CS7643: Deep Learning
Fall 2020
HW1 Solutions

Ruize Yang

September 9, 2020

1.

For i,j, $\frac{\partial s_i}{\partial z_j} = \frac{1}{\sum_k e^{z_k}} \frac{\partial e^{z_i}}{\partial z_j} - e^{z_i} \frac{1}{(\sum_k e^{z_k})^2} \frac{\partial \sum_k e^{z_k}}{\partial z_j}$

If $i = j$: $\frac{\partial s_i}{\partial z_j} = \frac{e^i \sum_{k \neq i} e^{z_k}}{(\sum_k e^{z_k})^2} = s_i(1 - s_i)$

If $i \neq j$: $\frac{\partial s_i}{\partial z_j} = -\frac{e^{z_i} e^{z_j}}{(\sum_k e^{z_k})^2} = -s_i s_j$

$(\frac{\partial s}{\partial z})_{ij} = \begin{cases} s_i(1 - s_i) & i = j \\ -s_i s_j & i \neq j \end{cases}$

2.
g has a local minimum at $w^t = (w_i^t, ..., w_n^t) \Rightarrow$ for small δ where $|\delta| < \gamma$, $g((w_1^t, ..., w_i^t + \delta, ..., w_n^t)) - g(w^t) \geq 0$.

$$\frac{\partial g(w^t)}{\partial w_i} = \lim_{\delta \to 0^+} \frac{g((w_1^t, ..., w_i^t + \delta, ..., w_n^t)) - g(w^t)}{\delta} \geq 0$$

$$\frac{\partial g(w^t)}{\partial w_i} = \lim_{\delta \to 0^-} \frac{g((w_1^t, ..., w_i^t + \delta, ..., w_n^t)) - g(w^t)}{\delta} \leq 0$$

$$\Rightarrow \frac{\partial g(w^t)}{\partial w_i} = 0 \text{ for } i = 1, ..., n, \text{ gradient at } w^t = 0.$$

When g is viewed as single-variable function of one of the entries in w, at saddle point gradient $= 0$, some $g(w_i)$ could have local minimum, some $g(w_j)$ have local maximum, but it is not the minimum for all entries of w, and its not a local minimum.

3.
As g is differentiable and convex, then for any $x, y \in R^n$, $g(y) \geq g(x) + \nabla g(x)^T(y - x)$. Then for $\nabla g(w^*) = 0$, for any $x \in R^n$, $g(x) \geq g(w^*) + \nabla g(w^*)^T(x - w^*) = g(w^*)$, $\Rightarrow g(w^*)$ is global minimum.

4.
$f(w) = \frac{1}{2}(w-2)^2 + \frac{1}{2}(w+1)^2 = w^2 - w + \frac{5}{2}$
$f_1' = w - 2, f_2' = w + 1$, $f_1' < 0$ and $f_2' < 0 : w < -1$; $f_1' > 0$ and $f_2' > 0 : w > 2$
Moving to + direction (gradient is negative): $f(w+\eta) = (w^2 - w + \frac{5}{2}) + \eta(\eta + 2w - 1)$
$f(w+\eta) < f(w) \Rightarrow w < \frac{1}{2}(1-\eta) \Rightarrow w < \frac{1}{2}$
Moving to - direction (gradient is positive): $f(w+\eta) = (w^2 - w + \frac{5}{2}) + \eta(\eta - 2w + 1)$
$f(w-\eta) < f(w) \Rightarrow w > \frac{1}{2}(1+\eta) \Rightarrow w > \frac{1}{2}$
To make the objective function decrease: gradient is negative and $w < \frac{1}{2}$, or gradient is positive and $w > \frac{1}{2}$.
When $w > 2$ or $w < -1$, loss function decrease at every iteration. When w is between $(-1, 2)$, loss function may not decrease at every iteration.

5.

6.
If G is a DAG, then there is one node with no incoming edges, label it as v_1. Remove v_1 and all its outgoing edges, the resulting graph G_1 is also a DAG. There is one node in G_1 with no incoming edges, label it as v_2, remove v_2 and all its outgoing edges. Repeat until every node in G is labeled and removed, and get an ordering $\{v_1, ..., v_n\}$ of the nodes on G. As each time only remove the outgoing edges of one node, so for any edge e_{ij} from v_i to v_j, v_i is removed before v_j, so $i < j$ for any i, j. This ordering is a topological ordering.

7.
Assume G is not a DAG and there is one cycle. Let the edges of the cycle be $(v_{i1}, v_{i2}), (v_{i2}, v_{i3}), ..., (v_{ik}, v_{i1})$. In topological ordering, there is $v_{i1} < v_{i2} < v_{i3} < ... < v_{ik} < v_{i1}$, which is impossible. So there is no cycle in G, and G is a DAG.

8.
The authors optimize a model structure that could work well across a wide range of weights, instead of optimize the parameters given a model structure. They show that the model structure could be naturally capable of performing a given task. But for later steps in the algorithm where the evolving of networks is mostly refining, it seems not really different from setting some weights to zero in a fully connected network during training process, especially when they also use individually tuned weights in the paper.

9.
The paper show that some simple network at early steps can do well, one thing that can be further compared is the importance of training vs evolving in structure, e.g. is the champion network generated by algorithm still the best after conventional training?