NOVEL KERNEL MODELS AND UNIFORM CONVERGENCE BOUNDS FOR NEURAL NETWORKS BEYOND THE OVER-PARAMETERIZED REGIME

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents two models - called *global* and *local* models - of neural-networks applicable to neural networks of arbitrary width, depth and topology, assuming only finite-energy neural activations. The first model is *exact* (un-approximated) and *global* (applicable for arbitrary weights), casting the neural network in reproducing kernel Banach space (RKBS). This leads to a width-independent (under usual scaling) bound on the Rademacher complexity of neural networks in terms of the spectral-norm of the weight matrices, which is depth-independent for sufficiently small weights. For illustrative purposes we consider how this bound may be applied to untrained networks with LeCun, He and Glorot initialization, discuss their connect to width and depth dependence in the Rademacher complexity bound, and suggest a modified He initialization that gives a depth-independent Rademacher complexity bound whp. The second model is exact and *local*, casting the *change* in neural network function resulting from a bounded change in weights and biases (ie. a training step) in reproducing kernel Hilbert space (RKHS) with a well-defined local-intrinsic neural kernel (LiNK). The neural tangent kernel (NTK) is shown to be a first-order approximation of the LiNK, so the local model gives insight into how the NTK model may be generalized outside of the over-parameterized limit. Analogous to the global model, a bound on the Rademacher complexity of network adaptation is obtained from the local model. Throughout the paper (a) dense feed-forward ReLU networks and (b) residual networks (ResNet) are used as illustrative examples and to provide insight into their operation and properties.

1 INTRODUCTION

The application of reproducing kernel Hilbert space (RKHS (Aronszajn, 1950)) and reproducing kernel Banach space (RKBS (Lin et al., 2022; Der & Lee, 2007; Zhang et al., 2009; Zhang & Zhang, 2012; Song et al., 2013; Sriperumbudur et al., 2011; Xu & Ye, 2014)) theory to the study of neural networks has a long history (Neal, 1996; Weinan et al., 2019; Parhi & Nowak, 2021; Lee et al., 2018; Matthews et al., 2018; Rahimi & Benjamin, 2009; Bach, 2014; 2017; Daniely et al., 2016; Daniely, 2017; Cho & Saul, 2009; Bartolucci et al., 2021; Spek et al., 2022). Neural tangent kernels (NTKs) are an exemplar of this approach, modeling training using a first-order (tangent) model. This approach has led to a wide body of work on convergence and generalization (Du et al., 2019b; Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2020; Zou & Gu, 2019; Arora et al., 2019b;a; Cao & Gu, 2019), mostly focused on the wide-network (over-parameterized) limit. In parallel, there is a growing body of work investigating the uniform convergence, complexity, and capacity of neural networks under various regimes (Neyshabur et al., 2015; 2018; 2019; 2017; Harvey et al., 2017; Bartlett et al., 2017; Golowich et al., 2018; Arora et al., 2018; Allen-Zhu et al., 2018; Dräxler et al., 2018; Li & Liang, 2018; Nagarajan & Kolter, 2019b; Zhou et al., 2019).

Nevertheless, as noted in eg. (Arora et al., 2019b; Lee et al., 2019; Bai & Lee, 2019), there are limits to the descriptive powers of NTK models. A gap has been observed between NTK-based predictions and actual performance (Arora et al., 2019b; Lee et al., 2019), and the validity of NTK models naturally breaks down outside of the over-parameterized limit, which has led to attempts to generalize NTK models outside of the over-parameterized - for example (Bell et al., 2023) use an

exact pathwise kernel, while (Shilton et al., 2023) presented an exact model for dense feedforward neural networks with smooth activations in RKBS, (Bartolucci et al., 2023; Sanders, 2020; Parhi & Nowak, 2021; Unser, 2021; 2019) have explored links to RKBS, and (Bai & Lee, 2019) explored higher-order approximations. However the assumptions made (smoothness, over-parameterization etc) and approximations used limit application and raise the question of whether it is possible to instead formulate *exact, non-approximate* models for neural networks that may be used for similar ends. With regard to uniform convergence, some authors such as (Nagarajan & Kolter, 2019a) have argued that such methods may be unable to explain generalization for neural networks at all due to the behavior of the Rademacher complexity (Bartlett & Mendelson, 2002; Steinwart & Christman, 2008).

In this paper we simultaneously address two questions, namely: (1) is it possible to formulate an exact (non-approximate) model for a wide class of neural networks, thereby avoiding entirely the question of gaps between real performance and model prediction; and (2) can such a model be used to derive general, non-vacuous, training-independent bounds on uniform convergence. We address both questions with the following contributions:

1. **Exact global model:** for arbitrary neural network topologies with arbitrary weights and biases and finite-energy activations, we construct a model that recasts neural networks and bilinear products in a reproducing kernel Banach space (RKBS). This leads to:
   (a) **Rademacher Complexity Bound:** using the global model, we bound the Rademacher complexity as a function of the spectral-norms of the weight matrices. We show that this bound is *width-independent* for standard weight-scaling, and depth-independent in the unbiased case for small weight matrices if the neural activations are L-Lipschitz. For networks satisfying this constraints, we prove that the Rademacher complexity is bounded as $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$ for training set size N.
   (b) **Rademacher Complexity of Randomly Initialized Networks:** we discuss the Rademacher complexity of randomly initialized networks, in particular following LeCun, He and Glorot initialization. We analyse the width- and depth- dependence of the Rademacher complexity bound for these initializations and present modified He and Glorot initializations for which $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$ with high probability.
2. **Exact local model:** again for arbitrary neural network topologies with arbitrary weights and biases and finite-energy activations, we construct a model that recasts the *change* in neural network operation due to a (spectral-norm) bounded change in weights and biases in a reproducing kernel Hilbert space (RKHS), with a locally-intrinsic neural kernel (LiNK) defined by the network topology, neural activations and initial weights. This leads to:
   (a) **Rademacher Complexity Bound:** analogous to the global model, the local model leads to a bound on the Rademacher complexity as a function of the spectral-norms of the *change* to the weight matrices.
   (b) **Connection with NTK:** we prove that the NTK is a first-order approximation LiNK in our local model, casting light on higher-order generalization of NTK models.

The paper is organized as follows. We first discuss the underlying scope and setting of the paper, and the relevant notions and notations used (section 2), as well as related work (section 3). The necessary mathematical background on Hermite polynomials is provided in section 4, including some discussion regarding how this will inform our contribution. We present out global model in section 5 before proceeding to use this model to derive bounds on Rademacher complexity in section 5.2. We finish by presenting our local model in section 6, which is applied to the problem of uniform convergence in section 6.2.

1.1 Mathematical Notation and Indexing Conventions

We use $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{N}_n = \{0, 1, 2, \ldots, n-1\}$, $\mathbb{Z}_+ = \{1, 2, 3, \ldots\}$. $|\mathbb{A}|$ is the number of elements in set $\mathbb{A}$. $\mathrm{Span}(\mathcal{X})$ is the linear span of $\mathcal{X}$. $[a]_+ = \max\{a, 0\}$. $f^{(n)}$ is the n^{th} derivative of f. $L^2(\mathbb{R}, e^{-x^2}) = \{\tau : \mathbb{R} \to \mathbb{R} \mid \int_\infty^\infty |\tau(\zeta)|^2 e^{-\zeta^2} d\zeta < \infty\}$ is the set of finite-energy functions. He_k are the (probabilist's) Hermite polynomials ($k \in \mathbb{N}$). $\mathrm{He}_k = He_k(0)$ are the Hermite numbers.

Column vectors are denoted $\mathbf{a}, \mathbf{b}, \ldots$, with elements $a_i, b_j, \ldots$. $|\mathbf{a}|$ and $\mathrm{sgn}(\mathbf{a})$ are the element-wise absolute and sign. $\|\mathbf{a}\|_2 = (\sum_i |a_i|^2)^{1/2}$ is the Euclidean norm. Matrices are denoted $\mathbf{A}, \mathbf{B}, \ldots$, with elements $A_{i,i'}$, rows $\mathbf{A}_{i:}$, and columns $\mathbf{A}_{:i'}$. $\mathbf{A} \odot \mathbf{B}$ is the Hadamard product. $\mathbf{A} \otimes \mathbf{B}$ is the

Kronecker product. $\mathbf{A}^{\otimes k} = \mathbf{A} \otimes \mathbf{A} \otimes {}^{k \text{ times}}\!\cdots\!\otimes \mathbf{A}$ is the Kronecker power. $\mathrm{diag}_i(\mathbf{A}_i)$ is a block-diagonal matrix with diagonal blocks $\mathbf{A}_i$. $\|\mathbf{A}\|_2 = \sup_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$ is the spectral norm.

$\langle \cdot, \cdot \rangle$ denotes an inner-product, where $\langle \mathbf{a}, \mathbf{a}' \rangle_{\mathbf{g}} = \sum_i g_i a_i a'_i$ for metric $\mathbf{g} > \mathbf{0}$, and unless otherwise stated $\langle \mathbf{a}, \mathbf{a}' \rangle = \langle \mathbf{a}, \mathbf{a}' \rangle_{\mathbf{1}}$. $\langle \cdot, \cdot]$ denotes a bi-linear product, where $\langle \mathbf{a}, \mathbf{a}']_{\mathbf{g}} = \sum_i g_i a_i a'_i$ for (indefinite) metric $\mathbf{g}$, and we follow the convention $\langle \mathbf{A}, \mathbf{a}']_{\mathbf{g}} = [\langle \mathbf{A}_{:j}, \mathbf{a}']_{\mathbf{g}}]_j$, $\langle \mathbf{A}, \mathbf{a}']_{\mathbf{G}} = [\langle \mathbf{A}_{:j}, \mathbf{a}']_{\mathbf{G}_{:j}}]_j$.

Indexing: $\widetilde{\jmath}, j \in \mathbb{N}_D$ index nodes in the computational graph. $\widetilde{\imath}_j \in \mathbb{N}_{\widetilde{H}^{[j]}}$ and $i_j \in \mathbb{N}_{H^{[j]}}$ index inputs and outputs to node j, respectively, where $\widetilde{H}^{[j]}$ and $H^{[j]}$ are the fan-in and fan-out of that node. $\mathbb{P}^{[j]}$ is the antecedent set of node j (set of nodes feeding into node j), and $p^{[j]} = |\mathbb{P}^{[j]}|$ the node's in-degree. Node $D-1$ is the output node, and $j = -1$ indicates the (virtual) input node.

2 Setting and Assumptions

In this paper we will deal with neural networks defined by directed acyclic graphs (aka computational skeletons) as per (Daniely et al., 2016), as this description is flexible enough to allow us to study both simple, fully-connected feed-forward networks such as ReLU and also non-trivial topologies such residual networks (ResNet). In this scheme networks are characterized by nodes and edges, ie.:

1. **Nodes:** a set of D nodes, indexed by $j \in \mathbb{N}_D$, where node $j = D-1$ is the output node and we include a virtual input node indexed as $j = -1$. A described shortly, a node j is characterized by a weight matrix $\mathbf{W}^{[j]} \in \mathbb{R}^{\widetilde{H}^{[j]} \times H^{[j]}}$ (sometimes split into individual sub-matrices $\mathbf{W}^{[\widetilde{\jmath},j]} \in \mathbb{R}^{H^{[\widetilde{\jmath}]} \times H^{[j]}}$ per incoming edge) and bias vector $\mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}}$.
2. **Edges:** a set of directed edges $(\widetilde{\jmath} \to j) \in (\mathbb{N}_D \cup \{-1\}) \times \mathbb{N}_D$ characterized by neural activation functions $\tau^{[\widetilde{\jmath},j]} : \mathbb{R} \to \mathbb{R}$, joining nodes to form a directed acyclic graph (DAG).

Given an input $\mathbf{x}$, data flows from the virtual input node $j = -1$, along the the edges and through the nodes of the DAG to the output node $j = D-1$, as defined by the recursive equation:

$$\mathbf{f}(\mathbf{x};\Theta) = \mathbf{x}^{[D-1]} \quad \left| \begin{array}{l} \tilde{\mathbf{x}}^{[j]} = \left[\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}, \tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} = \tau^{[\widetilde{\jmath},j]}(\mathbf{x}^{[\widetilde{\jmath}]}) \\ \mathbf{x}^{[j]} = \mathbf{W}^{[j]\mathrm{T}}\tilde{\mathbf{x}}^{[j]} + \gamma\mathbf{b}^{[j]} \\ \quad \left(= \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} + \gamma\mathbf{b}^{[j]}\right) \end{array} \right\} \begin{array}{l} \forall j \in \mathbb{N}_D, \widetilde{\jmath} \in \widetilde{\mathbb{P}}^{[j]} \\ \mathbf{x}^{[-1]} = \mathbf{x} \end{array} \tag{1}$$

For node j we define the antecedent set $\widetilde{\mathbb{P}}^{[j]} \subset \mathbb{N}_D \cup \{-1\}$, so the node has in-degree $\tilde{p}^{[j]} = |\widetilde{\mathbb{P}}^{[j]}|$, fan-out $H^{[j]}$ (with input dimension $H^{[-1]} = n$ and output dimension $H^{[D-1]} = m$) and fan-in $\widetilde{H}^{[j]} = \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} H^{[\widetilde{\jmath}]}$. In constructing our models we assume:

1. **Bounded inputs:** $\|\mathbf{x}\|_2 \le 1$.
2. **Finite weights and biases:** $\|\mathbf{W}^{[j]}\|_2, \|\mathbf{b}^{[j]}\|_2 < \infty$.
3. **Finite activations:** $\tau^{[\widetilde{\jmath},j]} \in L^2(\mathbb{R}, e^{-\zeta^2}) = \{\tau : \mathbb{R} \to \mathbb{R} \mid \int_{\infty}^{\infty} |\tau(\zeta)|^2 e^{-\zeta^2} d\zeta < \infty\}$.

We let $\Theta = \{\mathbf{W}^{[j]}, \mathbf{b}^{[j]} : j \in \mathbb{N}_D\}$ denote the collection of all weights and biases. We also denote the set of inputs satisfying the bounded input assumption $\mathbb{X}$, and the set of weights and biases satisfying the finiteness assumption $\mathbb{W}$, so $\mathbf{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{R}^m$. Given a training set $\mathcal{D} = \{(\mathbf{x}_{\{i\}}, \mathbf{y}_{\{i\}}) \in \mathbb{R}^n \times \mathbb{R}^m : i \in \mathbb{N}_N\}$ we assume the goal is to minimize the risk (for loss L, regularizer r, trade-off $\lambda \in \mathbb{R}_+$):

$$\Theta^\star = \operatorname*{argmin}_{\Theta\in\mathbb{W}} \textstyle\sum_i L\left(\mathbf{f}\left(\mathbf{x}_{\{i\}};\Theta\right) - \mathbf{y}_{\{i\}}\right) + \lambda r\left(\Theta\right) \tag{2}$$

When constructing and analysing the network for the global model, for all $j \in \mathbb{N}_D, \widetilde{\jmath} \in \mathbb{P}^{[j]}$ we find it convenient to define a nominal upper bounds $\mu^{[\widetilde{\jmath},j]}$ on the spectral norm of the weight matrices and $\beta^{[j]}$ on the Euclidean norm of the biases:

$$\left\|\mathbf{W}^{[\widetilde{\jmath},j]}\right\|_2 \le \mu^{[\widetilde{\jmath},j]} \text{ and } \left\|\mathbf{b}^{[j]}\right\|_2 \le \beta^{[j]} \tag{3}$$

Similarly for the local model (and in the analysis of the global model) we find it convenient to define a nominal bound $\mu^{[j]}$ so that $\|\mathbf{W}^{[j]}\|_2 \le \mu^{[j]}$. Building these into the model helps us to simplify our results later. *It is important to note that these are convenience factors only and may be as large as necessary to satisfy (3). The only restriction we make here is that $\mu^{[\widetilde{\jmath},j]}$ and $\beta^{[j]}$ must be finite.*

For example for randomly-initialized, untrained neural networks we may derive appropriate, high-probability upper bounds $\mu^{[\widetilde{\jmath},j]}$ and $\beta^{[j]}$ by considering the distribution from which the weights and biases are drawn. In the usual case $W^{[j]}_{\widetilde{\imath}_j,i_j} \sim \mathcal{N}(0,\sigma^{[j]2})$ we have that $\|\mathbf{W}^{[\widetilde{\jmath},j]}_{:i_j}\|_2^2 \sim \sigma^{[j]2}\chi^2_{H^{[\widetilde{\jmath}]}}$, so:

$$\left\|\mathbf{W}^{[\widetilde{\jmath},j]}_{:i_j}\right\|_2^2 \leq \sigma^{[j]2}\left(H^{[\widetilde{\jmath}]} + 2\sqrt{H^{[\widetilde{\jmath}]}}\ln\left(\frac{DH^{[j]}}{2\epsilon}\right) + 2\ln\left(\frac{DH^{[j]}}{\epsilon}\right)\right) \tag{4}$$

whp $\geq 1-\epsilon$ simultaneously $\forall j, i_j$ (see eg. (Laurent & Massart, 2000, Lemma 1, pg 1325)). From this we may derive bounds for several standard initialization schemes, for example:

1. **LeCun** ($\sigma^{[j]2}=\frac{1}{H^{[j]}}$): $\mu^{[\widetilde{\jmath},j]2} = \frac{H^{[\widetilde{\jmath}]}}{H^{[j]}} + \frac{2\sqrt{H^{[\widetilde{\jmath}]}}}{H^{[j]}}\ln\left(\frac{DH^{[j]}}{2\epsilon}\right) + \frac{2}{H^{[j]}}\ln\left(\frac{DH^{[j]}}{\epsilon}\right)$.
2. **He** ($\sigma^{[j]2}=\frac{1}{\widetilde{H}^{[j]}}$): $\mu^{[\widetilde{\jmath},j]2} = \frac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} + \frac{2\sqrt{H^{[\widetilde{\jmath}]}}}{\widetilde{H}^{[j]}}\ln\left(\frac{DH^{[j]}}{2\epsilon}\right) + \frac{2}{\widetilde{H}^{[j]}}\ln\left(\frac{DH^{[j]}}{\epsilon}\right)$.
3. **Glorot** ($\sigma^{[j]2}=\frac{1}{H^{[j]}+\widetilde{H}^{[j]}}$): $\mu^{[\widetilde{\jmath},j]2} = \frac{H^{[\widetilde{\jmath}]}}{H^{[j]}+\widetilde{H}^{[j]}} + \frac{2\sqrt{H^{[\widetilde{\jmath}]}}}{H^{[j]}+\widetilde{H}^{[j]}}\ln\left(\frac{DH^{[j]}}{2\epsilon}\right) + \frac{2}{H^{[j]}+\widetilde{H}^{[j]}}\ln\left(\frac{DH^{[j]}}{\epsilon}\right)$.

To illustrate our results we use the following network topologies (Glorot et al., 2011; He et al., 2016):

1. **Feedforward ReLU:** fully connected, unbiased, layerwise, feedforward, ReLU activations: $\widetilde{\mathbb{P}}^{[j]}=\{j-1\}$, $\gamma=0\ \forall j$; $\tau^{[j-1,j]}(\zeta)=[\zeta]_+\ \forall j\neq 0, D-1$; $\tau^{[-1,0]}(\zeta)=\tau^{[D-2,D-1]}(\zeta)=\zeta$.
2. **Residual Network (ResNet):** unbiased, alternating ReLU/skip network with $D\in 2\mathbb{Z}_+$: $\widetilde{\mathbb{P}}^{[j]}=\{j-1\}\ \forall j$ even; $\widetilde{\mathbb{P}}^{[j]}=\{j-1, j-2\}$, $\mathbf{W}^{[j]}=[\mathbf{W}^{[j]}_C; \frac{1}{2}\mathbf{I}]\ \forall j$ odd; $\gamma=0\ \forall j$; $\tau^{[j-1,j]}(\zeta)=[\zeta]_+\ \forall j\neq 0, D-1$; $\tau^{[j-2,j]}(\zeta)=\zeta\ \forall j$ odd; $\tau^{[-1,0]}(\zeta)=\tau^{[D-2,D-1]}(\zeta)=\zeta$.

3 Related Work

The use of kernel methods to model neural networks dates to at least (Neal, 1996). Fixing all weights and biases except for node j and defining feature map $\boldsymbol{\varphi}^{[j]}(\mathbf{x})=[\tilde{\mathbf{x}}^{[j]};\gamma]$, the network can be written as $\mathbf{f}(\mathbf{x};\Theta)=\mathbf{q}^{[j]}([\mathbf{W}^{[j]\mathrm{T}};\mathbf{b}^{[j]}]\boldsymbol{\varphi}^{[j]}(\mathbf{x}),\mathbf{x})$ for fixed $\mathbf{q}^{[j]}$, and (2) becomes kernel regression with NNGP (neural-network Guassian process) kernel:

$$\begin{aligned}\Sigma^{[j]}(\mathbf{x},\mathbf{x}') &= \mathbb{E}_k\left[\varphi_k^{[j]}(\mathbf{x})\,\varphi_k^{[j]}(\mathbf{x}')\right] = \mathbb{E}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left[\Sigma^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}')\right]\\ \Sigma^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') &= \gamma^2 + \mathbb{E}_{i_{\widetilde{\jmath}}}\left[\tilde{x}^{[\widetilde{\jmath},j]}_{i_{\widetilde{\jmath}}}\tilde{x}'^{[\widetilde{\jmath},j]}_{i_{\widetilde{\jmath}}}\right]\end{aligned} \tag{5}$$

In the wide limit, for suitable initialization, (5) is deterministic (dependent on the distribution $\Theta\sim\nu$), and it can be demonstrated that $\mathbf{x}^{[j]}(\cdot)\sim\mathrm{GP}(0,\Sigma^{[j]})$ (Neal, 1996; Lee et al., 2018; Matthews et al., 2018; Garriga-Alonso et al., 2019; Novak et al., 2019). This is the NNGP model, and may be used to eg. derive insights into the types of function the network is best suited to model. The NNGP kernel for our ReLU example is the arc-cosine kernel (Cho & Saul, 2009).

To model training, neural tangent kernels (NTKs (Jacot et al., 2018; Arora et al., 2019b)) form the basis of a first-order approximation of the behavior of a neural network as the weights and biases vary about their initialization Θ, i.e. $\mathbf{f}(\mathbf{x};\Theta+\Delta\Theta)\approx\mathbf{f}(\mathbf{x};\Theta)+\Delta\Theta^{\mathrm{T}}\nabla_\Theta\mathbf{f}(\mathbf{x};\Theta)$. Training is cast as kernel regression using the neural tangent kernel (NTK), recursively defined as:

$$\left.\begin{aligned}K_{\mathrm{NTK}}(\mathbf{x},\mathbf{x}') &= \mathbb{E}_k\left[\nabla_{\Theta_k}\mathbf{f}(\mathbf{x};\Theta)^{\mathrm{T}}\nabla_{\Theta_k}\mathbf{f}(\mathbf{x}';\Theta)\right] = K^{[D-1]}_{\mathrm{NTK}}(\mathbf{x},\mathbf{x}')\\ K^{[j]}_{\mathrm{NTK}}(\mathbf{x},\mathbf{x}') &= \Sigma^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}') + \mathbb{E}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left[\theta^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}')\,K^{[\widetilde{\jmath}]}_{\mathrm{NTK}}(\mathbf{x},\mathbf{x}')\right]\\ \theta^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') &= \mathbb{E}_{i_{\widetilde{\jmath}}}\left[\left\langle\tau^{[\widetilde{\jmath},j](1)}\left(x^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right)\tau^{[\widetilde{\jmath},j](1)}\left(x'^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right)\right\rangle\right]\ \forall\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}\end{aligned}\right\}\ \forall j \tag{6}$$

and $K^{[-1]}_{\mathrm{NTK}}(\mathbf{x},\mathbf{x}')=0$. The NTK model is accurate for small variations in weights/biases (the lazy regime). In the infinitely wide limit the NTK is deterministic, and weights/biases remain close to their initial values, leading to the gradient flow model where weights flow rather than change in discrete steps during training. This approach gives insight in areas including convergence and generalization (Du et al., 2019b; Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2020; Zou & Gu, 2019; Arora et al., 2019b;a; Cao & Gu, 2019).

Beyond this first-order model, while NTKs have made significant progress, a gap has been observed

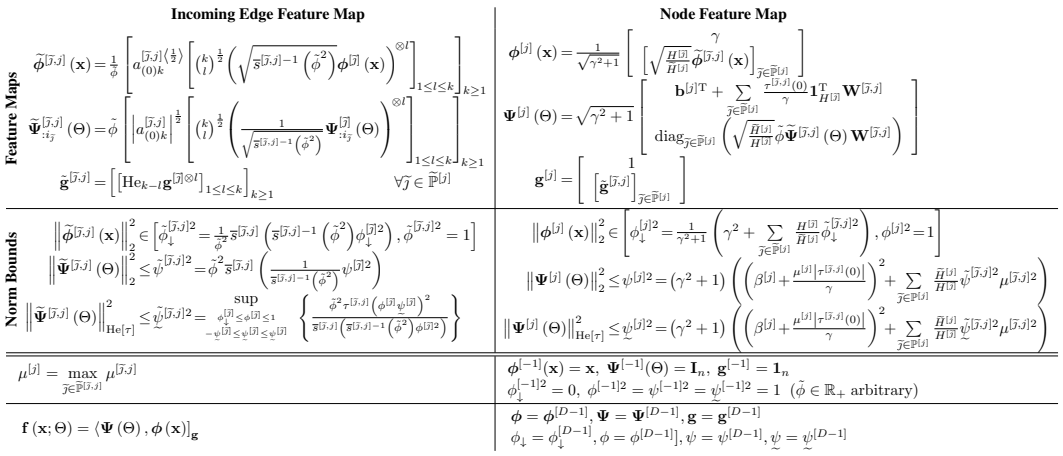

Figure 1: Recursive definition of the global dual and bounds. See theorem 1, section 5 for details.

between NTK-based predictions and actual performance (Arora et al., 2019b; Lee et al., 2019). One approach to bridging this gap is to construct higher-order or exact models. Works in this direction include (Bai & Lee, 2019), which presented a higher-order approximation; (Bell et al., 2023), which used a pathwise kernel; and (Shilton et al., 2023), which used an RKBS model.[1]

4 Hermite Representation of Neural Activations

The (probabilist's) Hermite polynomials are given by (Abramowitz et al., 1972; Morse & Feshbach, 1953; Olver et al., 2010; Courant & Hilbert, 1937) $He_k(\zeta) = (-1)^k e^{\zeta^2/2} \frac{d^k}{d\zeta^k} e^{-\zeta^2/2}\ \forall k \in \mathbb{N}$ and form an orthogonal basis of $L^2(\mathbb{R}, e^{-x^2})$. Both models we present here make use of the Hermite transform of the activations. For all edges $(\widetilde{\jmath}, j)$ in the network we define centered activations:

$$\overline{\tau}^{[\widetilde{\jmath},j]}(\zeta;\xi) = \tau^{[\widetilde{\jmath},j]}(\xi+\zeta) - \tau^{[\widetilde{\jmath},j]}(\xi)$$

which is simply a shifted form of the original activation $\tau^{[\widetilde{\jmath},j]}$. By assumption $\tau^{[\widetilde{\jmath},j]} \in L^2(\mathbb{R}, e^{-\zeta^2})$, so $\overline{\tau}^{[j',j]}(\cdot,\xi) \in L^2(\mathbb{R}, e^{-\zeta^2})$ and hence the Hermite transform exists and is denoted:

$$\begin{aligned}\overline{\tau}^{[\widetilde{\jmath},j]}(\zeta;\xi) &= \textstyle\sum_{k\geq 0} a_{(\xi)k}^{[\widetilde{\jmath},j]} He_k(\zeta) \\ &= \textstyle\sum_{k\geq 1} a_{(\xi)k}^{[\widetilde{\jmath},j]} \sum_{l=1}^{k} \binom{k}{l} \mathrm{He}_{k-l}\zeta^l\end{aligned} \tag{7}$$

where:

$$a_{(\xi)k}^{[\widetilde{\jmath},j]} = \tfrac{1}{\sqrt{2\pi k!}} \textstyle\int_{-\infty}^{\infty} \overline{\tau}^{[\widetilde{\jmath},j]}(\zeta;\xi)\, He_k(\zeta)\, e^{-\zeta^2/2} d\zeta \tag{8}$$

and $\mathrm{He}_k = He_k(0)$ are the (probabilist's) Hermite numbers. Note that in the second form of $\overline{\tau}^{[\widetilde{\jmath},j]}$ we use the additivity properties of the Hermite polynomials (see Appendix A).

For linear activations $\tau(\zeta) = \zeta$ then trivially $a_{(\xi)k} = \delta_{k,0}$. For ReLU activations $\tau(\zeta) = [\zeta]_+$, and as shown in Appendix A.3 (the general case $\forall \xi \in \mathbb{R}$ is more complex - see Appendix A.3):

$$a_{(0)k} = \begin{cases} \frac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!} & \text{if } k = 2p, p \in \mathbb{Z}_+ \\ \frac{1}{2}\delta_{k,1} & \text{otherwise} \end{cases} \tag{9}$$

5 Global Dual Model in Reproducing Kernel Banach Space

In this section we derive a dual model for the network described, where by global we mean not constructed about some weight initialization. Our derivation is similar to (Shilton et al., 2023), but

[1] In a similar vein, (Bartolucci et al., 2023; Sanders, 2020; Parhi & Nowak, 2021; Unser, 2021; 2019) explore the link to RKBS theory, though excepting (Unser, 2019) they only consider shallow networks.

based on a Hermite polynomial expansion rather than a Taylor series, making it applicable to a wider range of activation functions with fewer caveats. Our key result for this section is:[2]

Theorem 1. *Let* $\mathbf{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{R}^m$ *be a neural network (1) satisfying our assumptions,* $\tilde{\phi} \in \mathbb{R}_+$*. Then:*

$$\mathbf{f}(\mathbf{x}; \Theta) = \langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\mathbf{x})]_{\mathbf{g}} \tag{10}$$

with feature maps $\boldsymbol{\Psi} : \mathbb{W} \to \mathcal{W} = \overline{\text{span}}(\boldsymbol{\Psi}(\mathbb{W}))$ *and* $\boldsymbol{\phi} : \mathbb{X} \to \mathcal{X} = \overline{\text{span}}(\boldsymbol{\phi}(\mathbb{X}))$ *and metric* $\mathbf{g}$ *defined in Figure 1, where* $\|\boldsymbol{\Psi}(\Theta)\|_2 \leq \psi$ *and* $\phi_\downarrow \leq \|\boldsymbol{\phi}(\mathbf{x})\|_2 \leq \phi = 1 \; \forall \Theta \in \mathbb{W}, \mathbf{x} \in \mathbb{X}$*. Moreover:*

$$\|\mathbf{f}(\mathbf{x}; \Theta)\|_2 \leq \|\boldsymbol{\Psi}(\Theta)\|_{\text{He}[\tau]} \|\boldsymbol{\phi}(\mathbf{x})\|_2, \quad \|\boldsymbol{\Psi}\|^2_{\text{He}[\tau]} = \sup_{\mathbf{x} \in \mathbb{X}} \frac{\left\| \langle \boldsymbol{\Psi}, \boldsymbol{\phi}(\mathbf{x})]_{\mathbf{g}} \right\|_2^2}{\|\boldsymbol{\phi}(\mathbf{x})\|_2} \tag{11}$$

where $\|\boldsymbol{\Psi}(\Theta)\|_{\text{He}[\tau]} \leq \underset{\sim}{\psi} \; \forall \Theta \in \mathbb{W}$*, as per definitions in Figure 1.*

See Appendix B for a proof of this theorem. Intuitively, this result may be derived recursively, starting from the input node and progressing to the output, using the Hermite expansion of the activation for the edges. The operator-norm based bound (11) is included here due to the fact that the indefinite metric prevents us from naively bound $\|\mathbf{f}(\mathbf{x}; \Theta)\|_2$ in terms of $\phi\psi$ using the Cauchy-Schwarz inequality (as may be required e.g. when bounding Rademacher complexity).

Note that the norm-bound in Theorem 1 is defined in terms of the magnitude functions for each edge:

$$\overline{s}^{[\widetilde{\jmath},j]}(\zeta) = \textstyle\sum_{k \geq 0} \left| a^{[\widetilde{\jmath},j]}_{(0)k} \right| (1+\zeta)^k - \sum_{k \geq 0} \left| a^{[\widetilde{\jmath},j]}_{(0)k} \right| \tag{12}$$

The magnitude functions converge everywhere, are origin-crossing, monotonically increasing and superadditive on $\mathbb{R}^+$ - for details see Appendix A.2. For linear activations $\overline{s}(\zeta) = \zeta$, and for ReLU activations, as shown in Appendix A.3:

$$\overline{s}(\zeta) = \tfrac{1}{2}\zeta \left(\text{erfi}\left(\tfrac{1+\zeta}{\sqrt{2}} \right) + 1 \right) + \tfrac{1}{\sqrt{2\pi}} \left(e^{\frac{1}{2}} - e^{\frac{1}{2}(1+\zeta)^2} \right) + \tfrac{1}{2} \left(\text{erfi}\left(\tfrac{1+\zeta}{\sqrt{2}} \right) - \text{erfi}\left(\tfrac{1}{\sqrt{2}} \right) \right) \tag{13}$$

5.1 Implication: Neural Networks in Reproducing Kernel Banach Space

A reproducing kernel Banach space is defined as:

Definition 1 (Reproducing kernel Banach space (RKBS))**.** A reproducing kernel Banach space on a set $\mathbb{X}$ is a Banach space $\mathcal{F}$ of functions $\mathbf{f} : \mathbb{X} \to \mathbb{Y}$ for which the point evaluation functionals $\boldsymbol{\delta}_{\mathbf{x}}(\mathbf{f}) = \mathbf{f}(\mathbf{x})$ on $\mathcal{F}$ are continuous (i.e. $\forall \mathbf{x} \in \mathbb{X} \; \exists C_{\mathbf{x}} \in \mathbb{R}_+$ such that $\|\boldsymbol{\delta}_{\mathbf{x}}(\mathbf{f})\|_2 \leq C_{\mathbf{x}} \|\mathbf{f}\|_{\mathcal{F}} \; \forall \mathbf{f} \in \mathcal{F}$).

This definition is somewhat generic, so (Lin et al., 2022)[3] study the special case:

$$\mathcal{B} = \left\{ \mathbf{f}(\cdot; \Theta) = \langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\Phi}(\cdot)]_{\mathcal{W} \times \mathcal{X}} \middle| \, \Theta \in \mathbb{W} \right\} \tag{14}$$

where $\boldsymbol{\Phi} : \mathbb{X} \to \mathcal{X}$ is a data feature map, $\boldsymbol{\Psi} : \mathbb{W} \to \mathcal{W}$ is a weight feature map, $\mathcal{X}$ and $\mathcal{W}$ are Banach spaces, and $\langle \cdot, \cdot]_{\mathcal{W} \times \mathcal{X}} : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^m$ is continuous. The following result follows from theorem 1:[4]

Theorem 2. *The set* $\mathcal{F} = \{\mathbf{f}(\cdot; \Theta) : \mathbb{R}^n \to \mathbb{R}^m | \Theta \in \mathbb{W}\}$ *of networks (1) satisfying our assumptions forms a RKBS of form (14), where* $\|\mathbf{f}(\cdot; \Theta)\|_{\mathcal{F}} = \|\boldsymbol{\Psi}(\Theta)\|_{\text{He}[\tau]} \leq \underset{\sim}{\psi}$ *and* $C_{\mathbf{x}} = \|\boldsymbol{\phi}(\mathbf{x})\|_2 \leq \phi$*.*

5.2 Application: Bounding Rademacher Complexity

The global dual formulation may be used to bound Rademacher complexity, which in turn bounds the uniform convergence properties of the network class (Bartlett & Mendelson, 2002; Steinwart & Christman, 2008) (that is, the rate at which the empirical risk approaches the actual risk as a function of dataset size N). Assuming $\mathbf{x} \sim \nu$, the Rademacher complexity is defined as $\mathcal{R}_N(\mathcal{F}) = \mathbb{E}_\nu \mathbb{E}_\epsilon [\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i \in \mathbb{N}_N} \epsilon_i f(\mathbf{x}_i)]$ for Rademacher random variables $\epsilon_i \in \{\pm 1\}$. We have:

[2] When reading the norm-bounds on the feature maps in this theorem it is important to recall that $\beta^{[j]}$ and $\mu^{[\widetilde{\jmath},j]}$ are convenience factors representing an upper bounds on the value of $\|\mathbf{b}^{[j]}\|_2$ and $\|\mathbf{W}^{[\widetilde{\jmath},j]}\|_2$, respectively. We assume these are finite, but in general their value will depend on weight-initialization, dataset complexity and regularization (if any is used).

[3] See e.g. (Der & Lee, 2007; Lin et al., 2022; Zhang et al., 2009; Zhang & Zhang, 2012; Song et al., 2013; Sriperumbudur et al., 2011; Xu & Ye, 2014) for other perspectives.

[4] Note that the RKBS defined in theorem 2 is non-reflexive, which appears to rule out a trivial representor theory based on this dual in the global case (Lin et al., 2022).

Theorem 3. *The set $\mathcal{F} = \{f(\cdot;\Theta) : \mathbb{R}^n \to \mathbb{R} | \Theta \in \mathbb{W}\}$ of networks (1) satisfying our assumptions has Rademacher complexity bounded by $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}} \phi \underset{\sim}{\psi} = \frac{1}{\sqrt{N}} \underset{\sim}{\psi}$ (definitions as per Figure 1).*

The proof follows the usual template (see e.g. (Bartlett & Mendelson, 2002)) using our feature map, with (11) used instead of the Cauchy-Schwarz inequality (Appendix D). Intuititvely, we may think of the recursive bounds $\underset{\sim}{\tilde{\psi}}^{[j]}$, $\underset{\sim}{\psi}^{[\widetilde{\jmath},j]}$ on the weight feature map in terms of signal flow in a electrical circuit, precisely (with reference to figure 1):

1. A signal $\underset{\sim}{\psi}^{[-1]} = 1$ enters the network at the input node $\widetilde{\jmath} = -1$.
2. The outgoing edge from node $(\widetilde{\jmath}, j)$ amplifies this signal:

$$\underset{\sim}{\tilde{\psi}}^{[\widetilde{\jmath},j]} = \sqrt{\sup_{\substack{\phi^{[\widetilde{\jmath}]} \leq \phi^{[\widetilde{\jmath}]} \leq 1 \\ -\underset{\sim}{\psi}^{[\widetilde{\jmath}]} \leq \underset{\sim}{\psi}^{[\widetilde{\jmath}]} \leq \underset{\sim}{\psi}^{[\widetilde{\jmath}]}}} \left\{ \frac{\tilde{\phi}^2 \tau^{[\widetilde{\jmath},j]} \left(\phi^{[\widetilde{\jmath}]} \underset{\sim}{\psi}^{[\widetilde{\jmath}]}\right)^2}{\overline{s}^{[\widetilde{\jmath},j]} \left(\overline{s}^{[\widetilde{\jmath},j]-1} \left(\tilde{\phi}^2\right) \phi^{[\widetilde{\jmath}]2}\right)} \right\}} \tag{15}$$

3. Subsequent nodes j combine signals from incoming edges $(\widetilde{\jmath}, j)$ into an offset weighted sum:

$$\underset{\sim}{\psi}^{[j]2} = \left(\gamma^2 + 1\right) \left(\left(\beta^{[j]} + \tfrac{1}{\gamma} \mu^{[j]} \left| \tau^{[\widetilde{\jmath},j]}(0) \right| \right)^2 + \textstyle\sum_{\widetilde{\jmath} \in \mathbb{P}^{[j]}} \frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}} \underset{\sim}{\tilde{\psi}}^{[\widetilde{\jmath},j]2} \mu^{[\widetilde{\jmath},j]2} \right) \tag{16}$$

4. The signal propogates (steps 2-3) to the output $D-1$. The overall output is $\underset{\sim}{\psi}^{[D-1]}$.

For Lipschitz activations (ie. most activations) we can simplify step 2 with the following theorem:

Theorem 4. *For L-Lipschitz neural activations, in the limit $\tilde{\psi} \to 0_+$ (recall that $\tilde{\psi} \in \mathbb{R}_+$), we have that $\underset{\sim}{\psi}^{[j]2} = (\gamma^2+1)((\beta^{[j]} + \frac{1}{\gamma}\mu^{[j]}|\tau^{[\widetilde{\jmath},j]}(0)|)^2 + L^2 \sum_{\widetilde{\jmath} \in \mathbb{P}^{[j]}} \frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}} \underset{\sim}{\psi}^{[\widetilde{\jmath}]2} \mu^{[\widetilde{\jmath},j]2}) \; \forall j$, which in the ubiased case this simplifies to $\underset{\sim}{\psi}^{[j]2} = L^2 \sum_{\widetilde{\jmath} \in \mathbb{P}^{[j]}} \frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}} \underset{\sim}{\psi}^{[\widetilde{\jmath}]2} \mu^{[\widetilde{\jmath},j]2}$.*

from which we obtain the corollary for unbiased networks:[5]

Corollary 5. *Let $\mathcal{F} = \{f(\cdot;\Theta) : \mathbb{R}^n \to \mathbb{R} | \Theta \in \mathbb{W}\}$ be the set of networks (1) with zero bias $\gamma = \tau^{[\widetilde{\jmath},j]}(0) = 0$, and $\mu^{[\widetilde{\jmath},j]2} \leq \frac{H^{[\widetilde{\jmath}]}}{L^2 \widetilde{H}^{[\widetilde{\jmath}]}} \; \forall j, \widetilde{\jmath} \in \mathbb{P}^{[j]}$, Rademacher complexity is bounded $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$.*

This shows that *Rademacher complexity is depth-independent for sufficiently small network weights, but exponential in depth (the longest path in the network) for large weights*; and that width dependence will scale with the product of $\mu^{[\widetilde{\jmath},j]}$ along the longest path. To gain further insight it is worth considering the Rademacher complexity of unbiased, randomly networks $W^{[j]}_{\widetilde{\imath}_j, i_j} \sim \mathcal{N}(0, \sigma^{[j]2})$. We will consider the depth and width dependence of the complexity bound separately:

- **Depth dependence:** from (4) and corollary 5, the Rademacher complexity will be depth-independent and satisfy $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$ whp $\geq 1-\epsilon$ if:

$$\sigma^{[j]2} \leq \frac{1}{L^2 \left(\widetilde{H}^{[j]} + 2 \frac{\widetilde{H}^{[j]}}{\sqrt{H^{[\widetilde{\jmath}]}}} \ln\left(\frac{D H^{[j]}}{2\epsilon}\right) + 2 \frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}} \ln\left(\frac{D H^{[j]}}{\epsilon}\right) \right)} \tag{17}$$

  Note that this is a modified He initialization accounting for neural activation slope (through L) and correction terms for network topology $\frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}$, node count D and fan-out $H^{[j]}$. A similar modified Glorot initialization follows trivially.
- **Width dependence:** ignoring depth, we observe that for Glorot (and modified He/Glorot) initialization $\mathbb{R}_N(\mathcal{F}) \sim \mathcal{O}(1)$. LeCun and He initialization behave similarly if $H^{[j]} \asymp \widetilde{H}^{[j]}$, but LeCun initialization may scale arbitrarily at the output node ($H^{[D-1]} = m$, while $\widetilde{H}^{[j]}$ is arbitrary), and He initialization may be analogously badly behaved at the input.

With regard to our ReLU and ResNet examples (see section 2), both will have Rademacher complexity $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$ if the spectral norm of all weight matrices is less than 1 (ReLU) or $\frac{1}{2}$ (ResNet). This will also hold for random ReLU/ResNet networks whp for modified He (17) (or Glorot) initialization.

[5] In general the condition in the corollary is $(\gamma^2+1)((\beta^{[j]} + \frac{1}{\gamma}\mu^{[j]}|\tau^{[\widetilde{\jmath},j]}(0)|)^2 + L^2 \sum_{\widetilde{\jmath} \in \mathbb{P}^{[j]}} \frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}} \mu^{[\widetilde{\jmath},j]2}) \leq 1$.

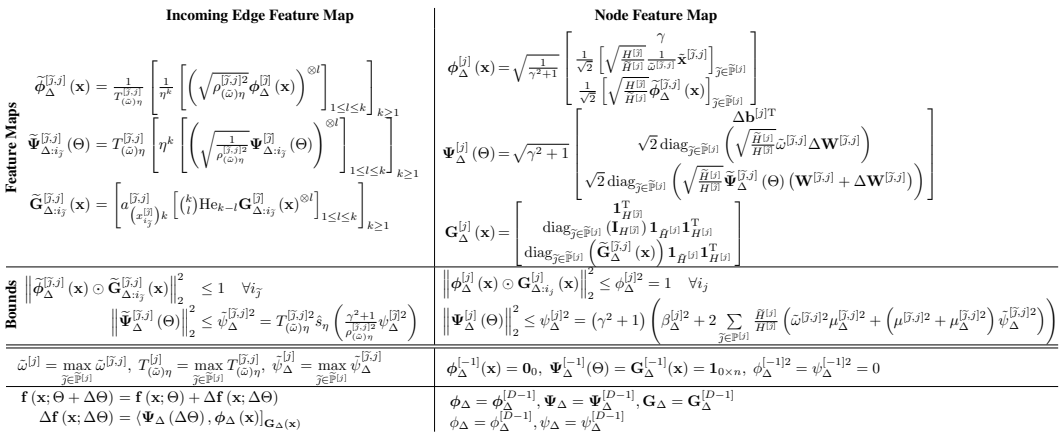

Figure 2: Recursive definition of local dual and bounds. See theorem 6, section 6 for details.

More generally in uniform convergence analysis we must consider how the weight-norm $\mu^{[\widetilde{\jmath},j]}$ evolves or increases during training. It is difficult to draw firm conclusions about this without delving into the specifics of training, however in the lazy regime, or otherwise given sufficiently strong regularization, we would expect that this norm-bound should remain close to its initialization value, potentially indicating good uniform-convergence behavior for a wide class of neural networks.

6 Local Dual Model in Reproducing Kernel Hilbert Space

When considering training or network adaptation it is better to model the change in the network rather than the network in-toto. To this end, in this section we present an exact (non-approximate) local RKHS model. Let Θ be the initial weight and biases and $\Delta\Theta$ the change in weights and biases - so $\Delta\Theta$ might be a training step, multiple steps, or even the complete training process after random initialization. Let $\tilde{\mathbf{x}}^{[j]}, \mathbf{x}^{[j]}$ denote the pre-activation (input) and post-activation (output) of node j with initial weights Θ given input $\mathbf{x}$; and $\Delta\tilde{\mathbf{x}}^{[j]}, \Delta\mathbf{x}^{[j]}$ the change in same due to the change in weights $\Delta\Theta$. The change in network operation is denoted $\Delta\mathbf{f} : \mathbb{X} \times \mathbb{W}_\Delta \to \mathbb{R}^m$:

$$\mathbf{f}\left(\mathbf{x}; \Theta + \Delta\Theta\right) = \mathbf{f}\left(\mathbf{x}; \Theta\right) + \Delta\mathbf{f}\left(\mathbf{x}; \Delta\Theta\right), \tag{18}$$

For the purposes of this analysis we augment our previous assumptions with in section 2 with:

4. **Bounded activation:** $\|\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\|_2 \leq \tilde{\omega}^{[\widetilde{\jmath},j]}\ \forall \widetilde{\jmath} \in \widetilde{\mathbb{P}}^{[j]}$ (note that $\tilde{\omega}^{[\widetilde{\jmath},j]} \leq \tilde{\phi}^{[\widetilde{\jmath},j]} \underset{\sim}{\psi}^{[\widetilde{\jmath},j]}$).
5. **Bounded weight and bias steps:** $\|\Delta\mathbf{W}^{[\widetilde{\jmath},j]}\|_2 \leq \mu_\Delta^{[\widetilde{\jmath},j]} \leq \mu^{[\widetilde{\jmath},j]}$, $\|\Delta\mathbf{b}^{[j]}\|_2 \leq \beta_\Delta^{[j]}$.

satisfying (20) (details in and after Theorem 6). Note that, unlike the parameters $\mu^{[\widetilde{\jmath},j]}, \beta^{[j]}$ in our global model which may be arbitrarily large and as such do not place any restriction on the network which may be modeled, the parameters $\mu_\Delta^{[\widetilde{\jmath},j]}, \beta_\Delta^{[j]}$ bounding step must satisfy (20) and are constrained themselves and subsequently constrain the size of step that can be modeled by the local model.

We define $\mathbb{W}_\Delta$ to be the set of all weight-steps satisfying these assumptions. The parameter $\tilde{\omega}^{[\widetilde{\jmath},j]}$ is a bound on the magnitude of the output of edge $(\widetilde{\jmath} \to j)$ in our initial network $\forall \mathbf{x} \in \mathbb{X}$. With this prequel, we have the following local analogue of theorem 1 (see proof in Appendix C):[6]

Theorem 6. *Let $\Delta\mathbf{f} : \mathbb{X} \times \mathbb{W}_\Delta \to \mathbb{R}^m$ be the change in neural network operation (18). Then:*

$$\Delta\mathbf{f}\left(\mathbf{x}; \Delta\Theta\right) = \left\langle \boldsymbol{\Psi}_\Delta\left(\Delta\Theta\right), \boldsymbol{\phi}_\Delta\left(\mathbf{x}\right)\right\rangle_{\mathbf{G}_\Delta(\mathbf{x})} \tag{19}$$

with feature maps $\boldsymbol{\phi}_\Delta : \mathbb{X} \to \mathcal{X}_\Delta = \overline{\text{span}}(\boldsymbol{\phi}_\Delta(\mathbb{X}))$, $\boldsymbol{\Psi}_\Delta : \mathbb{W}_\Delta \to \mathcal{W}_\Delta = \overline{\text{span}}(\boldsymbol{\Psi}_\Delta(\mathbb{W}_\Delta))$ and metric $\mathbf{G}_\Delta(\mathbf{x})$ as per Figure 2, where $\|\boldsymbol{\phi}_\Delta(\mathbf{x}) \odot \mathbf{G}_{\Delta:i_{D-1}}(\mathbf{x})\|_2 \leq \phi_\Delta\ \forall i_{D-1}$ and $\|\boldsymbol{\Psi}_\Delta(\Delta\Theta)\|_2 \leq$

[6]The decision to use a position dependent metric $\mathbf{G}_\Delta$ here is largely stylistic. We could of course absorb $\mathbf{G}_\Delta$ into $\boldsymbol{\phi}_\Delta$ without substantively changing our results.

$\psi_\Delta\ \forall \mathbf{x} \in \mathbb{X}$, $\Delta\Theta \in \mathbb{W}_\Delta$. *Moreover* $\|\mathbf{\Psi}_\Delta(\Delta\Theta)\|_2 \leq \psi_\Delta = S_\eta^2 < 1$ *if* $\forall j$*:*

$$\mu_\Delta^{[\widetilde{\jmath},j]2} + \tfrac{1}{2\tilde{p}^{[\widetilde{\jmath}]}\tilde{\omega}^{[\widetilde{\jmath}]2}}\beta_\Delta^{[j]2} \leq \tfrac{u^{[j]2}}{4\tilde{p}^{[\widetilde{\jmath}]}\tilde{\omega}^{[\widetilde{\jmath}]2}} \ : \quad u^{[j]2} = \min_{\widetilde{\jmath}: j \in \widetilde{\mathbb{P}}^{[\widetilde{\jmath}]}} \rho^{[j,\widetilde{\jmath}]2}_{(\tilde{\omega}^{[j,\widetilde{\jmath}]})\eta} \left\{ R_\eta^2, \hat{s}_\eta^{-1}\left(\tfrac{u^{[\widetilde{\jmath}]2}}{8\tilde{p}^{[\widetilde{\jmath}]}\mu^{[\widetilde{\jmath}]2}}\right)\right\} \tag{20}$$

We emphasise that this exactly models the change in the neural network without approximation so long as conditions in the theorem are met. This is in contrast to the NTK model, which is a first-order approximation whose accuracy will decrease as the step-size increases (e.g. for narrower networks).

In constructing Theorem 6 we use the *rectified activation functions* and their envelopes, respectively:

$$\begin{aligned} \hat{\tau}_\eta^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') &= \textstyle\sum_{k\geq 1} \frac{a_{(\xi)^k}^{[\widetilde{\jmath},j]} a_{(\xi')^k}^{[\widetilde{\jmath},j]}}{\eta^{2k}} \sum_{l=1}^k \binom{k}{l}^2 \mathrm{He}_{k-l}^2 \zeta^l \\ \hat{\tau}_\eta^{[\widetilde{\jmath},j]}\left(\zeta;\omega^{[\widetilde{\jmath},j]}\right) &= \sup_{|\xi|,|\xi'|\leq\omega^{[\widetilde{\jmath},j]}} \hat{\tau}_\eta^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') = \sup_{|\xi|\leq\omega^{[\widetilde{\jmath},j]}} \hat{\tau}_\eta^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi) \end{aligned} \tag{21}$$

where $\xi, \xi' \in [-\omega^{[\widetilde{\jmath},j]}, \omega^{[\widetilde{\jmath},j]}]$ are the centers of the rectified activation functions (the initial activation for some input about which our model is constructed) and $\eta \in (0,1)$ is fixed. Unlike the magnitude functions, the rectified activations have a finite ROC $|\hat{\tau}_\eta^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi')| \leq T^{[\widetilde{\jmath},j]2}_{(\xi,\xi')\eta}\ \forall|\zeta| \leq \rho^{[\widetilde{\jmath},j]2}_{(\xi,\xi')\eta}$, and likewise $|\hat{\tau}_\eta^{[\widetilde{\jmath},j]}(\zeta;\omega^{[\widetilde{\jmath},j]})| \leq T^{[\widetilde{\jmath},j]2}_{(\omega^{[\widetilde{\jmath},j]})\eta}\ \forall|\zeta| \leq \rho^{[\widetilde{\jmath},j]2}_{(\omega^{[\widetilde{\jmath},j]})\eta}$ (see Appendix A.2). The rectified activation envelopes are origin crossing, monotonically increasing and superadditive. We also define:

$$\hat{s}_\eta(\zeta) = \textstyle\sum_{k\geq 1}\eta^{2k}\sum_{1\leq l\leq k}\zeta^l = \frac{\zeta}{1-\zeta}\left(\frac{\eta^2}{1-\eta^2} - \frac{\zeta\eta^2}{1-\zeta\eta^2}\right)$$

which converges as given $\forall|\zeta| < R_\eta^2 < 1$, whereon $|\hat{s}_\eta(\zeta)| \leq S_\eta^2 = \frac{R_\eta^2}{1-R_\eta^2}(\frac{\eta^2}{1-\eta^2} - \frac{\eta^2 R_\eta^2}{1-\eta^2 R_\eta^2})$.

It is difficult to obtain a closed-form expression for the rectified activation or its envelope for the ReLU, but they are relatively straightforward to calculate, as are their convergence bounds. Figure 3 in Appendix A.3 shows a sample of various rectified activations for the ReLU with different centers.

6.1 Implication: Neural Network Change in Reproducing Kernel Hilbert Space

A vector-valued (v-v) reproducing kernel Hilbert space is defined as follows (Aronszajn, 1950; Steinwart & Christman, 2008; Shawe-Taylor & Cristianini, 2004; Mercer, 1909; Micchelli & Pontil, 2005; Caponnetto & De Vito, 2007; Reisert & Burkhardt, 2007; Carmeli et al., 2005; Schwartz, 1964):

Definition 2 (Reproducing kernel Hilbert space (RKHS))**.** A v-v reproducing kernel Hilbert space $\mathcal{H}$ on a set $\mathbb{X}$ is a Hilbert space $\mathcal{F}$ of functions $\mathbf{f} : \mathbb{X} \to \mathbb{R}^m$ for which the point evaluation functionals $\boldsymbol{\delta}_\mathbf{x}(\mathbf{f}) = \mathbf{f}(\mathbf{x})$ on $\mathcal{F}$ are continuous ($\forall \mathbf{x} \in \mathbb{X}\ \exists C_\mathbf{x} \in \mathbb{R}_+$ s.t. $\|\boldsymbol{\delta}_\mathbf{x}(\mathbf{f})\|_2 \leq C_\mathbf{x}\|\mathbf{f}\|_\mathcal{F}\ \forall \mathbf{f} \in \mathcal{F}$).

For an RKHS, Reisz representor theory implies that $\forall \mathbf{x} \in \mathbb{X}\ \exists$ unique $\mathbf{K}_\mathbf{x} \in \mathcal{F} \times \mathbb{R}^m$ such that $\langle \mathbf{f}(\mathbf{x}), \mathbf{v}\rangle = \langle \mathbf{f}, \mathbf{K}_\mathbf{x}\mathbf{v}\rangle_\mathcal{H}\ \forall \mathbf{v} \in \mathbb{R}^m$. From this, the kernel $\mathbf{K} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^{m\times m}$ is defined as:

$$\mathbf{K}(\mathbf{x},\mathbf{x}') = \left[\left\langle \mathbf{K}_\mathbf{x}\boldsymbol{\delta}^{[D-1]}_{(i_{D-1})}, \mathbf{K}_{\mathbf{x}'}\boldsymbol{\delta}^{[D-1]}_{(i'_{D-1})}\right\rangle_\mathcal{H}\right]_{i_{D-1},i'_{D-1}}, \quad \text{where } \boldsymbol{\delta}^{[j]}_{(k)} = [\delta_{k,i_j}]_{i_j}$$

Moore-Aronszajn theorem allows us to run the argument in reverse: any symmetric, positive definite $\mathbf{K} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^{m\times m}$ uniquely defines an RKHS, $\mathcal{H}_\mathbf{K}$ for which $\mathbf{K}$ is the kernel. From theorem 6:

Theorem 7. *The set* $\mathcal{F}_\Delta = \{\Delta\mathbf{f}(\cdot;\Delta\Theta) : \mathbb{R}^n \to \mathbb{R}^m | \Delta\Theta \in \mathbb{W}_\Delta\}$ *of changes in network behavior satisfying our assumptions, including the bound, lies in an RKHS* $\mathcal{H}_\mathbf{K}$ *(that is,* $\mathcal{F}_\Delta \subset \mathcal{H}_\mathbf{K}$*) with kernel* $\mathbf{K} = \mathbf{I}_m K_{\text{LiNK}}$*, where* $K_{\text{LiNK}} = K^{[D-1]}$*, is the* Local-intrinsic Neural Kernel (LiNK)*,* $\forall j$*:*

$$K^{[j]}(\mathbf{x},\mathbf{x}') = \tfrac{p^{[j]}}{\gamma^2+1} \mathop{\mathbb{E}}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left[\tfrac{H^{[\widetilde{\jmath}]}}{\tilde{\omega}^{[\widetilde{\jmath},j]2}}\Sigma^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') + \tfrac{H^{[\widetilde{\jmath}]}}{T^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta}} \mathop{\mathbb{E}}_{i_{\widetilde{\jmath}}}\left[\hat{\tau}_\eta^{[\widetilde{\jmath},j]}\left(\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta} K^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}'); x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}, x_{i_{\widetilde{\jmath}}}^{\prime[\widetilde{\jmath}]}\right)\right]\right] \tag{22}$$

and $K^{[-1]}(\mathbf{x},\mathbf{x}') = 0$ *and* $\Sigma^{[j]}(\mathbf{x},\mathbf{x}')$ *is the NNGP kernel. Moreover:*

$$\begin{aligned} \lim_{\eta\to 1} \mathbb{E}_{i_{\widetilde{\jmath}}}\left[\hat{\tau}_\eta^{[\widetilde{\jmath},j]}\left(\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta} K^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}'); x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}, x_{i_{\widetilde{\jmath}}}^{\prime[\widetilde{\jmath}]}\right)\right] &= \sum_{q\geq 1} \theta_q^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}')\left(\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta} K^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}')\right)^q \\ \theta_q^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') &= \mathbb{E}_{i_{\widetilde{\jmath}}}\left[\tfrac{1}{q!}\tau^{[\widetilde{\jmath},j](q)}\left(x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\right)\tfrac{1}{q!}\tau^{[\widetilde{\jmath},j](q)}\left(x_{i_{\widetilde{\jmath}}}^{\prime[\widetilde{\jmath}]}\right)\right] \end{aligned} \tag{23}$$

(here $\theta_q^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}')$ *is the raw covariance of the* q^{th} *derivative of link* $(\widetilde{\jmath},j)$*'s activation given* $\mathbf{x}$*,* $\mathbf{x}'$*.)*

The proof is requres two steps - step one is to apply the kernel trick (after some preliminaries), while step two uses Mertens' theorem to obtain the final result - see Appendix C.3 for details. We observe that the NTK is essentially (with some additional scaling factors) a first-order (in q) approximation of the LiNK - if we take the limit $\eta \to 1$ then to first order the LiNK is approximately:

$$K^{[j]}(\mathbf{x},\mathbf{x}') \approx \frac{p^{[j]}}{\gamma^2+1} \mathop{\mathbb{E}}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \left[\frac{H^{[\widetilde{\jmath}]}}{\widetilde{\omega}^{[\widetilde{\jmath},j]2}} \Sigma^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') + \frac{\rho_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2} H^{[\widetilde{\jmath}]}}{T_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2}} \theta_1^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') K^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}') \right]$$

recalling $K_{\text{LiNK}} = K^{[D-1]}$, where:

$$\theta_1^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') = \mathbb{E}_{i_{\widetilde{\jmath}}} \left[\tau^{[\widetilde{\jmath},j](1)}\left(x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\right) \tau^{[\widetilde{\jmath},j](1)}\left(x_{i_{\widetilde{\jmath}}}'^{[\widetilde{\jmath}]}\right) \right]$$

which is essentially the NTK (6) with some additional scale factors. Assuming random initialization the LiNK is well-defined for almost all $\mathbf{x} \in \mathbb{X}$ if $\tau^{[\widetilde{\jmath},j]} \in \mathcal{C}^\infty$ for almost all $\mathbf{x} \in \mathbb{X}$. Note however that $\mathcal{F}_\Delta \subset \mathcal{H}_\mathbf{K}$ - ie. $\mathcal{F}_\Delta$ is not an RKHS in general, but rather a subspace inside of one. Nor can we meaningfully replace $\mathcal{F}_\Delta$ with its span or completion, as this will contain elements that do not correspond to physically realizable networks. This is clear from Figure 2, where the weight-feature map $\mathbf{\Psi}_\Delta$ maps the network weights onto a (non-flat) subspace of $\ell^2(\mathbb{N})^m$, no column of which coincides with the subspace of same onto which $\boldsymbol{\phi}_\Delta$ maps input space. Thus in general the LiNK cannot be naively used as a basis for a representor theory in terms of the training dataset.

6.2 Application: Bounding Rademacher Complexity for Adaptation

Like the global model, an obvious application of the local dual model is the bounding of Rademacher complexity. The following result may be viewed as the local analogue of our previous bound:

Theorem 8. *The set $\mathcal{F}_\Delta = \{\Delta f(\cdot;\Delta\Theta) : \mathbb{R}^n \to \mathbb{R} | \Delta\Theta \in \mathbb{W}_\Delta\}$ of change in neural-network operation satisfying (20) has Rademacher complexity $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}} \phi_\Delta \psi_\Delta$ (defined in Figure 2).*

The proof follows the template of (Bartlett & Mendelson, 2002) using the local feature map and the Cauchy-Schwarz inequality (see Appendix D). Assuming an unbiased network $\phi_\Delta = 1$, and the Rademacher complexity bound is determined by the recursive equation:

$$\psi_\Delta^{[j]2} = (\gamma^2+1)\left(\beta_\Delta^{[j]2} + 2\sum_{\widetilde{\jmath}\in\mathbb{P}^{[j]}} \frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}} \left(\tilde{\omega}^{[\widetilde{\jmath},j]2}\mu_\Delta^{[\widetilde{\jmath},j]2} + \left(\mu^{[\widetilde{\jmath},j]2} + \mu_\Delta^{[\widetilde{\jmath},j]2}\right) T_{(\tilde{\omega})\eta}^{[\widetilde{\jmath},j]2} \hat{s}_\eta\left(\frac{\gamma^2+1}{\rho_{(\tilde{\omega})\eta}^{[\widetilde{\jmath},j]2}}\psi_\Delta^{[\widetilde{\jmath}]2}\right)\right)\right)$$

The width-dependence of the bound is dependent on the width-dependence of $\mu^{[\widetilde{\jmath},j]}$ and $\mu_\Delta^{[j]}$, but unfortunately there appears to be an unavoidably exponential depth-dependancy not present in the global model as $\hat{s}_\eta$ is positive and increasing on $\mathbb{R}^+$. In future work we hope to use this theorem to explain how methods such as LoRA (Hu et al., 2021) achieve better performance in terms of uniform convergence properties with restricted weight update rank (and hence the spectral norm of the weight-matrix changes). Moreover it may be interesting in future investigation to explore if spectral analysis of the LiNK could be used to bound *local* Rademacher complexity (Cortes et al., 2013; Bartlett et al., 2005), as previous investigations in RKHS using this approach give bounds up to $\mathcal{O}(\frac{1}{N})$.

7 Conclusions and Future Directions

In this paper we have presented two models of neural networks and neural network training for neural network of arbitrary width, depth and topology. First we presented an exact (non-approximated) RKBS model of the overall network in the form of a bilinear product between a data- and weight-feature map. We have used this model to construct a bound on Rademacher complexity, and for Lipschitz activations we have given conditions under which the Rademacher complexity is depth-independent, and how diffierent initialization schemes can achieve $\mathbb{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$. The second model we have presented models the *change* in the neural network due to a bounded change in weights and biases. This model cast the change in RKHS with the local-intrinsic neural kernel (LiNK). We have shown that this can be used to bound Rademacher complexity for network adaptation. We have also discussed the role of weight initialization and implications for feedforward ReLU networks and residual networks (ResNets), and presented the local intrinsic neural kernel for the ResNet.

REFERENCES

Milton Abramowitz, Irene A. Stegun, and Donald A. McQuarrie. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1972.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, Jan–Jun 1950.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of ICML*, 2018.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019a.

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*, pp. 8139–8148, 2019b.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Francis R. Bach. Breaking the curse of dimensionality with convex neural networks. *CoRR*, abs/1412.8690, 2014. URL `http://arxiv.org/abs/1412.8690`.

Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces. *arXiv preprint arXiv:2109.09710*, 2021.

Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, January 2023.

Brian Bell, Michaela Geyer, Juston Moore, David Glickenstein, and Amanda Fernandez. An exact kernel equivalence for finite classification models. In *TAG-ML*, 2023.

John P. Boyd. The rate of convergence of hermite function series. *Mathematics of Computation*, 35 (152):1309–1316, October 1980.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in neural information processing systems*, volume 32, 2019.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found Comput Math*, 7:331–368, 2007.

C. Carmeli, E. De Vito, and A. Toigo. Reproducing kernel hilbert spaces and mercer theorem. Technical Report arvXiv:math.FA/0504071, arXiv, 2005.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In Bengio Y., Schuurmans D., Lafferty J. D., C. K. I. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems 22*, pp. 342–350. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf.

Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. *Advances in neural information processing systems*, 26, 2013.

R. Courant and D. Hilbert. *Methods of Mathematical Physics*. John Wiley and sons, New York, 1937.

Amit Daniely. Sgd learns the conjugate kernel class of the network. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2422–2430. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6836-sgd-learns-the-conjugate-kernel-class-of-the-network.pdf.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2253–2261. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6427-toward-deeper-understanding-of-neural-networks-the-power-of-initialization-and-a-dual-view-on-expressivity.pdf.

Ricky Der and Danial Lee. Large-margin classification in banach spaces. In *Proceedings of the JMLR Workshop and Conference 2: AISTATS2007*, pp. 91–98, 2007.

Felix Dräxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *Conference on Learning Representations*, 2019b.

Adria Garriga-Alonso, Carl E. Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, pp. 1–16, May 2019.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *COLT*, 2018.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, London, 2000.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Einar Hille. Contributions to the theory of Hermitian series. II. The representation problem. *Trans. Amer. Math. Soc.*, 47:80–94, 1940.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

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

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *In International Conference on Learning Representations*, 2018.

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. *Advances in neural information processing systems*, 32, 2019.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, 2018.

Rongrong Lin, Haizhang Zhang, and Jun Zhang. On reproducing kernel banach spaces: Generic definitions and unified framework of constructions. *Acta Mathematica Sinica, English Series*, 2022.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv e-prints*, 2018.

James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Transactions of the Royal Society of London*, 209(A), 1909.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Philip M. Morse and Herman Feshbach. *Methods of Theoretical Physics*. McGraw-Hill, 1953.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dé Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11615–11626. Curran Associates, Inc., 2019a. URL http://papers.nips.cc/paper/9336-uniform-convergence-may-be-unable-to-explain-generalization-in-deep-learning.pdf.

Vaishnavh Nagarajan and Zico Kolter. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations (ICLR)*, 2019b.

Radford M. Neal. *Priors for infinite networks*, pp. 29–53. Springer, 1996.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of Conference on Learning Theory*, pp. 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5949–5958, 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of ICLR*, 2018.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *Proceedings of ICLR*, 2019.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Dan Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations, ICLR 2019*, 2019.

Frank W. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, USA, 1st edition, 2010. ISBN 0521140633.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.*, 22(43):1–40, 2021.

Ali Rahimi and Recht Benjamin. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 21*, pp. 1313–1320. Curran Associates, Inc., 2009.

Marco Reisert and Hans Burkhardt. Learning equivariant functions with matrix valued kernels. *Journal of Machine Learning Research*, 8(15):385–408, 2007.

Koen Sanders. Neural networks as functions parameterized by measures: Representer theorems and approximation benefits. Master's thesis, Eindhoven University of Technology, 2020.

Laurent Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'analyse mathématique*, 13:115–256, 1964.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Alistair Shilton, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Gradient descent in neural networks as sequential learning in reproducing kernel banach space. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31435–31488. PMLR, 23–29 Jul 2023.

Guohui Song, Haizhang Zhang, and Fred J. Hickernell. Reproducing kernel banach spaces with the ℓ^1 norm. *Applied and Computational Harmonic Analysis*, 34(1):96–116, Jan 2013.

Len Spek, Tjeerd Jan Heeringa, and Christoph Brune. Duality for neural networks through reproducing kernel banach spaces. *arXiv preprint arXiv:2211.05020*, 2022.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. Lanckriet. Learning in hilbert vs. banach spaces: A measure embedding viewpoint. In *Advances in Neural Information Processing Systems*, pp. 1773–1781, 2011.

Ingo Steinwart and Andreas Christman. *Support Vector Machines*. Springer, 2008.

Michael Unser. A representer theorem for deep neural networks. *J. Mach. Learn. Res.*, 20(110):1–30, 2019.

Michael Unser. A unifying representer theorem for inverse problems and machine learning. *Foundations of Computational Mathematics*, 21(4):941–960, 2021.

E. Weinan, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

Yuesheng Xu and Qi Ye. Generalized mercer kernels and reproducing kernel banach spaces. *arXiv preprint arXiv:1412.8663*, 2014.

Haizhang Zhang and Jun Zhang. Regularized learning in banach spaces as an optimization problem: representer theorems. *Journal of Global Optimization*, 54(2):235–250, Oct 2012.

Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Nonvacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in neural information processing systems*, volume 32, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

A PROPERTIES OF HERMITE POLYNOMIALS

The (probabilist's) Hermite polynomials are given by (Abramowitz et al., 1972; Morse & Feshbach, 1953; Olver et al., 2010; Courant & Hilbert, 1937):

$$He_k(\zeta) = (-1)^k e^{\frac{\zeta^2}{2}} \frac{d^k}{d\zeta^k} e^{-\frac{\zeta^2}{2}} \qquad \forall k \in \mathbb{N}$$

or, explicitly:

$$He_k(\zeta) = k! \sum_{0 \le 2p \le k} \frac{(-1)^p}{2^p p!(k-2p)!} \zeta^{k-2p} \qquad \forall k \in \mathbb{N} \tag{24}$$

and form an orthogonal basis of $L^2(\mathbb{R}, e^{-x^2})$. Thus for any $f \in L^2(\mathbb{R}, e^{-x^2})$ there exist Hermite coefficients $a_0, a_1, \ldots \in \mathbb{R}$ (i.e. the Hermite transform of f) so that:

$$f(\zeta) = \sum_{k \in \mathbb{N}} a_k He_k(\zeta) \quad \forall \zeta \in \mathbb{R}$$

where:

$$a_k = \frac{1}{k!\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\zeta) e^{-\frac{\zeta^2}{2}} He_k(\zeta) \, d\zeta$$

This series representation converges everywhere on the real line. Moreover (Hille, 1940; Boyd, 1980) this series converges on a strip $\mathbb{S}_\rho = \{z \in \mathbb{C} : -\rho < \mathrm{Im}(z) < \rho\}$ of width ρ about the real axis in the complex plane, where:[7]

$$\rho = -\limsup_{k \to \infty} \frac{1}{\sqrt{2k+1}} \log\left(\left|\frac{a_k}{\sqrt{k!\sqrt{\pi}}}\right|\right) \tag{25}$$

The Hermite numbers derive from the Hermite polynomials:[8]

$$\mathrm{He}_k \triangleq He_k(0) = \begin{cases} 0 & \text{if } k \text{ odd} \\ \frac{k!}{(\frac{k}{2})!}\left(-\frac{1}{2}\right)^{\frac{k}{2}} & \text{if } k \text{ even} \end{cases}$$

It is well known that (see e.g. (Morse & Feshbach, 1953)):

$$He_k(\zeta + \xi) = \sum_{0 \le l \le k} \binom{k}{l} He_{k-l}(\zeta) \xi^l$$

and so:

$$He_k(\zeta) = \sum_{0 \le l \le k} \binom{k}{l} \mathrm{He}_{k-l} \zeta^l$$

It follows that, taking care not to change or order of summation (remember this is an alternating series):

$$f(\zeta) = \sum_{k=0}^{\infty} a_k \sum_{l=0}^{k} \binom{k}{l} \mathrm{He}_{k-l} \zeta^l$$

Next we derive a helpful property involving *rectified* Hermite expansions. Let $f \in L^2(\mathbb{R}, e^{-x^2})$, $f(0) = 0$, then:

$$f(x) = \sum_{k=1}^{\infty} a_k He(x) = \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} \binom{k}{l} \mathrm{He}_{k-l} x^l$$

Denoting the imaginary element $\mathtt{i}$:

$$\begin{aligned} f(\mathtt{i}x) &= \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} \binom{k}{l} \mathrm{He}_{k-l} (\mathtt{i}x)^l \\ &= \sum_{k=1}^{\infty} a_k \sum_{l=0}^{k-1} \binom{k}{k-l} \mathrm{He}_l (\mathtt{i}x)^{k-l} \\ &= \sum_{k=1}^{\infty} \mathtt{i}^k a_k \sum_{l=0}^{k-1} \binom{k}{k-l} \left(\mathtt{i}^l \mathrm{He}_l\right) x^{k-l} \end{aligned}$$

[7] Note that (Hille, 1940; Boyd, 1980) use the normalized physicist's Hermite polynomials. The additional scale factor here arises in the translation to the un-normalized probabilist's Hermite polynomials used here.

[8] Typically the Hermite numbers are defined from the physicist's Hermite polynomials, but as we use the Probabilist's form we find these more convenient.

Recall that $\mathrm{He}_l = 0$ for $l = 1, 3, 5, \ldots$, and $\mathrm{sgn}(\mathrm{He}_{2p}) = (-1)^p$. Therefore

$$\begin{aligned} f(\mathtt{i}x) &= \sum_{k=1}^{\infty} \mathtt{i}^k a_k \sum_{l=0}^{k-1} \binom{k}{k-l} |\mathrm{He}_l| \, x^{k-l} \\ &= \sum_{k=1}^{\infty} \mathtt{i}^k a_k \sum_{l=1}^{k} \binom{k}{l} |\mathrm{He}_{k-l}| \, x^l \end{aligned}$$

and so:

$$\begin{aligned} \mathrm{Im}\left(f(\mathtt{i}x)\right) &= \sum_{k=1,3,5,\ldots} (-1)^{\lfloor \frac{k}{2} \rfloor} a_k \sum_{l=1}^{k} \binom{k}{l} |\mathrm{He}_{k-l}| \, x^l \\ \mathrm{Re}\left(f(\mathtt{i}x)\right) &= \sum_{k=2,4,6,\ldots} (-1)^{\lfloor \frac{k}{2} \rfloor} a_k \sum_{l=1}^{k} \binom{k}{l} |\mathrm{He}_{k-l}| \, x^l \end{aligned}$$

$$\mathrm{Im}\left(f(\mathtt{i}x)\right) + \mathrm{Re}\left(f(\mathtt{i}x)\right) = \sum_{k=1}^{\infty} (-1)^{\lfloor \frac{k}{2} \rfloor} a_k \sum_{l=1}^{k} \binom{k}{l} |\mathrm{He}_{k-l}| \, x^l \tag{26}$$

Finally we make some observations regarding derivatives that will be required later. Let $f \in L^2(\mathbb{R}, e^{-x^2})$, $f(0) = 0$. Then:

$$f(x) = \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} \binom{k}{l} \mathrm{He}_{k-l} x^l$$

and subsequently:

$$\begin{aligned} f^{(1)}(x) &= \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} l \binom{k}{l} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} l \tfrac{k!}{l!(k-l)!} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} k \tfrac{(k-1)!}{(l-1)!(k-l)!} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} k \binom{k-1}{l-1} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=0}^{\infty} (k+1) a_{k+1} \sum_{l=0}^{k-1} \binom{k}{l} \mathrm{He}_{k-l} x^l \end{aligned}$$

and:

$$\begin{aligned} f^{(2)}(x) &= \sum_{k=0}^{\infty} (k+1) a_{k+1} \sum_{l=1}^{k-1} (l-1) \binom{k}{l} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=0}^{\infty} (k+1) a_{k+1} \sum_{l=1}^{k-1} (l-1) \tfrac{k!}{l!(k-l)!} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=0}^{\infty} (k+1) a_{k+1} \sum_{l=1}^{k-1} k \tfrac{(k-1)!}{(l-l)!(k-l)!} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=1}^{\infty} (k+1) a_{k+1} \sum_{l=1}^{k-1} k \tfrac{(k-1)!}{(l-l)!(k-l)!} \mathrm{He}_{k-l} x^{l-1} \\ &= \sum_{k=0}^{\infty} (k+1)(k+2) a_{k+2} \sum_{l=0}^{k-2} \binom{k}{l} \mathrm{He}_{k-l} x^l \end{aligned}$$

and so on to:

$$f^{(n)}(x) = \sum_{k=0}^{\infty} \tfrac{(k+n)!}{k!} a_{k+n} \sum_{l=0}^{k-n} \binom{k}{l} \mathrm{He}_{k-l} x^l$$

A.1 Activation Functions

Following the previous method we introduce our notation for the activation functions. Recall $\tau^{[\widetilde{\jmath},j]} \in L^2(\mathbb{R}, e^{-x^2})$ by assumption. Subsequently $\overline{\tau}^{[\widetilde{\jmath},j]} \in L^2(\mathbb{R}, e^{-x^2})$, where:

$$\begin{aligned}\overline{\tau}^{[\widetilde{\jmath},j]}(\zeta;\xi) = \tau^{[\widetilde{\jmath},j]}(\xi+\zeta) - \tau^{[\widetilde{\jmath},j]}(\xi) &= \sum_{k\in\mathbb{N}} a^{[\widetilde{\jmath},j]}_{(\xi)k} He_k(\zeta) \quad \forall \zeta \in \mathbb{R}_+ \\ &= \sum_{k=0}^{\infty} a^{[\widetilde{\jmath},j]}_{(\xi)k} \sum_{l=0}^{k} \tbinom{k}{l} \mathrm{He}_{k-l}\zeta^l \\ &= \sum_{k=1}^{\infty} a^{[\widetilde{\jmath},j]}_{(\xi)k} \sum_{l=1}^{k} \tbinom{k}{l} \mathrm{He}_{k-l}\zeta^l\end{aligned} \tag{27}$$

(in the final step we use that $\overline{\tau}^{[\widetilde{\jmath},j]}(0;\xi) = 0$) with coefficients:

$$a^{[\widetilde{\jmath},j]}_{(\xi)k} = \tfrac{1}{k!\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{\tau}^{[\widetilde{\jmath},j]}(\zeta;\xi)\, e^{-\frac{\zeta^2}{2}} He_k(\zeta)\, d\zeta$$

which converges on a strip $\mathbb{S}_{\rho^{[\widetilde{\jmath},j]}_{(\xi)}} = \{z \in \mathbb{C} : -\rho^{[\widetilde{\jmath},j]}_{(\xi)} < \mathrm{Im}(z) < \rho^{[\widetilde{\jmath},j]}_{(\xi)}\}$ of width $\rho^{[\widetilde{\jmath},j]}_{(\xi)}$ about the real axis in the complex plane, where:

$$\rho^{[\widetilde{\jmath},j]}_{(\xi)} = -\limsup_{k\to\infty} \tfrac{1}{\sqrt{2k+1}} \log\left(\left|\tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k}}{\sqrt{k!\sqrt{\pi}}}\right|\right)$$

A.2 Rectified Activation Functions

Recall that the rectified activation functions are defined as:

$$\hat{\tau}^{[\widetilde{\jmath},j]}_{\eta}(\zeta;\xi,\xi') = \sum_{k=1}^{\infty} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k} a^{[\widetilde{\jmath},j]}_{(\xi')k}}{\eta^{2k}} \sum_{l=1}^{k} \tbinom{k}{l}^2 \mathrm{He}^2_{k-l}\zeta^l$$

where $\eta \in (0,1)$ is fixed. To understand the convergence of this function, observe that:

$$\hat{\tau}^{[\widetilde{\jmath},j]}_{\eta}(\zeta;\xi,\xi') \le \max_{\xi''\in\{\xi,\xi'\}} \sum_{k=1}^{\infty}\sum_{l=1}^{k} \left|\tfrac{a^{[\widetilde{\jmath},j]}_{(\xi'')k}}{\eta^k}\tbinom{k}{l}\mathrm{He}_{k-l}\sqrt{\zeta^l}\right|^2$$

which is the 2-norm of a sequence. Hence:

$$\hat{\tau}^{[\widetilde{\jmath},j]}_{\eta}(\zeta;\xi,\xi') \le \max_{\xi''\in\{\xi,\xi'\}} \left(\sum_{k=1}^{\infty}\sum_{l=1}^{k} \left|\tfrac{\left|a^{[\widetilde{\jmath},j]}_{(\xi'')k}\right|}{\eta^k}\tbinom{k}{l}\left|\mathrm{He}_{k-l}\right|\sqrt{\zeta^l}\right|\right)^2$$

Thus is suffices to study the convergence of:

$$\gamma^{[\widetilde{\jmath},j]}(\lambda;\xi) = \sum_{k=1}^{\infty} \tfrac{\left|a^{[\widetilde{\jmath},j]}_{(\xi)k}\right|}{\eta^k} \sum_{l=1}^{k} \tbinom{k}{l} \left|\mathrm{He}_{k-l}\right| \lambda^l$$

which in turn bounds:

$$\hat{\tau}^{[\widetilde{\jmath},j]}_{\eta}(\zeta;\xi,\xi') \le \max\left\{\gamma^{[\widetilde{\jmath},j]}\left(\sqrt{\zeta};\xi\right)^2, \gamma^{[\widetilde{\jmath},j]}\left(\sqrt{\zeta};\xi'\right)^2\right\}$$

Using (26):

$$\gamma^{[\widetilde{\jmath},j]}(\lambda;\xi) = \mathrm{Re}\left(\overline{\gamma}^{[\widetilde{\jmath},j]}(\mathtt{i}\lambda;\xi)\right) + \mathrm{Im}\left(\overline{\gamma}^{[\widetilde{\jmath},j]}(\mathtt{i}\lambda;\xi)\right)$$

where:

$$\gamma^{[\widetilde{\jmath},j]}(\lambda;\xi) = \sum_{k=1}^{\infty} \tfrac{\left|a^{[\widetilde{\jmath},j]}_{(\xi)k}\right|}{\eta^k} \sum_{l=1}^{k} \tbinom{k}{l} \mathrm{He}_{k-l}\lambda^l$$

which, using (25), is convergent for:

$$|\lambda| < L^{[\widetilde{\jmath},j]}_{(\xi)} < -\limsup_{k\to\infty} \tfrac{1}{\sqrt{2k+1}} \left(\ln\left|\tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k}}{\eta^k}\right|\right)$$

whereon:

$$\left|\gamma^{[\widetilde{\jmath},j]}(\lambda;\xi)\right| \le M^{[\widetilde{\jmath},j]}_{(\xi)} = \gamma^{[\widetilde{\jmath},j]}\left(L^{[\widetilde{\jmath},j]}_{(\xi)};\xi\right)$$

and we conclude that $\hat{\tau}_{\eta}^{[\widetilde{j},j]}(\zeta;\xi,\xi')$ is convergent for:

$$|\zeta| \leq \rho_{(\xi)\eta}^{[\widetilde{j},j]2} = \min\left\{-\limsup_{k\to\infty}\frac{1}{\sqrt{2k+1}}\ln\left|\frac{a_{(\xi)k}^{[\widetilde{j},j]}}{\eta^k}\right|, -\limsup_{k\to\infty}\frac{1}{\sqrt{2k+1}}\ln\left|\frac{a_{(\xi)k}^{[\widetilde{j},j]}}{\eta^k}\right|\right\}^2$$

whereon:

$$\hat{\tau}_{\eta}^{[\widetilde{j},j]}\left(\zeta;\xi,\xi'\right) \leq T_{(\xi)\eta}^{[\widetilde{j},j]2} = \hat{\tau}_{\eta}^{[\widetilde{j},j]}\left(\rho_{(\xi)\eta}^{[\widetilde{j},j]2};\xi,\xi'\right)$$

The envelope is convergent for:

$$|\zeta| < \rho_{(\omega)\eta}^{[\widetilde{j},j]2} = \inf_{|\xi|,|\xi'|\leq\omega}\rho_{(\xi,\xi')\eta}^{[\widetilde{j},j]2}$$

whereon:

$$T_{(\omega)\eta}^{[\widetilde{j},j]2} = \hat{\tau}_{\eta}^{[\widetilde{j},j]}\left(\rho_{(\omega)\eta}^{[\widetilde{j},j]2};\omega\right)$$

Finally:

$$\begin{aligned}\hat{s}_{\eta}\left(\zeta\right) &= \sum_{k=1}^{\infty}\eta^{2k}\sum_{l=1}^{k}\zeta^{l} = \tfrac{\zeta}{1-\zeta}\sum_{k=1}^{\infty}\eta^{2k}\left(1-\zeta^{k}\right) = \tfrac{\zeta}{1-\zeta}\left(\sum_{k=1}^{\infty}\eta^{2k} - \sum_{k=1}^{\infty}(\zeta\eta^2)^{k}\right)\\ &= \tfrac{\zeta}{1-\zeta}\left(\tfrac{\eta^2}{1-\eta^2} - \tfrac{\zeta\eta^2}{1-\zeta\eta^2}\right)\end{aligned}$$

is convergent $\forall|\zeta| < R_{\eta}^2 < 1$, with max value of $S_{\eta}^2 = \frac{R_{\eta}^2}{1-R_{\eta}^2}(\frac{\eta^2}{1-\eta^2} - \frac{\eta^2 R_{\eta}^2}{1-\eta^2 R_{\eta}^2})$ thereon.

A.3 ReLU Activation Function

In this section we derive the Hermite-polynomial expansion of the centered ReLU activation function:

$$\begin{aligned}\overline{\tau}^{[\mathrm{ReLU}]}\left(\zeta;\xi\right) &= \tau^{[\mathrm{ReLU}]}\left(\xi+\zeta\right) - \tau^{[\mathrm{ReLU}]}\left(\xi\right)\\ &= \left\{\begin{array}{ll}\zeta+\xi & \text{if } \zeta > -\xi\\ 0 & \text{otherwise}\end{array}\right. - \left[\xi\right]_{+}\\ &= \sum_{k=0}^{\infty} a_{(\xi)k}^{[\widetilde{j},j]} He_k\left(\zeta\right)\end{aligned}$$

We find it convenient to work in terms of the physicists Hermite polynomials H_k to suit (Gradshteyn & Ryzhik, 2000). Using this:

$$\begin{aligned}a_{(\xi)k}^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{2\pi k!}}\int_{-\xi}^{\infty}\left(\zeta+\xi\right)e^{-\frac{\zeta^2}{2}}He_k\left(\zeta\right)d\zeta - \tfrac{1}{\sqrt{2\pi k!}}\int_{-\infty}^{\infty}\left[\xi\right]_{+}e^{-\frac{\zeta^2}{2}}He_k\left(\zeta\right)d\zeta\\ &= \tfrac{1}{\sqrt{2\pi k!}}\int_{-\xi}^{\infty}\left(\zeta+\xi\right)e^{-\frac{\zeta^2}{2}}\tfrac{1}{\sqrt{2}^k}H_k\left(\tfrac{\zeta}{\sqrt{2}}\right)d\zeta - \tfrac{1}{\sqrt{2\pi k!}}\int_{-\infty}^{\infty}\left[\xi\right]_{+}e^{-\frac{\zeta^2}{2}}\tfrac{1}{\sqrt{2}^k}H_k\left(\tfrac{\zeta}{\sqrt{2}}\right)d\zeta\\ &= \sqrt{\tfrac{2}{\pi}}\tfrac{1}{k!}\int_{-\sqrt{2}\frac{\xi}{\sqrt{2}}}^{\infty}\left(\tfrac{\zeta}{\sqrt{2}}+\tfrac{\xi}{\sqrt{2}}\right)e^{-\left(\frac{\zeta}{\sqrt{2}}\right)^2}\tfrac{1}{\sqrt{2}^k}H_k\left(\tfrac{\zeta}{\sqrt{2}}\right)d\tfrac{\zeta}{\sqrt{2}} - \sqrt{\tfrac{2}{\pi}}\tfrac{1}{k!}\int_{-\infty}^{\infty}\left[\tfrac{\xi}{\sqrt{2}}\right]_{+}e^{-\left(\frac{\zeta}{\sqrt{2}}\right)^2}\tfrac{1}{\sqrt{2}^k}H_k\left(\tfrac{\zeta}{\sqrt{2}}\right)d\tfrac{\zeta}{\sqrt{2}}\\ &= \sqrt{\tfrac{2}{\pi}}\tfrac{1}{\sqrt{2}^k k!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}\left(\zeta+\tfrac{\xi}{\sqrt{2}}\right)e^{-\zeta^2}H_k\left(\zeta\right)d\zeta - \sqrt{\tfrac{2}{\pi}}\tfrac{1}{\sqrt{2}^k k!}\int_{-\infty}^{\infty}\left[\tfrac{\xi}{\sqrt{2}}\right]_{+}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta\\ &= \sqrt{\tfrac{2}{\pi}}\tfrac{1}{\sqrt{2}^k k!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}\zeta H_k\left(\zeta\right)d\zeta + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^k k!}\xi\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta - \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^k k!}\left[\xi\right]_{+}\int_{-\infty}^{\infty}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta\end{aligned}$$

Using the recursion and derivative properties, for $k > 1$:

$$\begin{aligned}\zeta H_k\left(\zeta\right) &= \tfrac{1}{2}H_{k+1}\left(\zeta\right) + \tfrac{1}{2}H_k'\left(\zeta\right)\\ &= \tfrac{1}{2}H_{k+1}\left(\zeta\right) + kH_{k-1}\left(\zeta\right)\end{aligned}$$

and hence, using (Gradshteyn & Ryzhik, 2000, (7.373)):

$$\begin{aligned}a_{(\xi)k}^{[\mathrm{ReLU}]} &= \tfrac{k+1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{k+1}(k+1)!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_{k+1}\left(\zeta\right)d\zeta + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{k-1}(k-1)!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_{k-1}\left(\zeta\right)d\zeta + \ldots\\ &\quad \ldots + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^k k!}\xi\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta - \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^k k!}\left[\xi\right]_{+}\int_{-\infty}^{\infty}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta\\ &= \tfrac{k+1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{k+1}(k+1)!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_{k+1}\left(\zeta\right)d\zeta + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{k-1}(k-1)!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_{k-1}\left(\zeta\right)d\zeta + \ldots\\ &\quad \ldots + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^k k!}\xi\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta - \delta_{k,2p}\tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{2p}(2p)!}\left[\xi\right]_{+}\int_{-\infty}^{\infty}e^{-\zeta^2}H_{2p}\left(\zeta\right)d\zeta\\ &= \tfrac{k+1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{k+1}(k+1)!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_{k+1}\left(\zeta\right)d\zeta + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^{k-1}(k-1)!}\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_{k-1}\left(\zeta\right)d\zeta + \ldots\\ &\quad \ldots + \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2}^k k!}\xi\int_{-\frac{\xi}{\sqrt{2}}}^{\infty}e^{-\zeta^2}H_k\left(\zeta\right)d\zeta\end{aligned}$$

Using (Gradshteyn & Ryzhik, 2000, (7.373)) we have:

$$\begin{aligned}
a_{(\xi)k}^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2^{k+1}(k+1)!}}(k+1)\left(e^{-\frac{\xi^2}{2}}H_k\left(-\tfrac{\xi}{\sqrt{2}}\right)-e^{-\frac{\infty^2}{2}}H_k\left(\tfrac{\infty}{\sqrt{2}}\right)\right)+\dots\\
&\quad\dots\tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2^{k-1}(k-1)!}}\left(e^{-\frac{\xi^2}{2}}H_{k-2}\left(-\tfrac{\xi}{\sqrt{2}}\right)-e^{-\frac{\infty^2}{2}}H_{k-2}\left(\tfrac{\infty}{\sqrt{2}}\right)\right)+\dots\\
&\quad\dots\tfrac{1}{\sqrt{\pi}}\tfrac{1}{\sqrt{2^k k!}}\xi\left(e^{-\frac{\xi^2}{2}}H_{k-1}\left(-\tfrac{\xi}{\sqrt{2}}\right)-e^{-\frac{\infty^2}{2}}H_{k-1}\left(\tfrac{\infty}{\sqrt{2}}\right)\right)\\
&= \tfrac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}\left(\tfrac{k+1}{\sqrt{2^k(k+1)!}}H_k\left(-\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{k-2}(k-1)!}}H_{k-2}\left(-\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{k-1}k!}}\xi H_{k-1}\left(-\tfrac{\xi}{\sqrt{2}}\right)\right)
\end{aligned}$$

If $k=2p$ and $p>0$ then, noting that $H_k(0)=\sqrt{2}^k\mathrm{He}_k$:

$$\begin{aligned}
a_{(\xi)2p}^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}\left(\tfrac{1}{\sqrt{2^{2p}(2p+1)!}}(2p+1)H_{2p}\left(-\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{2p-2}(2p-1)!}}H_{2p-2}\left(-\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{2p-1}(2p)!}}\xi H_{2p-1}\left(-\tfrac{\xi}{\sqrt{2}}\right)\right)\\
&= \tfrac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}\left(\tfrac{1}{\sqrt{2^{2p}(2p+1)!}}(2p+1)H_{2p}\left(\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{2p-2}(2p-1)!}}H_{2p-2}\left(\tfrac{\xi}{\sqrt{2}}\right)-\tfrac{1}{\sqrt{2^{2p-1}(2p)!}}\xi H_{2p-1}\left(\tfrac{\xi}{\sqrt{2}}\right)\right)\\
&= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!}e^{-\frac{\xi^2}{2}}\left(\tfrac{(-1)^{p+1}(2p-1)p!}{(2p)!}H_{2p}\left(\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{(-1)^{p+1}2p!}{(2(p-1))!}H_{2p-2}\left(\tfrac{\xi}{\sqrt{2}}\right)-\tfrac{(-1)^{p+1}(p-1)!}{\sqrt{2}(2(p-1))!}\xi H_{2p-1}\left(\tfrac{\xi}{\sqrt{2}}\right)\right)\\
&\left(= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!}e^{-\frac{\xi^2}{2}}\left(\tfrac{(-1)^{p+1}(2p-1)p!}{(2p)!}H_{2p}(0)+\tfrac{(-1)^{p+1}2p!}{(2(p-1))!}H_{2p-2}(0)\right),\ \text{if } \xi=0\right)\\
&\left(= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!}e^{-\frac{\xi^2}{2}}\left(\tfrac{(-1)^{p+1}(2p-1)p!}{(2p)!}2^p\mathrm{He}_{2p}+\tfrac{(-1)^{p+1}2p!}{(2(p-1))!}2^{p-1}\mathrm{He}_{2p-2}\right),\ \text{if } \xi=0\right)\\
&\left(= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!}e^{-\frac{\xi^2}{2}}\left(\tfrac{(-1)^{p+1}(2p-1)p!}{(2p)!}2^p\tfrac{(-1)^p(2p)!}{2^p p!}+\tfrac{(-1)^{p+1}2p!}{(2(p-1))!}2^{p-1}\tfrac{(-1)^{p+1}(2p-2)!}{(p-1)!2^{p-1}}\right),\ \text{if } \xi=0\right)\\
&\left(= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!},\ \text{if } \xi=0\right)
\end{aligned}$$

If $k=2p+1$ and $p>0$ then:

$$\begin{aligned}
a_{(\xi)2p+1}^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}\left(\tfrac{1}{\sqrt{2^{2p+1}(2p+2)!}}(2p+2)H_{2p+1}\left(-\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{2p-1}(2p)!}}H_{2p-1}\left(-\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{2p-2}(2p-1)!}}\xi H_{2p-2}\left(-\tfrac{\xi}{\sqrt{2}}\right)\right)\\
&= \tfrac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}\left(-\tfrac{1}{\sqrt{2^{2p+1}(2p+2)!}}(2p+2)H_{2p+1}\left(\tfrac{\xi}{\sqrt{2}}\right)-\tfrac{1}{\sqrt{2^{2p-1}(2p)!}}H_{2p-1}\left(\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2^{2p-2}(2p-1)!}}\xi H_{2p-2}\left(\tfrac{\xi}{\sqrt{2}}\right)\right)\\
&= \tfrac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}\left(-\tfrac{\sqrt{2}}{2^{p+1}(2p+1)!}H_{2p+1}\left(\tfrac{\xi}{\sqrt{2}}\right)-\tfrac{\sqrt{2}}{2^p(2p)!}H_{2p-1}\left(\tfrac{\xi}{\sqrt{2}}\right)+\tfrac{1}{2^{p-1}(2p-1)!}\xi H_{2p-2}\left(\tfrac{\xi}{\sqrt{2}}\right)\right)\\
&(=0 \text{ if } \xi=0)
\end{aligned}$$

For the cases $k=0,1$ we use the result:

$$\textstyle\int_a^b x^m e^{-x^2}dx = \frac{1}{2}\Gamma\left(\frac{m+1}{2},a^2\right)-\frac{1}{2}\Gamma\left(\frac{m+1}{2},b^2\right)$$

and so:

$$\textstyle\int_a^\infty x^m e^{-x^2}dx = \frac{1}{2}\Gamma\left(\frac{m+1}{2},a^2\right)$$

In the case $k=0$:

$$\begin{aligned}
a_{(\xi)0}^{[\mathrm{ReLU}]} &= \sqrt{\tfrac{2}{\pi}}\textstyle\int_{-\frac{\xi}{\sqrt{2}}}^\infty \zeta e^{-\zeta^2}d\zeta+\frac{1}{\sqrt{\pi}}\xi\int_{-\frac{\xi}{\sqrt{2}}}^\infty e^{-\zeta^2}d\zeta-[\xi]_+\\
&= \tfrac{1}{\sqrt{2\pi}}\Gamma\left(1,\tfrac{\xi^2}{2}\right)+\tfrac{1}{2\sqrt{\pi}}\xi\Gamma\left(\tfrac{1}{2},\tfrac{\xi^2}{2}\right)-[\xi]_+\\
&\left(= \tfrac{1}{\sqrt{2\pi}}\text{ if } \xi=0\right)
\end{aligned}$$

and in the case $k=1$:

$$\begin{aligned}
a_{(\xi)1}^{[\mathrm{ReLU}]} &= \tfrac{2}{\sqrt{\pi}}\textstyle\int_{-\frac{\xi}{\sqrt{2}}}^\infty \zeta^2 e^{-\zeta^2}d\zeta+\sqrt{\frac{2}{\pi}}\xi\int_{-\frac{\xi}{\sqrt{2}}}^\infty \zeta e^{-\zeta^2}d\zeta\\
&= \tfrac{1}{\sqrt{\pi}}\Gamma\left(\tfrac{3}{2},\tfrac{\xi^2}{2}\right)+\tfrac{1}{\sqrt{2\pi}}\xi\Gamma\left(1,\tfrac{\xi^2}{2}\right)\\
&\left(= \tfrac{1}{2}\text{ if } \xi=0\right)
\end{aligned}$$

Next we derive the magnitude functions for the ReLU. Using integration by parts, we see that:

$$\begin{aligned}
\tfrac{1}{\sqrt{2\pi}}\textstyle\int_c^x \frac{1}{\zeta^2}\left(e^{\frac{1}{2}\zeta^2}-1\right)d\zeta &= \tfrac{1}{\sqrt{2\pi}}\tfrac{1}{\sqrt{2}}\int_c^x\tfrac{2}{\zeta^2}\left(e^{\frac{1}{2}\zeta^2}-1\right)d\tfrac{\zeta}{\sqrt{2}}\\
&= \tfrac{1}{\sqrt{2\pi}}\tfrac{1}{\sqrt{2}}\int_{\frac{c}{\sqrt{2}}}^{\frac{x}{\sqrt{2}}}\tfrac{1}{\zeta^2}\left(e^{\zeta^2}-1\right)d\zeta\\
&= -\tfrac{1}{\sqrt{2\pi}}\tfrac{1}{x}\left(e^{\frac{1}{2}x^2}-1\right)+\tfrac{1}{\sqrt{2\pi}}\tfrac{1}{c}\left(e^{\frac{1}{2}c^2}-1\right)+\tfrac{1}{\sqrt{\pi}}\int_{\frac{c}{\sqrt{2}}}^{\frac{x}{\sqrt{2}}}e^{\zeta^2}d\zeta\\
&= -\tfrac{1}{\sqrt{2\pi}}\tfrac{1}{x}\left(e^{\frac{1}{2}x^2}-1\right)+\tfrac{1}{\sqrt{2\pi}}\tfrac{1}{c}\left(e^{\frac{1}{2}c^2}-1\right)+\tfrac{1}{2}\tfrac{2}{\sqrt{\pi}}\int_{\frac{c}{\sqrt{2}}}^{\frac{x}{\sqrt{2}}}e^{\zeta^2}d\zeta\\
&= -\tfrac{1}{\sqrt{2\pi}}\tfrac{1}{x}\left(e^{\frac{1}{2}x^2}-1\right)+\tfrac{1}{2}\mathrm{erfi}\left(\tfrac{x}{\sqrt{2}}\right)-\tfrac{1}{2}\left(\mathrm{erfi}\left(\tfrac{c}{\sqrt{2}}\right)-\tfrac{1}{\sqrt{2\pi}}\tfrac{2}{c}\left(e^{\frac{1}{2}c^2}-1\right)\right)
\end{aligned}$$

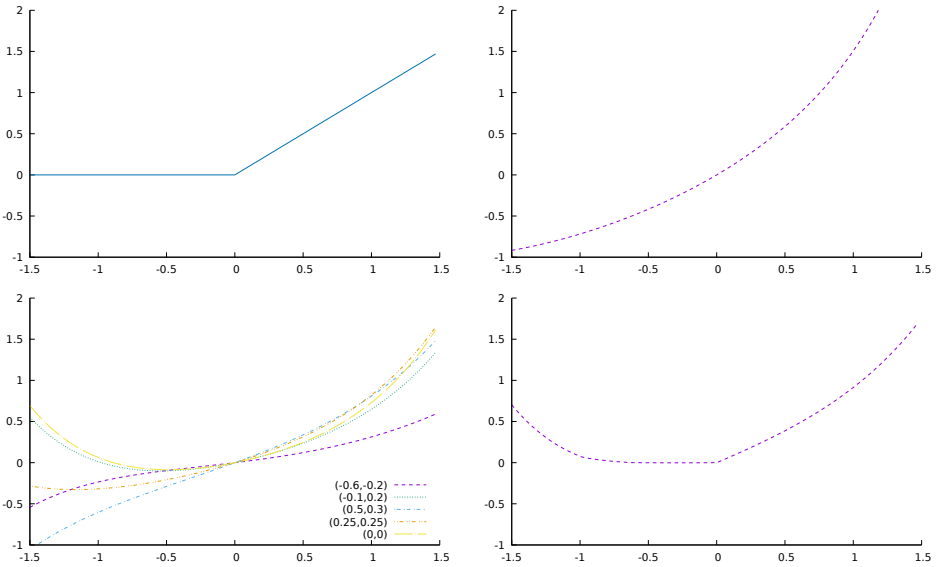

Figure 3: The ReLU magnitude activation and associated functions - ReLU activation $\tau^{[\mathrm{ReLU}]}$ (top left), magnitude $\overline{s}^{[\mathrm{ReLU}]}$ (top right), rectified activation $\hat{\tau}_{\eta}^{[\mathrm{ReLU}]}$ for various ξ, ξ' pairs (bottom left) and its envelope (bottom right) for $\xi, \xi' \in [-2, 2]$. Note that $\eta = 0.75$ for all plots.

So:

$$\begin{aligned}
\sum_{k=1}^{\infty} \left| a_k^{[\mathrm{ReLU}]} \right| x^k &= \tfrac{1}{2}x + \tfrac{1}{\sqrt{2\pi}} \sum_{p=1}^{\infty} \tfrac{x^{2p}}{(2p-1)2^p p!} \\
&= \tfrac{1}{2}x + \tfrac{1}{\sqrt{2\pi}} x \sum_{p=1}^{\infty} \tfrac{x^{2p-1}}{(2p-1)2^p p!} \\
&= \tfrac{1}{2}x + \tfrac{1}{\sqrt{2\pi}} x \int_c^x \left(\tfrac{\partial}{\partial \zeta} \sum_{p=1}^{\infty} \tfrac{\zeta^{2p-1}}{(2p-1)2^p p!} \right) d\zeta \\
&= \tfrac{1}{2}x + \tfrac{1}{\sqrt{2\pi}} x \int_c^x \left(\sum_{p=1}^{\infty} \tfrac{\zeta^{2p-2}}{2^p p!} \right) d\zeta \\
&= \tfrac{1}{2}x + \tfrac{1}{2\sqrt{2\pi}} x \int_c^x \left(\sum_{p=1}^{\infty} \tfrac{1}{p!} \left(\tfrac{1}{2}\zeta^2 \right)^{p-1} \right) d\zeta \\
&= \tfrac{1}{2}x + \tfrac{1}{\sqrt{2\pi}} x \int_c^x \tfrac{1}{\zeta^2} \left(\sum_{p=1}^{\infty} \tfrac{1}{p!} \left(\tfrac{1}{2}\zeta^2 \right)^{p} \right) d\zeta \\
&= \tfrac{1}{2}x + \tfrac{1}{\sqrt{2\pi}} x \int_c^x \tfrac{1}{\zeta^2} \left(e^{\frac{1}{2}\zeta^2} - 1 \right) d\zeta \\
&= \tfrac{1}{2}x \left(\mathrm{erfi}\left(\tfrac{x}{\sqrt{2}} \right) + 1 - \mathrm{erfi}\left(\tfrac{c}{\sqrt{2}} \right) + \tfrac{1}{\sqrt{2\pi}} \tfrac{2}{c} \left(e^{\frac{1}{2}c^2} - 1 \right) \right) + \tfrac{1}{\sqrt{2\pi}} \left(1 - e^{\frac{1}{2}x^2} \right)
\end{aligned}$$

Select c so that the first derivative is $\frac{1}{2}x$:

$$-\mathrm{erfi}\left(\tfrac{c}{\sqrt{2}} \right) + \tfrac{1}{\sqrt{2\pi}} \tfrac{2}{c} \left(e^{\frac{1}{2}c^2} - 1 \right) = 0 \text{ if } c = 0$$

Hence:

$$\begin{aligned}
\overline{s}^{[\mathrm{ReLU}]}(x) &\triangleq \sum_{k=1}^{\infty} \left| a_k^{[\mathrm{ReLU}]} \right| (1+x)^k - \sum_{k=1}^{\infty} \left| a_k^{[\mathrm{ReLU}]} \right| \\
&= \tfrac{1}{2}(1+x) \left(\mathrm{erfi}\left(\tfrac{1+x}{\sqrt{2}} \right) + 1 \right) + \tfrac{1}{\sqrt{2\pi}} \left(1 - e^{\frac{1}{2}(1+x)^2} \right) - \tfrac{1}{2} \left(\mathrm{erfi}\left(\tfrac{1}{\sqrt{2}} \right) + 1 \right) - \tfrac{1}{\sqrt{2\pi}} \left(1 - e^{\frac{1}{2}} \right) \\
&= \tfrac{1}{2}x \left(\mathrm{erfi}\left(\tfrac{1+x}{\sqrt{2}} \right) + 1 \right) + \tfrac{1}{\sqrt{2\pi}} \left(e^{\frac{1}{2}} - e^{\frac{1}{2}(1+x)^2} \right) + \tfrac{1}{2} \left(\mathrm{erfi}\left(\tfrac{1+x}{\sqrt{2}} \right) - \mathrm{erfi}\left(\tfrac{1}{\sqrt{2}} \right) \right)
\end{aligned} \tag{28}$$

B PROOF OF THE GLOBAL DUAL MODEL

Here we prove the validity of the global dual model presented in the paper. Recall that the dual model has the form (Theorem 1, equation (10)):

$$\mathbf{f}(\mathbf{x};\Theta) = \langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\mathbf{x})]_{\mathbf{g}} \tag{29}$$

where, as per Figure 1, $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{[D-1]}$, $\boldsymbol{\phi} = \boldsymbol{\phi}^{[D-1]}$, $\mathbf{g} = \mathbf{g}^{[D-1]}$ and, given the base case $\boldsymbol{\Psi}^{[-1]}(\Theta) = \mathbf{1}_n$, $\boldsymbol{\phi}^{[-1]}(\mathbf{x}) = \mathbf{x}$, $\mathbf{g}^{[-1]} = \mathbf{1}_n$, the recursive definition of the feature maps and metric is proposed:

$$\begin{aligned}
\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}_{:i_{\widetilde{\jmath}}}(\Theta) &= \tilde{\phi}\left[\left|a^{[\widetilde{\jmath},j]}_{(0)k}\right|^{\frac{1}{2}}\left[\binom{k}{l}^{\frac{1}{2}}\left(\sqrt{\frac{1}{\bar{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}}\boldsymbol{\Psi}^{[\widetilde{\jmath}]}_{:i_{\widetilde{\jmath}}}(\Theta)\right)^{\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1} \\
\widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}(\mathbf{x}) &= \frac{1}{\tilde{\phi}}\left[a^{[\widetilde{\jmath},j]\langle\frac{1}{2}\rangle}_{(0)k}\left[\binom{k}{l}^{\frac{1}{2}}\left(\sqrt{\bar{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\mathbf{x})\right)^{\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1} \\
\tilde{\mathbf{g}}^{[\widetilde{\jmath},j]} &= \left[\left[\mathrm{He}_{k-l}\mathbf{g}^{[\widetilde{\jmath}]\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1}
\end{aligned} \tag{30}$$

$\forall \widetilde{\jmath} \in \widetilde{\mathbb{P}}^{[j]}$ (the feature map transforms associated with the edges of the graph) and:

$$\begin{aligned}
\boldsymbol{\Psi}^{[j]}(\Theta) &= \sqrt{\gamma^2+1}\begin{bmatrix} \mathbf{b}^{[j]\mathrm{T}} + \boldsymbol{\upsilon}^{[j]\mathrm{T}}_{\tau} \\ \underset{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}{\mathrm{diag}}\left(\sqrt{\frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}}\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}(\Theta)\,\mathbf{W}^{[\widetilde{\jmath},j]}\right)\end{bmatrix} \\
\boldsymbol{\phi}^{[j]}(\mathbf{x}) &= \sqrt{\frac{1}{\gamma^2+1}}\begin{bmatrix} \gamma \\ \left[\sqrt{\frac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}}\widetilde{\boldsymbol{\Phi}}^{[\widetilde{\jmath},j]}(\mathbf{x})\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\end{bmatrix} \\
\mathbf{g}^{[j]} &= \begin{bmatrix} 1 \\ \left[\tilde{\mathbf{g}}^{[\widetilde{\jmath},j]}\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\end{bmatrix}
\end{aligned} \tag{31}$$

(the feature map transforms associated with the nodes of the graph) where $\boldsymbol{\upsilon}^{[j]}_{\tau} = \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \frac{\tau^{[\widetilde{\jmath},j]}(0)}{\gamma}\mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\mathbf{1}_{H^{[\widetilde{\jmath}]}}$. Our approach to demonstrating that this is true is to prove that, given some input $\mathbf{x} \in \mathbb{X}$ then, for all edges $(\widetilde{\jmath} \to j)$:

$$\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} - \mathbf{1}_{H^{[\widetilde{\jmath}]}}\tau^{[\widetilde{\jmath},j]}(0) = \left\langle \widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}(\Theta), \widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}(\mathbf{x}) \right]_{\tilde{\mathbf{g}}^{[\widetilde{\jmath},j]}} \tag{32}$$

and likewise for all nodes j:

$$\mathbf{x}^{[j]} = \left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\mathbf{x}) \right]_{\mathbf{g}^{[j]}} \tag{33}$$

Base case: By the definition of the base case, we have:

$$\mathbf{x}^{[-1]} = \left\langle \boldsymbol{\Psi}^{[-1]}(\Theta), \boldsymbol{\phi}^{[-1]}(\mathbf{x}) \right]_{\mathbf{g}^{[-1]}} = \mathbf{x}$$

Node case: Assume (32) is true. Then, using (31), we have that:

$$\begin{aligned}
\left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\mathbf{x}) \right]_{\mathbf{g}^{[j]}} &= \gamma\mathbf{b}^{[j]} + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\mathbf{1}_{H^{[\widetilde{\jmath}]}}\tau^{[\widetilde{\jmath},j]}(0) + \mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} - \mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\mathbf{1}_{H^{[\widetilde{\jmath}]}}\tau^{[\widetilde{\jmath},j]}(0)\right) \\
&= \gamma\mathbf{b}^{[j]} + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} \\
&= \mathbf{x}^{[j]}
\end{aligned}$$

Edge case: Assume (33) is true. Then, using (30), we have that:

$$\begin{aligned}
\left\langle \widetilde{\mathbf{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right), \widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\rangle_{\tilde{\mathbf{g}}^{[\widetilde{\jmath},j]}} &= \left[\left\langle \widetilde{\mathbf{\Psi}}^{[\widetilde{\jmath},j]}_{:i_{\widetilde{\jmath}}}\left(\Theta\right), \widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\rangle_{\tilde{\mathbf{g}}^{[\widetilde{\jmath},j]}}\right]_{i_{\widetilde{\jmath}}} \\
&= \left[\sum_{k\geq 1} a^{[\widetilde{\jmath},j]}_{(0)k} \sum_{1\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l}\left\langle \mathbf{\Psi}^{[j]}_{:i_{\widetilde{\jmath}}}\left(\Theta\right)^{\otimes l}, \boldsymbol{\phi}^{[j]}\left(\mathbf{x}\right)^{\otimes l}\right\rangle_{\mathbf{g}^{[j]\otimes l}}\right]_{i_{\widetilde{\jmath}}} \\
&= \left[\sum_{k\geq 1} a^{[\widetilde{\jmath},j]}_{(0)k} \sum_{1\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l}\left\langle \mathbf{\Psi}^{[j]}_{:i_{\widetilde{\jmath}}}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\mathbf{x}\right)\right\rangle^{l}_{\mathbf{g}^{[j]}}\right]_{i_{\widetilde{\jmath}}} \\
&= \left[\sum_{k\geq 1} a^{[\widetilde{\jmath},j]}_{(0)k} \sum_{1\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l} x^{[\widetilde{\jmath}]l}_{i_{\widetilde{\jmath}}}\right]_{i_{\widetilde{\jmath}}} \\
&= \left[\sum_{k\geq 0} a^{[\widetilde{\jmath},j]}_{(0)k} \sum_{0\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l} x^{[\widetilde{\jmath}]l}_{i_{\widetilde{\jmath}}} - \sum_{k\geq 0} a^{[\widetilde{\jmath},j]}_{(0)k} \sum_{0\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l} 0^{l}\right]_{i_{\widetilde{\jmath}}} \\
&= \left[\tau^{[\widetilde{\jmath},j]}\left(x^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right) - \tau^{[\widetilde{\jmath},j]}\left(0\right)\right]_{i_{\widetilde{\jmath}}} \\
&= \tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} - \mathbf{1}_{H^{[\widetilde{\jmath}]}} \tau^{[\widetilde{\jmath},j]}\left(0\right)
\end{aligned}$$

The desired result (29) then follows by identifying the output node $j = D - 1$.

B.1 Norm-Bounds for the Global Dual Model

Our proof of the norm-bounds of the global model follows the same model as our proof of the validity of said model. We want to prove the bounds

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\|^2_2 &\in \left[\tilde{\phi}^{[\widetilde{\jmath},j]2}_{\downarrow} = \tfrac{1}{\tilde{\phi}^2}\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\phi^{[\widetilde{\jmath}]2}_{\downarrow}\right), \tilde{\phi}^{[\widetilde{\jmath},j]2} = 1\right] \\
\left\|\widetilde{\mathbf{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\right\|^2_2 &\leq \tilde{\psi}^{[\widetilde{\jmath},j]2} = \tilde{\phi}^2\overline{s}^{[\widetilde{\jmath},j]}\left(\tfrac{1}{\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}\psi^{[\widetilde{\jmath}]2}\right) \\
\left\|\widetilde{\mathbf{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\right\|^2_{\mathrm{He}[\tau]} &\leq \underset{\sim}{\tilde{\psi}}^{[\widetilde{\jmath},j]2} = \sup_{\substack{\phi^{[\widetilde{\jmath}]}_{\downarrow}\leq\phi^{[\widetilde{\jmath}]}\leq 1 \\ -\underset{\sim}{\psi}^{[\widetilde{\jmath}]}\leq\underset{\sim}{\psi}^{[\widetilde{\jmath}]}\leq\underset{\sim}{\psi}^{[\widetilde{\jmath}]}}} \left\{\frac{\tilde{\phi}^2\tau^{[\widetilde{\jmath},j]}\left(\phi^{[\widetilde{\jmath}]}\underset{\sim}{\psi}^{[\widetilde{\jmath}]}\right)^2}{\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\phi^{[\widetilde{\jmath}]2}\right)}\right\}
\end{aligned} \tag{34}$$

$$\begin{aligned}
\left\|\boldsymbol{\phi}^{[j]}\left(\mathbf{x}\right)\right\|^2_2 &\in \left[\phi^{[j]2}_{\downarrow} = \tfrac{1}{\gamma^2+1}\left(\gamma^2 + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}\tilde{\phi}^{[\widetilde{\jmath},j]2}_{\downarrow}\right), \phi^{[j]2} = 1\right] \\
\left\|\mathbf{\Psi}^{[j]}\left(\Theta\right)\right\|^2_2 &\leq \psi^{[j]2} = \left(\gamma^2+1\right)\left(\left(\beta^{[j]} + \tfrac{\mu^{[j]}\left|\tau^{[\widetilde{\jmath},j]}(0)\right|}{\gamma}\right)^2 + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}\tilde{\psi}^{[\widetilde{\jmath},j]2}\mu^{[\widetilde{\jmath},j]2}\right) \\
\left\|\mathbf{\Psi}^{[j]}\left(\Theta\right)\right\|^2_{\mathrm{He}[\tau]} &\leq \underset{\sim}{\psi}^{[j]2} = \left(\gamma^2+1\right)\left(\left(\beta^{[j]} + \tfrac{\mu^{[j]}\left|\tau^{[\widetilde{\jmath},j]}(0)\right|}{\gamma}\right)^2 + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}\underset{\sim}{\tilde{\psi}}^{[\widetilde{\jmath},j]2}\mu^{[\widetilde{\jmath},j]2}\right)
\end{aligned} \tag{35}$$

with the base-cases $\|\boldsymbol{\phi}^{[-1]}(\mathbf{x})\|^2_2 \in [\phi^{[-1]2}_{\downarrow} = 0, \phi^{[-1]2} = 1]$, $\|\mathbf{\Psi}^{[-1]}(\Theta)\|^2_2 \leq \psi^{[-1]2} = 1$ and $\|\mathbf{\Psi}^{[-1]}(\Theta)\|^2_{\mathrm{He}[\tau]} \leq \underset{\sim}{\psi}^{[-1]2} = 1$. We proceed as follows:

Base case: By the definition of the base case, using our assumptions, we have:

$$\left\|\boldsymbol{\phi}^{[-1]}\left(\mathbf{x}\right)\right\|^2_2 = \|\mathbf{x}\|^2_2 \leq 1 = \phi^{[-1]2}$$

$$\left\|\mathbf{\Psi}^{[-1]}\left(\Theta\right)\right\|^2_2 = \|\mathbf{I}_n\|^2_2 \leq 1 = \psi^{[-1]2}$$

$$\left\|\mathbf{\Psi}^{[-1]}\left(\Theta\right)\right\|^2_{\mathrm{He}[\tau]} = \sup_{\mathbf{x}\in\mathbb{X}}\left\|\left\langle \frac{\boldsymbol{\phi}^{[-1]}(\mathbf{x})}{\left\|\boldsymbol{\phi}^{[-1]}(\mathbf{x})\right\|_2}, \mathbf{\Psi}^{[-1]}\left(\Theta\right)\right\rangle_{\mathbf{g}}\right\|^2_2 = \sup_{\mathbf{x}\in\mathbb{X}}\left\|\frac{\boldsymbol{\phi}^{[-1]}(\mathbf{x})}{\left\|\boldsymbol{\phi}^{[-1]}(\mathbf{x})\right\|_2}\right\|^2_2 \leq 1 = \underset{\sim}{\psi}^{[-1]2}$$

Node case: Assume (34) is true. Then, using our assumptions, we have that:

$$\begin{aligned}
\left\|\boldsymbol{\phi}^{[j]}\left(\mathbf{x}\right)\right\|_2^2 &= \tfrac{1}{\gamma^2+1}\left\|\begin{bmatrix} \gamma \\ \left[\sqrt{\frac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}}\widetilde{\boldsymbol{\Phi}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]} \end{bmatrix}\right\|_2^2 \\
&= \tfrac{1}{\gamma^2+1}\left(\gamma^2+\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}\tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}\left\|\widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\|_2^2\right) \\
&\in \left[\phi_{\downarrow}^{[j]2} = \tfrac{1}{\gamma^2+1}\left(\gamma^2+\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}\tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}\tilde{\phi}_{\downarrow}^{[\widetilde{\jmath},j]2}\right), \phi^{[j]2}=1\right] \\
\left\|\boldsymbol{\Psi}^{[j]}\left(\Theta\right)\right\|_2^2 &= \left(\gamma^2+1\right)\left\|\begin{bmatrix} \mathbf{b}^{[j]\mathrm{T}}+\boldsymbol{\upsilon}_{\tau}^{[j]\mathrm{T}} \\ \underset{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}{\operatorname{diag}}\left(\sqrt{\frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}}\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath},j]}\right) \end{bmatrix}\right\|_2^2 \\
&\leq \left(\gamma^2+1\right)\left(\left(\left\|\mathbf{b}^{[j]}\right\|_2+\left\|\boldsymbol{\upsilon}_{\tau}^{[j]}\right\|_2\right)^2+\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}\tfrac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}\left\|\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath},j]}\right\|_2^2\right) \\
&\leq \left(\gamma^2+1\right)\left(\left(\beta^{[j]}+\tfrac{\mu^{[j]}\left|\tau^{[\widetilde{\jmath},j]}(0)\right|}{\gamma}\right)^2+\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}\tfrac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}\widetilde{\psi}^{[\widetilde{\jmath},j]2}\mu^{[\widetilde{\jmath},j]2}\right) = \psi^{[j]2} \\
\left\|\boldsymbol{\Psi}^{[j]}\left(\Theta\right)\right\|_{\mathrm{He}[\tau]}^2 &= \left(\gamma^2+1\right)\left\|\begin{bmatrix} \mathbf{b}^{[j]\mathrm{T}}+\boldsymbol{\upsilon}_{\tau}^{[j]\mathrm{T}} \\ \underset{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}{\operatorname{diag}}\left(\sqrt{\frac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}}\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath},j]}\right) \end{bmatrix}\right\|_2^2 \\
&\leq \left(\gamma^2+1\right)\left(\left(\left\|\mathbf{b}^{[j]}\right\|_2+\left\|\boldsymbol{\upsilon}_{\tau}^{[j]}\right\|_2\right)^2+\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}\tfrac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}\left\|\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath},j]}\right\|_{\mathrm{He}[\tau]}^2\right) \\
&\leq \left(\gamma^2+1\right)\left(\left(\beta^{[j]}+\tfrac{\mu^{[j]}\left|\tau^{[\widetilde{\jmath},j]}(0)\right|}{\gamma}\right)^2+\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}[j]}\tfrac{\widetilde{H}^{[j]}}{H^{[\widetilde{\jmath}]}}\underset{\sim}{\psi}^{[\widetilde{\jmath},j]2}\mu^{[\widetilde{\jmath},j]2}\right) = \underset{\sim}{\psi}^{[j]2}
\end{aligned}$$

Edge case: Assume (35) is true. Then, using our assumptions and the definition (12), and using that the magnitude function is increasing on $\mathbb{R}_+$, we have that:

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\phi}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\|_2^2 &= \tfrac{1}{\tilde{\phi}^2}\left\|\left[a_{(0)k}^{[\widetilde{\jmath},j]\langle\frac{1}{2}\rangle}\left[\tbinom{k}{l}^{\frac{1}{2}}\left(\sqrt{\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}\boldsymbol{\phi}^{[\widetilde{\jmath}]}\left(\mathbf{x}\right)\right)^{\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1}\right\|_2^2 \\
&= \tfrac{1}{\tilde{\phi}^2}\sum_{k\geq 1}\left|a_{(0)k}^{[\widetilde{\jmath},j]}\right|\sum_{1\leq l\leq k}\tbinom{k}{l}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}\left(\mathbf{x}\right)\right\|_2^2\right)^l \\
&= \tfrac{1}{\tilde{\phi}^2}\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}\left(\mathbf{x}\right)\right\|_2^2\right) \\
&\in \left[\tfrac{1}{\tilde{\phi}^2}\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\phi_{\downarrow}^{[\widetilde{\jmath}]2}\right), \tfrac{1}{\tilde{\phi}^2}\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\phi^{[\widetilde{\jmath}]2}\right)\right] \\
&\in \left[\tfrac{1}{\tilde{\phi}^2}\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\phi_{\downarrow}^{[\widetilde{\jmath}]2}\right) = \tilde{\phi}_{\downarrow}^{[\widetilde{\jmath},j]2}, \tilde{\phi}^{[\widetilde{\jmath},j]2}=1\right] \\
\left\|\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\right\|_2^2 &= \tilde{\phi}^2\max_{i_{\widetilde{\jmath}}}\left\|\left[\left|a_{(0)k}^{[\widetilde{\jmath},j]}\right|^{\frac{1}{2}}\left[\tbinom{k}{l}^{\frac{1}{2}}\left(\sqrt{\frac{1}{\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}}\boldsymbol{\Psi}_{:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right)^{\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1}\right\|_2^2 \\
&= \tilde{\phi}^2\max_{i_{\widetilde{\jmath}}}\overline{s}^{[\widetilde{\jmath},j]}\left(\tfrac{1}{\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}\left\|\boldsymbol{\Psi}_{:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_2^2\right) \\
&= \tilde{\phi}^2\overline{s}^{[\widetilde{\jmath},j]}\left(\tfrac{1}{\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_2^2\right) \\
&\leq \tilde{\phi}^2\overline{s}^{[\widetilde{\jmath},j]}\left(\tfrac{1}{\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)}\psi^{[\widetilde{\jmath}]2}\right) \\
\left\|\widetilde{\boldsymbol{\Psi}}^{[\widetilde{\jmath},j]}\left(\Theta\right)\right\|_{\mathrm{He}[\tau]}^2 &\leq \underset{\sim}{\tilde{\psi}}^{[\widetilde{\jmath},j]2} = \sup_{\substack{\phi_{\downarrow}^{[\widetilde{\jmath}]}\leq\phi^{[\widetilde{\jmath}]}\leq 1 \\ -\underset{\sim}{\psi}^{[\widetilde{\jmath}]}\leq\underset{\sim}{\psi}^{[\widetilde{\jmath}]}\leq\underset{\sim}{\psi}^{[\widetilde{\jmath}]}}}\left\{\frac{\tilde{\phi}^2\tau^{[\widetilde{\jmath},j]}\left(\phi^{[\widetilde{\jmath}]}\underset{\sim}{\psi}^{[\widetilde{\jmath}]}\right)^2}{\overline{s}^{[\widetilde{\jmath},j]}\left(\overline{s}^{[\widetilde{\jmath},j]-1}\left(\tilde{\phi}^2\right)\phi^{[\widetilde{\jmath}]2}\right)}\right\}
\end{aligned}$$

where the final bound in this sequence is simply the definition of the norm in question with the range of the supremum expanded to the known bound on this range.

The desired result follows by identifying the output node $j = D-1$. We note that the bounds $\tilde{\phi} \in \mathbb{R}_+$ may be chosen arbitrarily here.

C Proof of the Local Dual Model

We now repeat the proof from the previous section B for the local model. Recall that the dual model has the form (Theorem 1, equation (10)):

$$\begin{aligned}\Delta\mathbf{f}\left(\mathbf{x};\Delta\Theta\right) &= \left\langle \boldsymbol{\Psi}_{\Delta}\left(\Delta\Theta\right), \boldsymbol{\phi}_{\Delta}\left(\mathbf{x}\right)\right]_{\mathbf{G}_{\Delta}(\mathbf{x})} \\ &= \left[\left\langle \boldsymbol{\Psi}_{\Delta:i_{D-1}}\left(\Delta\Theta\right), \boldsymbol{\phi}_{\Delta}\left(\mathbf{x}\right)\right]_{\mathbf{G}_{\Delta:i_{D-1}}(\mathbf{x})}\right]_{i_{D-1}}\end{aligned} \tag{36}$$

where, as per Figure 2, $\boldsymbol{\Psi}_{\Delta} = \boldsymbol{\Psi}_{\Delta}^{[D-1]}$, $\boldsymbol{\phi}_{\Delta} = \boldsymbol{\phi}_{\Delta}^{[D-1]}$, $\mathbf{G}_{\Delta} = \mathbf{G}_{\Delta}^{[D-1]}$ and, given the base case $\boldsymbol{\Psi}_{\Delta}^{[-1]}(\Delta\Theta) = \mathbf{0}_{0\times n}$, $\boldsymbol{\phi}_{\Delta}^{[-1]}(\mathbf{x}) = \mathbf{0}_0$ $\mathbf{G}_{\Delta}^{[-1]}(\mathbf{x}) = \mathbf{1}_{0\times n}$ the recursive definition of the feature maps is proposed as:

$$\begin{aligned}\widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right) &= \left[\frac{1}{\eta^k}\left[\left(\sqrt{\frac{\rho_{(\tilde{\omega})\eta}^{[\widetilde{\jmath},j]2}}{\gamma^2+1}}\boldsymbol{\phi}_{\Delta}^{[\widetilde{\jmath}]}\left(\mathbf{x}\right)\right)^{\otimes l}\right]_{1\le l\le k}\right]_{k\ge 1} \\ \widetilde{\boldsymbol{\Psi}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}\left(\Theta\right) &= \left[\eta^k\left[\left(\sqrt{\frac{\gamma^2+1}{\rho_{(\tilde{\omega})\eta}^{[\widetilde{\jmath},j]2}}}\boldsymbol{\Psi}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right)^{\otimes l}\right]_{1\le l\le k}\right]_{k\ge 1} \\ \widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right) &= \left[a_{\binom{x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}}{k}}^{[\widetilde{\jmath},j]}\left[\binom{k}{l}\mathrm{He}_{k-l}\mathbf{G}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\left(\mathbf{x}\right)^{\otimes l}\right]_{1\le l\le k}\right]_{k\ge 1}\end{aligned} \tag{37}$$

$\forall\widetilde{\jmath} \in \widetilde{\mathbb{P}}^{[j]}$ (the feature map transforms associated with the edges of the graph) and:

$$\begin{aligned}\boldsymbol{\phi}_{\Delta}^{[j]}\left(\mathbf{x}\right) &= \begin{bmatrix}\gamma \\ \frac{1}{\sqrt{2\tilde{p}^{[j]}}}\left[\frac{1}{\tilde{\omega}^{[\widetilde{\jmath},j]}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \\ \frac{1}{\sqrt{2\tilde{p}^{[j]}}}\left[\frac{1}{\tilde{\psi}_{\Delta}^{[\widetilde{\jmath},j]}}\widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\end{bmatrix} \\ \boldsymbol{\Psi}_{\Delta}^{[j]}\left(\Theta\right) &= \begin{bmatrix}\Delta\mathbf{b}^{[j]\mathrm{T}} \\ \sqrt{2\tilde{p}^{[j]}}\operatorname*{diag}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\tilde{\omega}^{[\widetilde{\jmath},j]}\mathbf{I}_{H^{[\widetilde{\jmath}]}}\right)\Delta\mathbf{W}^{[j]} \\ \sqrt{2\tilde{p}^{[j]}}\operatorname*{diag}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\tilde{\psi}_{\Delta}^{[\widetilde{\jmath},j]}\widetilde{\boldsymbol{\Psi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\Theta\right)\right)\left(\mathbf{W}^{[j]}+\Delta\mathbf{W}^{[j]}\right)\end{bmatrix} \\ \mathbf{G}_{\Delta}^{[j]}\left(\mathbf{x}\right) &= \begin{bmatrix}\mathbf{1}_{H^{[\widetilde{\jmath}]}}^{\mathrm{T}} \\ \operatorname*{diag}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\mathbf{I}_{H^{[\widetilde{\jmath}]}}\right)\mathbf{1}_{\tilde{H}^{[j]}}\mathbf{1}_{H^{[j]}}^{\mathrm{T}} \\ \operatorname*{diag}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\widetilde{\mathbf{G}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right)\mathbf{1}_{\tilde{H}^{[j]}}\mathbf{1}_{H^{[j]}}^{\mathrm{T}}\end{bmatrix}\end{aligned} \tag{38}$$

(the feature map transforms associated with the nodes of the graph). Our approach to demonstrating that this is true is to prove that, given some input $\mathbf{x} \in \mathbb{X}$ then, for all edges $(\widetilde{\jmath} \to j)$:

$$\Delta\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} = \left\langle \widetilde{\boldsymbol{\Psi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\Delta\Theta\right), \widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right]_{\widetilde{\mathbf{G}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x})} \tag{39}$$

and likewise for all nodes j:

$$\Delta\mathbf{x}^{[j]} = \left\langle \boldsymbol{\Psi}_{\Delta}^{[j]}\left(\Delta\Theta\right), \boldsymbol{\phi}_{\Delta}^{[j]}\left(\mathbf{x}\right)\right]_{\mathbf{G}_{\Delta}^{[j]}(\mathbf{x})} \tag{40}$$

Base case: By the definition of the base case, we have:

$$\Delta\mathbf{x}^{[-1]} = \left\langle \boldsymbol{\Psi}_{\Delta}^{[-1]}\left(\Delta\Theta\right), \boldsymbol{\phi}_{\Delta}^{[-1]}\left(\mathbf{x}\right)\right\rangle_{\mathbf{G}_{\Delta}^{[-1]}(\mathbf{x})} = \mathbf{0}_n$$

Node case: Assume (39) is true. Then, using (38), we have that:

$$\begin{aligned}
&\left\langle \boldsymbol{\Psi}_{\Delta}^{[j]}\left(\Delta\Theta\right), \boldsymbol{\phi}_{\Delta}^{[j]}\left(\mathbf{x}\right)\right\rangle_{\mathbf{G}_{\Delta}^{[j]}(\mathbf{x})} = \gamma\Delta\mathbf{b}^{[j]} + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\Delta\mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} + \mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\Delta\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} + \Delta\mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\Delta\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\right)\\
&= \gamma\left(\mathbf{b}^{[j]} + \Delta\mathbf{b}^{[j]}\right) + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\left(\mathbf{W}^{[\widetilde{\jmath},j]} + \Delta\mathbf{W}^{[\widetilde{\jmath},j]}\right)^{\mathrm{T}}\left(\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]} + \Delta\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\right) - \gamma\mathbf{b}^{[j]} - \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\mathbf{W}^{[\widetilde{\jmath},j]\mathrm{T}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\\
&= \left(\mathbf{x}^{[j]} + \Delta\mathbf{x}^{[j]}\right) - \mathbf{x}^{[j]} = \Delta\mathbf{x}^{[j]}
\end{aligned}$$

Edge case: Assume (40) is true. Then, using (37) and (27), we have that:

$$\begin{aligned}
\left\langle \widetilde{\boldsymbol{\Psi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\Delta\Theta\right), \widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\rangle_{\widetilde{\mathbf{G}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x})} &= \left[\sum_{k\geq 1} a_{\binom{x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}}{k}}^{[\widetilde{\jmath},j]2} \sum_{1\leq l\leq k}\binom{k}{l}\mathrm{He}_{k-l}\left\langle \boldsymbol{\Psi}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\left(\Delta\Theta\right), \boldsymbol{\phi}_{\Delta}^{[\widetilde{\jmath}]}\left(\mathbf{x}\right)\right\rangle_{\widetilde{\mathbf{G}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x})}^{l}\right]_{i_{\widetilde{\jmath}}}\\
&= \left[\sum_{k\geq 1} a_{\binom{x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}}{k}}^{[\widetilde{\jmath},j]} \sum_{1\leq l\leq k}\binom{k}{l}\mathrm{He}_{k-l}\Delta x_{i_{\widetilde{\jmath}}}^{[j]l}\right]_{i_{\widetilde{\jmath}}}\\
&= \left[\overline{\tau}^{[\widetilde{\jmath},j]}\left(\Delta x_{i_{\widetilde{\jmath}}}^{[j]}; x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\right)\right]_{i_{\widetilde{\jmath}}}\\
&= \left[\tau^{[\widetilde{\jmath},j]}\left(x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]} + \Delta x_{i_{\widetilde{\jmath}}}^{[j]}\right) - \tau^{[\widetilde{\jmath},j]}\left(x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\right)\right]_{i_{\widetilde{\jmath}}}\\
&= \Delta\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}
\end{aligned}$$

The desired result (36) then follows by identifying the output node $j = D-1$.

C.1 Norm-Bounds for the Local Dual Model

Our proof of the norm-bounds of the local model follows the same model as our proof of the validity of said model. We want to prove the bounds

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\odot\widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\|_2^2 &\leq \tilde{\phi}_{\Delta}^{[\widetilde{\jmath},j]2} = T_{(\tilde{\omega}^{[\widetilde{\jmath},j]})\eta}^{[\widetilde{\jmath},j]2}\forall i_{\widetilde{\jmath}}\\
\left\|\widetilde{\boldsymbol{\Psi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\Delta\Theta\right)\right\|_2^2 &\leq \tilde{\psi}_{\Delta}^{[\widetilde{\jmath},j]2} = \hat{s}_{\eta}\left(\frac{\gamma^2+1}{\rho_{(\tilde{\omega}^{[\widetilde{\jmath},j]})\eta}^{[\widetilde{\jmath},j]2}}\psi_{\Delta}^{[\widetilde{\jmath}]2}\right)
\end{aligned} \tag{41}$$

$$\begin{aligned}
\left\|\boldsymbol{\phi}_{\Delta}^{[j]}\left(\mathbf{x}\right)\odot\mathbf{G}_{\Delta:i_j}^{[j]}\left(\mathbf{x}\right)\right\|_2^2 &\leq \phi_{\Delta}^{[j]2} = \gamma^2 + 1\forall i_j\\
\left\|\boldsymbol{\Psi}_{\Delta}^{[j]}\left(\Theta\right)\right\|_2^2 &\leq \psi_{\Delta}^{[j]2} = \beta_{\Delta}^{[j]2} + 2\tilde{p}^{[j]}\left(\mu_{\Delta}^{[j]2}\tilde{\omega}^{[j]2} + \left(\mu^{[j]2} + \mu_{\Delta}^{[j]2}\right)\tilde{\psi}_{\Delta}^{[j]2}\right)
\end{aligned} \tag{42}$$

with the base-cases $\|\boldsymbol{\phi}_{\Delta}^{[-1]}(\mathbf{x})\odot\mathbf{G}_{\Delta:i_{-1}}^{[-1]}(\mathbf{x})\|_2^2 \leq \phi_{\Delta}^{[-1]2} = 0$ and $\|\boldsymbol{\Psi}_{\Delta}^{[-1]}(\Delta\Theta)\|_2^2 \leq \psi_{\Delta}^{[-1]2} = 0$. We proceed as follows:

Base case: By the definition of the base case, using our assumptions, we have:

$$\left\|\boldsymbol{\phi}_{\Delta}^{[-1]}\left(\mathbf{x}\right)\odot\mathbf{G}_{\Delta:i_{-1}}^{[-1]}\left(\mathbf{x}\right)\right\|_2^2 = 0 = \phi^{[-1]2}\forall i_{-1}$$

$$\left\|\boldsymbol{\Psi}_{\Delta}^{[-1]}\left(\Delta\Theta\right)\right\|_2^2 = 0 = \psi^{[-1]2}$$

Node case: Assume (41) is true. Then, using our assumptions, we have that:

$$\begin{aligned}
\left\|\boldsymbol{\phi}_{\Delta}^{[j]}\left(\mathbf{x}\right)\odot\mathbf{G}_{\Delta:i_j}^{[j]}\left(\mathbf{x}\right)\right\|_2^2 &= \gamma^2 + \tfrac{1}{2\tilde{p}^{[j]}}\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\tfrac{1}{\tilde{\omega}^{[\widetilde{\jmath},j]}}\left\|\mathbf{x}^{[\widetilde{\jmath},j]}\right\|_2^2 + \tfrac{1}{2\tilde{p}^{[j]}}\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\tfrac{1}{\tilde{\psi}_{\Delta}^{[\widetilde{\jmath},j]}}\left\|\widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\odot\widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}\left(\mathbf{x}\right)\right\|_2^2\\
&\leq \gamma^2 + 1 = \phi_{\Delta}^{[j]2}
\end{aligned}$$

$$\begin{aligned}
\left\|\boldsymbol{\Psi}_{\Delta}^{[j]}\left(\Delta\Theta\right)\right\|_2^2 &= \left\|\Delta\mathbf{b}^{[j]\mathrm{T}}\right\|_2^2 + 2\tilde{p}^{[j]} \max_{\widetilde{j}\in\widetilde{\mathbb{P}}^{[j]}} \tilde{\omega}^{[\widetilde{j},j]2} \left\|\Delta\mathbf{W}^{[j]}\right\|_2^2 + 2\tilde{p}^{[j]} \max_{\widetilde{j}\in\widetilde{\mathbb{P}}^{[j]}} \tilde{\psi}_{\Delta}^{[\widetilde{j},j]2} \left\|\mathbf{W}^{[j]} + \Delta\mathbf{W}^{[j]}\right\|_2^2 \\
&\leq \beta_{\Delta}^{[j]2} + 2\tilde{p}^{[j]} \max_{\widetilde{j}\in\widetilde{\mathbb{P}}^{[j]}} \tilde{\omega}^{[\widetilde{j},j]2} \mu_{\Delta}^{[j]2} + 2\tilde{p}^{[j]} \max_{\widetilde{j}\in\widetilde{\mathbb{P}}^{[j]}} \tilde{\psi}_{\Delta}^{[\widetilde{j},j]2} \left(\mu^{[j]2} + \mu_{\Delta}^{[j]2}\right) \\
&= \beta_{\Delta}^{[j]2} + 2\tilde{p}^{[j]} \tilde{\omega}^{[j]2} \mu_{\Delta}^{[j]2} + 2\tilde{p}^{[j]} \tilde{\psi}_{\Delta}^{[j]2} \left(\mu^{[j]2} + \mu_{\Delta}^{[j]2}\right) \\
&= \beta_{\Delta}^{[j]2} + 2\tilde{p}^{[j]} \left(\mu_{\Delta}^{[j]2} \tilde{\omega}^{[j]2} + \left(\mu^{[j]2} + \mu_{\Delta}^{[j]2}\right) \tilde{\psi}_{\Delta}^{[j]2}\right) = \psi_{\Delta}^{[j]2}
\end{aligned}$$

Edge case: Assume (42) is true. Then, using our assumptions and the definition (21) of the rectified activation function and it's increasing (on $\mathbb{R}_+$) envelope, we have that:

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{j},j]}\left(\mathbf{x}\right) \odot \widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j},j]}\left(\mathbf{x}\right)\right\|_2^2 &= \left\| \left[\frac{a^{[\widetilde{j},j]2}\left(x_{i_{\widetilde{j}}}^{[\widetilde{j}]}\right)_k}{\eta^k} \left[\binom{k}{l} \mathrm{He}_{k-l}\left(\frac{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}}{\gamma^2+1} \right)^{\frac{l}{2}} \boldsymbol{\phi}_{\Delta}^{[\widetilde{j}]}\left(\mathbf{x}\right)^{\otimes l} \odot \mathbf{G}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\mathbf{x}\right)^{\otimes l} \right]_{1\leq l\leq k} \right]_{k\geq 1} \right\|_2^2 \\
&= \left\| \left[\frac{a^{[\widetilde{j},j]}\left(x_{i_{\widetilde{j}}}^{[\widetilde{j}]}\right)_k}{\eta^k} \left[\binom{k}{l} \mathrm{He}_{k-l}\left(\left(\frac{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}}{\gamma^2+1} \right)^{\frac{l}{2}} \boldsymbol{\phi}_{\Delta}^{[\widetilde{j}]}\left(\mathbf{x}\right) \odot \mathbf{G}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\mathbf{x}\right) \right)^{\otimes l} \right]_{1\leq l\leq k} \right]_{k\geq 1} \right\|_2^2 \\
&= \sum_{k\geq 1} \frac{a^{[\widetilde{j},j]2}\left(x_{i_{\widetilde{j}}}^{[\widetilde{j}]}\right)_k}{\eta^{2k}} \sum_{1\leq l\leq k} \binom{k}{l}^2 \mathrm{He}_{k-l}^2\left(\frac{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}}{\gamma^2+1} \left\|\boldsymbol{\phi}_{\Delta}^{[\widetilde{j}]}\left(\mathbf{x}\right) \odot \mathbf{G}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\mathbf{x}\right)\right\|_2^2 \right)^l \\
&= \hat{\tau}_{\eta}^{[\widetilde{j},j]}\left(\frac{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}}{\gamma^2+1} \left\|\boldsymbol{\phi}_{\Delta}^{[\widetilde{j}]}\left(\mathbf{x}\right) \odot \mathbf{G}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\mathbf{x}\right)\right\|_2^2 ; x_{i_{\widetilde{j}}}^{[\widetilde{j}]}, x_{i_{\widetilde{j}}}^{[\widetilde{j}]} \right) \\
&\leq \hat{\tau}_{\eta}^{[\widetilde{j},j]}\left(\frac{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}}{\gamma^2+1} \left\|\boldsymbol{\phi}_{\Delta}^{[\widetilde{j}]}\left(\mathbf{x}\right) \odot \mathbf{G}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\mathbf{x}\right)\right\|_2^2 ; \tilde{\omega}^{[\widetilde{j},j]} \right) \\
&\leq \hat{\tau}_{\eta}^{[\widetilde{j},j]}\left(\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2} ; \tilde{\omega}^{[\widetilde{j},j]} \right) \\
&\leq T_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2} = \tilde{\phi}_{\Delta}^{[\widetilde{j},j]2}
\end{aligned}$$

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\Psi}}_{\Delta}^{[\widetilde{j},j]}\left(\Delta\Theta\right)\right\|_2^2 &= \max_{i_{\widetilde{j}}} \left\| \left[\eta^k \left[\left(\frac{\gamma^2+1}{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}} \right)^{\frac{l}{2}} \left(\boldsymbol{\Psi}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\Delta\Theta\right)\right)^{\otimes l} \right]_{1\leq l\leq k} \right]_{k\geq 1} \right\|_2^2 \\
&= \max_{i_{\widetilde{j}}} \hat{s}_{\eta}\left(\frac{\gamma^2+1}{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}} \left\|\boldsymbol{\Psi}_{\Delta:i_{\widetilde{j}}}^{[\widetilde{j}]}\left(\Delta\Theta\right)\right\|_2^2 \right) \\
&= \hat{s}_{\eta}\left(\frac{\gamma^2+1}{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}} \left\|\boldsymbol{\Psi}_{\Delta}^{[\widetilde{j}]}\left(\Delta\Theta\right)\right\|_2^2 \right) \\
&\leq \hat{s}_{\eta}\left(\frac{\gamma^2+1}{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}} \psi_{\Delta}^{[\widetilde{j}]2} \right) = \tilde{\psi}_{\Delta}^{[\widetilde{j},j]2}
\end{aligned} \tag{43}$$

The desired result follows by identifying the output node $j = D - 1$.

C.2 Convergence Region of Local Dual Model

Here we prove the bounds on the change in weights and biases for which the local dual model holds as state in theorem 6. The goal is to place bounds on $\mu_{\Delta}^{[j]}, \beta_{\Delta}^{[j]}$ to ensure that:

$$\frac{1}{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}} \left\|\boldsymbol{\Psi}_{\Delta}^{[\widetilde{j}]}\left(\Delta\Theta\right)\right\|_2^2 \leq \frac{1}{\rho_{(\tilde{\omega}^{[\widetilde{j},j]})\eta}^{[\widetilde{j},j]2}} \psi_{\Delta}^{[\widetilde{j}]2} \leq R_{\eta}^2$$

$\forall \widetilde{j} \in \widetilde{\mathbb{P}}^{[j]}, \forall j$ in (43). This result suffices to ensure that $\|\widetilde{\boldsymbol{\Psi}}_{\Delta}^{[\widetilde{j},j]}(\Delta\Theta) \odot \widetilde{\mathbf{G}}_{:i_j}^{[\widetilde{j},j]}\|_2^2 \leq 1$ converges, and subsequently that both feature maps are convergent, allowing us to conclude that the model is well-defined (convergent) and enabling e.g. our bound on Rademacher complexity to be derived.

We aim to find the largest possible $\mu_\Delta^{[\widetilde{\jmath},j]}, \beta_\Delta^{[j]}$ such that:

$$\frac{1}{\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega}^{[\widetilde{\jmath},j]})\eta}} \psi_\Delta^{[\widetilde{\jmath}]2} \leq R_\eta^2$$

$\forall j : \widetilde{\jmath} \in \widetilde{\mathbb{P}}^{[j]}$. So we require that:

$$\psi_\Delta^{[\widetilde{\jmath}]2} \leq \min_{j:\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega}^{[\widetilde{\jmath},j]})\eta} R_\eta^2 \tag{44}$$

Recall that:

$$\psi_\Delta^{[\widetilde{\jmath}]2} = \left(\gamma^2+1\right)\left(\beta_\Delta^{[\widetilde{\jmath}]2} + 2\sum_{q\in\mathbb{P}^{[\widetilde{\jmath}]}} \frac{\widetilde{H}^{[\widetilde{\jmath}]}}{H^{[q]}}\left(\mu_\Delta^{[q,\widetilde{\jmath}]2}\tilde{\omega}^{[q,\widetilde{\jmath}]2} + 2\mu^{[q,\widetilde{\jmath}]2}\tilde{\psi}_\Delta^{[q,\widetilde{\jmath}]2}\right)\right)$$

Thus our condition becomes:

$$\left(\beta_\Delta^{[\widetilde{\jmath}]2} + 2\sum_{q\in\mathbb{P}^{[\widetilde{\jmath}]}} \frac{\widetilde{H}^{[\widetilde{\jmath}]}}{H^{[q]}}\mu_\Delta^{[q,\widetilde{\jmath}]2}\tilde{\omega}^{[q,\widetilde{\jmath}]2}\right) + 4\sum_{q\in\mathbb{P}^{[\widetilde{\jmath}]}} \frac{\widetilde{H}^{[\widetilde{\jmath}]}}{H^{[q]}}\mu^{[q,\widetilde{\jmath}]2}\tilde{\psi}_\Delta^{[q,\widetilde{\jmath}]2} \leq \min_{j:\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \frac{\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega}^{[\widetilde{\jmath},j]})}}{\gamma^2+1} R_\eta^2$$

Selecting $d^{[\widetilde{\jmath}]} \in (0,1)\ \forall j$ we can simplify the requirement to:

$$\left(\gamma^2+1\right)\left(\beta_\Delta^{[\widetilde{\jmath}]2} + 2p^{[\widetilde{\jmath}]}\mu_\Delta^{[\widetilde{\jmath}]2}\tilde{\omega}^{[\widetilde{\jmath}]2}\right) \leq d^{[\widetilde{\jmath}]}\psi_\Delta^{[\widetilde{\jmath}]2}$$

and:

$$4p^{[\widetilde{\jmath}]}\mu^{[\widetilde{\jmath}]2}\tilde{\psi}_\Delta^{[\widetilde{\jmath}]2} \leq \left(1-d^{[\widetilde{\jmath}]}\right) \min_{j:\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \frac{\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega}^{[\widetilde{\jmath},j]})}}{\gamma^2+1} R_\eta^2$$

Re-ordering, we find:

$$4\tilde{p}^{[\widetilde{\jmath}]}\mu^{[\widetilde{\jmath}]2}\hat{s}_\eta\left(\frac{1}{\rho^{[\widetilde{\jmath}',\widetilde{\jmath}]2}_{(\tilde{\omega}^{[\widetilde{\jmath}',\widetilde{\jmath}]})\eta}}\psi_\Delta^{[\widetilde{\jmath}']2}\right) \leq \left(1-d^{[\widetilde{\jmath}]}\right)\psi_\Delta^{[\widetilde{\jmath}]2}\ \ \forall\widetilde{\jmath}' \in \widetilde{\mathbb{P}}^{[\widetilde{\jmath}]}$$

and after cleaning up and re-indexing:

$$\psi_\Delta^{[j]2} \leq \rho^{[j,\widetilde{\jmath}]2}_{(\tilde{\omega}^{[j,\widetilde{\jmath}]})\eta}\hat{s}_\eta^{-1}\left(\frac{1}{4\tilde{p}^{[\widetilde{\jmath}]}\mu^{[\widetilde{\jmath}]2}}\left(1-d^{[\widetilde{\jmath}]}\right)\psi_\Delta^{[\widetilde{\jmath}]2}\right)\ \ \forall\widetilde{\jmath} : j \in \widetilde{\mathbb{P}}^{[\widetilde{\jmath}]}$$

so, overall:

$$\psi_\Delta^{[j]2} \leq \min_{\widetilde{\jmath}:j\in\widetilde{\mathbb{P}}^{[\widetilde{\jmath}]}} \rho^{[j,\widetilde{\jmath}]2}_{(\tilde{\omega}^{[j,\widetilde{\jmath}]})\eta}\left\{R_\eta^2, \hat{s}_\eta^{-1}\left(\frac{1}{4\tilde{p}^{[\widetilde{\jmath}]}\mu^{[\widetilde{\jmath}]2}}\left(1-d^{[\widetilde{\jmath}]}\right)\psi_\Delta^{[\widetilde{\jmath}]2}\right)\right\}$$

and so sufficient conditions are:

$$\left(\beta_\Delta^{[j]2} + 2p^{[j]}\tilde{\omega}^{[j]2}\mu_\Delta^{[j]2}\right) \leq d^{[j]}u^{[j]2}$$

where:

$$u^{[j]2} = \min_{\widetilde{\jmath}:j\in\widetilde{\mathbb{P}}^{[\widetilde{\jmath}]}} \rho^{[j,\widetilde{\jmath}]2}_{(\tilde{\omega}^{[j,\widetilde{\jmath}]})\eta}\left\{R_\eta^2, \hat{s}_\eta^{-1}\left(\frac{1}{4\tilde{p}^{[\widetilde{\jmath}]}\mu^{[\widetilde{\jmath}]2}}\left(1-d^{[\widetilde{\jmath}]}\right)\psi_\Delta^{[\widetilde{\jmath}]2}\right)\right\}$$

and the final result follows by setting $d^{[j]} = \frac{1}{2}\ \forall j$.

C.3 RKHS Form of Local Model

Our goal in this section is to derive the LiNK kernel from the Local dual model. For reason that will become apparent we find it convenient to work with a slight variant of the local dual feature map with minor re-scaling, namely:

$$\begin{aligned}
\widetilde{\boldsymbol{\phi}}_\Delta^{[\widetilde{\jmath},j]}(\mathbf{x}) &= \frac{1}{T^{[\widetilde{\jmath},j]}_{(\tilde{\omega})\eta}}\left[\frac{1}{\eta^k}\left[\left(\sqrt{\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta}}\boldsymbol{\phi}_\Delta^{[\widetilde{\jmath}]}(\mathbf{x})\right)^{\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1} \\
\widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}(\mathbf{x}) &= \left[a^{[\widetilde{\jmath},j]}_{\binom{x^{[\widetilde{\jmath}]}}{i_{\widetilde{\jmath}}}k}\left[\binom{k}{l}\mathrm{He}_{k-l}\mathbf{G}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}(\mathbf{x})^{\otimes l}\right]_{1\leq l\leq k}\right]_{k\geq 1}
\end{aligned} \tag{45}$$

$\forall \widetilde{\jmath} \in \widetilde{\mathbb{P}}^{[j]}$ and:

$$\begin{aligned}\boldsymbol{\phi}_{\Delta}^{[j]}(\mathbf{x}) &= \sqrt{\tfrac{1}{\gamma^2+1}}\begin{bmatrix} \gamma \\ \frac{1}{\sqrt{2}}\left[\sqrt{\frac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}}\frac{1}{\tilde{\omega}^{[\widetilde{\jmath},j]}}\tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \\ \frac{1}{\sqrt{2}}\left[\sqrt{\frac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}}}\widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x})\right]_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\end{bmatrix} \\ \mathbf{G}_{\Delta}^{[j]}(\mathbf{x}) &= \begin{bmatrix} \mathbf{1}_{H^{[\widetilde{\jmath}]}}^{\mathrm{T}} \\ \underset{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}{\operatorname{diag}}\left(\mathbf{I}_{H^{[\widetilde{\jmath}]}}\right)\mathbf{1}_{\tilde{H}^{[j]}}\mathbf{1}_{H^{[j]}}^{\mathrm{T}} \\ \underset{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}{\operatorname{diag}}\left(\widetilde{\mathbf{G}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x})\right)\mathbf{1}_{\tilde{H}^{[j]}}\mathbf{1}_{H^{[j]}}^{\mathrm{T}}\end{bmatrix}\end{aligned} \tag{46}$$

We have already demonstrated that the necessary conditions for functions in our space to like in an RKHS, so it remains derive the kernel representing this space. To this end we recall that Reisz representor theory implies that $\forall \mathbf{x} \in \mathbb{X}\ \exists$ unique $\mathbf{K}_{\mathbf{x}} \in \mathcal{F} \times \mathbb{R}^m$ such that $\langle \mathbf{f}(\mathbf{x}), \mathbf{v}\rangle = \langle \mathbf{f}, \mathbf{K}_{\mathbf{x}}\mathbf{v}\rangle_{\mathcal{H}}$ $\forall \mathbf{v} \in \mathbb{R}^m$; from which the kernel may be obtained using:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \left[\left\langle \mathbf{K}_{\mathbf{x}}\boldsymbol{\delta}_{(i_{D-1})}^{[D-1]}, \mathbf{K}_{\mathbf{x}'}\boldsymbol{\delta}_{(i'_{D-1})}^{[D-1]}\right\rangle_{\mathcal{H}}\right]_{i_{D-1},i'_{D-1}}$$

where $\boldsymbol{\delta}_{(k)}^{[j]} = [\delta_{k,i_j}]_{i_j}$. Thus our first task is to find such a $\mathbf{K}_{\mathbf{x}}$. Recall:

$$\Delta\mathbf{f}(\mathbf{x}) = \left(\boldsymbol{\Psi}_{\Delta}^{[D-1]}(\Theta) \odot \mathbf{G}_{\Delta}^{[D-1]}(\mathbf{x})\right)^{\mathrm{T}} \boldsymbol{\phi}_{\Delta}^{[D-1]}(\mathbf{x})$$

and so:

$$\begin{aligned}\langle \Delta\mathbf{f}(\mathbf{x}), \mathbf{v}\rangle &= \Delta\mathbf{f}(\mathbf{x})^{\mathrm{T}}\mathbf{v} \\ &= \boldsymbol{\phi}_{\Delta}^{[D-1]}(\mathbf{x})^{\mathrm{T}}\left(\boldsymbol{\Psi}_{\Delta}^{[D-1]}(\Theta) \odot \mathbf{G}_{\Delta}^{[D-1]}(\mathbf{x})\right)\mathbf{v} \\ &= \textstyle\sum_{k,i'_{D-1}} \phi_{\Delta k}^{[D-1]}(\mathbf{x})\left(\Psi_{\Delta k,i'_{D-1}}^{[D-1]}(\Theta)\, G_{\Delta k,i'_{D-1}}^{[D-1]}(\mathbf{x})\right) v_{i'_{D-1}} \\ &= \textstyle\sum_{k,i'_{D-1}} \Psi_{\Delta k,i'_{D-1}}^{[D-1]}(\Theta)\left(\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,i'_{D-1}}^{[D-1]}(\mathbf{x})\, v_{i'_{D-1}}\right) \\ &= \textstyle\sum_{k,i'_{D-1}} \Psi_{\Delta k,i'_{D-1}}^{[D-1]}(\Theta)\left(\sum_{i''_{D-1}} \delta_{i'_{D-1},i''_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,i''_{D-1}}^{[D-1]}(\mathbf{x})\, v_{i''_{D-1}}\right) \\ &= \langle \mathbf{f}, \mathbf{K}_{\mathbf{x}}\mathbf{v}\rangle\end{aligned}$$

where we have denoted:

$$\begin{aligned}\mathbf{f} &= \left[\Psi_{\Delta k,i_{D-1}}^{[D-1]}(\Theta)\right]_{(k,i_{D-1})} \\ \mathbf{K}_{\mathbf{x}} &= \left[\delta_{i_{D-1},i'_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,i'_{D-1}}^{[D-1]}(\mathbf{x})\right]_{(k,i_{D-1}),i'_{D-1}}\end{aligned}$$

We may therefore proceed to derive the kernel as follows, letting $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ be the standard inner product:

$$\begin{aligned}K_{i_{D-1},i'_{D-1}}(\mathbf{x}, \mathbf{x}') &= \left[\left\langle \left[\delta_{\tilde{i}'_{D-1},\tilde{i}''_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,\tilde{i}''_{D-1}}^{[D-1]}(\mathbf{x})\right]_{(k,\tilde{i}'_{D-1}),\tilde{i}''_{D-1}} \left[\delta_{i_{D-1},\tilde{i}''_{D-1}}\right]_{\tilde{i}''_{D-1}},\right.\right. \\ &\qquad \left.\left.\left[\delta_{\tilde{i}'_{D-1},\tilde{i}''_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x}')\, G_{\Delta k,\tilde{i}''_{D-1}}^{[D-1]}(\mathbf{x}')\right]_{(k,\tilde{i}'_{D-1}),\tilde{i}''_{D-1}} \left[\delta_{i'_{D-1},\tilde{i}''_{D-1}}\right]_{\tilde{i}''_{D-1}}\right\rangle_{\mathcal{H}}\right]_{i_{D-1},i'_{D-1}} \\ &= \left[\left\langle \left[\delta_{\tilde{i}'_{D-1},i_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,i_{D-1}}^{[D-1]}(\mathbf{x})\right]_{(k,\tilde{i}'_{D-1})},\right.\right. \\ &\qquad \left.\left.\left[\delta_{\tilde{i}'_{D-1},i'_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x}')\, G_{\Delta k,i'_{D-1}}^{[D-1]}(\mathbf{x}')\right]_{(k,\tilde{i}'_{D-1})}\right\rangle_{\mathcal{H}}\right]_{i_{D-1},i'_{D-1}} \\ &= \left[\sum_{k,\tilde{i}'_{D-1}} \delta_{\tilde{i}'_{D-1},i_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,i_{D-1}}^{[D-1]}(\mathbf{x})\, \delta_{\tilde{i}'_{D-1},i'_{D-1}}\phi_{\Delta k}^{[D-1]}(\mathbf{x}')\, G_{\Delta k,i'_{D-1}}^{[D-1]}(\mathbf{x}')\right]_{i_{D-1},i'_{D-1}} \\ &= \left[\delta_{i_{D-1},i'_{D-1}}\sum_{k}\phi_{\Delta k}^{[D-1]}(\mathbf{x})\, G_{\Delta k,i_{D-1}}^{[D-1]}(\mathbf{x})\,\phi_{\Delta k}^{[D-1]}(\mathbf{x}')\, G_{\Delta k,i'_{D-1}}^{[D-1]}(\mathbf{x}')\right]_{i_{D-1},i'_{D-1}} \\ &= \delta_{i_{D-1},i'_{D-1}}K_{i_{D-1}}^{[D-1]}(\mathbf{x}, \mathbf{x}')\end{aligned}$$

where:

$$
\begin{aligned}
K_{i_j}^{[j]}(\mathbf{x},\mathbf{x}') &= \left\langle \boldsymbol{\phi}_{\Delta}^{[j]}(\mathbf{x}) \odot \mathbf{G}_{\Delta:i_j}^{[j]}(\mathbf{x}), \boldsymbol{\phi}_{\Delta}^{[j]}(\mathbf{x}') \odot \mathbf{G}_{\Delta:i_j}^{[j]}(\mathbf{x}') \right\rangle \\
&= \tfrac{1}{\gamma^2+1}\left(\gamma^2 + \tfrac{1}{2}\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{\widetilde{\omega}^{[\widetilde{\jmath},j]2}} \left\langle \tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}, \tilde{\mathbf{x}}'^{[\widetilde{\jmath},j]} \right\rangle + \ldots \right. \\
&\qquad \left. \ldots + \tfrac{1}{2}\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \sum_{i_{\widetilde{\jmath}}} \left\langle \widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x}) \odot \widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}(\mathbf{x}), \widetilde{\boldsymbol{\phi}}_{\Delta}^{[\widetilde{\jmath},j]}(\mathbf{x}') \odot \widetilde{\mathbf{G}}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath},j]}(\mathbf{x}') \right\rangle \right) \\
&= \tfrac{1}{\gamma^2+1}\left(\gamma^2 + \tfrac{1}{2}\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{\widetilde{\omega}^{[\widetilde{\jmath},j]2}} \left\langle \tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}, \tilde{\mathbf{x}}'^{[\widetilde{\jmath},j]} \right\rangle + \ldots \right. \\
&\left. \ldots + \tfrac{1}{2}\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{T_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2}} \sum_{i_{\widetilde{\jmath}}} \sum_{k\geq 1} \frac{a_{\left(x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\right)_k}^{[\widetilde{\jmath},j]} a_{\left(x_{i_{\widetilde{\jmath}}}'^{[\widetilde{\jmath}]}\right)_k}^{[\widetilde{\jmath},j]}}{\eta^{2k}} \sum_{1\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l}^2 \left(\rho_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2} \left\langle \boldsymbol{\phi}_{\Delta}^{[\widetilde{\jmath}]}(\mathbf{x}) \odot \mathbf{G}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}(\mathbf{x}), \boldsymbol{\phi}_{\Delta}^{[\widetilde{\jmath}]}(\mathbf{x}') \odot \mathbf{G}_{\Delta:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}(\mathbf{x}') \right\rangle\right)^l \right) \\
&= \tfrac{1}{\gamma^2+1}\left(\gamma^2 + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{\widetilde{\omega}^{[\widetilde{\jmath},j]2}} \left\langle \tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}, \tilde{\mathbf{x}}'^{[\widetilde{\jmath},j]} \right\rangle + \ldots \right. \\
&\left. \ldots + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{T_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2}} \sum_{i_{\widetilde{\jmath}}} \sum_{k\geq 1} \frac{a_{\left(x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}\right)_k}^{[\widetilde{\jmath},j]} a_{\left(x_{i_{\widetilde{\jmath}}}'^{[\widetilde{\jmath}]}\right)_k}^{[\widetilde{\jmath},j]}}{\eta^{2k}} \sum_{1\leq l\leq k} \binom{k}{l} \mathrm{He}_{k-l}^2 \left(\rho_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2} K_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}')\right)^l \right) \\
&= \tfrac{1}{\gamma^2+1}\left(\gamma^2 + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{\widetilde{\omega}^{[\widetilde{\jmath},j]2}} \left\langle \tilde{\mathbf{x}}^{[\widetilde{\jmath},j]}, \tilde{\mathbf{x}}'^{[\widetilde{\jmath},j]} \right\rangle + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{T_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2}} \sum_{i_{\widetilde{\jmath}}} \hat{\tau}_{\eta}^{[\widetilde{\jmath},j]}\left(\rho_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2} K_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}'); x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}, x_{i_{\widetilde{\jmath}}}'^{[\widetilde{\jmath}]}\right)\right)
\end{aligned}
$$

So, noting that this is independent of i_j, we find that $\mathbf{f} \in \mathcal{H}_{\mathbf{K}}$, where $\mathbf{K} = \mathbf{I}K_{\mathrm{LiNK}}$, where $K_{\mathrm{LiNK}} = K_{\mathrm{LiNK}}^{[D-1]}$ is defined recursively:

$$
K_{\mathrm{LiNK}}^{[j]}(\mathbf{x},\mathbf{x}') = \tfrac{1}{\gamma^2+1}\left(\sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{\omega}^{[\widetilde{\jmath},j]2}} \Sigma^{[\widetilde{\jmath},j]}(\mathbf{x},\mathbf{x}') + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{T_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2}} \sum_{i_{\widetilde{\jmath}}} \hat{\tau}_{\eta}^{[\widetilde{\jmath},j]}\left(\rho_{(\widetilde{\omega})\eta}^{[\widetilde{\jmath},j]2} K_{\mathrm{LiNK}}^{[\widetilde{\jmath}]}(\mathbf{x},\mathbf{x}'); x_{i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}, x_{i_{\widetilde{\jmath}}}'^{[\widetilde{\jmath}]}\right)\right)
$$

with $K_{\mathrm{LiNK}}^{[-1]}(\mathbf{x},\mathbf{x}') = 0$.

Recall that:

$$
\hat{\tau}_{\eta}^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') = \sum_{k=1}^{\infty} \frac{a_{(\xi)k}^{[\widetilde{\jmath},j]} a_{(\xi')k}^{[\widetilde{\jmath},j]}}{\eta^{2k}} \sum_{l=1}^{k} \binom{k}{l}^2 \mathrm{He}_{k-l}^2 \zeta^l
$$

We aim to express:

$$
\hat{\tau}_{\eta}^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') = \sum_{q} \hat{\tau}_{\eta,q}^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi')
$$

where $\hat{\tau}_{\eta,q}^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') = b(\xi,\xi')\zeta^q$. As noted previously, if:

$$
f(x) = \sum_{k=0}^{\infty} a_k \sum_{l=0}^{k} \binom{k}{l} \mathrm{He}_{k-l} x^l
$$

then:

$$
f^{(q)}(x) = \sum_{k=0}^{\infty} \tfrac{(k+q)!}{k!} a_{k+q} \sum_{l=0}^{k-q} \binom{k}{l} \mathrm{He}_{k-l} x^l
$$

Working toward the general case:

$$
\begin{aligned}
\hat{\tau}_{\eta,1}^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') &= \zeta \sum_{k=1}^{\infty} \frac{a_{(\xi)k}^{[\widetilde{\jmath},j]} a_{(\xi')k}^{[\widetilde{\jmath},j]}}{\eta^{2k}} \binom{k}{1}^2 \mathrm{He}_{k-1}^2 \\
&= \zeta \sum_{k,k'=0}^{\infty} \delta_{k,k'} \frac{a_{(\xi)k+1}^{[\widetilde{\jmath},j]}}{\eta^{k+1}} \binom{k+1}{1} \mathrm{He}_k \frac{a_{(\xi')k'+1}^{[\widetilde{\jmath},j]}}{\eta^{k'+1}} \binom{k'+1}{1} \mathrm{He}_{k'}
\end{aligned}
$$

and:

$$
\begin{aligned}
\hat{\tau}_{\eta,2}^{[\widetilde{\jmath},j]}(\zeta;\xi,\xi') &= \zeta^2 \sum_{k=2}^{\infty} \frac{a_{(\xi)k}^{[\widetilde{\jmath},j]} a_{(\xi')k}^{[\widetilde{\jmath},j]}}{\eta^{2k}} \binom{k}{2}^2 \mathrm{He}_{k-2}^2 \\
&= \zeta^2 \sum_{k,k'=0}^{\infty} \delta_{k,k'} \frac{a_{(\xi)k+2}^{[\widetilde{\jmath},j]}}{\eta^{k+2}} \binom{k+2}{2} \mathrm{He}_k \frac{a_{(\xi')k'+2}^{[\widetilde{\jmath},j]}}{\eta^{k'+2}} \binom{k'+2}{2} \mathrm{He}_{k'}
\end{aligned}
$$

and so on, so:

$$\begin{aligned}
\hat{\tau}^{[\widetilde{\jmath},j]}_{\eta,q}\left(\zeta;\xi,\xi'\right) &= \zeta^q \sum_{k=q}^{\infty} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k} a^{[\widetilde{\jmath},j]}_{(\xi')k}}{\eta^{qk}} \binom{k}{q}^2 \mathrm{He}^2_{k-q} \\
&= \zeta^q \sum_{k,k'=0}^{\infty} \delta_{k,k'} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k+q}}{\eta^{k+q}} \binom{k+q}{q} \mathrm{He}_k \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi')k'+q}}{\eta^{k'+q}} \binom{k'+q}{q} \mathrm{He}_{k'} \\
&= \tfrac{1}{q!^2} \zeta^q \sum_{k,k'=0}^{\infty} \delta_{k,k'} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k+q}}{\eta^{k+q}} \tfrac{(k+q)!}{k!} \mathrm{He}_k \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi')k'+q}}{\eta^{k'+q}} \tfrac{(k'+q)!}{k'!} \mathrm{He}_{k'} \\
&= \tfrac{1}{q!^2} \zeta^q \sum_{k,k'=0}^{\infty} \delta_{k,k'} \sum_{l=0}^{k-q} \sum_{l'=0}^{k'-q} \delta_{l,l'} \left(\tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k+q}}{\eta^{k+q}} \tfrac{(k+q)!}{k!} \binom{k}{l} \mathrm{He}_{k-l} 0^l \right) \left(\tfrac{a^{[\widetilde{\jmath},j]}_{(\xi')k'+q}}{\eta^{k'+q}} \tfrac{(k'+q)!}{k'!} \binom{k'}{l'} \mathrm{He}_{k'-l'} 0^{l'} \right) \\
&= \zeta^q \left(\tfrac{1}{q!} \sum_{k=0}^{\infty} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k+q}}{\eta^{k+q}} \tfrac{(k+q)!}{k!} \sum_{l=0}^{k-q} \binom{k}{l} \mathrm{He}_{k-l} 0^l \right) \cdot \left(\tfrac{1}{q!} \sum_{k'=0}^{\infty} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi')k'+q}}{\eta^{k'+q}} \tfrac{(k'+q)!}{k'!} \sum_{l'=0}^{k'-q} \binom{k'}{l'} \mathrm{He}_{k'-l'} 0^{l'} \right) \\
&= \tfrac{1}{q!} \overline{\tau}^{[\widetilde{\jmath},j](q)}_{\eta}\left(0;\xi\right) \tfrac{1}{q!} \overline{\tau}^{[\widetilde{\jmath},j](q)}_{\eta}\left(0;\xi'\right) \zeta^q
\end{aligned}$$

where $\cdot$ is the Cauchy product and in the final step we have used Mertens' theorem for Cauchy products, and:

$$\overline{\tau}^{[\widetilde{\jmath},j]}_{\eta}\left(\zeta;\xi\right) = \sum_{k=1}^{\infty} \tfrac{a^{[\widetilde{\jmath},j]}_{(\xi)k}}{\eta^k} \sum_{l=1}^{k} \binom{k}{l} \mathrm{He}_{k-l} \zeta^l$$

Finally, we see that, in the limit $\eta \to 1$:

$$\hat{\tau}^{[\widetilde{\jmath},j]}_{\eta,q}\left(\zeta;\xi,\xi'\right) = \tfrac{1}{q!} \overline{\tau}^{[\widetilde{\jmath},j](q)}\left(0;\xi\right) \tfrac{1}{q!} \overline{\tau}^{[\widetilde{\jmath},j](q)}\left(0;\xi'\right) \zeta^q$$

which leads to the simplified form of the LiNK:

$$\begin{aligned}
K^{[j]}_{\mathrm{LiNK}}\left(\mathbf{x},\mathbf{x}'\right) = \tfrac{1}{\gamma^2+1} \Bigg(& \sum_{\widetilde{\jmath} \in \widetilde{\mathbb{P}}[j]} \tfrac{H^{[\widetilde{\jmath}]}}{\tilde{\omega}^{[\widetilde{\jmath},j]2}} \Sigma^{[\widetilde{\jmath},j]}\left(\mathbf{x},\mathbf{x}'\right) + \dots \\
& \dots + \sum_{\widetilde{\jmath} \in \widetilde{\mathbb{P}}[j]} \tfrac{H^{[\widetilde{\jmath}]}}{\widetilde{H}^{[j]}} \tfrac{1}{T^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta}} \sum_{q=1}^{\infty} \sum_{i_{\widetilde{\jmath}}} \tfrac{1}{q!} \overline{\tau}^{[\widetilde{\jmath},j](q)}\left(0; x^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right) \tfrac{1}{q!} \overline{\tau}^{[\widetilde{\jmath},j](q)}\left(0; x'^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right) \left(\rho^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta} K^{[\widetilde{\jmath}]}_{\mathrm{LiNK}}\left(\mathbf{x},\mathbf{x}'\right) \right)^q \Bigg)
\end{aligned}$$

with $K^{[-1]}_{\mathrm{LiNK}}(\mathbf{x},\mathbf{x}') = 0$. With some cleanup:

$$\begin{aligned}
K^{[j]}_{\mathrm{LiNK}}\left(\mathbf{x},\mathbf{x}'\right) = \tfrac{1}{\gamma^2+1} \Bigg(& \mathbb{E}_{\widetilde{\jmath} \in \widetilde{\mathbb{P}}[j]} \left[\tfrac{H^{[\widetilde{\jmath}]}}{\tilde{\omega}^{[\widetilde{\jmath},j]2}} \Sigma^{[\widetilde{\jmath},j]}\left(\mathbf{x},\mathbf{x}'\right) \right] + \dots \\
& \dots + \mathbb{E}_{\widetilde{\jmath} \in \widetilde{\mathbb{P}}[j]} \left[\tfrac{H^{[\widetilde{\jmath}]}}{T^{[\widetilde{\jmath},j]2}_{(\tilde{\omega})\eta}} \sum_{q=1}^{\infty} \mathbb{E}_{i_{\widetilde{\jmath}}} \left[\tfrac{1}{q!} \tau^{[\widetilde{\jmath},j](q)}\left(x^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right) \tfrac{1}{q!} \tau^{[\widetilde{\jmath},j](q)}\left(x'^{[\widetilde{\jmath}]}_{i_{\widetilde{\jmath}}}\right) \right] K^{[\widetilde{\jmath}]}_{\mathrm{LiNK}}\left(\mathbf{x},\mathbf{x}'\right)^q \right] \Bigg)
\end{aligned}$$

with $K^{[-1]}_{\mathrm{LiNK}}(\mathbf{x},\mathbf{x}') = 0$.

D Rademacher Complexity Bounds - Proof of Theorems 3, 4 and 8

In this supplementary we prove theorems relating to the Rademacher complexity of neural networks for the global and local models.

Theorem 3 *The set $\mathcal{F} = \{f(\cdot;\Theta) : \mathbb{R}^n \to \mathbb{R} | \Theta \in \mathbb{W}\}$ of networks (1) satisfying our assumptions has Rademacher complexity bounded by $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}} \phi \underset{\sim}{\psi}$ (definitions as per Figure 1).*

Proof. Let $\mathcal{F}$ be the set of attainable neural networks (scalar output) and ϵ a Rademacher random

variable. Let $\mathbf{x} \sim \nu$. Then the Rademacher complexity is bounded as:

$$\begin{aligned}
\mathcal{R}_N(\mathcal{F}) &\triangleq \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \tfrac{1}{N} \textstyle\sum_k \epsilon_k f(\mathbf{x}_k) \right] \\
&=^{\text{Global Dual}} \tfrac{1}{N} \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Theta \in \mathbb{W}} \textstyle\sum_k \epsilon_k \left\langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\mathbf{x}_k) \right]_{\mathbf{g}} \right] \\
&= \tfrac{1}{N} \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Theta \in \mathbb{W}} \left\langle \boldsymbol{\Psi}(\Theta), \textstyle\sum_k \epsilon_k \boldsymbol{\phi}(\mathbf{x}_k) \right]_{\mathbf{g}} \right] \\
&\leq^{\text{Operator norm}} \tfrac{1}{N} \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Theta \in \mathbb{W}} \|\boldsymbol{\Psi}(\Theta)\|_{\mathrm{He}[\tau]} \sqrt{\left\| \textstyle\sum_k \epsilon_k \boldsymbol{\phi}(\mathbf{x}_k) \right\|_2^2} \right] \\
&\leq^{\text{Norm Bound}} \tfrac{\underset{\sim}{\psi}}{N} \mathbb{E}_\nu \left[\mathbb{E}_\epsilon \sqrt{\left\| \textstyle\sum_k \epsilon_k \boldsymbol{\phi}(\mathbf{x}_k) \right\|_2^2} \right] \\
&\leq^{\text{Jensen}} \tfrac{\underset{\sim}{\psi}}{N} \mathbb{E}_\nu \left[\sqrt{\mathbb{E}_\epsilon \left\| \textstyle\sum_k \epsilon_k \boldsymbol{\phi}(\mathbf{x}_k) \right\|_2^2} \right] \\
&=^{\{\mathbb{E}_\epsilon \epsilon_k \epsilon_l = \delta_{k,l}\}} \tfrac{\underset{\sim}{\psi}}{N} \mathbb{E}_\nu \left[\sqrt{\textstyle\sum_k \|\boldsymbol{\phi}(\mathbf{x}_k)\|_2^2} \right] \\
&\leq^{\text{Norm Bound}} \tfrac{\underset{\sim}{\psi}}{N} \mathbb{E}_\nu \left[\sqrt{N \phi^2} \right] = \tfrac{\phi \underset{\sim}{\psi}}{\sqrt{N}}
\end{aligned}$$

□

Theorem 4 *Let $\mathcal{F} = \{f(\cdot;\Theta) : \mathbb{R}^n \to \mathbb{R} | \Theta \in \mathbb{W}\}$ be the set of unbiased networks (1) with L-Lipschitz activations satisfying our assumptions. Then $\mathcal{R}_N(\mathcal{F}) \leq \max_{\mathcal{S} \in \mathbb{S}} \prod_{j \in \mathcal{S}} L^2 \tilde{p}^{[j]} \mu^{[j]}$, where $\mathbb{S}$ is the set of all input-output paths in the network graph.*

Proof. Observe from Figure 1 that:

$$\tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath},j]2} \leq \frac{\phi^{[\tilde{\jmath}]2}}{\overline{\mathbf{s}}^{[\tilde{\jmath},j]} \left(\overline{\mathbf{s}}^{[\tilde{\jmath},j]-1} \left(\tilde{\phi}^{[\tilde{\jmath},j]2} \right) \phi^{[\tilde{\jmath}]2} \right)} L^2 \tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath}]2}$$

Note that the numerator of the factional part is linearly increasing while the denominator is superlinearly increasing, so we may pessimise this bound as:

$$\begin{aligned}
\tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath},j]2} &\leq \frac{\phi^{[\tilde{\jmath}]2}}{\overline{\mathbf{s}}^{[\tilde{\jmath},j]} \left(\overline{\mathbf{s}}^{[\tilde{\jmath},j]-1} \left(\tilde{\phi}^{[\tilde{\jmath},j]2} \right) \phi^{[\tilde{\jmath}]2} \right)} L^2 \tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath}]2} \\
&\leq \lim_{\phi^{[\tilde{\jmath}]} \to 0} \frac{\phi^{[\tilde{\jmath}]2}}{\overline{\mathbf{s}}^{[\tilde{\jmath},j]} \left(\overline{\mathbf{s}}^{[\tilde{\jmath},j]-1} \left(\tilde{\phi}^{[\tilde{\jmath},j]2} \right) \phi^{[\tilde{\jmath}]2} \right)} L^2 \tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath}]2} = \frac{1}{a_0^{[\tilde{\jmath},j]} \overline{\mathbf{s}}^{[\tilde{\jmath},j]-1} \left(\tilde{\phi}^{[\tilde{\jmath},j]2} \right)} L^2 \tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath}]2}
\end{aligned}$$

As $\tilde{\phi}^{[\tilde{\jmath},j]2}$ is a free parameter here we can let $\tilde{\phi}^{[\tilde{\jmath},j]2} \to 0$, in which limit $\tilde{\phi}^{[\tilde{\jmath},j]2} \underset{\sim}{\psi}^{[\tilde{\jmath},j]2} \leq L^2 \underset{\sim}{\psi}^{[\tilde{\jmath}]2} \leq L^2 p^{[\tilde{\jmath}]} \mu^{[\tilde{\jmath}]2} \sum_{\tilde{\tilde{\jmath}} \in \tilde{\mathbb{P}}^{[\tilde{\jmath}]}} \tilde{\phi}^{[\tilde{\tilde{\jmath}},\tilde{\jmath}]2} \underset{\sim}{\psi}^{[\tilde{\tilde{\jmath}},\tilde{\jmath}]2}$. The result follows recursively, then upper bounding with the pathwise product. □

Theorem 8 *The set $\mathcal{F}_\Delta = \{\Delta f(\cdot;\Delta\Theta) : \mathbb{R}^n \to \mathbb{R} | \Delta\Theta \in \mathbb{W}_\Delta\}$ of change in neural-network operation satisfying (20) has Rademacher complexity $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}} \phi_\Delta \psi_\Delta$ (defined in Figure 2).*

Proof. Let $\mathcal{F}_\Delta$ be the set of attainable changes in neural networks (scalar output) and ϵ a Rademacher

random variable. Let $\mathbf{x} \sim \nu$. Then the Rademacher complexity is bounded as:

$$
\begin{aligned}
\mathcal{R}_N\left(\mathcal{F}_\Delta\right) &\triangleq \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Delta f \in \mathcal{F}} \frac{1}{N} \sum_k \epsilon_k \Delta f\left(\mathbf{x}_k\right)\right] \\
&=^{\text{Local Dual}} \frac{1}{N} \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Delta\Theta \in \mathbb{W}_\Delta} \sum_k \epsilon_k \left\langle \boldsymbol{\Psi}_\Delta\left(\Delta\Theta\right), \boldsymbol{\phi}_\Delta\left(\mathbf{x}_k\right)\right]_{\mathbf{g}}\right] \\
&= \frac{1}{N} \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Delta\Theta \in \mathbb{W}_\Delta} \left\langle \boldsymbol{\Psi}_\Delta\left(\Delta\Theta\right), \sum_k \epsilon_k \boldsymbol{\phi}_\Delta\left(\mathbf{x}_k\right)\right]_{\mathbf{g}}\right] \\
&\leq^{\text{Cauchy Schwarz}} \frac{1}{N} \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Delta\Theta \in \mathbb{W}_\Delta} \left\|\boldsymbol{\Psi}_\Delta\left(\Delta\Theta\right)\right\|_2 \sqrt{\left\|\sum_k \epsilon_k \boldsymbol{\phi}_\Delta\left(\mathbf{x}_k\right)\right\|_2^2}\right] \\
&\leq^{\text{Norm Bound}} \frac{\psi_\Delta}{N} \mathbb{E}_\nu \left[\mathbb{E}_\epsilon \sqrt{\left\|\sum_k \epsilon_k \boldsymbol{\phi}_\Delta\left(\mathbf{x}_k\right)\right\|_2^2}\right] \\
&\leq^{\text{Jensen}} \frac{\psi_\Delta}{N} \mathbb{E}_\nu \left[\sqrt{\mathbb{E}_\epsilon \left\|\sum_k \epsilon_k \boldsymbol{\phi}_\Delta\left(\mathbf{x}_k\right)\right\|_2^2}\right] \\
&=^{\{\mathbb{E}_\epsilon \epsilon_k \epsilon_l = \delta_{k,l}\}} \frac{\psi_\Delta}{N} \mathbb{E}_\nu \left[\sqrt{\sum_k \left\|\boldsymbol{\phi}_\Delta\left(\mathbf{x}_k\right)\right\|_2^2}\right] \\
&\leq^{\text{Norm Bound}} \frac{\psi_\Delta}{N} \mathbb{E}_\nu \left[\sqrt{N \phi_\Delta^2}\right] = \frac{\phi_\Delta \psi_\Delta}{\sqrt{N}}
\end{aligned}
$$

□