Back Propagation Through Time

Yongqiang Huang

November 2019

1 Scalar form

First we look at the over-simplified scalar version where all variables are one-dimensional. Let our network be a binary classifier:

$$s_t = tanh(Ux_t + Ws_{t-1}) \tag{1}$$
$$\hat{y}_t = sigm(Vs_t) \tag{2}$$
$$E_t = -y_t \log \hat{y}_t \tag{3}$$

Assume $t \geq 1$ and s_0 is given, that is s_0 does not depend on any other variable. We want to compute $\frac{\partial E_t}{\partial V}$, $\frac{\partial E_t}{\partial W}$, and $\frac{\partial E_t}{\partial U}$. First, let us compute $\frac{\partial E_t}{\partial V}$.

$$\frac{\partial E_t}{\partial V} = \frac{\partial E_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial V} \tag{4}$$

Then, we compute $\frac{\partial E_t}{\partial W}$:

$$\frac{\partial E_t}{\partial W} = \frac{\partial E_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial s_t} \frac{\partial s_t}{\partial W} \tag{5}$$

Notice from Eq. (1) that s_t is a function of both W and s_{t-1}. Before going forward, we need to review the concept of total derivative, which says the following. If function $z = f(u, v)$ where $u = u(x)$, $v = v(x)$, then the total derivative of z with respect to x is

$$\frac{dz}{dx} = \frac{\partial z}{\partial u} \frac{du}{dx} + \frac{\partial z}{\partial v} \frac{dv}{dx} \tag{6}$$

Notice the definition does not confine the exact form of $f(\cdot)$, it does not have to be addition, nor does it have to be multiplication, or specifically anything else. As long as it involves different functions of x, the functions would contribute to the total derivative linearly with the same weight 1. Okay, back to what we were saying. $s_t = tanh(Ux_t + Ws_{t-1})$, here W is a function of W itself, and s_{t-1} is a function of W, so using the definition of the total derivative in Eq. (6), we have

$$\frac{\partial s_t}{\partial W} = \frac{\partial s_t}{\partial W} \frac{\partial W}{\partial W} + \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial W} \tag{7}$$

Hence the derivative is recursive, and it stops at s_1 which depends on s_0. For example, $\frac{\partial s_3}{\partial W}$ expands to

$$\frac{\partial s_3}{\partial W} = \frac{\partial s_3}{\partial W} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial W} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial s_1}\frac{\partial s_1}{\partial W} \tag{8}$$

$$= \frac{\partial s_3}{\partial s_3}\frac{\partial s_3}{\partial W} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial W} + \frac{\partial s_3}{\partial s_1}\frac{\partial s_1}{\partial W} \tag{9}$$

$$= \sum_{1\le k\le 3} \frac{\partial s_3}{\partial s_k}\frac{\partial s_k}{\partial W} \tag{10}$$

The expansion shows that the $\frac{\partial s_3}{\partial W}$ goes through s_3, s_2, and s_1, in every s for which W is an input. We can generalize the formula of s_3 to s_t:

$$\frac{\partial s_t}{\partial W} = \sum_{1\le k\le t} \frac{\partial s_t}{\partial s_k}\frac{\partial s_k}{\partial W} \tag{11}$$

$$= \frac{\partial s_t}{\partial s_t}\frac{\partial s_t}{\partial W} + \frac{\partial s_t}{\partial s_{t-1}}\frac{\partial s_{t-1}}{\partial W} + \frac{\partial s_t}{\partial s_{t-2}}\frac{\partial s_{t-2}}{\partial W} + \cdots + \frac{\partial s_t}{\partial s_1}\frac{\partial s_1}{\partial W} \tag{12}$$

Let's take a closer look at each individual term $\frac{\partial s_t}{\partial s_k}\frac{\partial s_k}{\partial W}$. If we expand it, we have

$$\frac{\partial s_t}{\partial s_k}\frac{\partial s_k}{\partial W} = \frac{\partial s_t}{\partial s_{t-1}}\frac{\partial s_{t-1}}{\partial s_{t-2}}\frac{\partial s_{t-2}}{\partial s_{t-3}}\cdots\frac{\partial s_{k+1}}{\partial s_k}\frac{\partial s_k}{\partial W} \tag{13}$$

$$= W \cdot W \cdot W \cdots \frac{\partial s_k}{\partial W} \tag{14}$$

$$= W^{t-k}\frac{\partial s_k}{\partial W} \tag{15}$$

Thus we can rewrite Eq. (11) as

$$\frac{\partial s_t}{\partial W} = \sum_{1\le k\le t} W^{t-k}\frac{\partial s_k}{\partial W} \tag{16}$$

Lastly, we compute $\frac{\partial E_t}{\partial U}$. U is an input to s_t, which means we will treat U similarly to the way we treat W:

$$\frac{\partial E_t}{\partial U} = \frac{\partial E_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial s_t}\frac{\partial s_t}{\partial U} \tag{17}$$

where

$$\frac{\partial s_t}{\partial U} = \sum_{1\le k\le t} W^{t-k}\frac{\partial s_k}{\partial U} \tag{18}$$

2 Vector form, single data point

Now that we are comfortable with the scalar version, we are proceeding to the vector version, which requires us to define the network again, a little differently:

$$z_t = Ux_t + Ws_{t-1} \tag{19}$$
$$s_t = \tanh(z_t) \tag{20}$$
$$g_t = Vs_t \tag{21}$$
$$\hat{y}_t = \text{softmax}(g_t) \tag{22}$$
$$L_t = -y_t^\top \log(\hat{y}_t) \tag{23}$$

The dimensions of the variables are

$$x_t : M \times 1 \tag{24}$$
$$z_t, s_t : D \times 1 \tag{25}$$
$$g_t, \hat{y}_t, y_t : C \times 1 \tag{26}$$
$$U : D \times M \tag{27}$$
$$W : D \times D \tag{28}$$
$$V : C \times D \tag{29}$$

L_t is the loss at time t, which can also be written using summation:

$$L_t = -\sum_{i=1}^{C} y_{t,i} \log(\hat{y}_{t,i}) \tag{30}$$

The back propagation starts with $\frac{\partial L_t}{\partial \hat{y}_t}$, which has dimension $1 \times C$: we have only one output variable L_t and C input variables $\hat{y}_{t,1}, \hat{y}_{t,2}, \cdots, \hat{y}_{t,C}$.

$$\frac{\partial L_t}{\partial \hat{y}_t} = [\frac{\partial L_{t,1}}{\partial \hat{y}_{t,1}}, \frac{\partial L_{t,2}}{\partial \hat{y}_{t,2}}, \cdots, \frac{\partial L_{t,C}}{\partial \hat{y}_{t,C}}] \tag{31}$$

If we consider $\log() = \ln()$, i.e. natual logarithm, then

$$\frac{\partial L_t}{\partial \hat{y}_t} = [-\frac{y_{t,1}}{\hat{y}_{t,1}}, -\frac{y_{t,2}}{\hat{y}_{t,2}}, \cdots, -\frac{y_{t,C}}{\hat{y}_{t,C}}] \tag{32}$$

Next, we compute $\frac{\partial L_t}{\partial g_t}$:

$$\frac{\partial L_t}{\partial g_t} = \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial g_t} \tag{33}$$

Let's get the dimension right. $\frac{\partial L_t}{\partial g_t}$ should also have dimension $1 \times C$, and $\frac{\partial \hat{y}_t}{\partial g_t}$ should have dimension $C \times C$. The relation between $\hat{y}_t$ and g_t is as follows:

$$\hat{y}_{t,i} = \frac{\exp(g_{t,i})}{\sum_{i=j}^{C} \exp(g_{t,j})} \tag{34}$$

and the derivative

$$\frac{\partial \hat{y}_{t,i}}{\partial g_{t,j}} = \begin{cases} \hat{y}_{t,i}(1-\hat{y}_{t,i}) & \text{if } i = j \\ -\hat{y}_{t,i}\hat{y}_{t,j} & \text{if } i \neq j \end{cases} \tag{35}$$

or

$$\frac{\partial \hat{y}_{t,i}}{\partial g_{t,j}} = \hat{y}_{t,i}(\mathbf{1}(i=j) - \hat{y}_{t,j}) \tag{36}$$

where

$$\mathbf{1}(i=j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{37}$$

$\frac{\partial \hat{y}_{t,i}}{\partial g_{t,j}}$ is an element representation of matrix $\frac{\partial \hat{y}_t}{\partial g_t}$, it shows everything we need to know about matrix $\frac{\partial \hat{y}_t}{\partial g_t}$. Now let's compute $\frac{\partial L_t}{\partial g_t}$. To do that we only need to know its elements:

$$\frac{\partial L_t}{\partial g_{t,i}} = \sum_{k=1}^{C} \frac{\partial L_t}{\partial \hat{y}_{t,k}} \frac{\partial \hat{y}_{t,k}}{\partial g_{t,i}} \tag{38}$$

Let's pause here and see what it means. L_t goes to $g_{t_{t,i}}$ through $\hat{y}_t$, and more specifically, through *every* element of $\hat{y}_t$. Therefore, we need to consider *every* element of $\hat{y}_t$. Let's continue.

$$\begin{aligned} \frac{\partial L_t}{\partial g_{t,i}} &= -\sum_{k=1}^{C} \frac{y_{t,k}}{\hat{y}_{t,k}} \hat{y}_{t,k}(\mathbf{1}(k=i) - \hat{y}_{t,i}) && (39) \\ &= -\sum_{k=1}^{C} y_{t,k}(\mathbf{1}(k=i) - \hat{y}_{t,i}) && (40) \\ &= \sum_{k=1}^{C} y_{t,k}\hat{y}_{t,i} - y_{t,i} && (41) \\ &= (\sum_{k=1}^{C} y_{t,k})\hat{y}_{t,i} - y_{t,i} && (42) \\ \frac{\partial L_t}{\partial g_{t,i}} &= \hat{y}_{t,i} - y_{t,i} && (43) \end{aligned}$$

From Eq. (42) to Eq. (43), we assume y_t is a one-hot vector and therefore $\sum_{k=1}^{C} y_{t,k} = 1$. In vector form:

$$\frac{\partial L_t}{\partial g_t} = (\hat{y}_t - y_t)^\top \tag{44}$$

Now we are ready to compute the gradient of our first weight matrix V:

$$\frac{\partial L_t}{\partial V} = \frac{\partial L_t}{\partial g_t} \frac{\partial g_t}{\partial V} \tag{45}$$

Now there is a problem: what is $\frac{\partial g_t}{\partial V}$? Apparently its dimension should be $C \times (C \times D)$, which is intimidating, and that is why we will *not* directly compute it.

After all, we only care about $\frac{\partial L_t}{\partial V}$, which only requires that we know its element $\frac{\partial L_t}{\partial V_{i,j}}$. This is so important that it deserves our pausing here to strengthen it:

We *never* specifically write out a tensor, which has more than two dimensions. Rather, we represent the tensor using only its elements, which is equivalent. Let the input be x, and the output be y, where

$$y = Wx \tag{46}$$

and let the final loss be L, which is a scalar. We strive to always compute $\frac{\partial L}{\partial y}$ first, which for now, we assume to be a vector. Then we compute $\frac{\partial L}{\partial W}$ by computing $\frac{\partial L}{\partial W_{i,j}}$. We always keep a scalar(loss)-to-variable at hand before we go further in the back propagation. By the way, this makes the back-propagating process sequential.

Back to the derivation, L_t goes to $V_{i,j}$ through every single element in g_t, of which the consequence we sum up:

$$\frac{\partial L_t}{\partial V_{i,j}} = \sum_{k=1}^{C} \frac{\partial L_t}{\partial g_{t,k}} \frac{\partial g_{t,k}}{\partial V_{i,j}} \tag{47}$$

To know $\frac{\partial g_{t,k}}{\partial V_{i,j}}$, we need to know how g_t is computed from V. A row in g_t only uses the same row in V:

$$g_{t,i} = \sum_{d=1}^{D} V_{i,d} s_{t,d}, \quad (i = 1, 2, \ldots, C) \tag{48}$$

Therefore if we look at one row in g_t and a different row in V, the derivative will be 0. To summarize:

$$\begin{align}
\frac{\partial g_{t,k}}{\partial V_{i,j}} &= \begin{cases} s_{t,j} & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \tag{49} \\
&= \mathbf{1}(i = k) s_{t,j} \tag{50}
\end{align}$$

Now let's get back to solving $\frac{\partial L_t}{\partial V_{i,j}}$.

$$\begin{align}
\frac{\partial L_t}{\partial V_{i,j}} &= \sum_{k=1}^{C} \frac{\partial L_t}{\partial g_{t,k}} \frac{\partial g_{t,k}}{\partial V_{i,j}} \tag{51} \\
&= \sum_{k=1}^{C} (\hat{y}_{t,k} - y_{t,k}) \mathbf{1}(i = k) s_{t,j} \tag{52} \\
&= (\hat{y}_{t,i} - y_{t,i}) s_{t,j} \tag{53}
\end{align}$$

Thus, we can obtain the vector form of the gradient, which is an outer product:

$$\frac{\partial L_t}{\partial V} = (\hat{y}_t - y_t) s_t^\top \tag{54}$$

Next we compute $\frac{\partial L_t}{\partial W}$, which is

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial s_t}\frac{\partial s_t}{\partial W} \tag{55}$$

$\frac{\partial L_t}{\partial s_t}$ is computed by

$$\frac{\partial L_t}{\partial s_t} = \frac{\partial L_t}{\partial g_t}\frac{\partial g_t}{\partial s_t} = (\hat{y}_t - y_t)^\top V \tag{56}$$

$\frac{\partial s_t}{\partial W}$ involves recursive computation. Different from the scalar case, we will not write out the recursion naively, because it will be multi-dimensional. We will, as previously mentioned, write the recursion as part of the computation of a scalar-to-matrix chain:

$$\frac{\partial L_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_t}\frac{\partial s_t}{\partial s_k}\frac{\partial s_k}{\partial W} \tag{57}$$

Let's break the above equation apart and make it more specific. First we look at $\frac{\partial s_t}{\partial s_k}$:

$$\begin{aligned} \frac{\partial s_t}{\partial s_k} &= \frac{\partial s_t}{\partial s_{t-1}}\frac{\partial s_{t-1}}{\partial s_{t-2}}\cdots\frac{\partial s_{k+1}}{\partial s_k} && (58)\\ &= \prod_{i=k+1}^{t}\frac{\partial s_i}{\partial s_{i-1}} && (59)\\ &= \prod_{i=k+1}^{t}\frac{\partial s_i}{\partial z_i}\frac{\partial z_i}{\partial s_{i-1}} && (60)\\ & && (61) \end{aligned}$$

Since tanh() is element-wise, $\frac{\partial s_i}{\partial z_i}$ is a diagonal matrix:

$$\frac{\partial s_i}{\partial z_i} = \mathrm{diag}((1 - s_{i,1}^2), (1 - s_{i,2}^2), \cdots, (1 - s_{i,D}^2)) \tag{62}$$

Thus

$$\frac{\partial s_t}{\partial s_k} = \prod_{i=k+1}^{t}\frac{\partial s_i}{\partial z_i}W = W^{t-k}\prod_{i=k+1}^{t}\frac{\partial s_i}{\partial z_i} \tag{63}$$

The immediate derivative of s_k with respect to W is

$$\frac{\partial s_k}{\partial W} = \frac{\partial s_k}{\partial z_k}\frac{\partial z_k}{\partial W} \tag{64}$$

Thus the specific form of $\frac{\partial L_t}{\partial W}$ is

$$\frac{\partial L_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_t} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial W} \tag{65}$$

$$= \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_t} W^{t-k} \left(\prod_{i=k+1}^{t} \frac{\partial s_i}{\partial z_i} \right) \frac{\partial s_k}{\partial z_k} \frac{\partial z_k}{\partial W} \tag{66}$$

$$\frac{\partial L_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_t} W^{t-k} \left(\prod_{i=k}^{t} \frac{\partial s_i}{\partial z_i} \right) \frac{\partial z_k}{\partial W} \tag{67}$$

The gradient of U is very similar to that of W:

$$\frac{\partial L_t}{\partial U} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_t} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial U} \tag{68}$$

$$= \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_t} W^{t-k} \left(\prod_{i=k}^{t} \frac{\partial s_i}{\partial z_i} \right) \frac{\partial z_k}{\partial U} \tag{69}$$

One detail that will be used in the implementation of the above is, given $y = Wx$ and $\frac{\partial L}{\partial y}$, what is $\frac{\partial L}{\partial W}$? The answer is

$$\frac{\partial L}{\partial W} = (x \cdot \frac{\partial L}{\partial y})^\top \tag{70}$$

The final gradient with respect to the weights are:

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial V} \tag{71}$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \tag{72}$$

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial U} \tag{73}$$

3 Vector form, multiple data points