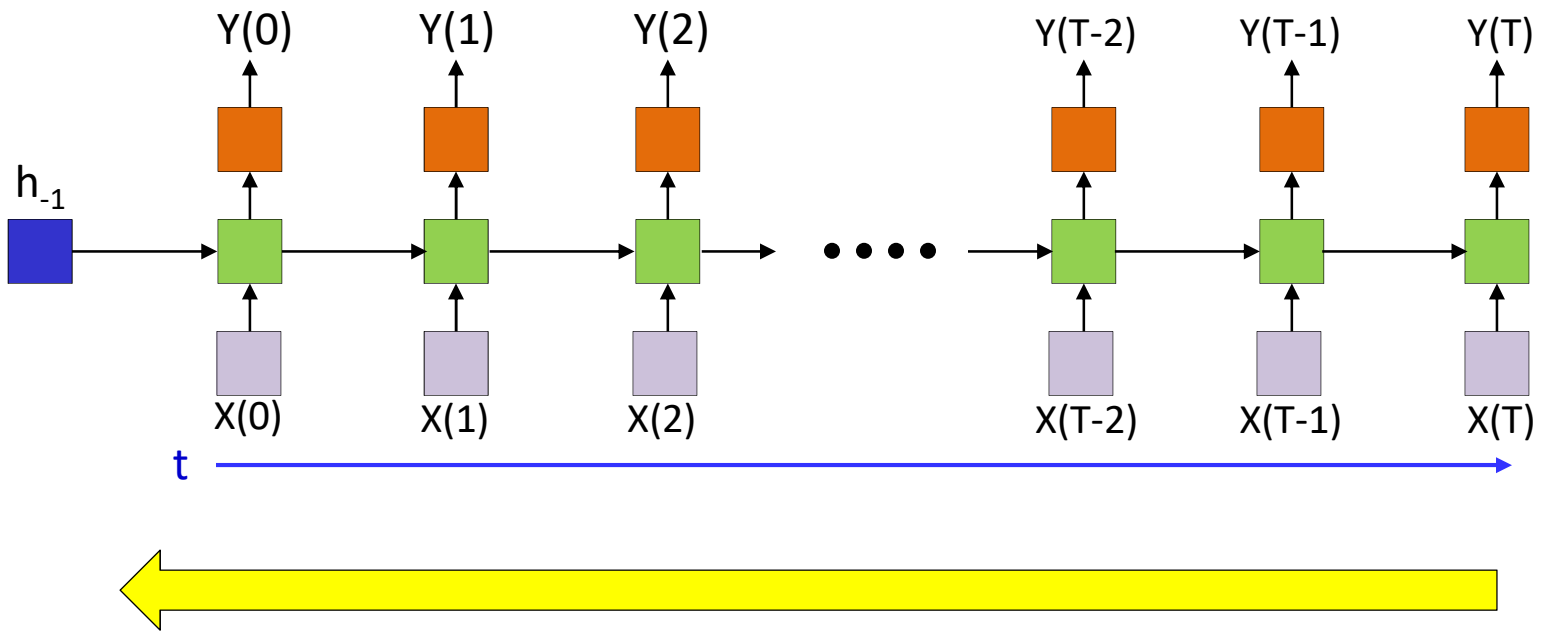
1. Let the output $Y(t)$ is an M-dimensional vector.
2. Then the pre-activation value $Z_i^{(2)}(t)$ of the i^{th} node in the output layer at time t is computed as follow. Note that it takes $\mathbf{h}(t)$ as input.

$$Z_i^{(2)}(t) = \sum_{j=1}^{N_1} w_{ji}^{(2)} h_j(t) + b_i^{(2)}, \ i = 1 \ldots M$$

3. Output of i^{th} node in the output layer is found by applying activation function $f_i^{(2)}$ on $Z_i^{(2)}(t)$.

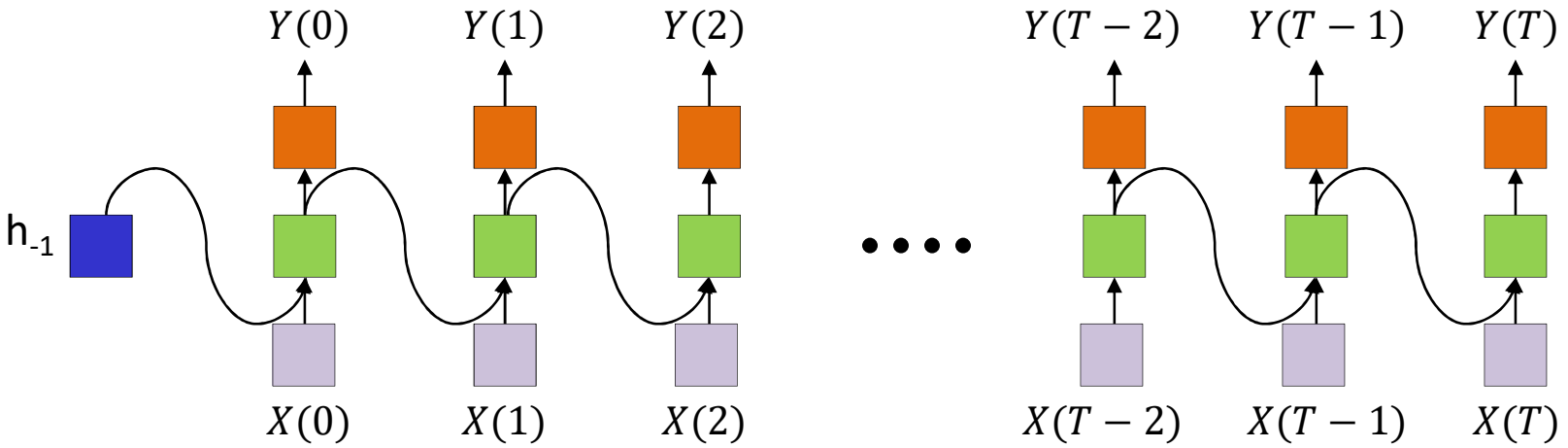
$$Y_i(t) = f_i^{(2)}\left(Z_1^{(2)}(t), \ldots, Z_M^{(2)}(t)\right), \ i = 1 \ldots M$$

Training: Computing gradients



- For each training input:
- Backward pass: Compute gradients via backpropagation
  - *Back Propagation Through Time*

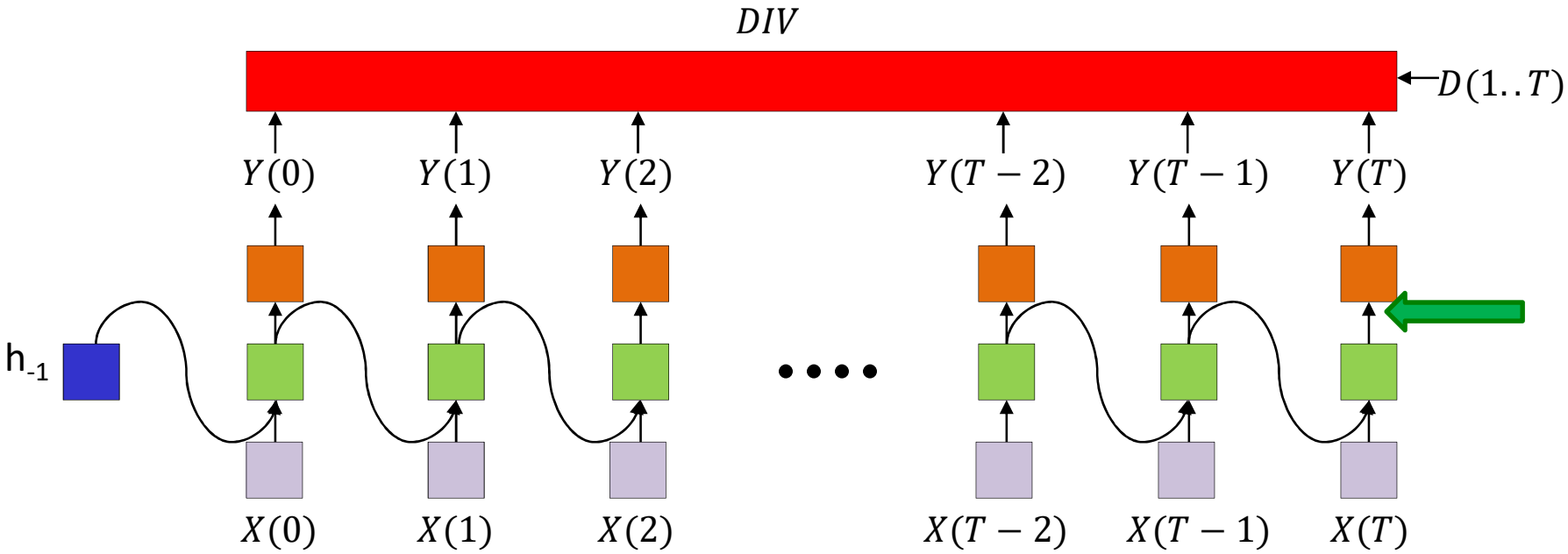
Back Propagation Through Time



Will only focus on *one* training instance

All subscripts represent *components* and not training instance index

Back Propagation Through Time



First step of backprop: Compute $\frac{dDIV}{dY_i(T)}$ for all i

$$\nabla_{Z^{(2)}(T)} DIV = \nabla_{Y(T)} DIV \nabla_{Z^{(2)}(T)} Y(T)$$

Vector output activation

$$\frac{dDIV}{dZ_i^{(2)}(T)} = \frac{dDIV}{dY_i(T)} \frac{dY_i(T)}{dZ_i^{(2)}(T)}$$

OR

$$\frac{dDIV}{dZ_i^{(2)}(T)} = \sum_j \frac{dDIV}{dY_j(T)} \frac{dY_j(T)}{dZ_i^{(2)}(T)}$$

BPTT: Backward Pass (Compute $\frac{\partial DIV}{\partial Z_i^{(2)}(T)}$)

1. We know that $Z_i^{(2)}(T) = \sum_{j=1}^{N_1} w_{ji}^{(2)} h_j(T) + b_i^{(2)},\ i = 1 \ldots M$
2. **Case 1: Vector output activation**
   - $Y_k(T) = f_k^{(2)}\left(Z_1^{(2)}(T), \ldots, Z_M^{(2)}(T)\right),\ k = 1 \ldots M$
   - An example of such an $f_k^{(2)}$ is softmax function, i.e.

     $$f_k^{(2)}\left(Z_1^{(2)}(T), \ldots, Z_M^{(2)}(T)\right) = \frac{\exp\left(Z_k^{(2)}(T)\right)}{\sum_{j=1}^{M} \exp\left(Z_j^{(2)}(T)\right)},\ k = 1 \ldots M.$$

   - In this case, $\frac{\partial DIV}{\partial Z_i^{(2)}(T)} = \sum_{j=1}^{M} \frac{\partial DIV}{\partial Y_j(T)} \frac{\partial Y_j(T)}{\partial Z_i^{(2)}(T)}$
3. **Case 2: Scalar output activation**
   - $Y_k(T) = f_k^{(2)}\left(Z_k^{(2)}(T)\right),\ k = 1 \ldots M$
   - An example of such an $f_k^{(2)}$ is tanh function
   - In this case, $\frac{\partial DIV}{\partial Z_i^{(2)}(T)} = \frac{\partial DIV}{\partial Y_i(T)} \frac{\partial Y_i(T)}{\partial Z_i^{(2)}(T)}$

Case 1: Vector output activation

$$\nabla_{Z^{(2)}(T)} DIV = \left[\sum_{j=1}^{M} \frac{\partial DIV}{\partial Y_j(T)} \frac{\partial Y_j(T)}{\partial Z_1^{(2)}(T)} \cdots\cdots \sum_{j=1}^{M} \frac{\partial DIV}{\partial Y_j(T)} \frac{\partial Y_j(T)}{\partial Z_M^{(2)}(T)}\right]$$

$$= \left[\frac{\partial DIV}{\partial Y_1(T)} \cdots\cdots \frac{\partial DIV}{\partial Y_M(T)}\right] \begin{bmatrix} \frac{\partial Y_1(T)}{\partial Z_1^{(2)}(T)} & \frac{\partial Y_1(T)}{\partial Z_2^{(2)}(T)} & \cdots & \frac{\partial Y_1(T)}{\partial Z_M^{(2)}(T)} \\ \frac{\partial Y_2(T)}{\partial Z_1^{(2)}(T)} & \frac{\partial Y_2(T)}{\partial Z_2^{(2)}(T)} & \cdots & \frac{\partial Y_2(T)}{\partial Z_M^{(2)}(T)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_M(T)}{\partial Z_1^{(2)}(T)} & \frac{\partial Y_2(T)}{\partial Z_2^{(2)}(T)} & \cdots & \frac{\partial Y_M(T)}{\partial Z_M^{(2)}(T)} \end{bmatrix}$$

$$= \nabla_{Y(T)} DIV \;\; \nabla_{Z^{(2)}(T)} Y(T)$$

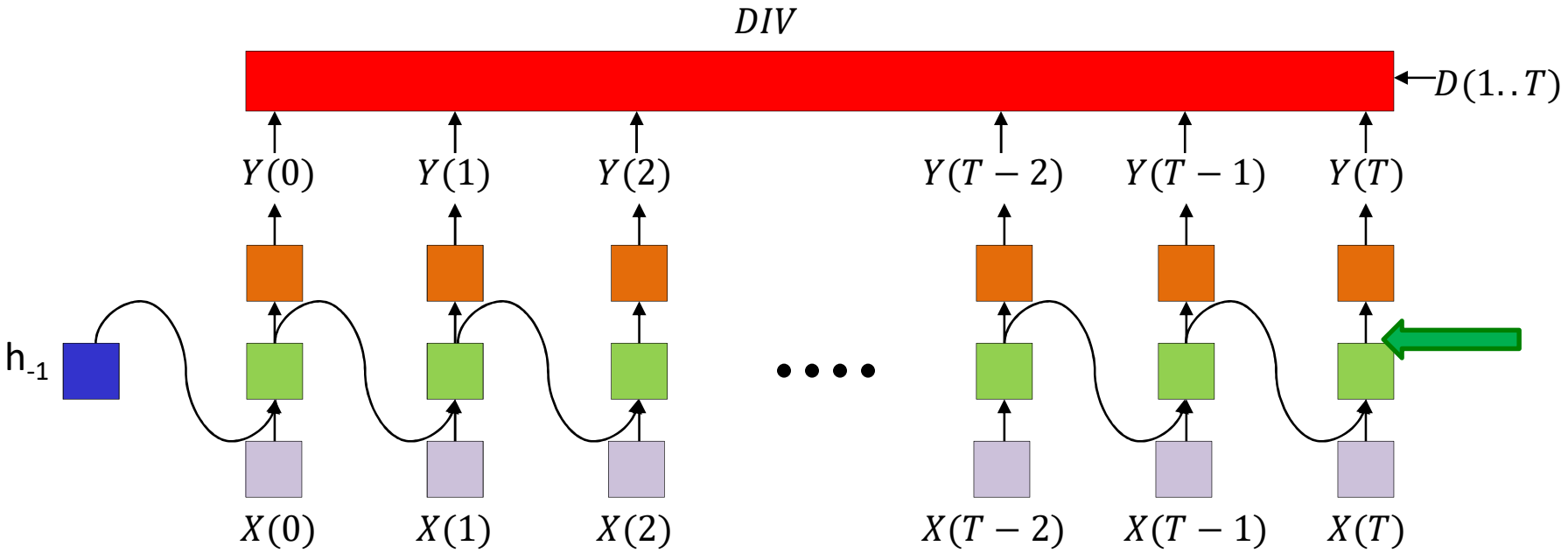
Case 2: Scalar output activation

$$\nabla_{Z^{(2)}(T)} DIV = \left[\frac{\partial DIV}{\partial Y_1(T)} \frac{\partial Y_1(T)}{\partial Z_1^{(2)}(T)} \cdots\cdots \frac{\partial DIV}{\partial Y_M(T)} \frac{\partial Y_M(T)}{\partial Z_M^{(2)}(T)} \right]$$

$$= \left[\frac{\partial DIV}{\partial Y_1(T)} \cdots\cdots \frac{\partial DIV}{\partial Y_M(T)} \right] \begin{bmatrix} \frac{\partial Y_1(T)}{\partial Z_1^{(2)}(T)} & 0 & \cdots & 0 \\ 0 & \frac{\partial Y_2(T)}{\partial Z_2^{(2)}(T)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial Y_M(T)}{\partial Z_M^{(2)}(T)} \end{bmatrix}$$

$$= \nabla_{Y(T)} DIV \;\; \nabla_{Z^{(2)}(T)} Y(T)$$

Back Propagation Through Time



$$\frac{dDIV}{dh_i(T)} = \sum_j \frac{dDIV}{dZ_j^{(2)}(T)} \frac{dZ_j^{(2)}(T)}{dh_i(T)} = \sum_j w_{ij}^{(2)} \frac{dDIV}{dZ_j^{(2)}(T)}$$

$\frac{dDIV}{dY_i(T)}$ for all i

$$\frac{dDIV}{dZ_i^{(2)}(T)} = \frac{dDiv(T)}{dY_i(T)} \frac{dY_i(T)}{dZ_i^{(2)}(T)}$$

$$\nabla_{h(T)} DIV = \nabla_{Z^{(2)}(T)} DIV \; W^{(2)}$$

Compute $\nabla_{\mathbf{h}(T)} DIV = \left[\frac{\partial DIV}{\partial h_1(T)} \quad \cdots \quad \frac{\partial DIV}{\partial h_{N_1}(T)} \right]$

1. Note that pre-activation value of j^{th} node in the output layer at time T is $Z_j^{(2)}(T) = \sum_{k=1}^{N_1} w_{kj}^{(2)} h_k(T) + b_j^{(2)}$.
2. $h_i(T)$ is an input to all the $Z_1^{(2)}(T), \ldots, Z_M^{(2)}(T)$. Thus, we see that,

$$\frac{\partial DIV}{\partial h_i(T)} = \sum_{j=1}^{M} \frac{\partial DIV}{\partial Z_j^{(2)}(T)} \frac{\partial Z_j^{(2)}(T)}{\partial h_i(T)} = \sum_{j=1}^{M} w_{ij}^{(2)} \frac{\partial DIV}{\partial Z_j^{(2)}(T)}$$

3. Computing $\nabla_{\mathbf{h}(T)} DIV$ now, we get,

$$\nabla_{\mathbf{h}(T)} DIV = \left[\frac{\partial DIV}{\partial h_1(T)} \quad \cdots\cdots \quad \frac{\partial DIV}{\partial h_{N_1}(T)} \right]$$

$$= \left[\sum_{j=1}^{M} \frac{\partial DIV}{\partial Z_j^{(2)}(T)} \frac{\partial Z_j^{(2)}(T)}{\partial h_1(T)} \quad \cdots\cdots \quad \sum_{j=1}^{M} \frac{\partial DIV}{\partial Z_j^{(2)}(T)} \frac{\partial Z_j^{(2)}(T)}{\partial h_{N_1}(T)} \right]$$

$$= \left[\frac{\partial DIV}{\partial Z_1^{(2)}(T)} \quad \cdots \quad \frac{\partial DIV}{\partial Z_M^{(2)}(T)} \right] \begin{bmatrix} \frac{\partial Z_1^{(2)}(T)}{\partial h_1(T)} & \frac{\partial Z_1^{(2)}(T)}{\partial h_2(T)} & \cdots & \frac{\partial Z_1^{(2)}(T)}{\partial h_{N_1}(T)} \\ \frac{\partial Z_2^{(2)}(T)}{\partial h_1(T)} & \frac{\partial Z_2^{(2)}(T)}{\partial h_2(T)} & \cdots & \frac{\partial Z_2^{(2)}(T)}{\partial h_{N_1}(T)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Z_M^{(2)}(T)}{\partial h_1(T)} & \frac{\partial Z_2(T)}{\partial h_2(T)} & \cdots & \frac{\partial Z_M^{(2)}(T)}{\partial h_{N_1}(T)} \end{bmatrix}$$

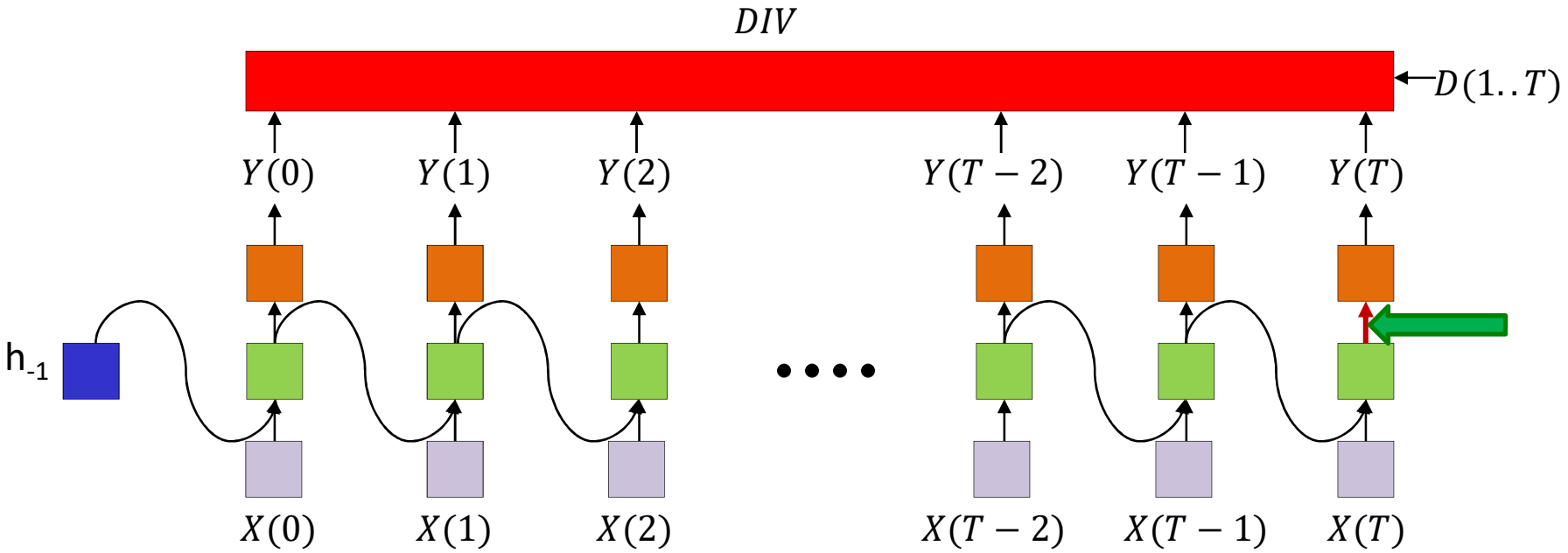
Compute $\nabla_{\mathbf{h}(T)} DIV = \left[\frac{\partial DIV}{\partial h_1(T)} \quad \cdots \quad \frac{\partial DIV}{\partial h_{N_1}(T)} \right]$

$$\nabla_{\mathbf{h}(T)} DIV = \nabla_{Z^{(2)}(T)} DIV \;\; \nabla_{\mathbf{h}(T)} Z^{(2)}(T)$$

$$= \left[\frac{\partial DIV}{\partial Z_1^{(2)}(T)} \quad \cdots \quad \frac{\partial DIV}{\partial Z_M^{(2)}(T)} \right] \begin{bmatrix} w_{11}^{(2)} & w_{21}^{(2)} & \cdots & w_{N_1 1}^{(2)} \\ w_{12}^{(2)} & w_{22}^{(2)} & \cdots & w_{N_1 2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1M}^{(2)} & w_{2M}^{(2)} & \cdots & w_{N_1 M}^{(2)} \end{bmatrix}$$

$$= \nabla_{Z^{(2)}(T)} DIV \;\; W^{(2)}$$

Back Propagation Through Time



$$\frac{dDIV}{dZ_i^{(2)}(T)} = \frac{dDiv(T)}{dY_i(T)} \frac{dY_i(T)}{dZ_i^{(2)}(T)} \qquad \frac{dDIV}{dh_i(T)} = \sum_j w_{ij}^{(2)} \frac{dDIV}{dZ_j^{(2)}(T)}$$

$$\nabla_{W^{(2)}} DIV = h(T) \nabla_{Z^{(2)}(T)} DIV$$

$$\frac{dDIV}{dw_{ij}^{(2)}} = \frac{dDIV}{dZ_j^{(2)}(T)} h_i(T)$$

Compute $\nabla_{W^{(2)}} DIV$

1. Note that $W^{(2)}$ is an $M \times N_1$ weight matrix connecting hidden layer to the output layer.
2. $w_{ij}^{(2)}$ is connected to pre-activation value of j^{th} output node at T as $Z_j^{(2)}(T) = \sum_{k=1}^{N_1} w_{kj}^{(2)} h_k(T) + b_j^{(2)}$.
3. Thus,

$$\frac{\partial DIV}{\partial w_{ij}^{(2)}} = \frac{\partial DIV}{\partial Z_j^{(2)}(T)} \frac{\partial Z_j^{(2)}(T)}{\partial w_{ij}^{(2)}} = \frac{\partial DIV}{\partial Z_j^{(2)}(T)} h_i(T)$$

4. Computing $\nabla_{W^{(2)}} DIV$

$$\nabla_{W^{(2)}} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial w_{11}^{(2)}} & \frac{\partial DIV}{\partial w_{12}^{(2)}} & \cdots & \frac{\partial DIV}{\partial w_{1M}^{(2)}} \\ \frac{\partial DIV}{\partial w_{21}^{(2)}} & \frac{\partial DIV}{\partial w_{22}^{(2)}} & \cdots & \frac{\partial DIV}{\partial w_{2M}^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial w_{N_1 1}^{(2)}} & \frac{\partial DIV}{\partial w_{N_1 2}^{(2)}} & \cdots & \frac{\partial DIV}{\partial w_{N_1 M}^{(2)}} \end{bmatrix}$$

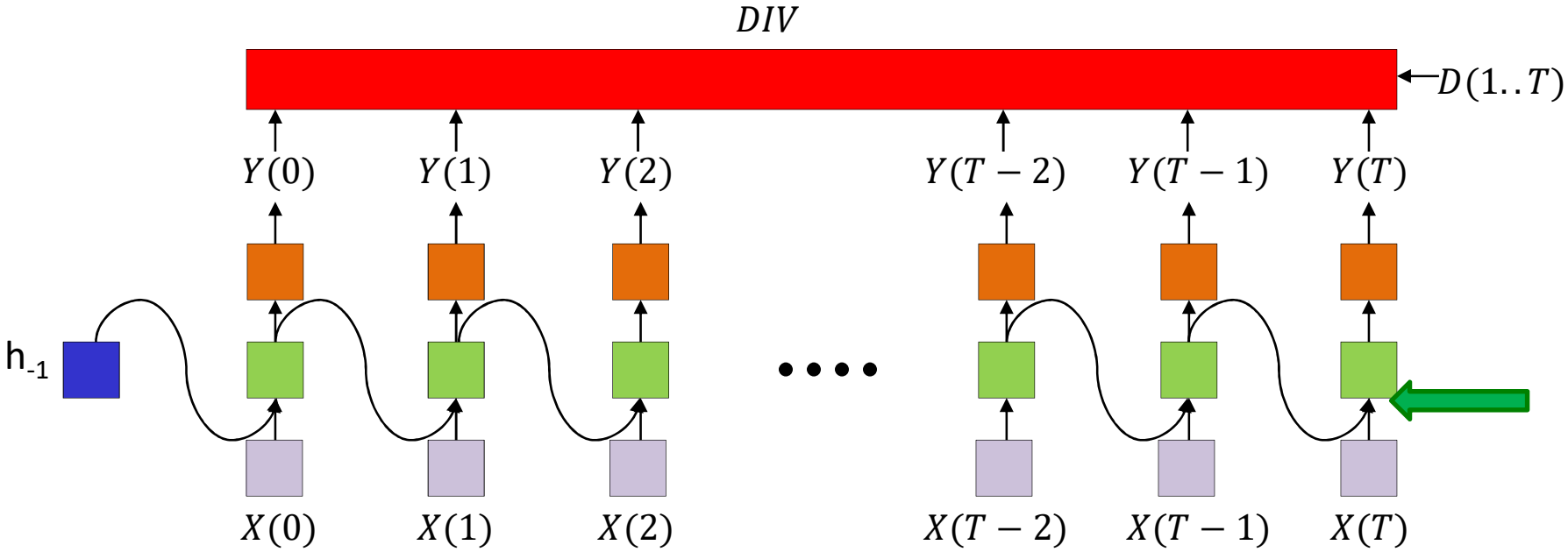
Compute $\nabla_{W^{(2)}} DIV$

$$\nabla_{W^{(2)}} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(2)}(T)} h_1(T) & \frac{\partial DIV}{\partial Z_2^{(2)}(T)} h_1(T) & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T)} h_1(T) \\ \frac{\partial DIV}{\partial Z_1^{(2)}(T)} h_2(T) & \frac{\partial DIV}{\partial Z_2^{(2)}(T)} h_2(T) & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T)} h_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial Z_1^{(2)}(T)} h_{N_1}(T) & \frac{\partial DIV}{\partial Z_2^{(2)}(T)} h_{N_1}(T) & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T)} h_{N_1}(T) \end{bmatrix}$$

$$= \begin{bmatrix} h_1(T) \\ h_2(T) \\ \vdots \\ h_{N_1}(T) \end{bmatrix} \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(2)}(T)} & \frac{\partial DIV}{\partial Z_2^{(2)}(T)} & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T)} \end{bmatrix}$$

$$= \mathbf{h}(T) \nabla_{Z^{(2)}} DIV$$

Back Propagation Through Time



$$\nabla_{Z^{(1)}(T)} DIV = \nabla_{h(T)} DIV \, \nabla_{Z^{(1)}(T)} h(T)$$

$$\frac{dDIV}{dZ_i^{(1)}(T)} = \frac{dDIV}{dh_i(T)} \frac{dh_i(T)}{dZ_i^{(1)}(T)}$$

$$\frac{dDIV}{dZ_i^{(2)}(T)} = \frac{dDIV}{dY_i(T)} \frac{dY_i(T)}{dZ_i^{(2)}(T)}$$

$$\frac{dDIV}{dh_i(T)} = \sum_j w_{ij}^{(2)} \frac{dDIV}{dZ_j^{(2)}(T)}$$

$$\frac{dDIV}{dw_{ij}^{(2)}} = \frac{dDIV}{dZ_j^{(2)}(T)} h_i(T)$$

Compute $\nabla_{Z^{(1)}(T)} DIV$

1. $Z_i^{(1)}(T)$ is the pre-activation value of i^{th} hidden node at time T.
2. Note that $h_i(T) = Z_i^{(1)}(T)$.
3. Thus, $\frac{\partial DIV}{\partial Z_i^{(1)}(T)} = \frac{\partial DIV}{\partial h_i(T)} \frac{\partial h_i(T)}{\partial Z_i^{(1)}(T)}$.
4. Computing $\nabla_{Z^{(1)}(T)} DIV$

$$\nabla_{Z^{(1)}(T)} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T)} & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial DIV}{\partial h_1(T)} \frac{\partial h_1(T)}{\partial Z_1^{(1)}(T)} & \frac{\partial DIV}{\partial h_2(T)} \frac{\partial h_2(T)}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial DIV}{\partial h_{N_1}(T)} \frac{\partial h_{N_1}(T)}{\partial Z_{N_1}^{(1)}(T)} \end{bmatrix}$$

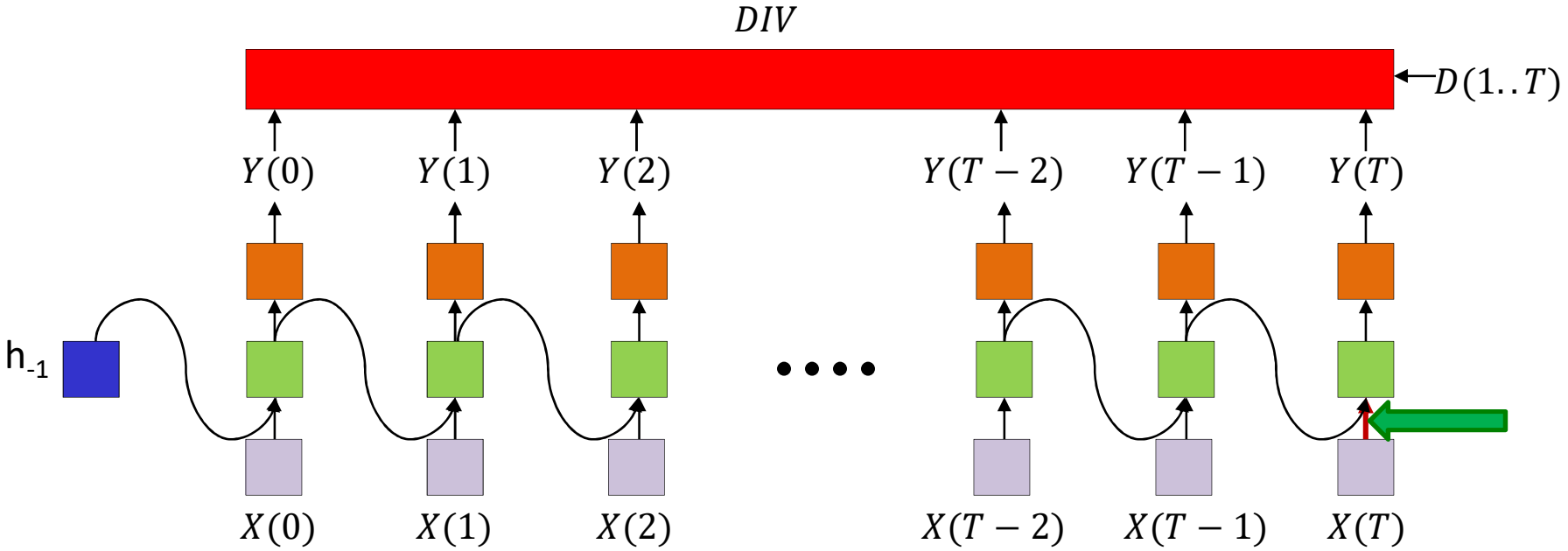
$$= \begin{bmatrix} \frac{\partial DIV}{\partial h_1(T)} & \frac{\partial DIV}{\partial h_2(T)} & \cdots & \frac{\partial DIV}{\partial h_{N_1}(T)} \end{bmatrix} \begin{bmatrix} \frac{\partial h_1(T)}{\partial Z_1^{(1)}(T)} & \frac{\partial h_1(T)}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial h_1(T)}{\partial Z_{N_1}^{(1)}(T)} \\ \frac{\partial h_2(T)}{\partial Z_1^{(1)}(T)} & \frac{\partial h_2(T)}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial h_2(T)}{\partial Z_{N_1}^{(1)}(T)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_{N_1}(T)}{\partial Z_1^{(1)}(T)} & \frac{\partial h_{N_1}(T)}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial h_{N_1}(T)}{\partial Z_{N_1}^{(1)}(T)} \end{bmatrix}$$

Compute $\nabla_{Z^{(1)}(T)} DIV$

$$\nabla_{Z^{(1)}(T)} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial h_1(T)} & \frac{\partial DIV}{\partial h_2(T)} & \cdots & \frac{\partial DIV}{\partial h_{N_1}(T)} \end{bmatrix} \begin{bmatrix} \frac{\partial h_1(T)}{\partial Z_1^{(1)}(T)} & 0 & \cdots & 0 \\ 0 & \frac{\partial h_2(T)}{\partial Z_2^{(1)}(T)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial h_{N_1}(T)}{\partial Z_{N_1}^{(1)}(T)} \end{bmatrix}$$

$$= \nabla_{\mathbf{h}(T)} DIV \;\; \nabla_{Z^{(1)}(T)} \mathbf{h}(T)$$

Back Propagation Through Time



$$\nabla_{W^{(1)}} DIV = X(T) \nabla_{Z^{(1)}(T)} DIV$$

$$\frac{dDIV}{dw_{ij}^{(1)}} = \frac{dDIV}{dZ_j^{(1)}(T)} X_i(T)$$

Compute $\nabla_{W^{(1)}} DIV$

1. Note that $W^{(1)}$ is $N_1 \times N$ weight matrix between the input layer and the hidden layer.
2. $w_{ij}^{(1)}$ is connected to $Z_j^{(1)}(T)$ as follows.
   $Z_j^{(1)}(T) = \sum_{k=1}^{N} w_{kj}^{(1)} X_k(T) + \sum_{k=1}^{N_1} w_{kj}^{(11)} h_k(T-1) + b_j^{(1)}$
3. Thus, $\frac{\partial DIV}{\partial w_{ij}^{(1)}} = \frac{\partial DIV}{\partial Z_j^{(1)}(T)} \frac{\partial Z_j^{(1)}(T)}{\partial w_{ij}^{(1)}} = \frac{\partial DIV}{\partial Z_j^{(1)}(T)} X_i(T)$
4. Computing $\nabla_{W^{(1)}} DIV$

$$\nabla_{W^{(1)}} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial w_{11}^{(1)}} & \frac{\partial DIV}{\partial w_{12}^{(1)}} & \cdots & \frac{\partial DIV}{\partial w_{1N_1}^{(1)}} \\ \frac{\partial DIV}{\partial w_{21}^{(1)}} & \frac{\partial DIV}{\partial w_{22}^{(1)}} & \cdots & \frac{\partial DIV}{\partial w_{2N_1}^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial w_{N1}^{(1)}} & \frac{\partial DIV}{\partial w_{N2}^{(1)}} & \cdots & \frac{\partial DIV}{\partial w_{NN_1}^{(1)}} \end{bmatrix}$$

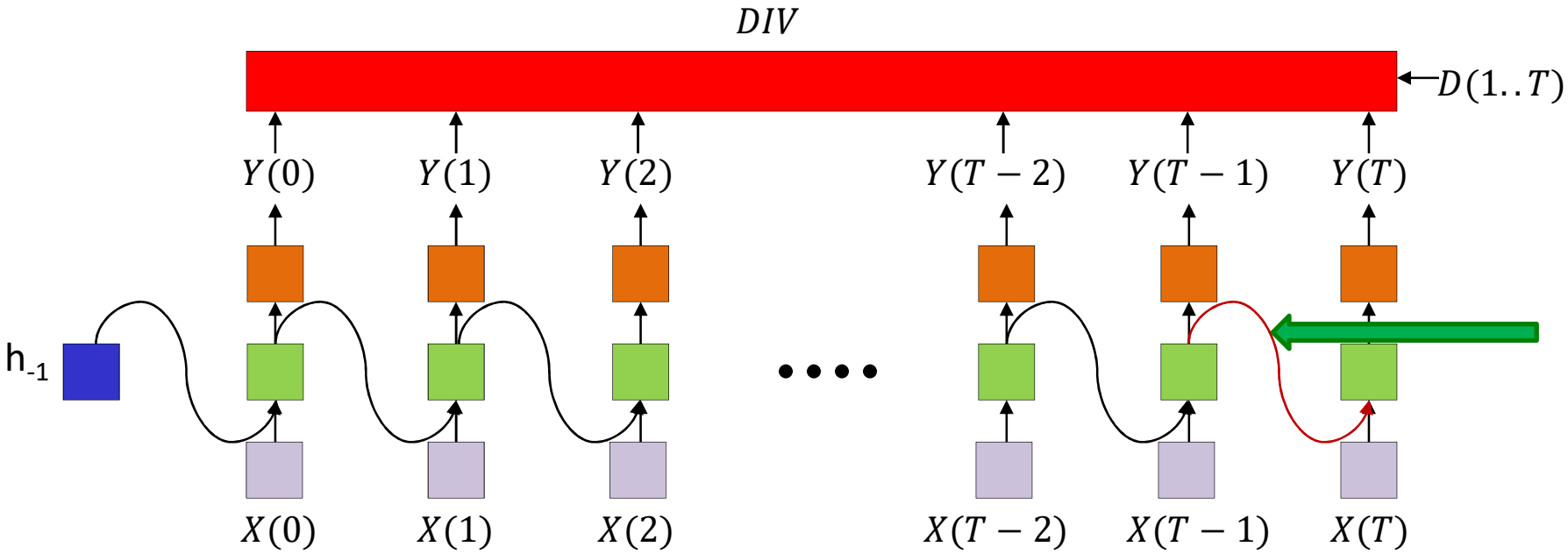
Compute $\nabla_{W^{(1)}} DIV$

$$\nabla_{W^{(1)}} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T)} X_1(T) & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} X_1(T) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} X_1(T) \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T)} X_2(T) & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} X_2(T) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} X_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T)} X_N(T) & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} X_N(T) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} X_N(T) \end{bmatrix}$$

$$= \begin{bmatrix} X_1(T) \\ X_2(T) \\ \vdots \\ X_N(T) \end{bmatrix} \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T)} & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} \end{bmatrix}$$

$$= X(T)\, \nabla_{Z^{(1)}(T)} DIV$$

Back Propagation Through Time



$$\nabla_{W^{(11)}} DIV = h(T-1)\nabla_{Z^{(1)}(T)} DIV$$

$$\frac{dDIV}{dw_{ij}^{(1)}} = \frac{dDIV}{dZ_j^{(1)}(T)} X_i(T)$$

$$\frac{dDIV}{dw_{ij}^{(11)}} = \frac{dDIV}{dZ_j^{(1)}(T)} h_i(T-1)$$

Compute $\nabla_{W^{(11)}} DIV$

1. Note that $W^{(11)}$ is $N_1 \times N_1$ weight matrix between $\mathbf{h}(T-1)$ and $\mathbf{h}(T)$.
2. $w_{ij}^{(11)}$ is connected to $Z_j^{(1)}(T)$ as follows.
   $Z_j^{(1)}(T) = \sum_{k=1}^{N} w_{kj}^{(1)} X_k(T) + \sum_{k=1}^{N_1} w_{kj}^{(11)} h_k(T-1) + b_j^{(1)}$
3. Thus, $\frac{\partial DIV}{\partial w_{ij}^{(11)}} = \frac{\partial DIV}{\partial Z_j^{(1)}(T)} \frac{\partial Z_j^{(1)}(T)}{\partial w_{ij}^{(11)}} = \frac{\partial DIV}{\partial Z_j^{(1)}(T)} h_i(T-1)$.
4. Computing $\nabla_{W^{(1)}} DIV$

$$\nabla_{W^{(11)}} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial w_{11}^{(11)}} & \frac{\partial DIV}{\partial w_{12}^{(11)}} & \cdots & \frac{\partial DIV}{\partial w_{1N_1}^{(11)}} \\ \frac{\partial DIV}{\partial w_{21}^{(11)}} & \frac{\partial DIV}{\partial w_{22}^{(11)}} & \cdots & \frac{\partial DIV}{\partial w_{2N_1}^{(11)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial w_{N_1 1}^{(11)}} & \frac{\partial DIV}{\partial w_{N_1 2}^{(11)}} & \cdots & \frac{\partial DIV}{\partial w_{N_1 N_1}^{(11)}} \end{bmatrix}$$

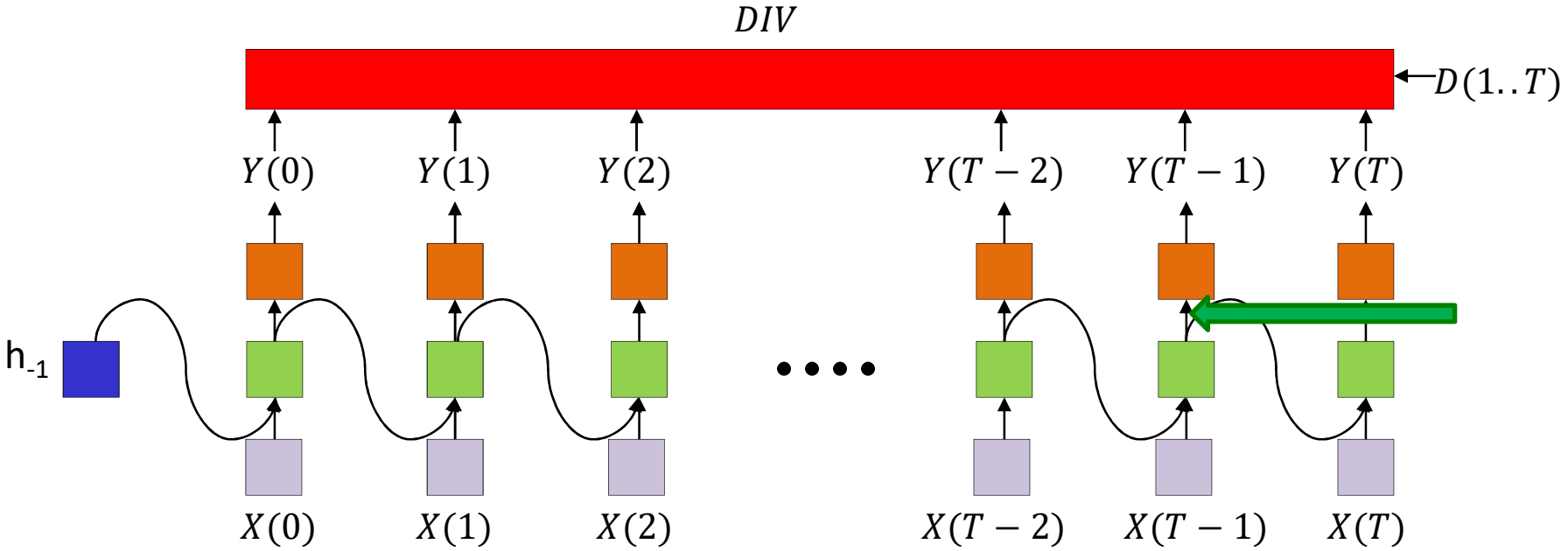
Compute $\nabla_{W^{(11)}} DIV$

$$\nabla_{W^{(11)}} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T)} h_1(T-1) & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} h_1(T-1) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} h_1(T-1) \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T)} h_2(T-1) & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} h_2(T-1) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} h_2(T-1) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T)} h_{N_1}(T-1) & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} h_{N_1}(T-1) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} h_{N_1}(T-1) \end{bmatrix}$$

$$= \begin{bmatrix} h_1(T-1) \\ h_2(T-1) \\ \vdots \\ h_{N_1}(T-1) \end{bmatrix} \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T)} & \frac{\partial DIV}{\partial Z_2^{(1)}(T)} & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} \end{bmatrix}$$

$$= \mathbf{h}(T-1)\, \nabla_{Z^{(1)}(T)} DIV$$

Back Propagation Through Time



$$\nabla_{Z^{(2)}(T-1)} DIV = \nabla_{Y(T-1)} DIV \, \nabla_{Z^{(2)}(T)} Y(T-1)$$

Vector output activation

$$\frac{dDIV}{dZ_i^{(2)}(T-1)} = \frac{dDIV}{dY_i(T-1)} \frac{dY_i(T-1)}{dZ_i^{(2)}(T-1)}$$

OR

$$\frac{dDIV}{dZ_i^{(2)}(T-1)} = \sum_j \frac{dDIV}{dY_j(T-1)} \frac{dY_j(T-1)}{dZ_i^{(2)}(T-1)}$$

Case 1: Vector output activation

$$\nabla_{Z^{(2)}(T-1)} DIV = \left[\frac{\partial DIV}{\partial Z_1^{(2)}(T-1)} \quad \cdots \quad \frac{\partial DIV}{\partial Z_M^{(2)}(T-1)}\right]$$

$$= \left[\sum_{j=1}^{M} \frac{\partial DIV}{\partial Y_j(T-1)} \frac{\partial Y_j(T-1)}{\partial Z_1^{(2)}(T-1)} \quad \cdots\cdots \quad \sum_{j=1}^{M} \frac{\partial DIV}{\partial Y_j(T-1)} \frac{\partial Y_j(T-1)}{\partial Z_M^{(2)}(T-1)}\right]$$

$$= \left[\frac{\partial DIV}{\partial Y_1(T-1)} \quad \cdots\cdots \quad \frac{\partial DIV}{\partial Y_M(T-1)}\right] \begin{bmatrix} \frac{\partial Y_1(T-1)}{\partial Z_1^{(2)}(T-1)} & \frac{\partial Y_1(T-1)}{\partial Z_2^{(2)}(T-1)} & \cdots & \frac{\partial Y_1(T-1)}{\partial Z_M^{(2)}(T-1)} \\ \frac{\partial Y_2(T-1)}{\partial Z_1^{(2)}(T-1)} & \frac{\partial Y_2(T-1)}{\partial Z_2^{(2)}(T-1)} & \cdots & \frac{\partial Y_2(T-1)}{\partial Z_M^{(2)}(T-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_M(T-1)}{\partial Z_1^{(2)}(T-1)} & \frac{\partial Y_M(T-1)}{\partial Z_2^{(2)}(T-1)} & \cdots & \frac{\partial Y_M(T-1)}{\partial Z_M^{(2)}(T-1)} \end{bmatrix}$$

$$= \nabla_{Y(T-1)} DIV \;\; \nabla_{Z^{(2)}(T-1)} Y(T-1)$$

Compute $\nabla_{Z^{(2)}(T-1)} DIV$

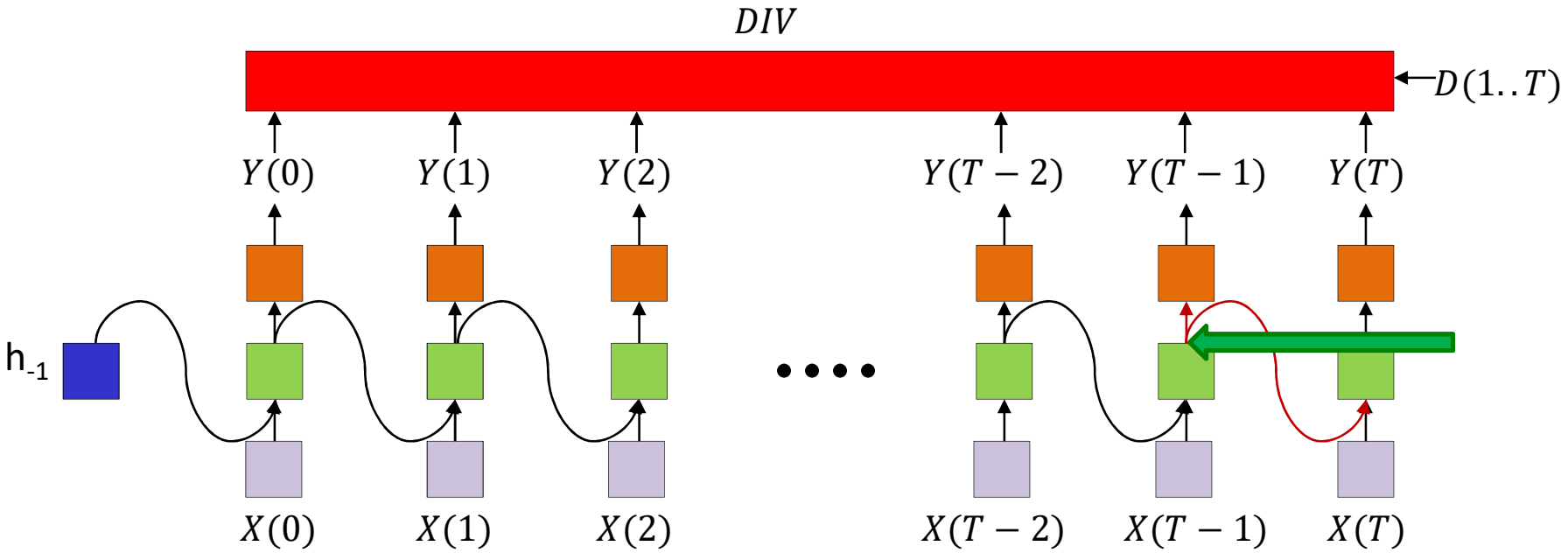
Case 2: Scalar output activation

$$\nabla_{Z^{(2)}(T-1)} DIV = \left[\frac{\partial DIV}{\partial Y_1(T-1)} \frac{\partial Y_1(T-1)}{\partial Z_1^{(2)}(T-1)} \quad \cdots\cdots \quad \frac{\partial DIV}{\partial Y_M(T-1)} \frac{\partial Y_M(T-1)}{\partial Z_M^{(2)}(T-1)} \right]$$

$$= \left[\frac{\partial DIV}{\partial Y_1(T-1)} \quad \cdots\cdots \quad \frac{\partial DIV}{\partial Y_M(T-1)} \right] \begin{bmatrix} \frac{\partial Y_1(T-1)}{\partial Z_1^{(2)}(T-1)} & 0 & \cdots & 0 \\ 0 & \frac{\partial Y_2(T-1)}{\partial Z_2^{(2)}(T-1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial Y_M(T-1)}{\partial Z_M^{(2)}(T-1)} \end{bmatrix}$$

$$= \nabla_{Y(T-1)} DIV \;\; \nabla_{Z^{(2)}(T-1)} Y(T-1)$$

Back Propagation Through Time



$$\frac{dDIV}{dh_i(T-1)} = \sum_j w_{ij}^{(2)} \frac{dDIV}{dZ_j^{(2)}(T-1)} + \sum_j w_{ij}^{(11)} \frac{dDIV}{dZ_j^{(1)}(T)}$$

$$\nabla_{h(T-1)} DIV = \nabla_{Z^{(2)}(T-1)} DIV \; W^{(2)} + \nabla_{Z^{(1)}(T)} DIV \; W^{(11)}$$

Compute $\nabla_{\mathbf{h}(T-1)} DIV$

1. $h_i(T-1)$ is an input to all the $Z_1^{(2)}(T-1), \ldots, Z_M^{(2)}(T-1)$ where $Z_j^{(2)}(T-1) = \sum_{k=1}^{N_1} w_{kj}^{(2)} h_k(T-1) + b_j^{(2)}$.
2. $h_i(T-1)$ is an input to all the $Z_1^{(1)}(T), \ldots, Z_{N_1}^{(1)}(T)$ where $Z_j^{(1)}(T) = \sum_{k=1}^{N} w_{kj}^{(2)} X_k(T) + \sum_{k=1}^{N_1} w_{kj}^{(11)} h_k(T-1) + b_j^{(1)}$.
3. Thus, we see that,

$$\frac{\partial DIV}{\partial h_i(T-1)} = \sum_{j=1}^{M} \frac{\partial DIV}{\partial Z_j^{(2)}(T-1)} \frac{\partial Z_j^{(2)}(T-1)}{\partial h_i(T-1)} + \sum_{j=1}^{N_1} \frac{\partial DIV}{\partial Z_j^{(1)}(T)} \frac{\partial Z_j^{(1)}(T)}{\partial h_i(T-1)}$$

$$= \sum_{j=1}^{M} w_{ij}^{(2)} \frac{\partial DIV}{\partial Z_j^{(2)}(T-1)} + \sum_{j=1}^{N_1} w_{ij}^{(11)} \frac{\partial DIV}{\partial Z_j^{(1)}(T)}$$

4. Computing $\nabla_{\mathbf{h}(T-1)} DIV = \begin{bmatrix} \frac{\partial DIV}{\partial h_1(T-1)} & \cdots & \frac{\partial DIV}{\partial h_{N_1}(T-1)} \end{bmatrix}$

$$\nabla_{\mathbf{h}(T-1)} DIV = \begin{bmatrix} \sum_{j=1}^{M} \frac{\partial DIV}{\partial Z_j^{(2)}(T-1)} \frac{\partial Z_j^{(2)}(T-1)}{\partial h_1(T-1)} & \cdots & \sum_{j=1}^{M} \frac{\partial DIV}{\partial Z_j^{(2)}(T-1)} \frac{\partial Z_j^{(2)}(T-1)}{\partial h_{N_1}(T-1)} \end{bmatrix}$$

$$+ \begin{bmatrix} \sum_{j=1}^{N_1} \frac{\partial DIV}{\partial Z_j^{(1)}(T)} \frac{\partial Z_j^{(1)}(T)}{\partial h_1(T-1)} & \cdots & \sum_{j=1}^{N_1} \frac{\partial DIV}{\partial Z_j^{(1)}(T)} \frac{\partial Z_j^{(1)}(T)}{\partial h_{N_1}(T-1)} \end{bmatrix}$$

Compute $\nabla_{\mathbf{h}(T-1)} DIV$

$$= \left[\frac{\partial DIV}{\partial Z_1^{(2)}(T-1)} \quad \cdots \quad \frac{\partial DIV}{\partial Z_M^{(2)}(T-1)} \right] \begin{bmatrix} \frac{\partial Z_1^{(2)}(T-1)}{\partial h_1(T-1)} & \frac{\partial Z_1^{(2)}(T-1)}{\partial h_2(T-1)} & \cdots & \frac{\partial Z_1^{(2)}(T-1)}{\partial h_{N_1}(T-1)} \\ \frac{\partial Z_2^{(2)}(T-1)}{\partial h_1(T-1)} & \frac{\partial Z_2^{(2)}(T-1)}{\partial h_2(T-1)} & \cdots & \frac{\partial Z_2^{(2)}(T-1)}{\partial h_{N_1}(T-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Z_M^{(2)}(T-1)}{\partial h_1(T-1)} & \frac{\partial Z_M^{(2)}(T-1)}{\partial h_2(T-1)} & \cdots & \frac{\partial Z_M^{(2)}(T-1)}{\partial h_{N_1}(T-1)} \end{bmatrix}$$

$$+ \left[\frac{\partial DIV}{\partial Z_1^{(1)}(T)} \quad \cdots \quad \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T)} \right] \begin{bmatrix} \frac{\partial Z_1^{(1)}(T)}{\partial h_1(T-1)} & \frac{\partial Z_1^{(1)}(T)}{\partial h_2(T-1)} & \cdots & \frac{\partial Z_1^{(1)}(T)}{\partial h_{N_1}(T-1)} \\ \frac{\partial Z_2^{(1)}(T)}{\partial h_1(T-1)} & \frac{\partial Z_2^{(1)}(T)}{\partial h_2(T-1)} & \cdots & \frac{\partial Z_2^{(1)}(T)}{\partial h_{N_1}(T-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Z_{N_1}^{(1)}(T)}{\partial h_1(T-1)} & \frac{\partial Z_{N_1}^{(1)}(T)}{\partial h_2(T-1)} & \cdots & \frac{\partial Z_{N_1}^{(1)}(T)}{\partial h_{N_1}(T-1)} \end{bmatrix}$$

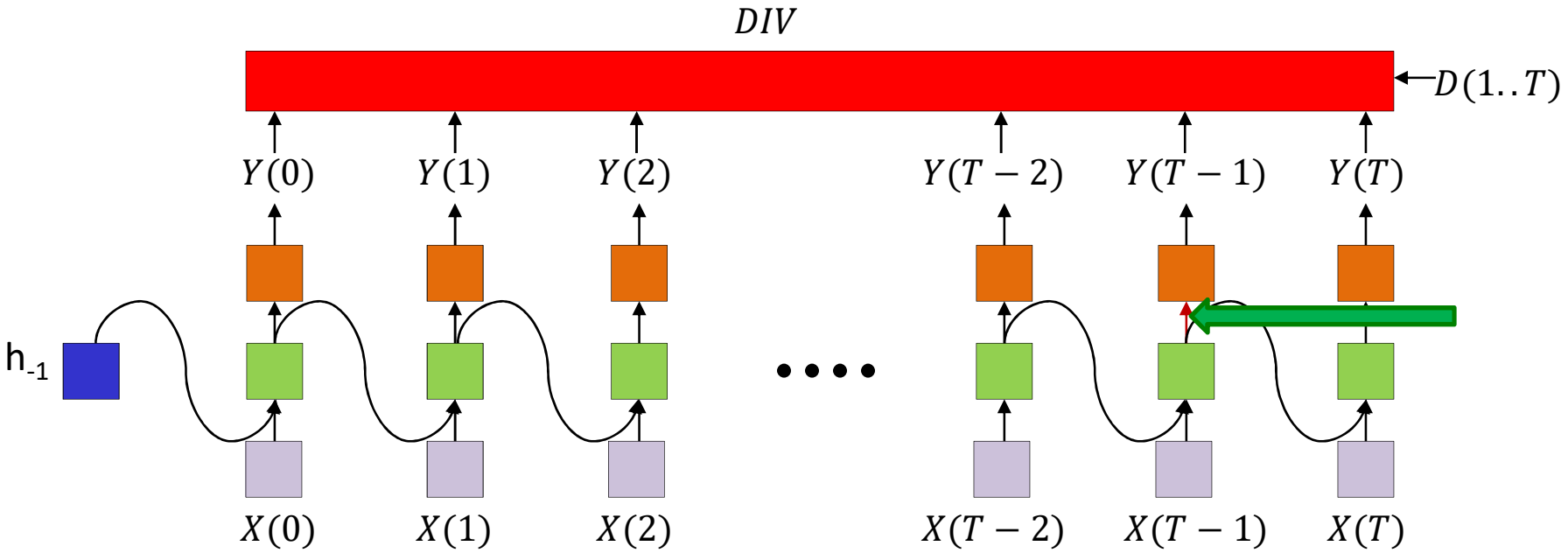
Compute $\nabla_{\mathbf{h}(T-1)} DIV$

$$= \nabla_{Z^{(2)}(T-1)} DIV \begin{bmatrix} w_{11}^{(2)} & w_{21}^{(2)} & \cdots & w_{N_1 1}^{(2)} \\ w_{12}^{(2)} & w_{22}^{(2)} & \cdots & w_{N_1 2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1M}^{(2)} & w_{2M}^{(2)} & \cdots & w_{N_1 M}^{(2)} \end{bmatrix}$$

$$+ \nabla_{Z^{(1)}(T)} DIV \begin{bmatrix} w_{11}^{(11)} & w_{21}^{(11)} & \cdots & w_{N_1 1}^{(11)} \\ w_{12}^{(11)} & w_{22}^{(11)} & \cdots & w_{N_1 2}^{(11)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1N_1}^{(11)} & w_{2N_1}^{(11)} & \cdots & w_{N_1 N_1}^{(11)} \end{bmatrix}$$

$$= \nabla_{Z^{(2)}(T-1)} DIV \; W^{(2)} + \nabla_{Z^{(1)}(T)} DIV \; W^{(11)}$$

Back Propagation Through Time



$$\frac{dDIV}{dh_i(T-1)} = \sum_j w_{ij}^{(2)} \frac{dDIV}{dZ_j^{(2)}(T-1)} + \sum_j w_{ij}^{(11)} \frac{dDIV}{dZ_j^{(1)}(T)}$$

Note the addition

$$\frac{dDIV}{dw_{ij}^{(2)}} \mathrel{+}= \frac{dDIV}{dZ_j^{(2)}(T-1)} h_i(T-1)$$

$$\nabla_{W^{(2)}} DIV \mathrel{+}= h(T-1) \nabla_{Z^{(2)}(T-1)} DIV$$

Updating $\nabla_{W^{(2)}} DIV$

$$\frac{\partial DIV}{\partial w_{ij}^{(2)}}_{|new} = \frac{\partial DIV}{\partial w_{ij}^{(2)}}_{|old} + \frac{\partial DIV}{\partial Z_j^{(2)}(T-1)} \frac{\partial Z_j^{(2)}(T-1)}{\partial w_{ij}^{(2)}}$$

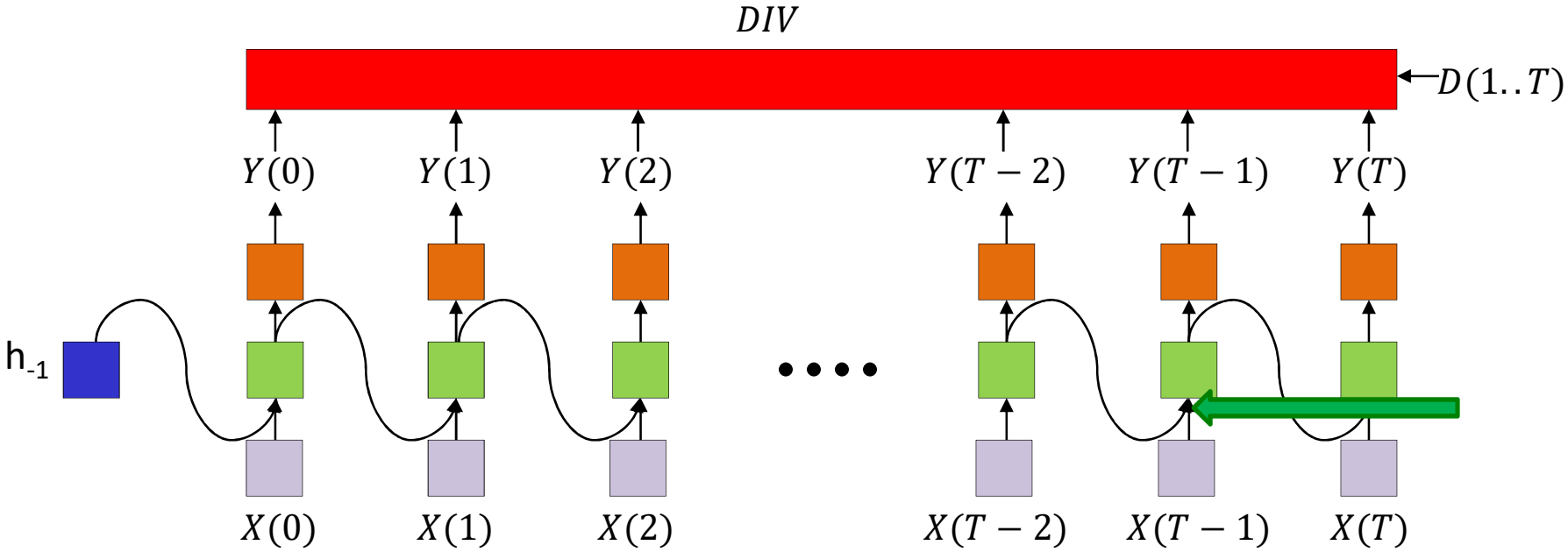
$$= \frac{\partial DIV}{\partial Z_j^{(2)}(T)} h_i(T) + \frac{\partial DIV}{\partial Z_j^{(2)}(T-1)} h_i(T-1)$$

$$\nabla_{W^{(2)}} DIV_{|new} = \nabla_{W^{(2)}} DIV_{|old}$$

$$+ \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(2)}(T-1)} h_1(T-1) & \frac{\partial DIV}{\partial Z_2^{(2)}(T-1)} h_1(T-1) & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T-1)} h_1(T-1) \\ \frac{\partial DIV}{\partial Z_1^{(2)}(T-1)} h_2(T-1) & \frac{\partial DIV}{\partial Z_2^{(2)}(T-1)} h_2(T-1) & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T-1)} h_2(T-1) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial Z_1^{(2)}(T-1)} h_{N_1}(T-1) & \frac{\partial DIV}{\partial Z_2^{(2)}(T-1)} h_{N_1}(T-1) & \cdots & \frac{\partial DIV}{\partial Z_M^{(2)}(T-1)} h_{N_1}(T-1) \end{bmatrix}$$

$$= \mathbf{h}(T) \nabla_{Z^{(2)}(T)} DIV + \mathbf{h}(T-1) \nabla_{Z^{(2)}(T-1)} DIV$$

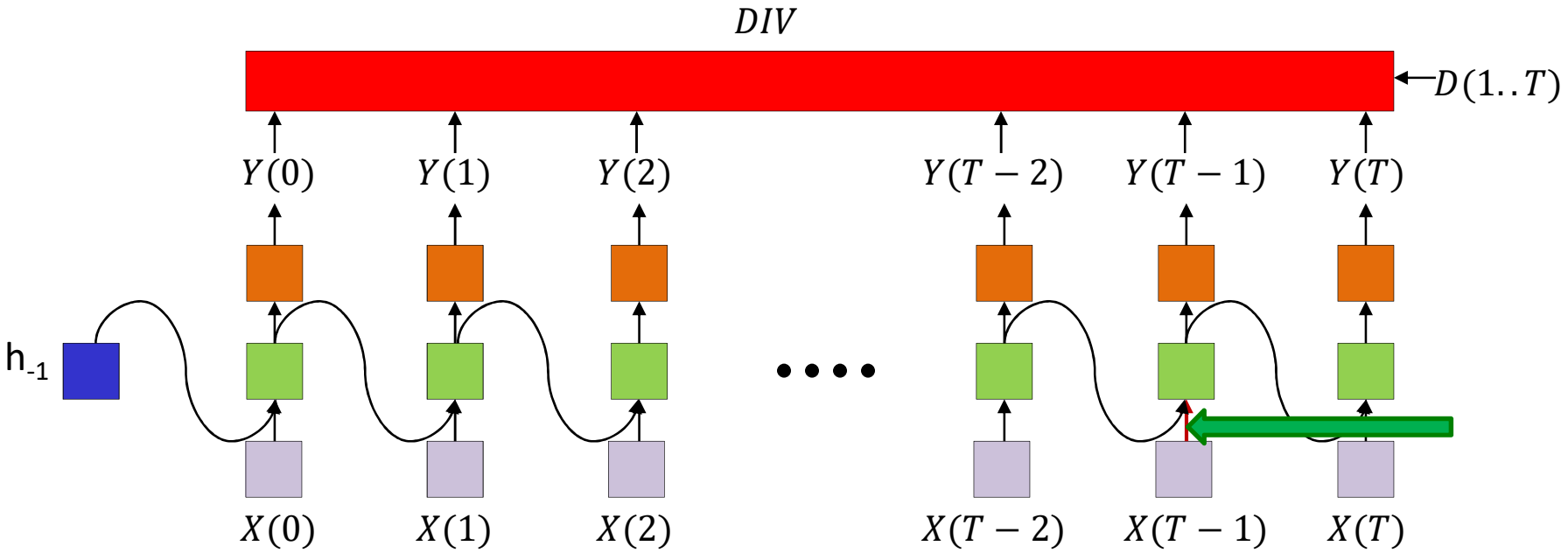
Back Propagation Through Time



$$\frac{dDIV}{dZ_i^{(1)}(T-1)} = \frac{dDIV}{dh_i(T-1)} \frac{dh_i(T-1)}{dZ_i^{(1)}(T-1)}$$

$$\nabla_{Z^{(1)}\,(T-1)} DIV = \nabla_{h(T-1)} DIV \, \nabla_{Z^{(1)}(T-1)} h(T-1)$$

Back Propagation Through Time



$$\frac{dDIV}{dZ_i^{(1)}(T-1)} = \frac{dDIV}{dh_i(T-1)} \frac{dh_i(T-1)}{dZ_i^{(1)}(T-1)}$$

$$\frac{dDIV}{dw_{ij}^{(1)}} \mathrel{+}= \frac{dDIV}{dZ_j^{(1)}(T-1)} X_i(T-1)$$

Note the addition

$$\nabla_{W^{(1)}} DIV \mathrel{+}= X(T-1) \nabla_{Z^{(1)}(T-1)} DIV$$

Updating $\nabla_{W^{(1)}} DIV$

$$\frac{\partial DIV}{\partial w_{ij}^{(1)}}_{|new} = \frac{\partial DIV}{\partial w_{ij}^{(1)}}_{|old} + \frac{\partial DIV}{\partial Z_j^{(1)}(T-1)} \frac{\partial Z_j^{(1)}(T-1)}{\partial w_{ij}^{(1)}}$$

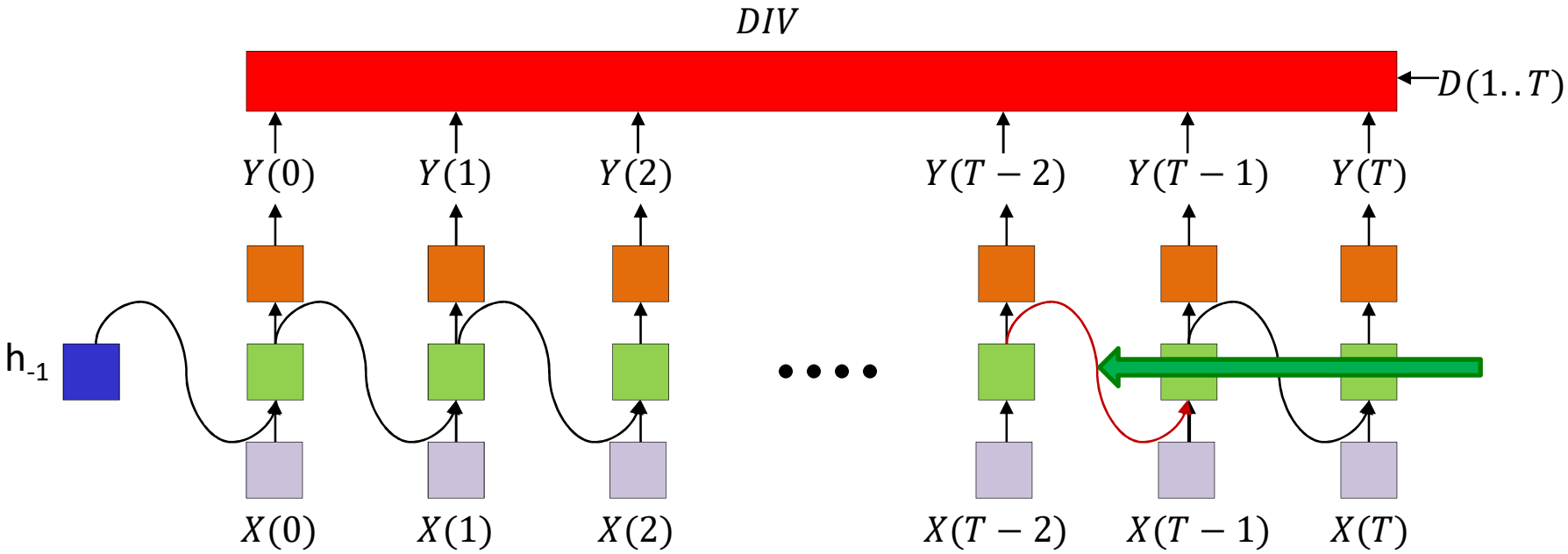
$$= \frac{\partial DIV}{\partial Z_j^{(1)}(T)} X_i(T) + \frac{\partial DIV}{\partial Z_j^{(1)}(T-1)} X_i(T-1)$$

$$\nabla_{W^{(1)}} DIV_{|new} = \nabla_{W^{(1)}} DIV_{|old}$$

$$+ \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T-1)} X_1(T-1) & \frac{\partial DIV}{\partial Z_2^{(1)}(T-1)} X_1(T-1) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T-1)} X_1(T-1) \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T-1)} X_2(T-1) & \frac{\partial DIV}{\partial Z_2^{(1)}(T-1)} X_2(T-1) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T-1)} X_2(T-1) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T-1)} X_N(T-1) & \frac{\partial DIV}{\partial Z_2^{(1)}(T-1)} X_N(T-1) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T-1)} X_N(T-1) \end{bmatrix}$$

$$= X(T) \nabla_{Z^{(1)}(T)} DIV + X(T-1) \nabla_{Z^{(1)}(T-1)} DIV$$

Back Propagation Through Time



$$\frac{dDIV}{dw_{ij}^{(1)}} \mathrel{+}= \frac{dDIV}{dZ_j^{(1)}(T-1)} X_i(T-1)$$

Note the addition

$$\frac{dDIV}{dw_{ij}^{(11)}} \mathrel{+}= \frac{dDIV}{dZ_j^{(1)}(T-1)} h_i(T-2)$$

$$\nabla_{W^{(11)}} DIV \mathrel{+}= h(T-2) \nabla_{Z^{(1)}(T-1)} DIV$$

Updating $\nabla_{W^{(11)}} DIV$

$$\frac{\partial DIV}{\partial w_{ij}^{(11)}}_{|new} = \frac{\partial DIV}{\partial w_{ij}^{(11)}}_{|old} + \frac{\partial DIV}{\partial Z_j^{(1)}(T-1)} \frac{\partial Z_j^{(1)}(T-1)}{\partial w_{ij}^{(11)}}$$

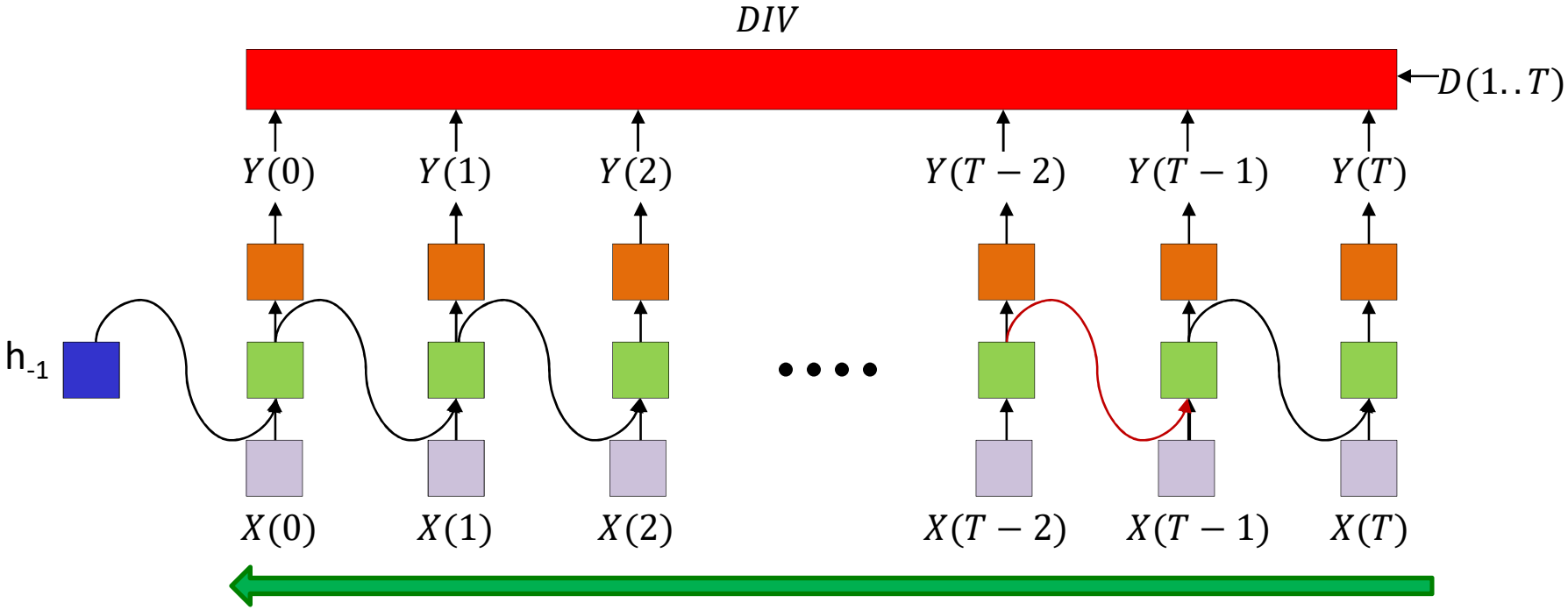
$$= \frac{\partial DIV}{\partial Z_j^{(1)}(T)} h_i(T-1) + \frac{\partial DIV}{\partial Z_j^{(1)}(T-1)} h_i(T-2)$$

$$\nabla_{W^{(11)}} DIV_{|new} = \nabla_{W^{(11)}} DIV_{|old}$$

$$+ \begin{bmatrix} \frac{\partial DIV}{\partial Z_1^{(1)}(T-1)} h_1(T-2) & \frac{\partial DIV}{\partial Z_2^{(1)}(T-1)} h_1(T-2) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T-1)} h_1(T-2) \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T-1)} h_2(T-2) & \frac{\partial DIV}{\partial Z_2^{(1)}(T-1)} h_2(T-2) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T-1)} h_2(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial DIV}{\partial Z_1^{(1)}(T-1)} h_{N_1}(T-2) & \frac{\partial DIV}{\partial Z_2^{(1)}(T-1)} h_{N_1}(T-2) & \cdots & \frac{\partial DIV}{\partial Z_{N_1}^{(1)}(T-1)} h_{N_1}(T-2) \end{bmatrix}$$

$$= \mathbf{h}(T-1) \nabla_{Z^{(1)}(T)} DIV + \mathbf{h}(T-2) \nabla_{Z^{(1)}(T-1)} DIV$$

Back Propagation Through Time

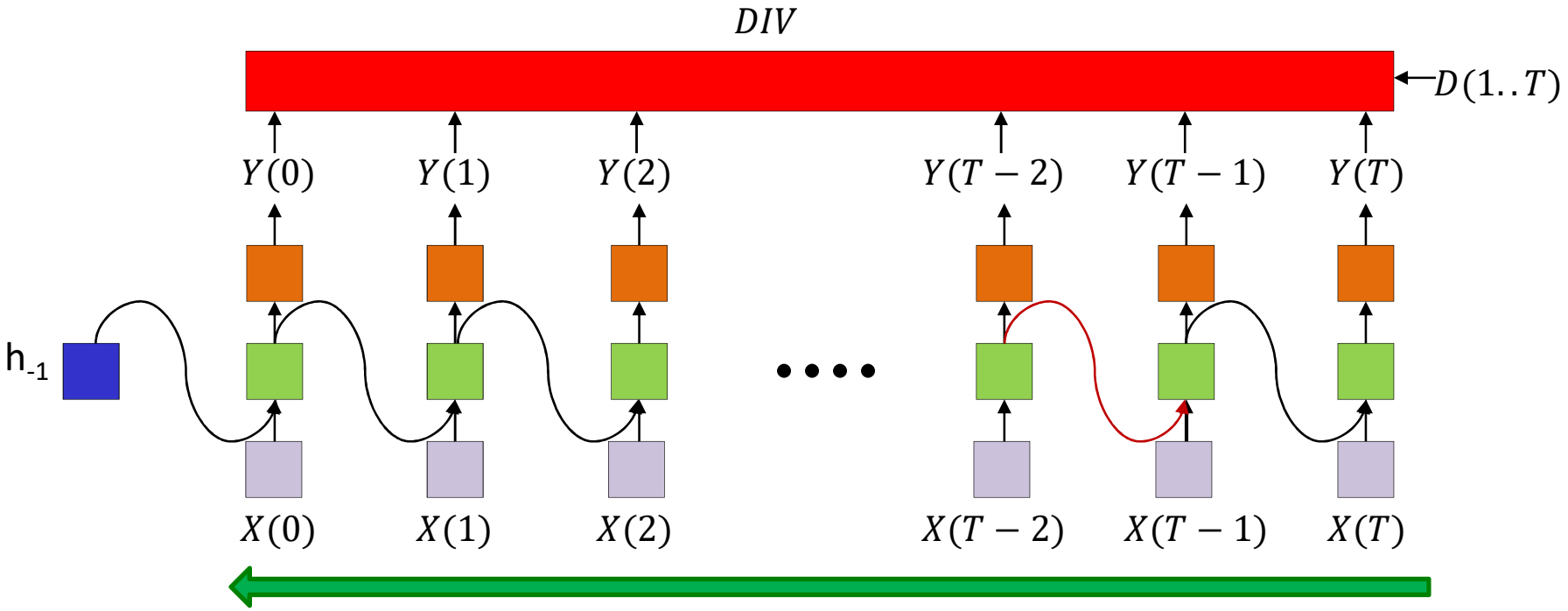


Continue computing derivatives going backward through time until..

$$\frac{dDIV}{dh_i(-1)} = \sum_j w_{ij}^{(11)} \frac{dDIV}{dZ_j^{(1)}(0)}$$

$$\nabla_{h_{-1}} DIV = \nabla_{Z^{(1)}(0)} DIV W^{(11)}$$

Back Propagation Through Time

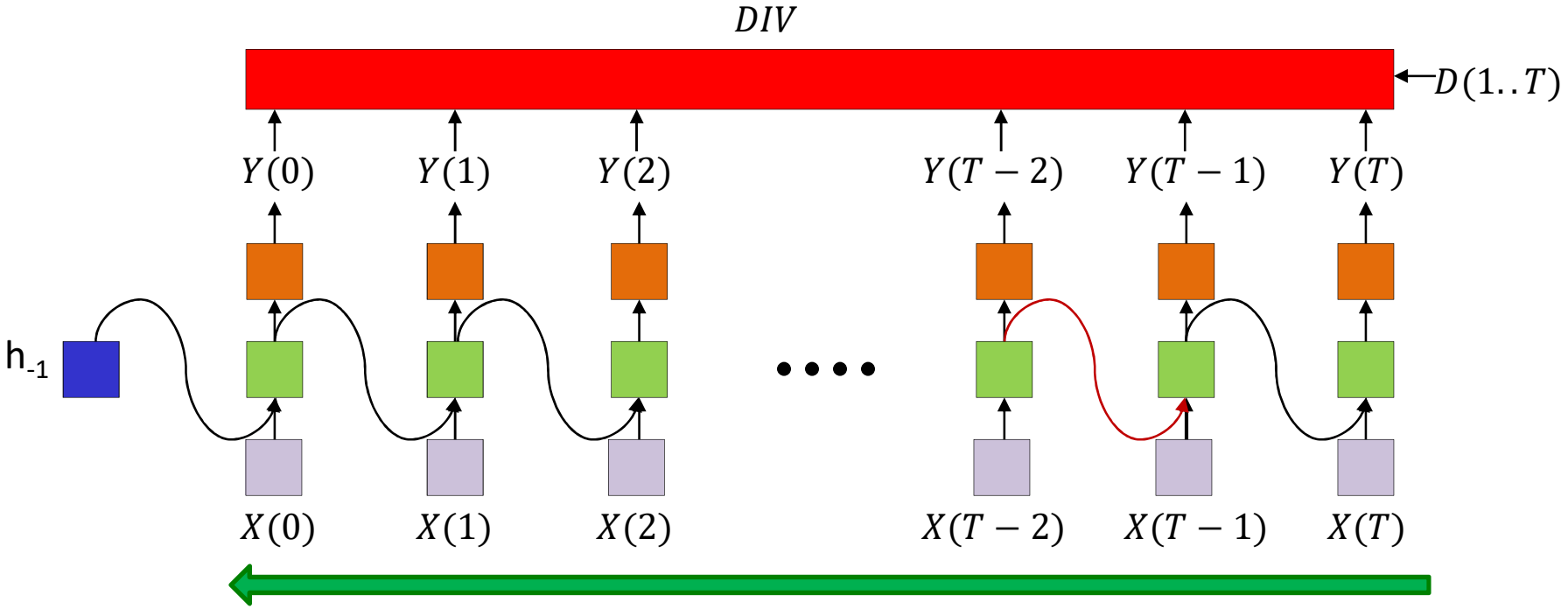


$$\frac{dDIV}{dh_i^{(k)}(t)} = \sum_j w_{i,j}^{(k+1)} \frac{dDIV}{dZ_j^{(k+1)}(t)} + \sum_j w_{i,j}^{(k,k)} \frac{dDIV}{dZ_j^{(k)}(t+1)}$$

Not showing derivatives at output neurons

$$\frac{dDIV}{dZ_i^{(k)}(t)} = \frac{dDIV}{dh_i^{(k)}(t)} f_k'\left(Z_i^{(k)}(t)\right)$$

Back Propagation Through Time

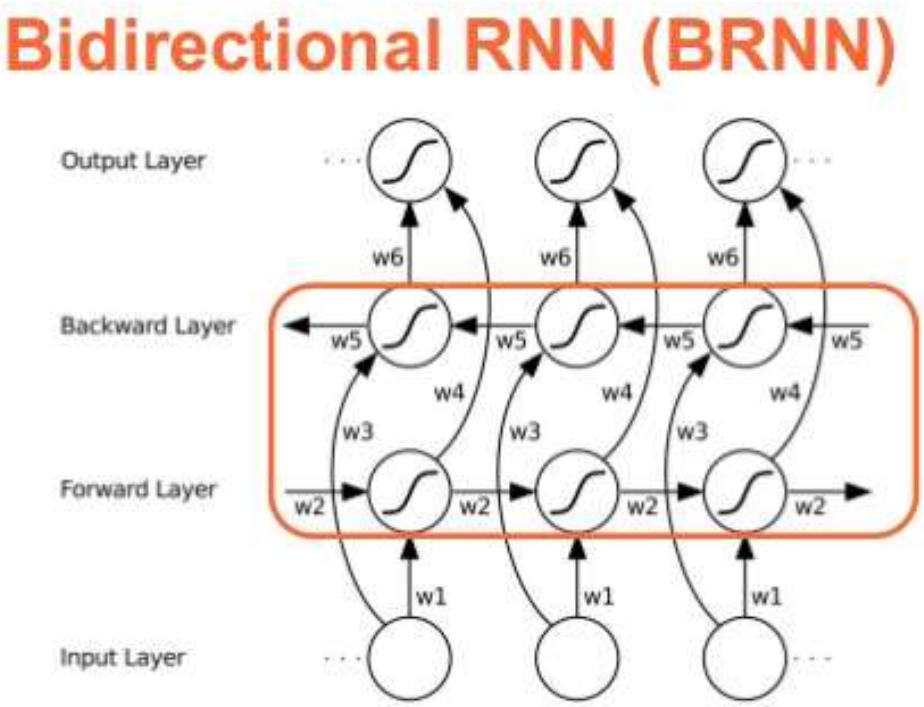


$$\frac{dDIV}{dh_i(-1)} = \sum_j w_{ij}^{(11)} \frac{dDIV}{dZ_j^{(1)}(0)}$$

$$\frac{dDIV}{dw_{ij}^{(1)}} = \sum_t \frac{dDIV}{dZ_j^{(1)}(t)} X_i(t)$$

$$\frac{dDIV}{dw_{ij}^{(11)}} = \sum_t \frac{dDIV}{dZ_j^{(1)}(t)} h_i(t-1)$$

Extensions to the RNN: *Bidirectional RNN*



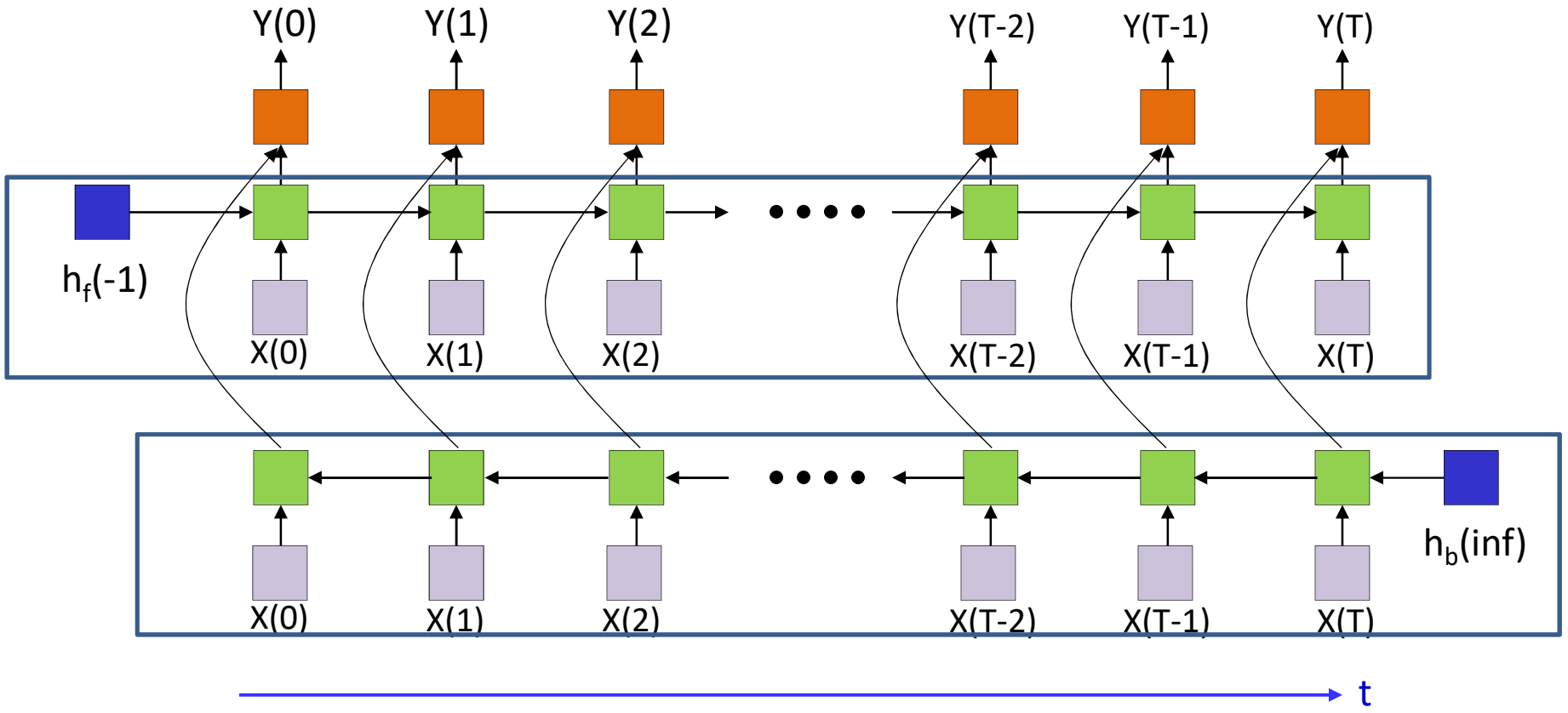
Must learn weights w2, w3, w4 & w5; in addition to w1 & w6.

Proposed by Schuster and Paliwal 1997

Alex Graves, "Supervised Sequence Labelling with Recurrent Neural Networks"

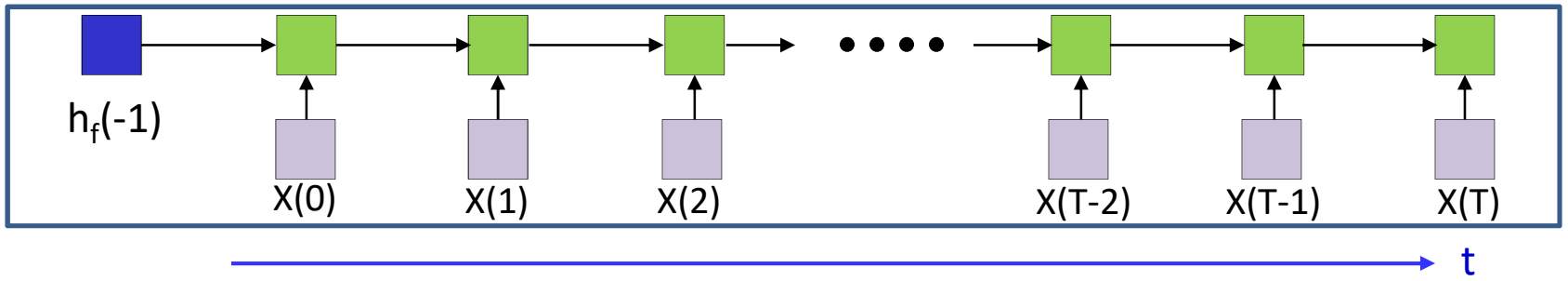
- RNN with both forward and backward recursion
  - Explicitly models the fact that just as the future can be predicted from the past, the past can be deduced from the future

Bidirectional RNN



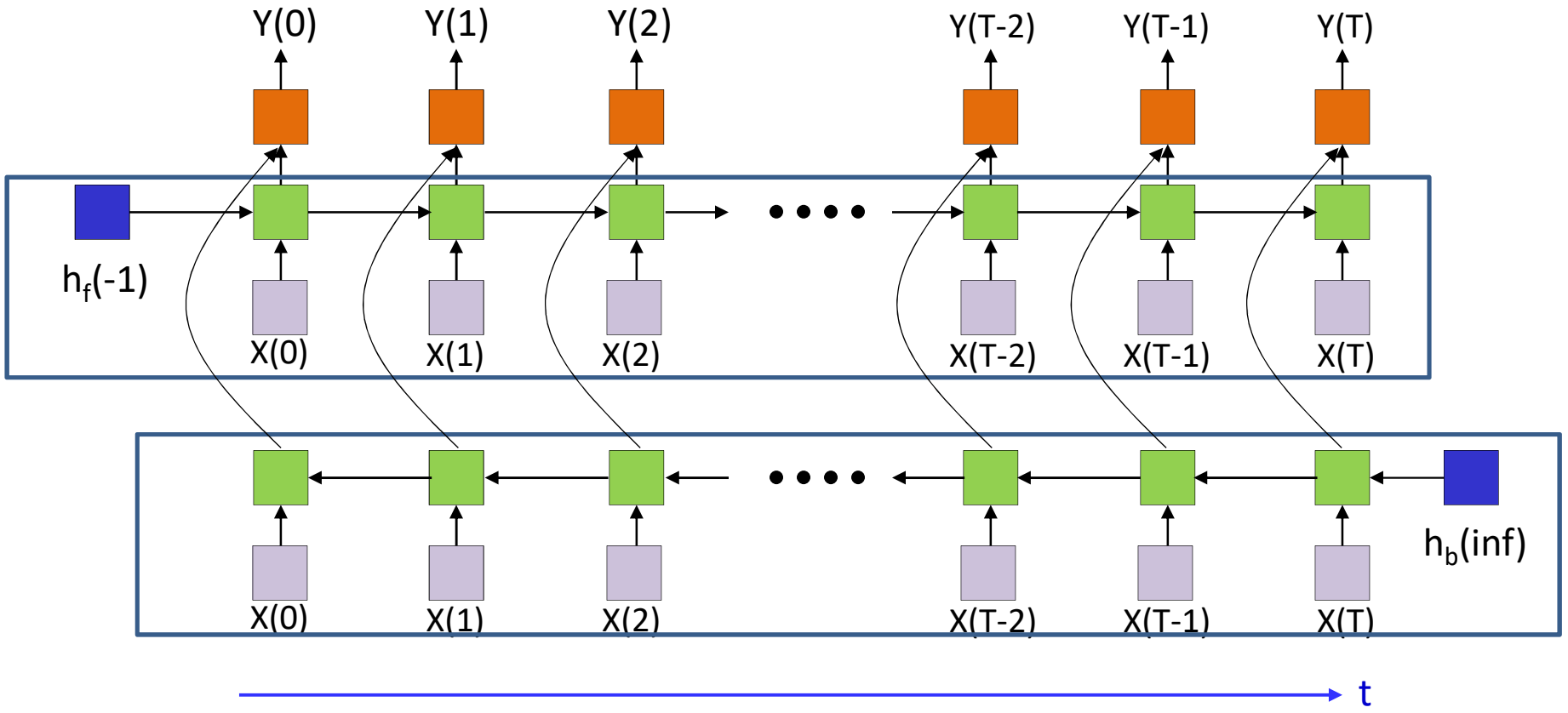
- A forward net process the data from t=0 to t=T
- A backward net processes it backward from t=T down to t=0

Bidirectional RNN: Processing an input string



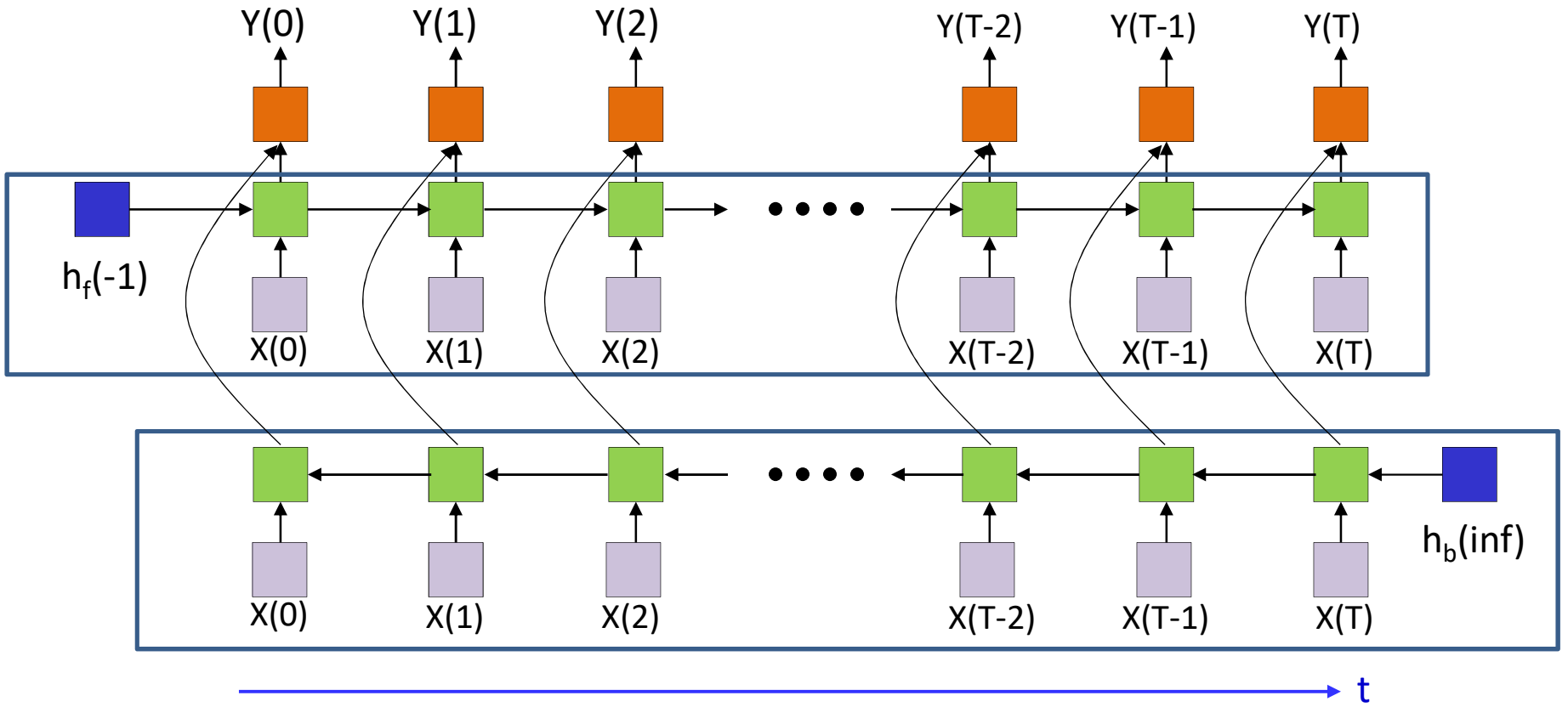
- The forward net process the data from t=0 to t=T
  - Only computing the hidden states, initially

Bidirectional RNN: Processing an input string



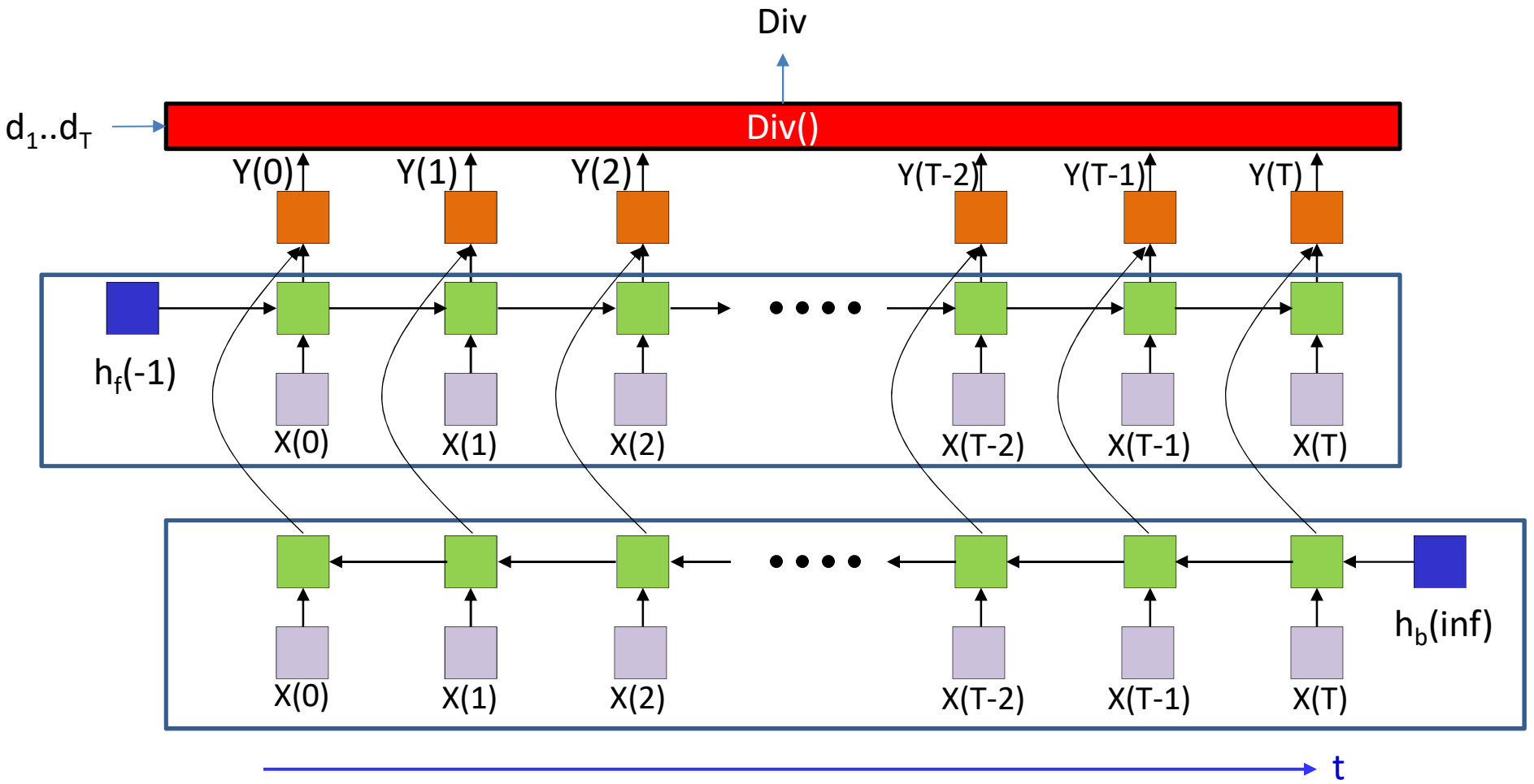
- The computed states of both networks are used to compute the final output at each time

Backpropagation in BRNNs



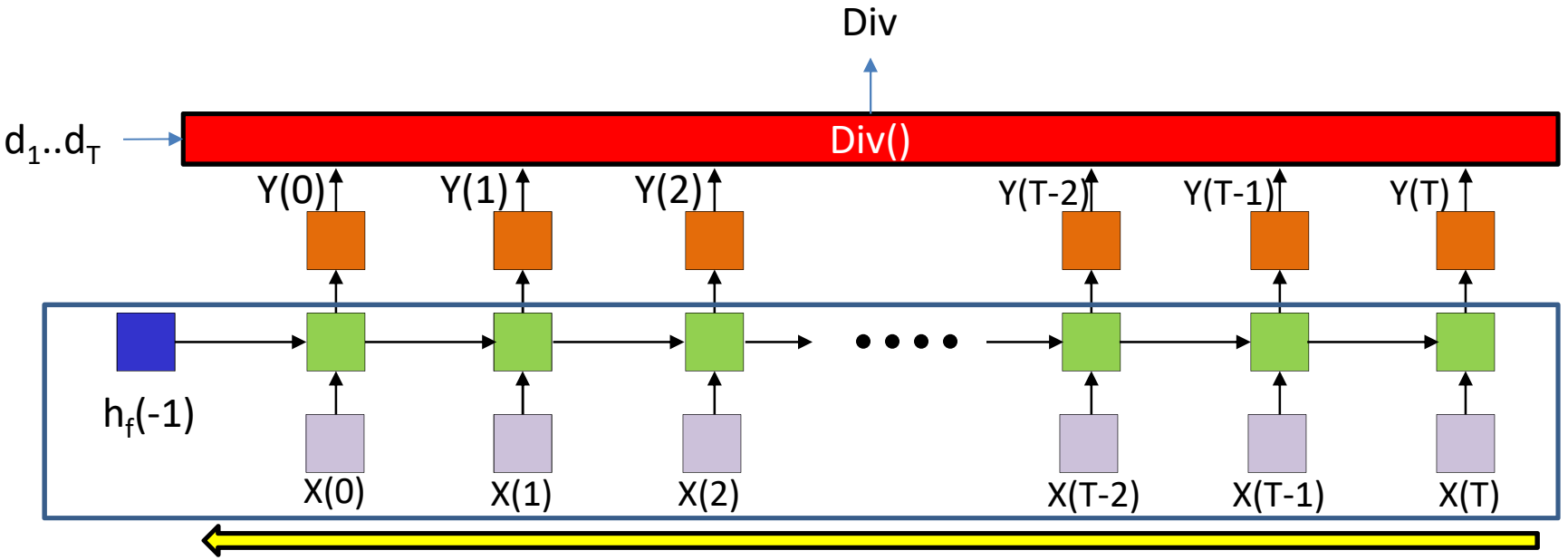
- Forward pass: Compute both forward and backward networks and final output

Backpropagation in BRNNs



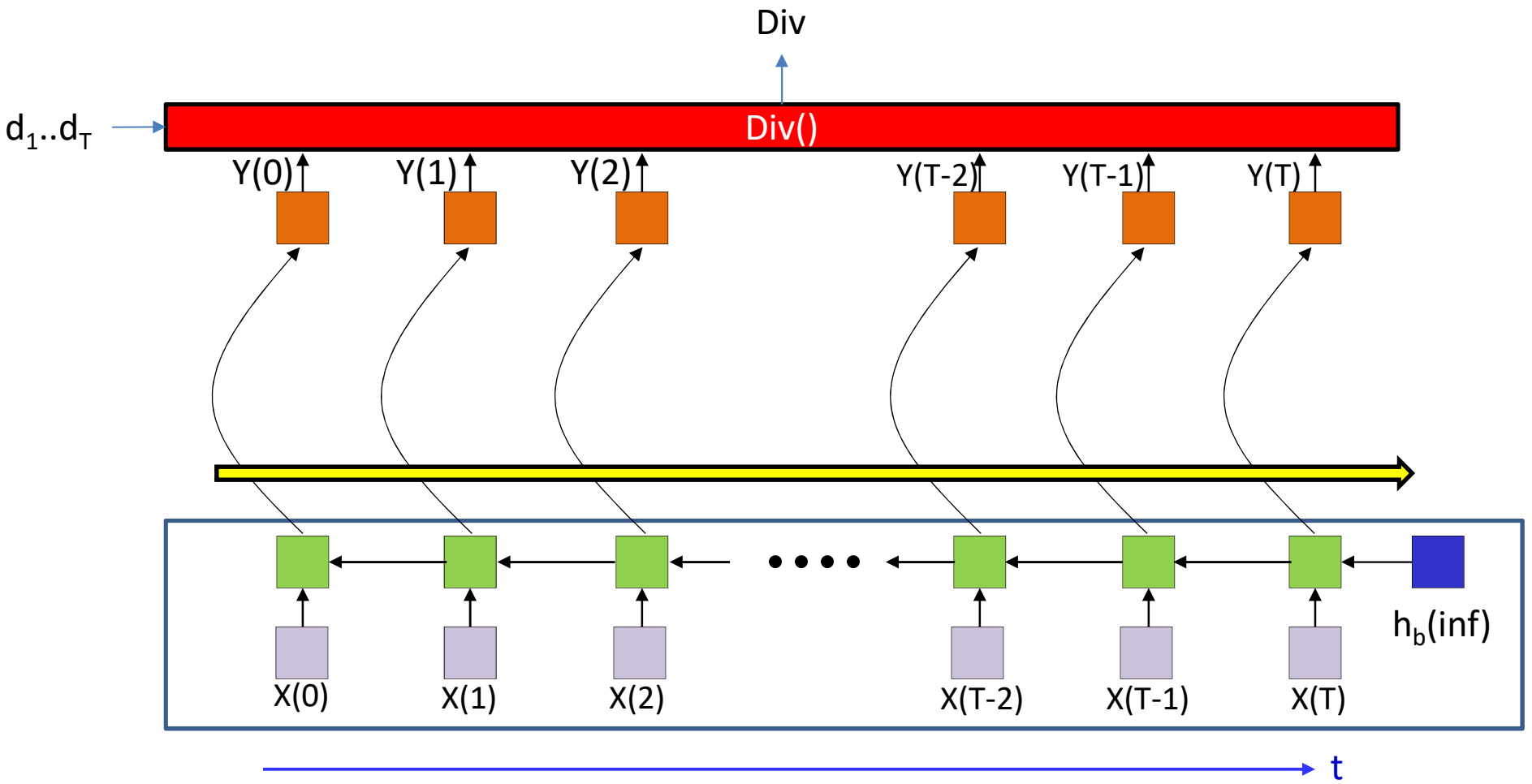
- Backward pass: Define a divergence from the desired output

Backpropagation in BRNNs



- Backward pass: Define a divergence from the desired output
- Separately perform back propagation on both nets
  - **From t=T down to t=0 for the forward net**

Backpropagation in BRNNs



- Backward pass: Define a divergence from the desired output
- Separately perform back propagation on both nets
  - From t=T down to t=0 for the forward net
  - **From t=0 up to t=T for the backward net**

Need for Bidirectional RNN

1. Consider the following three tasks of filling in the blanks in a text:
   - I am _____
   - I am _____ very hungry.
   - I am _____ very hungry, I could eat half a pig.
2. We might fill the blanks with very different words such as "happy", "not", and "very".
3. End of the phrase (if available) conveys significant information about which word to pick.
4. A sequence model should be capable of taking advantage of this information.
5. For instance, to recognize whether "Green" refers to "Mr. Green" or to the color, longer-range context is equally vital.