ECE/CS/ME 539 – Fall 2024 — Activity Solution 30

1. Backpropagation Through Time

Consider a recurrent linear system with state $h_t \in \mathbb{R}$ and inputs $x_t \in \mathbb{R}$ defined by the equation

$$h_t = ah_{t-1} + bx_t$$

where a is a state transition coefficient and b is an input transformation coefficient. Suppose that we unroll this system for T time steps $x_1, \ldots, x_T$ using a known initial state h_0, and that we want to find the coefficients a and b so that the final state h_T matches a specific target value y by minimizing the mean squared error.

$$L = \frac{1}{2}(h_T - y)^2$$

(a) Derive $\frac{\partial L}{\partial h_T}$.

$$\frac{\partial L}{\partial h_T} = \frac{\partial}{\partial h_T}\left(\frac{1}{2}(h_T - y)^2\right) = (h_T - y)\frac{\partial (h_T - y)}{\partial h_T} = h_T - y$$

(b) Show that $\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_T} a^{T-t}$.

Using the chain rule, we know that

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_t}.$$

Thus, we have the following sequence of gradients:

$$\frac{\partial L}{\partial h_T} = h_T - y$$

$$\frac{\partial L}{\partial h_{T-1}} = \frac{\partial L}{\partial h_T} a$$

$$\frac{\partial L}{\partial h_{T-2}} = \frac{\partial L}{\partial h_{T-1}} a = \frac{\partial L}{\partial h_T} a^2$$

$$\frac{\partial L}{\partial h_{T-3}} = \frac{\partial L}{\partial h_{T-2}} a = \frac{\partial L}{\partial h_T} a^3$$

$$\cdots$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_T} a^{T-t}$$

(c) Derive $\frac{\partial L}{\partial a}$ and $\frac{\partial L}{\partial b}$.

Given:

We have a recurrent linear system defined by:

$$h_t = ah_{t-1} + bx_t$$

for $t = 1, 2, \ldots, T$, with initial state h_0 given.

Our loss function is:

$$L = \frac{1}{2}(h_T - y)^2$$

We are to find expressions for the gradients $\frac{\partial L}{\partial a}$ and $\frac{\partial L}{\partial b}$.

Step 1: Compute $\frac{\partial L}{\partial h_T}$

As derived in part (a):

$$\frac{\partial L}{\partial h_T} = h_T - y$$

Step 2: Compute $\frac{\partial h_t}{\partial a}$ and $\frac{\partial h_t}{\partial b}$

Computing $\frac{\partial h_t}{\partial a}$:

We need to compute $\frac{\partial h_t}{\partial a}$ for $t = 1, 2, \ldots, T$.

Starting with the recurrence relation:

$$h_t = ah_{t-1} + bx_t$$

Taking derivative with respect to a:

$$\frac{\partial h_t}{\partial a} = h_{t-1} + a\frac{\partial h_{t-1}}{\partial a}$$

Let us denote:

$$s_t = \frac{\partial h_t}{\partial a}$$

Then the recursive equation becomes:

$$s_t = h_{t-1} + as_{t-1}$$

with the base case:

$$s_0 = \frac{\partial h_0}{\partial a} = 0 \quad \text{(since } h_0 \text{ is given and does not depend on } a\text{)}$$

Unrolling the recursion:

First few terms:

- $s_1 = h_0 + as_0 = h_0 + 0 = h_0$
- $s_2 = h_1 + as_1 = h_1 + ah_0$
- $s_3 = h_2 + as_2 = h_2 + ah_1 + a^2h_0$

Continuing this pattern, we find:

$$s_t = h_{T-1} + ah_{T-2} + a^2 h_{T-3} + \cdots + a^{T-1} h_0$$

Summation form:

$$s_t = \sum_{t=1}^{T} a^{t-1} h_{T-t} = \sum_{t=1}^{T} a^{T-t} h_{t-1}$$

Computing $\frac{\partial h_t}{\partial b}$:

Similarly, taking derivative with respect to b:

$$\frac{\partial h_t}{\partial b} = x_t + a\frac{\partial h_{t-1}}{\partial b}$$

Let us denote:

$$r_t = \frac{\partial h_t}{\partial b}$$

Then:

$$r_t = x_t + ar_{t-1}$$

with the base case:

$$r_0 = \frac{\partial h_0}{\partial b} = 0 \quad \text{(since } h_0 \text{ does not depend on } b\text{)}$$

Unrolling the recursion:

First few terms:

- $r_1 = x_1 + a \cdot 0 = x_1$
- $r_2 = x_2 + ax_1$
- $r_3 = x_3 + ax_2 + a^2 x_1$

Summation form:

$$r_t = \sum_{t=1}^{T} a^t x_{T-t}$$

Step 3: Compute $\frac{\partial L}{\partial a}$ and $\frac{\partial L}{\partial b}$

Using the chain rule:

$$\frac{\partial L}{\partial a} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial a}$$

Similarly:

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial b}$$

Substituting the expressions:

Computing $\frac{\partial L}{\partial a}$:

$$\frac{\partial L}{\partial a} = \sum_{t=1}^{T} (h_T - y)\, s_t = (h_T - y) \sum_{t=1}^{T} a^{t-1} h_{T-t} = (h_T - y) \sum_{t=1}^{T} a^{T-t} h_{t-1}$$

Computing $\frac{\partial L}{\partial b}$:

Similarly, the derivative of h_T with respect to b is:

$$\frac{\partial h_T}{\partial b} = \sum_{t=1}^{T} a^{T-t} x_t$$

Thus:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial h_T} \frac{\partial h_T}{\partial b} = (h_T - y) \left(\sum_{t=1}^{T} a^{T-t} x_t \right)$$

(d) Suppose that $a < 1$. Discuss what happens to $\frac{\partial L}{\partial h_t}$ if $T \gg t$. What if $a > 1$?

As we proved in part (b),

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_T} a^{T-t}.$$

When $a < 1$, $\frac{\partial L}{\partial h_t}$ approaches zero. This is the *gradient vanishing problem.*

When $a > 1$, $\frac{\partial L}{\partial h_t}$ becomes very large. This is the *gradient explosion problem.*

(e) This problem assumes that both the inputs and states are scalars. In RNNs, we usually have inputs and hidden state vectors, in which case the transition weights A and input transformation weights B are matrices (not scalars). What are the conditions on A or B that would lead to similar issues as those identified in part (d)?

In the vectorized case, a becomes the eigenvalues of A. Gradient vanishing occurs if all eigenvalues are less than one in magnitude ($|\lambda| < 1$). Gradient explosion occurs if any eigenvalue exceeds one in magnitude ($|\lambda| > 1$).