Gradient Descent in Neural Networks as Sequential Learning in Reproducing Kernel Banach Space

Alistair Shilton [1] **Sunil Gupta** [1] **Santu Rana** [1] **Svetha Venkatesh** [1]

Abstract

The study of Neural Tangent Kernels (NTKs) has provided much needed insight into convergence and generalization properties of neural networks in the over-parametrized (wide) limit by approximating the network using a first-order Taylor expansion with respect to its weights in the neighborhood of their initialization values. This allows neural network training to be analyzed from the perspective of reproducing kernel Hilbert spaces (RKHS), which is informative in the over-parametrized regime, but a poor approximation for narrower networks as the weights change more during training. Our goal is to extend beyond the limits of NTK toward a more general theory. We construct an exact power-series representation of the neural network in a finite neighborhood of the initial weights as an inner product of two feature maps, respectively from data and weight-step space, to feature space, allowing neural network training to be analyzed from the perspective of reproducing kernel *Banach* space (RKBS). We prove that, regardless of width, the training sequence produced by gradient descent can be exactly replicated by regularized sequential learning in RKBS. Using this, we present novel bound on uniform convergence where the iterations count and learning rate play a central role, giving new theoretical insight into neural network training.

1. Introduction

The remarkable progress made in neural networks in recent decades has led to an explosion in their adoption in a wide swathe of applications. However this widespread success has also left unanswered questions, the most obvious of which is why non-convex, massively over-parameterized networks are able to perform so much better than predicted by traditional machine learning theory.

Neural tangent kernels represent an attempt to answer this question. As per (Jacot et al., 2018; Arora et al., 2019b), during training, the evolution of an over-parameterized neural network follows the kernel gradient of the functional cost with respect to a neural tangent kernel (NTK). It was shown that, for a sufficiently wide network with random weight initialization, the NTK is effectively fixed, and results from machine learning in reproducing kernel Hilbert space (RKHS) can thus be brought to bear on the problem. This has led to a plethora of results analysing the convergence (Du et al., 2019b; Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2020; Zou & Gu, 2019) and generalization (Arora et al., 2019b;a; Cao & Gu, 2019) properties of neural networks.

Despite their successes, NTK models are not without problems. As noted in (Bai & Lee, 2019), the expressive power of the linear approximation used by NTK is limited to that of the corresponding, randomized feature space or RKHS, as evidenced by the observed gap between NTK predictions and actual performance. To break out of this regime, (Bai & Lee, 2019) proposed using a second or higher-order approximation of the network. Moreover it is natural to ask how well a linear approximation of the behaviour, constructed on the assumption of small weight-steps, will scale to larger weight steps in narrower networks.

To overcome these difficulties we replace the Taylor approximation used in NTK with an exact power series representation of the neural network in a finite neighbourhood around the initial weights. We demonstrate that this leads to a representation as an inner product between two feature maps, from data and weight-step space, respectively. This structure underlies the construction of reproducing kernel Banach spaces (RKBS, (Lin et al., 2022)), allowing us to go on to show an equivalence between back-propagation and sequential learning in RKBS, which is similar to NTK but without the constraints of linearity, allowing us to derive new bounds on uniform convergence for networks of arbitrary width.

[1]Applied Artificial Intelligence Institute ($\mathrm{A}^2\mathrm{I}^2$), Deakin University, Geelong, Australia. Correspondence to: Alistair Shilton <alistair.shilton@deakin.edu.au>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Related Work

There has been a significant amount of work looking at uniform convergence behaviour of networks of different types using variety of assumptions during training (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, 2019a;b; Zhou et al., 2019).

The study of the connection between kernel methods and neural networks has a long history. (Neal, 1996) demonstrated that, in the infinite-width limit, iid randomly initialized single-layer networks converge to draws from a Gaussian process. This was extended to multi-layered neural networks in (Lee et al., 2018; Matthews et al., 2018) by assuming random weights up to (but not including) the output layer. Other works deriving approximate kernels by assuming random weights include (Rahimi & Benjamin, 2009; Bach, 2014; 2017; Daniely et al., 2016; Daniely, 2017).

Neural tangent kernels (Jacot et al., 2018; Arora et al., 2019b) are a more recent development. The basis of NTK is to approximate the behaviour of neural network (for a given input $\mathbf{x}$) as the weights and biases vary about some initial values using a first-order Taylor approximation. This approximation is linear in the change in weights, and the coefficients of this approximation are functions of $\mathbf{x}$ and may therefore be treated as a feature map, making the model amenable to the kernel trick and subsequent analysis in terms of RKHS theory. This approach may be generalized to higher order approximations (Bai & Lee, 2019), but the size of change in the weights that can be approximated remains limited except in the over-parametrized limit, where the variation of the weights becomes small.

Arc-cosine kernels (Cho & Saul, 2009) work on a similar premise. For activation functions of the form $\tau(\xi) = (\xi)_+^p$, $p \in \mathbb{N}$, in the infinite-width limit, arc-cosine kernels capture the feature map of the network. Depth is achieved by composition of kernels. However once again this approach is restricted to networks of infinite width, whereas our approach works for arbitrary networks.

Finally there has been some very recent work (Bartolucci et al., 2021; Sanders, 2020; Parhi & Nowak, 2021; Unser, 2021; 2019) in a similar vein to the current work, seeking to connect neural networks to RKBS theory. However (Bartolucci et al., 2021; Sanders, 2020; Parhi & Nowak, 2021; Unser, 2021) consider only 1 and 2 layer networks (we consider networks of arbitrary depth), and more generally no equivalence is established between the weight-steps found by back-propagation and those found by regularized learning in RKBS.

3. Notations

Let $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{N}_n = \{0, 1, \ldots, n-1\}$. Vectors and matrices are denoted $\mathbf{a}$ and $\mathbf{A}$, respectively, with elements a_i, $A_{i,i'}$, and columns $\mathbf{A}_{:i}$ indexed by $i, i' \in \mathbb{N}$. We define:

$$\|\mathbf{x}\|_p = (\textstyle\sum_i |x_i|^p)^{1/p}, \; \|\mathbf{A}\|_{p,q} = \|[\|\mathbf{A}_{:i}\|_p]_i\|_q$$
$$\|\mathbf{x}\|_\infty = \max_i\{|x_i|\}, \; \|\mathbf{x}\|_{-\infty} = \min_i\{|x_i|\}$$

$\forall p, q \in [-\infty, 0) \cup (0, \infty]$, which are norms if $p, q \in [1, \infty]$. The Frobenius norm and inner product are $\|\cdot\|_F = \|\cdot\|_{2,2}$, $\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathrm{Tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})$. The Kronecker and Hadamard product are $\mathbf{a} \otimes \mathbf{b}$, $\mathbf{a} \odot \mathbf{b}$. The Kronecker and Hadamard powers are $\mathbf{a}^{\otimes c} = \mathbf{a} \otimes \mathbf{a} \otimes \overset{c \text{ times}}{\cdots} \otimes \mathbf{a}$, $\mathbf{a}^{\odot c} = \mathbf{a} \odot \mathbf{a} \odot \overset{c \text{ times}}{\cdots} \odot \mathbf{a}$. The elementwise absolute and sign are $|\mathbf{a}|$, $\mathrm{sgn}(\mathbf{a})$. Finally, for vectors $\mathbf{a}, \mathbf{b}$ we let $\boldsymbol{\varrho}(\mathbf{a}, \mathbf{b}) = [\, a_0(\mathbf{b}^{\otimes 1})^{\mathrm{T}}, a_1(\mathbf{b}^{\otimes 2})^{\mathrm{T}}, \ldots\,]^{\mathrm{T}}$. $\mathrm{diag}(\mathbf{A})$ is a vector containing the diagonal elements of $\mathbf{A}$, and conversely $\mathrm{diag}(\mathbf{a})$ is a diagonal matrix with diagonal elements from $\mathbf{a}$.

We study fully connected D-layer neural networks $\mathbf{f} : (\mathbb{X} \subset \mathbb{R}^n) \to (\mathbb{Y} \subset \mathbb{R}^m)$ with layer widths $H^{[j]}$ (and $H^{[-1]} = n$) trained on a training set $\{(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}) \in \mathbb{X} \times \mathbb{Y} : k \in \mathbb{N}_N\}$. We use index range conventions $k \in \mathbb{N}_N$, $j \in \mathbb{N}_D$ and $i_{j+1} \in \mathbb{N}_{H^{[j]}}$, and for clarity we write:

Relating to: Training vector k ↘ Layer j ↓ The r^{th} derivative ↙

$$X^{\{k\}[j](r)}$$

Variable name ↗ Variable indices ↖

so for example $\mathbf{W}^{[j]}$ is the weight matrix for layer j, $\mathbf{x}^{\{k\}[j]}$ is the input (image) to layer j of the network given network input $\mathbf{x}^{\{k\}}$, and $f^{(2)}(z)$ is the 2^{nd} derivative of $f(z)$. With regard to training, $X_{\mathcal{O}}$ means "value of variable X before iteration" and X_Δ means "change in X due to iteration". Finally, where relevant, we use a superscript $X^{\boxtimes}$ to indicate that X relates to gradient descent (back-propagation), and $X^\bullet$ if X relates to RKBS regularized risk minimization.

4. Background

4.1. Reproducing Kernel Banach Space

A reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) is a Hilbert space $\mathcal{H}$ of functions $f : \mathbb{X} \to \mathbb{Y}$ for which the point evaluation functionals $\delta_{\mathbf{x}}(f) = f(\mathbf{x})$ are continuous. Thus, applying the Riesz representor theorem, there exists a kernel K such that:

$$f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$

Subsequently $K(\mathbf{x}, \mathbf{x}') = \langle K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot) \rangle$ and, by the Moore-Aronszajn theorem, K is uniquely defined by $\mathcal{H}$ and vice-versa. K is called the reproducing kernel, and the corresponding RKHS is denoted $\mathcal{H}_K$. RKHS based approaches

Table 1: Summary of the construction of reproducing kernel Banach space as per (Lin et al., 2022).

	Notation in present Paper	Notation used in (Lin et al., 2022)
Data space:	$\mathbb{X} \subset \mathbb{R}^n$	Ω_1 (input space)
Weight-step space:	$\mathbb{W}_\mathcal{O} \subset \prod_{j\in\mathbb{N}_D} \mathbb{R}^{H^{[j-1]}\times H^{[j]}} \times \mathbb{R}^{H^{[j]}}$	Ω_2 (weight space)
Data Feature map:	$\mathbf{\Phi}_\mathcal{O} : \mathbb{X} \to \mathcal{X}_\mathcal{O} \subset \mathbb{R}^{\infty\times m}$	$\Phi_1 : \Omega_1 \to \mathcal{W}_1$
Weight-step feature map:	$\mathbf{\Psi}_\mathcal{O} : \mathbb{W}_\mathcal{O} \to \mathcal{W}_\mathcal{O} \subset \mathbb{R}^{\infty\times m}$	$\Phi_2 : \Omega_2 \to \mathcal{W}_2$
Data Banach space:	$\mathcal{X}_\mathcal{O} = \mathrm{span}(\mathbf{\Phi}_\mathcal{O}(\mathbb{X}))$, $\|\cdot\|_{\mathcal{X}_\mathcal{O}} = \|\cdot\|_F$	$\mathcal{W}_1$ with norm $\|\cdot\|_{\mathcal{W}_1}$
Weight-step Banach space:	$\mathcal{W}_\mathcal{O} = \mathrm{span}(\mathbf{\Psi}_\mathcal{O}(\mathbb{W}_\mathcal{O}))$, $\|\cdot\|_{\mathcal{W}_\mathcal{O}} = \|\cdot\|_F$	$\mathcal{W}_2$ with norm $\|\cdot\|_{\mathcal{W}_2}$
Bilinear form:	$\langle \mathbf{\Omega}, \mathbf{\Xi} \rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}} = \mathrm{diag}\left(\mathbf{\Omega}^{\mathrm{T}}\mathbf{\Xi}\right)$	$\langle \cdot,\cdot\rangle_{\mathcal{W}_1\times\mathcal{W}_2} : \mathcal{W}_1 \times \mathcal{W}_2 \to \mathbb{Y}$

have gained popularity as they are well suited to many aspects of machine learning (Steinwart & Christman, 2008; Shawe-Taylor & Cristianini, 2004). The inner product structure enables the kernel trick, and the kernel is readily understood as a similarity measure. Furthermore, the structure of RKHSs has led to a rich framework of complexity analysis and generalization bounds (Steinwart & Christman, 2008; Shawe-Taylor & Cristianini, 2004). More recently neural tangent kernels were introduced (Jacot et al., 2018), allowing RKHS theory to be applied to the neural network training in the over-parametrized regime.

In an effort to introduce a richer set of geometrical structures into RKHS theory, reproducing kernel Banach spaces (RKBSs) generalize RKHSs by starting with a Banach space of functions (Der & Lee, 2007; Zhang et al., 2009; Song et al., 2013; Xu & Ye, 2014; Lin et al., 2022) etc. Precisely:

Definition 1 (Reproducing kernel Banach space (RKBS - (Lin et al., 2022)))**.** A reproducing kernel Banach space $\mathcal{B}$ on a set $\mathbb{X}$ is a Banach space of functions $f : \mathbb{X} \to \mathbb{Y}$ such that every point evaluation $\delta_\mathbf{x} : \mathcal{B} \to \mathbb{Y}$, $\mathbf{x} \in \mathbb{X}$, on $\mathcal{B}$ is continuous (so $\forall \mathbf{x} \in \mathbb{X}\ \exists C_\mathbf{x} \in \mathbb{R}_+$ such that $|\delta_\mathbf{x}(f)| = |f(\mathbf{x})| \leq C_\mathbf{x}\|f\|_\mathcal{B}\ \forall f \in \mathcal{B}$).

There are several distinct approaches to RKBS construction. In the present context however we find the approach of (Lin et al., 2022, Theorem 2.1) most convenient. Given the components outlined in Table 1, and assuming that $\mathbf{\Phi}_\mathcal{O}(\mathbb{X})$ is dense in $\mathcal{X}_\mathcal{O}$ and that $\mathbf{\Psi}_\mathcal{O}(\mathbb{W}_\mathcal{O})$ is dense in $\mathcal{W}_\mathcal{O}$, we define the reproducing kernel Banach space $\mathcal{B}_\mathcal{O}$ on $\mathbb{X}$ as:

$$\begin{array}{c}\mathcal{B}_\mathcal{O} = \left\{ \langle \mathbf{\Phi}_\mathcal{O}(\cdot), \mathbf{\Omega}\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}} \middle| \mathbf{\Omega} \in \mathcal{W}_\mathcal{O}\right\} \\ \text{where } \left\|\langle \mathbf{\Phi}_\mathcal{O}(\cdot), \mathbf{\Omega}\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}}\right\|_{\mathcal{B}_\mathcal{O}} = \|\mathbf{\Omega}\|_{\mathcal{W}_\mathcal{O}}\end{array} \tag{1}$$

with reproducing Banach kernel:

$$K_\mathcal{O}(\mathbf{x}, \mathbf{W}_\Delta) = \langle \mathbf{\Phi}_\mathcal{O}(\mathbf{x}), \mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}} \tag{2}$$

5. Setup and Assumptions

We assume a fully-connected, D-layer feedforward neural network $\mathbf{f} : (\mathbb{X} \subseteq \mathbb{R}^n) \to (\mathbb{Y} \subseteq \mathbb{R}^m)$ with layers of widths $H^{[0]}, H^{[1]}, \ldots, H^{[D-1]}$, where $H^{[D-1]} = m$ and we define $H^{[-1]} = n$. We assume layer $j \in \mathbb{N}_D$ ($j \in \mathbb{N}_D$ throughout) contains only neurons with activation function $\tau^{[j]} : \mathbb{R} \to \mathbb{R}$. The network is defined recursively $\forall j \in \mathbb{N}_D$:

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathbf{x}^{[D]} \in \mathbb{R}^{H^{[D-1]}} \\ \mathbf{x}^{[j+1]} &= \tau^{[j]}(\tilde{\mathbf{x}}^{[j]}) \in \mathbb{R}^{H^{[j]}} \\ \tilde{\mathbf{x}}^{[j]} &= \tfrac{1}{\sqrt{H^{[j]}}}\mathbf{W}^{[j]\mathrm{T}}\mathbf{x}^{[j]} + \alpha^{[j]}\mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}} \\ \mathbf{x}^{[0]} &= \mathbf{x} \in \mathbb{X} \subset \mathbb{R}^{H^{[-1]}}\ (H^{[-1]} = n) \end{aligned} \tag{3}$$

where $\mathbf{W}^{[j]} \in \mathbb{R}^{H^{[j-1]}\times H^{[j]}}$ and $\mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}}$ are weights and biases, which we summarise as $\mathbf{W} \in \mathbb{W}$, and $\alpha^{[j]} \in \mathbb{R}_+$ are fixed. The set of functions of this form is denoted $\mathcal{F}$.

We assume the goal of training is to take a training set and find weights and biases to minimize the empirical risk:

$$\begin{aligned} \mathbf{W}^\star &= \underset{\mathbf{W}\in\mathbb{W}}{\operatorname{argmin}}\, R_E(\mathbf{W}, \mathbb{D}) \\ R_E(\mathbf{W}, \mathbb{D}) &= \textstyle\sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_\mathbf{W}\left(\mathbf{x}^{\{k\}}\right)\right) \end{aligned} \tag{4}$$

where $\mathbf{f}_\mathbf{W}$ is a network of the form (3) with weights and biases $\mathbf{W}$, $\mathbb{D} = \{(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}) \in \mathbb{X} \times \mathbb{Y} : k \in \mathbb{N}_N\}$ is a training set ($k \in \mathbb{N}_N$ throughout), and $E : \mathbb{X}\times\mathbb{Y}\times\mathbb{R}^m \to \mathbb{R}$ is an error function defining the purpose of the network.

We make the following technical assumptions:

1. Input space: $\mathbb{X} = [-1, 1]^n$.
2. Error function: $E : \mathbb{X} \times \mathbb{Y} \times \mathbb{R}^m \to \mathbb{R}$ is $\mathcal{C}^1$ and L_E-Lipschitz in its third argument.
3. Activation functions: for all $j \in \mathbb{N}_D$, $\tau^{[j]} : \mathbb{R} \to [-1, 1]$ is bounded, $\mathcal{C}^\infty$, and has a power-series representations with region of convergence (ROC) at least $\rho^{[j]} \in \mathbb{R}_+$ around all $z \in \mathbb{R}$.
4. Weight non-triviality: for all $j \in \mathbb{N}_D$, $\mathbf{W}^{[j]} \neq \mathbf{0}$ at all times during training.[1]

[1] Networks that do not meet this requirement have a constant output independent of input $\mathbf{x}$. We do not consider this a restrictive assumption as it is highly unlikely that a randomly initialized network trained with a typical training set will ever reach this state.

5. Weight initialization: we assume LeCun initialization, so for all $j \in \mathbb{N}_D$, $W^{[j]}_{i_j, i_{j+1}}, b^{[j]}_{i_{j+1}} \sim \mathcal{N}(0,1)$.
6. Training: we assume training is gradient descent (back-propagation) with learning rate $\eta \in \mathbb{R}_+$.

These assumptions are necessary to make our analysis possible. In practice we may extend our analysis to unbounded activation functions by relaxing assumption 3 to $\tau^{[j]} : \mathbb{R} \to [-M^{[j]}, M^{[j]}]$, where $M^{[j]}$ exceeds the largest possible input to neurons in layer j (for example assuming weights are bounded in prior layers and using the maximum output $M^{[j-1]}$ of the previous layer). Note, however, that this will restrict the size of weight step (due to one iteration of back-propagation) that can be modelled by our method, which will scale inversely with $M^{[j]}$ - see appendix for full details. The continuity assumption in assumption 3 is more difficult to overcome, as this is fundamental for our analysis. Thus our method cannot be applied to e.g. ReLU (it may be possible to overcome this restriction by using an (arbitrarily accurate) polynomial approximation, but it is unclear whether this is mathematically reasonable, particularly when a weight-step traverses the point of non-smoothness).

5.1. Back-Propagation Training

As stated above, we assume the network is trained using back-propagation (gradient descent) (Goodfellow et al., 2016). This is an iterative approach. An iteration starts with initial weights and biases $\mathbf{W}_{\mathcal{O}} \in \mathbb{W}$. A weight-step:

$$\mathbf{W}^{\boxast}_{\Delta} = -\eta \, \tfrac{\partial}{\partial \mathbf{W}} \textstyle\sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right)\Big|_{\mathbf{W}=\mathbf{W}_{\mathcal{O}}}$$

is calculated, and weights and biases are updated as $\mathbf{W} = \mathbf{W}_{\mathcal{O}} + \mathbf{W}^{\boxast}_{\Delta}$. Our notational convention for activations before an iteration, and the subsequent change due to a weight step, are given in figure 1. The weight-step is (Goodfellow et al., 2016) (see appendix B for a derivation):

$$\begin{aligned} \mathbf{W}^{[j]\boxast}_{\Delta:i_{j+1}} &= -\frac{\eta}{\sqrt{H^{[D-1]} H^{[D-2]} \ldots H^{[j+1]}}} \sum_k \gamma^{\{k\}[j]}_{\mathcal{O} i_{j+1}} \frac{\mathbf{x}^{\{k\}[j]}_{\mathcal{O}}}{\sqrt{H^{[j]}}} \\ b^{[j]\boxast}_{\Delta:i_{j+1}} &= -\frac{\eta}{\sqrt{H^{[D-1]} H^{[D-2]} \ldots H^{[j+1]}}} \sum_k \gamma^{\{k\}[j]}_{\mathcal{O} i_{j+1}} \alpha^{[j]} \end{aligned} \tag{5}$$

for all $j \in \mathbb{N}_D, i_{j+1}$ where, recursively $\forall j \in \mathbb{N}_{D-1}$:

$$\begin{aligned} \gamma^{\{k\}[j-1]}_{\mathcal{O} i_j} &= \sum_{i_{j+1}} \gamma^{\{k\}[j]}_{\mathcal{O} i_{j+1}} W^{[j]}_{\mathcal{O} i_j, i'_{j+1}} \tau^{[j-1](1)}\left(\tilde{x}^{\{k\}[j-1]}_{\mathcal{O} i_j}\right) \\ \gamma^{\{k\}[D-1]}_{\mathcal{O} i_D} &= \nabla_{i_D} E\left(\ldots^{\{k\}}\right) \tau^{[D-1](1)}\left(\tilde{x}^{\{k\}[D-1]}_{\mathcal{O} i_D}\right) \end{aligned}$$

Note that the change in bias $\mathbf{b}^{[j]}_{\Delta}$ is proportional to $\alpha^{[j]}$.

6. Analysis of a Single Iteration

In the first phase of our analysis we consider the change in neural network behaviour resulting from a small in weights and biases (a weight-step). The overall training sequence is readily extrapolated from this as per section 7. Our first goal is to rewrite the neural network after a training iteration as:

$$\mathbf{f}\left(\mathbf{x}\right) = \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right) + \mathbf{f}_{\Delta}\left(\mathbf{x}\right) \tag{6}$$

where $\mathbf{f}_{\mathcal{O}} = \mathbf{f}_{\mathbf{W}_{\mathcal{O}}} : (\mathbb{X} \subset \mathbb{R}^n) \to (\mathbb{Y} \subset \mathbb{R}^m)$ is the neural network before the iteration and $\mathbf{f}_{\Delta} : (\mathbb{X} \subset \mathbb{R}^n) \to \mathbb{R}^m$ is the change in network behaviour due to the change $\mathbf{W} = \mathbf{W}_{\mathcal{O}} \to \mathbf{W} = \mathbf{W}_{\mathcal{O}} + \mathbf{W}_{\Delta}$ in weights and biases for this iteration, as detailed in Figure 1, so that:

$$\mathbf{f}_{\Delta}\left(\mathbf{x}\right) = \left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \tag{7}$$

where:

$$\begin{aligned} \mathbf{\Phi}_{\mathcal{O}} &: \mathbb{X} \to \mathcal{X}_{\mathcal{O}} = \mathrm{span}\left(\mathbf{\Phi}_{\mathcal{O}}\left(\mathbb{X}\right)\right) \subset \mathbb{R}^{\infty \times m} \\ \mathbf{\Psi}_{\mathcal{O}} &: \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} = \mathrm{span}\left(\mathbf{\Psi}_{\mathcal{O}}\left(\mathbb{W}_{\mathcal{O}}\right)\right) \subset \mathbb{R}^{\infty \times m} \end{aligned} \tag{8}$$

are feature maps determined entirely by the structure of the network (number and width of layers, activation functions) and the initial weights and biases $\mathbf{W}_{\mathcal{O}}$; and $\langle \mathbf{\Xi}, \mathbf{\Omega} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \mathrm{diag}\left(\mathbf{\Xi}^{\mathrm{T}} \mathbf{\Omega}\right)$ is a bilinear form. Subsequently our second goal is to derive kernels and norms from these feature maps to allow us to study their convergence properties, and we finish by proving an equivalence between the weight-step $\mathbf{W}^{\boxast}_{\Delta}$ due to a single step of back-propagation and the analogous weight-step $\mathbf{W}^{\bullet}_{\Delta}$ that minimizes the (RKBS) regularized risk.

6.1. Contribution 1: Feature-Map Expansion

In this section we derive appropriate feature maps to express the change in neural network behaviour for a finite weight-step. Our approach is simple in principle but technical, so details are reserved for appendix B. Roughly speaking however, we begin by noting that, for a smooth activation function $\tau^{[j]} : \mathbb{R} \to \mathbb{R}$, $z \in \mathbb{R}$ and finite-dimensional vectors $\mathbf{c}, \mathbf{c}'$ whose inner product lies in the radius of convergence $\rho^{[j]}$ (so that $|\langle \mathbf{c}, \mathbf{c} \rangle| < \rho^{[j]}$), the power-series representation of $\tau^{[j]}$ about z can be written

$$\tau^{[j]}\left(z + \langle \mathbf{c}, \mathbf{c} \rangle\right) = \tau^{[j]}\left(z\right) + \left\langle \boldsymbol{\varrho}\left(\mathbf{g}^{[j]}\left(z\right), \mathbf{c}\right), \boldsymbol{\varrho}\left(\mathbf{1}_{\infty}, \mathbf{c}'\right)\right\rangle$$

where:

$$\begin{aligned} \boldsymbol{\varrho}(\mathbf{a}, \mathbf{d}) &= \left[\, a_0 \mathbf{d}^{\otimes 1 \mathrm{T}} \; a_1 \mathbf{d}^{\otimes 2 \mathrm{T}} \; a_2 \mathbf{d}^{\otimes 3 \mathrm{T}} \; \ldots \,\right]^{\mathrm{T}} \\ \mathbf{g}^{[j]}(z) &= \left[\, \tfrac{1}{1!} \tau^{[j](1)}(z) \; \tfrac{1}{2!} \tau^{[j](2)}(z) \; \tfrac{1}{3!} \tau^{[j](3)}(z) \; \ldots \,\right]^{\mathrm{T}} \end{aligned}$$

Given an input $\mathbf{x}$, starting at layer 0 and working forward, and with reference to Figure 1, we can write the change $\mathbf{x}^{[1]}_{\Delta}$ in the output of layer 0 due to the weight-step $\mathbf{W}_{\Delta}$ as:

$$\mathbf{x}^{[1]}_{\Delta} = \left[\left\langle \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}\left(\mathbf{x}\right), \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}\left(\mathbf{W}_{\Delta}\right)\right\rangle\right]_{i_1}$$

where:

$$\begin{aligned} \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{x}) &= \boldsymbol{\varrho}\left(\mathbf{g}^{[0]}\left(\tilde{x}^{[0]}_{\mathcal{O} i_1}\right), \mu^{[0]}_{i_1} \begin{bmatrix} \frac{1}{\sqrt{2}} \alpha^{[0]} \\ \frac{1}{\sqrt{2 H^{[0]}}} \mathbf{x}_{\mathcal{O}} \end{bmatrix}\right) \\ \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{W}_{\Delta}) &= \boldsymbol{\varrho}\left(\mathbf{1}_{\infty}, \frac{1}{\mu^{[0]}_{i_1}} \begin{bmatrix} \sqrt{2} b^{[0]}_{\Delta i_1} \\ \sqrt{2} \mathbf{W}^{[0]}_{\Delta:i_1} \end{bmatrix}\right) \end{aligned} \tag{9}$$

$$
\left(\begin{array}{l} \textbf{Before Iteration} \\ \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right) = \mathbf{x}_{\mathcal{O}}^{[D]} \in \mathbb{Y} \\ \mathbf{x}_{\mathcal{O}}^{[j+1]} = \tau^{[j]}\left(\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}\right) \\ \tilde{\mathbf{x}}_{\mathcal{O}}^{[j]} = \frac{1}{\sqrt{H^{[j]}}}\mathbf{W}_{\mathcal{O}}^{[j]\mathrm{T}}\mathbf{x}_{\mathcal{O}}^{[j]} + \alpha^{[j]}\mathbf{b}_{\mathcal{O}}^{[j]} \\ \mathbf{x}_{\mathcal{O}}^{[0]} = \mathbf{x} \in \mathbb{X} \end{array} \right) + \left(\begin{array}{l} \textbf{Weight-step Change} \\ \mathbf{f}_{\Delta}\left(\mathbf{x}\right) = \mathbf{f}\left(\mathbf{x}\right) - \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right) \\ \mathbf{x}_{\Delta}^{[j+1]} = \mathbf{x}^{[j+1]} - \mathbf{x}_{\mathcal{O}}^{[j+1]} \\ \tilde{\mathbf{x}}_{\Delta}^{[j]} = \tilde{\mathbf{x}}^{[j]} - \tilde{\mathbf{x}}_{\mathcal{O}}^{[j]} \\ \mathbf{x}_{\Delta}^{[0]} = \mathbf{x} \in \mathbf{0}_n \end{array} \right) = \left(\begin{array}{l} \textbf{After Iteration} \\ \mathbf{f}(\mathbf{x}) = \mathbf{x}^{[D]} \in \mathbb{Y} \\ \mathbf{x}^{[j+1]} = \tau^{[j]}\left(\tilde{\mathbf{x}}^{[j]}\right) \\ \tilde{\mathbf{x}}^{[j]} = \frac{1}{\sqrt{H^{[j]}}}\mathbf{W}^{[j]\mathrm{T}}\mathbf{x}^{[j]} + \alpha^{[j]}\mathbf{b}^{[j]} \\ \mathbf{x}^{[0]} = \mathbf{x} \in \mathbb{X} \end{array} \right)
$$

Figure 1: Definition of terms for neural network before and after an iteration.

where we note that both feature maps have a finite radius of convergence. The feature maps are parameterised by the scale factors $\mu_{i_1}^{[0]} \in \mathbb{R}_+$ whose role is mainly technical, insofar as they will allow us to show equivalence between RKBS regularized risk minimization and back-propagation.[2] Their exact value (beyond existence) is unimportant here.

The process is repeated for subsequent layers (see appendix B for details). After working through all layers:

$$\mathbf{f}_{\Delta}\left(\mathbf{x}\right) = \left\langle \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}} \tag{10}$$

where $\boldsymbol{\Phi}_{\mathcal{O}} = \boldsymbol{\Phi}_{\mathcal{O}}^{[D-1]}$, $\boldsymbol{\Psi}_{\mathcal{O}} = \boldsymbol{\Psi}_{\mathcal{O}}^{[D-1]}$, and $\forall j \in \mathbb{N}_D \backslash \{0\}$:

$$
\begin{aligned}
\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x}) &= \boldsymbol{\varrho}\left(\mathbf{g}^{[j]}\left(\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}\right), \mu_{i_{j+1}}^{[j]} \begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[j]} \\ \frac{1}{\sqrt{H^{[j]}}}\mathbf{x}_{\mathcal{O}}^{[j]} \end{bmatrix} \\ \left[\frac{\omega_{i_j}^{[j]} W_{\mathcal{O}i_j,i_{j+1}}^{[j]}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]}\sqrt{H^{[j]}}}\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right]_{i_j} \\ \left[\frac{\omega_{i_j}^{[j]}}{\sqrt{H^{[j]}}}\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right)\right]_{i_j} \end{bmatrix} \right) \\
\boldsymbol{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{W}_{\Delta}) &= \boldsymbol{\varrho}\left(\mathbf{1}_{\infty}, \frac{1}{\mu_{i_{j+1}}^{[j]}} \begin{bmatrix} \begin{bmatrix} \sqrt{2}b_{\Delta i_{j+1}}^{[j]} \\ \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \end{bmatrix} \\ \left[\frac{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]}}{\omega_{i_j}^{[j]}}\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_{\Delta}\right)\right]_{i_j} \\ \left[\frac{W_{\Delta i_j,i_{j+1}}^{[j]}}{\omega_{i_j}^{[j]}}\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_{\Delta}\right)\right]_{i_j} \end{bmatrix} \right)
\end{aligned} \tag{11}
$$

recursively $\forall j \in \mathbb{N}_D \backslash \{0\}$ which are parameterised by scale factors $\mu_{i_{j+1}}^{[j]} \in \mathbb{R}_+$ and shadow weights $\omega_{i_j}^{[j]}, \widetilde{\omega}_{i_j,i_{j+1}}^{[j]} \in \mathbb{R}_+$, which play a role in the equivalence between RKBS regularized risk minimization and back-propagation (otherwise their exact values are unimportant).

6.2. Contribution 2: Induced Kernels and Norms

In the previous section we established that, as a result of a single weight-step $\mathbf{W}_{\Delta}$, we can write:

$$
\begin{aligned}
\mathbf{f}\left(\mathbf{x}\right) &= \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right) + \mathbf{f}_{\Delta}\left(\mathbf{x}\right) \\
\mathbf{f}_{\Delta}\left(\mathbf{x}\right) &= \left\langle \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}
\end{aligned}
$$

where $\mathbf{f}_{\mathcal{O}} : \mathbb{X} \to \mathbb{Y}$ is the neural network pre-iteration and the feature maps $\boldsymbol{\Phi}_{\mathcal{O}}$, $\boldsymbol{\Psi}_{\mathcal{O}}$ are feature maps (9-11). Using these, we induce kernels on $\mathbb{X}$, $\mathbb{W}_{\mathcal{O}}$ using the kernel trick:

$$
\begin{aligned}
\mathbf{K}_{\mathcal{X}_{\mathcal{O}}}\left(\mathbf{x}, \mathbf{x}'\right) &= \left\langle \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}'\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{X}_{\mathcal{O}}} \\
&= \boldsymbol{\Phi}_{\mathcal{O}}^{\mathrm{T}}\left(\mathbf{x}\right)\boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}'\right) \\
\mathbf{K}_{\mathcal{W}_{\mathcal{O}}}\left(\mathbf{W}_{\Delta}, \mathbf{W}'_{\Delta}\right) &= \left\langle \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right), \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}'_{\Delta}\right)\right\rangle_{\mathcal{W}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}} \\
&= \boldsymbol{\Psi}_{\mathcal{O}}^{\mathrm{T}}\left(\mathbf{W}_{\Delta}\right)\boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}'_{\Delta}\right)
\end{aligned}
$$

We call these kernels *neural neighbourhood kernels* (NNK) as they describe the similarity structure in the finite neighbourhood of the $\mathbf{W}_{\mathcal{O}}$ (c/f NTK, which is the behaviour tangent to, or in the infinitesimal neighbourhood of, $\mathbf{W}_{\mathcal{O}}$). These matrix-valued kernels are symmetric and positive definite by construction, and could potentially be used (transfered) in support vector machines (SVMs) or similar kernel-based methods, measuring similarity on $\mathbb{X}$ and $\mathbb{W}_{\mathcal{O}}$, respectively. Similarly, we induce a Banach kernel:

$$
\begin{aligned}
\mathbf{K}_{\mathcal{O}}\left(\mathbf{x}, \mathbf{W}'_{\Delta}\right) &= \left\langle \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}'_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}} \\
&= \mathrm{diag}\left(\mathbf{f}_{\Delta}\left(\mathbf{x}\right)\right)
\end{aligned} \tag{12}
$$

which is trivially the change in the network output $\mathbf{f}_{\Delta}(\mathbf{x})$ under weight-step $\mathbf{W}_{\Delta}$ for input vector $\mathbf{x}$, diagonalised.

The precise form of the neural neighbourhood kernels is rather complicated (the derivation is straightforward but resulting recursive equation is very long). The NNK is the (un-approximated) analogue of the NTK. However while neural networks evolve - to first order - in the RKHS defined by the NTK, the same is not true of the NNK. Indeed, it is not difficult to see that *RKHS regularization using the NNK will always result in a weight vector that is* not *the image of a weight-step under* $\boldsymbol{\Psi}_{\mathcal{O}}$ *(i.e. RKHS theory is insufficient, and we need RKBS theory to proceed).* Thus, as the NNK is not the main focus of our paper will not reproduce them in the body of the paper - the interested reader can find them in appendix C. Using these induced kernels we may obtain expressions for the norms of the images of $\mathbb{X}$ and $\mathbb{W}_{\mathcal{O}}$ in feature space:

$$
\begin{aligned}
\left\|\boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right)\right\|_{\mathcal{X}_{\mathcal{O}}}^2 &= \left\|\boldsymbol{\Phi}_{\mathcal{O}}^{[D-1]}\left(\mathbf{x}\right)\right\|_F^2 \\
&= \mathrm{Tr}\left(\mathbf{K}_{\mathcal{X}_{\mathcal{O}}}\left(\mathbf{x}, \mathbf{x}\right)\right) \\
\left\|\boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\|_{\mathcal{W}_{\mathcal{O}}}^2 &= \left\|\boldsymbol{\Psi}_{\mathcal{O}}^{[D-1]}\left(\mathbf{W}_{\Delta}\right)\right\|_F^2 \\
&= \mathrm{Tr}\left(\mathbf{K}_{\mathcal{W}_{\mathcal{O}}}\left(\mathbf{W}_{\Delta}, \mathbf{W}_{\Delta}\right)\right)
\end{aligned} \tag{13}
$$

[2]See appendix for more discussion.

Once again, the precise form of these expressions is complicated (the derivation is straightforward but the answer is very long - details can be found in appendix C). The importance of these norms lies in deriving conditions on convergence of the feature maps. Defining the helper function:

$$\overline{\sigma}^{[j]}(\zeta) = \max_{z \in \mathbb{R}_+ \cup \{0\}} \left\{ \sum_{l=1}^{\infty} \left(\tfrac{1}{l!}\right)^2 \tau^{[j](l)}(z)\, \tau^{[j](l)}(z)\, \zeta^l \right\} \tag{14}$$

and the constants:

$$\begin{aligned} s^{[j]2} &= \begin{cases} \frac{1}{2}\alpha^{[j]2} + \frac{1}{2}\frac{H^{[-1]}}{H^{[0]}} & \text{if } j = 0 \\ \frac{1}{2}\alpha^{[j]2} + \frac{H^{[j-1]}}{H^{[j]}} & \text{otherwise} \end{cases} \\ t_{\Delta i_{j+1}}^{[j]2} &= \begin{cases} \left[2b_{\Delta i_1}^{[0]2} + 2\left\|\mathbf{W}_{\Delta:i_1}^{[0]}\right\|_2^2\right]_{i_1} & \text{if } j = 0 \\ \left[2b_{\Delta i_{j+1}}^{[j]2} + \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2\right]_{i_{j+1}} & \text{otherwise} \end{cases} \end{aligned} \tag{15}$$

which are, loosely speaking, surrogates for, respectively, the expansion/contraction (fanout) in width from layer j to layer $j-1$ and the size of the weight-step at layer j; in appendix C.6 we derive the following Lemmas (see corresponding proofs of Theorems 5 and 6 in the appendix):

Lemma 1. *Let $\epsilon_\phi^{[j]} \in (0,1)\ \forall j \in \mathbb{N}_D$. For a given neural network and initial weights $\mathbf{W}_\mathcal{O}$ define $\frac{\phi^{[j]}}{H^{[j]}} = \overline{\sigma}^{[j]}((1 - \epsilon_\phi^{[j]})\sqrt{\rho^{[j]}})\ \forall j \in \mathbb{N}_D$. If the scale factors satisfy:*

$$\mu_{i_{j+1}}^{[j]2} \leq \begin{cases} \dfrac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\rho^{[j]\frac{1}{2}}} + \frac{1}{H^{[j]}}\sum_{i_j}\omega_{i_j}^{[j]2}\left(\frac{W_{\mathcal{O}i_j,i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\frac{\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\rho^{[j]\frac{1}{2}}}} & \text{if } j>0 \\ \dfrac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} & \text{if } j=0 \end{cases}$$

$\forall j \in \mathbb{N}_D, i_{j+1}$ then $\left\|\mathbf{\Phi}_\mathcal{O}^{[j]}(\mathbf{x})\right\|_F^2 \leq \phi^{[j]}\ \forall \mathbf{x} \in \mathbb{X}$.

Lemma 2. *Let $\epsilon_\psi^{[j]} \in (0,1)\ \forall j \in \mathbb{N}_D$. For a given neural network and initial weights $\mathbf{W}_\mathcal{O}$ and weight-step $\mathbf{W}_\Delta$, if:*

$$\begin{aligned} & t_{\Delta i_1}^{[0]2} \leq \left(1 - \epsilon_\psi^{[0]}\right)\mu_{i_1}^{[0]2} \\ & t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2}{\omega_{i_j}^{[j]2}}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2}\right) \leq \left(1 - \epsilon_\psi^{[j]}\right)\mu_{i_{j+1}}^{[j]2} \end{aligned}$$

$\forall j \in \mathbb{N}_D, i_{j'+1}$ then $\left\|\mathbf{\Psi}_\mathcal{O}^{[j]}(\mathbf{W}_\Delta)\right\|_F^2 \leq H^{[j]}\frac{1-\epsilon_\psi^{[j]}}{\epsilon_\psi^{[j]}}$.

Lemma 1 allows us to place bounds on the scale factors to ensure that the feature map $\mathbf{\Phi}_\mathcal{O}^{[j]} : \mathbb{X} \to \mathcal{X}_\mathcal{O}$ is finite (well defined) for all $\mathbf{x} \in \mathbb{X}$. Lemma 2 is similar, but rather than bounding the scale factors it takes these as given (by Lemma 1) and places bounds on the size of the weight-step $\mathbf{W}_\Delta$ for which the feature map is $\mathbf{\Psi}_\mathcal{O}^{[j]} : \mathbb{W} \to \mathcal{W}_\mathcal{O}$ is finite (well defined). Taken together, therefore, they give some bound on the size of weight-step that can be modelled for a give neural network structure and initial weights $\mathbf{W}_\mathcal{O}$. However they say nothing directly about the shadow weights. In the next section we use these Lemmas to establish a link between the weight-step generated by gradient descent and learning in RKBS, as well as clarifying the size of weight-step which we can model using our construct.

6.3. Contribution 3: Equivalence of Gradient Descent and regularized Risk Minimization in RKBS

In our previous contributions we showed that the change $\mathbf{f}_\Delta$ in neural network behaviour can be represented as by the form (10) with feature maps (9,11), derived kernels from these feature maps and gave bounds on the size of the weight-step for which this representation is valid. In this section we use these results to establish a link between gradient descent learning in neural networks and regularized risk minimization in reproducing kernel Banach space (RKBS).

To begin, it is not difficult to see that the feature maps $\mathbf{\Phi}_\mathcal{O} : \mathbb{X} \to \mathcal{X}_\mathcal{O}$ and $\mathbf{\Psi}_\mathcal{O} : \mathbb{W}_\mathcal{O} \to \mathcal{W}_\mathcal{O}$ define a RKBS $\mathcal{B}_\mathcal{O}$ imbued with $\|\cdot\|_{\mathcal{B}_\mathcal{O}} = \|\cdot\|_F$ using (1) with the reproducing Banach kernel (12) deriving from (2):[3]

$$\begin{aligned} \mathbf{K}_\mathcal{O}(\mathbf{x}, \mathbf{W}'_\Delta) &= \langle \mathbf{\Phi}_\mathcal{O}(\mathbf{x}), \mathbf{\Psi}_\mathcal{O}(\mathbf{W}'_\Delta)\rangle_{\mathcal{X}_\mathcal{O} \times \mathcal{W}_\mathcal{O}} \\ &= \operatorname{diag}(\mathbf{f}_\Delta(\mathbf{x})) \end{aligned}$$

in terms of which the change $\mathbf{f}_\Delta^{\circledast}$ in the network's behaviour due to a back-progation iteration may be written:

$$\mathbf{f}_\Delta^{\circledast}(\cdot) = \mathbf{K}_\mathcal{O}\left(\cdot, \mathbf{W}_\Delta^{\circledast}\right)\mathbf{1}$$

For a given neural network with initial weights $\mathbf{W}_\mathcal{O}$, we assume that the weight-step is chosen using gradient descent.[4] An alternative approach might be to select a weight-step to minimize the regularized risk in RKBS, specifically:

$$\begin{aligned} \mathbf{W}_\Delta^{\bullet} &= \operatorname*{argmin}_{\mathbf{W}_\Delta \in \mathbb{W}_\mathcal{O}} R_\lambda(\mathbf{W}_\Delta) \\ R_\lambda(\mathbf{W}_\Delta) &= \lambda \|\mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\|_{\mathcal{W}_\mathcal{O}} + R_E(\mathbf{W}_\Delta + \mathbf{W}_\mathcal{O}, \mathbb{D}) \end{aligned} \tag{16}$$

where we call λ the trade-off coefficient. Larger trade-off coefficients favour smaller weight-steps, and vice-versa. The advantage of this form over the back-propagation derived weight-step is that we can directly apply complexity bounds etc. from RKBS theory, and then extend to the complete training process. This motivates us to ask:

For a given neural network with initial weights

[3] In the appendix we prove that these maps satisfy the relevant density requirements.

[4] Note that the training set $\mathbb{D}$ here is for this iteration only and may be a random subset of a larger training set.

and biases $\mathbf{W}_{\mathcal{O}}$, let $\mathbf{W}_{\Delta}^{\boxast}$ be the back-propagation weight-step (gradient descent with learning rate η) defined by (5), and let $\mathbf{W}_{\Delta}^{\bullet}$ be a weight-step solving the regularized risk minimization problem (16). Given the gradient-descent derived weight-step $\mathbf{W}_{\Delta}^{\boxast}$, can we select scale factors, shadow weights and trade-off parameter λ (as a function of $\mathbf{W}_{\Delta}^{\boxast}$) that would guarantee that $\mathbf{W}_{\Delta}^{\bullet} = \mathbf{W}_{\Delta}^{\boxast}$?

If the answer is yes (which we demonstrate) then we can gain understanding of back-propagation by analysing (16). Now, the solution to (16) must satisfy first-order optimality conditions (assuming differentiability for simplicity), so:

$$\tfrac{\partial}{\partial \mathbf{W}_{\Delta}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta}^{\bullet} \right) \right\|_{\mathcal{W}_{\mathcal{O}}} = \tfrac{-1}{\lambda} \tfrac{\partial}{\partial \mathbf{W}_{\Delta}} R_E(\mathbf{W}_{\Delta}^{\bullet} + \mathbf{W}_{\mathcal{O}}, \mathbb{D})$$

Note that *if* the gradient of the regularization term satisfies:

$$\tfrac{\partial}{\partial \mathbf{W}_{\Delta}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta}^{\boxast} \right) \right\|_{\mathcal{W}_{\mathcal{O}}} = \nu \mathbf{W}_{\Delta}^{\boxast}$$

for some $\nu \in \mathbb{R}_+$, and $\frac{1}{\lambda} = \eta\nu$, then $\mathbf{W}_{\Delta}^{\bullet} = \mathbf{W}_{\Delta}^{\boxast}$. Thus the question of whether there exists scaling factors, shadow weights and λ such that the regularized risk minimization weight-step corresponds to the gradient-descent weight-step for a specified learning rate η can be answered in the affirmative by proving the existence of canonical scalings, which we define as follows:

Definition 2 (Canonical Scaling)**.** For a given neural network, initial weights $\mathbf{W}_{\mathcal{O}}$ and weight step $\mathbf{W}_{\Delta}^{\boxast}$ generated by back-propagation, we define a *canonical scaling* to be a set of scaling factors and shadow weights for which $\frac{\partial}{\partial \mathbf{W}_{\Delta}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta}^{\boxast} \right) \right\|_{\mathcal{W}_{\mathcal{O}}} = \nu \mathbf{W}_{\Delta}^{\boxast}$ for $\nu \in \mathbb{R}_+$, and $\left\| \boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}^{\boxast}) \right\|_{\mathcal{W}_{\mathcal{O}}}, \left\| \boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x}) \right\|_{\mathcal{X}_{\mathcal{O}}} < \infty \; \forall \mathbf{x} \in \mathbb{X}$.

To prove the existence of canonical scaling and thus the key connection between gradient descent (back-propagation) and learning in RKBS, using (15), defining:

$$t_{\Delta i_{j+1}}^{[j]\boxast 2} = \begin{cases} 2b_{\Delta i_1}^{[0]\boxast 2} + 2\left\| \mathbf{W}_{\Delta:i_1}^{[0]\boxast} \right\|_2^2 & \text{if } j = 0 \\ 2b_{\Delta i_{j+1}}^{[j]\boxast 2} + 2\left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]\boxast} \right\|_2^2 & \text{otherwise} \end{cases}$$

$\forall j \in \mathbb{N}_D, i_{j+1}$, in the appendix we prove the following key result based on Lemmas 1 and 2:

Theorem 1. *Let $\epsilon, \chi \in (0, 1)$. For a given neural network with initial weights $\mathbf{W}_{\mathcal{O}}$, let $\mathbf{W}_{\Delta}^{\boxast}$ be the weight-step for this derived from back-propagation, assuming wlog that $\alpha^{[j]} \in \mathbb{R}_+$ is chosen such that $\forall j \in \mathbb{N}_{D-1}$:*[5]

$$\left\| \mathbf{W}_{\Delta}^{[j+1]\boxast} \right\|_F^2 = \chi \left\| \mathbf{t}_{\Delta}^{[j]\boxast} \right\|_\infty^2$$

[5]Note that $b_{\Delta i_{j+1}}^{[j]\boxast}$ is proportional to $\alpha^{[j]}$, so we can always increase $t_{\Delta i_{j+1}}^{[j]\boxast 2}$ to ensure the condition holds by adjusting $\alpha^{[j]}$.

Let $\epsilon_\psi = 1 - \frac{1}{1-\chi} \frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}} \|\mathbf{t}_{\Delta}^{[D-1]\boxast}\|_\infty^2$ and:[6]

$$\epsilon_\phi^{[j]} = 1 - \tfrac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \Bigg(\ldots$$

$$\ldots \frac{\frac{1}{1-\chi} \frac{\chi^{D-1-j} \epsilon_\psi \left(1-\epsilon_\phi^{[j+1]}\right) \sqrt{\rho^{[j+1]}}}{\left\| \mathbf{t}_{\Delta}^{[j+1]\boxast} \right\|_\infty^2} - \frac{\chi^{D-1-j} \epsilon_\psi}{1-\chi^{D-1-j} \epsilon_\psi} s^{[j+1]2}}{\max\limits_{i_{j+2}} \left\{ \frac{\frac{1}{H^{[j+1]}} \left\| \mathbf{W}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2}{\left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_\infty^2} + \frac{H^{[j]}}{H^{[j+1]}} \right\}} \Bigg)$$

$\forall j \in \mathbb{N}_{D-1}$, where $\epsilon_\phi^{[D-1]} = \epsilon$. If the weight-step satisfies $\|\mathbf{t}_{\Delta}^{[j]\boxast}\|_\infty^2 < B^{[j]2} \; \forall j \in \mathbb{N}_D$, where:

$$B^{[j]2} = \begin{cases} \dfrac{(1-\chi)^2}{\left(\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}} \right)} & \textit{if } j = D-1 \\ \dfrac{(1-\chi)^2 \left(1-\chi^{D-j-1}\epsilon_\psi\right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} \right)} & \textit{if } j < D-1 \end{cases} \tag{17}$$

then there exists of a canonical scaling:

$$\tfrac{\partial}{\partial \mathbf{W}_{\Delta}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta} \right) \right\|_{\mathcal{W}_{\mathcal{O}}} \Big|_{\mathbf{W}_{\Delta} = \mathbf{W}_{\Delta}^{\boxast}} = \nu \mathbf{W}_{\Delta}^{\boxast}$$
$$\nu = \tfrac{4}{\left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2} \tfrac{(1-\epsilon_\psi)(1-\chi)}{(\epsilon_\psi - \chi)^2} < \infty$$

where $\|\boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x})\|_F^2 \leq H^{[D-1]} \overline{\sigma}^{[D-1]}((1-\epsilon)\sqrt{\rho^{[D-1]}}) \; \forall \mathbf{x} \in \mathbb{X}$ and $\|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta})\|_F^2 \leq H^{[D-1]} \frac{1-\epsilon_\psi}{\epsilon_\psi}$.

This is proven as corollary 10 in the appendix. This theorem tells us that, for any set of initial weights $\mathbf{W}_{\mathcal{O}}$, for a sufficiently small weight-step $\mathbf{W}_{\Delta}^{\boxast}$ generated by back propagation, there exists a canonical scaling - i.e. a set of scaling factors and shadow weights such that the gradient-descent weight-step is exactly equivalent to the step generated by regularized RKBS learning using an appropriate trade-off parameter λ. Note that:

- The maximum step-size $B^{[j]}$ in layer j (up to near-identity scaling terms) is determined by the inverse fanout $1/s^{[j]}$ (which scales roughly as $H^{[j]}/H^{[j-1]}$) and the scaled radius of convergence $(1-\epsilon_\phi^{[j]})\sqrt{\rho^{[j]}}$ (which scales roughly inverse the the weight-step in subsequent layers). This bound will tend to get smaller as we move from the output layer back toward the input, but so too will the weight-steps in many cases due to the problem of vanishing gradients.
- The trade-off coefficient (degree of regularization) required by this canonical scaling is:

$$\lambda = \tfrac{1}{\eta\nu} = \tfrac{B^{[D-1]2}}{4\eta} \left(1 - \left(\tfrac{\left\| \mathbf{t}_{\Delta}^{[j]\boxast} \right\|_\infty}{B^{[D-1]}} \right)^2 \right)^2$$

[6]The positivity of ϵ_ψ is due to the constraints on $\|\mathbf{t}_{\Delta}^{[D-1]\boxast}\|_\infty^2$.

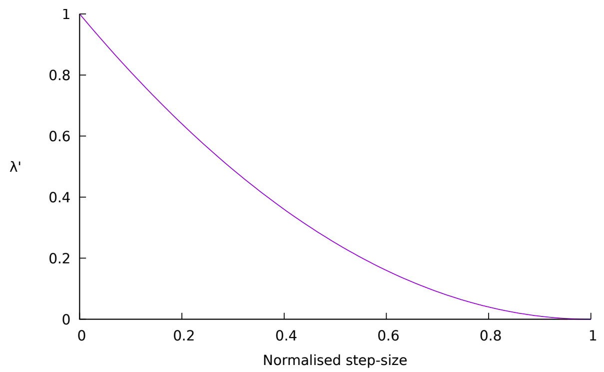

Figure 2: Normalized Canonical trade-off coefficient $\lambda' = \frac{4\eta}{B^{[D-1]2}}\lambda$ vs normalized gradient-descent step-size $\frac{\|\mathbf{t}_\Delta^{[j]\boxast}\|_\infty}{B^{[D-1]}}$.

as shown in figure 2. Note that (a) the degree of regularization required to generate an equivalent RKBS weight-step is inversely proportional to the learning rate used to generate the original back-propagation weight-step and (b) larger gradient-descent weight-steps are equivalent to less regularized RKBS weight-steps, as might be expected.

7. Application - Rademacher Complexity

Having established an exact representation of neural networks in finite neighbourhoods of weights and biases, established the link with RKBS theory and demonstrated that a gradient descent step is equivalent to a regularized step in RKBS by appropriate, a-posterior selection of scale factors and shadow weights (canonical scaling), we now consider an application of this framework to uniform convergence analysis using Rademacher complexity. The Rademacher complexity of a set $\mathcal{G}$ of real-valued functions is a measure of its capacity. Assuming training vectors $\mathbf{x}_i \sim \nu$ and Rademacher random variables $\epsilon_i \in \{-1, 1\}$, the Rademacher complexity of $\mathcal{G}$ is (Mendelson, 2003):

$$\mathcal{R}_N\left(\mathcal{G}\right) = \mathbb{E}_{\nu,\epsilon}\left[\sup_{f\in\mathcal{G}}\left|\frac{1}{N}\sum_i \epsilon_i f\left(\mathbf{x}_i\right)\right|\right]$$

This may be used in uniform convergence analysis to bound how quickly the empirical risk converges to the expected risk, typically of the form:

$$\left|\overline{R}\left(g\right) - R_E\left(g\right)\right| \leq c\mathcal{R}_N\left(\mathcal{G}\right) + \text{excess risk}$$

The following theorem demonstrates how our framework may be used to bound Rademacher complexity for a scalar-output neural network:

Theorem 2. *Let $\epsilon, \chi \in (0,1)$ and for a given neural network with initial weights $\mathbf{W}_\mathcal{O}$, and let $\mathbf{W}_\Delta^\boxast$ be the weight-step for this derived from back-propagation satisfying the conditions set out in corollary 10. Then $f_\Delta^\boxast \in \mathcal{F}^\bullet$, where the Rademacher complexity of $\mathcal{F}^\bullet$ is bounded as:*

$$\frac{\mathcal{R}_N(\mathcal{F}^\bullet)}{H^{[D-1]}} \leq \sqrt{\frac{\overline{\sigma}^{[D-1]}\left((1-\epsilon)\sqrt{\rho^{[D-1]}}\right)}{N}\frac{\left\|\mathbf{t}_\Delta^{[D-1]\boxast}\right\|_\infty^2}{\frac{1}{1-\chi}B^{[D-1]2}-\left\|\mathbf{t}_\Delta^{[D-1]\boxast}\right\|_\infty^2}}$$

The proof of this theorem is a straightforward application of theorem 1 (see proof of theorem 11 in the appendix for details). Apart from the usual $\frac{1}{N}$ scaling this theorem is very different from typical bounds on Rademacher complexity, as it directly bounds the complexity using the size of the weight-step for a single iteration of gradient descent. To gain more insight into this bound we take the following steps:

1. By the properties of Rademacher complexity (Bartlett & Mendelson, 2002, Theorem 12) the complexity of the trained neural network after T iterations may be bounded by a simple summation of the bounds on each step, so, assuming training is stopped at T iterations:

$$\frac{\mathcal{R}_N(\mathcal{F}^\bullet)}{H^{[D-1]}} \leq T\sqrt{\frac{\overline{\sigma}^{[D-1]}\left((1-\epsilon)\sqrt{\rho^{[D-1]}}\right)}{\frac{1}{1-\chi}N}\frac{\left\|\mathbf{t}_\Delta^{[D-1]\boxast}\right\|_\infty^2}{B^{[D-1]2}-\left\|\mathbf{t}_\Delta^{[D-1]\boxast}\right\|_\infty^2}}$$

   This supports the practice of using early-stopping to prevent overfitting in neural networks.[7]

2. The size of the back-propagation weight-step scales proportionally with the learning rate, so for sufficiently small learning rates we find, using (17):[8]

$$\frac{\mathcal{R}_N(\mathcal{F}^\bullet)}{H^{[D-1]}} \lessapprox T\sqrt{\frac{1}{N}\frac{\overline{\sigma}^{[D-1]}\left((1-\epsilon)\sqrt{\rho^{[D-1]}}\right)}{(1-\chi)(1-\epsilon)\sqrt{\rho^{[D-1]}}}}\left\|\mathbf{t}_\Delta^{[D-1]\boxast}\right\|_\infty s^{[D-1]}$$

   However as the learning rate increases the bound will increase at an accelerating rate.[9] This supports the idea that lower learning rates act as a form of regularization in neural network training.

3. Assuming $\tau^{[D-1]}$ is L_τ-Lipchitz, recalling the form of the back-propagation step (5) and using the assumption that the error function is L_E-Lipschitz, we see:

$$\begin{aligned}\left\|\mathbf{W}_{\Delta:i_D}^{[D-1]\boxast}\right\|_2^2 &\leq \eta^2 L_E^2 L_\tau^2 N^2 \tfrac{H^{[D-2]}}{H^{[D-1]}} \\ \left|b_{\Delta:i_D}^{[D-1]\boxast}\right|^2 &\leq \eta L_E^2 L_\tau^2 N^2 \alpha^{[D-1]2}\end{aligned}$$

   and hence:

$$\begin{aligned}s^{[D-1]2} &= \alpha^{[D-1]2} + \tfrac{H^{[D-2]}}{H^{[D-1]}} \\ \left\|\mathbf{t}_\Delta^{[j]\boxast}\right\|_\infty^2 &\leq 2\eta L_E^2 L_\tau^2 N^2\left(2\alpha^{[D-1]2} + \tfrac{H^{[D-2]}}{H^{[D-1]}}\right)\end{aligned}$$

   Thus, assuming $\frac{H^{[D-2]}}{H^{[D-1]}} \gg \max\{1, \alpha^{[D-1]2}\}$, by (15):

[7] We suspect that a more careful analysis accounting for the overlap between the RKBSs for each step may give sub-linear dependence on T, but this is beyond the scope of the present work.

[8] We use the approximation $\frac{x}{a+x} \approx \frac{x}{a}$ for $0 < x \ll a$ here.

[9] Eventually the bound will become meaningless as the step-size exceeds the size we can model using an RKBS.

$$\frac{\mathcal{R}_N(\mathcal{F}^\bullet)}{H^{[D-1]}} \lessapprox T\sqrt{\frac{2}{N}\frac{\overline{\sigma}^{[D-1]}\left((1-\epsilon)\sqrt{\rho^{[D-1]}}\right)}{(1-\chi)(1-\epsilon)\sqrt{\rho^{[D-1]}}}\eta L_E L_\tau N\frac{H^{[D-2]}}{H^{[D-1]}}}$$

4. Finally, following standard practice and scaling the learning rate as $\eta = \frac{s}{N\sqrt{H^{[D-2]}}}$, $0 < s \ll 1$, we obtain:

$$\mathcal{R}_N\left(\mathcal{F}^\bullet\right) \lessapprox sT\sqrt{\frac{2H^{[D-2]}}{N}\frac{\overline{\sigma}^{[D-1]}\left((1-\epsilon)\sqrt{\rho^{[D-1]}}\right)}{(1-\chi)(1-\epsilon)\sqrt{\rho^{[D-1]}}}}L_E L_\tau$$

Thus the complexity bound scales approximately as $\mathcal{O}(\sqrt{\frac{H^{[D-2]}}{N}})$ in the small learning rate limit. The impact of layer widths preceding this and the depth of the network are not apparent in the bound, which is perhaps unsurprising as we may view these layers are essentially a feature map feeding into the output layer, and hence, provided this map is sufficiently complex ((Kidger & Lyons, 2020) would appear to suggest that non-polynomial should suffice, but the extension to here is speculative), we would expect that the output dimension of the feature map should play a pivotal role compared to the internal details of the feature map itself. It is possible that the widths of earlier layers and network depth would be indirectly present as bounding the learning rate (s in our assumptions) for which this analysis holds, as may be glimpsed in theorem 1, but it is not clear how such factors could be formally brought forward in our bound.

Finally, the role of the helper function $\overline{\sigma}^{[D-1]}$ here depends explicitly on the form of the activation function $\tau^{[D-1]}$ in the output layer. We finish by considering two special cases:

- Linear output neuron: if we let $\rho^{[D-1]}$ be sufficiently large that it exceeds the output of the network for any training input then the linear activation $\tau^{[D-1]}(\zeta) = \zeta$ satisfies our assumptions in the relaxed case discussed in section 5.[10] Moreover for this choice $\overline{\sigma}^{[D-1]}(\zeta) = \zeta$ and $L_\tau = 1$, and hence:

$$\mathcal{R}_{\mathrm{lin}N}\left(\mathcal{F}^\bullet\right) \lessapprox sT\sqrt{\frac{2}{1-\chi}\frac{H^{[D-2]}}{N}}L_E \tag{18}$$

- Tanh output neuron: if $\tau^{[D-1]} = \tanh$ then $\rho^{[D-1]} = \frac{\pi}{2}$ and $L_\tau = 1$. The form of $\sigma_{z,z'}^{[D-1]}$ is non-trivial (see appendix F for details), but it is not difficult to see that $\overline{\sigma}^{[D-1]} = \sigma_{0,0}^{[D-1]}$ and, using the power-series about 0:

$$\overline{\sigma}_{\tanh}^{[D-1]}\left(\zeta\right) = \sum_{m=1}^{\infty}\left(\frac{2^{2m}\left(2^{2m}-1\right)B_{2m}}{(2m)!}\right)^2\zeta^{2m-1}$$

where B_{2m} are the Bernoulli numbers. Hence:

$$\mathcal{R}_{\tanh N}\left(\mathcal{F}^\bullet\right) \lessapprox sT\sqrt{\frac{2}{1-\chi}\frac{H^{[D-2]}}{N}\frac{\overline{\sigma}_{\tanh}^{[D-1]}\left((1-\epsilon)\sqrt{\frac{\pi}{2}}\right)}{(1-\epsilon)\sqrt{\frac{\pi}{2}}}}L_E$$

[10]Note that the output layer of the network does not feed in to subsequent layers, so the range of output from this layer can vary arbitrarily without pushing the power-series expansion of neurons in the subsequent layer outside of their ROC. Thus we can let the input and output of linear output layer neuron be arbitrarily large without compromising the requirements of our model.

We note that in general the Rademacher complexity bound will simply be the bound for linear output neurons (18) multiplied by an activation-function dependent scaling factor. Thus asymptotically these bounds are independent of the output layer activation function.

8. Conclusions and Future Directions

In this paper we have established a connection between neural network training using gradient descent and regularized learning in reproducing kernel Banach space. We have introduced an exact representation of the behaviour of neural networks as the weights and biases are varied in a finite neighbourhood of some initial weights and biases in terms of an inner product of two feature maps, one from data space to feature space, the other from weight-step space to feature space. Using this, we showed that the change in neural network behaviour due to a single iteration of back-propagation lies in a reproducing kernel Banach space, and moreover that the weight-step found by back-propagation can be exactly replicated through regularized risk minimization in RKBS. Subsequently we presented an upper bound on the Rademacher complexity of neural networks applicable to both the over- and under-parametrized regimes, and discussed how this bound depends on learning rate, dataset size, network width and the number of training iterations used.

With regard to future work we foresee a number of useful directions. First, the analysis should be extended to non-smooth activation functions such as ReLU, presumably by modifying the feature map using a representation other than a power series expansion. Second, the precise influence of the learning rate needs to be further explicated, along with other details of Theorem 1. With regard to the neural neighbourhood kernels themselves, it would be helpful if these could be reduced to closed form to allow them to be used in practice.[11] Finally, more work is needed to understand the impact of the depth of the network on this theory.

Acknowledgement:

This research was partially funded by the Australian Government through the Australian Research Council (ARC). Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

This research was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210102798). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

[11]A Taylor approximation the NNKs is readily obtained, but it is difficult to say how accurate this may be without further analysis.

References

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

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

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

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of ICML*, 2018.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pp. 8139–8148, 2019b.

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

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

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

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

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

Bartolucci, F., De Vito, E., Rosasco, L., and Vigogna, S. Understanding neural networks with reproducing kernel banach spaces. *arXiv preprint arXiv:2109.09710*, 2021.

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

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In Y., B., D., S., D., L. J., Williams, C. K. I., and Culotta, A. (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`.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853, International Convention Centre, Sydney, Australia, Aug 2017. PMLR.

Cortes, C. and Vapnik, V. Support vector networks. *Machine Learning*, 20(3):273–297, 1995.

Cristianini, N. and Shawe-Taylor, J. *An Introductino to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2005.

Daniely, A. Sgd learns the conjugate kernel class of the network. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (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`.

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (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`.

D'Aurizio, J. Taylor series expansion of tanh x. Mathematics Stack Exchange `https://math.stackexchange.com/q/1052926` (version: 2014-12-05), 2014.

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

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

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *Conference on Learning Representations*, 2019b.

Genton, M. G. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *COLT*, 2018.

Gönen, M. and Alpaydin, E. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

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

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

Herbrich, R. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, 2002.

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

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. In *Conference on learning theory*, pp. 2306–2327. PMLR, 2020.

Knapp, M. P. Sines and cosines of angles in arithmetic progression. *Mathematics magazine*, 82(5):371, 2009.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *In International Conference on Learning Representations*, 2018.

Li, C., Venturi, L., and Xu, R. Learning the kernel for classification and regression. *arXiv preprint arXiv:1712.08597*, 2017.

Li, Y. and Liang, Y. 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.

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

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *arXiv e-prints*, 2018.

Mendelson, S. A few notes on statistical learning theory. In Mendelson, S. and Smola, A. J. (eds.), *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia, February 11–22, 2002 Revised Lectures*, pp. 1–40. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12 (2):181–198, March 2001.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In Wallach, H., Larochelle, H., Beygelzimer, A., dé Buc, F., Fox, E., and Garnett, R. (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.

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

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

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

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5949–5958, 2017.

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

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *Proceedings of ICLR*, 2019.

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

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

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

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

Smola, A. J. and Schölkopf, B. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998. Technical Report 1064, GMD First, April 1997.

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

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

Steinwart, I. and Christman, A. *Support Vector Machines*. Springer, 2008.

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

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

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

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

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

Zhou, W., Veitch, V., Austern, M., Adams, R. P., and Orbanz, P. Nonvacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations (ICLR)*, 2019.

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

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

A. Derivations and Proofs.

Our goal in this supplement is to present all derivations and proofs relevant to our paper, and also any additional material and description that may be useful.

We assume a fully-connected, D-layer feedforward neural network $\mathbf{f} : (\mathbb{X} \subseteq \mathbb{R}^n) \to (\mathbb{Y} \subseteq \mathbb{R}^m)$ with layers of widths $H^{[0]}, H^{[1]}, \ldots, H^{[D-1]}$, where $H^{[D-1]} = m$ and we define $H^{[-1]} = n$. We assume layer $j \in \mathbb{N}_D$ (we use the convention $j \in \mathbb{N}_D$ throughout) is made up of neurons with the same activation function $\tau^{[j]} : \mathbb{R} \to \mathbb{R}$. The network is defined recursively:

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathbf{x}^{[D]} \in \mathbb{R}^{H^{[D-1]}} \\ \mathbf{x}^{[j+1]} &= \tau^{[j]}\left(\tilde{\mathbf{x}}^{[j]}\right) \in \mathbb{R}^{H^{[j]}} \\ \tilde{\mathbf{x}}^{[j]} &= \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}^{[j]\mathrm{T}} \mathbf{x}^{[j]} + \alpha^{[j]} \mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}} \qquad \forall j \in \mathbb{N}_D \\ \mathbf{x}^{[0]} &= \mathbf{x} \in \mathbb{X} \subset \mathbb{R}^{H^{[-1]}} \; (H^{[-1]} = n) \end{aligned} \tag{19}$$

where $\mathbf{W}^{[j]} \in \mathbb{R}^{H^{[j-1]} \times H^{[j]}}$ and $\mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}}$ are weights and biases, and $\alpha^{[j]} \in \mathbb{R}_+$ is a constant we will use later. We define the set of neural networks taking the form (19) as:

$$\mathcal{F} = \{ \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m | \, \mathbf{f} \text{ has form (19) given } \mathbf{W} \in \mathbb{W} \} \tag{20}$$

where $\mathbf{W}$ summarises the weights and biases in (19):

$$\mathbf{W} = \left(\left(\mathbf{W}^{[j]}, \mathbf{b}^{[j]} \right) \middle| \, j \in \mathbb{N}_D \right) \in \mathbb{W} = \prod_{j \in \mathbb{N}_D} \left(\mathbb{R}^{H^{[j-1]} \times H^{[j]}} \times \mathbb{R}^{H^{[j]}} \right)$$

Typically, the goal in neural network training is to take a training set and find weights and biases to minimise some measure of mismatch between the true training labels and the network's predictions. For simplicity let us assume we wish to minimise:

$$\mathbf{f}^\star = \operatorname*{argmin}_{\mathbf{f} \in \mathcal{F}} \textstyle\sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}\left(\mathbf{x}^{\{k\}} \right) \right) \tag{21}$$

or, equivalently:

$$\mathbf{W}^\star = \operatorname*{argmin}_{\mathbf{W} \in \mathbb{W}} \textstyle\sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}} \right) \right) \tag{22}$$

where $\mathbf{f}_{\mathbf{W}}$ is a network of the form (19) with weights and biases $\mathbf{W}$, $\mathbb{D} = \{ (\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}) \in \mathbb{X} \times \mathbb{Y} : k \in \mathbb{N}_N \}$ is a training set (we use the convention $k \in \mathbb{N}_N$ throughout), and $E : \mathbb{X} \times \mathbb{Y} \times \mathbb{R}^m \to \mathbb{R}$ is an error function defining the purpose of the network.

We make the following assumptions:

1. Input space: we assume $\mathbb{X} = [-M^{[-1]}, M^{[-1]}]^n$.
2. Error function: we assume the error function $E : \mathbb{X} \times \mathbb{Y} \times \mathbb{R}^m \to \mathbb{R}$ is $\mathcal{C}^1$ and L_E-Lipschitz in its third argument.
3. Activation functions: we assume the activation functions $\tau^{[j]} : \mathbb{R} \to [-M^{[j]}, M^{[j]}]$ are bounded and $\mathcal{C}^\infty$, and that $\tau^{[j]}$ has a power-series representations with region of convergence (ROC) at least $\rho^{[j]} \in \mathbb{R}_+$ around z for all $z \in \mathbb{R}$, $j \in \mathbb{N}_D$.
4. Weight non-triviality: we assume $\mathbf{W}^{[j]} \neq \mathbf{0}$ for all $j \in \mathbb{N}_D$ at all times during training.[12]
5. Weight initialization: we assume that initially $W^{[j]}_{i_j, i_{j+1}}, b^{[j]}_{i_{j+1}} \sim \mathcal{N}(0, 1)$ (LeCun initialization).
6. Training: we assume the network is trained using the back-propagation with learning rate $\eta \in \mathbb{R}_+$.

[12]Note that networks that do not meet this requirement have a constant output independent of input $\mathbf{x}$. We do not consider this a restrictive assumption as it is highly unlikely that a randomly initialised network trained with a typical training set will ever reach this state.

A.1. Review of Back-propagation

We now give a brief review of back-propagation training, which is a systematic implementation of gradient descent on the weights and biases $\mathbf{W}$ to solve (22). Let us consider a single training iteration, where we start with initial weights and biases $\mathbf{W}_{\mathcal{O}} \in \mathbb{W}$ and calculate a weight-step $\mathbf{W}_{\Delta}^{\boxast}$ so that, after this iteration, $\mathbf{W} = \mathbf{W}_{\mathcal{O}} + \mathbf{W}_{\Delta}^{\boxast}$, where:

$$\mathbf{W}_{\Delta}^{\boxast} = -\eta \frac{\partial}{\partial \mathbf{W}} \sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right)\big|_{\mathbf{W}=\mathbf{W}_{\mathcal{O}}}$$

Denote the network activation prior to the iteration given input $\mathbf{x}$ as:

$$\begin{aligned}
\mathbf{f}_{\mathcal{O}}(\mathbf{x}) &= \mathbf{x}_{\mathcal{O}}^{[D]} \\
\mathbf{x}_{\mathcal{O}}^{[j+1]} &= \tau^{[j]}\left(\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}\right) \\
\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]} &= \frac{1}{\sqrt{H^{[j]}}} \mathbf{W}_{\mathcal{O}}^{[j]\mathrm{T}} \mathbf{x}_{\mathcal{O}}^{[j]} + \alpha^{[j]} \mathbf{b}_{\mathcal{O}}^{[j]} \\
\mathbf{x}_{\mathcal{O}}^{[0]} &= \mathbf{x} \in \mathbb{X}
\end{aligned} \tag{23}$$

for all layers $j \in \mathbb{N}_D$. The back-propagation iteration for layer $D-1$ is a gradient descent step with learning rate η, namely:

$$\begin{aligned}
W_{\Delta i_{D-1}, i_D}^{[D-1]\boxast} &= -\eta \frac{\partial}{\partial W_{i_{D-1}, i_D}^{[D-1]}} \sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right)\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \sum_{k, i'_D} \nabla_{i'_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right) \frac{\partial}{\partial W_{i_{D-1}, i_D}^{[D-1]}} f_{i'_D}\left(\mathbf{x}^{\{k\}}\right)\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \sum_{k, i'_D} \nabla_{i'_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right) \frac{\partial}{\partial W_{i_{D-1}, i_D}^{[D-1]}} x_{i'_D}^{\{k\}[D]}\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \sum_{k, i'_D} \nabla_{i'_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right) \frac{\partial}{\partial W_{i_{D-1}, i_D}^{[D-1]}} \tau^{[D-1]}\left(\tilde{x}_{i'_D}^{\{k\}[D-1]}\right)\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \sum_{k, i'_D} \nabla_{i'_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right) \tau^{[D-1](1)}\left(\tilde{x}_{i'_D}^{\{k\}[D-1]}\right) \frac{\partial}{\partial W_{i_{D-1}, i_D}^{[D-1]}} \tilde{x}_{i'_D}^{\{k\}[D-1]}\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \frac{1}{\sqrt{H^{[D-1]}}} \sum_k \gamma_{\mathcal{O} i_D}^{\{k\}[D-1]} x_{\mathcal{O} i_{D-1}}^{\{k\}[D-1]}
\end{aligned}$$

where $\tau^{(r)}(z) = \frac{\partial^r}{\partial^r z} \tau(z)$ and:

$$\gamma_{\mathcal{O} i_D}^{\{k\}[D-1]} = \nabla_{i_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right) \tau^{[D-1](1)}\left(\tilde{x}_{\mathcal{O} i_D}^{\{k\}[D-1]}\right)$$

Subsequently:

$$\begin{aligned}
W_{\Delta i_{D-2}, i_{D-1}}^{[D-2]\boxast} &= -\eta \frac{\partial}{\partial W_{i_{D-2}, i_{D-1}}^{[D-2]}} \sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}\left(\mathbf{x}^{\{k\}}\right)\right)\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \sum_{k, i'_D} \gamma_{\mathcal{O} i'_D}^{\{k\}[D-1]} \frac{\partial}{\partial W_{i_{D-2}, i_{D-1}}^{[D-2]}} \tilde{x}_{i'_D}^{\{k\}[D-1]}\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \frac{1}{\sqrt{H^{[D-1]}}} \sum_{k, i'_D, i'_{D-1}} \gamma_{\mathcal{O} i'_D}^{\{k\}[D-1]} W_{\mathcal{O} i'_{D-1}, i'_D}^{[D-1]} \frac{\partial}{\partial W_{i_{D-2}, i_{D-1}}^{[D-2]}} x_{i'_{D-1}}^{\{k\}[D-1]}\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \frac{1}{\sqrt{H^{[D-1]}}} \sum_{k, i'_D, i'_{D-1}} \gamma_{\mathcal{O} i'_D}^{\{k\}[D-1]} W_{\mathcal{O} i'_{D-1}, i'_D}^{[D-1]} \frac{\partial}{\partial W_{i_{D-2}, i_{D-1}}^{[D-2]}} \tau^{[D-2]}\left(\tilde{x}_{i'_{D-1}}^{\{k\}[D-2]}\right)\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \frac{1}{\sqrt{H^{[D-1]}}} \sum_{k, i'_D, i'_{D-1}} \gamma_{\mathcal{O} i'_D}^{\{k\}[D-1]} W_{\mathcal{O} i'_{D-1}, i'_D}^{[D-1]} \tau^{[D-2](1)}\left(\tilde{x}_{i'_{D-1}}^{\{k\}[D-2]}\right) \frac{\partial}{\partial W_{i_{D-2}, i_{D-1}}^{[D-2]}} \tilde{x}_{i'_{D-1}}^{\{k\}[D-2]}\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \frac{1}{\sqrt{H^{[D-1]}}} \sum_{k, i'_{D-1}} \gamma_{\mathcal{O} i'_{D-1}}^{\{k\}[D-2]} \frac{\partial}{\partial W_{i_{D-2}, i_{D-1}}^{[D-2]}} \tilde{x}_{i'_{D-1}}^{\{k\}[D-2]}\Big| \mathbf{W} = \mathbf{W}_{\mathcal{O}} \\
&= -\eta \frac{1}{\sqrt{H^{[D-1]} H^{[D-2]}}} \sum_k \gamma_{\mathcal{O} i_{D-1}}^{\{k\}[D-2]} x_{\mathcal{O} i_{D-2}}^{\{k\}[D-2]}
\end{aligned}$$

where:

$$\gamma_{\mathcal{O} i_{D-1}}^{\{k\}[D-2]} = \sum_{i_D} \gamma_{\mathcal{O} i_D}^{\{k\}[D-1]} W_{\mathcal{O} i_{D-1}, i_D}^{\{k\}[D-1]} \tau^{[D-2](1)}\left(\tilde{x}_{\mathcal{O} i_{D-1}}^{\{k\}[D-2]}\right)$$

and so on through all layers. Summarising, for all $j \in \mathbb{N}_D$ (and using the MATLAB notation $\mathbf{A}_{:i}$ for column i of matrix $\mathbf{A}$):

$$\begin{aligned}
\mathbf{W}_{\Delta : i_{j+1}}^{[j]\boxast} &= -\eta \frac{1}{\sqrt{H^{[D-1]} H^{[D-2]} \ldots H^{[j+1]}}} \sum_k \gamma_{\mathcal{O} i_{j+1}}^{\{k\}[j]} \frac{1}{\sqrt{H^{[j]}}} \mathbf{x}_{\mathcal{O}}^{\{k\}[j]} \\
b_{\Delta : i_{j+1}}^{[j]\boxast} &= -\eta \frac{1}{\sqrt{H^{[D-1]} H^{[D-2]} \ldots H^{[j+1]}}} \sum_k \gamma_{\mathcal{O} i_{j+1}}^{\{k\}[j]} \alpha^{[j]}
\end{aligned} \tag{24}$$

where, recursively:

$$\begin{aligned}\gamma_{\mathcal{O}i_j}^{\{k\}[j-1]} &= \sum_{i_{j+1}} \gamma_{\mathcal{O}i_{j+1}}^{\{k\}[j]} W_{\mathcal{O}i_j,i'_{j+1}}^{[j]} \tau^{[j-1](1)} \left(\tilde{x}_{\mathcal{O}i_j}^{\{k\}[j-1]}\right) \\ \gamma_{\mathcal{O}i_D}^{\{k\}[D-1]} &= \nabla_{i_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}\left(\mathbf{x}^{\{k\}}\right)\right) \tau^{[D-1](1)} \left(\tilde{x}_{\mathcal{O}i_D}^{\{k\}[D-1]}\right)\end{aligned} \tag{25}$$

B. Dual Form of a Neural Network Step

Our first goal is to rewrite the neural network after a training iteration as:

$$\mathbf{f}\left(\mathbf{x}\right) = \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right) + \mathbf{f}_{\Delta}\left(\mathbf{x}\right)$$

where $\mathbf{f}_{\mathcal{O}} : (\mathbb{X} \subset \mathbb{R}^n) \to (\mathbb{Y} \subset \mathbb{R}^m)$ is the neural network before the iteration and $\mathbf{f}_{\Delta} : (\mathbb{X} \subset \mathbb{R}^n) \to \mathbb{R}^m$ is the change in network behaviour due to the change $\mathbf{W}_{\Delta} \in \mathbb{W}_{\mathcal{O}}$ in weights and biases for this iteration, so that:

$$\mathbf{f}_{\Delta}\left(\mathbf{x}\right) = \left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$$

where:

$$\begin{aligned}\mathbf{\Phi}_{\mathcal{O}} &: \mathbb{X} \to \mathcal{X}_{\mathcal{O}} = \mathrm{span}\left(\mathbf{\Phi}_{\mathcal{O}}\left(\mathbb{X}\right)\right) \subset \mathbb{R}^{\infty \times m} \\ \mathbf{\Psi}_{\mathcal{O}} &: \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} = \mathrm{span}\left(\mathbf{\Psi}_{\mathcal{O}}\left(\mathbb{W}_{\mathcal{O}}\right)\right) \subset \mathbb{R}^{\infty \times m}\end{aligned}$$

are feature maps;

$$\mathbb{W}_{\mathcal{O}} \subset \prod_{j \in \mathbb{N}_D} \left(\mathbb{R}^{H^{[j-1]} \times H^{[j]}} \times \mathbb{R}^{H^{[j]}}\right)$$

and $\langle \cdot, \cdot \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} : \mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}} \to \mathbb{R}^m$ is the bilinear form:

$$\left\langle \mathbf{\Xi}, \mathbf{\Omega} \right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \mathrm{diag}\left(\mathbf{\Xi}^{\mathrm{T}} \mathbf{\Omega}\right)$$

Moreover we aim to show that the feature maps are entirely defined by:

1. The structure of the network - that is, the number of layers, their widths and activation functions.
2. The weights and biases $\mathbf{W}_{\mathcal{O}} \in \mathbb{W}$ before the iteration.

We will then use this to construct appropriate norms for the weight feature space $\mathcal{W}_{\mathcal{O}}$ and data feature space $\mathcal{X}_{\mathcal{O}}$, which will allow us to prove that $\mathbf{f}_{\Delta}$ lies in a reproducing kernel Banach space, and analyse the complexity and convergence of the neural network.

B.1. Preliminary: Power Series Notation

The approach taken here is direct - we construct a power series expansion of the network without truncation and rearrange the terms to separate the resulting summation into the desired form. To make the process easier we define:

$$\boldsymbol{\varrho}\left(\mathbf{a}, \mathbf{d}\right) = \begin{bmatrix} a_0 \mathbf{d}^{\otimes 1} \\ a_1 \mathbf{d}^{\otimes 2} \\ \vdots \end{bmatrix}$$

noting that this is linear in the elementwise (Hadamard) product, so:

$$\boldsymbol{\varrho}\left(\mathbf{a}, \mathbf{d}\right) \odot \boldsymbol{\varrho}\left(\mathbf{a}', \mathbf{d}'\right) = \boldsymbol{\varrho}\left(\mathbf{a} \odot \mathbf{a}', \mathbf{d} \odot \mathbf{d}'\right)$$

Using this notation it is straightforward to obtain the following useful shorthand for the multi-dimensional Taylor expansion (using the smoothness assumption) of $\tau^{[j]} : \mathbb{R} \to \mathbb{R}$ about $z \in \mathbb{R}$ for $\mathbf{c}, \mathbf{c}' \in \mathbb{R}^p$, $|\langle \mathbf{c}, \mathbf{c}' \rangle| \leq \rho^{[j]}$:

$$\tau^{[j]}\left(z + \langle \mathbf{c}, \mathbf{c}' \rangle\right) = \tau^{[j]}\left(z\right) + \left\langle \boldsymbol{\varrho}\left(\mathbf{g}^{[j]}\left(z\right), \mathbf{c}\right), \boldsymbol{\varrho}\left(\mathbf{1}_{\infty}, \mathbf{c}'\right), \right\rangle \tag{26}$$

where the derivatives of $\tau^{[j]}$ at z are encoded as:

$$\mathbf{g}^{[j]}\left(z\right) = \begin{bmatrix} \frac{1}{1!} \tau^{[j](1)}\left(z\right) \\ \frac{1}{2!} \tau^{[j](2)}\left(z\right) \\ \frac{1}{3!} \tau^{[j](3)}\left(z\right) \\ \vdots \end{bmatrix}$$

and $\tau^{[j](r)}(z) = \frac{\partial^r}{\partial z^r} \tau^{[j]}(z)$.

B.2. Feature Maps for Neural Networks

We represent the operation before and after the iteration, as well as the change due to the iteration, as:

$$\begin{array}{c}\left(\begin{array}{l}\qquad\quad\textbf{Before Iteration}\\ \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right)=\mathbf{x}_{\mathcal{O}}^{[D]}\in\mathbb{Y}\\ \mathbf{x}_{\mathcal{O}}^{[j+1]}=\tau^{[j]}\left(\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}\right)\\ \tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}=\frac{1}{\sqrt{H^{[j]}}}\mathbf{W}_{\mathcal{O}}^{[j]\mathrm{T}}\mathbf{x}_{\mathcal{O}}^{[j]}+\alpha^{[j]}\mathbf{b}_{\mathcal{O}}^{[j]}\\ \mathbf{x}_{\mathcal{O}}^{[0]}=\mathbf{x}\in\mathbb{X}\end{array}\right)+\left(\begin{array}{l}\qquad\textbf{Weight-step Change}\\ \mathbf{f}_{\Delta}\left(\mathbf{x}\right)=\mathbf{f}\left(\mathbf{x}\right)-\mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right)\\ \mathbf{x}_{\Delta}^{[j+1]}=\mathbf{x}^{[j+1]}-\mathbf{x}_{\mathcal{O}}^{[j+1]}\\ \tilde{\mathbf{x}}_{\Delta}^{[j]}=\tilde{\mathbf{x}}^{[j]}-\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}\\ \mathbf{x}_{\Delta}^{[0]}=\mathbf{x}\in\mathbf{0}_{n}\end{array}\right)=\left(\begin{array}{l}\qquad\qquad\textbf{After Iteration}\\ \mathbf{f}\left(\mathbf{x}\right)=\mathbf{x}^{[D]}\in\mathbb{Y}\\ \mathbf{x}^{[j+1]}=\tau^{[j]}\left(\tilde{\mathbf{x}}^{[j]}\right)\\ \tilde{\mathbf{x}}^{[j]}=\frac{1}{\sqrt{H^{[j]}}}\mathbf{W}^{[j]\mathrm{T}}\mathbf{x}^{[j]}+\alpha^{[j]}\mathbf{b}^{[j]}\\ \mathbf{x}^{[0]}=\mathbf{x}\in\mathbb{X}\end{array}\right)\\ \Uparrow\\ \left(\begin{array}{l}\mathbf{x}_{\Delta}^{[j+1]}=\tau^{[j]}\left(\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}+\tilde{\mathbf{x}}_{\Delta}^{[j]}\right)-\tau^{[j]}\left(\tilde{\mathbf{x}}_{\mathcal{O}}^{[j]}\right)\\ \tilde{\mathbf{x}}_{\Delta}^{[j]}=\frac{1}{\sqrt{H^{[j]}}}\mathbf{W}_{\Delta}^{[j]\mathrm{T}}\left(\mathbf{x}_{\mathcal{O}}^{[j]}+\mathbf{x}_{\Delta}^{[j]}\right)+\frac{1}{\sqrt{H^{[j]}}}\mathbf{W}_{\mathcal{O}}^{[j]\mathrm{T}}\mathbf{x}_{\Delta}^{[j]}+\alpha^{[j]}\mathbf{b}_{\Delta}^{[j]}\end{array}\right)\end{array}\tag{27}$$

Our goal is to write $\mathbf{f}_{\Delta}:\mathbb{X}\to\mathbb{R}^{m}$ as:

$$\mathbf{f}_{\Delta}\left(\mathbf{x}\right)=\left\langle\mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right),\mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}$$

where $\mathbf{\Phi}_{\mathcal{O}}:\mathbb{X}\to\mathcal{X}_{\mathcal{O}}$ and $\mathbf{\Psi}_{\mathcal{O}}:\mathbb{W}_{\mathcal{O}}\to\mathcal{W}_{\mathcal{O}}$ are feature maps and $\langle\mathbf{\Xi},\mathbf{\Omega}\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}=\mathrm{diag}(\mathbf{\Xi}^{\mathrm{T}}\mathbf{\Omega})$ is a bilinear form. Starting with the first layer of the network, it is not difficult to see that:

$$\begin{array}{rl}\tilde{x}_{i_1}^{[0]}&=\tilde{x}_{\mathcal{O}i_1}^{[0]}+\tilde{x}_{\Delta i_1}^{[0]}\\ \tilde{x}_{\mathcal{O}i_1}^{[0]}&=\frac{1}{\sqrt{H^{[0]}}}\mathbf{W}_{\mathcal{O}:i_1}^{[0]\mathrm{T}}\mathbf{x}_{\mathcal{O}}+\alpha^{[0]}b_{\mathcal{O}i_1}^{[0]}\\ \tilde{x}_{\Delta i_1}^{[0]}&=\frac{1}{\sqrt{H^{[0]}}}\mathbf{W}_{\Delta:i_1}^{[0]\mathrm{T}}\mathbf{x}_{\mathcal{O}}+\alpha^{[0]}b_{\Delta i_1}^{[0]}\\ &=\left\langle\mu_{i_1}^{[0]}\left[\begin{array}{c}\frac{1}{\sqrt{2}}\alpha^{[0]}\\ \frac{1}{\sqrt{H^{[0]}}}\frac{1}{\sqrt{2}}\mathbf{x}_{\mathcal{O}}\end{array}\right],\frac{1}{\mu_{i_1}^{[0]}}\left[\begin{array}{c}\sqrt{2}b_{\Delta i_1}^{[0]}\\ \sqrt{2}\mathbf{W}_{\Delta:i_1}^{[0]}\end{array}\right]\right\rangle\end{array}$$

where $\mu_{i_1}^{[0]}\in\mathbb{R}_{+}$ are parameters we call scale factors, and $\mathbf{A}_{:l}$ for column l of matrix $\mathbf{A}$. The scale factors will always cancel out in our representation of $\mathbf{f}$, but will be important later when we place norms on feature space. Using a power series expansion as per (26), if $\|\tilde{\mathbf{x}}_{\Delta}^{[0]}\|_{\infty}\leq\rho^{[0]}$:

$$\begin{array}{rl}\mathbf{x}^{[1]}&=\left[\tau^{[0]}\left(\tilde{x}_{i_1}^{[0]}\right)\right]_{i_1}=\left[\tau^{[0]}\left(\tilde{x}_{\mathcal{O}i_1}^{[0]}+\tilde{x}_{\Delta i_1}^{[0]}\right)\right]_{i_1}\\ &=\mathbf{x}_{\mathcal{O}}^{[1]}+\left[\left\langle\boldsymbol{\varrho}\left(\mathbf{g}^{[0]}\left(\tilde{x}_{\mathcal{O}i_1}^{[0]}\right),\mu_{i_1}^{[0]}\left[\begin{array}{c}\frac{1}{\sqrt{2}}\alpha^{[0]}\\ \frac{1}{\sqrt{H^{[0]}}}\frac{1}{\sqrt{2}}\mathbf{x}_{\mathcal{O}}\end{array}\right]\right),\boldsymbol{\varrho}\left(\mathbf{1}_{\infty},\frac{1}{\mu_{i_1}^{[0]}}\left[\begin{array}{c}\sqrt{2}b_{\Delta i_1}^{[0]}\\ \sqrt{2}\mathbf{W}_{\Delta:i_1}^{[0]}\end{array}\right]\right)\right\rangle\right]_{i_1}\end{array}$$

and hence:

$$\mathbf{x}_{\Delta}^{[1]}=\left[\left\langle\mathbf{\Phi}_{\mathcal{O}:i_1}^{[0]}\left(\mathbf{x}\right),\mathbf{\Psi}_{\mathcal{O}:i_1}^{[0]}\left(\mathbf{W}_{\Delta}\right)\right\rangle\right]_{i_1}$$

where:

$$\begin{array}{rl}\mathbf{\Phi}_{\mathcal{O}:i_1}^{[0]}\left(\mathbf{x}\right)&=\boldsymbol{\varrho}\left(\mathbf{g}^{[0]}\left(\tilde{x}_{\mathcal{O}i_1}^{[0]}\right),\mu_{i_1}^{[0]}\left[\begin{array}{c}\frac{1}{\sqrt{2}}\alpha^{[0]}\\ \frac{1}{\sqrt{H^{[0]}}}\frac{1}{\sqrt{2}}\mathbf{x}_{\mathcal{O}}\end{array}\right]\right)\\ \mathbf{\Psi}_{\mathcal{O}:i_1}^{[0]}\left(\mathbf{W}_{\Delta}\right)&=\boldsymbol{\varrho}\left(\mathbf{1}_{\infty},\frac{1}{\mu_{i_1}^{[0]}}\left[\begin{array}{c}\sqrt{2}b_{\Delta i_1}^{[0]}\\ \sqrt{2}\mathbf{W}_{\Delta:i_1}^{[0]}\end{array}\right]\right)\end{array}$$

Moving to the second layer, we see that:

$$\begin{aligned}
\tilde{x}^{[1]}_{i_2} &= \tilde{x}^{[1]}_{\mathcal{O}i_2} + \tilde{x}^{[1]}_{\Delta i_2} \\
\tilde{x}^{[1]}_{\mathcal{O}i_2} &= \tfrac{1}{\sqrt{H^{[1]}}} \mathbf{W}^{[1]\mathrm{T}}_{\mathcal{O}:i_2} \mathbf{x}^{[1]}_{\mathcal{O}} + \alpha^{[1]} b^{[1]}_{\mathcal{O}i_2} \\
\tilde{x}^{[1]}_{\Delta i_2} &= \tfrac{1}{\sqrt{H^{[1]}}} \mathbf{W}^{[1]\mathrm{T}}_{\Delta:i_2} \mathbf{x}^{[1]}_{\mathcal{O}} + \tfrac{1}{\sqrt{H^{[1]}}} \left(\mathbf{W}^{[1]\mathrm{T}}_{\Delta:i_2} + \mathbf{W}^{[1]\mathrm{T}}_{\mathcal{O}:i_2} \right) \mathbf{x}^{[1]}_{\Delta} + \alpha^{[1]} b^{[1]}_{\Delta i_2} \\
&= \tfrac{1}{\sqrt{H^{[1]}}} \mathbf{W}^{[1]\mathrm{T}}_{\Delta:i_2} \mathbf{x}^{[1]}_{\mathcal{O}} + \tfrac{1}{\sqrt{H^{[1]}}} \textstyle\sum_{i_1} \left(W^{[1]}_{\mathcal{O}i_1,i_2} + W^{[1]}_{\Delta i_1,i_2} \right) \left\langle \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{x}), \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{W}_\Delta) \right\rangle + \alpha^{[1]} b^{[1]}_{\Delta i_2} \\
&= \left\langle \mu^{[1]}_{i_2} \begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \alpha^{[1]} \\ \frac{1}{\sqrt{H^{[1]}}} \mathbf{x}^{[1]}_{\mathcal{O}} \end{bmatrix} \\ \begin{bmatrix} \frac{1}{\widetilde{\omega}^{[1]}_{i_1,i_2}} \frac{\omega^{[1]}_{i_1} W^{[1]}_{\mathcal{O}i_1,i_2}}{\sqrt{H^{[1]}}} \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{x}) \end{bmatrix}_{i_1} \\ \begin{bmatrix} \frac{\omega^{[1]}_{i_1}}{\sqrt{H^{[1]}}} \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{x}) \end{bmatrix}_{i_1} \end{bmatrix}, \frac{1}{\mu^{[1]}_{i_2}} \begin{bmatrix} \begin{bmatrix} \sqrt{2} b^{[1]}_{\Delta i_2} \\ \mathbf{W}^{[1]}_{\Delta:i_2} \end{bmatrix} \\ \begin{bmatrix} \widetilde{\omega}^{[1]}_{i_1,i_2} \frac{1}{\omega^{[1]}_{i_1}} \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{W}_\Delta) \end{bmatrix}_{i_1} \\ \begin{bmatrix} \frac{W^{[1]}_{\Delta i_1,i_2}}{\omega^{[1]}_{i_1}} \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{W}_\Delta) \end{bmatrix}_{i_1} \end{bmatrix} \right\rangle
\end{aligned}$$

where $\mu^{[1]}_{i_2} \in \mathbb{R}_+$ are scale factors and $\omega^{[1]}_{i_1}, \widetilde{\omega}^{[1]}_{i_1,i_2} \in \mathbb{R}^+$ are shadow weights (note that both the shadow weights and scale factors cancel in the inner product). Taking the Taylor expansion and using the properties of $\boldsymbol{\varrho}$ we see that, provided $\|\tilde{\mathbf{x}}^{[1]}_{\Delta}\|_\infty \leq \rho^{[1]}$:

$$\mathbf{x}^{[2]}_{\Delta} = \left[\left\langle \mathbf{\Phi}^{[1]}_{\mathcal{O}:i_2}(\mathbf{x}), \mathbf{\Psi}^{[1]}_{\mathcal{O}:i_2}(\mathbf{W}_\Delta) \right\rangle \right]_{i_2}$$

where:

$$\mathbf{\Phi}^{[1]}_{\mathcal{O}:i_2}(\mathbf{x}) = \boldsymbol{\varrho} \left(\mathbf{g}^{[1]} \left(\tilde{x}^{[1]}_{\mathcal{O}i_2} \right), \mu^{[1]}_{i_2} \begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \alpha^{[1]} \\ \frac{1}{\sqrt{H^{[1]}}} \mathbf{x}^{[1]}_{\mathcal{O}} \end{bmatrix} \\ \begin{bmatrix} \frac{1}{\widetilde{\omega}^{[1]}_{i_1,i_2}} \frac{\omega^{[1]}_{i_1} W^{[1]}_{\mathcal{O}i_1,i_2}}{\sqrt{H^{[1]}}} \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{x}) \end{bmatrix}_{i_1} \\ \begin{bmatrix} \frac{\omega^{[1]}_{i_1}}{\sqrt{H^{[1]}}} \mathbf{\Phi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{x}) \end{bmatrix}_{i_1} \end{bmatrix} \right)$$

$$\mathbf{\Psi}^{[1]}_{\mathcal{O}:i_2}(\mathbf{W}_\Delta) = \boldsymbol{\varrho} \left(\mathbf{1}_\infty, \frac{1}{\mu^{[1]}_{i_2}} \begin{bmatrix} \begin{bmatrix} \sqrt{2} b^{[1]}_{\Delta i_2} \\ \mathbf{W}^{[1]}_{\Delta:i_2} \end{bmatrix} \\ \begin{bmatrix} \widetilde{\omega}^{[1]}_{i_1,i_2} \frac{1}{\omega^{[1]}_{i_1}} \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{W}_\Delta) \end{bmatrix}_{i_1} \\ \begin{bmatrix} \frac{W^{[1]}_{\Delta i_1,i_2}}{\omega^{[1]}_{i_1}} \mathbf{\Psi}^{[0]}_{\mathcal{O}:i_1}(\mathbf{W}_\Delta) \end{bmatrix}_{i_1} \end{bmatrix} \right)$$

Continuing, it is not difficult to see that, for any $j \in \mathbb{N}_D$, if:

$$\mathbf{x}^{[j]}_{\Delta} = \left[\left\langle \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x}), \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta) \right\rangle \right]_{i_j}$$

then, with scale factors $\mu^{[j]}_{i_{j+1}} \in \mathbb{R}_+$ and shadow weights $\omega^{[j]}_{i_j}, \widetilde{\omega}^{[j]}_{i_j,i_{j+1}} \in \mathbb{R}_+$:

$$\begin{aligned}
\tilde{x}^{[j]}_{i_{j+1}} &= \tilde{x}^{[j]}_{\mathcal{O}i_{j+1}} + \tilde{x}^{[j]}_{\Delta i_{j+1}} \\
\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}} &= \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}^{[j]\mathrm{T}}_{\mathcal{O}:i_{j+1}} \mathbf{x}^{[j]}_{\mathcal{O}} + \alpha^{[j]} b^{[j]}_{\mathcal{O}i_{j+1}} \\
\tilde{x}^{[j]}_{\Delta i_{j+1}} &= \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}^{[j]\mathrm{T}}_{\Delta:i_{j+1}} \mathbf{x}^{[j]}_{\mathcal{O}} + \tfrac{1}{\sqrt{H^{[j]}}} \left(\mathbf{W}^{[j]\mathrm{T}}_{\Delta:i_{j+1}} + \mathbf{W}^{[j]\mathrm{T}}_{\mathcal{O}:i_{j+1}} \right) \mathbf{x}^{[j]}_{\Delta} + \alpha^{[j]} b^{[j]}_{\Delta i_{j+1}} \\
&= \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}^{[j]\mathrm{T}}_{\Delta:i_{j+1}} \mathbf{x}^{[j]}_{\mathcal{O}} + \tfrac{1}{\sqrt{H^{[j]}}} \textstyle\sum_{i_j} \left(W^{[j]}_{\mathcal{O}i_j i_{j+1}} + W^{[j]}_{\Delta i_j i_{j+1}} \right) \left\langle \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x}), \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta) \right\rangle + \alpha^{[j]} b^{[j]}_{\Delta i_{j+1}} \\
&= \left\langle \mu^{[j]}_{i_{j+1}} \begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \alpha^{[j]} \\ \frac{1}{\sqrt{H^{[j]}}} \mathbf{x}^{[j]}_{\mathcal{O}} \end{bmatrix} \\ \begin{bmatrix} \frac{1}{\widetilde{\omega}^{[j]}_{i_j,i_{j+1}}} \frac{\omega^{[j]}_{i_j} W^{[j]}_{\mathcal{O}i_j,i_{j+1}}}{\sqrt{H^{[j]}}} \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x}) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{\omega^{[j]}_{i_j}}{\sqrt{H^{[j]}}} \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x}) \end{bmatrix}_{i_j} \end{bmatrix}, \frac{1}{\mu^{[j]}_{i_{j+1}}} \begin{bmatrix} \begin{bmatrix} \sqrt{2} b^{[j]}_{\Delta i_{j+1}} \\ \mathbf{W}^{[j]}_{\Delta:i_{j+1}} \end{bmatrix} \\ \begin{bmatrix} \widetilde{\omega}^{[j]}_{i_j,i_{j+1}} \frac{1}{\omega^{[j]}_{i_j}} \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{W^{[j]}_{\Delta i_j,i_{j+1}}}{\omega^{[j]}_{i_j}} \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta) \end{bmatrix}_{i_j} \end{bmatrix} \right\rangle
\end{aligned}$$

so, provided $\|\tilde{\mathbf{x}}_\Delta^{[j]}\|_\infty \leq \rho^{[j]}$, we have:

$$\mathbf{x}_\Delta^{[j+1]} = \left[\left\langle \mathbf{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{x}\right), \mathbf{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta\right)\right\rangle\right]_{i_{j+1}}$$

where:

$$\mathbf{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{x}\right) = \boldsymbol{\varrho}\left(\mathbf{g}^{[j]}\left(\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}\right), \mu_{i_{j+1}}^{[j]}\begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[j]} \\ \frac{1}{\sqrt{H^{[j]}}}\mathbf{x}_{\mathcal{O}}^{[j]} \end{bmatrix} \\ \begin{bmatrix} \frac{1}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]}}\frac{\omega_{i_j}^{[j]}W_{\mathcal{O}i_j,i_{j+1}}^{[j]}}{\sqrt{H^{[j]}}}\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{\omega_{i_j}^{[j]}}{\sqrt{H^{[j]}}}\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right) \end{bmatrix}_{i_j} \end{bmatrix}\right)$$

$$\mathbf{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta\right) = \boldsymbol{\varrho}\left(\mathbf{1}_\infty, \frac{1}{\mu_{i_{j+1}}^{[j]}}\begin{bmatrix} \begin{bmatrix} \sqrt{2}b_{\Delta i_{j+1}}^{[j]} \\ \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \end{bmatrix} \\ \begin{bmatrix} \widetilde{\omega}_{i_j,i_{j+1}}^{[j]}\frac{1}{\omega_{i_j}^{[j]}}\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{W_{\Delta i_j,i_{j+1}}^{[j]}}{\omega_{i_j}^{[j]}}\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right) \end{bmatrix}_{i_j} \end{bmatrix}\right)$$

Thus, by induction, so long as $\|\tilde{\mathbf{x}}_\Delta^{[j]}\|_\infty \leq \rho^{[j]}\ \forall j$:

$$\begin{aligned} \mathbf{f}\left(\mathbf{x}\right) &= \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right) + \mathbf{f}_\Delta\left(\mathbf{x}\right) \\ \mathbf{f}_\Delta\left(\mathbf{x}\right) &= \left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right), \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_\Delta\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}} \end{aligned} \tag{28}$$

where:

$$\begin{aligned} \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right) &= \mathbf{\Phi}_{\mathcal{O}}^{[D-1]}\left(\mathbf{x}\right) \\ \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_\Delta\right) &= \mathbf{\Psi}_{\mathcal{O}}^{[D-1]}\left(\mathbf{W}_\Delta\right) \end{aligned}$$

and, working recursively:

$$\begin{aligned} \mathbf{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{x}\right) &= \begin{cases} \boldsymbol{\varrho}\left(\mathbf{g}^{[j]}\left(\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}\right), \mu_{i_{j+1}}^{[j]}\begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[j]} \\ \frac{1}{\sqrt{H^{[j]}}}\mathbf{x}_{\mathcal{O}}^{[j]} \end{bmatrix} \\ \begin{bmatrix} \frac{1}{\widetilde{\omega}_{i_j i_{j+1}}^{[j]}}\frac{\omega_{i_j}^{[j]}W_{\mathcal{O}i_j,i_{j+1}}^{[j]}}{\sqrt{H^{[j]}}}\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{\omega_{i_j}^{[j]}}{\sqrt{H^{[j]}}}\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right) \end{bmatrix}_{i_j} \end{bmatrix}\right) & \text{if } j > 0 \\ \boldsymbol{\varrho}\left(\mathbf{g}^{[0]}\left(\tilde{x}_{\mathcal{O}i_1}^{[0]}\right), \mu_{i_1}^{[0]}\begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[0]} \\ \frac{1}{\sqrt{H^{[0]}}}\frac{1}{\sqrt{2}}\mathbf{x}_{\mathcal{O}} \end{bmatrix}\right) & \text{if } j = 0 \end{cases} \\ \mathbf{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta\right) &= \begin{cases} \boldsymbol{\varrho}\left(\mathbf{1}_\infty, \frac{1}{\mu_{i_{j+1}}^{[j]}}\begin{bmatrix} \begin{bmatrix} \sqrt{2}b_{\Delta i_{j+1}}^{[j]} \\ \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \end{bmatrix} \\ \begin{bmatrix} \widetilde{\omega}_{i_j,i_{j+1}}^{[j]}\frac{1}{\omega_{i_j}^{[j]}}\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{W_{\Delta i_j,i_{j+1}}^{[j]}}{\omega_{i_j}^{[j]}}\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right) \end{bmatrix}_{i_j} \end{bmatrix}\right) & \text{if } j > 0 \\ \boldsymbol{\varrho}\left(\mathbf{1}_\infty, \frac{1}{\mu_{i_1}^{[0]}}\begin{bmatrix} \sqrt{2}b_{\Delta i_1}^{[0]} \\ \sqrt{2}\mathbf{W}_{\Delta:i_1}^{[0]} \end{bmatrix}\right) & \text{if } j = 0 \end{cases} \end{aligned} \tag{29}$$

$\forall j \in \mathbb{N}_D$, with scale factors $\mu_{i_{j+1}}^{[j]} \in \mathbb{R}_+$ and shadow weights $\omega_{i_j}^{[j]}, \widetilde{\omega}_{i_j,i_{j+1}}^{[j]} \in \mathbb{R}_+$, $\forall j \in \mathbb{N}_D$.

To summarise, we have decomposed the neural network into the desired form (28) with the following components:

- A term $\mathbf{f}_{\mathcal{O}}\left(\mathbf{x}\right)$ which is the network prior to the iteration evaluated on $\mathbf{x}$.

- A feature map $\boldsymbol{\Phi}_{\mathcal{O}} : \mathbb{X} \to \mathcal{X}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X})) \subset \mathbb{R}^{\infty \times m}$ from data input space to data feature space that is dependent only on the structure of the network (number of layers, their widths and the neuron types) and the weights and biases $\mathbf{W}_{\mathcal{O}}$ prior to the iteration.
- A feature map $\boldsymbol{\Psi}_{\mathcal{O}} : \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})) \subset \mathbb{R}^{\infty \times m}$ from weight-step input space to weight-step feature space that is dependent only on the structure of the network (number of layers and their widths).

B.3. Density of Feature Maps

In this section we prove a key property of the feature maps that will be required to prove that $\mathbf{f}_{\Delta}$ lies in the reproducing kernel Banach space. As a preliminary we show a consequence of the non-triviality assumption, namely:

Lemma 3. *Let $\mathbf{f}_{\mathcal{O}} \in \mathcal{F}$ be a neural network satisfying the non-triviality assumption. Then $\mathbf{f}_{\Delta}(\mathbf{x})$ varies non-trivially with $\mathbf{x}$ (that is, it is non-constant).*

Proof. Recall that:

$$\begin{aligned} \tilde{\mathbf{x}}_{\Delta}^{[j]} &= \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}_{\Delta}^{[j]\mathrm{T}} \mathbf{x}_{\mathcal{O}}^{[j]} + \tfrac{1}{\sqrt{H^{[j]}}} \left(\mathbf{W}_{\mathcal{O}}^{[j]} + \mathbf{W}_{\Delta}^{[j]} \right)^{\mathrm{T}} \mathbf{x}_{\Delta}^{[j]} + \alpha^{[j]} \mathbf{b}_{\Delta}^{[j]} \\ &= \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}_{\Delta}^{[j]\mathrm{T}} \left(\mathbf{x}_{\mathcal{O}}^{[j]} + \mathbf{x}_{\Delta}^{[j]} \right) + \tfrac{1}{\sqrt{H^{[j]}}} \mathbf{W}_{\mathcal{O}}^{[j]\mathrm{T}} \mathbf{x}_{\Delta}^{[j]} + \alpha^{[j]} \mathbf{b}_{\Delta}^{[j]} \end{aligned}$$

we want to know the conditions under which $\tilde{\mathbf{x}}_{\Delta}^{[D-1]}$ (and hence $\mathbf{f}_{\Delta}(\mathbf{x})$) is a constant, independent of $\mathbf{x}$. Considering instead $\tilde{\mathbf{x}}_{\Delta}^{[j]}$, the only component in the above expression that depends on $\mathbf{x}$ is $\mathbf{x}_{\Delta}^{[j]}$, so $\tilde{\mathbf{x}}_{\Delta}^{[j]}$ is constant if either $\mathbf{W}_{\Delta}^{[j]} = -\mathbf{W}_{\mathcal{O}}^{[j]}$ or $\mathbf{x}_{\Delta}^{[j]}$ is constant independent of $\mathbf{x}$. So, recursing, we find that $\tilde{\mathbf{x}}_{\Delta}^{[D-1]}$ is constant (independent of $\mathbf{x}$) iff $\mathbf{W}_{\Delta}^{[j]} = -\mathbf{W}_{\mathcal{O}}^{[j]}$ for some $j \in \mathbb{N}_D$. However, recall that our non-triviality assumption explicitly rules out this case (which corresponds to trivial neural network post iteration), so $\mathbf{W}_{\Delta}^{[j]} \neq -\mathbf{W}_{\mathcal{O}}^{[j]}$ for all $j \in \mathbb{N}_D$, which suffices to demonstrate the desired result. □

Having addressed this preliminary, we now move to the main result for this section:[13]

Lemma 4 (Density of Feature Maps)**.** *The linear span of $\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X})$ is dense in $\mathcal{X}_{\mathcal{O}}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$; and the linear span of $\boldsymbol{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})$ is dense in $\mathcal{W}_{\mathcal{O}}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}}$, where $\langle \boldsymbol{\Omega}, \boldsymbol{\Xi} \rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}} = \langle \boldsymbol{\Xi}, \boldsymbol{\Omega} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}^{\mathrm{T}}$.*

Proof. By definition (Lin et al., 2022), the linear span $\mathrm{span}(\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X}))$ of $\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X})$ is dense in $\mathcal{X}_{\mathcal{O}}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$ if for any $\boldsymbol{\Omega}_{\Delta} \in \mathcal{W}_{\mathcal{O}}$, the statement:

$$\langle \boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x}), \boldsymbol{\Omega}_{\Delta} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \mathbf{0} \quad \forall \mathbf{x} \in \mathbb{X}$$

implies that $\boldsymbol{\Omega}_{\Delta} = \mathbf{0}$. By definition $\mathcal{W}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}}))$ so that $\boldsymbol{\Omega}_{\Delta} = \sum_l \tilde{\alpha}^{\{l\}} \mathbf{W}_{\Delta}^{\{l\}}$ for some $\tilde{\alpha}^{\{0\}}, \tilde{\alpha}^{\{1\}}, \ldots \in \mathbb{R}$, $\mathbf{W}_{\Delta}^{\{0\}}, \mathbf{W}_{\Delta}^{\{1\}}, \ldots \in \mathbb{W}_{\mathcal{O}}$, and hence:

$$\langle \boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x}), \boldsymbol{\Omega}_{\Delta} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \textstyle\sum_l \tilde{\alpha}^{\{l\}} \left\langle \boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x}), \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}^{\{l\}}\right) \right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \textstyle\sum_l \tilde{\alpha}^{\{l\}} \mathbf{f}_{\Delta}^{\{l\}}(\mathbf{x})$$

where we denote by $\mathbf{f}_{\Delta}^{\{l\}}$ the change in $\mathbf{f}$ due to weight-step $\mathbf{W}_{\Delta}^{\{l\}}$. By our preliminary this is $\mathbf{0}$ for all $\mathbf{x} \in \mathbb{X}$ iff $\tilde{\alpha}^{\{0\}} = \tilde{\alpha}^{\{l\}} = \ldots = 0$, so $\boldsymbol{\Omega}_{\Delta} = \mathbf{0}$, which proves the required density property. Similarly, $\mathrm{span}(\boldsymbol{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}}))$ is dense in $\mathcal{W}_{\mathcal{O}}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}}$ if for any $\boldsymbol{\Xi}_{\Delta} \in \mathcal{X}_{\mathcal{O}}$, the observation that:

$$\langle \boldsymbol{\Xi}_{\Delta}, \boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \mathbf{0} \quad \forall \mathbf{W}_{\Delta} \in \mathbb{W}$$

implies $\boldsymbol{\Xi}_{\Delta} = \mathbf{0}$. By definition $\mathcal{X}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X}))$ so that $\boldsymbol{\Xi}_{\Delta} = \sum_l \tilde{\beta}^{\{l\}} \mathbf{x}^{\{l\}}$ for some $\tilde{\beta}^{\{0\}}, \tilde{\beta}^{\{1\}}, \ldots \in \mathbb{R}$, $\mathbf{x}^{\{0\}}, \mathbf{x}^{\{1\}}, \ldots \in \mathbb{X}$, and hence:

$$\langle \boldsymbol{\Xi}_{\Delta}, \boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \textstyle\sum_l \tilde{\beta}^{\{l\}} \left\langle \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}^{\{l\}}\right), \boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}) \right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \textstyle\sum_l \tilde{\beta}^{\{l\}} \mathbf{f}_{\Delta}\left(\mathbf{x}^{\{l\}}\right)$$

But again, by our preliminary, this is $\mathbf{0}$ for all $\mathbf{W}_{\Delta} \in \mathbb{W}_{\mathcal{O}}$ iff $\tilde{\beta}^{\{0\}} = \tilde{\beta}^{\{1\}} = \ldots = 0$, so $\boldsymbol{\Xi}_{\Delta} = \mathbf{0}$, completing the proof. □

[13] See (Lin et al., 2022) for discussion of density used here.

Remark 1. An objection may be raised that the non-triviality assumption is arbitrary and in any case not guaranteed to hold, bringing the above result into doubt. However, considering that neural networks are usually initialised with non-trivial weights and biases (or, to be precise, the chance that any one layer has all zero weights and biases drawn from e,g. a uniform or normal distribution is 0), and that training data is usually non-trivial itself (that is, the targets vary and not just the inputs, so that the trivial case has no special significance in terms of potential optimality giving a fixed output), it seems highly unlikely that the weights and biases would return to a trivial (all weights and biases 0 in at least one layer) state over any finite number of back-propagation iterations - indeed we conjecture that a result saying that the probability of encountering trivial networks weights and biases during training is precisely 0 given random weight initialization and noise-affected data drawn from a distribution should exist, but unfortunately we have been unable to obtain a proof, so this remains speculative for now.

C. Induced Kernels and Induced Norms

In the previous section we established that, as a result of a single iteration (step-change in weights and biases), the operation of the neural network can be written:

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}_{\mathcal{O}}(\mathbf{x}) + \mathbf{f}_{\Delta}(\mathbf{x})$$

where $\mathbf{f}_{\mathcal{O}} : \mathbb{X} \to \mathbb{Y}$ is the neural network before the iteration and $\mathbf{f}_{\Delta} : \mathbb{X} \to \mathbb{Y}_{\mathcal{O}}$ is the change in network behaviour due to the change $\mathbf{W}_{\Delta} \in \mathbb{W}_{\mathcal{O}}$ in weights and biases for this iteration. Moreover we showed that:

$$\mathbf{f}_{\Delta}(\mathbf{x}) = \langle \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}), \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$$

where:

$$\begin{aligned} \mathbf{\Phi}_{\mathcal{O}} &: \mathbb{X} \to \mathcal{X}_{\mathcal{O}} = \mathrm{span}(\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})) \subset \mathbb{R}^{\infty \times m} \\ \mathbf{\Psi}_{\mathcal{O}} &: \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} = \mathrm{span}(\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})) \subset \mathbb{R}^{\infty \times m} \end{aligned}$$

are feature maps, and $\langle \cdot, \cdot \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} : \mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}} \to \mathbb{R}^m$ is the bilinear form:

$$\langle \mathbf{\Xi}, \mathbf{\Omega} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \mathrm{diag}\left(\mathbf{\Xi}^{\mathrm{T}} \mathbf{\Omega}\right)$$

In this section we will place (induce) kernels and norms on $\mathbb{X}_{\mathcal{O}}$ and $\mathbb{W}_{\mathcal{O}}$, which will in turn allow us to constrain the space of changes in neural network behaviour:

$$\mathcal{F}_{\mathcal{O}} = \{ \mathbf{f}_{\Delta} : \mathbb{X} \to \mathbb{Y}_{\mathcal{O}} | \, \mathbf{f}_{\Delta} \text{ as above} \}$$

using Hölder's inequality:

$$\| \mathbf{f}_{\Delta}(\mathbf{x}) \|_2 \leq \| \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}) \|_{\mathcal{X}_{\mathcal{O}}} \| \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}) \|_{\mathcal{W}_{\mathcal{O}}}$$

C.1. Induced Kernels

We begin by using the feature maps and the kernel trick to induce kernels on $\mathbb{X}$ and $\mathbb{W}_{\mathcal{O}}$, specifically:

$$\begin{aligned} \mathbf{K}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') &= \langle \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}), \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}') \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}} \\ \mathbf{K}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_{\Delta}, \mathbf{W}'_{\Delta}) &= \langle \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}), \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}'_{\Delta}) \rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \end{aligned}$$

where we define the bilinear forms:

$$\begin{aligned} \langle \mathbf{\Xi}, \mathbf{\Xi}' \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}} &= \mathbf{\Xi}^{\mathrm{T}} \mathbf{\Xi}' \\ \langle \mathbf{\Omega}, \mathbf{\Omega}' \rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} &= \mathbf{\Omega}^{\mathrm{T}} \mathbf{\Omega}' \end{aligned}$$

The matrix-valued kernels $\mathbf{K}_{\mathcal{X}_{\mathcal{O}}}$ and $\mathbf{K}_{\mathcal{W}_{\mathcal{O}}}$ are positive-definite (Mercer) by construction. Apart from their theoretical use, these could potentially be used (transfered) for support vector machines (SVMs), Gaussian Processes (GP) or similar kernel-based methods, measuring similarity on $\mathbb{X}$ and $\mathbb{W}_{\mathcal{O}}$, respectively.

For all layers $j \in \mathbb{N}_D$ we define:

$$\begin{aligned} \mathbf{K}^{[j]}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') &= \left\langle \mathbf{\Phi}^{[j]}_{\mathcal{O}}(\mathbf{x}), \mathbf{\Phi}^{[j]}_{\mathcal{O}}(\mathbf{x}') \right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}} \\ \mathbf{K}^{[j]}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_{\Delta}, \mathbf{W}'_{\Delta}) &= \left\langle \mathbf{\Psi}^{[j]}_{\mathcal{O}}(\mathbf{W}_{\Delta}), \mathbf{\Psi}^{[j]}_{\mathcal{O}}(\mathbf{W}'_{\Delta}) \right\rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \end{aligned}$$

Recalling (29) we see that, for all $j \in \mathbb{N}_D \backslash \{0\}$:

$$
\begin{aligned}
K^{[j]}_{\mathcal{X}\mathcal{O} i_{j+1}, i'_{j+1}}\left(\mathbf{x}, \mathbf{x}'\right) = \sum_l \varrho_l \Bigg(& \mathbf{g}^{[j]}\left(\tilde{x}^{[j]}_{\mathcal{O} i_{j+1}}\right) \odot \mathbf{g}^{[j]}\left(\tilde{x}'^{[j]}_{\mathcal{O} i'_{j+1}}\right), \ldots \\
& \ldots \mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}} \begin{bmatrix} \begin{bmatrix} \frac{1}{2}\alpha^{[j]2} \\ \frac{1}{H^{[j]}} \mathbf{x}^{[j]}_{\mathcal{O}} \odot \mathbf{x}^{[j]}_{\mathcal{O}} \end{bmatrix} \\ \begin{bmatrix} \frac{\omega^{[j]2}_{i_j} W^{[j]}_{\mathcal{O} i_j, i_{j+1}} W^{[j]}_{\mathcal{O} i_j, i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} H^{[j]}} \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{x}\right) \odot \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{x}'\right) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{x}\right) \odot \mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{x}'\right) \end{bmatrix}_{i_j} \end{bmatrix} \Bigg) \\
K^{[j]}_{\mathcal{W}\mathcal{O} i_{j+1}, i'_{j+1}}\left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta\right) = \sum_l \varrho_l \Bigg(& \mathbf{1}_\infty, \ldots \\
& \ldots \frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \begin{bmatrix} \begin{bmatrix} 2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} \\ \mathbf{W}^{[j]}_{\Delta:i_{j+1}} \odot \mathbf{W}'^{[j]}_{\Delta:i'_{j+1}} \end{bmatrix} \\ \begin{bmatrix} \frac{\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}}}{\omega^{[j]2}_{i_j}} \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{W}_\Delta\right) \odot \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{W}'_\Delta\right) \end{bmatrix}_{i_j} \\ \begin{bmatrix} \frac{W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}}}{\omega^{[j]2}_{i_j}} \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{W}_\Delta\right) \odot \mathbf{\Psi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{W}'_\Delta\right) \end{bmatrix}_{i_j} \end{bmatrix} \Bigg)
\end{aligned}
$$

and:

$$
\begin{aligned}
K^{[0]}_{\mathcal{X}\mathcal{O} i_1, i'_1}\left(\mathbf{x}, \mathbf{x}'\right) &= \sum_l \varrho_l \left(\mathbf{g}^{[0]}\left(\tilde{x}^{[0]}_{\mathcal{O} i_1}\right) \odot \mathbf{g}^{[0]}\left(\tilde{x}'^{[0]}_{\mathcal{O} i'_1}\right), \mu^{[0]}_{i_1} \mu^{[0]}_{i'_1} \begin{bmatrix} \frac{1}{2}\alpha^{[0]2} \\ \frac{1}{H^{[0]}} \frac{1}{2} \mathbf{x}^{[0]}_{\mathcal{O}} \odot \mathbf{x}^{[0]}_{\mathcal{O}} \end{bmatrix} \right) \\
K^{[0]}_{\mathcal{W}\mathcal{O} i_1, i'_1}\left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta\right) &= \sum_l \varrho_l \left(\mathbf{1}_\infty, \frac{1}{\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1}} \begin{bmatrix} 2 b^{[0]}_{\Delta i_1} b'^{[0]}_{\Delta i'_1} \\ 2 \mathbf{W}^{[0]}_{\Delta:i_1} \odot \mathbf{W}'^{[0]}_{\Delta:i'_1} \end{bmatrix} \right)
\end{aligned}
$$

Subsequently $\forall j \in \mathbb{N}_D \backslash \{0\}$:

$$
\begin{aligned}
K^{[j]}_{\mathcal{X}\mathcal{O} i_{j+1}, i'_{j+1}}(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} \left(\tfrac{1}{l!}\right)^2 \tau^{[j](l)}\left(\tilde{x}^{[j]}_{\mathcal{O} i_{j+1}}\right) \tau^{[j](l)}\left(\tilde{x}'^{[j]}_{\mathcal{O} i'_{j+1}}\right) \Bigg(\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}} \Bigg(& \left(\tfrac{1}{2}\alpha^{[j]2} + \tfrac{1}{H^{[j]}} \left\langle \mathbf{x}^{[j]}_{\mathcal{O}}, \mathbf{x}'^{[j]}_{\mathcal{O}} \right\rangle \right) + \ldots \\
& \sum_{i_j} \frac{1}{\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}}} \frac{\omega^{[j]2}_{i_j} W^{[j]}_{\mathcal{O} i_j, i_{j+1}} W^{[j]}_{\mathcal{O} i_j, i'_{j+1}}}{H^{[j]}} K^{[j-1]}_{\mathcal{X}\mathcal{O} i_j, i_j}\left(\mathbf{x}, \mathbf{x}'\right) + \sum_{i_j} \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} K^{[j-1]}_{\mathcal{X}\mathcal{O} i_j, i_j}\left(\mathbf{x}, \mathbf{x}'\right) \Bigg) \Bigg)^l \\
K^{[j]}_{\mathcal{W}\mathcal{O} i_{j+1}, i'_{j+1}}\left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta\right) = \sum_{l=1}^{\infty} \Bigg(\frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \Bigg(& \left(2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} + \left\langle \mathbf{W}^{[j]}_{\Delta:i_{j+1}}, \mathbf{W}'^{[j]}_{\Delta:i'_{j+1}} \right\rangle \right) + \ldots \\
& \sum_{i_j} \widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} \frac{1}{\omega^{[j]2}_{i_j}} K^{[j-1]}_{\mathcal{W}\mathcal{O} i_j, i_j}\left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta\right) + \sum_{i_j} \frac{W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}}}{\omega^{[j]2}_{i_j}} K^{[j-1]}_{\mathcal{W}\mathcal{O} i_j, i_j}\left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta\right) \Bigg) \Bigg)^l
\end{aligned}
$$

and:

$$
\begin{aligned}
K^{[0]}_{\mathcal{X}\mathcal{O} i_1, i'_1}\left(\mathbf{x}, \mathbf{x}'\right) &= \sum_{l=1}^{\infty} \left(\tfrac{1}{l!}\right)^2 \tau^{[0](l)}\left(\tilde{x}^{[0]}_{\mathcal{O} i_1}\right) \tau^{[0](l)}\left(\tilde{x}'^{[0]}_{\mathcal{O} i'_1}\right) \left(\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1} \left(\tfrac{1}{2}\alpha^{[0]2} + \tfrac{1}{2} \tfrac{1}{H^{[0]}} \left\langle \mathbf{x}^{[0]}_{\mathcal{O}}, \mathbf{x}'^{[0]}_{\mathcal{O}} \right\rangle \right) \right)^l \\
K^{[0]}_{\mathcal{W}\mathcal{O} i_1, i'_1}\left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta\right) &= \sum_{l=1}^{\infty} \left(\frac{1}{\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1}} \left(2 b^{[0]}_{\Delta i_1} b'^{[0]}_{\Delta i'_1} + 2 \left\langle \mathbf{W}^{[0]}_{\Delta:i_1}, \mathbf{W}'^{[0]}_{\Delta:i'_1} \right\rangle \right) \right)^l
\end{aligned}
$$

We define:

$$
\begin{aligned}
\theta\left(\zeta\right) &= \sum_{l=1}^{\infty} \zeta^l = \tfrac{\zeta}{1-\zeta} \quad \forall \zeta \in (-1, 1) \\
\sigma^{[j]}_{z,z'}\left(\zeta\right) &= \sum_{l=1}^{\infty} \left(\tfrac{1}{l!}\right)^2 \tau^{[j](l)}\left(z\right) \tau^{[j](l)}\left(z'\right) \zeta^l
\end{aligned}
\tag{30}
$$

$\forall j \in \mathbb{N}_D$, $z \in \mathbb{R}$, noting that:

$$\theta^{-1}(\zeta) = \tfrac{\zeta}{1+\zeta} \quad \forall \zeta \in (1, \infty) \tag{31}$$

Note that $\sigma^{[j]}_{z,z'}$ are derived from the power-series representation of the activation functions about z in such a manner that the constant term is removed, and, if $z = z'$, so too are all signs, making $\sigma^{[j]}_{z,z}$ an increasing function.[14] Using this definition and observation, provided the argument remains in the ROC of all relevant series, we see that:

$$\begin{aligned} \mathbf{K}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') &= \mathbf{K}^{[D-1]}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') \\ \mathbf{K}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) &= \mathbf{K}^{[D-1]}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) \end{aligned} \tag{32}$$

where, recursively $\forall j \in \mathbb{N}_D$:

$$\begin{aligned} &K^{[j]}_{\mathcal{X}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{x}, \mathbf{x}') = \ldots \\ &\ldots \begin{cases} \sigma^{[j]}_{\tilde{x}^{[j]}_{\mathcal{O} i_{j+1}}, \tilde{x}'^{[j]}_{\mathcal{O} i'_{j+1}}} \left(\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}} \left(\begin{array}{r} \left(\frac{1}{2}\alpha^{[j]2} + \frac{1}{H^{[j]}} \left\langle \mathbf{x}^{[j]}_{\mathcal{O}}, \mathbf{x}'^{[j]}_{\mathcal{O}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\frac{W^{[j]}_{\mathcal{O} i_j, i_{j+1}} W^{[j]}_{\mathcal{O} i_j, i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}}} + 1 \right) \ldots \\ \ldots \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} K^{[j-1]}_{\mathcal{X}_{\mathcal{O}} i_j, i_j}(\mathbf{x}, \mathbf{x}') \end{array} \right) \right) & \text{if } j > 0 \\ \sigma^{[0]}_{\tilde{x}^{[0]}_{\mathcal{O} i_1}, \tilde{x}'^{[0]}_{\mathcal{O} i'_1}} \left(\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1} \left(\frac{1}{2}\alpha^{[0]2} + \frac{1}{2} \frac{1}{H^{[0]}} \left\langle \mathbf{x}^{[0]}_{\mathcal{O}}, \mathbf{x}'^{[0]}_{\mathcal{O}} \right\rangle \right) \right) & \text{if } j = 0 \end{cases} \\ &K^{[j]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) = \ldots \\ &\ldots \begin{cases} \theta \left(\frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \left(\begin{array}{r} \left(2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} + \left\langle \mathbf{W}^{[j]}_{\Delta : i_{j+1}}, \mathbf{W}'^{[j]}_{\Delta : i'_{j+1}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \ldots \\ \ldots \frac{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}} \end{array} \right) \right) & \text{if } j > 0 \\ \theta \left(\frac{1}{\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1}} \left(2 b^{[0]}_{\Delta i_1} b'^{[0]}_{\Delta i'_1} + 2 \left\langle \mathbf{W}^{[0]}_{\Delta : i_1}, \mathbf{W}'^{[0]}_{\Delta : i'_1} \right\rangle \right) \right) & \text{if } j = 0 \end{cases} \end{aligned} \tag{33}$$

C.2. Induced Kernel Gradients

Later in the paper we will require the gradients of induced kernels. Recalling (32), (33) and applying the chain rule we see that:

$$\begin{aligned} \nabla_{\mathbf{x}} \mathbf{K}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') &= \left[\left[\frac{\partial}{\partial x_{i''_1}} \nabla_{\mathbf{x}} K^{[D-1]}_{\mathcal{X}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{x}, \mathbf{x}') \right]_{i''_1} \right]_{i_{j+1}, i'_{j+1}} \\ \nabla_{\mathbf{W}^{[j']}_\Delta} \mathbf{K}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) &= \left[\left[\frac{\partial}{\partial W^{[j']}_{\Delta i''_{j'}, i''_{j'+1}}} K^{[D-1]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) \right]_{i''_{j'}, i''_{j'+1}} \right]_{i_{j+1}, i'_{j+1}} \\ \nabla_{\mathbf{b}^{[j']}_\Delta} \mathbf{K}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) &= \left[\left[\frac{\partial}{\partial b^{[j']}_{\Delta i''_{j'+1}}} K^{[D-1]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) \right]_{i''_{j'+1}} \right]_{i_{j+1}, i'_{j+1}} \end{aligned} \tag{34}$$

[14]As an aside, we note that the power series representations of the activation functions $\tau^{[j]}$ about any z represent the same underlying function, minus $\tau^{[j]}(0)$, and so in principle we can start with a single such representation and use analytic continuation and reconstruct $\tau^{[j]}$ everywhere. Unfortunately the same is not true of $\sigma^{[j]}_{z,z'}$, which is perhaps unfortunate in this context.

where, recursively $\forall j \in \mathbb{N}_D$:

$$\frac{\partial}{\partial x_{i_1''}} K^{[j]}_{\mathcal{X}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{x}, \mathbf{x}') = \ldots$$
$$\ldots \begin{cases} \sigma^{[j](1)}_{\tilde{x}^{[j]}_{\mathcal{O} i_{j+1}}, \tilde{x}'^{[j]}_{\mathcal{O} i'_{j+1}}} \left(\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}} \left(\begin{array}{c} \left(\frac{1}{2}\alpha^{[j]2} + \frac{1}{H^{[j]}} \left\langle \mathbf{x}^{[j]}_{\mathcal{O}}, \mathbf{x}'^{[j]}_{\mathcal{O}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\frac{W^{[j]}_{\mathcal{O} i_j, i_{j+1}} W^{[j]}_{\mathcal{O} i_j, i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}}} + 1 \right) \ldots \\ \ldots \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} K^{[j-1]}_{\mathcal{X}_{\mathcal{O}} i_j, i_j}(\mathbf{x}, \mathbf{x}') \end{array} \right) \right) \ldots & \text{if } j > 0 \\ \quad \ldots \mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}} \sum_{i_j} \left(\frac{W^{[j]}_{\mathcal{O} i_j, i_{j+1}} W^{[j]}_{\mathcal{O} i_j, i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}}} + 1 \right) \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} \frac{\partial}{\partial x_{i_1''}} K^{[j-1]}_{\mathcal{X}_{\mathcal{O}} i_j, i_j}(\mathbf{x}, \mathbf{x}') & \\ \sigma^{[0](1)}_{\tilde{x}^{[0]}_{\mathcal{O} i_1}, \tilde{x}'^{[0]}_{\mathcal{O} i'_1}} \left(\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1} \left(\frac{1}{2}\alpha^{[0]2} + \frac{1}{2} \frac{1}{H^{[0]}} \langle \mathbf{x}, \mathbf{x}' \rangle \right) \right) \mu^{[0]}_{i_1} \mu^{[0]}_{i'_1} \frac{1}{H^{[0]}} \delta_{i_1, i_1''} \frac{1}{2} x'_{i'_1} & \text{if } j = 0 \end{cases} \tag{35}$$

$$\frac{\partial}{\partial W^{[j']}_{\Delta i''_{j'}, i''_{j'+1}}} K^{[j]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) = \ldots$$
$$\ldots \begin{cases} \theta^{(1)} \left(\frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \left(\begin{array}{c} \left(2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} + \left\langle \mathbf{W}^{[j]}_{\Delta : i_{j+1}}, \mathbf{W}'^{[j]}_{\Delta : i'_{j+1}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \ldots \\ \ldots \frac{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}} \end{array} \right) \right) \ldots & \text{if } j > j' \\ \quad \ldots \frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \ldots & \\ \quad \ldots \frac{1}{\omega^{[j]2}_{i_j}} \frac{\partial}{\partial W^{[j']}_{\Delta i''_{j'}, i''_{j'+1}}} K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) & \\ \theta^{(1)} \left(\frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \left(\begin{array}{c} \left(2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} + \left\langle \mathbf{W}^{[j]}_{\Delta : i_{j+1}}, \mathbf{W}'^{[j]}_{\Delta : i'_{j+1}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \ldots \\ \ldots \frac{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}} \end{array} \right) \right) \ldots & \text{if } j = j' > 0 \\ \quad \ldots \frac{1}{\mu^{[j]}_{i''_{j+1}} \mu^{[j]}_{i'_{j+1}}} \delta_{i_{j+1}, i''_{j+1}} \left(1 + \frac{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i''_j, i''_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta)}{\omega^{[j]2}_{i''_j}} \right) W'^{[j]}_{\Delta i''_j, i'_{j+1}} & \\ \theta^{(1)} \left(\frac{1}{\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1}} \left(2 b^{[0]}_{\Delta i_1} b'^{[0]}_{\Delta i'_1} + 2 \left\langle \mathbf{W}^{[0]}_{\Delta : i_1}, \mathbf{W}'^{[0]}_{\Delta : i'_1} \right\rangle \right) \right) \frac{1}{\mu^{[0]}_{i''_1} \mu^{[0]}_{i'_1}} \delta_{i_1, i''_1} 2 W'^{[0]}_{\Delta i''_0, i'_1} & \text{if } j = j' = 0 \end{cases} \tag{36}$$

$$
\begin{aligned}
&\frac{\partial}{\partial b^{[j']}_{\Delta i''_{j'+1}}} K^{[j]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}} \left(\mathbf{W}_\Delta, \mathbf{W}'_\Delta \right) = \ldots \\
&\ldots \begin{cases}
\theta^{(1)} \left(\frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \left(\begin{array}{l} \left(2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} + \left\langle \mathbf{W}^{[j]}_{\Delta : i_{j+1}}, \mathbf{W}'^{[j]}_{\Delta : i'_{j+1}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \ldots \\ \qquad \ldots \frac{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}} \end{array} \right) \right) \ldots & \\
\qquad \ldots \frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \frac{1}{\omega^{[j]2}_{i_j}} \ldots & \text{if } j > j' \\
\qquad\qquad \ldots \frac{\partial}{\partial b^{[j']}_{\Delta i''_{j'+1}}} K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) & \\
\theta^{(1)} \left(\frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \left(\begin{array}{l} \left(2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} + \left\langle \mathbf{W}^{[j]}_{\Delta : i_{j+1}}, \mathbf{W}'^{[j]}_{\Delta : i'_{j+1}} \right\rangle \right) + \ldots \\ \ldots \sum_{i_j} \left(\widetilde{\omega}^{[j]}_{i_j, i_{j+1}} \widetilde{\omega}^{[j]}_{i_j, i'_{j+1}} + W^{[j]}_{\Delta i_j, i_{j+1}} W'^{[j]}_{\Delta i_j, i'_{j+1}} \right) \ldots \\ \qquad \ldots \frac{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j, i_j}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}} \end{array} \right) \right) \ldots & \text{if } j = j' > 0 \\
\qquad\qquad \ldots \frac{1}{\mu^{[j]}_{i''_{j+1}} \mu^{[j]}_{i'_{j+1}}} \delta_{i_{j+1}, i''_{j+1}} 2 b'^{[j]}_{\Delta i'_{j+1}} & \\
\theta^{(1)} \left(\frac{1}{\mu^{[0]}_{i_1} \mu^{[0]}_{i'_1}} \left(2 b^{[0]}_{\Delta i_1} b'^{[0]}_{\Delta i'_1} + 2 \left\langle \mathbf{W}^{[0]}_{\Delta : i_1}, \mathbf{W}'^{[0]}_{\Delta : i'_1} \right\rangle \right) \right) \frac{1}{\mu^{[0]}_{i''_1} \mu^{[0]}_{i'_1}} \delta_{i_1, i''_1} 2 b'^{[0]}_{\Delta i'_1} & \text{if } j = j' = 0
\end{cases}
\end{aligned}
\tag{37}
$$

C.3. Induced Norms

We now consider the norms induced by the feature maps. Precisely, we wish to impose norms on $\mathcal{X}_{\mathcal{O}} \subset \mathbb{R}^{\infty \times m}$ and $\mathcal{W}_{\mathcal{O}} \subset \mathbb{R}^{\infty \times m}$ and calculate $\|\mathbf{\Phi}_{\mathcal{O}}(\mathbf{x})\|_{\mathcal{X}_{\mathcal{O}}}, \|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_{\mathcal{W}_{\mathcal{O}}}$ from this, which will subsequently provide the *induced* norms we will require when constructing the reproducing kernel Banach space for $\mathbf{f}_\Delta$. In principle we could apply an arbitrary norm to these spaces. However, recall that:

$$\mathbf{f}_\Delta \left(\mathbf{x} \right) = \left\langle \mathbf{\Phi}_{\mathcal{O}} \left(\mathbf{x} \right), \mathbf{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$$

and hence:

$$\left\| \mathbf{f}_\Delta \left(\mathbf{x} \right) \right\|_2 = \left\langle \mathbf{\Phi}_{\mathcal{O}} \left(\mathbf{x} \right), \mathbf{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\rangle_F$$

so a natural choice is a pair of dual norms - that is, a pair of norms $\| \cdot \|_{\mathcal{X}_{\mathcal{O}}}$ and $\| \cdot \|_{\mathcal{W}_{\mathcal{O}}}$ such that:

$$\|\mathbf{\Xi}\|_{\mathcal{X}_{\mathcal{O}}} = \sup_{\mathbf{\Omega} \in \mathcal{W}_{\mathcal{O}}} \left\{ \left\langle \mathbf{\Xi}, \mathbf{\Omega} \right\rangle_F \middle| \, \|\mathbf{\Omega}\|_{\mathcal{W}_{\mathcal{O}}} \leq 1 \right\}$$

and hence by Hölder's inequality:

$$\left\| \mathbf{f}_\Delta \left(\mathbf{x} \right) \right\|_2 \leq \left\| \mathbf{\Phi}_{\mathcal{O}} \left(\mathbf{x} \right) \right\|_{\mathcal{X}_{\mathcal{O}}} \left\| \mathbf{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}}$$

In the present context we find it convenient to use the Frobenius (elementwise Euclidean) norms:

$$
\begin{aligned}
\|\mathbf{\Xi}\|^2_{\mathcal{X}_{\mathcal{O}}} &= \|\mathbf{\Xi}\|^2_F = \textstyle\sum_{i_D} \|\mathbf{\Xi}_{:i_D}\|^2_2 \\
\|\mathbf{\Omega}\|^2_{\mathcal{W}_{\mathcal{O}}} &= \|\mathbf{\Omega}\|^2_F = \textstyle\sum_{i_D} \|\mathbf{\Omega}_{:i_D}\|^2_2
\end{aligned}
$$

and so:

$$
\begin{aligned}
\left\| \mathbf{\Phi}_{\mathcal{O}} \left(\mathbf{x} \right) \right\|^2_{\mathcal{X}_{\mathcal{O}}} &= \mathrm{Tr} \left(\mathbf{K}_{\mathcal{X}_{\mathcal{O}}} \left(\mathbf{x}, \mathbf{x} \right) \right) \\
\left\| \mathbf{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|^2_{\mathcal{W}_{\mathcal{O}}} &= \mathrm{Tr} \left(\mathbf{K}_{\mathcal{W}_{\mathcal{O}}} \left(\mathbf{W}_\Delta, \mathbf{W}_\Delta \right) \right)
\end{aligned}
$$

Before proceeding we define:

$$s^{[j]2} = \begin{cases} \frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{H^{[-1]}}{H^{[0]}} M^{[-1]2} & \text{if } j = 0 \\ \frac{1}{2}\alpha^{[j]2} + \frac{H^{[j-1]}}{H^{[j]}} M^{[j-1]2} & \text{otherwise} \end{cases}$$
$$t_{\Delta i_{j+1}}^{[j]2} = \begin{cases} \left[2b_{\Delta i_1}^{[0]2} + 2\left\|\mathbf{W}_{\Delta:i_1}^{[0]}\right\|_2^2\right]_{i_1} & \text{if } j = 0 \\ \left[2b_{\Delta i_{j+1}}^{[j]2} + \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2\right]_{i_{j+1}} & \text{otherwise} \end{cases} \quad \forall j \in \mathbb{N}_D, i_{j+1}$$

where, roughly speaking, $t^{[j-1]2}$ represents the magnitude of the weight-step for layer j and $s^{[j]2}$ represents the upper bound on the image of a vector $\mathbf{x} \in \mathbb{X}$ at the input to layer j. Using this notation:

$$\begin{aligned} \|\mathbf{\Phi}_{\mathcal{O}}(\mathbf{x})\|_{\mathcal{X}_{\mathcal{O}}}^2 &= \textstyle\sum_{i_D} \left\|\mathbf{\Phi}_{\mathcal{O}:i_D}^{[D-1]}(\mathbf{x})\right\|_2^2 \\ \|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_{\mathcal{W}_{\mathcal{O}}}^2 &= \textstyle\sum_{i_D} \left\|\mathbf{\Psi}_{\mathcal{O}:i_D}^{[D-1]}(\mathbf{W}_\Delta)\right\|_2^2 \end{aligned} \tag{38}$$

where, recursively $\forall j \in \mathbb{N}_D$:

$$\begin{aligned} \left\|\mathbf{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x})\right\|_2^2 &= \begin{cases} \sigma_{\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}, \tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}}^{[j]} \left(\mu_{i_{j+1}}^{[j]2} \left(\begin{array}{l} \left(\frac{1}{2}\alpha^{[j]2} + \frac{1}{H^{[j]}}\left\|\mathbf{x}_{\mathcal{O}}^{[j]}\right\|_2^2\right) + \ldots \\ \quad \ldots \sum_{i_j} \left(1 + \frac{W_{\mathcal{O}i_j,i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}\right) \frac{\omega_{i_j}^{[j]2}}{H^{[j]}} \left\|\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right\|_2^2 \end{array} \right)\right) & \text{if } j > 0 \\ \sigma_{\tilde{x}_{\mathcal{O}i_1}^{[0]}, \tilde{x}_{\mathcal{O}i_1}^{[0]}}^{[0]} \left(\mu_{i_1}^{[0]2}\left(\frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{1}{H^{[0]}}\left\|\mathbf{x}_{\mathcal{O}}^{[0]}\right\|_2^2\right)\right) & \text{if } j = 0 \end{cases} \\ \left\|\mathbf{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{W}_\Delta)\right\|_2^2 &= \begin{cases} \theta\left(\frac{1}{\mu_{i_{j+1}}^{[j]2}}\left(t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2}\right)\frac{\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2}{\omega_{i_j}^{[j]2}}\right)\right) & \text{if } j > 0 \\ \theta\left(\frac{1}{\mu_{i_1}^{[0]2}} t_{\Delta i_1}^{[0]2}\right) & \text{if } j = 0 \end{cases} \end{aligned} \tag{39}$$

C.4. Induced Norm Gradients

Later in the paper we will require the gradients of induced norms. Recalling (38), (39) and applying the chain rule we see that:

$$\begin{aligned} \nabla_{\mathbf{x}} \|\mathbf{\Phi}_{\mathcal{O}}(\mathbf{x})\|_{\mathcal{X}_{\mathcal{O}}}^2 &= \textstyle\sum_{i_D} \left[\frac{\partial}{\partial x_{i_1''}} \left\|\mathbf{\Phi}_{\mathcal{O}:i_D}^{[D-1]}(\mathbf{x})\right\|_2^2 \right]_{i_1''} \\ \nabla_{\mathbf{W}_\Delta^{[j']}} \|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_{\mathcal{W}_{\mathcal{O}}}^2 &= \textstyle\sum_{i_D} \left[\frac{\partial}{\partial W_{\Delta i_{j'}'', i_{j'+1}''}^{[j']}} \left\|\mathbf{\Psi}_{\mathcal{O}:i_D}^{[D-1]}(\mathbf{W}_\Delta)\right\|_2^2 \right]_{i_{j'}'', i_{j'+1}''} \\ \nabla_{\mathbf{b}_\Delta^{[j']}} \|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_{\mathcal{W}_{\mathcal{O}}}^2 &= \textstyle\sum_{i_D} \left[\frac{\partial}{\partial b_{\Delta i_{j'}''}^{[j']}} \left\|\mathbf{\Psi}_{\mathcal{O}:i_D}^{[D-1]}(\mathbf{W}_\Delta)\right\|_2^2 \right]_{i_{j'+1}''} \end{aligned} \tag{40}$$

where, recursively $\forall j \in \mathbb{N}_D$:

$$\begin{aligned} &\frac{\partial}{\partial x_{i_1'}} \left\|\mathbf{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x})\right\|_2^2 = \ldots \\ &\ldots \begin{cases} \sigma_{\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}, \tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}}^{[j](1)} \left(\mu_{i_{j+1}}^{[j]2} \left(\begin{array}{l} \left(\frac{1}{2}\alpha^{[j]2} + \frac{1}{H^{[j]}}\left\|\mathbf{x}_{\mathcal{O}}^{[j]}\right\|_2^2\right) + \ldots \\ \quad \ldots \sum_{i_j} \left(\frac{W_{\mathcal{O}i_j,i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}} + 1\right) \frac{\omega_{i_j}^{[j]2}}{H^{[j]}} \left\|\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right\|_2^2 \end{array} \right)\right) \ldots & \text{if } j > 0 \\ \qquad \ldots \mu_{i_{j+1}}^{[j]2} \sum_{i_j} \left(\frac{W_{\mathcal{O}i_j,i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}} + 1\right) \frac{\omega_{i_j}^{[j]2}}{H^{[j]}} \frac{\partial}{\partial x_{i_1'}} \left\|\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right\|_2^2 & \\ \sigma_{\tilde{x}_{\mathcal{O}i_1}^{[0]}, \tilde{x}_{\mathcal{O}i_1}^{[0]}}^{[0](1)} \left(\mu_{i_1}^{[0]2}\left(\frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{1}{H^{[0]}}\|\mathbf{x}\|_2^2\right)\right) \mu_{i_1}^{[0]2} \frac{1}{H^{[0]}} \delta_{i_1,i_1'} x_{i_1'} & \text{if } j = 0 \end{cases} \end{aligned} \tag{41}$$

$$\frac{\partial}{\partial W^{[j']}_{\Delta i'_{j'}, i'_{j'+1}}} \left\| \boldsymbol{\Psi}^{[j]}_{\mathcal{O}:i_{j+1}} \left(\mathbf{W}_\Delta \right) \right\|_2^2 = \ldots$$
$$\ldots \begin{cases} \theta^{(1)} \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} \left(t^{[j]}_{\Delta i_{j+1}} + \sum_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j, i_{j+1}} + W^{[j]2}_{\Delta i_j, i_{j+1}} \right) \frac{\left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j} (\mathbf{W}_\Delta) \right\|_2^2}{\omega^{[j]2}_{i_j}} \right) \right) \ldots \\ \quad \ldots \frac{1}{\mu^{[j]2}_{i_{j+1}}} \sum_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j, i_{j+1}} + W^{[j]2}_{\Delta i_j, i_{j+1}} \right) \frac{1}{\omega^{[j]2}_{i_j}} \frac{\partial}{\partial W^{[j']}_{\Delta i'_{j'}, i'_{j'+1}}} \left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j} \left(\mathbf{W}_\Delta \right) \right\|_2^2 & \text{if } j > j' \\ \theta^{(1)} \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} \left(t^{[j]}_{\Delta i_{j+1}} + \sum_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j, i_{j+1}} + W^{[j]2}_{\Delta i_j, i_{j+1}} \right) \frac{\left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j} (\mathbf{W}_\Delta) \right\|_2^2}{\omega^{[j]2}_{i_j}} \right) \right) \ldots \\ \quad \ldots \frac{1}{\mu^{[j]2}_{i'_{j+1}}} \delta_{i_{j+1}, i'_{j+1}} 2 \left(1 + \frac{\left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i'_j} (\mathbf{W}_\Delta) \right\|_2^2}{\omega^{[j]2}_{i'_j}} \right) W^{[j]}_{\Delta i'_j, i_{j+1}} & \text{if } j = j' > 0 \\ \theta^{(1)} \left(\frac{1}{\mu^{[0]2}_{i_1}} t^{[0]}_{\Delta i_1} \right) \frac{1}{\mu^{[0]2}_{i'_1}} \delta_{i_1, i'_1} 4 W^{[0]}_{\Delta i'_0, i_1} & \text{if } j = j' = 0 \end{cases} \tag{42}$$

$$\frac{\partial}{\partial b^{[j']}_{\Delta i'_{j'+1}}} \left\| \boldsymbol{\Psi}^{[j]}_{\mathcal{O}:i_{j+1}} \left(\mathbf{W}_\Delta \right) \right\|_2^2 = \ldots$$
$$\ldots \begin{cases} \theta^{(1)} \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} \left(t^{[j]}_{\Delta i_{j+1}} + \sum_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j, i_{j+1}} + W^{[j]2}_{\Delta i_j, i_{j+1}} \right) \frac{\left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j} (\mathbf{W}_\Delta) \right\|_2^2}{\omega^{[j]2}_{i_j}} \right) \right) \ldots \\ \quad \ldots \frac{1}{\mu^{[j]2}_{i_{j+1}}} \sum_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j, i_{j+1}} + W^{[j]2}_{\Delta i_j, i_{j+1}} \right) \frac{1}{\omega^{[j]2}_{i_j}} \frac{\partial}{\partial b^{[j']}_{\Delta i'_{j'+1}}} \left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j} \left(\mathbf{W}_\Delta \right) \right\|_2^2 & \text{if } j > j' \\ \theta^{(1)} \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} \left(t^{[j]}_{\Delta i_{j+1}} + \sum_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j, i_{j+1}} + W^{[j]2}_{\Delta i_j, i_{j+1}} \right) \frac{\left\| \boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j} (\mathbf{W}_\Delta) \right\|_2^2}{\omega^{[j]2}_{i_j}} \right) \right) \ldots \\ \quad \ldots \frac{1}{\mu^{[j]2}_{i'_{j+1}}} \delta_{i_{j+1}, i'_{j+1}} 4 b^{[j]}_{\Delta i_{j+1}} & \text{if } j = j' > 0 \\ \theta^{(1)} \left(\frac{1}{\mu^{[0]2}_{i_1}} t^{[0]}_{\Delta i_1} \right) \frac{1}{\mu^{[0]2}_{i'_1}} \delta_{i_1, i'_1} 4 b^{[0]}_{\Delta i_1} & \text{if } j = j' = 0 \end{cases} \tag{43}$$

C.5. Properties of θ and σ

Here we present an important preliminary property of the θ and and $\sigma_{z,z'}$ functions. Recall that:

$$\begin{aligned} \theta\left(\zeta\right) &= \sum_{l=1}^{\infty} \zeta^l = \tfrac{\zeta}{1-\zeta} \quad \forall \zeta \in (-1,1) \\ \sigma^{[j]}_{z,z'}\left(\zeta\right) &= \sum_{l=1}^{\infty} \left(\tfrac{1}{l!}\right)^2 \tau^{[j](l)}\left(z\right) \tau^{[j](l)}\left(z'\right) \zeta^l \end{aligned}$$

where:

$$\begin{aligned} \theta^{-1}\left(\zeta\right) &= \tfrac{\zeta}{1+\zeta} \\ \theta^{(1)}\left(\zeta\right) &= \tfrac{1}{(1-\zeta)^2} \\ \theta^{(1)-1}\left(\zeta\right) &= \tfrac{\sqrt{\zeta}-1}{\sqrt{\zeta}} \end{aligned}$$

We note that, for $z \in \mathbb{R}$, $j \in \mathbb{N}_D$:

- $\sigma^{[j]}_{z,z} : (-\sqrt{\rho^{[j]}}, \sqrt{\rho^{[j]}}) \to \mathbb{R}$ is, in general, unbounded, smooth and non-Lipchitz.
- $\sigma^{[j]}_{z,z} : [0, \sqrt{\rho^{[j]}}) \to \mathbb{R}_+$ is, in general, unbounded, smooth, strictly increasing and non-Lipschitz.
- $\theta : (-1,1) \to \mathbb{R}$ is unbounded, smooth and non-Lipchitz.

- $\theta : [0, 1) \to [0, \infty)$ is unbounded, smooth, strictly increasing and non-Lipschitz.
- $\theta^{-1} : [0, \infty) \to [0, 1)$ is bounded, smooth, strictly increasing and Lipschitz.
- $\theta^{(1)} : [0, 1) \to [1, \infty)$ is unbounded, smooth, strictly increasing and non-Lipschitz.
- $\theta^{(1)-1} : [1, \infty) \to [0, 1)$ is bounded, smooth, strictly increasing and Lipschitz.

The unbounded, non-Lipschitz nature of $\sigma^{[j]}_{z,z}$ and θ function is problematic when analysing the convergence properties of our feature maps. However if we restrict the domains of $\sigma^{[j]}_{z,z}$ and θ by defining $\epsilon^{[j]}_{\psi}, \epsilon^{[j]}_{\phi} \in (0, 1)\ \forall j \in \mathbb{N}_D$ and:

$$\overline{\sigma}^{[j]}(\zeta) = \max_{z \in \mathbb{R}_+ \cup \{0\}} \left\{ \sigma^{[j]}_{z,z}(\zeta) \right\}$$

then:

- $\overline{\sigma}^{[j]} : [0, (1 - \epsilon^{[j]}_{\phi})\sqrt{\rho^{[j]}}] \to [0, \frac{\phi^{[j]}}{H^{[j]}}] \subset \mathbb{R}_+$ is bounded, smooth, strictly increasing and $L^{[j]}_{\phi}$-Lipschitz, where:

$$\begin{aligned} \tfrac{\phi^{[j]}}{H^{[j]}} &= \overline{\sigma}^{[j]} \left((1 - \epsilon^{[j]}_{\phi})\sqrt{\rho^{[j]}} \right) \\ L^{[j]}_{\phi} &= \nabla_{+1} \overline{\sigma}^{[j]} \left(\left(1 - \epsilon^{[j]}_{\phi} \right)) \sqrt{\rho^{[j]}} \right) \end{aligned}$$

  and we note that $\frac{\phi^{[j]}}{H^{[j]}} \to \infty$ as $\epsilon^{[j]}_{\phi} \to 0$.

- $\overline{\sigma}^{[j]-1} : [0, \frac{\phi^{[j]}}{H^{[j]}}] \to [0, (1 - \epsilon^{[j]}_{\phi})\sqrt{\rho^{[j]}}]$ is bounded, smooth and strictly increasing.
- $\theta : [0, 1 - \epsilon^{[j]}_{\psi}] \to [0, \frac{\psi^{[j]}}{H^{[j]}}] \subset \mathbb{R}_+$ is bounded, smooth, strictly increasing and $L^{[j]}_{\psi}$-Lipschitz, where:

$$\begin{aligned} \tfrac{\psi^{[j]}}{H^{[j]}} &= \theta \left(1 - \epsilon^{[j]}_{\psi} \right) = \tfrac{1 - \epsilon^{[j]}_{\psi}}{\epsilon^{[j]}_{\psi}} \\ L^{[j]}_{\psi} &= \tfrac{\partial}{\partial \zeta} \theta \left(1 - \epsilon^{[j]}_{\psi} \right) = \tfrac{1}{\epsilon^{[j]2}_{\psi}} \end{aligned}$$

  and we note that $\frac{\psi^{[j]}}{H^{[j]}} \to \infty$ as $\epsilon^{[j]}_{\psi} \to 0$.

- $\theta^{-1} : [0, \frac{\psi^{[j]}}{H^{[j]}}] \to [0, 1 - \epsilon^{[j]}_{\psi}]$ is bounded, smooth and strictly increasing.

This restriction in domain will be used in the following section when we analyse the convergence and finiteness of our induced norms. We also find it useful to define the related functions:

$$\begin{aligned} \kappa(\zeta) &= \zeta \theta^{(1)}(\zeta) = \tfrac{\zeta}{(1-\zeta)^2} \\ \kappa^{-1}(\zeta) &= \tfrac{2\zeta + 1 - \sqrt{4\zeta + 1}}{2\zeta} \\ \lambda^{-1}(\zeta) &= \tfrac{1}{\zeta} \kappa^{-1}(\zeta) = \tfrac{2\zeta + 1 - \sqrt{4\zeta + 1}}{2\zeta^2} \\ \lambda(\zeta) &= \tfrac{1 - \sqrt{y}}{y} \end{aligned}$$

where we note that:

- $\kappa : [0, 1) \to [0, \infty)$ is strictly increasing.
- $\kappa^{-1} : [0, \infty) \to [0, 1)$ is strictly increasing.
- $\lambda : (0, 1] \to [0, \infty)$ is strictly decreasing.
- $\lambda^{-1} : [0, \infty) \to (0, 1]$ is strictly decreasing.

C.6. Convergence Conditions

We finish this section by considering the issue of the convergence of the induced norms, noting that these results also apply to the induced kernels as:

$$\begin{aligned}\det\left(\mathbf{K}_{\mathcal{X}_\mathcal{O}}\left(\mathbf{x},\mathbf{x}'\right)\right) &\leq \left(\frac{\operatorname{Tr}\left(\mathbf{K}_{\mathcal{X}_\mathcal{O}}\left(\mathbf{x},\mathbf{x}'\right)\right)}{m}\right)^m \leq \left(\frac{\left\langle \boldsymbol{\Phi}_\mathcal{O}(\mathbf{x}),\boldsymbol{\Phi}_\mathcal{O}(\mathbf{x}')\right\rangle_F}{m}\right)^m \leq \left(\frac{\|\boldsymbol{\Phi}_\mathcal{O}(\mathbf{x})\|_{\mathcal{X}_\mathcal{O}}\left\|\boldsymbol{\Phi}_\mathcal{O}(\mathbf{x}')\right\|_{\mathcal{X}_\mathcal{O}}}{m}\right)^m \\ \det\left(\mathbf{K}_{\mathcal{W}_\mathcal{O}}\left(\mathbf{W}_\Delta,\mathbf{W}'_\Delta\right)\right) &\leq \left(\frac{\operatorname{Tr}\left(\mathbf{K}_{\mathcal{W}_\mathcal{O}}\left(\mathbf{W}_\Delta,\mathbf{W}'_\Delta\right)\right)}{m}\right)^m \leq \left(\frac{\left\langle \boldsymbol{\Psi}_\mathcal{O}(\mathbf{W}_\Delta),\boldsymbol{\Psi}_\mathcal{O}(\mathbf{W}'_\Delta)\right\rangle_F}{m}\right)^m \leq \ldots \\ &\qquad\ldots\left(\frac{\|\boldsymbol{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\|_{\mathcal{W}_\mathcal{O}}\left\|\boldsymbol{\Psi}_\mathcal{O}(\mathbf{W}'_\Delta)\right\|_{\mathcal{W}_\mathcal{O}}}{m}\right)^m\end{aligned}$$

and using the positive definiteness of the induced kernels. Whereas for the feature representation of $\mathbf{f}_\Delta$ it sufficed to ensure that $\|\tilde{\mathbf{x}}_\Delta^{[j]}\|_\infty \leq \rho^{[j]}$ for all i_{j+1}, j, we cannot use this directly here as we the feature map $\boldsymbol{\Phi}$ does not have access to $\mathbf{W}_\Delta$ (which is indeterminant in this context in any case), and likewise the feature map $\boldsymbol{\Psi}$ does not have access to $\mathbf{x}$.

Recall our definition:

$$\begin{aligned}s^{[j]2} &= \begin{cases} \frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{H^{[-1]}}{H^{[0]}}M^{[-1]2} & \text{if } j=0 \\ \frac{1}{2}\alpha^{[j]2} + \frac{H^{[j-1]}}{H^{[j]}}M^{[j-1]2} & \text{otherwise}\end{cases} \\ t_{\Delta i_{j+1}}^{[j]2} &= \begin{cases} \left[2b_{\Delta i_1}^{[0]2} + 2\left\|\mathbf{W}_{\Delta:i_1}^{[0]}\right\|_2^2\right]_{i_1} & \text{if } j=0 \\ \left[2b_{\Delta i_{j+1}}^{[j]2} + \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2\right]_{i_{j+1}} & \text{otherwise}\end{cases} \qquad \forall j\in\mathbb{N}_D, i_{j+1}\end{aligned}$$

Using this notation, the following theorems present conditions for the convergence (finiteness) of the induced norms for all $\mathbf{x}\in\mathbb{X}$ and given weight-step $\mathbf{W}_\Delta$:

Theorem 5. *Let $\epsilon_\phi^{[j]} \in (0,1)\ \forall j\in\mathbb{N}_D$ and for a given neural network and initial weights $\mathbf{W}_\mathcal{O}$ define:*

$$\frac{\phi^{[j]}}{H^{[j]}} = \overline{\sigma}^{[j]}\left(\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}\right) \quad \forall j\in\mathbb{N}_D$$

If the scale factors satisfy:

$$\begin{aligned}\mu_{i_1}^{[0]2} &\leq \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} && \forall i_1 \\ \mu_{i_{j+1}}^{[j]2} &\leq \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\sum_{i_j}\omega_{i_j}^{[j]2}\left(\frac{W_{\mathcal{O}i_ji_{j+1}}^{[j]2}}{\tilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\frac{\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}} && \forall j\in\mathbb{N}_D\backslash\{0\}, i_{j+1}\end{aligned}$$

then $\left\|\boldsymbol{\Phi}_\mathcal{O}^{[j]}(\mathbf{x})\right\|_F^2 \leq \phi^{[j]}\ \forall\mathbf{x}\in\mathbb{X},\ j\in\mathbb{N}_D$.

Proof. Recalling (38), (39), we have that:

$$\begin{aligned}&\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x})\right\|_2^2 = \ldots \\ &\ldots\begin{cases}\sigma_{\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]},\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}}^{[j]}\left(\mu_{i_{j+1}}^{[j]2}\left(\left(\alpha^{[j]2}+\frac{2}{H^{[j]}}\left\|\mathbf{x}_\mathcal{O}^{[j]}\right\|_2^2\right)+\sum_{i_j}\frac{\omega_{i_j}^{[j]2}}{H^{[j]}}\left(\frac{W_{\mathcal{O}i_ji_{j+1}}^{[j]2}}{\tilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right\|_2^2\right)\right) & \text{if } j>0 \\ \sigma_{\tilde{x}_{\mathcal{O}i_1}^{[0]},\tilde{x}_{\mathcal{O}i_1}^{[0]}}^{[0]}\left(\mu_{i_1}^{[0]2}\left(\alpha^{[0]2}+\frac{1}{H^{[0]}}\left\|\mathbf{x}_\mathcal{O}^{[0]}\right\|_2^2\right)\right) & \text{if } j=0\end{cases}\end{aligned}$$

where as discussed in section C.5 on the restricted range $\sigma_{z,z}^{[j]} : [0,(1-\epsilon_\phi^{[j]})\sqrt{\rho^{[j]}}] \to [0,\phi^{[j]}]$ is bounded, increasing and Lipschitz for all $z\in\mathbb{R}$. By our assumptions we have that:

$$\begin{aligned}\alpha^{[0]2}+\frac{1}{H^{[0]}}\left\|\mathbf{x}_\mathcal{O}^{[0]}\right\|_2^2 &\leq s^{[0]2} \\ \alpha^{[j]2}+\frac{2}{H^{[j]}}\left\|\mathbf{x}_\mathcal{O}^{[j]}\right\|_2^2 &\leq s^{[j]2} \quad \forall j\in\mathbb{N}_D\backslash\{0\}\end{aligned}$$

and hence:

$$\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{x}\right)\right\|_2^2 \leq \begin{cases} \sigma_{\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]},\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}}^{[j]}\left(\mu_{i_{j+1}}^{[j]2}\left(s^{[j]2}+\sum_{i_j}\frac{\omega_{i_j}^{[j]2}}{H^{[j]}}\left(\frac{W_{\mathcal{O}i_j i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right)\right\|_2^2\right)\right) & \text{if } j>0 \\ \sigma_{\tilde{x}_{\mathcal{O}i_1}^{[0]},\tilde{x}_{\mathcal{O}i_1}^{[0]}}^{[0]}\left(\mu_{i_1}^{[0]2}s^{[0]2}\right) & \text{if } j=0 \end{cases}$$

We will construct sufficient conditions for convergence by bounding this bound. Starting with the input layer, the convergence of $\|\boldsymbol{\Phi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{x})\|_2^2$ is assured if the argument of $\sigma_{\tilde{x}_{\mathcal{O}i_1}^{[0]},\tilde{x}_{\mathcal{O}i_1}^{[0]}}^{[0]}$ lies in the restricted ROC, i.e.:

$$\mu_{i_1}^{[0]2}s^{[0]2} \leq \left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}} \quad \forall i_1 \tag{44}$$

and moreover if this condition is met then:

$$\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_1}^{[0]}\left(\mathbf{x}\right)\right\|_2^2 \leq \frac{\phi^{[0]}}{H^{[0]}} \quad \forall i_1 \tag{45}$$

For layer $j\in\mathbb{N}_D\backslash\{0\}$, convergence of $\|\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x})\|_2^2$ is assured if the argument of $\sigma_{\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]},\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}}^{[j]}$ lies in the restricted ROC:

$$\mu_{i_{j+1}}^{[j]2}\left(s^{[j]2}+\sum_{i_j}\frac{\omega_{i_j}^{[j]2}}{H^{[j]}}\left(\frac{W_{\mathcal{O}i_j i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{x}\right)\right\|_2^2\right) \leq \left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}} \quad \forall i_{j+1} \tag{46}$$

and moreover if this condition is met then:

$$\left\|\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{x}\right)\right\|_2^2 \leq \frac{\phi^{[j]}}{H^{[j]}} \quad \forall i_{j+1}$$

Assuming layers $j'<j$ satisfy (44), (46) then convergence of $\|\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x})\|_2^2$ is assured if:

$$\mu_{i_{j+1}}^{[j]2}\left(s^{[j]2}+\sum_{i_j}\frac{\omega_{i_j}^{[j]2}}{H^{[j]}}\left(\frac{W_{\mathcal{O}i_j i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\frac{\phi^{[j-1]}}{H^{[j-1]}}\right) \leq \left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}} \quad \forall i_{j+1}$$

or, sufficiently:

$$\mu_{i_{j+1}}^{[j]2} \leq \frac{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}{s^{[j]2}+\frac{1}{H^{[j]}}\sum_{i_j}\omega_{i_j}^{[j]2}\left(\frac{W_{\mathcal{O}i_j i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\frac{\phi^{[j-1]}}{H^{[j-1]}}} \quad \forall j\in\mathbb{N}_D\backslash\left\{0\right\}, i_{j+1}$$

which completes the proof. □

Theorem 6. *Let $\epsilon_\psi^{[j]}\in(0,1)\ \forall j\in\mathbb{N}_D$ and for a given neural network and initial weights $\mathbf{W}_\mathcal{O}$ define:*

$$\frac{\psi^{[j]}}{H^{[j]}} = \theta\left(1-\epsilon_\psi^{[j]}\right) = \frac{1-\epsilon_\psi^{[j]}}{\epsilon_\psi^{[j]}}$$

For a weight-step $\mathbf{W}_\Delta$, if:

$$t_{\Delta i_1}^{[0]2} \leq \left(1-\epsilon_\psi^{[0]}\right)\mu_{i_1}^{[0]2} \quad \forall i_1$$

$$t_{\Delta i_{j+1}}^{[j]2}+\sum_{i_j}\frac{\left\|\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2}{\omega_{i_j}^{[j]2}}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}+W_{\Delta i_j,i_{j+1}}^{[j]2}\right) \leq \left(1-\epsilon_\psi^{[j]}\right)\mu_{i_{j+1}}^{[j]2} \quad \forall j\in\mathbb{N}_D\backslash\left\{0\right\}, i_{j+1}$$

then $\left\|\boldsymbol{\Psi}_\mathcal{O}^{[j]}(\mathbf{W}_\Delta)\right\|_F^2 \leq \psi^{[j]}$ for all $j\in\mathbb{N}_D$.

Proof. Recalling (38), (39), we have that:

$$\left\|\boldsymbol{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta\right)\right\|_2^2 = \ldots$$

$$\ldots\begin{cases} \theta\left(\frac{1}{\mu_{i_{j+1}}^{[j]2}}\left(t_{\Delta i_{j+1}}^{[j]2}+\sum_{i_j}\frac{1}{\omega_{i_j}^{[j]2}}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}+W_{\Delta i_j,i_{j+1}}^{[j]2}\right)\left\|\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\right)\right) & \text{if } j>0 \\ \theta\left(\frac{1}{\mu_{i_1}^{[0]2}}t_{\Delta i_1}^{[0]2}\right) & \text{if } j=0 \end{cases}$$

where as discussed in section C.5 on the restricted range $\theta : [0, (1-\epsilon_\psi^{[j]})] \to [0, \psi^{[j]}]$ is bounded, increasing and $\frac{1}{\epsilon_\psi^{[j]2}}$-Lipschitz. Starting at layer 0 we see that if:

$$\frac{t_{\Delta i_1}^{[0]2}}{\mu_{i_1}^{[0]2}} \leq 1 - \epsilon_\psi^{[0]}$$

Then:

$$\left\| \boldsymbol{\Psi}_{\mathcal{O}:i_1}^{[0]} \left(\mathbf{W}_\Delta \right) \right\|_2^2 \leq \frac{\psi^{[0]}}{H^{[0]}} \quad \forall i_1$$

Moreover for layer j, if:

$$\frac{1}{\mu_{i_{j+1}}^{[j]2}} \left(t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\| \boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta) \right\|_2^2}{\omega_{i_j}^{[j]2}} \left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2} \right) \right) \leq 1 - \epsilon_\psi^{[j]} \quad \forall i_{j+1}$$

Then:

$$\left\| \boldsymbol{\Psi}_{\mathcal{O}:i_2}^{[1]} \left(\mathbf{W}_\Delta \right) \right\|_2^2 \leq \frac{\psi^{[j]}}{H^{[j]}} \quad \forall i_2$$

which gives our sufficient conditions in the general case:

$$\begin{aligned}
\mu_{i_1}^{[0]2} &\geq \frac{t_{\Delta i_1}^{[0]2}}{1-\epsilon_\psi^{[0]}} && \forall i_1 \\
\mu_{i_{j+1}}^{[j]2} &\geq \frac{1}{1-\epsilon_\psi^{[j]}} \left(t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\| \boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta) \right\|_2^2}{\omega_{i_j}^{[j]2}} \left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2} \right) \right) && \forall j \in \mathbb{N}_D \setminus \{0\}, i_{j+1}
\end{aligned}$$

which completes the proof. □

Note that this theorem may be equivalently stated using a bound on the scale factors:

$$\begin{aligned}
\mu_{i_1}^{[0]2} &\geq \frac{1}{1-\epsilon_\psi^{[0]}} t_{\Delta i_1}^{[0]2} \quad \forall i_1 \\
\mu_{i_{j+1}}^{[j]2} &\geq \frac{1}{1-\epsilon_\psi^{[j]}} \left(t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\| \boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta) \right\|_2^2}{\omega_{i_j}^{[j]2}} \left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2} \right) \right) \quad \forall j \in \mathbb{N}_D \setminus \{0\}, i_{j+1}
\end{aligned}$$

rather than as a bound on the weight-step. Thus theorems 5 and 6 give some insight into the role of the scale factors in our scheme. Loosely speaking, the scale factors must be chosen as a tradeoff between the convergence of the feature maps $\boldsymbol{\Phi}_\mathcal{O}$ and $\boldsymbol{\Psi}_\mathcal{O}$, and these theorems how this trade-off may be tuned, and to what extent. In particular:

- The upper bounds on the scale factors given in theorem 5 get larger (less strict) as the "size" of $\mathbb{X}$ and and the range of outputs of all layers of the network, as measured by $s^{[j]}$, gets smaller, diverging to ∞ (unbounded) in the limit $s^{[j]} \to 0^+$. Of course the factors $s^{[j]}$ are determined by the structure of the network, the range of the dataset $\mathbb{X}$ and our choice of α, so in a practice this upper bound on the scale factors is fixed.
- The lower bounds on the scale factors given in theorem 6 get smaller (less strict) as the size of the weight-step, plus a factor dependent on $\widetilde{\omega}_{i_j,i_{j+1}}^{[j]}$, gets smaller, decreasing to 0 (unbounded) as these go to zero. Thus we see that the scale factors are effectively bounded below by the convergence conditions on the feature map $\boldsymbol{\Psi}_\mathcal{O}$, with the bound being determined by the size of the weight-step. Unlike the upper bound given in theorem 5, we can make the lower bound arbitrarily small by requiring that the weight-step and offset be sufficiently small.

The interaction of these two observations - the upper and lower bounds on the scale factors - determines what weight-steps we can model over the set of all inputs $\mathbb{X}$. Furthermore, the presence of the scale factors in the lower bounds in theorem 6, gives an indication of how they may be chosen. To be precise, for the bounds to make sense we need that:

- The shadow weights $\widetilde{\omega}_{i_j,i_{j+1}}^{[j]}$ must be of the same (or lower) order (scale) as the weight-step to ensure that, for a sufficiently small weight-step, the lower bound on the scale factors becomes lower than the upper bound on the scale factors.

- The shadow weights $\omega_{i_j}^{[j]}$ must be of the same (or greater) order (scale) as the norm $\|\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\|_2^2$ to cancel out its influence on the lower bound on the scale factors.

We make this intuition concrete by combining these theorems to obtain sufficient conditions on the size of the weight-step, the shadow weights and the scale factors to ensure that both $\boldsymbol{\Phi}_\mathcal{O}$ and $\boldsymbol{\Psi}_\mathcal{O}$ are convergent for all $\mathbf{x} \in \mathbb{X}$:

Theorem 7. *Let $\widetilde{\epsilon}, \epsilon_\phi^{[j]}, \epsilon_\psi \in (0,1)\ \forall j \in \mathbb{N}_D$ and for a given neural network and initial weights $\mathbf{W}_\mathcal{O}$ define:*

$$\epsilon_\psi^{[j]} = \begin{cases} (1-\widetilde{\epsilon}) \dfrac{1-\epsilon_\psi^{[j+1]}}{\frac{H^{[j]}}{H^{[j+1]}} \frac{\phi^{[j]}}{H^{[j]}} \frac{\left\|\mathbf{W}_\mathcal{O}^{[j+1]}\right\|_{2,\infty}^2}{\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}} + (1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j+1]}\right) \frac{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+2}}^{[j+1]}\right\|_{-\infty}^2}{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+2}}^{[j+1]}\right\|_{\infty}^2}} \dfrac{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+2}}^{[j+1]}\right\|_{-\infty}^2}{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+2}}^{[j+1]}\right\|_{\infty}^2} & \textit{if } j < D-1 \\ \epsilon_\psi & \textit{otherwise} \end{cases} \quad \forall j \in \mathbb{N}_D$$

recursively, and subsequently:

$$\begin{aligned} \frac{\phi^{[j]}}{H^{[j]}} &= \overline{\sigma}^{[j]}\left(\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}\right) \\ \frac{\psi^{[j]}}{H^{[j]}} &= \theta\left(1-\epsilon_\psi^{[j]}\right) = \frac{1-\epsilon_\psi^{[j]}}{\epsilon_\psi^{[j]}} \end{aligned} \qquad \forall j \in \mathbb{N}_D$$

$$\left(\text{Equivalently:} \quad \frac{\psi^{[j]}}{H^{[j]}} = (1-\widetilde{\epsilon}) \frac{1-\epsilon_\psi^{[j+1]}}{\frac{H^{[j]}}{H^{[j+1]}} \frac{\phi^{[j]}}{H^{[j]}} \frac{\left\|\mathbf{W}_\mathcal{O}^{[j+1]}\right\|_{2,\infty}^2}{\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}}} \frac{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+2}}^{[j+1]}\right\|_{-\infty}^2}{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+2}}^{[j+1]}\right\|_{\infty}^2} \qquad \forall j \in \mathbb{N}_{D-1}\right)$$

For a weight-step $\mathbf{W}_\Delta$, if the shadow weights satisfy:

$$\begin{aligned} \omega_{i_j}^{[j]2} &\in \left[\left\|\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2, \frac{\psi^{[j-1]}}{H^{[j-1]}}\right] & \forall j \in \mathbb{N}_D, i_j \\ \left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 &\leq \widetilde{\epsilon} \frac{1-\epsilon_\psi^{[j]}}{H^{[j-1]}\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)}{\left\|\mathbf{W}_\mathcal{O}^{[j]}\right\|_{2,\infty}^2}\right)} & \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1} \end{aligned}$$

and the weight-step satisfies:

$$t_{\Delta i_1}^{[0]2} \leq \frac{1-\epsilon_\psi^{[0]}}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \qquad \forall i_1$$

$$t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\|\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2}{\omega_{i_j}^{[j]2}}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2}\right) \leq \frac{1-\epsilon_\psi^{[j]}}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_{-\infty}^2} + 1\right)\frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)}{\left\|\mathbf{W}_\mathcal{O}^{[j]}\right\|_{2,\infty}^2}}$$

$\forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$, and the scale factors satisfy:

$$\mu_{i_1}^{[0]2} \in \left[\frac{1}{1-\epsilon_\psi^{[0]}} t_{\Delta i_1}^{[0]2}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)}\right] \qquad \forall i_1$$

$$\mu_{i_{j+1}}^{[j]2} \in \left[\frac{1}{1-\epsilon_\psi^{[j]}}\left(t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\|\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2}{\omega_{i_j}^{[j]2}}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2}\right)\right), \ldots\right.$$
$$\left.\ldots \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_{-\infty}^2} + 1\right)\frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)}{\left\|\mathbf{W}_\mathcal{O}^{[j]}\right\|_{2,\infty}^2}}\right] \qquad \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$$

then $\left\|\boldsymbol{\Phi}_\mathcal{O}^{[j]}(\mathbf{x})\right\|_F^2 \leq \phi^{[j]}\ \forall \mathbf{x} \in \mathbb{X}$, $j \in \mathbb{N}_D$ and $\left\|\boldsymbol{\Psi}_\mathcal{O}^{[j]}(\mathbf{W}_\Delta)\right\|_F^2 \leq \psi^{[j]}$ for all $j \in \mathbb{N}_D$.

Proof. Combining theorems 5 and 6, we see that for convergence in both $\mathbf{\Phi}_{\mathcal{O}}$, $\mathbf{\Psi}_{\mathcal{O}}$ we require that:

$$\begin{array}{rll} t_{\Delta i_1}^{[0]2} \leq & \frac{\left(1-\epsilon_{\psi}^{[0]}\right)\left(1-\epsilon_{\phi}^{[0]}\right)\sqrt{\rho^{[0]}}}{s^{[0]2}} & \forall i_1 \\ t_{\Delta i_{j+1}}^{[j]2} + \sum_{i_j} \frac{\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right\|_2^2}{\omega_{i_j}^{[j]2}} \left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]2}\right) \leq & \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}{s^{[j]2}+\frac{1}{H^{[j]}}\sum_{i_j}\omega_{i_j}^{[j]2}\left(\frac{W_{\mathcal{O}i_ji_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+1\right)\frac{\phi^{[j-1]}}{H^{[j-1]}}} & \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1} \end{array}$$

Using our condition $\omega_{i_j}^{[j]2} \in \left[\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\|_2^2, \frac{\psi^{[j-1]}}{H^{[j-1]}}\right]$ then it suffices that:

$$\begin{array}{rll} t_{\Delta i_1}^{[0]2} \leq & \frac{\left(1-\epsilon_{\psi}^{[0]}\right)\left(1-\epsilon_{\phi}^{[0]}\right)\sqrt{\rho^{[0]}}}{s^{[0]2}} & \forall i_1 \\ t_{\Delta i_{j+1}}^{[j]2} + H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 + \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 \leq & \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}{s^{[j]2}+\frac{1}{H^{[j]}}\left(\sum_{i_j}\frac{W_{\mathcal{O}i_ji_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+H^{[j-1]}\right)\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}} & \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1} \end{array}$$

We require that this may be satisfied for a sufficiently small weight step, so $\forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$:

$$\begin{aligned} & H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 \leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}{s^{[j]2}+\frac{1}{H^{[j]}}\left(\sum_{i_j}\frac{W_{\mathcal{O}i_ji_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}}+H^{[j-1]}\right)\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}} \\ \Rightarrow\quad & H^{[j-1]} \leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}{s^{[j]2}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2+\frac{1}{H^{[j]}}\left(\frac{1}{\chi_{i_{j+1}}^{[j]}}\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2+H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2\right)\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}} \\ \Rightarrow\quad & H^{[j-1]}s^{[j]2}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 + \frac{H^{[j-1]}}{H^{[j]}}\left(\frac{1}{\chi_{i_{j+1}}^{[j]}}\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2 + H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2\right)\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}} \leq \ldots \\ & \qquad \ldots\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}} \\ \Rightarrow\quad & H^{[j-1]}\left(s^{[j]2}+\frac{H^{[j-1]}}{H^{[j]}}\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}\right)\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 \leq \ldots \\ & \qquad \ldots\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}} - \frac{H^{[j-1]}}{H^{[j]}}\frac{1}{\chi_{i_{j+1}}^{[j]}}\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}} \\ \Rightarrow\quad & \left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 \leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}-\frac{H^{[j-1]}}{H^{[j]}}\frac{1}{\chi_{i_{j+1}}^{[j]}}\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{H^{[j-1]}\left(s^{[j]2}+\frac{H^{[j-1]}}{H^{[j]}}\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}\right)} \\ \Rightarrow \text{(suff.)}\quad & \left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2 \leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}-\frac{H^{[j-1]}}{H^{[j]}}\frac{1}{\chi_{i_{j+1}}^{[j]}}\left\|\mathbf{W}_{\mathcal{O}}^{[j]}\right\|_{2,\infty}^2\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{H^{[j-1]}\left(s^{[j]2}+\frac{H^{[j-1]}}{H^{[j]}}\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}\right)} \end{aligned}$$

where:

$$\chi_{i_{j+1}}^{[j]} = \frac{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_{-\infty}^2}{\left\|\widetilde{\boldsymbol{\omega}}_{:i_{j+1}}^{[j]}\right\|_\infty^2}$$

To be well-defined, we require that the right-hand-side is positive, for which it suffices that:

$$\frac{\psi^{[j-1]}}{H^{[j-1]}} = (1-\widetilde{\epsilon})\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}\frac{H^{[j]}}{H^{[j-1]}}\frac{\chi_{i_{j+1}}^{[j]}}{\left\|\mathbf{W}_{\mathcal{O}}^{[j]}\right\|_{2,\infty}^2}\frac{H^{[j-1]}}{\phi^{[j-1]}}$$

which we may rewrite in terms of $\epsilon_{\psi}^{[j-1]}$:

$$\begin{aligned} \epsilon_{\psi}^{[j-1]} &= \theta^{-1}\left((1-\widetilde{\epsilon})\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}\frac{H^{[j]}}{H^{[j-1]}}\frac{\chi_{i_{j+1}}^{[j]}}{\left\|\mathbf{W}_{\mathcal{O}}^{[j]}\right\|_{2,\infty}^2}\frac{H^{[j-1]}}{\phi^{[j-1]}}\right) \\ &= \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}{\frac{H^{[j-1]}}{H^{[j]}}\frac{\left\|\mathbf{W}_{\mathcal{O}}^{[j]}\right\|_{2,\infty}^2}{\chi_{i_{j+1}}^{[j]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}+(1-\widetilde{\epsilon})\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}} \end{aligned}$$

in which case our sufficient conditions become:

$$\left\| \widetilde{\boldsymbol{\omega}}^{[j]}_{:i_{j+1}} \right\|_\infty^2 \leq \widetilde{\epsilon} \frac{\left(1-\epsilon_\psi^{[j]}\right)}{H^{[j-1]} \left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + (1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right) \frac{\chi^{[j]}_{i_{j+1}}}{\left\|\mathbf{W}^{[j]}_{\mathcal{O}}\right\|_{2,\infty}^2} \right)}$$

and:

$$t^{[0]2}_{\Delta i_1} \leq \frac{\left(1-\epsilon_\psi^{[0]}\right)}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \qquad \forall i_1$$

$$t^{[j]2}_{\Delta i_{j+1}} + \sum_{i_j} \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i_j}} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}}\right) \leq \frac{\left(1-\epsilon_\psi^{[j]}\right)}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \left(\frac{\left\|\mathbf{W}^{[j]}_{\mathcal{O}:i_{j+1}}\right\|_2^2}{H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}^{[j]}_{:i_{j+1}}\right\|_{-\infty}^2} + 1\right) \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)\chi^{[j]}_{i_{j+1}}}{\left\|\mathbf{W}^{[j]}_{\mathcal{O}}\right\|_{2,\infty}^2}}$$

$\forall j \in \mathbb{N}_D \setminus \{0\}, i_{j+1}$. In terms of the scale factors, recalling theorems 5 and 6, we require that (sufficiently):

$$\mu^{[0]2}_{i_1} \in \left[\frac{1}{1-\epsilon_\psi^{[0]}} t^{[0]2}_{\Delta i_1}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \right] \qquad \forall i_1$$

$$\mu^{[j]2}_{i_{j+1}} \in \left[\frac{1}{1-\epsilon_\psi^{[j]}} \left(t^{[j]2}_{\Delta i_{j+1}} + \sum_{i_j} \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i_j}} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}}\right)\right), \ldots \right.$$

$$\left. \ldots \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \left(\frac{\left\|\mathbf{W}^{[j]}_{\mathcal{O}:i_{j+1}}\right\|_2^2}{H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}^{[j]}_{:i_{j+1}}\right\|_{-\infty}^2} + 1\right) \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)\chi^{[j]}_{i_{j+1}}}{\left\|\mathbf{W}^{[j]}_{\mathcal{O}}\right\|_{2,\infty}^2}} \right] \qquad \forall j \in \mathbb{N}_D \setminus \{0\}, i_{j+1}$$

Noting that $\chi^{[j]}_{i_{j+1}} \leq 1$, we can further tighten our constraints to get sufficient conditions:

$$\left\| \widetilde{\boldsymbol{\omega}}^{[j]}_{:i_{j+1}} \right\|_\infty^2 \leq \widetilde{\epsilon} \frac{1-\epsilon_\psi^{[j]}}{H^{[j-1]} \left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)}{\left\|\mathbf{W}^{[j]}_{\mathcal{O}}\right\|_{2,\infty}^2} \right)}$$

$$t^{[0]2}_{\Delta i_1} \leq \frac{1-\epsilon_\psi^{[0]}}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \qquad \forall i_1$$

$$t^{[j]2}_{\Delta i_{j+1}} + \sum_{i_j} \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i_j}} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}}\right) \leq \frac{1-\epsilon_\psi^{[j]}}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \left(\frac{\left\|\mathbf{W}^{[j]}_{\mathcal{O}:i_{j+1}}\right\|_2^2}{H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}^{[j]}_{:i_{j+1}}\right\|_{-\infty}^2} + 1\right) \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)}{\left\|\mathbf{W}^{[j]}_{\mathcal{O}}\right\|_{2,\infty}^2}}$$

$\forall j \in \mathbb{N}_D \setminus \{0\}, i_{j+1}$ and

$$\mu^{[0]2}_{i_1} \in \left[\frac{1}{1-\epsilon_\psi^{[0]}} t^{[0]2}_{\Delta i_1}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \right] \qquad \forall i_1$$

$$\mu^{[j]2}_{i_{j+1}} \in \left[\frac{1}{1-\epsilon_\psi^{[j]}} \left(t^{[j]2}_{\Delta i_{j+1}} + \sum_{i_j} \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i_j}} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}}\right)\right), \ldots \right.$$

$$\left. \ldots \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \left(\frac{\left\|\mathbf{W}^{[j]}_{\mathcal{O}:i_{j+1}}\right\|_2^2}{H^{[j-1]}\left\|\widetilde{\boldsymbol{\omega}}^{[j]}_{:i_{j+1}}\right\|_{-\infty}^2} + 1\right) \frac{(1-\widetilde{\epsilon})\left(1-\epsilon_\psi^{[j]}\right)}{\left\|\mathbf{W}^{[j]}_{\mathcal{O}}\right\|_{2,\infty}^2}} \right] \qquad \forall j \in \mathbb{N}_D \setminus \{0\}, i_{j+1}$$

which completes the proof. ☐

We finish this section by considering a special case that will be important shortly:

Theorem 8. *Let* $\epsilon_{\phi}^{[j]}, \epsilon_{\psi}^{[j]}, \delta \in (0,1)\ \forall j \in \mathbb{N}_D$ *and for a given neural network with initial weights* $\mathbf{W}_{\mathcal{O}}$ *and weight-step* $\mathbf{W}_{\Delta}$*, using the shadow weights:*

$$\begin{array}{rl} \omega_{i_j}^{[j]2} &= \left\| \mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_{\Delta}\right)\right\|_2^2 \\ \widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} &= \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 - W_{\Delta i_j,i_{j+1}}^{[j]2} \text{ if } j>0 \end{array} \qquad \forall j \in \mathbb{N}_D, i_j, i_{j+1}$$

where:

$$\begin{array}{rl} \frac{\phi^{[j]}}{H^{[j]}} &= \overline{\sigma}^{[j]}\left(\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}\right) \\ \frac{\psi^{[j]}}{H^{[j]}} &= \theta\left(1-\epsilon_{\psi}^{[j]}\right) = \frac{1-\epsilon_{\psi}^{[j]}}{\epsilon_{\psi}^{[j]}} \end{array} \qquad \forall j \in \mathbb{N}_D$$

If the weight-step satisfies:

$$\begin{array}{rl} t_{\Delta i_1}^{[0]2} &\leq \frac{\left(1-\epsilon_{\psi}^{[0]}\right)}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_{\phi}^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \qquad \forall i_1 \\ t_{\Delta i_{j+1}}^{[j]2} + \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 &\leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)}{\frac{s^{[j]2}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}} \end{array}$$

$\forall j \in \mathbb{N}_D \backslash \left\{0\right\}, i_{j+1}$ *and the scale factors satisfy:*

$$\begin{array}{rl} \mu_{i_1}^{[0]2} &\in \left[\frac{1}{1-\epsilon_{\psi}^{[0]}} t_{\Delta i_1}^{[0]2}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_{\phi}^{[0]}\right)\sqrt{\rho^{[0]}}}\right)}\right] \qquad \forall i_1 \\ \mu_{i_{j+1}}^{[j]2} &\in \left[\frac{1}{1-\epsilon_{\psi}^{[j]}}\left(t_{\Delta i_{j+1}}^{[j]2} + \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2\right), \ldots \right. \\ & \left. \ldots \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}}\right] \qquad \forall j \in \mathbb{N}_D \backslash \left\{0\right\}, i_{j+1} \end{array}$$

then $\left\|\mathbf{\Phi}_{\mathcal{O}}^{[j]}(\mathbf{x})\right\|_F^2 \leq \phi^{[j]}\ \forall \mathbf{x} \in \mathbb{X}$, $j \in \mathbb{N}_D$ *and* $\left\|\mathbf{\Psi}_{\mathcal{O}}^{[j]}(\mathbf{W}_{\Delta})\right\|_F^2 \leq \psi^{[j]}$ *for all* $j \in \mathbb{N}_D$.

Proof. Our proof is a simple analogue of the proof of theorem refth:convergephipsi with adjustments for our pre-defined scale factors. Combining theorems 5 and 6, we see that for convergence in both $\mathbf{\Phi}_{\mathcal{O}}$, $\mathbf{\Psi}_{\mathcal{O}}$ we require that, using our definitions of the scale factors:

$$\begin{array}{rll} t_{\Delta i_1}^{[0]2} &\leq \frac{\left(1-\epsilon_{\psi}^{[0]}\right)\left(1-\epsilon_{\phi}^{[0]}\right)\sqrt{\rho^{[0]}}}{s^{[0]2}} & \forall i_1 \\ t_{\Delta i_{j+1}}^{[j]2} + \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 &\leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}{s^{[j]2} + \frac{1}{H^{[j]}}\sum_{i_j}\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_{\Delta})\right\|_2^2\left(\frac{W_{\mathcal{O}i_j i_{j+1}}^{[j]2}}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 - W_{\Delta i_j,i_{j+1}}^{[j]2}} + 1\right)\frac{\phi^{[j-1]}}{H^{[j-1]}}} & \forall j \in \mathbb{N}_D \backslash \left\{0\right\}, i_{j+1} \end{array}$$

and so it suffices that:

$$
\begin{aligned}
t_{\Delta i_1}^{[0]2} &\leq \frac{\left(1-\epsilon_\psi^{[0]}\right)}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \quad \forall i_1 \\
t_{\Delta i_{j+1}}^{[j]2} + \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 &\leq \frac{\left(1-\epsilon_\psi^{[j]}\right)}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{w}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_F^2 - \left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}}
\end{aligned}
$$

$\forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$. In terms of the scale factors, recalling theorems 5 and 6, we require that (sufficiently):

$$
\begin{aligned}
\mu_{i_1}^{[0]2} &\in \left[\frac{1}{1-\epsilon_\psi^{[0]}} t_{\Delta i_1}^{[0]2}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)}\right] \quad \forall i_1 \\
\mu_{i_{j+1}}^{[j]2} &\in \left[\frac{1}{1-\epsilon_\psi^{[j]}}\left(t_{\Delta i_{j+1}}^{[j]2} + \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta i_{j+1}}^{[j]}\right\|_2^2\right), \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{w}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_F^2 - \left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}}\right]
\end{aligned}
$$

$\forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$ which completes the proof. □

D. Canonical Scaling

In the next section we will be considering regularised risk minimization problems of the form:

$$
\begin{aligned}
&\mathbf{W}_\Delta^\bullet = \mathop{\mathrm{argmin}}_{\mathbf{W}_\Delta \in \mathbb{W}_\mathcal{O}} R_\lambda\left(\mathbf{W}_\Delta\right) \\
&R_\lambda\left(\mathbf{W}_\Delta\right) = \lambda h\left(\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}\right) + \sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_\mathcal{O}\left(\mathbf{x}^{\{k\}}\right) + \left\langle \mathbf{\Phi}_\mathcal{O}\left(\mathbf{x}^{\{k\}}\right), \mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}}\right)
\end{aligned}
\tag{47}
$$

The question we will be addressing is:

> *For a given neural network with initial weights and biases $\mathbf{W}_\mathcal{O}$, let $\mathbf{W}_\Delta^{\boxast}$ be the back-propagation weight-step (gradient descent with learning rate η) defined by (24), and let $\mathbf{W}_\Delta^\bullet$ be a weight-step solving the regularised risk minimization problem (47). Given the gradient-descent derived weight-step $\mathbf{W}_\Delta^{\boxast}$, is there a selection shadow weights and scaling factors, possibly dependent on $\mathbf{W}_\Delta^{\boxast}$, and regularization parameter λ that would ensure that $\mathbf{W}_\Delta^\bullet = \mathbf{W}_\Delta^{\boxast}$?*

If the answer is yes (which we demonstrate) then we can gain understanding of back-propagation by analysing (47). Now, the solution to (47) must satisfy first-order optimality conditions (we assume differentiability for simplicity here):

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{W}_\Delta}\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\bullet} &= -\frac{1}{\lambda}\frac{1}{h^{(1)}\left(\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}\right)} \ldots \\
&\ldots \frac{\partial}{\partial \mathbf{W}_\Delta}\sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_\mathcal{O}\left(\mathbf{x}^{\{k\}}\right) + \left\langle \mathbf{\Phi}_\mathcal{O}\left(\mathbf{x}^{\{k\}}\right), \mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}}\right)\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\bullet}
\end{aligned}
$$

Now, noting that the derivative of the second term in (47) corresponds to the gradient in back-propagation, if the gradient of the first (regularization) term satisfies:

$$
\frac{\partial}{\partial \mathbf{W}_\Delta}\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^{\boxast}} = \nu \mathbf{W}_\Delta^{\boxast}
$$

for some $\nu \in \mathbb{R}_+$, and:

$$
\lambda = \frac{1}{\eta\nu h^{(1)}\left(\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta^{\boxast}\right)\right\|_{\mathcal{W}_\mathcal{O}}\right)}
$$

then:

$$\mathbf{W}_{\Delta}^{\bullet} = \mathbf{W}_{\Delta}^{\boxast} = -\eta \frac{\partial}{\partial \mathbf{W}_{\Delta}} \sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}^{\{k\}}\right) + \left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}^{\{k\}}\right), \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}\right)\Big|_{\mathbf{W}_{\Delta} = \mathbf{W}_{\Delta}^{\boxast}}$$

Thus the question whether there exists scaling factors, shadow weights and λ such that the regularised risk minimization weight-step corresponds to the gradient-descent weight-step for a specified learning rate η can be answered in the affirmative by proving the existence of canonical scalings, which we define as follows:

Definition 3 (Canonical Scaling)**.** For a given neural network, initial weights $\mathbf{W}_{\mathcal{O}}$ and weight step $\mathbf{W}_{\Delta}^{\boxast}$ generated by back-propagation, we define a *canonical scaling* to be a set of shadow weights and scaling factors for which:

$$\frac{\partial}{\partial \mathbf{W}_{\Delta}} \left\| \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right) \right\|_{\mathcal{W}_{\mathcal{O}}} \Big|_{\mathbf{W}_{\Delta} = \mathbf{W}_{\Delta}^{\boxast}} = \nu \mathbf{W}_{\Delta}^{\boxast}$$

for some $\nu \in \mathbb{R}_+$, and $\left\| \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}^{\boxast}) \right\|_{\mathcal{W}_{\mathcal{O}}} < \infty$, $\left\| \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}) \right\|_{\mathcal{X}_{\mathcal{O}}} < \infty \; \forall \mathbf{x} \in \mathbb{X}$. We call the kernels and norms induced using a canonical scaling *canonical induced kernels* and *canonical induced norms*, respectively.

With regard to existence we have the following theorem that sets out sufficient conditions for the weight-step for a canonical scaling to exist. Defining $\delta \in (0,1)$ and:

$$t_{\Delta i_{j+1}}^{[j]\boxast 2} = \begin{cases} 2 b_{\Delta i_1}^{[0]\boxast 2} + 2 \left\| \mathbf{W}_{\Delta : i_1}^{[0]\boxast} \right\|_2^2 & \text{if } j = 0 \\ 2 b_{\Delta i_{j+1}}^{[j]\boxast 2} + \frac{2-\delta}{1-\delta} \left\| \mathbf{W}_{\Delta : i_{j+1}}^{[j]\boxast} \right\|_2^2 & \text{otherwise} \end{cases} \qquad \forall j \in \mathbb{N}_D, i_{j+1}$$

our central result is as follows:

Theorem 9. *Let $\epsilon_\phi, \epsilon_\psi, \delta, \chi \in (0,1)$ and for a given neural network with initial weights $\mathbf{W}_{\mathcal{O}}$, and let $\mathbf{W}_{\Delta}^{\boxast}$ be the weight-step for this derived from back-propagation. Let:*

$$\begin{aligned} 0 \leq \epsilon_\psi^{[j]} &\leq \begin{cases} 1 - \frac{s^{[D-1]2}}{(1-\epsilon_\phi)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2 & \textit{if } j = D-1 \\ \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_{\Delta}^{[j+1]\boxast} \right\|_F^2}{\left\| \mathbf{t}_{\Delta}^{[j]\boxast} \right\|_\infty^2} & \textit{otherwise} \end{cases} \\ \frac{\psi^{[j]}}{H^{[j]}} &= \theta\left(1 - \epsilon_\psi^{[j]}\right) = \frac{1-\epsilon_\psi^{[j]}}{\epsilon_\psi^{[j]}} \end{aligned}$$

$\forall j \in \mathbb{N}_D$, and:

$$\begin{aligned} 1 \geq \epsilon_\phi^{[j]} &\geq \begin{cases} \epsilon_\phi & \textit{if } j = D-1 \\ 1 - \frac{1}{\sqrt{\rho^{[D-2]}}} \overline{\sigma}^{[D-2]-1} \left(\frac{ \frac{1}{1-\chi} \frac{1-\epsilon_\psi^{[D-1]}}{\left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2} - \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} }{ \frac{1}{H^{[D-1]}} \max\limits_{i_D} \left\{ \frac{ \left\| \mathbf{W}_{\mathcal{O}:i_D}^{[D-1]} \right\|_2^2 }{ \frac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_\infty^2 } + H^{[D-2]} \right\} \frac{ \frac{\psi^{[D-2]}}{H^{[D-2]}} }{ \left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}} } } \right) & \textit{if } j = D-2 \\ 1 - \frac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \left(\frac{ \frac{1}{\left\| \mathbf{t}_{\Delta}^{[j+1]\boxast} \right\|_\infty^2} \left(1 - \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_{\Delta}^{[j+2]\boxast} \right\|_F^2}{\left\| \mathbf{t}_{\Delta}^{[j+1]\boxast} \right\|_\infty^2} \right) - \frac{s^{[j+1]2}}{\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}} }{ \frac{1}{H^{[j+1]}} \max\limits_{i_{j+2}} \left\{ \frac{ \left\| \mathbf{W}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2 }{ \frac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_\infty^2 } + H^{[j]} \right\} \frac{ \frac{\psi^{[j]}}{H^{[j]}} }{ \left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}} } } \right) & \textit{otherwise} \end{cases} \\ \frac{\phi^{[j]}}{H^{[j]}} &= \overline{\sigma}^{[j]} \left(\left(1 - \epsilon_\phi^{[j]}\right) \sqrt{\rho^{[j]}} \right) \end{aligned}$$

$\forall j \in \mathbb{N}_D$. For some $\alpha^{[j]} \in \mathbb{R}_+$ there exists $\varpi^{[j]} \in (0,1) \; \forall j \in \mathbb{N}_D \backslash \{0\}$ such that:[15]

$$\left\| \mathbf{W}_{\Delta}^{[j+1]\boxast} \right\|_F^2 = (1-\delta)\left(1 - \varpi^{[j+1]}\right) \left\| \mathbf{t}_{\Delta}^{[j]\boxast} \right\|_\infty^2$$

[15] Note that $b_{\Delta i_{j+1}}^{[j]\boxast}$ is proportional to $\alpha^{[j]}$, so we can always increase $t_{\Delta i_{j+1}}^{[j]\boxast 2}$ to ensure the condition holds by increasing $\alpha^{[j]}$ sufficiently.

If the weight-step satisfies:

$$\left\|\mathbf{t}_\Delta^{[j]\circledast}\right\|_\infty^2 < \begin{cases} \frac{(1-\chi)\left(1-\epsilon_\psi^{[D-1]}\right)}{\left(\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}\right)} & \textit{if } j = D-1 \\ \frac{\varpi^{[j+1]}\left(1-\epsilon_\psi^{[j]}\right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}\right)} & \textit{otherwise} \end{cases} \quad \forall j \in \mathbb{N}_D$$

then there exists a canonical scaling:

$$\frac{\partial}{\partial \mathbf{W}_\Delta} \left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \nu \mathbf{W}_\Delta^\circledast$$

where:

$$\nu = \frac{4}{\left\|\mathbf{t}_\Delta^{[D-1]\circledast}\right\|_\infty^2}\kappa\left(\frac{1-\epsilon_\psi^{[D-1]}}{1-\chi}\right)$$

and $\left\|\mathbf{\Phi}_\mathcal{O}^{[j]}(\mathbf{x})\right\|_F^2 \leq \phi^{[j]}$ $\forall \mathbf{x} \in \mathbb{X}$, $j \in \mathbb{N}_D$ *and* $\left\|\mathbf{\Psi}_\mathcal{O}^{[j]}(\mathbf{W}_\Delta)\right\|_F^2 \leq \psi^{[j]}$ *for all* $j \in \mathbb{N}_D$.

Proof. We aim to derive a canonical scalings for general, a-priori weight-steps $\mathbf{W}_\Delta^\circledast$. Recall from (40), (42) and (43) that:

$$\frac{\partial}{\partial \mathbf{W}_\Delta^{[j']}} \left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \sum_{i_D}\left[\frac{\partial}{\partial W_{\Delta i'_{j'}, i'_{j'+1}}^{[j']}}\left\|\mathbf{\Psi}_{\mathcal{O}:i_D}^{[D-1]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast}\right]_{i'_{j'}, i'_{j'+1}}$$

$$\frac{\partial}{\partial \mathbf{b}_\Delta^{[j']}} \left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \sum_{i_D}\left[\frac{\partial}{\partial b_{\Delta i'_{j'+1}}^{[j']}}\left\|\mathbf{\Psi}_{\mathcal{O}:i_D}^{[D-1]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast}\right]_{i'_{j'+1}}$$

where, recursively $\forall j \in \mathbb{N}_D$:

$$\frac{\partial}{\partial W_{\Delta i'_{j'}, i'_{j'+1}}^{[j']}}\left\|\mathbf{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \ldots$$

$$\ldots \begin{cases} \frac{1}{\mu_{i_{j+1}}^{[j]2}}\theta^{(1)}\left(\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{\mu_{i_{j+1}}^{[j]2}}\right)\sum_{i_j}\frac{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}+W_{\Delta i_j,i_{j+1}}^{[j]\circledast 2}}{\omega_{i_j}^{[j]2}}\left(\frac{\partial}{\partial W_{\Delta i'_{j'}, i'_{j'+1}}^{[j']}}\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast}\right) & \text{if } j > j' \\ \frac{1}{\mu_{i'_{j+1}}^{[j]2}}\theta^{(1)}\left(\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{\mu_{i_{j+1}}^{[j]2}}\right)\delta_{i_{j+1},i'_{j+1}}2\left(1+\frac{\left\|\mathbf{\Psi}_{\mathcal{O}:i'_j}^{[j-1]}(\mathbf{W}_\Delta^\circledast)\right\|_2^2}{\omega_{i'_j}^{[j]2}}\right)W_{\Delta i'_j,i_{j+1}}^{[j]\circledast} & \text{if } j = j' > 0 \\ \frac{1}{\mu_{i'_1}^{[0]2}}\theta^{(1)}\left(\frac{t_{\Delta i_1}^{[0]\circledast 2}}{\mu_{i_1}^{[0]2}}\right)\delta_{i_1,i'_1}4W_{\Delta i'_0,i_1}^{[0]\circledast} & \text{if } j = j' = 0 \end{cases}$$

$$\frac{\partial}{\partial b_{\Delta i'_{j'+1}}^{[j']}}\left\|\mathbf{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \ldots$$

$$\ldots \begin{cases} \frac{1}{\mu_{i_{j+1}}^{[j]2}}\theta^{(1)}\left(\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{\mu_{i_{j+1}}^{[j]2}}\right)\sum_{i_j}\frac{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}+W_{\Delta i_j,i_{j+1}}^{[j]\circledast 2}}{\omega_{i_j}^{[j]2}}\left(\frac{\partial}{\partial b_{\Delta i'_{j'+1}}^{[j']}}\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta\right)\right\|_2^2\Big|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast}\right) & \text{if } j > j' \\ \frac{1}{\mu_{i'_{j+1}}^{[j]2}}\theta^{(1)}\left(\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{\mu_{i_{j+1}}^{[j]2}}\right)\delta_{i_{j+1},i'_{j+1}}4b_{\Delta i_{j+1}}^{[j]\circledast} & \text{if } j = j' \end{cases}$$

and we have defined:

$$t_{\Delta i_{j+1}}^{[j]\circledast 2} = \begin{cases} \left[2b_{\Delta i_1}^{[0]\circledast 2} + 2\left\|\mathbf{W}_{\Delta:i_1}^{[0]\circledast}\right\|_2^2\right]_{i_1} & \text{if } j=0 \\ \left[2b_{\Delta i_{j+1}}^{[j]\circledast 2} + \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_2^2 + \sum_{i_j}\left(\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]\circledast 2}\right)\frac{\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta^\circledast)\right\|_2^2}{\omega_{i_j}^{[j]2}}\right]_{i_{j+1}} & \text{otherwise} \end{cases}$$

$\forall j \in \mathbb{N}_D, i_{j+1}$, so that:

$$\left\|\mathbf{\Psi}_{\mathcal{O}i_{j+1}}^{[j]}\left(\mathbf{W}_\Delta^\circledast\right)\right\|_2^2 = \theta\left(\frac{1}{\mu_{i_{j+1}}^{[j]}} t_{\Delta i_{j+1}}^{[j]\circledast 2}\right) = \frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{\mu_{i_{j+1}}^{[j]2} - t_{\Delta i_{j+1}}^{[j]\circledast 2}}$$

Note that the recursive fomula for the norm gradient with respect to a specific weight $W_{\Delta i'_{j'},i'_{j'+1}}^{[j']}$ or bias $b_{\Delta i'_{j'+1}}^{[j']}$ precisely mirrors the structure of the network, where the calculation begins at the ouput layer and recurses backwards along all possible paths to the neuron $i'_{j'+1}$ in layer j'. Indeed, if we define the set of all paths from any neuron in the output layer to neural $i'_{j'+1}$ in layer j' as:

$$\mathcal{P}_{i'_{j'+1}}^{[j']} = \left\{\mathtt{i} = (i_D, i_{D-1}, \ldots, i_{j'+1}) : i_{j+1} \in \mathbb{N}_{H^{[j]}} \forall j \in \mathbb{N}_D \backslash \mathbb{N}_{j'+1}, i_{j'+1} = i'_{j'+1}\right\}$$

then we can re-write the norm gradient as a pathwise sum:

$$\frac{\partial}{\partial W_{\Delta i'_{j'},i'_{j'+1}}^{[j']}} \left\|\mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_{\mathcal{O}}}^2\Bigg|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \ldots$$

$$\ldots\left(\sum_{\mathtt{i}\in\mathcal{P}_{i'_{j'+1}}^{[j']}} g_{i_{j'+1}}^{[j']\circledast} \prod_{j''=j'+1}^{D-1} g_{i_{j''+1}}^{[j'']\circledast} h_{i_{j''},i_{j''+1}}^{[j''-1,j'']\circledast}\right)\begin{cases} 2\left(1 + \frac{\left\|\mathbf{\Psi}_{\mathcal{O}:i'_j}^{[j-1]}(\mathbf{W}_\Delta^\circledast)\right\|_2^2}{\omega_{i'_j}^{[j]2}}\right) W_{\Delta i'_j,i_{j+1}}^{[j]\circledast} & \text{if } j'>0 \\ 4W_{\Delta i'_0,i_1}^{[0]\circledast} & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial b_{\Delta i'_{j'+1}}^{[j']}} \left\|\mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_{\mathcal{O}}}^2\Bigg|_{\mathbf{W}_\Delta=\mathbf{W}_\Delta^\circledast} = \left(\sum_{\mathtt{i}\in\mathcal{P}_{i'_{j'+1}}^{[j']}} g_{i_{j'+1}}^{[j']\circledast} \prod_{j''=j'+1}^{D-1} g_{i_{j''+1}}^{[j'']\circledast} h_{i_{j''},i_{j''+1}}^{[j''-1,j'']\circledast}\right) 4b_{\Delta i_{j+1}}^{[j]\circledast}$$

where each term in the sum of paths is the product of neuron costs:

$$g_{i_{j+1}}^{[j]\circledast} = \begin{cases} \frac{1}{\mu_{i_D}^{[D-1]2}}\theta^{(1)}\left(\frac{t_{\Delta i_D}^{[D-1]\circledast 2}}{\mu_{i_D}^{[D-1]2}}\right) & \text{if } j=D-1 \\ \frac{1}{\omega_{i_{j+1}}^{[j+1]2}}\frac{1}{\mu_{i_{j+1}}^{[j]2}}\theta^{(1)}\left(\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{\mu_{i_{j+1}}^{[j]2}}\right) & \text{otherwise} \end{cases}$$

and link costs:

$$h_{i_j,i_{j+1}}^{[j-1,j]\circledast} = \widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} + W_{\Delta i_j,i_{j+1}}^{[j]\circledast 2}$$

for all neurons and links between neurons on that path from the output layer back to neuron $i'_{j'+1}$ in layer j'..

Ignoring for the moment the question of convergence, we see that to obtain a canonical scaling we must use the scaling factors and shadow weights to "even out" the path-dependence of the neuron and link costs and flatten the dependence on the feature-map norm at the terminating neuron i_{j+1} in layer j. To this end it is straightforward to see that, if we select:

$$\begin{aligned} \omega_{i_j}^{[j]2} &= \left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_\Delta^\circledast\right)\right\|_2^2 \\ \widetilde{\omega}_{i_j,i_{j+1}}^{[j]2} &= \frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_2^2 - W_{\Delta i_j,i_{j+1}}^{[j]\circledast 2} \end{aligned} \quad \forall j \in \mathbb{N}_D, i_j, i_{j+1} \tag{48}$$

where $\delta \in (0,1)$ (remember that we are assuming the back-propagation weight-step $\mathbf{W}_\Delta^\circledast$ is calculated a-priori and then attempting to make the regularization-based weight-step $\mathbf{W}^{\bullet}_{\Delta}$ correspond to this) then:

$$\begin{aligned}
\frac{\partial}{\partial W^{[j']}_{\Delta i'_{j'}, i'_{j'+1}}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta} \right) \right\|^2_{\mathcal{W}_{\mathcal{O}}} \Bigg|_{\mathbf{W}_{\Delta} = \mathbf{W}^{\boxast}_{\Delta}} &= \left(\sum_{\mathtt{i} \in \mathcal{P}^{[j']}_{i'_{j'+1}}} g^{[j']\boxast}_{i_{j'+1}} \prod_{j''=j'+1}^{D-1} g^{[j'']\boxast}_{i_{j''+1}} h^{[j''-1,j'']\boxast}_{i_{j''}, i_{j''+1}} \right) 4 W^{[j]\boxast}_{\Delta i'_j, i_{j+1}} \\
\frac{\partial}{\partial b^{[j']}_{\Delta i'_{j'+1}}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta} \right) \right\|^2_{\mathcal{W}_{\mathcal{O}}} \Bigg|_{\mathbf{W}_{\Delta} = \mathbf{W}^{\boxast}_{\Delta}} &= \left(\sum_{\mathtt{i} \in \mathcal{P}^{[j']}_{i'_{j'+1}}} g^{[j']\boxast}_{i_{j'+1}} \prod_{j''=j'+1}^{D-1} g^{[j'']\boxast}_{i_{j''+1}} h^{[j''-1,j'']\boxast}_{i_{j''}, i_{j''+1}} \right) 4 b^{[j]\boxast}_{\Delta i_{j+1}}
\end{aligned}$$

and:

$$\begin{aligned}
g^{[j]\boxast}_{i_{j+1}} &= \begin{cases} \frac{1}{\mu^{[D-1]2}_{i_D}} \theta^{(1)} \left(\frac{t^{[D-1]\boxast 2}_{\Delta i_D}}{\mu^{[D-1]2}_{i_D}} \right) & \text{if } j = D-1 \\ \frac{1}{\left\| \boldsymbol{\Psi}^{[j]}_{\mathcal{O}:i_{j+1}} \left(\mathbf{W}^{\boxast}_{\Delta} \right) \right\|^2_2} \frac{1}{\mu^{[j]2}_{i_{j+1}}} \theta^{(1)} \left(\frac{t^{[j]\boxast 2}_{\Delta i_{j+1}}}{\mu^{[j]2}_{i_{j+1}}} \right) & \text{otherwise} \end{cases} \\
h^{[j-1,j]\boxast}_{i_j, i_{j+1}} &= \frac{1}{1-\delta} \left\| \mathbf{W}^{[j]\boxast}_{\Delta : i_{j+1}} \right\|^2_2
\end{aligned}$$

Using (39), we see that

$$\begin{aligned}
\left\| \boldsymbol{\Psi}^{[j]}_{\mathcal{O} i_{j+1}} \left(\mathbf{W}^{\boxast}_{\Delta} \right) \right\|^2_2 = \theta \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} t^{[j]\boxast 2}_{\Delta i_{j+1}} \right) &= \frac{t^{[j]\boxast 2}_{\Delta i_{j+1}}}{\mu^{[j]2}_{i_{j+1}} - t^{[j]\boxast 2}_{\Delta i_{j+1}}} \\
g^{[j]\boxast}_{i_{j+1}} &= \begin{cases} \frac{1}{\mu^{[D-1]2}_{i_D}} \theta^{(1)} \left(\frac{t^{[D-1]\boxast 2}_{\Delta i_D}}{\mu^{[D-1]2}_{i_D}} \right) & \text{if } j = D-1 \\ \frac{\mu^{[j]2}_{i_{j+1}} - t^{[j]\boxast 2}_{\Delta i_{j+1}}}{t^{[j]\boxast 2}_{\Delta i_{j+1}}} \frac{1}{\mu^{[j]2}_{i_{j+1}}} \theta^{(1)} \left(\frac{t^{[j]\boxast 2}_{\Delta i_{j+1}}}{\mu^{[j]2}_{i_{j+1}}} \right) & \text{otherwise} \end{cases}
\end{aligned}$$

$$t^{[j]\boxast 2}_{\Delta i_{j+1}} = \begin{cases} \left[2 b^{[0]\boxast 2}_{\Delta i_1} + 2 \left\| \mathbf{W}^{[0]\boxast}_{\Delta : i_1} \right\|^2_2 \right]_{i_1} & \text{if } j = 0 \\ \left[2 b^{[j]\boxast 2}_{\Delta i_{j+1}} + \frac{2-\delta}{1-\delta} \left\| \mathbf{W}^{[j]\boxast}_{\Delta : i_{j+1}} \right\|^2_2 \right]_{i_{j+1}} & \text{otherwise} \end{cases} \quad \forall j \in \mathbb{N}_D, i_{j+1}$$

Take as given than the scale factors (48) are used from here on. Having used our scale factors to even out the link costs in the norm gradient, our next task is to use the scale factors to even out the neuron costs and obtain a canonical scaling. Again neglecting (for the moment) the question of convergence, consider the gradients of the weights and biases in the output layer, which take the particularly simple form:

$$\begin{aligned}
\frac{\partial}{\partial W^{[D-1]}_{\Delta i'_{D-1}, i'_D}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta} \right) \right\|^2_{\mathcal{W}_{\mathcal{O}}} \Bigg|_{\mathbf{W}_{\Delta} = \mathbf{W}^{\boxast}_{\Delta}} &= g^{[D-1]\boxast}_{i'_D} 4 W^{[D-1]\boxast}_{\Delta i'_{D-1}, i_D} \\
\frac{\partial}{\partial b^{[D-1]}_{\Delta i'_D}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_{\Delta} \right) \right\|^2_{\mathcal{W}_{\mathcal{O}}} \Bigg|_{\mathbf{W}_{\Delta} = \mathbf{W}^{\boxast}_{\Delta}} &= g^{[D-1]\boxast}_{i'_D} 4 b^{[D-1]\boxast}_{\Delta i_D}
\end{aligned}$$

For canonical scaling, we must have that:

$$\frac{4}{\mu^{[D-1]2}_{i_D}} \theta^{(1)} \left(\frac{t^{[D-1]\boxast 2}_{\Delta i_D}}{\mu^{[D-1]2}_{i_D}} \right) = \frac{4}{t^{[D-1]\boxast 2}_{\Delta i_D}} \kappa \left(\frac{t^{[D-1]\boxast 2}_{\Delta i_D}}{\mu^{[D-1]2}_{i_D}} \right) = \nu$$

where we recall that:

$$\begin{aligned}
\kappa \left(\zeta \right) &= \zeta \theta^{(1)} \left(\zeta \right) = \frac{\zeta}{(1-\zeta)^2} \\
\kappa^{-1} \left(\zeta \right) &= \frac{2\zeta + 1 - \sqrt{4\zeta + 1}}{2\zeta}
\end{aligned}$$

and hence:

$$\mu^{[D-1]2}_{i_D} = \frac{t^{[D-1]\boxast 2}_{\Delta i_D}}{\kappa^{-1} \left(\frac{t^{[D-1]\boxast 2}_{\Delta i_D} \nu}{4} \right)} = \frac{2 \frac{\nu t^{[D-1]\boxast 2}_{\Delta i_D}}{4}}{2 \frac{\nu t^{[D-1]\boxast 2}_{\Delta i_D}}{4} + 1 - \sqrt{4 \frac{\nu t^{[D-1]\boxast 2}_{\Delta i_D}}{4} + 1}} t^{[D-1]\boxast 2}_{\Delta i_D}$$

Letting $\mu_{i_D}^{[D-1]2}$ take this value, consder the gradients of the weights and biases for layer $D-2$, which simplify to:

$$\begin{aligned}
\frac{\partial}{\partial W_{\Delta i'_{D-2}, i'_{D-1}}^{[D-2]}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}}^2 \Bigg|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^\boxast} &= \nu \left\| \mathbf{W}^{[D-1]} \right\|_F^2 g_{i_{D-1}}^{[D-2]\boxast} W_{\Delta i'_{D-2}, i_{D-1}}^{[D-2]\boxast} \\
\frac{\partial}{\partial b_{\Delta i'_{D-1}}^{[D-2]}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}}^2 \Bigg|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^\boxast} &= \nu \left\| \mathbf{W}^{[D-1]} \right\|_F^2 g_{i_{D-1}}^{[D-2]\boxast} b_{\Delta i_{D-1}}^{[D-2]\boxast}
\end{aligned}$$

and we see that, for a canonical scaling, we must have that:

$$\begin{aligned}
& \tfrac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[D-1]\boxast} \right\|_F^2 \frac{\mu_{i_{D-1}}^{[D-2]2} - t_{\Delta i_{D-1}}^{[D-2]2}}{t_{\Delta i_{D-1}}^{[D-2]2}} \frac{1}{\mu_{i_{D-1}}^{[D-2]2}} \frac{1}{\left(1 - \frac{t_{\Delta i_{D-1}}^{[D-2]\boxast 2}}{\mu_{i_{D-1}}^{[D-2]2}} \right)^2} = 1 \\
\Rightarrow \quad & \tfrac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[D-1]\boxast} \right\|_F^2 \frac{\mu_{i_{D-1}}^{[D-2]2} - t_{\Delta i_{D-1}}^{[D-2]2}}{t_{\Delta i_{D-1}}^{[D-2]2}} \frac{\mu_{i_{D-1}}^{[D-2]2}}{\left(\mu_{i_{D-1}}^{[D-2]2} - t_{\Delta i_{D-1}}^{[D-2]\boxast 2} \right)^2} = 1 \\
\Rightarrow \quad & \tfrac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[D-1]\boxast} \right\|_F^2 \mu_{i_{D-1}}^{[D-2]2} = t_{\Delta i_{D-1}}^{[D-2]2} \mu_{i_{D-1}}^{[D-2]2} - t_{\Delta i_{D-1}}^{[D-2]\boxast 4} \\
\Rightarrow \quad & \left(t_{\Delta i_{D-1}}^{[D-2]2} - \tfrac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[D-1]\boxast} \right\|_F^2 \right) \mu_{i_{D-1}}^{[D-2]2} = t_{\Delta i_{D-1}}^{[D-2]\boxast 4} \\
\Rightarrow \quad & \mu_{i_{D-1}}^{[D-2]2} = \frac{t_{\Delta i_{D-1}}^{[D-2]\boxast 2}}{t_{\Delta i_{D-1}}^{[D-2]2} - \frac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[D-1]\boxast} \right\|_F^2} t_{\Delta i_{D-1}}^{[D-2]\boxast 2}
\end{aligned}$$

For layer $D-3$, using the scale factors derived thus far:

$$\begin{aligned}
\frac{\partial}{\partial W_{\Delta i'_{D-3}, i'_{D-2}}^{[D-3]}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}}^2 \Bigg|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^\boxast} &= \nu \left\| \mathbf{W}^{[D-2]} \right\|_F^2 g_{i_{D-2}}^{[D-3]\boxast} W_{\Delta i'_{D-3}, i_{D-2}}^{[D-3]\boxast} \\
\frac{\partial}{\partial b_{\Delta i'_{D-2}}^{[D-3]}} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}}^2 \Bigg|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^\boxast} &= \nu \left\| \mathbf{W}^{[D-2]} \right\|_F^2 g_{i_{D-2}}^{[D-3]\boxast} b_{\Delta i_{D-2}}^{[D-3]\boxast}
\end{aligned}$$

and so on. Working back through all layers, therefore, we find that canonical scaling is attained if we select shadow weights:

$$\begin{aligned}
\omega_{i_j}^{[j]2} &= \left\| \boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]} \left(\mathbf{W}_\Delta^\boxast \right) \right\|_2^2 & \forall j \in \mathbb{N}_D, i_j \\
\widetilde{\omega}_{i_j, i_{j+1}}^{[j]2} &= \tfrac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]\boxast} \right\|_F^2 - W_{\Delta i_j, i_{j+1}}^{[j]\boxast 2} & \forall j \in \mathbb{N}_D \backslash \{0\}, i_j, i_{j+1}
\end{aligned} \tag{49}$$

and scale factors:

$$\begin{aligned}
\mu_{i_D}^{[D-1]2} &= \frac{t_{\Delta i_D}^{[D-1]\boxast 2}}{\kappa^{-1} \left(\frac{t_{\Delta i_D}^{[D-1]\boxast 2} \nu}{4} \right)} = \frac{2 \frac{\nu t_{\Delta i_D}^{[D-1]\boxast 2}}{4}}{2 \frac{\nu t_{\Delta i_D}^{[D-1]\boxast 2}}{4} + 1 - \sqrt{4 \frac{\nu t_{\Delta i_D}^{[D-1]\boxast 2}}{4} + 1}} t_{\Delta i_D}^{[D-1]\boxast 2} & \forall i_D \\
\mu_{i_{j+1}}^{[j]2} &= \frac{t_{\Delta i_{j+1}}^{[j]\boxast 2}}{t_{\Delta i_{j+1}}^{[j]\boxast 2} - \frac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2} t_{\Delta i_{j+1}}^{[j]\boxast 2} & \forall j \in \mathbb{N}_{D-1}, i_{j+1}
\end{aligned} \tag{50}$$

We also obtain our first additional constraint on the weight-step, namely, using the requisit positivity of $\mu_{i_{j+1}}^{[j]2}$:

$$\tfrac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2 \leq \left(1 - \varpi^{[j+1]} \right) \left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2 \tag{51}$$

Finally we must consider the question of convergence given scale factors and shadow weights (49), (50). Recall from theorem 8 that, given scale factors (49), using the definitions therein, if:

$$\begin{aligned}
t_{\Delta i_1}^{[0]2} &\leq \frac{\left(1 - \epsilon_\psi^{[0]} \right)}{\left(\frac{s^{[0]2}}{\left(1 - \epsilon_\phi^{[0]} \right) \sqrt{\rho^{[0]}}} \right)} \quad \forall i_1 \\
t_{\Delta i_{j+1}}^{[j]2} + \tfrac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \right\|_2^2 &\leq \frac{\left(1 - \epsilon_\psi^{[j]} \right)}{\frac{s^{[j]2}}{\left(1 - \epsilon_\phi^{[j]} \right) \sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}} \left(\frac{\left\| \mathbf{w}_{\mathcal{O}:i_{j+1}}^{[j]} \right\|_2^2}{\frac{1}{1-\delta} \left\| \mathbf{w}_{\Delta:i_{j+1}}^{[j]} \right\|_2^2 - \left\| \mathbf{w}_{\Delta:i_{j+1}}^{[j]} \right\|_\infty^2} + H^{[j-1]} \right) \frac{\frac{\psi^{[j-1]}}{H^{[j-1]}} \frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1 - \epsilon_\phi^{[j]} \right) \sqrt{\rho^{[j]}}}}
\end{aligned} \tag{52}$$

$\forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$, which we take as given, and:

$$\begin{array}{l}\mu_{i_1}^{[0]2} \in \left[\frac{1}{1-\epsilon_\psi^{[0]}} t_{\Delta i_1}^{[0]2}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}} \right)} \right] \quad \forall i_1 \\ \mu_{i_{j+1}}^{[j]2} \in \left[\frac{1}{1-\epsilon_\psi^{[j]}} \left(t_{\Delta i_{j+1}}^{[j]2} + \frac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \right\|_2^2 \right), \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}} \left(\frac{\left\| \mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]} \right\|_2^2}{\frac{1}{1-\delta}\left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]} \right\|_\infty^2} + H^{[j-1]} \right) \frac{\frac{\psi^{[j-1]}}{H^{[j-1]}} \frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}} \right]\end{array} \tag{53}$$

$\forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$ then the norms will convergence. We must show that the scale factors for our canonical scaling satisfy the bounds required for convergence. Consider first $\mu_{i_D}^{[D-1]2}$. In this case we require that, using (50):

$$\begin{array}{rl} & \frac{t_{\Delta i_D}^{[D-1]\boxast 2}}{\kappa^{-1}\left(\frac{t_{\Delta i_D}^{[D-1]\boxast 2}\nu}{4} \right)} \geq \frac{1}{1-\epsilon_\psi^{[D-1]}} t_{\Delta i_D}^{[D-1]\boxast 2} \quad \forall i_D \\ \Rightarrow & \frac{\left\| \mathbf{t}_{\Delta i_D}^{[D-1]\boxast} \right\|_\infty^2}{4} \nu \leq \kappa\left(1-\epsilon_\psi^{[D-1]}\right) = \frac{1-\epsilon_\psi^{[D-1]}}{\epsilon_\psi^{[D-1]2}} = \frac{1}{\epsilon_\psi^{[D-1]}} \frac{\psi^{[D-1]}}{H^{[D-1]}} \end{array}$$

and:

$$\begin{array}{rl} & \frac{t_{\Delta i_D}^{[D-1]\boxast 2}}{\kappa^{-1}\left(\frac{t_{\Delta i_D}^{[D-1]\boxast 2}\nu}{4} \right)} \leq \frac{1}{\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} + \frac{1}{H^{[D-1]}} \left(\frac{\left\| \mathbf{W}_{\mathcal{O}:i_D}^{[D-1]} \right\|_2^2}{\frac{1}{1-\delta}\left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_\infty^2} + H^{[D-2]} \right) \frac{\frac{\psi^{[D-2]}}{H^{[D-2]}} \frac{\phi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}} \quad \forall i_D \\ \Rightarrow & \frac{1}{H^{[D-1]}} \left(\frac{\left\| \mathbf{W}_{\mathcal{O}:i_D}^{[D-1]} \right\|_2^2}{\frac{1}{1-\delta}\left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_\infty^2} + H^{[D-2]} \right) \frac{\frac{\psi^{[D-2]}}{H^{[D-2]}} \frac{\phi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \leq \ldots \\ & \qquad \ldots \frac{1}{t_{\Delta i_D}^{[D-1]\boxast 2}} \kappa^{-1}\left(\frac{t_{\Delta i_D}^{[D-1]\boxast 2}\nu}{4} \right) - \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \quad \forall i_D \\ \Rightarrow & \frac{1}{H^{[D-1]}} \left(\frac{\left\| \mathbf{W}_{\mathcal{O}:i_D}^{[D-1]} \right\|_2^2}{\frac{1}{1-\delta}\left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_\infty^2} + H^{[D-2]} \right) \frac{\frac{\psi^{[D-2]}}{H^{[D-2]}} \frac{\phi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \leq \ldots \\ & \qquad \ldots \frac{1}{\left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2} \kappa^{-1}\left(\frac{\left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2 \nu}{4} \right) - \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \quad \forall i_D \end{array}$$

we maximise the right-side of this bound while satisfying the first bound by selecting:

$$\nu = \frac{4}{\left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2} \kappa\left(\frac{1-\epsilon_\psi^{[D-1]}}{1-\chi} \right)$$

so the conditions become (the first condition is the positivity of the right-side of the above, the second enforces the inequality):

$$\begin{array}{l} \epsilon_\psi^{[D-1]} < 1 - (1-\chi) \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \left\| t_{\Delta}^{[D-1]\boxast} \right\|_\infty^2 \\ \frac{\phi^{[D-2]}}{H^{[D-2]}} \leq \frac{\frac{1}{1-\chi} \frac{1-\epsilon_\psi^{[D-1]}}{\left\| \mathbf{t}_{\Delta}^{[D-1]\boxast} \right\|_\infty^2} - \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}}{\frac{1}{H^{[D-1]}} \max\limits_{i_D} \left\{ \frac{\left\| \mathbf{W}_{\mathcal{O}:i_D}^{[D-1]} \right\|_2^2}{\frac{1}{1-\delta}\left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_D}^{[D-1]\boxast} \right\|_\infty^2} + H^{[D-2]} \right\} \frac{\frac{\psi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}} \end{array}$$

or, equivalently:

$$\begin{aligned}
\nu &= \frac{4}{\left\|\mathbf{t}_\Delta^{[D-1]\circledast}\right\|_\infty^2} \kappa \left(\frac{1-\epsilon_\psi^{[D-1]}}{1-\chi}\right) \\
\epsilon_\psi^{[D-1]} &\in \left(0, 1-(1-\chi)\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}\left\|t_\Delta^{[D-1]\circledast}\right\|_\infty^2\right) \\
\epsilon_\phi^{[D-2]} &\in \left(1-\frac{1}{\sqrt{\rho^{[D-2]}}}\overline{\sigma}^{[D-2]-1}\left(\frac{\frac{1}{1-\chi}\frac{1-\epsilon_\psi^{[D-1]}}{\left\|\mathbf{t}_\Delta^{[D-1]\circledast}\right\|_\infty^2} - \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}}{\frac{1}{H^{[D-1]}}\max_{i_D}\left\{\frac{\left\|\mathbf{W}_{\mathcal{O}:i_D}^{[D-1]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_D}^{[D-1]\circledast}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_D}^{[D-1]\circledast}\right\|_\infty^2} + H^{[D-2]}\right\}\frac{\frac{\psi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}}\right), 1\right)
\end{aligned} \tag{54}$$

and we obtain our second additional constraint:

$$\left\|\mathbf{t}_\Delta^{[D-1]\circledast}\right\|_\infty^2 < \frac{(1-\chi)\left(1-\epsilon_\psi^{[D-1]}\right)}{\left(\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}\right)}$$

Next we consider layer $0 < j < D-1$. We require that, to satisfy (53), using (50):

$$\begin{aligned}
&\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{t_{\Delta i_{j+1}}^{[j]\circledast 2} - \frac{1}{1-\delta}\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2} t_{\Delta i_{j+1}}^{[j]\circledast 2} \geq \frac{1}{1-\epsilon_\psi^{[j]}} t_{\Delta i_{j+1}}^{[j]\circledast 2} \quad \forall i_{j+1} \\
\Rightarrow \quad & t_{\Delta i_{j+1}}^{[j]\circledast 2} \geq \frac{1}{1-\epsilon_\psi^{[j]}}\left(t_{\Delta i_{j+1}}^{[j]\circledast 2} - \frac{1}{1-\delta}\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2\right) \quad \forall i_{j+1} \\
\Rightarrow \quad & \left(\frac{1}{1-\epsilon_\psi^{[j]}} - 1\right) t_{\Delta i_{j+1}}^{[j]\circledast 2} \leq \frac{1}{1-\epsilon_\psi^{[j]}}\frac{1}{1-\delta}\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2 \quad \forall i_{j+1} \\
\Rightarrow \quad & \epsilon_\psi^{[j]} \leq \frac{1}{1-\delta}\frac{\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2}{t_{\Delta i_{j+1}}^{[j]\circledast 2}} \quad \forall i_{j+1}
\end{aligned}$$

and it suffices that:

$$\begin{aligned}
&\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2}}{t_{\Delta i_{j+1}}^{[j]\circledast 2} - \frac{1}{1-\delta}\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2} t_{\Delta i_{j+1}}^{[j]\circledast 2} \leq \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}} \quad \forall i_{j+1} \\
\Rightarrow \quad & \frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} + \frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} \leq \frac{t_{\Delta i_{j+1}}^{[j]\circledast 2} - \frac{1}{1-\delta}\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2}{t_{\Delta i_{j+1}}^{[j]\circledast 4}} \quad \forall i_{j+1} \\
\Rightarrow \quad & \frac{\phi^{[j-1]}}{H^{[j-1]}} \leq \frac{\frac{t_{\Delta i_{j+1}}^{[j]\circledast 2} - \frac{1}{1-\delta}\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2}{t_{\Delta i_{j+1}}^{[j]\circledast 4}} - \frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}}{\frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_\infty^2} + H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}} \quad \forall i_{j+1}
\end{aligned}$$

so it suffices to let:[16]

$$\begin{aligned}
\epsilon_\psi^{[j]} &\leq \frac{1}{1-\delta}\frac{\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2}{\left\|\mathbf{t}_\Delta^{[j]\circledast}\right\|_\infty^2} \\
\frac{\phi^{[j-1]}}{H^{[j-1]}} &\leq \frac{\frac{1}{\left\|\mathbf{t}_\Delta^{[j]\circledast}\right\|_\infty^2}\left(1-\frac{1}{1-\delta}\frac{\left\|\mathbf{W}_\Delta^{[j+1]\circledast}\right\|_F^2}{\left\|\mathbf{t}_\Delta^{[j]\circledast}\right\|_\infty^2}\right) - \frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}}{\frac{1}{H^{[j]}}\max_{i_{j+1}}\left\{\frac{\left\|\mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_2^2 - \left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\circledast}\right\|_\infty^2} + H^{[j-1]}\right\}\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}}
\end{aligned}$$

[16]Note that if $0 < a < x_{\min} \leq x$ then $\frac{x-a}{x^2}$ is minimised by setting $x = x_{\min}$.

or, equivalently:

$$\begin{aligned}
\epsilon_\psi^{[j]} &\in \left(0, \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2}{\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2} \right) \\
\epsilon_\phi^{[j-1]} &\in \left(1 - \frac{1}{\sqrt{\rho^{[j-1]}}} \overline{\sigma}^{[j-1]-1} \left(\frac{ \frac{1}{\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2} \left(1 - \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2}{\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2} \right) - \frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} }{ \frac{1}{H^{[j]}} \max\limits_{i_{j+1}} \left\{ \frac{\left\| \mathbf{W}_{\mathcal{O}:i_{j+1}}^{[j]} \right\|_2^2}{\frac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]\boxast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_{j+1}}^{[j]\boxast} \right\|_\infty^2} + H^{[j-1]} \right\} \frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} } \right), 1 \right)
\end{aligned}$$

where positivity compels us to further constrain the weight-step to ensure this is positive. Using (51), positivity requires our third additional constraint:

$$\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2 < \frac{\varpi^{[j+1]} \left(1-\epsilon_\psi^{[j]} \right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} \right)}$$

Finally we consider the case $j = 0$. We require that:

$$\begin{aligned}
&\frac{t_{\Delta i_1}^{[0]\boxast 2}}{t_{\Delta i_1}^{[0]\boxast 2} - \frac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2} t_{\Delta i_1}^{[0]\boxast 2} \geq \frac{1}{1-\epsilon_\psi^{[0]}} t_{\Delta i_1}^{[0]\boxast 2} \quad \forall i_1 \\
\Rightarrow \quad &\epsilon_\psi^{[0]} \leq \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2}{t_{\Delta i_1}^{[0]\boxast 2}} \quad \forall i_1
\end{aligned}$$

and:

$$\begin{aligned}
&\frac{t_{\Delta i_1}^{[0]\boxast 2}}{t_{\Delta i_1}^{[0]\boxast 2} - \frac{1}{1-\delta} \left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2} t_{\Delta i_1}^{[0]\boxast 2} \leq \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}} \right)} \quad \forall i_1 \\
\Rightarrow \quad &\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}} \leq \frac{1 - \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2}{t_{\Delta i_1}^{[0]\boxast 2}}}{t_{\Delta i_1}^{[0]\boxast 2}} \quad \forall i_1 \\
\Rightarrow \quad &\left\| \mathbf{t}_\Delta^{[0]\boxast} \right\|_\infty^2 \leq \frac{1 - \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2}{\left\| \mathbf{t}_\Delta^{[0]\boxast} \right\|_\infty^2}}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}} \right)}
\end{aligned}$$

so it suffices to let:

$$\epsilon_\psi^{[0]} \leq \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2}{\left\| \mathbf{t}_\Delta^{[0]\boxast} \right\|_\infty^2}$$

or, equivalently:

$$\epsilon_\psi^{[0]} \in \left(0, \frac{1}{1-\delta} \frac{\left\| \mathbf{W}_\Delta^{[1]\boxast} \right\|_F^2}{\left\| \mathbf{t}_\Delta^{[0]\boxast} \right\|_\infty^2} \right)$$

and our fourth additional constraint:

$$\left\| \mathbf{t}_\Delta^{[0]\boxast} \right\|_\infty^2 \leq \frac{\varpi^{[1]} \left(1-\epsilon_\psi^{[0]} \right)}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}} \right)}$$

There is one last technicality. Selecting

$$\epsilon_\psi^{[D-1]} \uparrow 1 - (1-\chi) \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \left\| t_\Delta^{[D-1]\boxast} \right\|_\infty^2$$

will result in:

$$\epsilon_\phi^{[D-2]} \uparrow 1$$

which is not allowed as $\epsilon_\phi^{[D-2]} < 1$ (that is, the range of $\epsilon_\phi^{[D-2]}$ becomes the empty set in the limit. Thus we tighten our contraints on $\epsilon_\psi^{[D-1]}$ to:

$$\epsilon_\psi^{[D-1]} \in \left(0, 1 - \frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \left\| t_\Delta^{[D-1]\circledast} \right\|_\infty^2 \right)$$

which completes the proof after some minor refactoring. ☐

From which we arrive at the corollary used in the main body of the paper:

Corollary 10. *Let $\epsilon, \chi \in (0,1)$ and for a given neural network with initial weights $\mathbf{W}_\mathcal{O}$, and let $\mathbf{W}_\Delta^\circledast$ be the weight-step for this derived from back-propagation. Let:*

$$\begin{aligned}
\epsilon_\psi &= \left(1 - \frac{1}{1-\chi} \frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\circledast} \right\|_\infty^2 \right) \\
\epsilon_\phi^{[j]} &= \begin{cases} \epsilon & \text{if } j = D-1 \\ 1 - \frac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \left(\frac{ \frac{1}{1-\chi} \frac{\chi^{D-1-j}\epsilon_\psi \left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}}{\left\| \mathbf{t}_\Delta^{[j+1]\circledast} \right\|_\infty^2} - \frac{\chi^{D-1-j}\epsilon_\psi}{1-\chi^{D-1-j}\epsilon_\psi} s^{[j+1]2} }{ \max_{i_{j+2}} \left\{ \frac{ \frac{1}{H^{[j+1]}} \left\| \mathbf{W}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2 }{ \frac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\circledast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\circledast} \right\|_\infty^2 } + \frac{H^{[j]}}{H^{[j+1]}} \right\} } \right) & \text{otherwise} \end{cases} \quad \forall j \in \mathbb{N}_D
\end{aligned}$$

For some $\alpha^{[j]} \in \mathbb{R}_+$:

$$\left\| \mathbf{W}_\Delta^{[j+1]\circledast} \right\|_F^2 = (1-\delta)\,\chi \left\| \mathbf{t}_\Delta^{[j]\circledast} \right\|_\infty^2$$

If the weight-step satisfies:

$$\left\| \mathbf{t}_\Delta^{[j]\circledast} \right\|_\infty^2 < \frac{(1-\chi)^2 \left(1 - \chi^{D-j-1} \mathbb{1}_{j<D-1} \epsilon_\psi\right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} \right)} \quad \forall j \in \mathbb{N}_D$$

then there exists of a canonical scaling:

$$\frac{\partial}{\partial \mathbf{W}_\Delta} \left\| \mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right) \right\|_{\mathcal{W}_\mathcal{O}} \Big|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^\circledast} = \nu \mathbf{W}_\Delta^\circledast$$

where:

$$\nu = \frac{4}{\left\| \mathbf{t}_\Delta^{[D-1]\circledast} \right\|_\infty^2} \kappa \left(\frac{1-\epsilon_\psi}{1-\chi} \right)$$

satisfying $\|\mathbf{\Phi}_\mathcal{O}(\mathbf{x})\|_F^2 \leq H^{[D-1]}\overline{\sigma}^{[D-1]}(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right)$ $\forall \mathbf{x} \in \mathbb{X}$, $\|\mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\|_F^2 \leq H^{[D-1]} \frac{1-\epsilon_\psi}{\epsilon_\psi}$.

Proof. We begin with theorem 9. Let $\varpi^{[j]} = 1 - \chi$ and $\epsilon_\phi^{[D-1]} = \epsilon_\phi$. The definitions in theorem 9 become:

$$\begin{aligned}
0 \leq \epsilon_\psi^{[j]} &\leq \chi^{D-1-j} \left(1 - \frac{s^{[D-1]2}}{(1-\epsilon_\phi)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\circledast} \right\|_\infty^2 \right) \\
1 \geq \epsilon_\phi^{[j]} &\geq \begin{cases} \epsilon_\phi & \text{if } j = D-1 \\ 1 - \frac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \left(\frac{ \frac{1}{1-\chi} \frac{1-\epsilon_\psi^{[j+1]}}{\left\| \mathbf{t}_\Delta^{[j+1]\circledast} \right\|_\infty^2} - \frac{s^{[j+1]2}}{\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}} }{ \max_{i_{j+2}} \left\{ \frac{ \frac{1}{H^{[j+1]}} \left\| \mathbf{W}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2 }{ \frac{1}{1-\delta} \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\circledast} \right\|_2^2 - \left\| \mathbf{W}_{\Delta:i_{j+2}}^{[j+1]\circledast} \right\|_\infty^2 } + \frac{H^{[j]}}{H^{[j+1]}} \right\} \frac{ \frac{\psi^{[j]}}{H^{[j]}} }{ \left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}} } } \right) & \text{otherwise} \end{cases}
\end{aligned}$$

$\forall j \in \mathbb{N}_D$, and:

$$\begin{aligned}
\frac{\psi^{[j]}}{H^{[j]}} &= \theta\left(1 - \epsilon_\psi^{[j]}\right) = \frac{1-\epsilon_\psi^{[j]}}{\epsilon_\psi^{[j]}} \\
\frac{\phi^{[j]}}{H^{[j]}} &= \overline{\sigma}^{[j]} \left(\left(1 - \epsilon_\phi^{[j]}\right) \sqrt{\rho^{[j]}} \right)
\end{aligned} \quad \forall j \in \mathbb{N}_D$$

and subsequently for some $\alpha^{[j]} \in \mathbb{R}_+$:

$$\left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2 = (1-\delta)\,\chi \left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2$$

The condition on the weight-step satisfies:

$$\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2 < \frac{(1-\chi)\left(1-\epsilon_\psi^{[j]}\right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}\right)} \quad \forall j \in \mathbb{N}_D$$

implying the existence of a canonical scaling:

$$\frac{\partial}{\partial \mathbf{W}_\Delta} \left\| \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_\Delta\right) \right\|_{\mathcal{W}_{\mathcal{O}}} \Big|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^{\boxast}} = \nu \mathbf{W}_\Delta^{\boxast}$$

where:

$$\nu = \frac{4}{\left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2} \kappa \left(\frac{1-\epsilon_\psi^{[D-1]}}{1-\chi} \right)$$

satisfying $\|\mathbf{\Phi}_{\mathcal{O}}(\mathbf{x})\|_F^2 \leq \phi^{[D-1]}$, $\|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_F^2 \leq \psi^{[D-1]}$. To further simplify the constraints, note that:

$$\begin{aligned}
1-\epsilon_\psi^{[j]} &= 1-\chi^{D-1-j} + \chi^{D-1-j} \frac{s^{[D-1]2}}{(1-\epsilon_\phi)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2 > 1-\chi^{D-1-j} \\
\frac{\psi^{[j]}}{H^{[j]}} &= \frac{1-\epsilon_\psi^{[j]}}{\epsilon_\psi^{[j]}} < \frac{1}{\epsilon_\psi^{[j]}} = \frac{1}{\chi^{D-1-j}} \frac{1}{1-\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi\right)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2}
\end{aligned}$$

so we can tighten the constraints on $\epsilon_\phi^{[j]}$ and $\|\mathbf{t}_\Delta^{[j]\boxast}\|_\infty^2$ to:

$$1 \geq \epsilon_\phi^{[j]} \geq \begin{cases} \epsilon_\phi & \text{if } j = D-1 \\ 1 - \frac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \left(\frac{\frac{1}{1-\chi} \frac{\epsilon_\psi^{[j]}\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}}{\left\| \mathbf{t}_\Delta^{[j+1]\boxast} \right\|_\infty^2} - \frac{\epsilon_\psi^{[j]}}{1-\epsilon_\psi^{[j]}} s^{[j+1]2}}{\max\limits_{i_{j+2}} \left\{ \frac{\frac{1}{H^{[j+1]}} \left\| \mathbf{w}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2}{\frac{1}{1-\delta} \left\| \mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_2^2 - \left\| \mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_\infty^2} + \frac{H^{[j]}}{H^{[j+1]}} \right\}} \right) & \text{otherwise} \end{cases}$$

$$\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2 < \frac{(1-\chi)\left(1-\chi^{D-1-j}\left(1-\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi\right)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2\right)\right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}\right)} \quad \forall j \in \mathbb{N}_D$$

Subsequently the definitions in theorem 9 can be tightened to, being careful to ensure that $\left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2$ is not overconstrained:

$$\begin{aligned}
\epsilon_\psi &= \left(1 - \frac{1}{1-\chi} \frac{s^{[D-1]2}}{(1-\epsilon_\phi)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2 \right) \\
\epsilon_\phi^{[j]} &= \begin{cases} \epsilon_\phi & \text{if } j = D-1 \\ 1 - \frac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \left(\frac{\frac{1}{1-\chi} \frac{\chi^{D-1-j}\epsilon_\psi\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}}{\left\| \mathbf{t}_\Delta^{[j+1]\boxast} \right\|_\infty^2} - \frac{\chi^{D-1-j}\epsilon_\psi}{1-\chi^{D-1-j}\epsilon_\psi} s^{[j+1]2}}{\max\limits_{i_{j+2}} \left\{ \frac{\frac{1}{H^{[j+1]}} \left\| \mathbf{w}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2}{\frac{1}{1-\delta} \left\| \mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_2^2 - \left\| \mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_\infty^2} + \frac{H^{[j]}}{H^{[j+1]}} \right\}} \right) & \text{otherwise} \end{cases}
\end{aligned} \quad \forall j \in \mathbb{N}_D$$

and subsequently for some $\alpha^{[j]} \in \mathbb{R}_+$:

$$\left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2 = (1-\delta)\,\chi \left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2$$

The condition on the weight-step satisfies (the first constraint is required to ensure that $\epsilon_\psi > 0$):

$$\left\| \mathbf{t}_{\Delta}^{[j]\boxast} \right\|_{\infty}^{2} < \frac{1-\chi}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} \right)} \left\{ \begin{array}{ll} 1 & \text{if } j = D-1 \\ \left(1 - \chi^{D-1-j}\epsilon_\psi\right) & \text{otherwise} \end{array} \right. \quad \forall j \in \mathbb{N}_D$$

implying the existence of a canonical scaling:

$$\frac{\partial}{\partial \mathbf{W}_\Delta} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}} \bigg|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^{\boxast}} = \nu \mathbf{W}_\Delta^{\boxast}$$

where:

$$\nu = \frac{4}{\left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2} \kappa \left(\frac{1-\epsilon_\psi}{1-\chi} \right)$$

satisfying $\|\boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x})\|_F^2 \leq H^{[D-1]} \overline{\sigma}^{[D-1]}(\left(1-\epsilon_\phi\right)\sqrt{\rho^{[D-1]}}\right)$, $\|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_F^2 \leq H^{[D-1]} \frac{1-\epsilon_\psi}{\epsilon_\psi}$.

At this point we have enough leeway in our construct to let $\delta \to 0$, and tidy up with $\epsilon_\phi = \epsilon$, so our constraints become:

$$\begin{aligned} \epsilon_\psi &= \left(1 - \frac{1}{1-\chi} \frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}} \left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2 \right) \\ \epsilon_\phi^{[j]} &= \left\{ \begin{array}{ll} \epsilon & \text{if } j = D-1 \\ 1 - \frac{1}{\sqrt{\rho^{[j]}}} \overline{\sigma}^{[j]-1} \left(\frac{ \frac{1}{1-\chi} \frac{\chi^{D-1-j}\epsilon_\psi \left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}}{\left\| \mathbf{t}_\Delta^{[j+1]\boxast} \right\|_\infty^2} - \frac{\chi^{D-1-j}\epsilon_\psi}{1-\chi^{D-1-j}\epsilon_\psi} s^{[j+1]2} }{ \max\limits_{i_{j+2}} \left\{ \frac{ \frac{1}{H^{[j+1]}} \left\| \mathbf{w}_{\mathcal{O}:i_{j+2}}^{[j+1]} \right\|_2^2 }{ \frac{1}{1-\delta} \left\| \mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_2^2 - \left\| \mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxast} \right\|_\infty^2 } + \frac{H^{[j]}}{H^{[j+1]}} \right\} } \right) & \text{otherwise} \end{array} \right. \quad \forall j \in \mathbb{N}_D \end{aligned}$$

and subsequently for some $\alpha^{[j]} \in \mathbb{R}_+$:

$$\left\| \mathbf{W}_\Delta^{[j+1]\boxast} \right\|_F^2 = (1-\delta)\,\chi \left\| \mathbf{t}_\Delta^{[j]\boxast} \right\|_\infty^2$$

The condition on the weight-step satisfies (the first constraint is required to ensure that $\epsilon_\psi > 0$):

$$\left\| \mathbf{t}_{\Delta}^{[j]\boxast} \right\|_{\infty}^{2} < \frac{1-\chi}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}} \right)} \left\{ \begin{array}{ll} 1 & \text{if } j = D-1 \\ \left(1 - \chi^{D-1-j}\epsilon_\psi\right) & \text{otherwise} \end{array} \right. \quad \forall j \in \mathbb{N}_D$$

implying the existence of a canonical scaling:

$$\frac{\partial}{\partial \mathbf{W}_\Delta} \left\| \boldsymbol{\Psi}_{\mathcal{O}} \left(\mathbf{W}_\Delta \right) \right\|_{\mathcal{W}_{\mathcal{O}}} \bigg|_{\mathbf{W}_\Delta = \mathbf{W}_\Delta^{\boxast}} = \nu \mathbf{W}_\Delta^{\boxast}$$

where:

$$\nu = \frac{4}{\left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2} \kappa \left(\frac{1-\epsilon_\psi}{1-\chi} \right)$$

satisfying $\|\boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x})\|_F^2 \leq H^{[D-1]} \overline{\sigma}^{[D-1]}(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right)$, $\|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\|_F^2 \leq H^{[D-1]} \frac{1-\epsilon_\psi}{\epsilon_\psi}$.

Finally, we need to ensure that ν is well defined, for which we require that $1-\epsilon_\psi < 1-\chi$ or, equivalently, $\epsilon_\psi > \chi$. For This is suffices that:

$$\left\| \mathbf{t}_\Delta^{[D-1]\boxast} \right\|_\infty^2 < \frac{1-\chi}{\left(\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}} \right)}$$

and so, tightening bounds slightly:

$$\left\|\mathbf{t}_{\Delta}^{[j]\circledast}\right\|_{\infty}^{2} < \frac{1-\chi}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}\right)} \begin{cases} (1-\chi) & \text{if } j = D-1 \\ \left(1-\chi^{D-1-j}\epsilon_{\psi}\right) & \text{otherwise} \end{cases} \quad \forall j \in \mathbb{N}_D$$

which completes the proof. □

Subsequently we obtain a bound on Rademacher complexity:

Theorem 11. *Let $\epsilon, \chi \in (0,1)$ and for a given neural network with initial weights $\mathbf{W}_{\mathcal{O}}$, and let $\mathbf{W}_{\Delta}^{\circledast}$ be the weight-step for this derived from back-propagation satisfying the conditions set out in corollary 10. Then $f_{\Delta}^{\circledast} \in \mathcal{F}^{\bullet}$, where the Rademacher complexity of $\mathcal{F}^{\bullet}$ is bounded as:*

$$\mathcal{R}_N\left(\mathcal{F}^{\bullet}\right) \leq H^{[D-1]} \sqrt{\frac{1}{N} \frac{\left\|\mathbf{t}_{\Delta}^{[D-1]\circledast}\right\|_{\infty}^{2}}{\frac{1}{1-\chi}\frac{(1-\chi)^2}{\left(\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}}\right)} - \left\|\mathbf{t}_{\Delta}^{[D-1]\circledast}\right\|_{\infty}^{2}} \bar{\sigma}^{[D-1]}\left(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right)}$$

Proof. By corollary 10 we know that, subject to assumptions on the size of the weight-step, there must exist scale factors, shadow weighs and regularization parameter λ such that the change in neural network operation $f_{\Delta}^{\circledast}$ due to back-propagation and the change in neural network operation $f_{\Delta}^{\bullet}$ resulting from minimization of the regularised risk R_{λ} will coincide, so $f_{\Delta}^{\circledast} = f_{\Delta}^{\bullet} \in \mathcal{B}_{\mathcal{O}}$. Fixing these parameters, moreover, we know from corollary 10 that $\|\mathbf{\Phi}_{\mathcal{O}}(\mathbf{x})\|_F^2 \leq H^{[D-1]}\bar{\sigma}^{[D-1]}(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right) \; \forall \mathbf{x} \in \mathbb{X}$, $\|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta})\|_F^2 \leq H^{[D-1]}\frac{1-\epsilon_{\psi}}{\epsilon_{\psi}}$, so:

$$f_{\Delta}^{\bullet} \in \mathcal{F}^{\bullet} = \left\{ \left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\cdot\right), \mathbf{\Omega}\right\rangle \middle| \, \|\mathbf{\Omega}\|_F^2 \leq H^{[D-1]}\frac{1-\epsilon_{\psi}}{\epsilon_{\psi}} \right\}$$

Hence for a Rademacher random variable ϵ the Rademacher complexity is bounded as follows:

$$\begin{aligned}
&\mathcal{R}_N\left(\mathcal{F}^{\bullet}\right) = \mathbb{E}_{\nu}\mathbb{E}_{\epsilon}\left[\sup_{f_{\Delta}\in\mathcal{F}^{\bullet}} \left|\tfrac{1}{N}\textstyle\sum_i \epsilon_i f\left(\mathbf{x}_i\right)\right|\right] \\
&= \mathbb{E}_{\nu}\mathbb{E}_{\epsilon}\left[\sup_{f_{\Delta}\in\mathcal{F}^{\bullet}} \left|\tfrac{1}{N}\textstyle\sum_i \epsilon_i \left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}_i\right), \mathbf{\Psi}\left(\mathbf{W}_{\Delta}^{\circledast}\right)\right\rangle\right|\right] \\
&\leq^{C.S.} \mathbb{E}_{\nu}\left[\tfrac{1}{N}\mathbb{E}_{\epsilon}\left[\sup_{f_{\Delta}\in\mathcal{F}^{\bullet}} \left\|\textstyle\sum_i \epsilon_i \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}_i\right)\right\|_F \left\|\mathbf{\Psi}\left(\mathbf{W}_{\Delta}^{\circledast}\right)\right\|_F\right]\right] \\
&\leq \mathbb{E}_{\nu}\left[\sqrt{\tfrac{1}{N}}\sqrt{H^{[D-1]}\tfrac{1-\epsilon_{\psi}}{\epsilon_{\psi}}\mathbb{E}_{\epsilon}\left[\left\|\tfrac{1}{N}\textstyle\sum_i \epsilon_i \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}_i\right)\right\|_F^2\right]}\right] \\
&\leq^{\text{Jensen}} \mathbb{E}_{\nu}\left[\sqrt{\tfrac{1}{N}H^{[D-1]}\tfrac{1-\epsilon_{\psi}}{\epsilon_{\psi}}}\sqrt{\tfrac{1}{N}\textstyle\sum_i \left\|\mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}_i\right)\right\|_F^2}\right] \\
&\leq H^{[D-1]}\sqrt{\tfrac{1}{N}\tfrac{1-\epsilon_{\psi}}{\epsilon_{\psi}}\bar{\sigma}^{[j]}\left(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right)}
\end{aligned}$$

independent of the data distribution ν.

Next recall the definitions from corollary 10:

$$\epsilon_{\psi} = 1 - \frac{1}{1-\chi}\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}}\left\|\mathbf{t}_{\Delta}^{[D-1]\circledast}\right\|_{\infty}^{2}$$

so that:

$$\frac{1-\epsilon_{\psi}}{\epsilon_{\psi}} = \frac{\left\|\mathbf{t}_{\Delta}^{[D-1]\circledast}\right\|_{\infty}^{2}}{\frac{1}{1-\chi}\frac{(1-\chi)^2}{\left(\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}}\right)} - \left\|\mathbf{t}_{\Delta}^{[D-1]\circledast}\right\|_{\infty}^{2}}$$

	Notation in present Paper	Notation used in (Lin et al., 2022)
Data space:	$\mathbb{X} \subset \mathbb{R}^n$	Ω_1 (input space)
Weight-step space:	$\mathbb{W}_{\mathcal{O}} \subset \prod_{j \in \mathbb{N}_D} \mathbb{R}^{H^{[j-1]} \times H^{[j]}} \times \mathbb{R}^{H^{[j]}}$	Ω_2 (weight space)
Data Feature map:	$\boldsymbol{\Phi}_{\mathcal{O}} : \mathbb{X} \to \mathcal{X}_{\mathcal{O}} \subset \mathbb{R}^{\infty \times m}$	$\Phi_1 : \Omega_1 \to \mathcal{W}_1$
Weight-step feature map:	$\boldsymbol{\Psi}_{\mathcal{O}} : \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} \subset \mathbb{R}^{\infty \times m}$	$\Phi_2 : \Omega_2 \to \mathcal{W}_2$
Data Banach space:	$\mathcal{X}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X})) \subset \mathbb{R}^{\infty \times m}$, where $\| \cdot \|_{\mathcal{X}_{\mathcal{O}}} = \| \cdot \|_F$	$\mathcal{W}_1$ with norm $\| \cdot \|_{\mathcal{W}_1}$
Weight-step Banach space:	$\mathcal{W}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})) \subset \mathbb{R}^{\infty \times m}$, where $\| \cdot \|_{\mathcal{W}_{\mathcal{O}}} = \| \cdot \|_F$	$\mathcal{W}_2$ with norm $\| \cdot \|_{\mathcal{W}_2}$
Bilinear form:	$\langle \cdot, \cdot \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} : \mathbb{R}^{\infty \times m} \times \mathbb{R}^{\infty \times m} \to \mathbb{R}^m$, $\langle \boldsymbol{\Omega}, \boldsymbol{\Xi} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \mathrm{diag}\left(\boldsymbol{\Omega}^{\mathrm{T}} \boldsymbol{\Xi} \right)$	$\langle \cdot, \cdot \rangle_{\mathcal{W}_1 \times \mathcal{W}_2} : \mathcal{W}_1 \times \mathcal{W}_2 \to \mathbb{Y}$

Figure 3: Summary of the construction of reproducing kernel Banach space as per (Lin et al., 2022).

and hence:

$$\mathcal{R}_N \left(\mathcal{F}^{\bullet} \right) \leq H^{[D-1]} \sqrt{ \frac{1}{N} \frac{ \left\| \mathbf{t}_{\Delta}^{[D-1]\circledast} \right\|_{\infty}^{2} }{ \frac{1}{1-\chi} \frac{(1-\chi)^2}{\left(\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}} \right)} - \left\| \mathbf{t}_{\Delta}^{[D-1]\circledast} \right\|_{\infty}^{2} } \bar{\sigma}^{[j]} \left((1-\epsilon) \sqrt{\rho^{[D-1]}} \right) }$$

□

E. Neural Networks and Reproducing Kernel Banach Spaces

A reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) is a Hilbert spaces $\mathcal{H}$ of functions $f : \mathbb{X} \to \mathbb{Y}$ for which the point evaluation functions are continuous. Thus, applying the Riesz representor theorem, there exists a kernel K such that:

$$f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$$

for all $f \in \mathcal{H}$. Subsequently $K(\mathbf{x}, \mathbf{x}') = \langle K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot) \rangle$ and, by the Moore-Aronszajn theorem, K is uniquely defined by $\mathcal{H}$ and vice-versa. K is called the reproducing kernel, and the corresponding RKHS is denoted $\mathcal{H}_K$. RKHSs have gained popularity in machine learning because they are well suited to many aspects of machine learning (Steinwart & Christman, 2008; Shawe-Taylor & Cristianini, 2004; Cortes & Vapnik, 1995; Chowdhury & Gopalan, 2017; Cristianini & Shawe-Taylor, 2005; Genton, 2001; Gönen & Alpaydin, 2011; Herbrich, 2002; Li et al., 2017; Müller et al., 2001; Smola & Schölkopf, 1998). The inner product structure enables the kernel trick, which is of great practical use, and the kernel itself is readily understood as a similarity measure. More importantly here, the structure of RKHSs has led to a rich framework of complexity analysis and generalization bounds (Steinwart & Christman, 2008; Shawe-Taylor & Cristianini, 2004) that form a foundation for this branch of machine learning, which more recently has been extended to neural networks through the theory of neural tangent kernels (Jacot et al., 2018).

Reproducing kernel Banach spaces (RKBSs) are a generalizationn of RKHSs which start with a Banach space of functions rather than a Hilbert space (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). Precisely:

Definition 4 (Reproducing kernel Banach space (RKBS, (Lin et al., 2022))). A reproducing kernel Banach space $\mathcal{B}$ on a set $\mathbb{X}$ is a Banach space of functions $f : \mathbb{X} \to \mathbb{Y}$ such that every point evaluation $\delta_{\mathbf{x}} : \mathcal{B} \to \mathbb{Y}$, $\mathbf{x} \in \mathbb{X}$, on $\mathcal{B}$ is continuous. That is, $\exists C_{\mathbf{x}} \in \mathbb{R}_+$ such that:

$$|\delta_{\mathbf{x}}(f)| = |f(\mathbf{x})| \leq C_{\mathbf{x}} \|f\|_{\mathcal{B}}$$

for all $f \in \mathcal{B}$.

This introduces a richer set of geometrical structures and allows for new and exciting extensions to the usual kernel framework, such as asymmetric kernels, sparse learning in Feature space, lasso in statistics and m-kernels. In the present context it allows us to extend NTKs from a first-order approximation in the overparametrised regime to an exact representation without the need for infinite width approximations. There are many approaches to RKBS theory in the literature, but in the present context we find the method of (Lin et al., 2022) most helpful.

In this formulation, reproducing kernels may be constructed from a set of basic ingredients that closely mirror our construction of $\mathbf{f}_\Delta$ in the preceding sections. In particular, as per (Lin et al., 2022), and in light of lemma 4, we can construct RKBS $\mathcal{B}_\mathcal{O}$ containing $\mathbf{f}_\Delta$ using the ingredients defined in Figure 3. This gives us the reproducing kernel Banach space $\mathcal{B}_\mathcal{O}$ on $\mathbb{X}$:

$$\mathcal{B}_\mathcal{O} = \left\{ \left\langle \mathbf{\Phi}_\mathcal{O}(\cdot), \mathbf{\Omega} \right\rangle_{\mathcal{X}_\mathcal{O} \times \mathcal{W}_\mathcal{O}} \middle| \mathbf{\Omega} \in \mathcal{W}_\mathcal{O} \right\}, \text{ where } \left\| \left\langle \mathbf{\Phi}_\mathcal{O}(\cdot), \mathbf{\Omega} \right\rangle_{\mathcal{X}_\mathcal{O} \times \mathcal{W}_\mathcal{O}} \right\|_{\mathcal{B}_\mathcal{O}} = \|\mathbf{\Omega}\|_{\mathcal{W}_\mathcal{O}} \tag{55}$$

the elements of $\mathcal{B}_\mathcal{O}$ include the functions $\mathbf{f}_\Delta : \mathbb{R}^n \to \mathbb{R}^m$ - that is, the change in the neural network behaviour due to a change $\mathbf{W}_\Delta \in \mathbb{W}_\mathcal{O}$ in weights and biases during one iteration, subject to convergence conditions - as well all linear combinations thereof (recalling that $\mathcal{W}_\mathcal{O} = \text{span}(\mathbf{\Psi}_\mathcal{O}(\mathbb{W}_\mathcal{O}))$). This has the reproducing kernel $\mathbf{K}_\mathcal{O} : \mathbb{X} \times \mathbb{W}_\mathcal{O} \to \mathbb{R}^{m \times m}$:

$$\mathbf{K}_\mathcal{O}(\mathbf{x}, \mathbf{W}_\Delta) = \text{diag}\left(\left\langle \mathbf{\Phi}_\mathcal{O}(\mathbf{x}), \mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta) \right\rangle_{\mathcal{X}_\mathcal{O} \times \mathcal{W}_\mathcal{O}} \right) = \text{diag}\left(\mathbf{f}_\Delta(\mathbf{x}) \right) \tag{56}$$

We also obtain the corresponding reproducing kernel Banach space $\mathcal{B}^*_\mathcal{O}$ on $\mathbb{W}_\mathcal{O}$:

$$\mathcal{B}^*_\mathcal{O} = \left\{ \left\langle \mathbf{\Xi}, \mathbf{\Psi}_\mathcal{O}(\cdot) \right\rangle_{\mathcal{X}_\mathcal{O} \times \mathcal{W}_\mathcal{O}} \middle| \mathbf{\Xi} \in \mathcal{X}_\mathcal{O} \right\}, \text{ where } \left\| \left\langle \mathbf{\Xi}, \mathbf{\Psi}_\mathcal{O}(\cdot) \right\rangle_{\mathcal{X}_\mathcal{O} \times \mathcal{W}_\mathcal{O}} \right\|_{\mathcal{B}^*_\mathcal{O}} = \|\mathbf{\Xi}\|_{\mathcal{X}_\mathcal{O}} \tag{57}$$

whose members are the evaluation functionals for $\mathcal{B}_\mathcal{O}$, which has the reproducing kernel $\mathbf{K}^*_\mathcal{O} : \mathbb{W}_\mathcal{O} \times \mathbb{X} \to \mathbb{R}^{m \times m}$ given by:

$$\mathbf{K}^*_\mathcal{O}(\mathbf{W}_\Delta, \mathbf{x}) = \mathbf{K}_\mathcal{O}(\mathbf{x}, \mathbf{W}_\Delta) = \text{diag}\left(\mathbf{f}_\Delta(\mathbf{x}) \right)$$

F. Case study - The tanh Activation Function

Let us consider the activation function:

$$\tau^{[j]}(\zeta) = \tanh(\zeta) \tag{58}$$

We wish to construct the corresponding $\sigma^{[j]}_{z,z'}$ for arbitrary $z, z' \in \mathbb{R}$. As we shown in section F.1, the Taylor expansion of $\tanh$ about an arbitrary $z \in \mathbb{R}$ is:

$$\tanh(z + \zeta) = \tanh(z) + \sum_{m=1}^{\infty} a^{\tanh}_{(z)m} \zeta^m$$

where:

$$a^{\tanh}_{(z)m} = 2\left(-\text{sgn}(z)\right)^m \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{z^2 + \left(\left(n+\frac{1}{2}\right)\pi\right)^2}} \right)^{m+1} \cos\left((m+1)\,\text{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{z} \right) \right)$$

which is valid for all $z, \zeta \in \mathbb{R}$ such that $|\zeta| \leq \frac{\pi}{2}$. Hence, by our definition (30), using this expansion:

$$\sigma^{[j]}_{(z,z')}(\zeta) = \sum_{m=1}^{\infty} a^{\tanh}_{(z)m} a^{\tanh}_{(z')m} \zeta^m \tag{59}$$

In the case $z = z' = 0$ this simplifies to:

$$a^{\tanh}_{(0)m} = \frac{2^{2m}\left(2^{2m}-1\right)B_{2m}}{(2m)!}$$

where B_{2m} are the Bernoulli numbers. So:

$$\sigma^{[j]}_{(0,0)}(\zeta) = \sum_{m=1}^{\infty} \left(\frac{2^{2m}\left(2^{2m}-1\right)B_{2m}}{(2m)!} \right)^2 \zeta^m$$

F.1. The Taylor Expansions of $\tanh$ About an Arbitrary Point

The Taylor expansion of $\tanh$ about 0 is well-known, specifically:

$$\tanh(\xi) = \sum_{m=1}^{\infty} \frac{2^{2m}\left(2^{2m}-1\right)B_{2m}}{(2m)!} \xi^{2m-1}$$

where B_{2n} is the Bernoulli number, which is valid for $|\xi| \leq \frac{\pi}{2}$. However in this paper we require the Taylor expansion of $\tanh$ about an *arbitrary* point z - however we have been unable to find such an expansion in the literature. In this section we obtain such an expansion using a method inspired by the approach given in the post (D'Aurizio, 2014).

We know that the Weierstrass product expansion of $\cosh$ is:

$$\begin{aligned}\cosh(z+\xi) &= \prod_{n=0}^{\infty}\left(1+\frac{(z+\xi)^2}{\left(n+\frac{1}{2}\right)^2\pi^2}\right)\\ &= \prod_{n=0}^{\infty}\left(1+\frac{(|z|+\mathrm{sgn}(z)\xi)^2}{\left(n+\frac{1}{2}\right)^2\pi^2}\right)\end{aligned}$$

for all $z, \xi \in \mathbb{R}$ (the "sign absorption" here will be important later in our derivation). Hence:

$$\log(\cosh(z+\xi)) = \sum_{n=0}^{\infty}\log\left(1+\frac{(|z|+\mathrm{sgn}(z)\xi)^2}{\left(n+\frac{1}{2}\right)^2\pi^2}\right)$$

and so, taking the derivative $\frac{\partial}{\partial\xi}$ on both sides:

$$\begin{aligned}\mathrm{sgn}(z)\tanh(z+\xi) &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{\frac{1}{\left(n+\frac{1}{2}\right)^2\pi^2}}{1+\frac{(|z|+\mathrm{sgn}(z)\xi)^2}{\left(n+\frac{1}{2}\right)^2\pi^2}}\\ &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)^2\pi^2+z^2+2|z|\,\mathrm{sgn}(z)\xi+\xi^2}\\ &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{1}{|z|+\mathtt{i}\left(n+\frac{1}{2}\right)\pi+\mathrm{sgn}(z)\xi}\,\frac{1}{|z|-\mathtt{i}\left(n+\frac{1}{2}\right)\pi+\mathrm{sgn}(z)\xi}\\ &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{1}{|z|+\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\,\frac{1}{|z|-\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\,\frac{1}{1+\frac{1}{|z|+\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\mathrm{sgn}(z)\xi}\,\frac{1}{1+\frac{1}{|z|-\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\mathrm{sgn}(z)\xi}\end{aligned}$$

where $\mathtt{i}=\sqrt{-1}$. Thus, for all $\xi^2 \leq z^2+(\frac{\pi}{2})^2$:

$$\begin{aligned}\mathrm{sgn}(z)\tanh(z+\xi) &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{1}{|z|+\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\,\frac{1}{|z|-\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\sum_{p,q=0}^{\infty}(-\mathrm{sgn}(z))^{p+q}\ldots\\ &\qquad\ldots\left(\frac{1}{|z|+\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\right)^p\left(\frac{1}{|z|-\mathtt{i}\left(n+\frac{1}{2}\right)\pi}\right)^q\xi^{p+q}\\ &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\sum_{p,q=0}^{\infty}(-\mathrm{sgn}(z))^{p+q}\ldots\\ &\qquad\ldots\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{p+q}{2}}e^{\mathtt{i}(p-q)\mathrm{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)}\xi^{p+q}\\ &= 2(|z|+\mathrm{sgn}(z)\xi)\sum_{n=0}^{\infty}\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\sum_{p,q=0}^{\infty}(-\mathrm{sgn}(z))^{p+q}\ldots\\ &\qquad\ldots\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{p+q}{2}}\cos\left((p-q)\,\mathrm{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\xi^{p+q}\end{aligned}$$

and so, re-arranging our indexing:

$$\begin{aligned}
\operatorname{sgn}(z)\tanh(z+\xi) &= 2\left(|z|+\operatorname{sgn}(z)\,\xi\right)\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\ldots \\
&\qquad \ldots\left(\sum_{n=0}^{\infty}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m}{2}+1}\left(\sum_{j=-m,-m+2,\ldots,m-2,m}\cos\left(j\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\right)\xi^m \\
&= 2\left(|z|+\operatorname{sgn}(z)\,\xi\right)\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\ldots \\
&\qquad \ldots\left(\sum_{n=0}^{\infty}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m}{2}+1}\left(\begin{cases} 2\sum\limits_{j=0,2,\ldots,m}\cos\left(j\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)-1 & \text{if } m \text{ even} \\ 2\sum\limits_{j=1,3,\ldots,m}\cos\left(j\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) & \text{if } m \text{ odd}\end{cases}\right)\right)\xi^m \\
&= 2\left(|z|+\operatorname{sgn}(z)\,\xi\right)\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\ldots \\
&\qquad \ldots\left(\sum_{n=0}^{\infty}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m}{2}+1}\left(\begin{cases} 2\sum\limits_{j=0}^{\frac{m}{2}}\cos\left(2j\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)-1 & \text{if } m \text{ even} \\ 2\sum\limits_{j=0}^{\frac{m-1}{2}}\cos\left((2j+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) & \text{if } m \text{ odd}\end{cases}\right)\right)\xi^m
\end{aligned}$$

Using the result (Knapp, 2009):

$$\sum_{j=0}^{N-1}\cos(a+jd) = \frac{\sin\left(\frac{Nd}{2}\right)}{\sin\left(\frac{d}{2}\right)}\cos\left(a+\frac{(N-1)d}{2}\right)$$

and the usual results for (co)sines of sums and differences, we have that:

$$\begin{aligned}
\sum_{j=0}^{\frac{m}{2}}\cos\left(2j\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) &= \frac{\sin\left(\left(\frac{m}{2}+1\right)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)}{\sin\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)}\cos\left(\frac{m}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&= \frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left(\frac{m+2}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\cos\left(\frac{m}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&= \frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left(\frac{m+1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\cos\left(\frac{m+1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&\quad +\frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left(\frac{1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\cos\left(\frac{1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&= \frac{1}{2}\frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\ldots \\
&\qquad \ldots+\frac{1}{2}\frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&= \frac{1}{2}\frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)+\frac{1}{2}
\end{aligned}$$

and:

$$\begin{aligned}
\sum_{j=0}^{\frac{m-1}{2}}\cos\left((2j+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) &= \frac{\sin\left(\left(\frac{m-1}{2}+1\right)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)}{\sin\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)}\cos\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)+\frac{m-1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&= \frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left(\frac{m+1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\cos\left(\frac{m+1}{2}\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&= \frac{1}{2}\frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)
\end{aligned}$$

so that:

$$\begin{aligned}
\operatorname{sgn}(z)\tanh(z+\xi) &= 2\left(|z|+\operatorname{sgn}(z)\,\xi\right)\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\ldots \\
&\quad \ldots\left(\sum_{n=0}^{\infty}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m}{2}+1}\frac{\sqrt{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}}{\left(n+\frac{1}{2}\right)\pi}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&= 2\left(|z|+\operatorname{sgn}(z)\,\xi\right)\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&= 2|z|\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&\quad +2\sum_{m=0}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m+1}\left(-\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+2}{2}}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{-\frac{1}{2}}\ldots\right. \\
&\quad \left.\ldots\sin\left((m+2)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)-\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^{m+1} \\
&= 2|z|\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{1}{2}}\sin\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&\quad +2|z|\sum_{m=1}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&\quad +2\sum_{m=1}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(-\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{-\frac{1}{2}}\ldots\right. \\
&\quad \left.\ldots\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)-\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^{m} \\
&= 2|z|\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{1}{2}}\sin\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right) \\
&\quad +2|z|\sum_{m=1}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&\quad -2\sum_{m=1}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\ldots\right. \\
&\quad \left.\ldots\sin\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{-\frac{1}{2}}\cos\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&\quad +2\sum_{m=1}^{\infty}\left(-\operatorname{sgn}(z)\right)^{m}\left(\sum_{n=0}^{\infty}\frac{1}{\left(n+\frac{1}{2}\right)\pi}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\ldots\right. \\
&\quad \left.\ldots\cos\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{-\frac{1}{2}}\sin\left(\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\
&= \text{continued next page}\ldots
\end{aligned}$$

$$\begin{aligned}
\ldots\text{from previous page} &= 2\left|z\right| \sum_{n=0}^{\infty} \tfrac{1}{\left(n+\frac{1}{2}\right)\pi} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{1}{2}} \sin\left(\operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right) \\
&\quad +2\left|z\right| \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \tfrac{1}{\left(n+\frac{1}{2}\right)\pi} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \sin\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right)\right) \xi^m \\
&\quad -2 \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \tfrac{1}{\left(n+\frac{1}{2}\right)\pi} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \ldots \right. \\
&\qquad\qquad \left. \ldots \sin\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right) \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{-\frac{1}{2}} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{1}{2}} |z| \right) \xi^m \\
&\quad +2 \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \tfrac{1}{\left(n+\frac{1}{2}\right)\pi} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \ldots \right. \\
&\qquad\qquad \left. \ldots \cos\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{z} \right)\right) \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{-\frac{1}{2}} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{1}{2}} \left(n+\tfrac{1}{2}\right)\pi \right) \xi^m \\
&= 2\left|z\right| \sum_{n=0}^{\infty} \tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \\
&\quad +2\left|z\right| \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \tfrac{1}{\left(n+\frac{1}{2}\right)\pi} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \sin\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right)\right) \xi^m \\
&\quad -2\left|z\right| \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \tfrac{1}{\left(n+\frac{1}{2}\right)\pi} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \sin\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right)\right) \xi^m \\
&\quad +2 \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \cos\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right)\right) \xi^m \\
&= 2\left|z\right| \sum_{n=0}^{\infty} \tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} + 2 \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m \left(\sum_{n=0}^{\infty} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \cos\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right)\right) \xi^m
\end{aligned}$$

and so:

$$\begin{aligned}
\tanh\left(z+\xi\right) = 2z \sum_{n=0}^{\infty} \tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} + 2\operatorname{sgn}\left(z\right) \sum_{m=1}^{\infty} \left(-\operatorname{sgn}\left(z\right)\right)^m & \left(\sum_{n=0}^{\infty} \left(\tfrac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \ldots \right. \\
& \left. \ldots \cos\left(\left(m+1\right) \operatorname{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|} \right)\right)\right) \xi^m
\end{aligned}$$

(As a quick sanity check, let $z \to 0^{\pm}$, so $\operatorname{atan}(\frac{(n+\frac{1}{2})\pi}{|z|}) \to \frac{\pi}{2}$ and $\operatorname{sgn}(z) = \pm 1$:

$$\begin{aligned}
\tanh\left(\xi\right) &= \pm 2 \sum_{m=1}^{\infty} \left(\mp 1\right)^m \left(\sum_{n=0}^{\infty} \left(\tfrac{1}{\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \cos\left(\left(m+1\right) \tfrac{\pi}{2} \right)\right) \xi^m \\
&= -2 \sum_{m \text{ odd}} \left(\sum_{n=0}^{\infty} \left(\tfrac{1}{\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{m+1}{2}} \cos\left(\left(m+1\right) \tfrac{\pi}{2} \right)\right) \xi^m \\
&= -2 \sum_{m=0}^{\infty} \left(\sum_{n=0}^{\infty} \left(\tfrac{1}{\left(n+\frac{1}{2}\right)^2\pi^2} \right)^{\frac{2m+2}{2}} \cos\left(\left(m+1\right) \pi \right)\right) \xi^{2m+1} \\
&= 2 \sum_{m=0}^{\infty} \tfrac{\left(-1\right)^m 2^{2m+2}}{\pi^{2m+2}} \left(\sum_{n=0}^{\infty} \tfrac{1}{\left(2n+1\right)^{2m+2}} \right) \xi^{2m+1} \\
&= 2 \sum_{m=0}^{\infty} \tfrac{\left(-1\right)^m 2^{2m+2}}{\pi^{2m+2}} \left(\sum_{n=1}^{\infty} \tfrac{1}{n^{2m+2}} - \sum_{n=1}^{\infty} \tfrac{1}{\left(2n\right)^{2m+2}} \right) \xi^{2m+1} \\
&= 2 \sum_{m=0}^{\infty} \tfrac{\left(-1\right)^m 2^{2m+2}}{\pi^{2m+2}} \left(1 - \tfrac{1}{4^{m+1}}\right) \zeta\left(2\left(m+1\right)\right) \xi^{2m+1} \\
&= \sum_{m=1}^{\infty} \tfrac{2^{2m}\left(2^{2m}-1\right)B_{2m}}{\left(2m\right)!} \xi^{2m-1}
\end{aligned}$$

where ζ is the Reimann-zeta function and B_{2n} the Bernoulli number, which is the usual Taylor expansion of $\tanh$ about 0). Continuing, using (Gradshteyn & Ryzhik, 2000, (1.421.2)):

$$\begin{aligned}\mathrm{sgn}\left(z\right)\tanh\left(z+\xi\right) &= \tfrac{8|z|}{\pi^2}\sum_{n=0}^{\infty}\tfrac{1}{\left(\frac{2z}{\pi}\right)^2+(2n+1)^2}+\ldots \\ &\quad\ldots+2\sum_{m=1}^{\infty}\left(\left(-\,\mathrm{sgn}\left(z\right)\right)^m\left(\tfrac{2}{\pi}\right)^{m+1}\sum_{n=0}^{\infty}\left(\tfrac{1}{\left(\frac{2z}{\pi}\right)^2+(2n+1)^2}\right)^{\frac{m+1}{2}}\cos\left(\left(m+1\right)\mathrm{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\ &= \tfrac{8|z|}{\pi^2}\sum_{n=1}^{\infty}\tfrac{1}{(2n-1)^2+\left(\frac{2z}{\pi}\right)^2}+2\sum_{m=1}^{\infty}\left(\left(-\,\mathrm{sgn}\left(z\right)\right)^m\sum_{n=0}^{\infty}\left(\tfrac{1}{z^2+\left(\left(n+\frac{1}{2}\right)\pi\right)^2}\right)^{\frac{m+1}{2}}\cos\left(\left(m+1\right)\mathrm{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m \\ &= \tanh\left(|z|\right)+2\sum_{m=1}^{\infty}\left(\left(-\,\mathrm{sgn}\left(z\right)\right)^m\sum_{n=0}^{\infty}\left(\tfrac{1}{\sqrt{z^2+\left(\left(n+\frac{1}{2}\right)\pi\right)^2}}\right)^{m+1}\cos\left(\left(m+1\right)\mathrm{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m\end{aligned}$$

and hence, using the fact that $\tanh$ is an odd function:

$$\tanh\left(z+\xi\right)=\tanh\left(z\right)+\textstyle\sum_{m=1}^{\infty}a^{\tanh}_{(z)m}\xi^m \tag{60}$$

where, using that $\cos$ and atan are odd:

$$a^{\tanh}_{(z)m}=2\left(\left(-\,\mathrm{sgn}\left(z\right)\right)^m\sum_{n=0}^{\infty}\left(\tfrac{1}{\sqrt{z^2+\left(\left(n+\frac{1}{2}\right)\pi\right)^2}}\right)^{m+1}\cos\left(\left(m+1\right)\mathrm{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{z}\right)\right)\right) \tag{61}$$

Thus we can write the helper function (59) for the tanh neuron as:

$$\sigma^{[j]}_{(z,z')}\left(\zeta\right)=\sum_{m=1}^{\infty}a^{\tanh}_{(z,z')m}\zeta^m \tag{62}$$

where:

$$\begin{aligned}a^{\tanh}_{(z,z')m}=4\,\mathrm{sgn}^m\left(zz'\right)\sum_{n,n'=0}^{\infty}\left(\tfrac{1}{\sqrt{z^2+\left(\left(n+\frac{1}{2}\right)\pi\right)^2}}\tfrac{1}{\sqrt{z'^2+\left(\left(n'+\frac{1}{2}\right)\pi\right)^2}}\right)^{m+1}\ldots \\ \ldots\cos\left(\left(m+1\right)\mathrm{atan}\left(\tfrac{\left(n+\frac{1}{2}\right)\pi}{z}\right)\right)\cos\left(\left(m+1\right)\mathrm{atan}\left(\tfrac{\left(n'+\frac{1}{2}\right)\pi}{z'}\right)\right)\end{aligned} \tag{63}$$