Exercise: Updating a parameter matrix

Yoshinari Fujinuma

April 3, 2017

1 List of Mathematical Notations

- x^t: input vector (e.g., word vector) at time t
- $y^t = (y_1^t, ..., y_n^t)$: gold data for time t. Assume y_k^t are probabilities.
- $p^t = (p_1^t, ..., p_n^t)$: prediction for time t. Assume p_k^t are probabilities.
- $s^t = (s_1^t, ..., s_n^t)$: state (vector) at time t.
- $h^t = (h_1^t, ..., h_n^t)$: output (vector) at time t. $h^t = o^t \odot \tanh(s^t)$
- $i^t = (i_1^t, ..., i_n^t)$: input gate (vector) at time t.
- $f^t = (f_1^t, ..., f_n^t)$: forget gate (vector) at time t.
- $o^t = (o_1^t, ..., o_n^t)$: output gate (vector) at time t.
- V_a: weight matrix for the input vector x
- U_a: weight matrix for the hidden state vector h
- a^t: candidate input vector a at time t (which will be multiplied by the input gate).
- L: error/cost function. Assume cross-entropy error in this exercise.

From Siddharth's 4th slide, we define a^t as

$$a^t = \tanh\big(V_a x^j + U_a h^{j-1} + b_a\big) \tag{1}$$

Note that the parameter matrix V_a is only used for computing a^t.

2 Update the parameter matrix V_a

Let's update one of the parameter matrices V_a using stochastic gradient descent i.e.

$$V_a = V_a + \eta \frac{\partial L}{\partial V_a}$$

where η is the learning rate (scalar) and $\frac{\partial L}{\partial V_a}$ is the derivative of L w.r.t V_a.

If we want to compute $\frac{\partial L}{\partial V_a}$, it is a scalar by matrix derivative. To make it easier to understand, let's lower the dimension and focus on the first element of a^t i.e. a_1^t.

Here are some points to consider for Equation 1:

- tanh and + are element-wise operations.
- $V_a x^j$ is a matrix-to-vector multiplication.

From the above two points, the first element of a^t i.e. a_1^t is

$$a_1^t = \tanh\big(V_{a1} x^j + U_{a1} h^{j-1} + b_1\big) \tag{2}$$

The derivative of L w.r.t the first row of V_a i.e. $\frac{\partial L}{\partial V_{a1}}$ is now scalar-to-vector derivative (which is a vector).

2.1 Preparation for Applying a Chain Rule

Assume the error function L at time t as a cross-entropy loss i.e.

$$L^t = -\sum_k y_k^t \log p_k^t$$

where p^t is the outcome of applying softmax function to a vector h^t i.e.

$$p_i^t = \frac{\exp\big(h_i^t\big)}{\sum_k \exp\big(h_k^t\big)},$$

$$h_i^t = o_i^t \tanh\big(s_i^t\big),$$

and

$$s_i^t = i_i^t a_i^t + f_i^t s_i^{t-1}.$$

Let's update the first row of a parameter matrix V_a i.e. V_{a1} using stochastic gradient descent:

$$V_{a1} = V_{a1} + \eta \frac{\partial L}{\partial V_{a1}}$$

where η is the learning rate (scalar) and $\frac{\partial L}{\partial V_{a1}}$ is the derivative of L w.r.t V_{a1}.

Now, our goal is to compute $\frac{\partial L}{\partial V_{a1}}$

2.2 Applying the Chain Rule using the Computational Graph

First, let's plot the computation graph. Take a look at Figure 1.

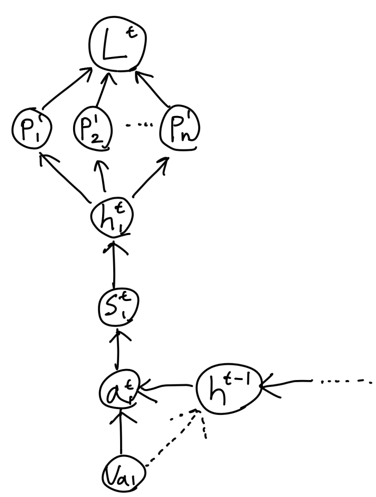

Figure 1: Simplified computational graph w.r.t V^{a1}

$$\frac{\partial L^t}{\partial V_{a1}} = \frac{\partial E^t}{\partial h_1^t} \frac{\partial h_1^t}{\partial s_1^t} \frac{\partial s_1^t}{\partial a_1^t} \left(\frac{\partial a_1^t}{\partial V_{a1}} + \frac{\partial a_1^t}{\partial h_1^{t-1}} \frac{\partial h_1^{t-1}}{\partial V_{a1}} \right) \tag{3}$$

3 Exercises

3.1 Exercise 1: Hand-computing the Partial Derivatives

Complete calculating the following derivatives by filling in the blank boxes.

Let's compute $\frac{\partial L^1}{\partial V_{a1}}$ and regard h^0 as the initial hidden state i.e. constant vector. Our goal is to compute the four partial derivatives:

$$\frac{\partial L^1}{\partial V_{a1}} = \frac{\partial L^1}{\partial h_1^1}\frac{\partial h_1^1}{\partial s_1^1}\frac{\partial s_1^1}{\partial a_1^1}\frac{\partial a_1^1}{\partial V_{a1}} \tag{4}$$

$$\tag{5}$$

Let's start from the easiest ones. Since $\frac{\partial s_1^1}{\partial a_1^1}$ and $\frac{\partial h_1^1}{\partial s_1^1}$ are scalar by scalar derivatives,

$$\frac{\partial s_1^1}{\partial a_1^1} = \frac{\partial}{\partial a_1^1}(i_1^1 a_1^1 + f_1^1 s_1^0) = \boxed{}$$

$$\frac{\partial h_1^1}{\partial s_1^1} = \frac{\partial}{\partial s_1^1}(o_1^1 \tanh(s_1^1)) = o_1^1(\boxed{})$$

which **matches to second to the last line in Siddharth's 10th slide**.

- **Hint**: $\frac{\partial}{\partial x}(\tanh(x)) = 1 - \tanh^2(x)$

Next, we consider the scalar by vector derivative $\frac{\partial a_1^1}{\partial V_{a1}}$. Recall that $\frac{\partial a_1^1}{\partial V_{a1}}$ is defined as

$$\frac{\partial a_1^1}{\partial V_{a1}} = (\frac{\partial a_1^1}{\partial V_{a11}}, \frac{\partial a_1^1}{\partial V_{a12}}, ..., \frac{\partial a_1^1}{\partial V_{a1n}}) \tag{6}$$

Let's consider the first element $\frac{\partial a_1^1}{\partial V_{a11}}$. Let q_1 be

$$q_1 = V_{a1}x^1 + U_{a1}h^0 + b_1. \tag{7}$$

NOTE: h^{t-1} is also dependent on V_a, so if we are computing derivatives for time $t \geq 2$, $\frac{\partial q_1}{\partial V_{a11}} = x_1^t + \frac{\partial q_1}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial V_{a11}}$. However, we are computing the derivative for $t = 1$, and h^0 is a constant vector, so we don't worry about it in this exercise.

$$\frac{\partial a_1^1}{\partial V_{a11}} = 1 - \tanh^2(q_1)\frac{\partial q_1}{\partial V_{a11}} = 1 - \tanh^2(q_1)\boxed{}$$

HINT: $V_{a1}x^1 = V_{a11}x_1^1 + V_{a12}x_2^1 + ... + V_{a1n}x_n^1 = \sum_{i=1}^n V_{a1i}x_i^1$

So thinking back into a vector form i.e. $\frac{\partial a_1^1}{\partial V_{a1}}$,

$$\frac{\partial a_1^1}{\partial V_{a1}} = (1 - \tanh^2(q))(\boxed{})^T \tag{8}$$

which **matches to the result from Siddharth's 14th slide**.

$\frac{\partial L^1}{\partial h_1^1}$ is a bit trickier since $p_1^1, ..., p_n^1$ all depend on h_1^1 due to the denominator $\sum_k \exp(h_k^1)$. I'll just post the result, but see the appendix if you are interested in how did this result pop up.

$$\frac{\partial L^1}{\partial h_1^1} = p_1^1 - y_1^1 \tag{9}$$

By gathering up all the calculated partial derivatives, we get

$$\frac{\partial L^1}{\partial V_{a1}} = \frac{\partial L^1}{\partial h_1^1}\frac{\partial h_1^1}{\partial s_1^1}\frac{\partial s_1^1}{\partial a_1^1}\frac{\partial a_1^1}{\partial V_{a1}} \tag{10}$$

$$= (p_1^1 - y_1^1)(o_1^1(\boxed{\qquad\qquad})(\boxed{\qquad})(1 - \tanh^2(q_1))(\boxed{\qquad})^T \tag{11}$$

3.2 Exercise 2

Let $p_1^1 = 1, y_1^1 = 0.5, o_1^1 = 1, s_1^0 = 1, i_1^1 = 1, V_{a1} = (1,1), U_{a1} = (1,1), x^1 = \begin{bmatrix}1\\1\end{bmatrix}, h^0 = \begin{bmatrix}1\\1\end{bmatrix}, b_1 = 1$. Compute the values of the following derivatives:

$$\frac{\partial L^1}{\partial h_1^1} =$$
$$\frac{\partial h_1^1}{\partial s_1^1} =$$
$$\frac{\partial s_1^1}{\partial a_1^1} =$$
$$\frac{\partial a_1^1}{\partial V_{a1}} =$$
$$\frac{\partial L^1}{\partial V_{a1}} = \frac{\partial L^1}{\partial h_1^1}\frac{\partial h_1^1}{\partial s_1^1}\frac{\partial s_1^1}{\partial a_1^1}\frac{\partial a_1^1}{\partial V_{a1}} =$$

HINT: Use $\tanh^2(5) \approx 0.9998$ and $\tanh^2(1.9998) \approx 0.9293$

3.3 Exercise 3

Confirm that your hand-computed derivative matches the result of the automatic differentiation by Theano.

4 Appendix: Calculation of $\frac{\partial L^1}{\partial h_1^1}$

$\frac{\partial L^1}{\partial h_1^1}$ is a bit trickier since $p_1^1, ..., p_n^1$ all depend on h_1^1 due to the denominator $\sum_k \exp(h_k^1)$.

$$\begin{aligned}
\frac{\partial L^1}{\partial h_1^1} &= \frac{\partial}{\partial h_1^1}(-\sum_k y_k^1 \log p_k^1) && (12)\\
&= -\sum_k (y_k^1 \frac{\partial \log p_k^1}{\partial h_1^1}) && (13)\\
&= -\sum_k (y_k^1 \frac{\partial \log p_k^1}{\partial p_k^1}\frac{\partial p_k^1}{\partial h_1^1}) && (14)\\
&= -\sum_k (y_k^1 \frac{1}{p_k^1}\frac{\partial p_k^1}{\partial h_1^1}) && (15)
\end{aligned}$$

$$\frac{\partial p_1^1}{\partial h_1^1} = \frac{\partial}{\partial h_1^1}(\frac{\exp(h_1^1)}{\sum_k \exp(h_k^1)}) = \frac{\exp(h_1^1)}{\sum_k \exp(h_k^1)} - \boxed{\phantom{\frac{\exp(h_1^1)}{\sum_k}}} = p_1^1(1 - \boxed{}) \qquad (16)$$

Hint: $\frac{\partial}{\partial h_1^1}(\sum_k exp(h_k^1))^{-1} = -\frac{exp(h_1^1)}{(\sum_k exp(h_k^1))^2} = -p_1^1(\frac{1}{\sum_k exp(h_k^1)})$

For $\alpha \neq 1$,

$$\begin{aligned}
\frac{\partial p_\alpha^1}{\partial h_1^1} &= \frac{\partial}{\partial h_1^1}(\frac{\exp(h_\alpha^1)}{\sum_k \exp(h_k^1)}) && (17)\\
&= \exp(h_\alpha^1)\frac{\partial}{\partial h_1^1}(\frac{1}{\sum_k \exp(h_k^1)}) && (18)\\
&= -\exp(h_\alpha^1)\frac{\exp(h_1^1)}{(\sum_k \exp(h_k^1))^2} && (19)\\
&= -p_\alpha^1 p_1^1 && (20)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial L^1}{\partial h_1^1} &= -(\frac{y_1^1}{p_1^1}p_1^1(1-p_1^1)) + (\sum_{k\neq 1}\frac{y_k^1}{p_k^1}p_k^1 p_1^1) && (21)\\
&= -(y_1^1(1-p_1^1)) + \sum_{k\neq 1}(y_k^1 p_1^1) && (22)\\
&= -y_1^1 + p_1^1 \sum_k (y_k^1) && (23)\\
&= -y_1^1 + p_1^1 \; (\because) \; y_k^1 \text{ is a probability.} && (24)
\end{aligned}$$