Neural Path Features and Neural Path Kernel : Understanding the role of gates in deep learning

Chandrashekar Lakshminarayanan* **and Amit Vikram Singh***
Indian Institute of Technology Palakkad
{chandru@iitpkd.ac.in,amitkvikram@gmail.com}

Abstract

Rectified linear unit (ReLU) activations can also be thought of as *gates*, which, either pass or stop their pre-activation input when they are *on* (when the pre-activation input is positive) or *off* (when the pre-activation input is negative) respectively. A deep neural network (DNN) with ReLU activations has many gates, and the on/off status of each gate changes across input examples as well as network weights. For a given input example, only a subset of gates are *active*, i.e., on, and the sub-network of weights connected to these active gates is responsible for producing the output. At randomised initialisation, the active sub-network corresponding to a given input example is random. During training, as the weights are learnt, the active sub-networks are also learnt, and potentially hold very valuable information.

In this paper, we analytically characterise the role of active sub-networks in deep learning. To this end, we encode the on/off state of the gates of a given input in a novel *neural path feature* (NPF), and the weights of the DNN are encoded in a novel *neural path value* (NPV). Further, we show that the output of network is indeed the inner product of NPF and NPV. The main result of the paper shows that the *neural path kernel* associated with the NPF is a fundamental quantity that characterises the information stored in the gates of a DNN. We show via experiments (on MNIST and CIFAR-10) that in standard DNNs with ReLU activations NPFs are learnt during training and such learning is key for generalisation. Furthermore, NPFs and NPVs can be learnt in two separate networks and such learning also generalises well in experiments. In our experiments, we observe that almost all the information learnt by a DNN with ReLU activations is stored in the gates - a novel observation that underscores the need to further investigate the role of gating in DNNs.

1 Introduction

We consider deep neural networks (DNNs) with rectified linear unit (ReLU) activations. A special property of ReLU activation is that it can be written as a product of its pre-activation input, say $q \in \mathbb{R}$ and a gating signal, $\gamma_r(q) = \mathbb{1}_{\{q>0\}}$, i.e., $\chi_r(q) = q \cdot \gamma_r(q)$. In what follows, we call χ_r as the ReLU activation and γ_r as the ReLU gate. While the weights (of a DNN) remain the same across input examples, the $1/0$ state of the gates (or simply gates) change across input examples. For each input example, there is a corresponding *active* sub-network consisting of those gates which are 1, and the weights which pass through such gates. This active sub-network can be said to hold the memory for a given input, i.e., only those weights that pass through such active gates contribute to the output. In this viewpoint, at random initialisation of the weights, for a given input example, a random sub-network is active and produces a random output. However, as the weights change during training (say via gradient descent), the gates change, and hence the active sub-networks corresponding to the various

*Both authors contributed equally.

Preprint. Under review.

input examples also change. At the end of training, for each input example, there is a learned active sub-network, and produces a learned output. Thus, the gates of a trained DNN could potentially contain valuable information. In this paper, we study the role of the gates, and the dynamics of the gates while training DNNs using gradient descent (GD). Our findings can be summarised in the following claims which we theoretically/experimentally justify in the paper:

Claim I (see Section 5): *Active sub-networks are fundamental entities in DNNs with ReLU activations.*

Claim II (see Section 6): *Learning of the active sub-networks during training is key for generalisation.*

Before we discuss "Claims I and II" in terms of our novel contributions in Section 1.2, we present the background of *neural tangent feature and kernel* (NTF and NTK) in Section 1.1.

Notation: We denote the set $\{1, \dots, n\}$ by $[n]$. For $x, y \in \mathbb{R}^m$, $\langle x, y \rangle = x^\top y$. The maximum and minimum eigenvalue of a real symmetric matrix A are denoted by $\rho_{\max}(A)$ and $\rho_{\min}(A)$. We consider fully-connected DNNs with w hidden units per layer and $d-1$ hidden layers. The output of the DNN for an input $x \in \mathbb{R}^{d_{in}}$ is denoted by $\hat{y}_\Theta(x) \in \mathbb{R}$, where $\Theta \in \mathbb{R}^{d_{net}}$ are the network weight ($d_{net} = d_{in}w + (d-2)w^2 + w$). We denote by $\Theta(l, j, i)$, the weight connecting the j^{th} hidden unit of layer $l-1$ to the i^{th} hidden unit of layer $l \in [d]$. $\Theta(1) \in \mathbb{R}^{w \times d_{in}}, \Theta(l) \in \mathbb{R}^{w \times w}, \forall l \in \{2, \dots, d-1\}, \Theta(d) \in \mathbb{R}^{w \times 1}$. The dataset is given by $(x_s, y_s)_{s=1}^n \in \mathbb{R}^{d_{in}} \times \mathbb{R}$. The loss function is given by $L_\Theta = \frac{1}{2}\sum_{s=1}^n (\hat{y}_\Theta(x_s) - y_s)^2$. We use $\nabla_\Theta(\cdot)$ stands for the gradient of $(\cdot)$ with respect to the network weights. We use vectorised notations $y = (y_s, s \in [n]), \hat{y}_\Theta = (\hat{y}_\Theta(x_s), s \in [n]) \in \mathbb{R}^n$ for the true and predicted outputs and $e_t = (\hat{y}_{\Theta_t} - y) \in \mathbb{R}^n$ for the error in the prediction. We use $\theta \in \Theta$ to denote single arbitrary weight, and $\partial_\theta(\cdot)$ to denote $\frac{\partial(\cdot)}{\partial \theta}$. $\Sigma \in \mathbb{R}^{n \times n}$ is the input Gram matrix with entries $\Sigma(s, s') = \langle x_s, x_{s'} \rangle$.

1.1 Background: Neural Tangent Feature and Kernel

The NTF and NTK machinery was developed in some of the recent works [9, 1, 4, 6] to understand optimisation and generalisation in DNNs trained using GD. For an input $x \in \mathbb{R}^{d_{in}}$, the NTF is given by $\psi_{x,\Theta} = \nabla_\Theta \hat{y}_\Theta(x) \in \mathbb{R}^{d_{net}}$, i.e., the gradient of the network output with respect to its weights. The NTK matrix K_Θ on the dataset is the $n \times n$ Gram matrix of the NTFs of the input examples, and is given by $K_\Theta(s, s') = \langle \psi_{x_s,\Theta}, \psi_{x_{s'},\Theta} \rangle, s, s' \in [n]$.

Proposition 1.1 (**Lemma 3.1** Arora et al. [2019])**.** *Consider the GD procedure to minimise the squared loss $L(\Theta)$ with infinitesimally small step-size: $\dot{\Theta}_t = -\nabla_\Theta L_{\Theta_t}$. If follows that the dynamics of the error term can be written as $\dot{e}_t = -K_{\Theta_t} e_t$.*

Prior works [9, 6, 1, 4] have studied DNNs trained using GD in the so called 'NTK regime', which occurs under appropriate randomised initialisation, and when the width of the DNN approaches infinity. The characterising property of the NTK regime is that as $w \to \infty$, $K_{\Theta_0} \to K^{(d)}$, and $K_{\Theta_t} \approx K_{\Theta_0}$, where $K^{(d)}$ (see (2) in Appendix A) is a deterministic matrix whose superscript (d) denotes the depth of the DNN. Arora et al. [2019] show that infinite width DNN trained using GD is equivalent to kernel regression with the limiting NTK matrix $K^{(d)}$ (and hence enjoys the generalisation ability of the limiting NTK matrix $K^{(d)}$). Further, Arora et al. [2019] propose a pure kernel method based on what they call the CNTK, which is the limiting NTK matrix $K^{(d)}$ for an infinite width convolutional neural network (CNN). Cao and Gu [2019] show that in the NTK regime, a DNN is almost a linear learner with the random NTFs at initialisation, and show a generalisation bound in the form of $\tilde{\mathcal{O}}\left(d \cdot \sqrt{y^\top \left(K^{(d)}\right)^{-1} y / n}\right)$[2].

Open Question: Arora et al. [2019] report a $5\% - 6\%$ performance gain of finite width CNNs (which do not operate in the NTK regime) over the exact CNTKs corresponding to infinite width CNNs, and infer that the study of DNNs in the NTK regime cannot fully explain the success of practical neural networks yet. Can we explain the reason for the performance gain of CNNs over CNTK?

[2] $a_t = \mathcal{O}(b_t)$ if $\limsup_{t \to \infty} |a_t / b_t| < \infty$, and $\tilde{\mathcal{O}}(\cdot)$ is used to hide logarithmic factors in $\mathcal{O}(\cdot)$.

1.2 Our Contributions

To the best of our knowledge, we are the first to analytically characterise the role played by active sub-networks in deep learning as presented in the 'Claims I and II'. The key contributions can be arranged into three landmarks as described below.

• The first step involves breaking a DNN into individual paths, and each path again into gates and weights. To this end, we encode the states of the gates in a novel *neural path feature* (NPF) and the weights in a novel *neural path value* (NPV) and express the output of the DNN as an inner product of NPF and NPV (see Section 2). In contrast to NTF/NTK which are *first-order* quantities (based on derivatives with respect to the weights), NPF and NPV are *zeroth-order* quantities. The kernel matrix associated to the NPFs namely the *neural path kernel* (NPK) matrix $H_\Theta \in \mathbb{R}^{n\times n}$ has a special structure, i.e., it can be written as a *Hadamard* product of the input Gram matrix, and a correlation matrix $\Lambda_\Theta \in \mathbb{R}^{n\times n}$, whose entries $\Lambda_\Theta(s, s')$ is equal to the total number of path in the sub-network that is active for both input examples $s, s' \in [n]$. With the Λ_Θ matrix we reach our first landmark.

• Second step is to characterise performance of active sub-networks in a 'stand alone' manner. To this end, we consider a new idealised setting namely fixed NPF (FNPF) setting, wherein, the NPFs are fixed (i.e., held constant) and only the NPV is learnt via gradient descent. In this setting, we show that (see Theorem 5.1), in the limit of infinite width and under randomised initialisation the NTK converges to a matrix $K^{(d)}_{\text{FNPF}} = \text{constant} \times H_{\text{FNPF}}$, where $H_{\text{FNPF}} \in \mathbb{R}^{n\times n}$ is the NPK matrix corresponding to the fixed NPFs. $K^{(d)}$ matrix of Jacot et al. [2018], Arora et al. [2019], Cao and Gu [2019] becomes the $K^{(d)}_{\text{FNPF}}$ matrix in the FNPF setting, wherein, we initialise the NPV statistically independent of the fixed NPFs (see Assumption 5.1). With Theorem 5.1, we reach our second landmark, i.e. we justify "Claim I", that active sub-networks are fundamental entities, which follows from the fact that $H_{\text{FNPF}} = \Sigma \odot \Lambda_{\text{FNPF}}$, where Λ_{FNFP} corresponds to the fixed NPFs.

• Third step is to show experimentally that sub-network learning happens in practice. We show that in finite width DNNs with ReLU activations, NPFs are learnt continuously during training, and such learning is key for generalisation. We observe that fixed NPFs obtained from the initial stages of training generalise poorly than CNTK (of Arora et al. [2019]), whereas, fixed NPFs obtained from later stages of training generalise better than CNTK and generalise as well as standard DNNs with ReLU. This throws light on the open question in Section 1.1, i.e., the difference between the NTK regime and the finite width DNNs is perhaps due to NPF learning. In finite width DNNs, NPFs are learnt during training and in the NTK regime no such feature learning happens during training (since $K^{(d)}$ is fixed). Since the NPFs completely encode the information pertaining to the active sub-networks, we complete our final landmark namely justification of "Claim II",

2 Neural Path Feature and Kernel: Encoding Gating Information

The gating property of the ReLU activation allows us to express the output of the DNN as a summation of the contribution of the individual paths, and paves a natural way to encode the $1/0$ states of the gates *without loss of information*. The contribution of a path is the product of the signal in its input node, the 'd' weights in the path and the '$(d-1)$' gates in the path. For an input $x \in \mathbb{R}^{d_{in}}$, and parameter $\Theta \in \mathbb{R}^{d_{net}}$, we encode the gating information in a novel *neural path feature* (NPF), $\phi_{x,\Theta} \in \mathbb{R}^P$ and the weights in a novel *neural path value* (NPV) $v_\Theta \in \mathbb{R}^P$, where, $P = d_{in}w^{(d-1)}$ is the total number of paths. The NPF co-ordinate of a path is the product of the signal at its input node and the gates in the path. The NPV co-ordinate of a path is the product of the weights in the paths. By stacking the NPFs of all the input examples we obtain the NPF matrix as $\Phi_\Theta = (\phi_{x_s,\Theta}, s \in [n]) \in \mathbb{R}^{P\times n}$. Then the input-output relationship of a DNN in vector form is given by:

$$\hat{y}_\Theta = \Phi_\Theta^\top v_\Theta, \tag{1}$$

where the NPF matrix Φ_Θ can also be interpreted as the *hidden feature matrix* which along with v_Θ is learnt during gradient descent on $\Theta \in \mathbb{R}^{d_{net}}$.

2.1 Paths, Neural Path Feature, Neural Path Value and Network Output

A path starts from an input node, passes through exactly one weight (and one hidden node) in each layer and ends at the output node. We have a total of $P = d_{in}w^{(d-1)}$ paths. Let us say that an

Input Layer	:	$z_{x,\Theta}(0)$	$=$	x
Pre-Activation	:	$q_{x,\Theta}(l,i)$	$=$	$\Theta(l,\cdot,i)^\top z_{x,\Theta}(l-1), l \in [d-1], i \in [w]$
Gating Values	:	$G_{x,\Theta}(l,i)$	$=$	$\gamma_r(q_{x,\Theta}(l,i)), l \in [d-1], i \in [w]$, where $\gamma_r(q) = \mathbb{1}_{\{q>0\}}$
Hidden Layer	:	$z_{x,\Theta}(l,i)$	$=$	$\chi_r(q_{x,\Theta}(l,i)) = q_{x,\Theta}(l,i) \cdot G_{x,\Theta}(l,i), l \in [d-1], i \in [w]$
Final Output	:	$\hat{y}_\Theta(x)$	$=$	$\Theta(d)^\top z_{x,\Theta}(d-1)$

Table 1: DNN with ReLU activation. Here, $x \in \mathbb{R}^{d_{in}}$ is the input to the DNN, and $\hat{y}_\Theta(x)$ is the output, 'q's are pre-activation inputs, 'z's are output of the hidden layers, 'G's are the gating values. $l \in [d-1]$ is the index of the layer, and $i \in [w]$ is the index of the hidden units in a layer.

enumeration of the paths is given by $[P] = \{1, \ldots, P\}$. Let $\mathcal{I}_l \colon [P] \to [w], l = 0, \ldots, d-1$ provide the index of the hidden unit through which a path p passes in layer l (with the convention that $\mathcal{I}_d(p) = 1, \forall p \in [P]$).

Definition 2.1. *Let $x \in \mathbb{R}^{d_{in}}$ be the input to the DNN. For this input,*

(i) The activity of a path p is given by : $A_\Theta(x,p) \stackrel{def}{=} \Pi_{l=1}^{d-1} G_{x,\Theta}(l, \mathcal{I}_l(p))$.

(ii) The neural path feature (NPF) is given by : $\phi_{x,\Theta} \stackrel{def}{=} (x(\mathcal{I}_0(p))A_\Theta(x_s,p), p \in [P]) \in \mathbb{R}^P$.

(iii) The neural path value (NPV) if given by : $v_\Theta \stackrel{def}{=} \left(\Pi_{l=1}^{d} \Theta(l, \mathcal{I}_{l-1}(p), \mathcal{I}_l(p)), p \in [P]\right) \in \mathbb{R}^P$.

A path p is active if all the gates in the paths are on.

Proposition 2.1. *The output of the network can be written as an inner product of the NPF and NPV, i.e.,* $\hat{y}_\Theta(x) = \langle \phi_{x,\Theta}, v_\Theta \rangle = \sum_{p \in [P]} x(\mathcal{I}_0(p)) A_\Theta(x,p) v_\Theta(p)$.

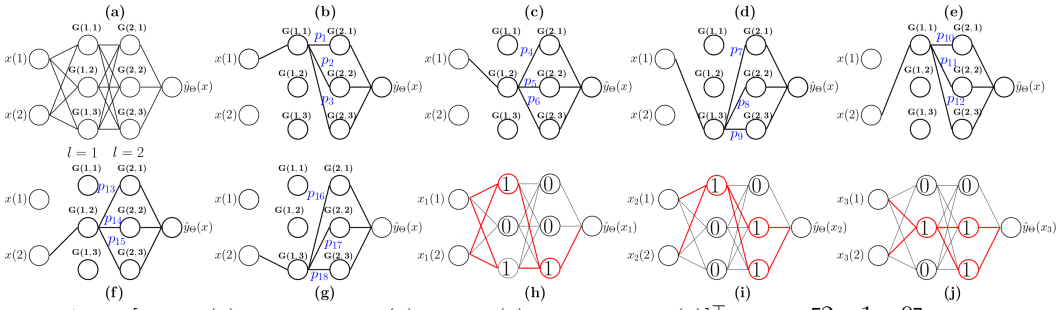

Figure 1: A toy illustration of gates, paths and active sub-networks. The cartoon (**a**) in the top left corner shows a DNN with 2 hidden layers, 6 ReLU gates $G(l,i), l = 1,2, i = 1,2,3$, 2 input nodes $x(1)$ and $x(2)$ and an output node $\hat{y}_\Theta(x)$. Cartoons (**b**) to (**g**) show the enumeration of the paths $p_1, \ldots, p_{18}$. Cartoons (**h**), (**i**) and (**j**) show hypothetical gates for 3 different hypothetical input examples $\{x_s\}_{s=1}^3 \in \mathbb{R}^2$. In each of the cartoons (**h**), (**i**) and (**j**), the $1/0$ inside the circles denotes the on/off state of the gates, and the bold paths/gates shown in red colour constitute the active sub-network for that particular input example. The NPFs are given by $\phi_x = [x(1)A(x,p_1), \ldots, x(1)A(x,p_9), x(2)A(x,p_{10}), \ldots, x(2)A(x,p_{18})]^\top$.Here, $\Lambda(1,2) = 1$ because paths p_3 and p_{12} are both active for input examples x_1 and x_2 and the input dimension is 2.

2.2 Neural Path Kernel : Similarity based on active sub-networks

Definition 2.2. *For input examples $s, s' \in [n]$, define* $\mathcal{A}_\Theta(s,s') \stackrel{def}{=} \{p \in [P] \colon A_\Theta(x_s,p) = A_\Theta(x_{s'},p) = 1\}$ *to be the set of 'active' paths for both s, s' and* $\Lambda_\Theta(s,s') \stackrel{def}{=} \frac{|\mathcal{A}_\Theta(s,s')|}{d_{in}}$.

Remark: Owing to the symmetry of a DNN, the same number of active paths start from any fixed input node. In Definition 2.2, Λ_Θ measures the size of the active sub-network as the total number of

active paths starting from any fixed input node. For examples $s, s' \in [n], s \neq s'$, $\Lambda_\Theta(s,s)$ is equal to the size of the sub-network active for s, and $\Lambda_\Theta(s,s')$ is equal to the size of the sub-network active for both s and s'. For an illustration of NPFs and Λ please see Figure 1.

Lemma 2.1. *Let $H_\Theta \stackrel{def}{=} \Phi_\Theta^\top \Phi_\Theta$ be the NPK matrix, and $\Lambda_\Theta \in \mathbb{R}^{n\times n}$ be as in Definition 2.2 . It follows that $H_\Theta = \Sigma \odot \Lambda_\Theta$, where $\odot$ is the Hadamard product, and Σ is the input Gram matrix.*

3 Dynamics of Gradient Descent with NPF and NPV Learning

In Section 2, we mentioned that during gradient descent, the DNN is learning a relation $\hat{y}_\Theta = \Phi_\Theta^\top v_\Theta$, i.e., both the NPFs and the NPV are learnt. In this section, we connect the newly defined quantities, i..e, Φ_Θ and v_Θ to the NTK matrix K_Θ (see Proposition 3.1), and re-write the gradient descent dynamics in Proposition 3.2 taking into account of NPF and NPV learning.

3.1 NPV and NPF Learning

Definition 3.1. *The gradient of the NPV of path p is defined as $\varphi_{p,\Theta} \stackrel{def}{=} (\partial_\theta v_\Theta(p), \theta \in \Theta) \in \mathbb{R}^{d_{net}}$.*

Remark The change of the NPV is given by $\dot{v}_{\Theta_t}(p) = \langle \varphi_{p,\Theta_t}, \dot{\Theta}_t \rangle$, where $\dot{\Theta}_t$ is the change of the weights. We now collect the gradients $\varphi_{p,\Theta}$ of all the paths to define a *value tangent kernel* (VTK).

Definition 3.2. *Let $\nabla_\Theta v_\Theta$ be a $d_{net} \times P$ matrix of NPV derivatives given by $\nabla_\Theta v_\Theta = (\varphi_{p,\Theta}, p \in [P])$. Define the VTK to be the $P \times P$ matrix given by $\mathcal{V}_\Theta = (\nabla_\Theta v_\Theta)^\top (\nabla_\Theta v_\Theta)$.*

Remark An important point to note here is that the VTK is a quantity that is dependent only on the weights. To appreciate the same, consider a deep linear network (DLN) [15, 5] which has identity activations, i.e., all the gates are 1 for all inputs, and weights. For a DLN and DNN with identical network architecture (i.e., w and d), and identical weights, $\mathcal{V}_\Theta$ is also identical. Thus, $\mathcal{V}_\Theta$ is the gradient based information that excludes the gating information.

The NPFs changes at those time instants when any one of the gates switches from 1 to 0 or from 0 to 1. In the time between two such switching instances, NPFs of all the input examples in the dataset remain the same, and between successive switching instances, the NPF of at least one of the input example in the dataset changes. In what follows, in Proposition 3.2 we re-write Proposition 1.1 taking into account the switching instances which we define in Definition 3.3.

Definition 3.3. *Define a sequence of monotonically increasing time instants $\{T_i\}_{i=0}^\infty$ (with $T_0 = 0$) to be 'switching' instants if $\phi_{x_s,\Theta_t} = \phi_{x_s,\Theta_{T_i}}, \forall s \in [n], \forall t \in [T_i, T_{i+1}), i = 0, \ldots, \infty$, and $\forall i = 0, \ldots, \infty \; \exists s(i) \in [n]$ such that $\phi_{x_{s(i)},\Theta_{T_i}} \neq \phi_{x_{s(i)},\Theta_{T_{i+1}}}$.*

3.2 Gradient Descent

Proposition 3.1. *The NTK is given by $K_\Theta = \Phi_\Theta^\top \mathcal{V}_\Theta \Phi_\Theta$.*

Remark K_{Θ_t} changes during training (i) continuously at all $t \geq 0$ due to $\mathcal{V}_{\Theta_t}$, and (ii) at switching instants $T_i, i = 0, \ldots, \infty$ due to the change in $\Phi_{\Theta_{T_i}}$. We now describe the gradient descent dynamics taking into the dynamics of the NPV and the NPFs.

Proposition 3.2. *Let $\{T_i\}_{i=0}^\infty$ be as in Definition 3.3. For $t \in [T_i, T_{i+1})$ and small step-size of GD:*

$$\begin{array}{lcll} \text{Weights Dynamics} & : & \dot{\Theta}_t & = & -\sum_{s=1}^n \psi_{x_s,\Theta_t} e_t(s) \\ \text{NPV Dynamics} & : & \dot{v}_{\Theta_t}(p) & = & \langle \varphi_{p,\Theta_t}, \dot{\Theta}_t \rangle, \forall p \in [P] \\ \text{Error Dynamics} & : & \dot{e}_t & = & -K_{\Theta_t} e_t, \text{ where } K_{\Theta_t} = \Phi_{\Theta_{T_i}}^\top \mathcal{V}_{\Theta_t} \Phi_{\Theta_{T_i}} \end{array}$$

Proposition 3.3. $\rho_{min}(K_\Theta) \leq \rho_{\min}(H_\Theta) \rho_{\max}(\mathcal{V}_\Theta)$.

Remark For the NTK to be well conditioned, it is necessary for the NPK to be well conditioned. This is quite intuitive, in that, the closer two inputs are, the closer are their NPFs, and it is harder to train the network to produce arbitrarily different outputs for such inputs that are very close to one another.

4 Deep Gated Networks: Decoupling Neural Path Feature and Value

In order to ascertain that NPF learning indeed makes a difference, we should measure the generalisation performance with and without NPF learning. This can be achieved by a deep gated network (see Figure 2 below for details) having two networks of identical architecture namely i) a feature network parameterised by $\Theta^{\mathrm{F}} \in \mathbb{R}^{d_{net}}$, that holds gating information, and hence the NPFs and ii) a value network that holds the NPVs parameterised by $\Theta^{\mathrm{V}} \in \mathbb{R}^{d_{net}}$. In what follows, we let $\Theta^{\text{DGN}} = (\Theta^{\mathrm{F}}, \Theta^{\mathrm{V}}) \in \mathbb{R}^{2d_{net}}$ to denote the combined parameters of a DGN. By making $\Theta^{\mathrm{F}} \in \mathbb{R}^{d_{net}}$ trainable/non-trainable, we can *enable/disable* the NPF gradient, which gives rise to the following two modes of operating a DGN:

1. **Fixed NPF (FNPF):** Here, $\Theta^{\mathrm{F}}_t = \Theta^{\mathrm{F}}_0, \forall t \geq 0$, i.e., $\Theta^{\mathrm{F}} \in \mathbb{R}^{d_{net}}$ is non-trainable. Thus the DGN learns the relation $\hat{y}_{\Theta^{\text{DGN}}} = \Phi^{\top}_{\Theta^{\mathrm{F}}_0} v_{\Theta^{\mathrm{V}}}$, where $\Phi_{\Theta^{\mathrm{F}}_0} \in \mathbb{R}^{P \times n}$ is a fixed NPF matrix, and $v_{\Theta^{\mathrm{V}}}$ is learned via gradient descent on $\Theta^{\mathrm{V}} \in \mathbb{R}^{d_{net}}$.

2. **Decoupled NPF Learning (DNPFL):** Here both $\Theta^{\mathrm{F}} \in \mathbb{R}^{d_{net}}$ and $\Theta^{\mathrm{V}} \in \mathbb{R}^{d_{net}}$ are trained, and the DGN learns the relation $\hat{y}_{\Theta^{\text{DGN}}} = \Phi^{\top}_{\Theta^{\mathrm{F}}} v_{\Theta^{\mathrm{V}}}$. In comparison to (1), here we have two parameters $\Theta^{\mathrm{F}} \in \mathbb{R}^{d_{net}}$ and $\Theta^{\mathrm{V}} \in \mathbb{R}^{d_{net}}$ as opposed to a single $\Theta \in \mathbb{R}^{d_{net}}$ in (1).

Note: FNPF and DNPFL are idealised modes to understand the role of gates, and not alternate proposals to replace standard DNNs with ReLU activations.

Layer	Feature Network (NPF)	Value Network (NPV)
Input	$z^{\mathrm{F}}_{x,t}(0) = x$	$z^{\mathrm{V}}_{x,t}(0) = x$
Activation	$q^{\mathrm{F}}_{x,t}(l) = \Theta^{\mathrm{F}}_t(l)^{\top} z^{\mathrm{F}}_{x,t}(l-1)$	$q^{\mathrm{V}}_{x,t}(l) = \Theta^{\mathrm{V}}_t(l)^{\top} z^{\mathrm{V}}_{x,t}(l-1)$
Hidden	$z^{\mathrm{F}}_{x,t}(l) = \chi^{\mathrm{F}}\left(q^{\mathrm{F}}_{x,t}(l)\right)$	$z^{\mathrm{V}}_{x,t}(l) = q^{\mathrm{V}}_{x,t}(l) \odot G_{x,t}(l)$
Output	None	$\hat{y}_t(x) = \Theta^{\mathrm{V}}(d)^{\top} z^{\mathrm{V}}_{x,t}(d-1)$
Gating Values:	$G_{x,t}(l) = \gamma_r\left(q^{\mathrm{F}}_{x,t}(l)\right)$	or $G_{x,t}(l) = \gamma_{sr}\left(q^{\mathrm{F}}_{x,t}(l)\right)$

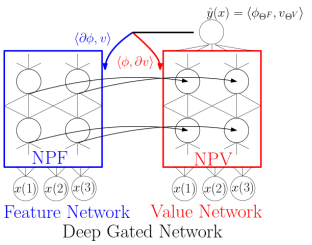

Figure 2: Deep gated network (DGN) setup. The pre-activations $q^{\mathrm{F}}_{x,t}(l)$ of layer $l \in [d-1]$ from the feature network are used to derive the gating values $G_{x,t}(l)$ of layer $l \in [d-1]$.

Proposition 4.1 (Gradient Dynamics in a DGN). *Let* $\psi^{F}_{x,\Theta^{DGN}} \overset{def}{=} \nabla_{\Theta^F} \hat{y}_{\Theta^{DGN}}(x) \in \mathbb{R}^{d_{net}}$, $\psi^{V}_{x,\Theta^{DGN}} \overset{def}{=} \nabla_{\Theta^V} \hat{y}_{\Theta^{DGN}}(x) \in \mathbb{R}^{d_{net}}$. *Let* $K^{V}_{\Theta^{DGN}}$ *and* $K^{F}_{\Theta^{DGN}}$ *be* $n \times n$ *matrices with entries* $K^{V}_{\Theta^{DGN}}(s,s') = \langle \psi^{V}_{x_s,\Theta^{DGN}}, \psi^{V}_{x_{s'},\Theta^{DGN}} \rangle$ *and* $K^{F}_{\Theta^{DGN}}(s,s') = \langle \psi^{F}_{x_s,\Theta^{DGN}}, \psi^{F}_{x_{s'},\Theta^{DGN}} \rangle$. *For infinitesimally small step-size of GD, the error dynamics in a DGN (in the DNPFL and FNPF modes) is given by:*

Dynamics	DNPFL			FNPF
Weight	$\dot{\Theta}^{\mathrm{V}}_t$	$=$	$-\sum_{s=1}^{n} \psi^{\mathrm{V}}_{x,\Theta^{\text{DGN}}_t} e_t(s), \dot{\Theta}^{\mathrm{F}}_t = -\sum_{s=1}^{n} \psi^{\mathrm{F}}_{x,\Theta^{\text{DGN}}_t} e_t(s)$	$\dot{\Theta}^{\mathrm{V}}_t$ same as (DNFPL), $\dot{\Theta}^{\mathrm{F}}_t = 0$
NPF	$\dot{\phi}_{x_s,\Theta^{\mathrm{F}}_t}(p)$	$=$	$x(\mathcal{I}_0(p)) \sum_{\theta^{\mathrm{F}} \in \Theta^{\mathrm{F}}} \partial_{\theta^{\mathrm{F}}} A_{\Theta^{\mathrm{F}}_t}(x_s,p) \dot{\theta}^{\mathrm{F}}_t, \forall p \in [P], s \in [n]$	$\dot{\phi}_{x_s,\Theta^{\mathrm{F}}_t}(p) = 0$
NPV	$\dot{v}_{\Theta^{\mathrm{V}}_t}(p)$	$=$	$\sum_{\theta^{\mathrm{V}} \in \Theta^{\mathrm{V}}} \partial_{\theta^{\mathrm{V}}} v_{\Theta^{\mathrm{V}}_t}(p) \dot{\theta}^{\mathrm{V}}_t, \forall p \in [P]$	$\dot{v}_{\Theta^{\mathrm{V}}_t}(p)$ same as DNPFL
Error	$\dot{e}_t$	$=$	$-\left(K^{\mathrm{V}}_{\Theta^{\text{DGN}}} + K^{\mathrm{F}}_{\Theta^{\text{DGN}}}\right) e_t$	$\dot{e}_t = -\left(K^{\mathrm{V}}_{\Theta^{\text{DGN}}}\right) e_t$

Remark: The gradient dynamics in a DGN specified in Proposition 4.1 is similar to the gradient dynamics in a DNN specified in Proposition 3.2. Important difference is that in a DGN there are $2d_{net}$ parameters, and hence the NTF $\psi_{x,\Theta} = (\psi^{\mathrm{F}}_{x,\Theta}, \psi^{\mathrm{V}}_{x,\Theta}) \in \mathbb{R}^{2d_{net}}$, wherein, $\psi^{\mathrm{V}}_{x,\Theta^{\text{DGN}}} \in \mathbb{R}^{d_{net}}$ flows through the value network and $\psi^{\mathrm{F}}_{x,\Theta^{\text{DGN}}} \in \mathbb{R}^{d_{net}}$ flows through the feature network.

5 Learning with Fixed NPFs: Role Of Active Sub-Networks

In this section, we provide theoretical justification for "Claim I", i.e., the active sub-networks are fundamental entities in DNNs.

Definition 5.1. *Define the measure of information stored in the gates of a* DNN *with parameter* $\bar{\Theta} \in \mathbb{R}^{d_{net}}$ *to be the generalisation performance of a* DGN *with identical architecture operated in the* FNPF *mode whose* $\Theta^{F}_0 = \bar{\Theta}$ *are non-trainable, and* $\Theta^{V} \in \mathbb{R}^{d_{net}}$ *are trained.*

Consider a DNN parameterised by $\bar{\Theta} \in \mathbb{R}^{d_{net}}$. At randomised initialisation, we can obtain random NPFs $\Phi_{\bar{\Theta}_0}$, and after training for T epochs, and we can obtain learnt NPFs $\Phi_{\bar{\Theta}_T}$. Thus, while measuring information in the gates of this trained DNN, as per Definition 5.1, we are retaining $\Phi_{\bar{\Theta}_T}$ by storing the weights as $\Theta_0^{\text{F}} = \bar{\Theta}_T$ in the feature network, and discarding $v_{\bar{\Theta}_T}$, and re-training Θ^{V} to learn a new relation $\hat{y}_{\Theta^{\text{DGN}}} = \Phi_{\Theta_0^{\text{F}}}^{\top} v_{\Theta^{\text{V}}} = \Phi_{\bar{\Theta}_T}^{\top} v_{\Theta^{\text{V}}}$. Similarly, in the case of random NPFs we are learning the relation, $\hat{y}_{\Theta^{\text{DGN}}} = \Phi_{\Theta_0^{\text{F}}}^{\top} v_{\Theta^{\text{V}}} = \Phi_{\bar{\Theta}_0}^{\top} v_{\Theta^{\text{V}}}$. In what follows, we use H_{FNFP} to refer to $H_{\Theta_0^{\text{F}}}$.

Assumption 5.1. *(i) $\Theta_0^V \in \mathbb{R}^{d_{net}}$ is statistically independent of the fixed NPFs (stored in $\Theta_0^F \in \mathbb{R}^{d_{net}}$ of the feature network), (ii) Θ_0^V are sampled i.i.d from symmetric Bernoulli over $\{-\sigma, +\sigma\}$.*

Theorem 5.1. *Under Assumption 5.1, as $w \to \infty$, $K_{\Theta_0^{DGN}} \to K_{FNPF}^{(d)} = d \cdot \sigma^{2(d-1)} H_{FNPF}$.*

• **Active Sub-Network:** From previous results Arora et al. [2019], it follows that as $w \to \infty$, the optimisation and generalisation properties of the fixed NPF learner can be tied down to the infinite width NTK of the FNPF learner $K_{\text{FNPF}}^{(d)}$ and hence to H_{FNPF} (treating $d\sigma^{2(d-1)}$ as a scaling factor). We can further breakdown $H_{\text{FNPF}} = \Sigma \odot \Lambda_{\text{FNPF}}$, where $\Lambda_{\text{FNPF}} = \Lambda_{\Theta_0^{\text{F}}}$. This justifies "Claim I".

• $K^{(d)}$ in prior works [9, 1, 4] essentially becomes $K_{\text{FNPF}}^{(d)}$ under Assumption 5.1. To understand this, let us consider a DNN with weights $\Theta \in \mathbb{R}^{d_{net}}$. From [9, 1, 4] it follows that under randomised initialisation of $\Theta_0 \in \mathbb{R}^{d_{net}}$ as $w \to \infty$ the NTK of the DNN $K_{\Theta_0} \to K^{(d)}$. The simplification of $K^{(d)}$ to $K_{\text{FNPF}}^{(d)}$ in Theorem 5.1 occurs when we copy these random NPFs corresponding to $\Theta_0 \in \mathbb{R}^{d_{net}}$ into the feature network and keep them fixed, i.e., $\Theta_t^{\text{F}} = \Theta_0^{\text{F}} = \Theta_0 \in \mathbb{R}^{d_{net}}, t \geq 0$, and train $\Theta^{\text{V}} \in \mathbb{R}^{d_{net}}$ with initialisation as per Assumption 5.1.

• **Choice of σ:** In the case of random NPFs obtained by initialising Θ_0^{F} at random by sampling from a symmetric distribution, we expect $\frac{w}{2}$ gates to be on every layer, so $\sigma = \sqrt{\frac{2}{w}}$ is a normalising choice, in that, the diagonal entries of $\sigma^{2(d-1)}\Lambda_{\text{FNPF}}(s,s) \approx 1$ in this case.

• We discuss a more detailed version of Theorem 5.1 in the Appendix, where we discuss the role of width and depth on a pure memorisation task.

6 Experiments: Fixed NPFs, NPF Learning and Verification of Claim II

In this section, we justify "Claim II", i.e., active sub-networks learning is key for generalisation. Since the active sub-network are encoded in the NPFs, we verify the claim by comparing different network settings which vary in their NPF learning capabilities. We resolve the open question of Arora et al. [2019] mentioned in Section 1.1, by providing an empirical explanation for the performance gain of finite width CNN over the pure kernel method based on the exact infinite width CNTK.

Networks for Comparison: The performance of the following networks on standard MNIST and CIFAR-10 datasets will be used for comparison: (i) fixed random (FRNPF): in the DGN, we randomly initialise both $\Theta_0^{\text{F}}, \Theta_0^{\text{V}}$, make Θ^{F} *non-trainable* and train only Θ^{V} , (ii) fixed learnt (FLNPF): we initialise Θ_0^{V} randomly, and copy weights from a pre-trained ReLU network (of identical architecture) into Θ_0^{F}. Similar to FR case, Θ^{F} is non-trainable and only Θ^{V} is trained (iii) decoupled learning (DNPFL): we randomly initialise both $\Theta_0^{\text{F}}, \Theta_0^{\text{V}}$, and train both Θ^{F} and Θ^{V}, (iv) ReLU: Standard DNNs/CNNs with ReLU. We will also use the numerical results reported in Arora et al. [2019].

1. **Finite Vs Infinite width alone is not enough to explain the performance gain of CNN:** Both FRNPF and ReLU are finite width networks. However, performance of FRNPF is approximately 67% which is worse than CNTK of Arora et al. [2019] whose performance is 77.43%, and the performance of our CNN architecture with *global-average-pooling* (GCONV in Table 2) is 80.34%. We trained FRNPF with independent initialisation (II), where Θ_0^{F} and Θ_0^{V} are statistically independent, and dependent initialisation (DI), where $\Theta_0^{\text{F}} = \Theta_0^{\text{V}}$. FRNPF (II) and FRNPF (DI) were close in our experiments (see columns 4 and 5 in Table 2). Further, both FRNPF (DI) and ReLU start with the same NTK matrix at initialisation. If finite width was the sole reason for the better performance then even FRNPF (II), (DI) should have performed better than CNTK in the experiments, and they did not. Thus, finite width alone does not explain the performance gain of CNN over CNTK.

2. **NPF Learning Vs No NPF Learning is key to explain the performance gain of CNN:** FLNPF with weights copied from a fully trained ReLU performs close to 79.68% which is almost as good as ReLU's 80.43% (see FLNP column in Table 2). Further, NPFs are learnt continuously during

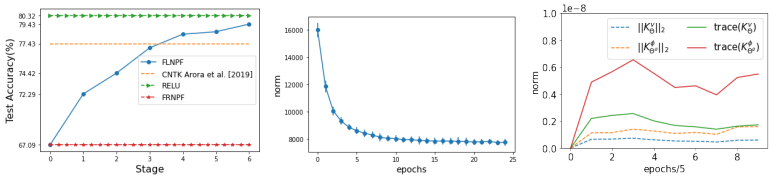

Figure 3: Dynamics of NPF Learning.

Arch	Optimiser	Dataset	FRNPF (II)	FRNPF (DI)	DNPFL	FLNPF	ReLU
FC	SGD	MNIST	95.85 ± 0.10	95.85 ± 0.17	97.86 ± 0.11	97.10 ± 0.09	97.85 ± 0.09
FC	Adam	MNIST	96.02 ± 0.13	96.09 ± 0.12	$\mathbf{98.22 \pm 0.05}$	$\mathbf{97.82 \pm 0.02}$	$\mathbf{98.14 \pm 0.07}$
VCONV	SGD	CIFAR-10	58.92 ± 0.62	58.83 ± 0.27	63.21 ± 0.07	63.06 ± 0.73	67.02 ± 0.43
VCONV	Adam	CIFAR-10	64.86 ± 1.18	64.68 ± 0.84	$\mathbf{69.45 \pm 0.76}$	$\mathbf{71.4 \pm 0.47}$	$\mathbf{72.43 \pm 0.54}$
GCONV	SGD	CIFAR-10	67.36 ± 0.56	66.86 ± 0.44	$\mathbf{74.57 \pm 0.43}$	$\mathbf{78.52 \pm 0.39}$	$\mathbf{78.90 \pm 0.37}$
GCONV	Adam	CIFAR-10	67.09 ± 0.58	67.08 ± 0.27	$\mathbf{77.12 \pm 0.19}$	$\mathbf{79.68 \pm 0.32}$	$\mathbf{80.43 \pm 0.35}$

Table 2: Shows the generalisation performance of different NPFs learning settings. The values in the table are averaged over 5 runs. Here, FC is a fully connected network with $w = 100$ and $d = 5$. VCONV and GCONV denote Vanilla CNN and CNN with GAP respectively. Please check Appendix B for details on architecture of VCONV and GCONV, and the hyper-parameters.

the training, and the performance gap between FRNPF and ReLU is continuous. In the case of CIFAR-10, we trained a GCONV network (parameterised by $\bar{\Theta}$) for 60 epochs, and we obtained 6 different weights at various *stages* of the training process. Stage 1: $\bar{\Theta}_{10}$, stage 2: $\bar{\Theta}_{20}$, stage 3: $\bar{\Theta}_{30}$, stage 4: $\bar{\Theta}_{40}$, stage 5: $\bar{\Theta}_{50}$, stage 6: $\bar{\Theta}_{60}$. We copy these weights obtained at various stages of training to setup 6 different FLNPFs, i.e., FLNPF-1 to FLNPF-6. We observe that the performance of FLNPF-1 to FLNPF-6 increases monotonically, with FLNPF-1 performing 72% which is better than FRNPF (i.e., 67.08%), and FLNPF-6 performing as well as ReLU (see left most plot in Figure 3). The performance of CNTK of Arora et al. [2019] is 77.43%. Thus, through its various stages, the FLNPF starts from below 77.43% and surpasses to reach 79.68%, which implies performance gain of CNN is due to learning of NPFs.

3. **Dynamics of active sub-networks during training:** We considered "Binary"-MNIST data set with two classes namely digits 4 and 7, with the labels taking values in $\{-1, +1\}$ and squared loss. We trained a fully connected (FC) network ($w = 100$, $d = 5$). Let $\widehat{H}_{\Theta_t} = \frac{1}{trace(H_{\Theta_t})} H_{\Theta_t}$ be the normalised NPK matrix. For a subset size, $n' = 200$ (100 examples per class) we plot $\nu_t = y^\top (\widehat{H}_{\Theta_t})^{-1} y$, (where $y \in \{-1, 1\}^{200}$ is the labelling function), and observe that ν_t reduces as training proceeds (see middle plot in Figure 3). Note that, $\nu_t = \sum_{i=1}^{n'} (u_{i,t}^\top y)^2 (\hat{\rho}_{i,t})^{-1}$, where $u_{i,t} \in \mathbb{R}^{n'}$ are the orthonormal eigenvectors of $\widehat{H}_{\Theta_t}$ and $\hat{\rho}_{i,t}, i \in [n']$ are the corresponding eigenvalues. Since $\sum_{i=1}^{n'} \hat{\rho}_{i,t} = 1$, the only way ν_t reduces is when more and more energy gets concentrated on $\hat{\rho}_{i,t}$s for which $(u_{i,t}^\top y)^2$s are also high. Since $H_{\Theta_t} = \Sigma \odot \Lambda_{\Theta_t}$, only Λ_{Θ_t} is learnt during training.

4. **Decoupled learning** of NPFs also performed better than FRNPFs (see column DNPFL in Table 2). This demonstrates the fact that NPFs can also be learnt in 'stand alone' manner. In this case, the NTK is given by $K_{\Theta^{\text{DGN}}} = K^{\text{V}}_{\Theta^{\text{DGN}}} + K^{\text{F}}_{\Theta^{\text{DGN}}}$. For MNIST, we compared $K^{\text{V}}_{\Theta^{\text{DGN}}}$ and $K^{\text{F}}_{\Theta^{\text{DGN}}}$ (calculated using 100 examples in total with 10 examples per each of the 10 classes) using their trace and Frobenius norms, and we observe that $K^{\text{V}}_{\Theta^{\text{DGN}}}$ and $K^{\text{F}}_{\Theta^{\text{DGN}}}$ are in the same scale (see right plot in Figure 3), which is perhaps pointing to the fact that both $K^{\text{F}}_{\Theta^{\text{DGN}}}$ and $K^{\text{F}}_{\Theta^{\text{DGN}}}$ are equally important for obtaining good generalisation performance. In DNPFL, we can separately study the kernel $K^{\text{F}}_{\Theta^{\text{DGN}}}$ responsible for NPF learning, an interesting future research direction.

7 Related Work

Jacot et al. [2018] showed the NTK to be the central quantity in the study of generalisation properties of infinite width DNNs. Jacot et al. [2019] identify two regimes that occur at initialisation in fully connected DNNs as the width increases to infinity namely i) *freeze:* here, the (scaled) NTK converges to a constant and hence leads to slow training, and ii) *chaos:* here, the NTK converges to Kronecker delta and hence hurts generalisation. Jacot et al. [2019] also suggest that for good generalisation

it is important to operate the DNNs at the edge of the freeze and the chaos regimes. Arora et al. [2019] proposed pure kernel method based on the infinite width CNTK (NTK of convolutional neural network) and showed that it out performed state-of-the-art kernel methods by 10%. Arora et al. [2019] also noted a performance gain (about $5-6\%$) of the CNNs over the CNTK. However, it was also noted by Arora et al. [2019], Lee et al. [2019] that random NTFs obtained from finite width neural networks do not perform as well as their limiting infinite width counterparts. Arora et al. [2019], Cao and Gu [2019] provided generalisation bounds with the NTK norm. Du et al. [2018] use the NTK to show that over-parameterised DNNs trained by gradient descent achieve zero training error. Du and Hu [2019], Shamir [2019], Saxe et al. [2013] studied deep linear networks. Since deep linear networks are special cases of deep gated networks, Theorem 5.1 of our paper also provides an expression for the NTK at initialisation of deep linear networks. To see this, in the case of deep linear networks, all the gates are always 1 for all input examples, and Λ_Θ will be a matrix whose entries will be $w^{(d-1)}$.

The results in our paper are complementary to the prior NTF/NTK based works, in that, the NPK and NPFs are zeroth order kernel and features respectively. In contrast, the NTF is the gradient of the network output with respect to the weights of the network and hence the NTF/NTK are essentially first order quantities. The fixed NPF regime is different from the NTK regime and the freeze/chaos regimes studied in prior works, in that, in the fixed NPF setting the gates are controlled by a separate feature network.

Gated linearity was studied recently by Fiat et al. [2019], where single layered gated networks were considered. In terms of the work in our paper, Fiat et al. [2019] consider the fixed NPF setting with random NPFs of a single layer network. In contrast to the work by Fiat et al. [2019], in this paper we considered DGN of depth d, and we also showed (using the DNPFL setting) that by gradient descent on the parameters of the feature and the value network we can learn the NPFs leading to better generalisation than learning with the fixed random NPFs. We believe that handling of depth d networks, identification and the use of novel quantities namely NPFs, NPK and, the role of NPF learning in generalisation amount to significant progress in comparison to Fiat et al. [2019].

The role of gates was also empirically studied by Srivastava et al. [2014], where the active sub-networks are called as *locally competetive* networks. They encode the active subnetwork information in a sub-mask which is bit string that encodes the $0/1$ state of the all the gates. The sub-masks were then visualised using t-SNE. The visualisation showed that the "subnetworks active for examples of the same class are much more similar to each other compared to the ones activated for the examples of different classes". Balestriero et al. [2018] show the connection between $\max$-affine linearity and DNN with ReLU activations. Neyshabur et al. [2015] used the notion of paths to define a *path-norm* based gradient descent procedure.

8 Conclusion

In this paper, we studied the role of active sub-networks in deep learning by encoding the gates in the neural path features. We showed that the neural path features are learnt during training and such learning is key for generalisation. In our experiments, we observed that almost all information of a trained DNN is stored in the neural path features. We conclude by saying that *understanding deep learning requires understanding neural path feature learning*.

References

[1] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

[2] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[3] Randall Balestriero et al. A spline theory of deep learning. In *International Conference on Machine Learning*, pages 374–383, 2018.

[4] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

[5] Simon S Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. *arXiv preprint arXiv:1901.08572*, 2019.

[6] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[7] Jonathan Fiat, Eran Malach, and Shai Shalev-Shwartz. Decoupling gating from linearity. *CoRR*, abs/1906.05032, 2019. URL `http://arxiv.org/abs/1906.05032`.

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[9] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[10] Arthur Jacot, Franck Gabriel, and Clément Hongler. Freeze and chaos for dnns: an ntk view of batch normalization, checkerboard and boundary effects. *arXiv preprint arXiv:1907.05715*, 2019.

[11] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8570–8581, 2019.

[12] Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

[13] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? *arXiv preprint arXiv:1911.13299*, 2019.

[14] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[15] Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pages 2691–2713, 2019.

[16] Rupesh Kumar Srivastava, Jonathan Masci, Faustino Gomez, and Jürgen Schmidhuber. Understanding locally competitive networks. *arXiv preprint arXiv:1410.1165*, 2014.

Appendix

A Expression for $K^{(d)}$

The $K^{(d)}$ matrix is computed by the recursion in (2).

$$\tilde{K}^{(1)}(s,s') = \Sigma^{(1)}(s,s') = \Sigma(s,s'), M^{(l)}_{ss'} = \begin{bmatrix} \Sigma^{(l)}(s,s) & \Sigma^{(l)}(s,s') \\ \Sigma^{(l)}(s',s) & \Sigma^{(l)}(s',s') \end{bmatrix} \in \mathbb{R}^2,$$

$$\Sigma^{(l+1)}(s,s') = 2 \cdot \mathbb{E}_{(q,q')\sim N(0,M^{(l)}_{ss'})}\left[\chi(q)\chi(q')\right], \hat{\Sigma}^{(l+1)}(s,s') = 2 \cdot \mathbb{E}_{(q,q')\sim N(0,M^{(l)}_{ss'})}\left[\partial\chi(q)\partial\chi(q')\right],$$

$$\tilde{K}^{(l+1)} = \tilde{K}^{(l)} \odot \hat{\Sigma}^{(l+1)} + \Sigma^{(l+1)}, K^{(d)} = \left(\tilde{K}^{(d)} + \Sigma^{(d)}\right)/2 \tag{2}$$

where $s, s' \in [n]$ are two input examples in the dataset, Σ is the data Gram matrix, $\partial\chi$ stands for the derivative of the activation function with respect to the pre-activation input, $N(0, M)$ stands for the mean-zero Gaussian distribution with co-variance matrix M.

B Experimental Setup

Dataset: We used standard datasets namely MNIST and CIFAR-10, with categorical cross entropy loss. We also used a 'Binary'-MNIST dataset, which is MNIST with only the two classes corresponding to digits 4 and 7, with label -1 for digit 4 and $+1$ for digit 7. For the 'Binary'-MNIST dataset, we used the squared loss.

Optimiser and Step-Size: We used stochastic gradient descent (SGD) and *Adam* as optimisers. In the case of SGD, we tried constant step-sizes in the set $\{0.1, 0.01, 0.001\}$ and chose the best. In the case of Adam the we used a constant step size of $3e^{-4}$. In both cases, we used batch size to be 32.

Network Architecture:

1. We used a fully connected (FC) DNN with $(w = 128, d = 5)$ for MNIST.

2. To train CIFAR-10, we used a *Vanilla* CNN architecture denoted by VCONV and a CNN architecture with *global-average-pooling* denoted by GCONV. VCONV is an architecture without pooling, residual connections, dropout or batch-normalisations, and is given by: input layer is $(32, 32, 3)$, followed by convolution layers with a stride of $(3, 3)$ and channels 64, 64, 128, 128 followed by a flattening to layer with 256 hidden units, followed by a fully connected layer with 256 units, and finally a 10 width soft-max layer to produce the final predictions. GCONV is same as VCONV with a *global-average-pooling* (GAP) layer at the boundary between the convolutional and fully connected layers.

Gating:

1. For both FRNPF, and FLNPF, we let $\chi^{\text{F}} = \chi_r$, and $G_{x,t}(l) = \gamma_r\left(q^{\text{F}}_{x,t}(l)\right)$.

2. In the case, DNPFL, we let $\chi^{\text{F}} = \chi_r$, and $G_{x,t}(l) = \gamma_{sr}\left(q^{\text{F}}_{x,t}(l)\right)$. Here $\gamma_{sr}(q) = \frac{1}{(1+\exp(-\beta\cdot q))}$ is a *soft-ReLU* gate which takes values in $(0, 1)$. In our experiments we used $\beta = 8$. The use of soft-ReLU makes it straightforward for the feature gradients to flow via the gating network.

Initialisation: In the case of FRNPF, we considered two possible initialisations namely i) *independent initialisation* (II), i.e., Θ^{F}_0 and Θ^{V}_0 are statistically independent, and ii) *dependent initialisation* (DI), i.e., $\Theta^{\text{F}}_0 = \Theta^{\text{V}}_0$, a case which mimics the NPFs and NPVs of a standard DNN with ReLU activations. In the case of FLNPF, $\Theta^{\text{F}}_0 = \bar{\Theta}$, where $\bar{\Theta}$ is the parameter of a pre-trained (at various stages of training) DNN with ReLU activations.

Epochs: All the models were trained close to 100% training accuracy. All the models took less than 100 epochs to train.

Reported Values: In order to obtain the values in Table 2, and in the left most plot of Figure 3 we used 5 runs. In each run, we took the best generalisation performance obtained in that run and then averaged the same over 5 runs.

C Applying Theorem 5.1 In Finite Width Case

In this section, we describe the technical step in applying Theorem 5.1 which requires $w \to \infty$ to measure the information in the gates of a DNN with finite width as per Definition 5.1. Since we are training only the value network in the FPNP mode of the DGN, it is possible to let the width of the value network alone go to ∞, while keeping the width of the feature network (which stores the fixed NPFs) finite. This is easily achieved by multiplying the width by a positive integer $m \in \mathbb{Z}_+$, and *padding* the gates 'm' times.

Definition C.1. *Define DGN$^{(m)}$ to be the DGN whose feature network is of width w and depth d, and whose value network is a fully connected network of width mw and depth d. The $mw(d-1)$ gating values are obtained by 'padding' the $w(d-1)$gating values of the width 'w', depth 'd' feature network 'm' times (see Figure 4, Table 3).*

Remark: DGN$^{(m)}$ has a total of $P^{(m)} = (mw)^{(d-1)}d_{in}$ paths. Thus, the NPF and NPV are quantities in $\mathbb{R}^{P^{(m)}}$. In what follows, we denote the NPF matrix of DGN$^{(m)}$ by $\Phi^{(m)}_{\Theta^{\text{F}}_0} \in \mathbb{R}^{P^{(m)} \times n}$, and use $H^{(m)}_{\text{FNPF}} = (\Phi^{(m)}_{\Theta^{\text{F}}_0})^\top \Phi^{(m)}_{\Theta^{\text{F}}_0}$.

Before we proceed to state the version of Theorem 5.1 for DGN$^{(m)}$, we will look at an equivalent definition for Λ_Θ (see Definition 2.2).

Definition C.2. *For input examples $s, s' \in [n]$ define*

1. $\tau_\Theta(s, s', l) \stackrel{def}{=} \sum_{i=1}^{w} G_{x_s,\Theta}(l,i) G_{x_{s'},\Theta}(l,i)$ *be the number of activations that are "on" for both inputs $s, s' \in [n]$ in layer $l \in [d-1]$.*

2. $\Lambda_\Theta(s, s') \stackrel{def}{=} \Pi_{l=1}^{d-1} \tau_\Theta(s, s', l)$.

Corollary C.1 (Corollary to Theorem 5.1)**.** *Under Assumption 5.1 with σ replaced by $\sigma_{(m)} = \sigma/\sqrt{m}$, as $m \to \infty$, $K_{\Theta_0^{DGN^{(m)}}} \to K^{(d)}_{FNPF} = d \cdot \sigma^{2(d-1)}_{(m)} H^{(m)}_{FNPF} = d \cdot \sigma^{2(d-1)} H_{FNPF}$.*

Proof. Let $\Lambda^{(m)}_{\text{FNPF}}$ and $\tau^{(m)}_{\text{FNPF}}$ be quantities associated with DGN$^{(m)}$. We know that $H^{(m)}_{\text{FNFP}} = \Sigma \odot \Lambda^{(m)}_{\text{FNPF}}$. Dropping the subscript FNPF to avoid notational clutter, we have

$$\begin{aligned}
\left(\sigma/\sqrt{m}\right)^{2(d-1)} \Lambda^{(m)}(s,s') &= \sigma^{2(d-1)} \frac{1}{m^{(d-1)}} \Pi_{l=1}^{d-1} \tau^{(m)}(s,s',l) \\
&= \sigma^{2(d-1)} \frac{1}{m^{(d-1)}} \Pi_{l=1}^{d-1} \left(m\tau(s,s',l)\right) \\
&= \sigma^{2(d-1)} \frac{1}{m^{(d-1)}} m^{(d-1)} \Pi_{l=1}^{d-1} \tau(s,s',l) \\
&= \sigma^{2(d-1)} \Pi_{l=1}^{d-1} \tau(s,s',l) \\
&= \sigma^{2(d-1)} \Lambda(s,s')
\end{aligned}$$

□

D Proofs of technical results

Proof of Proposition 1.1

Layer	Feature Network (NPF)	Value Network (NPV)
Input	$z^{\text{F}}_{x,t}(0) = x$	$z^{\text{V}}_{x,t}(0) = x$
Activation	$q^{\text{F}}_{x,t}(l) = \Theta^{\text{F}}_t(l)^\top z^{\text{F}}_{x,t}(l-1)$	$q^{\text{V}}_{x,t}(l) = \Theta^{\text{V}}_t(l)^\top z^{\text{V}}_{x,t}(l-1)$
Hidden	$z^{\text{F}}_{x,t}(l) = \chi^{\text{F}}\left(q^{\text{F}}_{x,t}(l)\right)$	$z^{\text{V}}_{x,t}(l) = q^{\text{V}}_{x,t}(l) \odot G_{x,t}(l)$
Output	None	$\hat{y}_t(x) = \Theta^{\text{V}}(d)^\top z^{\text{V}}_{x,t}(d-1)$
Gating Values: $G_{x,t}(l) = \gamma_r\left(q^{\text{F}}_{x,t}(l)\right)$ or $G_{x,t}(l) = \gamma_{sr}\left(q^{\text{F}}_{x,t}(l)\right)$		

Table 3: Deep Gated Network with padding. Here the gating values are padded, i.e., $G_{x,t}(l, kw+i) = G_{x,t}(l,i), \forall k = 0, 1, \ldots, m-1, i \in [w]$.

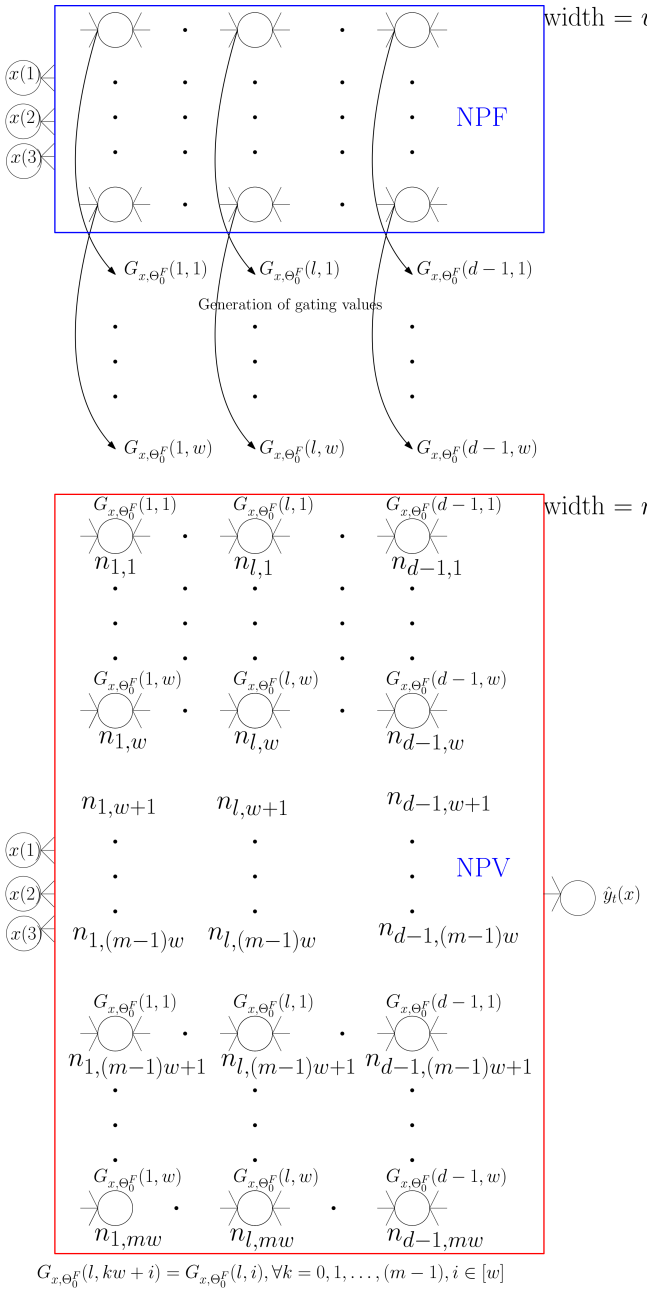

Figure 4: DGN$^{(m)}$ where the value network is of width mw and depth d. The gates are derived by padding the gating values obtained from the feature network 'm' times, i.e., $G_{x,t}(l, kw+i) = G_{x,t}(l,i), \forall k = 0,1,\ldots,m-1, i \in [w]$.

Proof. We know that $e_t = (e_t(s), s \in [n]) \in \mathbb{R}^n$, and $e_t(s) = \hat{y}_{\Theta_t}(x_s) - y(s)$. Now

$$\begin{aligned} L_{\Theta_t} &= \frac{1}{2}\sum_{s'=1}^{n}(\hat{y}_{\Theta_t} - y)^2 \\ &= \frac{1}{2}\sum_{s'=1}^{n} e_t^2 \\ \nabla_\Theta L_{\Theta_t} &= \sum_{s'=1}^{n} \nabla_\Theta \hat{y}_{\Theta_t}(x_{s'}) e_t(s') \\ \nabla_\Theta L_{\Theta_t} &= \sum_{s'=1}^{n} \psi_{x_{s'},\Theta_t} e_t(s') \end{aligned} \tag{3}$$

For gradient descent, $\dot{\Theta}_t = -\nabla_\Theta L_{\Theta_t}$, from (3) it follows that

$$\dot{\Theta}_t = -\sum_{s'=1}^{n} \psi_{x_{s'},\Theta_t} e_t(s') \tag{4}$$

Now $\dot{e}_t = \dot{\hat{y}}_{\Theta_t}$, and expanding $\dot{\hat{y}}_{\Theta_t}(x_s)$ for some $s \in [n]$, we have:

$$\begin{aligned}
\dot{\hat{y}}_{\Theta_t}(x_s) &= \frac{d\hat{y}_{\Theta_t}(x_s)}{dt} \\
&= \sum_{\theta\in\Theta} \frac{d\hat{y}_{\Theta_t}(x_s)}{d\theta}\frac{d\theta_t}{dt}, \text{ by expressing this summation as a dot product we obtain} \\
\dot{\hat{y}}_{\Theta_t}(x_s) &= \langle \psi_{x_s,\Theta_t}, \dot{\Theta}_t \rangle
\end{aligned} \tag{5}$$

We now use that fact that Θ_t is updated by gradient descent

$$\begin{aligned}
\dot{\hat{y}}_{\Theta_t}(x_s) &= -\langle \psi_{x_s,\Theta_t}, \sum_{s'=1}^{n} \psi_{x_{s'},\Theta_t} e_t(s') \rangle \\
&= -\sum_{s'=1}^{n} K_{\Theta_t}(s,s') e_t(s')
\end{aligned} \tag{6}$$

The proof is complete by recalling that $\hat{y}_{\Theta_t} = (\hat{y}_{\Theta_t}(x_s), s \in [n])$, and $\dot{e}_t = \dot{\hat{y}}_{\Theta_t}$. □

Proof of Proposition 2.1

Proof. Let $x \in \mathbb{R}^{d_{in}}$ be the input to the DNN and $\hat{y}_\Theta(x)$ be its output. The output can be written in terms of the final hidden layer output

$$\begin{aligned}
\hat{y}_\Theta(x) &= \Theta(d)^\top z_{x,\Theta}(d-1) \\
&= \sum_{j_{d-1}=1}^{w} \Theta(d, j_{d-1}, 1) z_{x,\Theta}(d-1, j_{d-1}) \\
&= \sum_{j_{d-1}=1}^{w} \Theta(d, j_{d-1}, 1) G_{x\Theta}(d-1, j_{d-1}) q_{x,\Theta}(d-1, j_{d-1})
\end{aligned} \tag{7}$$

Now $q_{x,\Theta}(d-1, j_{d-1})$ for a fixed j_{d-1} can again be expanded as

$$\begin{aligned}
q_{x,\Theta}(d-1, j_{d-1}) &= \sum_{j_{d-2}=1}^{w} \Theta(d, j_{d-2}, j_{d-1}) z_{x,\Theta}(d-2, j_{d-2}) \\
&= \sum_{j_{d-2}=1}^{w} \Theta(d-1, j_{d-2}, j_{d-1}) G_{x,\Theta}(d-2, j_{d-2}) q_{x,\Theta}(d-2, j_{d-2})
\end{aligned} \tag{8}$$

Now plugging in (8) in the expression in (7), we have

$$\begin{aligned}
\hat{y}_\Theta(x) &= \sum_{j_{d-1}=1}^{w} \Theta(d, j_{d-1}, 1) G_{x\Theta}(d-1, j_{d-1}) \left(\sum_{j_{d-2}=1}^{w} \Theta(d-1, j_{d-2}, j_{d-1}) G_{x,\Theta}(d-2, j_{d-2}) q_{x,\Theta}(d-2, j_{d-2}) \right) \\
&= \sum_{j_{d-1}, j_{d-2} \in [w]} G_{x,\Theta}(d-1, j_{d-1}) G_{x,\Theta}(d-2, j_{d-2}) \Theta(d, j_{d-1}, 1) \Theta(d-1, j_{d-2}, j_{d-1}) q_{x,\Theta}(d-2, j_{d-2})
\end{aligned} \tag{9}$$

By expanding q's for all the previous layers till the input layer we have

$$\sum_{j_d=1, j_{d-1}, \ldots, j_1 \in [w], j \in [d_{in}]} x(j) \Pi_{l=1}^{d-1} G_{x,\Theta}(l, j_l) \Pi_{l=1}^{d} \Theta(l, j_{l-1}, j_l)$$

□

Proof of Lemma 2.1

Proof.

$$\begin{aligned}\langle \phi_{x_s,\Theta}, \phi_{x_{s'},\Theta}\rangle &= \sum_{p\in[P]} x_s(\mathcal{I}_0(p))x_{s'}(\mathcal{I}_0(p))A_\Theta(x_s,p)A_\Theta(x_{s'},p)\\ &= \sum_{i=1}^{d_{in}} x_s(i)x_{s'}(i)\Lambda_\Theta(s,s')\\ &= \langle x_s, x_{s'}\rangle \cdot \Lambda_\Theta(s,s') \qquad (10)\end{aligned}$$

□

Proof of Proposition 3.1

Proof. Let $\Psi_\Theta = (\psi_{x_s,\Theta}, s\in[n]) \in \mathbb{R}^{d_{net}\times n}$ be the NTF matrix, then the NTK matrix is given by $K_{\Theta_t} = \Psi_{\Theta_t}^\top \Psi_{\Theta_t}$. Note that, $\hat{y}_\Theta(x_s) = \langle \phi_{x_s,\Theta}, v_\Theta\rangle = \langle v_\Theta, \phi_{x_s,\Theta}\rangle = v_\Theta^\top \phi_{x_s,\Theta}$. Now $\psi_{x_s,\Theta} = \nabla_\Theta v_\Theta \phi_{x_s,\Theta}$, and hence $\Psi = \nabla_\Theta v_\Theta \Phi_\Theta$. Hence, $K_{\Theta_t} = \Psi_{\Theta_t}^\top \Psi_{\Theta_t} = \Phi_\Theta^\top (\nabla_\Theta v_\Theta)^\top (\nabla_\Theta v_\Theta)\Phi_\Theta = \Phi_\Theta^\top \mathcal{V}_\Theta \Phi_\Theta$. □

Proof of Proposition 3.2

Proof. Follows in a similar manner as the proof of Proposition 1.1. □

Proof of Proposition 3.3

Proof. $\rho_{\min}(K_\Theta) = \min_{\substack{x\in\mathbb{R}^n\\ \|x\|_2=1}} x^\top K_\Theta x$. Let $x'\in\mathbb{R}^n$ such that $\|x'\|_2 = 1$ and $\rho_{\min}(K_\Theta) = {x'}^\top K_\Theta x'$. Now, let $y' = \Phi x'$. Then we have, $\rho_{\min}(K_\Theta) = {y'}^\top \mathcal{V}_\Theta y'$. Hence $\rho_{\min}(K_\Theta) \le \|y'\|_2^2 \rho_{\max}(\mathcal{V}_\Theta)$. Now, $\|y'\|_2^2 = {x'}^\top \Phi_\Theta^\top \Phi_\Theta x' \le \rho_{min}(H_\Theta)$. □

Proof of Proposition 4.1

Proof. Follows in a similar manner as proof of Proposition 1.1. □

Lemma D.1. *Let $\varphi_{p,\Theta}$ be as in Definition 3.1, under Assumption 5.1, for paths $p, p_1, p_2 \in \mathcal{P}, p_1 \neq p_2$, at initialisation we have (i) $\mathbb{E}\left[\langle \varphi_{p_1,\Theta_0^{\mathrm{v}}}, \varphi_{p_2,\Theta_0^{\mathrm{v}}}\rangle\right] = 0$, (ii) $\langle \varphi_{p,\Theta_0^{\mathrm{v}}}, \varphi_{p,\Theta_0^{\mathrm{v}}}\rangle = d\sigma^{2(d-1)}$.*

Proof.

$$\langle \varphi_{p_1,\Theta_0^{\mathrm{v}}}, \varphi_{p_2,\Theta_0^{\mathrm{v}}}\rangle = \sum_{\theta^{\mathrm{v}}\in\Theta^{\mathrm{v}}} \partial_{\theta^{\mathrm{v}}} v_{\Theta_0^{\mathrm{v}}}(p_1)\partial_{\theta^{\mathrm{v}}} v_{\Theta_0^{\mathrm{v}}}(p_2)$$

Let $p \rightsquigarrow (\cdot)$ denote the fact that path p passes through $(\cdot)$, and let $p \not\rightsquigarrow (\cdot)$ denote the fact that path p does not pass through $\not\rightsquigarrow$. Let $\theta^{\mathrm{V}} \in \Theta^{\mathrm{V}}$ be any weight such that $p \rightsquigarrow \theta^{\mathrm{V}}$, and w.l.o.g let θ^{V} belong to layer $l' \in [d]$. If either $p_1 \not\rightsquigarrow \theta^{\mathrm{V}}$ or $p_2 \not\rightsquigarrow \theta^{\mathrm{V}}$, then it follows that $\partial_{\theta^{\mathrm{v}}} v_{\Theta_0^{\mathrm{v}}}(p_1)\partial_{\theta^{\mathrm{v}}} v_{\Theta_0^{\mathrm{v}}}(p_2) = 0$. In the case when $p_1, p_2 \rightsquigarrow \theta^{\mathrm{V}}$, we have

$$\begin{aligned}&\mathbb{E}\left[\partial_{\theta^{\mathrm{v}}} v_{\Theta_0^{\mathrm{v}}}(p_1)\partial_{\theta^{\mathrm{v}}} v_{\Theta_0^{\mathrm{v}}}(p_2)\right]\\ &= \mathbb{E}\left[\prod_{\substack{l=1\\ l\neq l'}}^{d}\left(\Theta_0^{\mathrm{V}}(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1))\Theta_0^{\mathrm{V}}(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))\right)\right]\\ &= \prod_{\substack{l=1\\ l\neq l'}}^{d}\mathbb{E}\left[\Theta_0^{\mathrm{V}}(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1))\Theta_0^{\mathrm{V}}(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))\right]\end{aligned}$$

where the $\mathbb{E}[\cdot]$ moved inside the product because at initialisation the weights (of different layers) are independent of each other. Since $p_1 \neq p_2$, in one of the layers $\tilde{l} \in [d-1], \tilde{l} \neq l'$ they do not pass through the same weight, i.e., $\Theta_0^{\mathrm{V}}(\tilde{l}, \mathcal{I}_{\tilde{l}-1}(p_1), \mathcal{I}_{\tilde{l}}(p_1))$ and $\Theta_0^{\mathrm{V}}(\tilde{l}, \mathcal{I}_{\tilde{l}-1}(p_2), \mathcal{I}_{\tilde{l}}(p_2))$ are distinct weights. Using this fact

$$
\begin{aligned}
&\mathbb{E}\left[\partial_{\theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}(p_1) \partial_{\theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}(p_2)\right] \\
&= \prod_{\substack{l=1 \\ l \neq l', \tilde{l}}}^{d} \mathbb{E}\left[\Theta_0^{\mathrm{V}}(l, \mathcal{I}_{l-1}(p_1), \mathcal{I}_l(p_1)) \Theta_0^{\mathrm{V}}(l, \mathcal{I}_{l-1}(p_2), \mathcal{I}_l(p_2))\right] \\
&= \mathbb{E}\left[\Theta_0^{\mathrm{V}}(\tilde{l}, \mathcal{I}_{\tilde{l}-1}(p_1), \mathcal{I}_{\tilde{l}}(p_1))\right] \mathbb{E}\left[\Theta_0^{\mathrm{V}}(\tilde{l}, \mathcal{I}_{\tilde{l}-1}(p_2), \mathcal{I}_{\tilde{l}}(p_2))\right] \\
&= 0
\end{aligned}
$$

The proof of (ii) is complete by noting that $\sum_{\theta^{\mathrm{V}} \in \Theta^{\mathrm{V}}} \partial_{\theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}(p) \partial_{\theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}(p)$ has d non-zero terms for a single path p and at initialisation we have

$$
\begin{aligned}
&\partial_{\theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}(p) \partial_{\theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}(p) \\
&= \prod_{\substack{l=1 \\ l \neq l'}}^{d} {\Theta_0^{\mathrm{V}}}^2(l, \mathcal{I}_{l-1}(p), \mathcal{I}_l(p)) \\
&= \sigma^{2(d-1)}
\end{aligned}
$$

□

Detailed version of Theorem 5.1 with proof.

Theorem D.1. *Under Assumption 5.1, and* $\frac{4d}{w^2} < 1$ *it follows that*

$$
\begin{aligned}
\mathbb{E}\left[K_{\Theta_0^{DGN}}\right] &= d \cdot \sigma^{2(d-1)} H_{FNPF} \\
Var\left[K_{\Theta_0^{DGN}}(s, s')\right] &\leq O\left(d_{in}^2 \sigma^{4(d-1)} \max\{d^2 w^{2(d-2)+1}, d^3 w^{2(d-2)}\}\right)
\end{aligned}
$$

Proof. We have

$$
\begin{aligned}
\mathbb{E}\left[K_{\Theta_0^{\mathrm{DGN}}}\right] &= \mathbb{E}\left[\Phi_{\mathrm{FNPF}}^{\top} \mathcal{V}_{\Theta_0^{\mathrm{V}}} \Phi_{\mathrm{FNPF}}\right] \\
&= \mathbb{E}\left[\Phi_{\mathrm{FNPF}}^{\top} (\nabla_{\Theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}})^{\top} (\nabla_{\Theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}}) \Phi_{\mathrm{FNPF}}\right] \\
&= \Phi_{\mathrm{FNPF}}^{\top} \mathbb{E}\left[(\nabla_{\Theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}})^{\top} (\nabla_{\Theta^{\mathrm{V}}} v_{\Theta_0^{\mathrm{V}}})\right] \Phi_{\mathrm{FNPF}} \\
&\overset{(a)}{=} d \cdot \sigma^{2(d-1)} \Phi_{\mathrm{FNPF}}^{\top} \Phi_{\mathrm{FNPF}} \\
&= d \cdot \sigma^{2(d-1)} H_{\mathrm{FNPF}}
\end{aligned}
$$

where, (a) follows from Lemma D.1.

We now turn to the variance calculation. The idea is that we expand $Var[K_0(s, s')] = \mathbb{E}\left[K_0(s, s')^2\right] - \mathbb{E}[K_0(s, s')]^2$ and identify the terms which cancel due to subtraction and then bound the rest of the terms.

Notation: In what follows, we let K_0 to denote $K_{\Theta_0^{\mathrm{DGN}}}$ and drop superscript V from Θ_0^{V}, and subscript Θ_0^{V} from $v_{\Theta_0^{\mathrm{V}}}$. Further, we assume that the weights can be enumerated as $\theta(1), \ldots, \theta(d_{net})$. We also denote $p \rightsquigarrow (\cdot)$ to denote the fact that path p passes through $(\cdot)$ and $p \not\rightsquigarrow (\cdot)$ to denote the fact that path p does not pass through $(\cdot)$. We use a shortcut notation $A(s, p)$ instead of $A(x_s, p)$. In what follows, we let $x \in \mathbb{R}^{d_{in} \times n}$ to be the data matrix.

Let $\theta(m), m \in [d_{net}]$ belong to layer $l'(m)$, then

$$
\begin{aligned}
&\mathbb{E}\left[K_0(s,s')\right] \\
&= \sum_{m=1}^{d_{net}} \mathbb{E}\left[\left(\sum_{p_1\in[P]} x(\mathcal{I}_0(p_1),s)A_0(s,p_1)\frac{\partial v_0(p_1)}{\partial\theta(m)}\right)\left(\sum_{p_2\in[P]} x(\mathcal{I}_0(p_2),s)A_0(s',p_2)\frac{\partial v_0(p_2)}{\partial\theta(m)}\right)\right] \\
&= \sum_{m=1}^{d_{net}} \mathbb{E}\left[\sum_{p_1,p_2\in[P]} x(\mathcal{I}_0(p_1),s)A_0(s,p_1)\frac{\partial v_0(p_1)}{\partial\theta(m)}x(\mathcal{I}_0(p_2),s')A_0(s',p_2)\frac{\partial v_0(p_2)}{\partial\theta(m)}\right] \\
&\overset{(a)}{=} \sum_{m=1}^{d_{net}} \sum_{\substack{p_1,p_2\in[P]\\ p_1,p_2\rightsquigarrow\theta(m)}} x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)\mathbb{E}\Bigg[\prod_{\substack{l=1\\ l\neq l'(m)}}^{d-1}\Theta_0(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1)) \\
&\Theta_0(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))\Bigg] \\
&\overset{(b)}{=} \sum_{m=1}^{d_{net}} \sum_{\substack{p_1,p_2\in[P]\\ p_1,p_2\rightsquigarrow\theta(m)}} x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)\prod_{\substack{l=1\\ l\neq l'(m)}}^{d-1}\mathbb{E}\Bigg[\Theta_0(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1)) \\
&\Theta_0(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))\Bigg] \qquad (11)
\end{aligned}
$$

where (a) follows from the fact that for $p \not\rightsquigarrow \theta(m)$, $\frac{\partial v_0(p)}{\partial\theta(m)} = 0$, and (b) follows from the fact that at initialisation the layer weights are independent of each other. Note that the right hand side of (11) only terms with $p_1 = p_2$ will survive the expectation.

In the following expression in (12), note that only terms of the form $p_1 = p_2$ and $p_3 = p_4$ are non-zero.

$$
\begin{aligned}
&\mathbb{E}\left[K_0(s,s')\right]^2 = \\
&\Bigg(\sum_{m=1}^{d_{net}} \sum_{\substack{p_1,p_2\in[P]\\ p_1,p_2\rightsquigarrow\theta(m)}} x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)\prod_{\substack{l=1\\ l\neq l'(m)}}^{d-1}\mathbb{E}\Big[\Theta_0(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1)) \\
&\Theta_0(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))\Big]\Bigg)\times \\
&\Bigg(\sum_{m'=1}^{d_{net}} \sum_{\substack{p_3,p_4\in[P]\\ p_3,p_4\rightsquigarrow\theta(m')}} x(\mathcal{I}_0(p_3),s)A_0(s,p_3)x(\mathcal{I}_0(p_4),s')A_0(s',p_4)\prod_{\substack{l=1\\ l\neq l'(m')}}^{d-1}\mathbb{E}\Big[\Theta_0(l,\mathcal{I}_{l-1}(p_3),\mathcal{I}_l(p_3)) \\
&\Theta_0(l,\mathcal{I}_{l-1}(p_4),\mathcal{I}_l(p_4))\Big]\Bigg)
\end{aligned}
$$

$$\mathbb{E}\left[K_0(s,s')\right]^2 =$$
$$\sum_{m,m'=1}^{d_{net}} \sum_{\substack{p_1,p_2,p_3,p_4\in[P] \\ p_1,p_2\rightsquigarrow\theta(m) \\ p_3,p_4\rightsquigarrow\theta(m')}} \Bigg[\Bigg(x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)x(\mathcal{I}_0(p_3),s)$$
$$A_0(s,p_3)x(\mathcal{I}_0(p_4),s')A_0(s',p_4)\Bigg) \times \Bigg(\prod_{\substack{l=1 \\ l\neq l'(m') \\ l\neq l'(m)}}^{d-1} \mathbb{E}\left[\Theta_0(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1))\Theta_0(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))\right]$$
$$\mathbb{E}\left[\Theta_0(l,\mathcal{I}_{l-1}(p_3),\mathcal{I}_l(p_3))\Theta_0(l,\mathcal{I}_{l-1}(p_4),\mathcal{I}_l(p_4))\right]\Bigg)\times$$
$$\Bigg(\mathbb{E}\left[\Theta_0(l,\mathcal{I}_{l'(m')-1}(p_1),\mathcal{I}_{l'(m')}(p_1))\Theta_0(l,\mathcal{I}_{l'(m')-1}(p_2),\mathcal{I}_{l'(m')}(p_2))\right]\Bigg)\times$$
$$\Bigg(\mathbb{E}\left[\Theta_0(l,\mathcal{I}_{l'(m)-1}(p_3),\mathcal{I}_{l'(m)}(p_3))\Theta_0(l,\mathcal{I}_{l'(m)-1}(p_4),\mathcal{I}_{l'(m)}(p_4))\right]\Bigg)\Bigg] \tag{12}$$

In the expression in (13), paths p_1, p_2, p_3, p_4 do not have constraints, and can be distinct.

$$\mathbb{E}\left[K_0^2(s,s')\right] =$$
$$\sum_{m,m'=1}^{d_{net}} \sum_{\substack{p_1,p_2,p_3,p_4\in[P] \\ p_1,p_2\rightsquigarrow\theta(m) \\ p_3,p_4\rightsquigarrow\theta(m')}} \Bigg[\Bigg(x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)x(\mathcal{I}_0(p_3),s)$$
$$A_0(s,p_3)x(\mathcal{I}_0(p_4),s')A_0(s',p_4)\Bigg) \times \Bigg(\prod_{\substack{l=1 \\ l\neq l'(m') \\ l\neq l'(m)}}^{d-1} \mathbb{E}[\Theta_0(l,\mathcal{I}_{l-1}(p_1),\mathcal{I}_l(p_1))\Theta_0(l,\mathcal{I}_{l-1}(p_2),\mathcal{I}_l(p_2))$$
$$\Theta_0(l,\mathcal{I}_{l-1}(p_3),\mathcal{I}_l(p_3))\Theta_0(l,\mathcal{I}_{l-1}(p_4),\mathcal{I}_l(p_4))]\Bigg)\times$$
$$\Bigg(\mathbb{E}\left[\Theta_0(l,\mathcal{I}_{l'(m')-1}(p_1),\mathcal{I}_{l'(m')}(p_1))\Theta_0(l,\mathcal{I}_{l'(m')-1}(p_2),\mathcal{I}_{l'(m')}(p_2))\right]\Bigg)\times$$
$$\Bigg(\mathbb{E}\left[\Theta_0(l,\mathcal{I}_{l'(m)-1}(p_3),\mathcal{I}_{l'(m)}(p_3))\Theta_0(l,\mathcal{I}_{l'(m)-1}(p_4),\mathcal{I}_{l'(m)}(p_4))\right]\Bigg)\Bigg] \tag{13}$$

We now state the following facts/observations.

• *Fact 1:* Any term that survives the expectation (i.e., does not become 0) and participates in (13) is of the form $\sigma^{4(d-1)}\big(x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)x(\mathcal{I}_0(p_3),s)A_0(s,p_3)x(\mathcal{I}_0(p_4),s')A_0(s',p_4)\big)$, where p_1, p_2, p_3, p_4 are free variables. Any term that survives the expectation (i.e., does not become 0) and participates in participates in (12) is of the form $\sigma^{4(d-1)}\big(x(\mathcal{I}_0(p_1),s)A_0(s,p_1)x(\mathcal{I}_0(p_2),s')A_0(s',p_2)x(\mathcal{I}_0(p_3),s)A_0(s,p_3)x(\mathcal{I}_0(p_4),s')A_0(s',p_4)\big)$, where $p_1 = p_2, p_3 = p_4$.

• *Fact 2:* The number of paths through a particular weight $\theta(m)$ in one of the middle layers is $d_{in}w^{d-3}$. The number of paths through a particular weight $\theta(m)$ in the first layer is w^{d-2} . The number of paths through a particular weight $\theta(m)$ in the last layer is $d_{in}w^{d-2}$.

• *Fact 3:* Let $\mathcal{P}'$ be an arbitrary set of paths constrained to pass through some set of weights. Let $\mathcal{P}''$ be the set of paths obtained by adding an additional constraint that the paths also should pass through a particular weight say $\theta(m)$. Now, if $\theta(m)$ belongs to :

1. a middle layer, then $|\mathcal{P}''| = \frac{|\mathcal{P}'|}{w^2}$.

2. the first layer, then $|\mathcal{P}''| = \frac{|\mathcal{P}'|}{d_{in}w}$.

3. the last layer, then $|\mathcal{P}''| = \frac{|\mathcal{P}'|}{w}$.

• *Fact 4:* For any p_1, p_2, p_3, p_4 combination that survives the expectation in (13) can be written as

$$\Bigg(x(\mathcal{I}_0(p_1), s) A_0(s, p_1) x(\mathcal{I}_0(p_2), s') A_0(s', p_2) x(\mathcal{I}_0(p_3), s)$$

$$A_0(s, p_3) x(\mathcal{I}_0(p_4), s') A_0(s', p_4) \Bigg) \times$$

$$\Bigg(\prod_{\substack{l=1 \\ l \neq l'(m') \\ l \neq l'(m)}}^{d-1} \mathbb{E}[\Theta_0(l, \mathcal{I}_{l-1}(p_1), \mathcal{I}_l(p_1)) \Theta_0(l, \mathcal{I}_{l-1}(p_2), \mathcal{I}_l(p_2))$$

$$\Theta_0(l, \mathcal{I}_{l-1}(p_3), \mathcal{I}_l(p_3)) \Theta_0(l, \mathcal{I}_{l-1}(p_4), \mathcal{I}_l(p_4))] \Bigg) \times$$

$$\Bigg(\mathbb{E}\left[\Theta_0(l, \mathcal{I}_{l'(m')-1}(p_1), \mathcal{I}_{l'(m')}(p_1)) \Theta_0(l, \mathcal{I}_{l'(m')-1}(p_2), \mathcal{I}_{l'(m')}(p_2)) \right] \Bigg) \times$$

$$\Bigg(\mathbb{E}\left[\Theta_0(l, \mathcal{I}_{l'(m)-1}(p_3), \mathcal{I}_{l'(m)}(p_3)) \Theta_0(l, \mathcal{I}_{l'(m)-1}(p_4), \mathcal{I}_{l'(m)}(p_4)) \right] \Bigg)$$

where $\rho_a \rightsquigarrow \theta(m)$ and $\rho_b \rightsquigarrow \theta(m')$ are what we call as *base* (case) paths.

• *Fact 5:* For any given base paths ρ_a and ρ_b there could be multiple assignments possible for p_1, p_2, p_3, p_4.

• *Fact 6:* Terms in (13), wherein, the base case is generated as $p_1 = p_2 = \rho_a$ and $p_3 = p_4 = \rho_b$ (or $p_1 = p_2 = \rho_b$ and $p_3 = p_4 = \rho_a$), get cancelled with the corresponding terms in (12).

• *Fact 7:* When the bases paths ρ_a and ρ_b do not intersect (i.e., do not pass through the same weight in any one of the layers), the only possible assignment is $p_1 = p_2 = \rho_a$ and $p_3 = p_4 = \rho_b$ (or $p_1 = p_2 = \rho_b$ and $p_3 = p_4 = \rho_a$), and such terms are common in (13) and (12), and hence do not show up in the variance term.

• *Fact 7:* Let base paths ρ_a and ρ_b intersect/cross at layer $l_1, \ldots, l_k, k \in [d-1]$, and let $\rho_a = (\rho_a(1), \ldots, \rho_a(k+1))$ where $\rho_a(1)$ is a sub-path string from layer 1 to l_1, and $\rho_a(2)$ is the sub-path string from layer $l_1 + 1$ to l_2 and so on, and $\rho_a(k+1)$ is the sub-path string from layer $l_k + 1$ to the output node. Then the set of paths that can occur in $\mathbb{E}\left[K_0(s, s')^2\right]$ are of the form:

1. $p_1 = p_2 = \rho_a, p_3 = p_4 = \rho_b$ (or $p_1 = p_2 = \rho_b, p_3 = p_4 = \rho_a$) which get cancelled in the $\mathbb{E}\left[K_0(s, s')\right]^2$ term.
2. $p_1 = \rho_a$, $p_3 = \rho_b$, $p_2 = (\rho_b(1), \rho_a(2), \rho_a(3), \ldots, \rho_a(k+1))$, $p_4 = (\rho_a(1), \rho_b(2), \rho_b(3), \ldots, \rho_b(k+1))$, which are obtained by *splicing* the base paths in various combinations. Note that for such spliced paths $p_1 \neq p_2$ and $p_3 \neq p_4$ and hence do not occur in the expression for $\mathbb{E}\left[K_0(s, s')\right]^2$ in (12).

• *Fact 8:* For k crossings of the base paths there are 4^{k+1} splicings possible, and those many terms are extra in the $\mathbb{E}\left[K_0(s, s')^2\right]$ expression in (13), when compared to the $\mathbb{E}\left[K_0(s, s')\right]^2$ expression in (12).

Upper Bound: We now enumerate various possible crossings of the base paths, and calculate an upper bound for the magnitude of the contribution of 'spliced' terms to the variance term using the *Fact 1* to *Fact 8*. In short, we find an upper bound for the those terms that do not get cancelled in the variance calculation. Further, without loss of generality we drop $x(\mathcal{I}_0(p))$ and $A(\cdot, \cdot)$ terms in this upper calculation.

Case 1**:** $k = 1$ crossing, in either first or last layer. There are $d_{in} w$ weights in the first layer and w weights in the last layer. The number of base path combinations passing through the first layer is $w^{d-2} \times w^{d-2}$. The number of base path combinations passing through the last layer is

$(d_{in}w^{d-2}) \times (d_{in}w^{d-2})$. For each of these cases, m, m' could take $O(d^2)$ possible values. And the multiplication of the weights themselves contribute to $\sigma^{4(d-1)}$. Splicing of these base paths could be done in 4^2 ways. Putting them together we have

$$\begin{aligned}&\sigma^{4(d-1)} \times (w) \times (d_{in}^2 \times w^{d-2} \times w^{d-2}) \times d^2 \times 4^2 \\ &+ \sigma^{4(d-1)} \times (d_{in}w) \times (w^{d-2} \times w^{d-2}) \times d^2 \times 4^2 \\ &\leq 32 d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1}\end{aligned}$$

Case 2**:** $k = 1$ crossing, in one of the middle layers. There are $w^2(d-2)$ weights in the middle layers. The number of base path combinations that pass through a given weight in the middle layers is $(d_{in}w^{d-3}) \times (d_{in}w^{d-3})$. For each of these cases, m, m' could take $O(d^2)$ possible values. And the multiplication of the weights themselves contribute to $\sigma^{4(d-1)}$. Splicing of these base paths could be done in 4^2 ways. Putting them together we have

$$\sigma^{4(d-1)} \times w^2(d-2) \times (d_{in}^2 \times w^{d-3} \times w^{d-3}) \times d^2 \times 4^2 \leq 16 d_{in}^2 \sigma^{4(d-1)} d^3 w^{2(d-3)}$$

Case 3**:** $k = 2$ crossings, one in the first layer and other in the last layer. So, we have

$$\sigma^{4(d-1)}(d_{in}w \times w) \times (w^{(d-3)} \times w^{(d-3)}) d^2 \times 4^3 \leq (32 d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1}) \times (4w^{-1}),$$

Case 4**:** $k = 2$ crossings, first one in the first layer or the last layer, and the second one in the middle layer. This can be obtained by looking at the Case 1 and then adding the further restriction that the base paths should cross each other in the middle layer.

$$\begin{aligned}&32 d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1} \times w^2(d-2) \times w^{-2} \times w^{-2} \times 4 \\ &\leq (32 d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1}) \times (4dw^{-2})\end{aligned}$$

Case 5**:** $k = 2$ crossings, in the middle layer. This can be obtained by taking Case 2 and then adding the further restriction that the base paths should cross each other in the middle layer.

$$16 d_{in}^2 \sigma^{4(d-1)} d^3 w^{2(d-3)} \times w^2(d-2) \times w^{-2} \times w^{-2} \times 4 \leq (16 d_{in}^2 \sigma^{4(d-1)} d^3 w^{2(d-3)}) \times (4dw^{-2})$$

Case 6**:** $k = 3$ crossings, first one in the first layer or the last layer, and the other two in the middle layers. This can be obtained by considering Case 4 and then adding the further restriction that the base paths should cross each other in the middle layer.

$$(32 d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1}) \times (4dw^{-2}) \times (4dw^{-2})$$

Case 7**:** $k = 3$ crossings, first two in the first and last layers and the third one in the middle layers. This can be obtained by considering Case 3 and then adding the further restriction that the base paths should cross each other in the middle layer.

$$(32 d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1}) \times (4w^{-1}) \times (4dw^{-2})$$

Case 8**:** $k = 3$ crossings, in the middle layer. This can be obtained by considering Case 5 and then adding the further restriction that the base paths should cross each other in the middle layer.

$$(16 d_{in}^2 \sigma^{4(d-1)} d^3 w^{2(d-3)}) \times (4dw^{-2}) \times (4dw^{-2})$$

The cases can be extended in a similar way, increasing the number of crossings. Now, assuming $\frac{4d}{w^2} < 1$, the bounds in the various terms can be lumped together as below:

• We can add the bounds for Case 1, Case 4, Case 6 and other cases obtained by adding more crossings (one at a time) in the middle layer to Case 6. This gives rise to a term which is upper bounded by (for some constant $C > 0$):

$$C d_{in}^2 \sigma^{4(d-1)} d^2 w^{2(d-2)+1} \left(\frac{1}{1 - 4dw^{-2}} \right)$$

• We can add the bounds for Case 3, Case 7 and other cases obtained by adding more crossings (one at a time) in the middle layer to Case 6. This gives rise to a term which is upper bounded by

$$Cd_{in}^2\sigma^{4(d-1)}d^3w^{2(d-2)}\left(\frac{1}{1-4dw^{-2}}\right)$$

• We can add the bounds for Case 2, Case 5, Case 8 and other cases obtained by adding more crossings (one at a time) in the middle layer to Case 6. This gives rise to a term which is upper bounded by

$$Cd_{in}^2\sigma^{4(d-1)}d^2w^{2(d-2)}\left(\frac{1}{1-4dw^{-2}}\right)$$

Putting together we have the variance to be bounded by

$$Cd_{in}^2\sigma^{4(d-1)}\max\{d^2w^{2(d-2)+1}, d^3w^{2(d-2)}\},$$

for some constant $C > 0$. □

E DGN as a Lookup Table: Applying Theorem 5.1 to a pure memorisation task

In this section, we modify the DGN in Figure 2 into a memorisation network to solve a pure memorisation task. The objective of constructing the memorisation network is to understand the roles of depth and width in Theorem 5.1 in a simplified setting. In this setting, we show increasing depth till a point helps in training and increasing depth beyond it hurts training.

Definition E.1 (Memorisation Network/Task)**.** *Given a set of values $(y_s)_{s=1}^n \in \mathbb{R}$, a memorisation network (with weights $\Theta \in \mathbb{R}^{d_{net}}$) accepts $s \in [n]$ as its input and produces $\hat{y}_\Theta(s) \approx y_s$ as its output. The loss of the memorisation network is defined as $L_\Theta = \frac{1}{2}\sum_{s=1}^n(\hat{y}_\Theta(s) - y_s)^2$.*

Layer	Memorisation Network
Input	$z_t(0) = 1$
Activation	$q_{s,t}(l) = \Theta_t(l)^\top z_{s,t}(l-1)$
Hidden	$z_{s,t}(l) = q_{s,t}(l) \odot G_{s,t}(l)$
Output	$\hat{y}_t(s) = \Theta(d)^\top z_{s,t}(d-1)$

Table 4: Memorisation Network. The input is fixed and is equal to 1. All the internal variables depend on the index s and the parameter Θ_t. The gating values Gs are external and independent variables.

Fixed Random Gating: The memorisation network is described in Table 4. In a memorisation network, the gates are *fixed and random*, i.e., for each index $s \in [n]$, the gating values $G_{s,0}(l,i), \forall l \in [d-1], i \in [w]$ are sampled from $Ber(\mu), \mu \in (0,1)$ taking values in $\{0,1\}$, and kept fixed throughout training, i.e., $G_{s,t}(\cdot,\cdot) = G_{s,0}(\cdot,\cdot)\,\forall\, t \geq 0$. The input to the memorisation network is fixed as 1, and since the gating is fixed and random there is a separate random sub-network to memorise each target $y_s \in \mathbb{R}$. The memorisation network can be used to memorise the targets $(y_s)_{s=1}^n$ by training it using gradient descent by minimising the squared loss L_Θ. In what follows, we let K_0 and H_0 to be the NTK and NPK of the memorisation network at initialisation.

Performance of Memorisation Network: From Proposition 1.1 we know that as $w \to \infty$, the training error dynamics of the memorisation network follows:

$$\dot{e}_t = -K_0 e_t, \tag{14}$$

i.e., the spectral properties of K_0 (or H_0) dictates the rate of convergence of the training error to 0. In the case of the memorisation network with fixed and random gates, we can calculate $\mathbb{E}[K_0]$ explicitly.

Spectrum of H_0: The input Gram matrix Σ is a $n \times n$ matrix with all entries equal to 1 and its rank is equal to 1, and hence $H_0 = \Lambda_0$. We can now calculate the properties of Λ_0. It is easy to check that

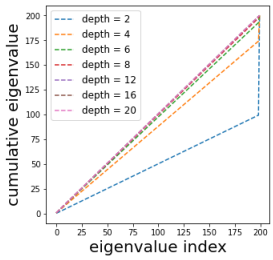

Figure 5: Ideal spectrum of $\mathbb{E}\left[K_0\right]/d$ for a memorisation network for $n = 200$.

$\mathbb{E}_\mu\left[\Lambda_0(s,s)\right] = (\mu w)^{(d-1)}, \forall s \in [n]$ and $\mathbb{E}_\mu\left[\Lambda_0(s,s')\right] = (\mu^2 w)^{(d-1)}, \forall s, s' \in [n]$. For $\sigma = \sqrt{\frac{1}{\mu w}}$, and $\mathbb{E}_\mu\left[K_0(s,s)/d\right] = 1$, and $\mathbb{E}_\mu\left[K_0(s,s')/d\right] = \mu^{(d-1)}$.

Why increasing depth till a point helps ? We have:

$$\frac{\mathbb{E}\left[K_0\right]}{d} = \begin{bmatrix} 1 & \mu^{d-1} & \dots & \mu^{d-1} & \dots \\ \dots & 1 & \dots & \mu^{d-1} & \dots \\ \dots & \mu^{d-1} & \dots & 1 & \dots \\ \dots & \mu^{d-1} & \dots & \mu^{d-1} & 1 \end{bmatrix} \tag{15}$$

i.e., all the diagonal entries are 1 and non-diagonal entries are μ^{d-1}. Now, let $\rho_i \geq 0, i \in [n]$ be the eigenvalues of $\frac{\mathbb{E}[K_0]}{d}$, and let $\rho_{\max}$ and $\rho_{\min}$ be the largest and smallest eigenvalues. One can easily show that $\rho_{\max} = 1 + (n-1)\mu^{d-1}$ and corresponds to the eigenvector with all entries as 1, and $\rho_{\min} = (1 - \mu^{d-1})$ repeats $(n-1)$ times, which corresponds to eigenvectors given by $[0, 0, \dots, \underbrace{1, -1}_{i \text{ and } i+1}, 0, 0, \dots, 0]^\top \in \mathbb{R}^n$ for $i = 1, \dots, n-1$. Note that as $d \to \infty$, $\rho_{\max}, \rho_{\min} \to 1$.

Why increasing depth beyond a point hurts? In Theorem D.1, note that for a fixed width w, as the depth increases the variance of the entries $K_0(s,s')$ deviates from its expected value $\mathbb{E}\left[K_0(s,s')\right]$. Thus the structure of the Gram matrix degrades from (15), leading to smaller eigenvalues.

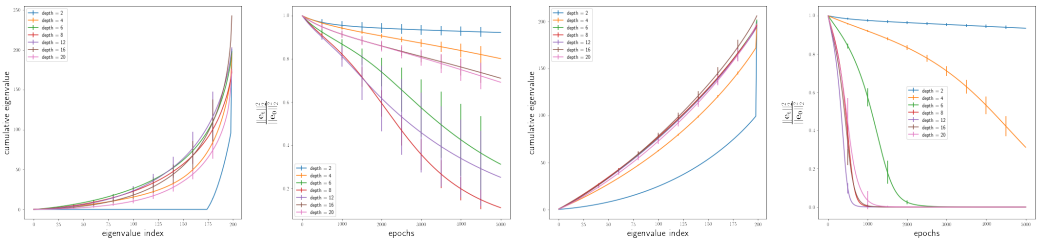

Figure 6: Shows the plots for the memorisation network with $\mu = \frac{1}{2}$ and $\sigma = \sqrt{\frac{2}{w}}$. The number of points to be memorised is $n = 200$. The left most plot shows the e.c.d.f for $w = 25$ and the second plot from the left shows the error dynamics during training for $w = 25$. The second plot from the right shows the e.c.d.f for $w = 500$ and the right most plot shows the error dynamics during training for $w = 500$. All plots are averaged over 10 runs.

E.1 Experiment

We set $n = 200$, and $y_s \sim \text{Uniform}[-1, 1]$. We look at the cumulative eigenvalue (e.c.d.f) obtained by first sorting the eigenvalues in ascending order then looking at their cumulative sum. The ideal behaviour (Figure 5) as predicted from theory is that for indices $k \in [n-1]$, the e.c.d.f should increase at a linear rate, i.e., the cumulative sum of the first k indices is equal to $k(1 - \mu^{d-1})$, and the difference between the last two indices is $1 + (n-1)\mu^{d-1}$. In Figure 6, we plot the actual e.c.d.f for various depths $d = 2, 4, 6, 8, 12, 16, 20$ and $w = 25, 500$ (first and third plots from the left in Figure 6).

Roles of depth and width: In order to compare how the rate of convergence varies with the depth, we set the step-size $\alpha = \frac{0.1}{\rho_{\max}}$, $w = 100$. We use the vanilla SGD-optimiser. Note the $\frac{1}{\rho_{\max}}$ in the stepsize, ensures that the uniformity of maximum eigenvalue across all the instances, and the convergence should be limited by the smaller eigenvalues. We also look at the convergence rate of the ratio $\frac{\|e_t\|_2^2}{\|e_0\|_2^2}$. We notice that for $w = 25$, increasing depth till $d = 8$ improves the convergence, however increasing beyond $d = 8$ worsens the convergence rate. For $w = 500$, increasing the depth till $d = 12$ improves convergence, and $d = 16, 20$ are worse than $d = 12$.