Homework 5 Solution

April 13, 2018

1

We apply the back-propagation from the output layer to the gates. To this end, we start from H_t. Since the hidden states and output are univariate variables, gradient of H up to time T can be written as:

$$\nabla H = \sum_{t=1}^{T} \frac{\partial l(Y_t, \hat{Y}_t)}{\partial H_t} = \sum_{t=1}^{T} \frac{\partial \hat{Y}_t}{\partial H_t} \nabla \hat{Y} \tag{1}$$

Since $\hat{Y}_t = \sigma(W^y H_t + W_0^y)$, at a specific time t, gradient of H_t can be written as:

$$\nabla H_t = \frac{\partial \hat{Y}_t}{\partial H_t} \nabla \hat{Y}_t = \hat{Y}_t(1 - \hat{Y}_t) W^y \nabla \hat{Y}_t \tag{2}$$

Then we can apply the chain rule to obtain the gradient of three gates.

$$\begin{aligned}
\nabla f_t &= \frac{\partial \hat{Y}_t}{\partial H_t} \frac{\partial H_t}{\partial C_t} \frac{\partial C_t}{\partial f_t} \nabla \hat{Y}_t && (3)\\
&= \hat{Y}_t(1 - \hat{Y}_t) W^y o_t [1 - \tanh(C_t)^2] C_{t-1} \nabla \hat{Y}_t && (4)\\
\nabla i_t &= \frac{\partial \hat{Y}_t}{\partial H_t} \frac{\partial H_t}{\partial C_t} \frac{\partial C_t}{\partial i_t} \nabla \hat{Y}_t && (5)\\
&= \hat{Y}_t(1 - \hat{Y}_t) W^y o_t [1 - \tanh(C_t)^2] \tilde{C}_t \nabla \hat{Y}_t && (6)\\
\nabla o_t &= \frac{\partial \hat{Y}_t}{\partial H_t} \frac{\partial H_t}{\partial o_t} \nabla \hat{Y}_t && (7)\\
&= \hat{Y}_t(1 - \hat{Y}_t) W^y \tanh(C_t) \nabla \hat{Y}_t && (8)
\end{aligned}$$

2

For a specific time t, gradient of weight W^{hi} is:

$$\begin{aligned}
\nabla W_t^{hi} &= \frac{\partial i_t}{\partial W^{hi}} \nabla i_t \\
&= i_t(1 - i_t) H_{t-1} \nabla i_t
\end{aligned} \tag{9}$$

while ∇i_t can be computed from (6). The total gradient of W^{hi} is aggregated as:

$$\nabla W^{hi} = \sum_{t=1}^{T} W_t^{hi} \tag{10}$$

3

If $o_t \approx 0$, $i_t \approx 0$ and $f_t \approx 1$, then $C_t \approx C_{t-1}$. So gradient $\nabla C_t \approx 0$, then $\nabla H_t \approx 0$, the gradient does not explode.