HW 1

Quentin Mot

2023-02-06

1. Collaborators:

I did not collaborate with anyone while preparing this assignment

2. Optimization

2.1

Recall that two vectors u, v are orthogonal if their dot product is 0.

Note that the tangent of the arbitrary curve $\boldsymbol{r}(t)$ is $[\frac{\partial x_1(t)}{\partial t}; ...; \frac{\partial x_d(t)}{\partial t}]$.

Then, the dot product of ∇f_0 and $\boldsymbol{r}(t)$ evaluated at t_0 is:

(1)
$$[\frac{\partial x_1(t)}{\partial t}\frac{\partial f}{\partial x_1}; ...; \frac{\partial x_d(t)}{\partial t}\frac{\partial f}{\partial x_d}]$$

But, since $x_i(t)$ evaluated at t_0 equals x_i for all $i \in 1, .., d$, by the chain rule (1) equals (2):

(2)
$$[\frac{\partial f}{\partial t}; ...; \frac{\partial f}{\partial t}]$$

But, t is just a scalar parameter for the curve $\boldsymbol{r}(t)$ that is not present in $f(x)$

Hence (2) equals (3):

(3)
$$[0; ...; 0]$$

Thus ∇f_0 and $\boldsymbol{r}(t)$ are orthogonal.

QED

In non technical terms, we just showed that taking small steps towards points in the level set i.e. towards more points in $\boldsymbol{r}$ by changing t does not change $f(x)$, that is it is orthogonal to the direction that would most change $f(x)$. This is important in the context of deep learning because otherwise moving around the level set would increase the value of f, which contradicts the definition of a level set and could lead to exploding or vanishing gradient problems.

2.2

Suppose that $g \in \mathbb{R}^n \to \mathbb{R}$ has a local minimum at some $\boldsymbol{w}^t$.

Then, by definition there exists some $\gamma > 0$ such that for all $\boldsymbol{w} \in \mathbb{R}^n$, the L2 norm of the difference $\boldsymbol{w}^t - \boldsymbol{w}$ is less than γ implies $g(\boldsymbol{w}^t) \leq g(\boldsymbol{w})$

Now suppose that the gradient of g at $\boldsymbol{w}^t \neq 0$. Then, moving away from $\boldsymbol{w}^t$ in the direction of the negative gradient to $\boldsymbol{w}^l$ allows for $g(\boldsymbol{w}^l) \leq g(\boldsymbol{w}^t)$.

But, that is a contradiction with the definition of a local minimum, i.e. we have just proved that a local minimum with nonzero gradient is not a local minimum.

Hence, the gradient of g at a local minimum $\boldsymbol{w}^t$ equals 0

If the gradient of$g$ at $\boldsymbol{w}^t = 0$, it is not necessarily true that $\boldsymbol{w}^t$ is a local minimum. Indeed, $\boldsymbol{w}^t = 0$ implies that $\boldsymbol{w}^t$ is either a local minimum or a local maximum. Via a simple counterexample, to the claim that if the gradient of g at $\boldsymbol{w}^t = 0$ then $\boldsymbol{w}^t$ is a local minimum, if $g(x) = -x^2$, then the gradient being equal to zero at $\boldsymbol{w}^t$ implies that $\boldsymbol{w}^t$ is a local (and, in this case, global) maximum.

Hence the gradient at g of $\boldsymbol{w}^t = 0$ does not necessarily imply $\boldsymbol{w}^t$ is a local minimum.

QED

2.3

Suppose that there exists a point $\boldsymbol{w}^m$ such that $g(\boldsymbol{w}^m) < g(\boldsymbol{w}^*)$

Then, from the convexity of g all points $w_t = t\boldsymbol{w}^* + (1-t)\boldsymbol{w}^m, t \in (0,1)$ are in the domain of g.

Then, from the above inequality:

$$g(w_t) \leq tg(\boldsymbol{x}^*) + (1-t)g(\boldsymbol{w}^m) < tg(\boldsymbol{w}^*) + (1-t)g(\boldsymbol{w}^*) = g(\boldsymbol{w}^*)$$

That is, $g(w_t) < g(\boldsymbol{w}^*)$

Recall the definition of a local minimum, specifically that there exists some $\gamma > 0$ such that for all $\boldsymbol{w} \in \mathbb{R}^n$, the L2 norm of the difference $\boldsymbol{w}^* - \boldsymbol{w}$ is less than γ implies $g(\boldsymbol{w}^*) \leq g(\boldsymbol{w})$.

Let t be sufficiently close to 1 such that the definition of a local minimum applies to the w_t in the neighborhood of $g(\boldsymbol{w}^*)$.

Then we have $g(w_t) \geq g(\boldsymbol{w}^*)$ and $g(w_t) < g(\boldsymbol{x}^*)$, a contradiction.

Hence $\boldsymbol{w}^m$ cannot exist and $\boldsymbol{w}^*$ is a global minimum

QED

2.4

So that we can use the product rule, break each term that will be differentiated in the resulting Jacobian into the form $\frac{e^{z_i}}{1}\frac{1}{\sum_k e^{z_k}}$.

When the partial derivative is respect to z_i, $\frac{\partial e^{z_i}}{\partial z_i} = e^{z_i}$. Otherwise, that is for $j \neq i$, $\frac{\partial e^{z_i}}{\partial z_j} = 0$

The partial derivative $\frac{\partial \frac{1}{\sum_k e^{z_k}}}{\partial z_i} = -(\frac{\partial \sum_k e^{z_k}}{\partial z_i})^{-2} = -(e^{z_i})^{-2}$ equals $-e^{-2z_i}$.

Then, the entries of the resulting Jacobian with rows and columns i, j that refer to s_i and z_j, respectively are as follows:

If $i = j$, then the entry equals $\frac{e^{z_i}}{\sum_k e^{z_k}} + \frac{e^{z_i}}{-e^{2z_i}} = \frac{e^{z_i}}{\sum_k e^{z_k}} - \frac{1}{e^{z_i}} = \frac{e^{2z_i}}{e^{z_i}\sum_k e^{z_k}} - \frac{\sum_k e^{z_k}}{e^{z_i}\sum_k e^{z_k}} = \frac{e^{2z_i} - \sum_k e^{z_k}}{e^{z_i}\sum_k e^{z_k}}$

Else (that is, $i \neq j$), the entry equals $-\frac{e^{z_i}}{e^{2z_j}}$

QED

2.5

Note that the definition of the $d-1$ simplex is simply a set of vectors in R^d whose entries are all nonnegative and whose sum equals 1.

Trivially, the entries s_i of the softmax function satisfy those conditions.

Now, let $s(\boldsymbol{x}) = \boldsymbol{y}$. Again, $\boldsymbol{y}$ trivially satisfies the necessary conditions.

Note that $-\boldsymbol{x}^T\boldsymbol{y} = -\sum_{k_1} \frac{z_{k_1} e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}}$ and that $H(\boldsymbol{y}) = -\sum_{k_1} \frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}} \log(\frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}})$

Then $A =: -\boldsymbol{x}^T\boldsymbol{y} - H(\boldsymbol{y}) = -\sum_{k_1} \frac{z_{k_1} e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}} - (-\sum_{k_1} \frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}} \log(\frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}})) = -\sum_{k_1} \frac{z_{k_1} e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}} + \sum_{k_1} \frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}} \log(\frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}})$

Factor out $(\sum_{k_2} e^{z_{k_2}})^{-1}$:

$$A = (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} e^{z_{k_1}} - \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}}))$$

Redistribut and realize that by definition the former term equals 1:

$$A = (\sum_{k_2} e^{z_{k_2}})^{-1} \sum_{k_1} e^{z_{k_1}} - (\sum_{k_2} e^{z_{k_2}})^{-1} \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}})$$

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1} \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\frac{e^{z_{k_1}}}{\sum_{k_2} e^{z_{k_2}}})$$

Note that from the logarithm quotient rule we can rewrite A as:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1} \sum_{k_1} z_{k_1} e^{z_{k_1}} (\log(e^{z_{k_1}}) - \log(\sum_{k_2} e^{z_{k_2}}))$$

Distribute the sum:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} z_{k_1} e^{z_{k_1}} \log(e^{z_{k_1}}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\sum_{k_2} e^{z_{k_2}}))$$

Use the logarithm power rule on the first term:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} \log(e^{z_{k_1} + z_{k_1} e^{z_{k_1}}}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\sum_{k_2} e^{z_{k_2}}))$$

Simplify, the first exponenent:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} \log(e^{z_{k_1}(1+e^{z_{k_1}})}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\sum_{k_2} e^{z_{k_2}}))$$

Apply the logarithm product rule to the first term:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\log(e^{\sum_{k_1} z_{k_1}(1+e^{z_{k_1}})}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} \log(\sum_{k_2} e^{z_{k_2}}))$$

Assume that all logarithms are the natural logarithm:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} z_{k_1}(1 + e^{z_{k_1}}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} \ln(\sum_{k_2} e^{z_{k_2}}))$$

Let $e^a = \sum_{k_2} e^{z_{k_2}}$, which is valid since $\sum_{k_2} e^{z_{k_2}}$ is a nonnegative real number. Then:

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} z_{k_1}(1 + e^{z_{k_1}}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} \ln(e^a))$$

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} z_{k_1}(1 + e^{z_{k_1}}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} a)$$

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} z_{k_1}(1 + e^{z_{k_1}}) - \sum_{k_1} z_{k_1} e^{z_{k_1}} a)$$

$$A = 1 - (\sum_{k_2} e^{z_{k_2}})^{-1}(\sum_{k_1} z_{k_1} + (1 - a)\sum_{k_1} z_{k_1} e^{z_{k_1}})$$

$$A = 1 - \frac{\sum_{k_1} z_{k_1}}{\sum_{k_2} e^{z_{k_2}}} + \frac{(1 - a)}{\sum_{k_2} e^{z_{k_2}}} \sum_{k_1} z_{k_1} e^{z_{k_1}}$$

Take the gradient of the above, that is ∇A, which will be a vector indexed by z_i. Then each entry $i \in 1, .., d$ equals:

$$\frac{\partial A}{\partial z_i} = \frac{\partial}{\partial z_i} \frac{\sum_{k_1} z_{k_1}}{\sum_{k_2} e^{z_{k_2}}} + \frac{\partial}{\partial z_i} \frac{(1 - a)}{\sum_{k_2} e^{z_{k_2}}} \sum_{k_1} z_{k_1} e^{z_{k_1}}$$

Similar to 2.4, the above simplifies to only the terms in the summations that contain z_i with the partial derivatives of all other terms equal to zero. Furthermore, also similar to 2.4 the first term simplifies via the product rule to:

$$\frac{\partial A}{\partial z_i} = \frac{1}{\sum_{k_2} e^{z_{k_2}}} - \frac{\sum_{k_1} z_{k_1}}{e^{2z_i}} + \frac{\partial}{\partial z_i} \frac{(1 - a)}{\sum_{k_2} e^{z_{k_2}}} \sum_{k_1} z_{k_1} e^{z_{k_1}}$$

Now, sequentially and pairwise using the product rule on the second term:

$$\frac{\partial A}{\partial z_i} = \frac{1}{\sum_{k_2} e^{z_{k_2}}} - \frac{\sum_{k_1} z_{k_1}}{e^{2z_i}} + (-\frac{1}{\sum_{k_2} e^{z_{k_2}}} \frac{\partial}{\partial z_i}(a) - \frac{(1 - a)}{e^{2z_i}}) \sum_{k_1} z_{k_1} e^{z_{k_1}} + \frac{(1 - a)}{\sum_{k_2} e^{z_{k_2}}}(z_i e^{z_i} + e^{z_i})$$

Simplifying and plugging a back in:

$$\frac{\partial A}{\partial z_i} = \frac{1}{\sum_{k_2} e^{z_{k_2}}} - \frac{\sum_{k_1} z_{k_1}}{e^{2z_i}} - \frac{1}{\sum_{k_2} e^{z_{k_2}}} \frac{\partial}{\partial z_i}(\ln(\sum_{k_2} e^{z_{k_2}})) - \frac{(1 - \ln(\sum_{k_2} e^{z_{k_2}}))}{e^{2z_i}} \sum_{k_1} z_{k_1} e^{z_{k_1}} + \frac{(1 - \ln(\sum_{k_2} e^{z_{k_2}}))}{\sum_{k_2} e^{z_{k_2}}}(z_i e^{z_i} + e^{z_i})$$

Factoring out $\frac{(1 - \ln(\sum_{k_2} e^{z_{k_2}}))}{\sum_{k_2} e^{z_{k_2}}}$:

$$\frac{\partial A}{\partial z_i} = \frac{1}{\sum_{k_2} e^{z_{k_2}}} - \frac{\sum_{k_1} z_{k_1}}{e^{2z_i}} + (-\frac{1}{\sum_{k_2} e^{z_{k_2}}} \frac{\partial}{\partial z_i}(\ln(\sum_{k_2} e^{z_{k_2}})) + \frac{(1 - \ln(\sum_{k_2} e^{z_{k_2}}))}{e^{2z_i}})((z_i e^{z_i} + e^{z_i}) - \sum_{k_1} z_{k_1} e^{z_{k_1}})$$

Well, I think I did something wrong! Hopefully this is worth some partial credit lol. Hope you're doing well! :)

This formal interpretation tells us that the softmax layer in a neural network maps the logits of the previous layers to values that sum to 1 while preserving their relative magnitude. Those values can be easily interpreted as probabilities for whatever classification or similar task the neural network is performing.

3.6

Prove by induction using the following lemma stated in the homework:

Lemma: If G is a DAG, then at least one node in G has no incoming edges.

Proof:

It is given that G is a DAG. Now, suppose the lemma is false and that each node has at least one incoming edge. Pick an arbitrary start node and switch the direction of every edge. Go from the start vertex to the next vertex, which must exist by the assumption that each node has at least one incoming (now outgoing) edge. Repeat until you find a cycle, which must exist from the pigeon hole principle since every node has at least one an outgoing edge to another vertex, that is there must be n edges where n is the number of vertices in G, but it is only possible to visit $n-1$ edges before a cycle must occur since the edge of the arbitrary start node is not visited when we visit the node. The, G has and does not have a cycle, which is a contradiction. Hence, the lemma is true.

Now, Let $n = 1$. Then G has a topological ordering, namely, G.

Let $n = k$. Assume that G has a topological ordering for all $n \in 1, .., k$

Let $n = k + 1$ Recall from the lemma that G must have at least one node with no incoming edges. Find that node. Then, deleting v from G yields another acyclic graph G' since deleting an node with no incoming edges cannot create a cycle.

From the inductive hypothesis, the new graph G' has n node and consequently a topological ordering.

To create a topological ordering for G, start with v and append the topological ordering for G'. Again, this is valid and does not create a cycle since v has no incoming language.

Hence, G with $n = k + 1$ has a topological ordering.

By the principle of strong mathematical induction, all DAGs with $n \geq 1$ vertices have topological orderings.

QED

3.7

Suppose that G has a topological ordering and that G is not a directed graph, that is that G contains a directed cycle.

Let v_i be the lowest indexed member of a cycle C in G and let v_j be the vertex immediately before v_i in the cycle. By construction, $i < j$.

This construction is consistent with the topological ordering of the graph, by which all edges in the graph are between v_k and v_l with $k < l$.

But, for the the cycle to exist there must be an edge from v_j to v_i, that is, an edge from j > i. That contradicts the aforementioned property of a topological ordering.

Hence, G cannot contain a directed cycle if G has a topological ordering. Thus, G is a directed acyclic graph

QED.

4.8

The key contribution of the paper is to show that network architectures themselves can encode solutions, with a related strength being that produced architectures that do encode solutions can often do better than their counterparts which do not use neural architecture search either before or after undergoing training themselves.

A strength of the paper is the visualizations such as interactive demos which allow for a better conceptual understanding of the material; yet another strength is the elucidation of Weight Agnostic Neural Network Search. Yet another strength notable to the author is the comparison of WANNs with different fixed topology neural networks. The use and focus on minimal architectures is also interesting and useful, especially as it pertaints to potential future interpretability research. The final strength I will mention is the interpretation of the small WANNs.

One of the weaknesses of the paper is that it does not sufficiently explain some of the baselines the WANNs were compared to and terms used that the reader may not be familiar with in the text. Another, perhaps more controversial, "weakness" of the paper is that it focuses on modeling the innate behavior exhibited by animals in certain domains, which is distinct from many AI tasks and goals.

4.9

My personal takeaways from the paper include that it was very interesting to learn about a the approaches to the neural architecture search as opposed to either hand tuned or other neural networks with fixed topology that I was familiar with before this class and have become more familiar with during the class so far. I think modeling animal behavior and genetic algorithms on evolution has been a particularly fruitful area of advancement in the long term, as discussed in lecture, but for reasons I still do not fully comprehend anad that I think the paper did not sufficiently express research has trended away from drawing inspiration from those examples.

I'm also very interested in the minimal architectures discovered and used to achieve near to or better than state of the art performance, since they were easily interpretable. I wonder if such methods could be used to find interpretable minimal architectures for NNs in other domains.