Question 1)

This paper is aimed at understanding the power law relationship between performance, trained data size and model parameter size. The authors identified two regimes: resolution limited and variance limited. In variance limited regimes, performance improves when increasing either dataset size or parameter size. Under this regime, assuming a large enough dataset and enough model parameters, training is limited by the variance of the data itself.

In resolution limited regimes, either the dataset size or the number of parameters is very large. In this regime, increasing the other parameter improves performance according to a power law. In this regime, the model is thought to better 'resolve' the data manifold, or underlying structure of the data, with more data points or parameters. This is similar to an image, where more pixels allow better resolution of the fine structure of the subject. With more parameters/data, the manifold can be better understood. The authors identified a duality between under parameterized and overparameterized models. This means that models that are under parameterized share the same scaling exponent with models that are over parameterized, meaning that over parameterized models scale the same with more data as under parameterized models scale with more data.

The key takeaway from the paper is that there is a direct relationship between model size and training dataset size.

One weakness of the paper is that it focuses on asymptotic behavior. While beneficial as a theoretical tool, it isn't necessarily practical in application. One interesting extension of this work could be identifying the relationship between computing cost and data manifold resolution. This could provide some insight into the energy cost of training a model.

Question 2)

This paper introduces Kolmogorov-Arnold Networks (KANs). These are an alternative to Multi-Layer Perceptrons (MLPs), that switch out linear activation functions for learnable activation functions. These replace parameters with a 1D spline function.

The principle behind KANs is the Kolmogorov-Arnold Theorem. This theorem states that bounded, multivariate continuous functions can be represented by a sum of individual 1D functions. Effectively each connection between nodes is one of these 1D functions, and the model aims to learn the connections between nodes rather than a weight.

KANs allow for fine tuning of individual connections and work well to approximate low dimensional functions. KANs also typically require much fewer parameters as each connection is more representative of the function to learn. Since each connection is 'learned', KANs can be adjusted dynamically, allowing the user some degree of fine tuning for each connection. Furthermore, KANs lend themselves to interpretability as opposed to MLPs which are often a black box.

KANs are at least as accurate as MLPs, with a smaller footprint. KANs scale better than MLPs as well, obeying better power laws than MLPs. The paper introduces sparsification, which is the idea of reducing connections in the network until the 'best' or 'most necessary' connections are left. Since KAN parameters are local, they are only non zero over a small region, training KANs means that one can target a specific region of the function to learn with their training data. This will not disrupt the parameters in other regions of the function, isolating the region of interest.

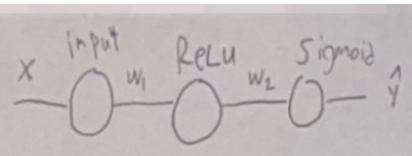
One drawback of using KANs is that training cannot use batch processing. Since each edge has its own parameter to learn, it is impossible to use batches to train the activation function itself.

This paper does a great job of presenting a novel idea. The authors do a good job of highlighting the interpretability of the model and cases where this interpretability outweighs the computational cost.

One future direction I would be interested in is trying to parallelize the training process. If you could group multiple edges by spline ‘type’, you might be able to train batches of edges at a time (Multi headed KAN). Another option is exploring initialization strategies for KANs. MLP initialization is fairly well understood, and it would be interesting to introduce similar ideas to KANs.

John Reilly

3a)



$\hat{y} = \sigma(w_2\, \text{ReLu}(w_1 x))$ $\quad l = -y \log \hat{y} - (1-y)\log(1-\hat{y})$

b) $\frac{\partial l}{\partial v} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial u} \frac{\partial u}{\partial v}$ $\quad (w_1, w_2) = (v, v)$

$u = v\,\text{ReLu}(vx)$

$\frac{\partial l}{\partial \hat{y}} = \frac{-y}{\hat{y}} + (1-y)\left(\frac{1}{1-\hat{y}}\right)$

$\frac{\partial \hat{y}}{\partial u} = \sigma(u)(1-\sigma(u))$

$\frac{\partial u}{\partial v} = 1\,\text{ReLu}(vx) + v\frac{\partial}{\partial v}\text{ReLu}(vx) = \begin{cases} 2vx, & vx > 0 \\ 0 & \text{else} \end{cases}$

$\frac{\partial l}{\partial v} = \left(\frac{-y}{\hat{y}} + (1-y)\frac{1}{1-\hat{y}}\right)\left(\sigma(u)(1-\sigma(u))\right)(2vx)$ if $vx > 0$

$= \left(\frac{-y(1-\hat{y})}{\hat{y}(1-\hat{y})} + \frac{\hat{y}(1-y)}{\hat{y}(1-\hat{y})}\right)\left(\hat{y}(1-\hat{y})\right) 2vx$

$= \left[-y(1-\hat{y}) + \hat{y}(1-y)\right] 2vx$

$= 2vx[-y + y\hat{y} + \hat{y} - y\hat{y}] = \boxed{2vx[\hat{y} - y] = \frac{\partial l}{\partial v}} = 0$ $\begin{cases} \hat{y} = y \to \frac{1}{1+e^{-v\,\text{ReLu}(vx)}} = y \\ 2vx = 0 \end{cases}$

if $vx \le 0$: $y = \hat{y} = 1/2$

if $vx > 0$:

$y(1+e^{-v^2x}) = 1$

$1 + e^{-v^2x} = \frac{1}{y}$

$e^{-v^2x} = \frac{1}{y} - 1$

$-v^2x = \ln\left(\frac{1}{y} - 1\right)$

$v^2 = \frac{-1}{x}\ln\left(\frac{1}{y} - 1\right)$

$v = \pm\sqrt{\frac{-1}{x}\ln\left(\frac{1}{y} - 1\right)}$

For real values: $\frac{-1}{x}\ln\left(\frac{1}{y} - 1\right) \ge 0$

$\frac{1}{y} - 1 \le 1$

$y \ge 1/2$

So $\boxed{v = \pm\sqrt{\frac{-1}{x}\ln\left(\frac{1}{y} - 1\right)},\ y \ge 1/2}$

$\frac{\partial^2 l}{\partial v^2} = \frac{\partial}{\partial v}(\hat{y} - y)(2vx) = 2vx\left(\frac{\partial}{\partial v}\hat{y} - y\right) + 2x(\hat{y} - y) = 2x\left[v\frac{\partial}{\partial v}\hat{y} + (\hat{y} - y)\right]$

$\frac{\partial \hat{y}}{\partial v} = \frac{\partial \hat{y}}{\partial u}\frac{\partial u}{\partial v}$

$\frac{\partial^2 l}{\partial v^2} = 2x\left[2v^2x(\hat{y}(1-\hat{y})) + (\hat{y} - y)\right]$

$\boxed{\frac{\partial^2 l}{\partial v^2} = 4v^2x^2(\hat{y}(1-\hat{y})) + 2x(\hat{y} - y)}$

To check convexity: Plug into Hessian:

$v = 0^+$: $0 + 2x(\hat{y} - y) = 2x\left(\frac{1}{2} - y\right) = x - 2xy = x(1-2y)$ → sign depends on x's sign

$\hat{y} = y$: $v = +ve$: $\frac{\partial^2 l}{\partial v^2} = 4\left(\frac{-1}{x}\ln\left(\frac{1}{y} - 1\right)\right)x^2(\hat{y}(1-\hat{y})) + 2x(\hat{y} - y)$

$= -4x\ln\left(\frac{1}{y} - 1\right)(\hat{y}(1-\hat{y})) = -4x\ln\left(\frac{1}{y} - 1\right)(\hat{y}(1-\hat{y}))$

c) $w_1 = v$, $w_2 = -v$

$\hat{y} = \sigma[-v\,\mathrm{ReLU}(vx)]$ $\quad q = -v\,\mathrm{ReLU}(vx)$

$$\frac{\partial l}{\partial v} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial q} \frac{\partial q}{\partial v}$$

$$\frac{\partial l}{\partial \hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$\frac{\partial \hat{y}}{\partial q} = \hat{y}(1-\hat{y})$$

$$\frac{\partial q}{\partial v} = -v\frac{\partial}{\partial v}\mathrm{ReLU}(vx) - \mathrm{ReLU}(vx) = \begin{cases} -2vx & \text{for } vx > 0 \\ 0 & \text{else} \end{cases}$$

$$\frac{\partial l}{\partial v} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}\,\hat{y}(1-\hat{y})(-2vx) = -2vx(\hat{y}-y) = 0$$

For $vx > 0$: $y = \hat{y}$: $\frac{1}{1+e^{v^2x}} = y$

$1 = y(1+e^{v^2x})$

$\frac{1}{y} - 1 = e^{v^2x}$

$\ln(\frac{1}{y}-1) = v^2x$

$\pm\sqrt{\frac{1}{x}\ln(\frac{1}{y}-1)} = v = v^*$

For real value:

$\ln(\frac{1}{y}-1) \geq 0$

$\frac{1}{y} - 1 \geq 1$

$\frac{1}{y} \geq 2$

$y \leq 1/2$

$$\frac{\partial^2 l}{\partial v^2} = \frac{\partial}{\partial v}(\hat{y}-y)(-2vx) - 2x(\hat{y}-y)$$

$$= -2x[(-2v^2x(\hat{y}(1-\hat{y})) + (\hat{y}-y)]$$

$$\frac{\partial^2 l}{\partial v^2} = 4v^2x^2(\hat{y}(1-\hat{y})) - 2x(\hat{y}-y)$$

Plugging into $\frac{\partial^2 l}{\partial v^2}$: $4x\ln(\frac{1}{y}-1)(\hat{y}(1-\hat{y}))$

So sign depends on: $4x\ln(\frac{1}{y}-1)$: $\ln(\frac{1}{y}-1) \geq 0$ by realness

So if $x > 0$: $\frac{\partial^2 l}{\partial v^2} > 0$

if $x < 0$: $\frac{\partial^2 l}{\partial v^2} < 0$

d) Refer to b + c. The sign of the Hessian depends on the input's sign X.

4a) $\mathcal{L}(y) \geq \mathcal{L}(x) + \langle \nabla \mathcal{L}(x), y-x \rangle + \frac{\mu}{2}\|x-y\|^2$

RHS is quadratic $\rightarrow$ cvx

min. w.r.t. $y \rightarrow \frac{\partial}{\partial y} = 0$

$0 + \nabla \mathcal{L}(x) - \mu\|x-y\| = 0$

$\frac{\nabla \mathcal{L}(x)}{\mu} = \|x-y\|$

$x - \frac{\nabla \mathcal{L}(x)}{\mu} = y$

$\mathcal{L}(y) = \mathcal{L}^* \geq \mathcal{L}(x) + \langle \nabla \mathcal{L}(x), -\frac{\nabla \mathcal{L}(x)}{\mu} \rangle + \frac{\mu}{2}\|\frac{\nabla \mathcal{L}(x)}{\mu}\|^2$

$\mathcal{L}^* \geq \mathcal{L}(x) - \|\nabla \mathcal{L}(x)\|^2 \frac{1}{\mu} + \frac{1}{2\mu}\|\nabla \mathcal{L}(x)\|^2$

$\frac{1}{2\mu} - \frac{1}{\mu} = \frac{1}{2\mu} - \frac{2}{2\mu} = \frac{-1}{2\mu}$

$\mathcal{L}^* \geq \mathcal{L}(x) - \frac{1}{2\mu}\|\nabla \mathcal{L}(x)\|^2$

$2\mu(\mathcal{L}^* - \mathcal{L}(x)) \geq -\|\nabla \mathcal{L}(x)\|^2 \Rightarrow \boxed{2\mu(\mathcal{L}(x) - \mathcal{L}^*) \leq \|\nabla \mathcal{L}(x)\|^2}$

6) $\mathcal{L}(w_{t+1}) \leq \mathcal{L}(w_t) + \nabla \mathcal{L}(w_t)^T (w_{t+1} - w_t) + \frac{L}{2}\|w_{t+1} - w_t\|^2$

$w_{t+1} = w_t - \frac{1}{L}\nabla \mathcal{L}(w_t)$

$\mathcal{L}(w_{t+1}) \leq \mathcal{L}(w_t) - \frac{1}{L}\|\nabla \mathcal{L}(w_t)\|^2 + \frac{L}{2}\|\frac{1}{L}\mathcal{L}(w_t)\|^2$

$\mathcal{L}(w_{t+1}) \leq \mathcal{L}(w_t) - \frac{1}{2L}\|\nabla \mathcal{L}(w_t)\|^2$ $\rightarrow$ From part a

$\leq \mathcal{L}(w_t) - \frac{1}{2L}(2\mu(\mathcal{L}(w_t) - \mathcal{L}^*))$

$\leq \mathcal{L}(w_t) - \frac{\mu}{L}(\mathcal{L}(w_t) - \mathcal{L}^*)$

$\mathcal{L}(w_{t+1}) - \mathcal{L}^* \leq \mathcal{L}(w_t) - \mathcal{L}^* - \frac{\mu}{L}(\mathcal{L}(w_t) - \mathcal{L}^*)$

$\mathcal{L}(w_{t+1}) - \mathcal{L}^* \leq (1 - \frac{\mu}{L})(\mathcal{L}(w_t) - \mathcal{L}^*)$

$\downarrow$ Recursive update

$\mathcal{L}(w_{t+1}) - \mathcal{L}^* \leq (1 - \frac{\mu}{L})^T(\mathcal{L}(w_T) - \mathcal{L}^*)$

$\leq (1 - \frac{\mu}{L})^T \Delta$

using $1 - x \leq e^{-x}$: $(1 - \frac{\mu}{L})^T \Delta \leq e^{-T\mu/L}\Delta$

c) assume $w^* + \tilde{w} \in W$

$\hat{w} \in W$

By definition, $\mathcal{L}(w^*) \le \mathcal{L}(\hat{w})$ for any $\hat{w} \in W$

So $\mathcal{L}(w^*) \le \mathcal{L}(\tilde{w})$

Similarly: $\mathcal{L}(\tilde{w}) + \frac{\lambda}{2}\|\tilde{w}\|^2 \le \mathcal{L}(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|^2$ for any $\hat{w} \in W$

So $\mathcal{L}(\tilde{w}) + \frac{\lambda}{2}\|\tilde{w}\|^2 \le \mathcal{L}(w^*) + \frac{\lambda}{2}\|w^*\|^2$

From L Lipschitz:

$$\mathcal{L}(w_1) \le \mathcal{L}(w_2) + \langle \nabla \mathcal{L}(w_2), w_1 - w_2 \rangle + \frac{L}{2}\|w_1 - w_2\|^2$$

let $w_2 = w^*$, $w_1 = 0$

$$\mathcal{L}(0) \le \mathcal{L}(w^*) + 0 + \frac{L}{2}\|w^*\|^2$$

$$\mathcal{L}(0) - \mathcal{L}^* \le \frac{L}{2}\|w^*\|^2$$

$$\mathcal{L}(w_1) - \mathcal{L}(w^*) \le \frac{L}{2}\|w^*\|^2$$

d) $\eta = \frac{1}{L+\lambda}$

From b: $\mathcal{L}(w_{t+1}) - \tilde{\mathcal{L}} \le e^{-T\left(\frac{\mu}{L}\right)}\Delta$ For μ strongly cvx w/ $\eta = 1/L$

$$\mathcal{L}(w_{t+1}) - \tilde{\mathcal{L}} \le e^{-T\left(\frac{\lambda}{L+\lambda}\right)}\Delta$$

From c: $\mathcal{L}(w_{t+1}) - \tilde{\mathcal{L}} \le e^{-T\left(\frac{\lambda}{L+\lambda}\right)}\Delta \le e^{-T\left(\frac{\lambda}{2L}\right)}\left(\frac{L}{2}\|w^*\|^2\right)$

$$\mathcal{L}(w_{t+1}) - \tilde{\mathcal{L}} \le e^{-T\left(\frac{\lambda}{L+\lambda}\right)}\left(\frac{L}{2}\|w^*\|^2\right)$$

$$\mathcal{L}(w_{t+1}) \le e^{-T\left(\frac{\lambda}{L+\lambda}\right)}\left(\frac{L}{2}\|w^*\|^2\right) + \tilde{\mathcal{L}} \le e^{-T\left(\frac{\lambda}{L+\lambda}\right)}\left(\frac{L}{2}\|w^*\|^2\right) + \mathcal{L}^* + \frac{\lambda}{2}\|w^*\|^2$$

$$\mathcal{L}(w_{t+1}) - \mathcal{L}^* \le e^{-T\left(\frac{\lambda}{L+\lambda}\right)}\left(\frac{L}{2}\|w^*\|^2\right) + \frac{\lambda}{2}\|w^*\|^2$$

4e) $\lambda \geq \frac{2L}{T}\log T \quad \lambda \leq L$

$$\mathcal{L}(w_{t+1}) - \mathcal{L}^* \leq \frac{L}{T}\log T \|w^*\|^2 + \frac{L}{2}\|w^*\|^2 e^{-\frac{2L}{T}\log T\left(\frac{1}{L+\frac{2L}{T}\log T}\right)}$$

$$\leq \frac{L}{T}\log T \|w^*\|^2 + \frac{L}{2}\|w^*\|^2 e^{-2L\log T\left(\frac{1}{LT+2L\log T}\right)}$$

$$\leq \frac{L}{T}\log T \|w^*\|^2 + \frac{L}{2}\|w^*\|^2 e^{-2\log T\left(\frac{T}{T+2\log T}\right)}$$

$$\leq \frac{L}{T}\log T \|w^*\|^2 + \frac{L}{2}\|w^*\|^2 \frac{1}{T^{2T/(T+2\log T)}}$$

as $T \to \infty$: $\frac{2T}{T+2\log T} \to 2$

$$\leq \frac{L}{T}\log T \|w^*\|^2 + \frac{L}{2T^2}\|w^*\|^2 \rightarrow \text{the two are almost equal}$$

T^2 goes to zero first. $\frac{\log T}{T}$ drives big O / upperbound

This shows that regularization introduces a $\log T$ term that grows slowly. This slows convergence, but the benefit of preventing overfitting and improving generalization is worth the cost.