Adrian Kukla CS 7643 Deep Learning HW1 Solutions

Qs 1

Since curve $r(t)$ is parameterized by t this means each component $x_i(t)$ changes as t changes $\forall i$

Since $r(t)$ lies on the level surface $Lf(x_0)$ for all $t \Rightarrow f(r(t)) = f(x_0)$

$\Rightarrow$ as t changes the value of f remains constant along the curve.

Differentiate both sides: to find how "f" changes along the curve differentiate both sides w.r.t. t

$$\frac{d}{dt} f(r(t)) = \frac{d}{dt} f(\underbrace{x_0}_{\text{constant} \Rightarrow \text{derivative is } 0}) = 0$$

Use chain rule

$$\frac{d}{dt} f(r(t)) = 0$$

$$\frac{d}{dt} f(r(t)) = \underbrace{\nabla f(r(t))}_{\text{gradient of } f \text{ at } r(t)} \cdot \underbrace{\frac{\partial r}{\partial t}}_{\text{tangent to the curve at } t}$$

At $t = t_0$, $r(t_0) = x_0$

$$\Rightarrow \nabla f(x_0) \cdot \left.\frac{dr}{dt}\right|_{t=t_0} = 0$$

$\Rightarrow$

$$\therefore \nabla f_0 \perp \left.\frac{\partial r}{\partial t}\right|_{t=t_0}$$

If the dot product of two vectors is zero, the vectors are orthogonal

Example - hill walking at a contour path of a mountain where the altitude doesn't change (i.e. a level surface). The ~~gradient~~ direction that will make one go uphill the steepest (i.e. the gradient) is always at a right angle to this path.

Gradients are key in deep learning because they explain how to change model parameters to minimize the loss function. The gradient points in the direction of the steepest change and orthogonality implies that as one moves along the level surface, the loss won't change. Understanding orthogonality can help design efficient optimization algorithms

Adrian Kukla DL HW1 Solutions

Q2 Prove local minimum implies zero gradient

g is differentiable at w_+ $\Rightarrow$ gradient $\nabla g(w_+)$ exists

Proof by contradiction $\rightarrow$ Assume that $\nabla g(w_+) \neq 0$ implying one value of the gradient is non-zero.

Evaluate directional derivative of g at w_+ in direction of $-\nabla g(w_+)$

$$\Rightarrow D_{-\nabla g(w_+)}\, g(w_+) = (-\nabla g(w_+)) \cdot \nabla g(w_+)$$
$$= -\|\nabla g(w_+)\|^2 < 0$$

Implies moving in direction of $-\nabla g(w_+)$ decreases function value g.

By definition $\exists\ \varepsilon > 0$ s.th. $h = -\varepsilon \nabla g(w_+)$

$$\Rightarrow g(w_+ + h) < g(w_+)$$

If ε is very small, then

$$\|h\|_2 = \varepsilon \|\nabla g(w_+)\|_2 < \gamma$$

$\Rightarrow$ $w_+ + h$ is within γ region.

This contradicts our initial assumption that $g(w_+) \leq g(w)$ $\forall w$ in the γ region.

$\Rightarrow$ $\nabla g(w_+) \neq 0$ must be false.

$$\therefore \nabla g(w_+) = 0$$

Showing the converse is not necessarily true

i.e. where the critical point is a saddle point.

Example

$$g(w) = w^3$$
$$g'(w) = 3w^2 = 0$$

$\Rightarrow$ $w = 0$ is a critical point

$$g''(w) = 6w = 0$$

$\Rightarrow w = 0$ is a saddle point point

as the second derivative is equal to zero

i.e. if gradient at a point is zero, the point is not necessarily a local minimum.

Adrian Kukla DL HW1 Solutions

Q3

A function g is convex if for $\forall\ w_1 \in \mathbb{R}^n$ & w_2 and $\forall\ \lambda \in [0,1]$

$$g(\lambda w_1 + (1-\lambda) w_2) \le \lambda g(w_1) + (1-\lambda) g(w_2)$$

Meaning line segment between two points on g lies above or on the graph

w^* is global minimum if $\forall\ w \in \mathbb{R}^n$, $g(w^*) \le g(w)$

Since g is convex, $\forall\ w \in \mathbb{R}^n$ & $\lambda \in [0,1]$ the inequality holds

$$g(\lambda w + (1-\lambda) w^*) \le \lambda g(w) + (1-\lambda) g(w^*)$$

$\Rightarrow$

$$\frac{g(\lambda w + (1-\lambda) w^*) - g(w^*)}{\lambda} \le g(w) - g(w^*)$$

As $\lambda \to 0^+$ we have the directional derivative of g at w^* on the left hand side of equation in the direction of $w - w^*$

$$\lim_{\lambda \to 0^+} \frac{g(\lambda w + (1-\lambda) w^*) - g(w^*)}{\lambda} = \nabla g(w^*) \cdot (w - w^*)$$

Since $\nabla g(w^*) = 0$

$\Rightarrow \quad 0 \le g(w) - g(w^*)$

$\Rightarrow \quad g(w^*) \le g(w) \quad \forall\ w \in \mathbb{R}^n$

$\therefore$ w^* is the global minimum of g.

ADRIAN KUKLA HW1 Solutions DL

Q4

Compute $\frac{\partial s_i}{\partial z_j}$ $\forall$ i, j i.e. the Jacobian matrix. $J_{ij} = \frac{\partial s_i}{\partial z_j}$

softmax function $s_i = \frac{e^{z_i}}{\sum_k e^{z_k}}$

Consider case when $i = j$ and when $i \neq j$.

Case $i = j$

$$\frac{\partial s_i}{\partial z_i} = \frac{e^{z_i} \cdot \sum_k e^{z_k} - e^{z_i} \cdot e^{z_i}}{(\sum_k e^{z_k})^2}$$

$$e^{z_i} \cdot \sum_k e^{z_k} - e^{z_i} \cdot e^{z_i} = e^{z_i} \left(\sum_k e^{z_k} - e^{z_i} \right)$$

$$\frac{\partial s_i}{\partial z_i} = \frac{e^{z_i} (\sum_k e^{z_k} - e^{z_i})}{(\sum_k e^{z_k})^2} = \frac{e^{z_i}}{\sum_k e^{z_k}} \cdot \frac{\sum_k e^{z_k} - e^{z_i}}{\sum_k e^{z_k}}$$

$$= s_i (1 - s_i)$$

Case $i \neq j$

$$\frac{\partial s_i}{\partial z_j} = \frac{0 \cdot \sum_k e^{z_k} - e^{z_i} \cdot e^{z_j}}{(\sum_k e^{z_k})^2} = \frac{-e^{z_i} \cdot e^{z_j}}{(\sum_k e^{z_k})^2}$$

$$= \frac{-e^{z_i}}{\sum_k e^{z_k}} \cdot \frac{e^{z_j}}{\sum_k e^{z_k}} = -s_i \cdot s_j$$

$\Rightarrow$

Jacobian matrix $J_{ij} : \frac{\partial s_i}{\partial z_j} = \begin{cases} s_i (1 - s_i) & \text{if } i = j \\ -s_i s_j & \text{if } i \neq j \end{cases}$

$$\frac{\partial s}{\partial z} = \text{diag}(s) - ss^T$$

diagonal matrix with elements of s on the diagonal

outer product of s with itself

Adrian Kukla HW1 Solutions DL

Qs 5

Construct the optimization problem as follows;

$$\underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \; -x^T y - H(y) \quad \text{subject to } 1^T y = 1, \; 0 \leq y_i \leq 1 \; \forall i$$

Lagrangian $L(y, \lambda, v) = -x^T y - H(y) + \lambda(1^T y - 1) - v^T y$

where v are non-negativity constraints $y_i \geq 0$

Apply KKT conditions

i.e. $\frac{\partial L}{\partial y_i} = -x_i + \log y_i + 1 + \lambda - v_i = 0$ using stationarity condition

$$\Rightarrow \log y_i = x_i - 1 - \lambda + v_i$$

$$\Rightarrow y_i = e^{x_i - 1 - \lambda + v_i}$$

$v_i \geq 0$ by KKT conditions

However $y_i > 0$ because y_i lies in the interior of the simplex.

$$\therefore v_i = 0$$

$$\Rightarrow y_i = e^{x_i - 1 - \lambda}$$

Using constant $1^T y = 1$

$$\sum_i y_i = \sum e^{x_i - 1 - \lambda} = 1$$

$$e^{-1-\lambda} = \frac{1}{\sum_i e^{x_i}}$$

$$= y_i = e^{x_i} \cdot \frac{1}{\sum_j x_j}$$

Softmax layer ensures the output is a valid prob. distr^n as it projects $x \in \mathbb{R}^d$ onto the interior of the simplex.

Which is the softmax function

$$y_i = s(x)_i = \frac{e^{x_i}}{\sum_j x_j}$$

This means that the softmax finds the most spread out probability distribution as $-H(y)$ wants y to have high entropy meaning it prefers y to be as uniform as possible.

The term $-x^T y$ encourages y to align with x. Softmax balances the two s.th. it leads to a prob. distr^n that encourages larger logits and preserves uniformity. So maximizes entropy.

Adrian Kukla CS 7643 HW1

Q6, Proof by Induction

If G has one node only, graph is trivially a DAG => valid topological ordering.

Assume that any DAG with k nodes has a topological ordering

Evaluate a DAG 'G' with k+1 nodes

Using lemma since G is a DAG it must have one node v with no incoming edges.

Built a new graph $G' = (V', E')$ where $V' \setminus \{v\}$ and E does not involve V

=> Since G is a DAG, G' is also a DAG as removing a node can't create cycles.

Using inductive hypothesis

G' (with k nodes) has a topological ordering $\{v_1, v_2, \dots, v_k\}$

Hence $\{v, v_1, v_2, \dots, v_k\}$ is a valid topological ordering for G

as:
- v has no incoming edges so okay to put first
- all edges in G' satisfy $i < j$
- all edges from v to a node in G' also satisfy $i < j$ since v is first node.

By induction if graph G is a DAG, then G has a topological ordering.

Adrian Kukla CS 7643 HW1

Q7

Proof by contradiction

Assume G has a topological order of G. By def^n/ $\forall$ edges $(v_i, v_j) \in E$, there is $i < j$

Suppose that G has a directed cycle (contradiction statement) s.th. $C = \{v_{k_1}, v_{k_2}, \dots, v_{k_m}, v_{k_1}\}$ where $(v_{k_m}, v_{k_1}) \in E$

However, since $\{v_1, v_2, \dots, v_n\}$ has a topological order, for each edge $(v_{k_i}, v_{k_{i+1}}) \in E$, there is $k_i < k_{i+1}$

Also for edge $(v_{k_m}, v_{k_1}) \in E$, there is $k_m < k_1$

From C,

$$k_1 < k_2 < \dots < k_m < k_1$$

But

$$k_1 < k_1 \text{ is impossible}$$

Conclusion: Assuming a directed cycle leads to a contradiction.

$\therefore$ G cannot contain any directed cycles.

CS7643

HW1 SP25

Adrian Kukla

Question 8

Briefly summarize the key contributions, strengths, and weaknesses of this paper.

The paper challenges the traditional belief that learning weight parameters via gradient descent is the primary driver of a neural network's predictive performance. It demonstrates that weight agnostic neural networks evolve architectures over time without relying on weight training.

Contributions

- Shows that specific architectures can perform well even with shared weights, highlighting the importance of network topology alongside weight optimization.
- Questions whether backpropagation is the only viable strategy for improving neural network performance, proposing a search-based alternative using NeuroEvolution of Augmented Topologies (NEAT).
- Suggests that genetic algorithms could identify architectures that generalize across tasks with minimal weight tuning.

Strengths

- The interactive version provides intuitive visualizations, making complex ideas more accessible.
- Innovatively challenges long-held assumptions, showing that architecture alone can encode strong inductive biases and improve performance, opening avenues for further research.
- Builds on principles inspired by natural behaviors, aligning with arguments like Zador's that backpropagation diverges from genetically inherited learning mechanisms.

Weaknesses

- The problems evaluated are narrow in scope, focusing on simple, structured environments, which may not translate well to unstructured tasks like speech recognition or image processing.
- Lacks a clear roadmap for integrating WANNs into modern deep learning frameworks for practical applications.
- No limitations section is included to critically analyze findings or address counterarguments.
- Discovered architectures might be computationally expensive and poorly suited for GPU acceleration.

Question 9

What is your personal takeaway from this paper? This could be expressed either in terms of relating the approaches adopted in this paper to your traditional understanding of learning parameterized models, or potential future directions of research in the area which the authors haven't addressed, or anything else that struck you as being noteworthy.

My personal takeaway has primarily to do with a life lesson, more than something specific to neural networks. It is that thinking out of the box, and not following a dogmatic world view is valuable and should be encouraged. To not blindly follow the standard practice inherited, and to be a conformist to traditional ways of thinking, but to evaluate new possibilities and be open minded.

In the context of neural networks I learned that it is also possible to tweak architectures and the neural network topology rather than sole use of the traditional backpropagation technique, which I personally thought was the' be-all end-all' of neural networks.

What interests me now is whether we can implement hybrid approaches investigating training weights combined with architectures that co-evolve with the trained weights.

NONE