CS446 / ECE 449: Machine Learning, Fall 2020, Homework 5

Name: Saud Alrasheed (Sauda2)

Worked individually

Problem (1):

Problem 1.a:

We have that,

$$L^{(t)} = -y^T \ln(\widehat{y}) = -\sum_{k=1}^{K} y_k \ln(p_k)$$

Where p_k is defined to be,

$$p_k = \frac{\exp^{z_k}}{\sum_j \exp^{z_j}}$$

From Homework 3, we know that,

$$i = k \implies \frac{\partial p_k}{\partial z_i} = p_k(1 - p_k)$$

$$i \neq k \implies \frac{\partial p_k}{\partial z_i} = -p_k p_i$$

Thus,

$$\frac{\partial L^{(t)}}{\partial V_{ij}} = -\sum_{k=1}^{K} y_k \frac{1}{p_k} \frac{\partial p_k}{\partial z_i} \frac{\partial z_i}{\partial V_i j} = -y_i h_j (1 - p_i) + \sum_{k \neq i} y_k p_i h_j$$

$$\frac{\partial L^{(t)}}{\partial V_{ij}} = (-y_i + \sum_k y_k p_i) h_j = h_j (p_i \sum_k y_k - y_i)$$

But, since $\sum_k y_k = 1$, we obtain the following,

$$\frac{\partial L^{(t)}}{\partial V_{ij}} = h_j(p_i - y_i)$$

Now, we use the same approach to find $\frac{\partial L^{(t)}}{\partial c_i}$. Therefore,

$$\frac{\partial L^{(t)}}{\partial c_i} = (p_i - y_i)$$

Thus, we are done!

Problem 1.b:

By the Multivariate chain rule we have that,

$$\frac{\partial L^{(t)}}{\partial U} = \sum_{k=1}^{t} \frac{\partial L^{(t)}}{\partial h^{(k)}} \frac{\partial h^{(k)}}{\partial U}$$

We will first try to understand $\frac{\partial L^{(t)}}{\partial h^{(k)}}$. We have that for $k = t$, we obtain the following expression,

$$\frac{\partial L^{(t)}}{\partial h^{(k)}} = \frac{\partial L^{(t)}}{\partial z^{(k)}} \frac{\partial z^{(k)}}{\partial h^{(k)}}$$

and for $k < t$, chain rule yields,

$$\frac{\partial L^{(t)}}{\partial h^{(k)}} = \frac{\partial L^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial h^{(t-2)}} \quad \dots \quad \frac{\partial h^{(k+2)}}{\partial h^{(k+1)}} \frac{\partial h^{(k+1)}}{\partial h^{(k)}}$$

Thus, one can write this expression as,

$$\frac{\partial L^{(t)}}{\partial h^{(k)}} = \frac{\partial L^{(t)}}{\partial h^{(t)}} \prod_{k<i\leq t} \frac{\partial h^{(i)}}{\partial h^{(i-1)}}$$

Now, we will try to find $\frac{\partial h^{(i)}}{\partial h^{(i-1)}}$, for all i. Clearly, one can use the chain rule to obtain the following,

$$\frac{\partial h^{(i)}}{\partial h^{(i-1)}} = diag(\sigma'(Wh^{(t-1)} + Ux^{(t)} + b))W$$

Since $h^{(t)} = \sigma(a)$, we have that $\frac{\partial \sigma}{\partial a} = 1 - (h^{(t)})^2$. Thus,

$$\frac{\partial h^{(i)}}{\partial h^{(i-1)}} = diag(1 - (h^{(i)})^2)W$$

$$\implies \frac{\partial L^{(t)}}{\partial h^{(k)}} = \frac{\partial L^{(t)}}{\partial h^{(t)}} \prod_{k<i\leq t} \frac{\partial h^{(i)}}{\partial h^{(i-1)}} = \frac{\partial L^{(t)}}{\partial h^{(t)}} W^{(t-k)} \prod_{k<i\leq t} diag(1 - (h^{(i)})^2)$$

Now, putting all peices together, we obtain the following,

$$\frac{\partial L^{(t)}}{\partial U} = \sum_{k=1}^{t} \frac{\partial L^{(t)}}{\partial h^{(t)}} W^{(t-k)} (\prod_{k<i\leq t} diag(1 - (h^{(i)})^2)) \frac{\partial h^{(k)}}{\partial U}$$

$$\implies \frac{\partial L^{(t)}}{\partial U} = \sum_{k=1}^{t} \frac{\partial L^{(t)}}{\partial h^{(t)}} W^{(t-k)} (\prod_{k<i\leq t} diag(1 - (h^{(i)})^2)) \; (1 - (h^{(k)})^2) x^{(k)}$$

$$\implies \frac{\partial L^{(t)}}{\partial U} = \sum_{k=1}^{t} (\widehat{y}^{(t)} - y^{(t)})V \; W^{(t-k)}(\prod_{k<i\leq t} diag(1-(h^{(i)})^2)) \; (1-(h^{(k)})^2)x^{(k)}$$

Now, we use the same approach to find $\frac{\partial L^{(t)}}{\partial b}$. Therefore,

$$\frac{\partial L^{(t)}}{\partial b} = \sum_{k=1}^{t} (\widehat{y}^{(t)} - y^{(t)})V \; W^{(t-k)}(\prod_{k<i\leq t} diag(1-(h^{(i)})^2)) \; (1-(h^{(k)})^2)$$

Problem 1.c:
Here, we will use the same approach as problem 1.b. Thus, we obtain,

$$\frac{\partial L^{(t)}}{\partial W} = \sum_{k=1}^{t} \frac{\partial L^{(t)}}{\partial h^{(k)}} \frac{\partial h^{(k)}}{\partial W} = \sum_{k=1}^{t} \frac{\partial L^{(t)}}{\partial h^{(t)}} \prod_{k<i\leq t} \frac{\partial h^{(i)}}{\partial h^{(i-1)}} \frac{\partial h^{(k)}}{\partial W}$$

$$\implies \frac{\partial L^{(t)}}{\partial W} = \sum_{k=1}^{t} \frac{\partial L^{(t)}}{\partial h^{(t)}} W^{(t-k)}(\prod_{k<i\leq t} diag(1-(h^{(i)})^2)) \; (1-(h^{(k)})^2)h^{(k-1)}$$

$$\implies \frac{\partial L^{(t)}}{\partial W} = \sum_{k=1}^{t} (\widehat{y}^{(t)} - y^{(t)})V \;\; W^{(t-k)}(\prod_{k<i\leq t} diag(1-(h^{(i)})^2)) \; (1-(h^{(k)})^2)h^{(k-1)}$$