神经网络中的梯度下降：作为再生核巴拿赫空间中的序贯学习

Alistair Shilton、Sunil Gupta、Santu Rana、Svetha Venkatesh{
alistair.shilton、sunil.gupta、santu.rana、svetha.venkatesh}@deakin.edu.au

2023年2月2日

arXiv:2302.00205v1 [stat.ML] 1 Feb 2023

摘要

神经切核（NTKs）的研究，通过在初始值附近对权重进行一阶泰勒展开来逼近网络，为理解过参数化（宽网络）极限下神经网络的收敛与泛化特性提供了亟需的洞见。这使得我们可以从再生核希尔伯特空间（RKHS）的角度分析神经网络训练；这种分析对于过参数化区域颇具启发性，但对于较窄的网络而言，由于训练过程中权重变化更大，它只能提供较差的逼近。我们的目标是突破神经切线核的局限，建立一种更一般的理论。我们构造了神经网络在初始权重有限邻域内的精确幂级数表示，将来自数据和权重步长空间的两个特征映射分别映射到特征空间，并通过二者的内积表示网络，从而能够从再生核巴拿赫空间（RKBS）的角度分析神经网络训练。我们证明，无论网络宽度如何，梯度下降生成的训练序列都可以通过再生核巴拿赫空间中的正则化序贯学习精确复现。基于此，我们提出了一种关于一致收敛的新界，其中迭代次数和学习率起着核心作用，为神经网络训练提供了新的理论见解。

1 引言

近几十年来，神经网络取得了显著进展，促使其在广泛的应用领域中迅速普及。然而，这种广泛的成功也留下了一些悬而未决的问题，其中最显而易见的是：为什么非凸、过度参数化的大型网络能够取得远超传统机器学习理论预测的表现?

神经切线核试图回答这个问题。根据 [Jacot 等人，2018，Arora 等人，2019b]，，在训练过程中，过度参数化神经网络的演化遵循相对于神经切线核（NTK）的泛函代价核梯度。研究表明，对于采用随机权重初始化且宽度足够大的网络，NTK 实际上是固定的，因此可以将再生核希尔伯特空间（RKHS）中的机器学习结果应用于这一问题。由此产生了大量分析神经网络收敛性[Du等人，2019年b，Allen-Zhu等人，2019年，Du等人，2019年a，Zou等人，2020年，Zou和Gu， 2019] 以及泛化能力 [Arora 等人，2019b,a，Cao和Gu， 2019] 的研究成果。

尽管取得了成功，神经切线核模型仍存在一些问题。正如 [Bai和Lee，2019], 所指出的，神经切线核所采用的线性近似，其表达能力仅限于相应的随机特征空间或再生核希尔伯特空间的表达能力，神经切线核预测与实际表现之间观察到的差距便是明证。为摆脱这一局限， [Bai和Lee，2019] 提出采用网络的二阶或更高阶近似。此外，一个自然的问题是：在较小权重步长的假设下构建的行为线性近似，在较窄的网络中扩展到较大权重步长时，效果会如何。

为克服这些困难，我们以神经网络在初始权重附近有限邻域内的精确幂级数表示，替代神经切线核中使用的泰勒近似。我们证明，这会使神经网络表示为两个特征映射之间的内积，二者分别来自数据空间和权重步长空间。这种结构构成了再生核巴拿赫空间（RKBS， [Lin等， 2022]）的构造基础。我们进而证明，反向传播与再生核巴拿赫空间中的序贯学习等价；这一结果与神经切线核类似，但不受线性性的约束，并使我们能够为任意宽度的网络推导出新的统一收敛界。

2 相关工作

已有大量研究探讨了采用不同训练假设的各类网络的一致收敛行为 [Neyshabur 等，2015、2018、2019、2017，Harvey 等，2017，Bartlett 等，2017，Golowich 等，2018，Arora 等，2018，Allen-Zhu 等，2018，Dr¨axler 等，2018，Li 和 Liang，2018，Nagarajan 和 Kolter，2019a、b，Zhou 等， 2019]。

关于核方法与神经网络之间联系的研究由来已久。 [Neal 1996]证明，在无限宽度极限下，独立同分布随机初始化的单层网络会收敛到高斯过程的样本。这一结果被扩展至多层神经网络： [Lee 等，2018、Matthews 等人， 2018] 假设除输出层之外的各层权重均为随机的。其他通过假设权重随机来推导近似核的研究包括 [Rahimi 和 Benjamin，2009，Bach，2014、2017，Daniely 等，2016，Daniely， 2017]。

	本文中的记号	Lin 等人 2022]使用的记号
数据空间:	$\mathbb{X} \subset \mathbb{R}^n$	Ω_1 （输入空间）
权重步长空间:	$\mathbb{W}_{\mathcal{O}} \subset \prod_{j\in\mathbb{N}_D} \mathbb{R}^{H^{[j-1]}\times H^{[j]}} \times \mathbb{R}^{H^{[j]}}$	Ω_2 （权重空间）
数据特征映射:	$\boldsymbol{\Phi}_{\mathcal{O}} : \mathbb{X} \to \mathcal{X}_{\mathcal{O}} \subset \mathbb{R}^{\infty\times m}$	$\Phi_1 : \Omega_1 \to \mathcal{W}_1$
权重步长特征映射:	$\boldsymbol{\Psi}_{\mathcal{O}} : \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} \subset \mathbb{R}^{\infty\times m}$	$\Phi_2 : \Omega_2 \to \mathcal{W}_2$
数据巴拿赫空间:	$\mathcal{X}_{\mathcal{O}} = \mathrm{span}(\boldsymbol{\Phi}_{\mathcal{O}}(\mathbb{X}))$, $\|\cdot\|_{\mathcal{X}_{\mathcal{O}}} = \|\cdot\|_F$	$\mathcal{W}_1$ 范数为 $\|\cdot\|_{\mathcal{W}_1}$
权重步巴拿赫空间:	$\mathcal{W}_{\mathcal{O}} =$ 张成空间$(\boldsymbol{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}}))$, $\|\cdot\|_{\mathcal{W}_{\mathcal{O}}} = \|\cdot\|_F$	$\mathcal{W}_2$ 范数为 $\|\cdot\|_{\mathcal{W}_2}$
双线性形式:	$\langle \boldsymbol{\Omega}\ \boldsymbol{\Xi}\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}} =$ 诊断 $(\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Xi})$	$\langle\cdot,\cdot\rangle_{\mathcal{W}_1\times\mathcal{W}_2} : \mathcal{W}_1 \times \mathcal{W}_2 \to \mathbb{Y}$

图1：根据 [Lin 等人 2022] 的工作，再生核巴拿赫空间构造的概要。

神经切核 [Jacot 等人，2018，Arora 等人，2019b]是较新的进展。神经切线核的基本思想是：对于给定输入$\mathbf{x}$，在权重和偏置围绕某些初始值变化时，用一阶泰勒近似来描述神经网络的行为。这种逼近关于权重变化是线性的，其系数是$\mathbf{x}$的函数，因此可将其视为特征映射，使模型适用于核技巧，并可进一步通过再生核希尔伯特空间理论进行分析。这种方法可以推广到更高阶的逼近 [Bai和Lee， 2019], 但可逼近的权重变化幅度仍然有限；过参数化极限是个例外，在此极限下权重的变化幅度会变小。

反余弦核 [Cho 和 Saul 2009] 的工作基于类似的前提。对于形式为 $\tau(\xi) = (\xi)_+^p$ 的激活函数 $p \in \mathbb{N}$，在无限宽度极限下，反余弦核能够捕捉网络的特征映射。通过核的复合实现深度。然而，这种方法再次受限于无限宽度的网络，而我们的方法适用于任意网络。

最后，最近出现了一些与本文思路相似的研究 [Bartolucci 等人，2021，Sanders，2020，Parhi 和 Nowak，2021，Unser，2021]，旨在将神经网络与再生核巴拿赫空间理论联系起来。然而，这些研究只考虑了 1 层和 2 层网络（我们考虑任意深度的网络），且没有证明通过反向传播得到的权重步长与通过再生核巴拿赫空间中的正则化学习得到的权重步长等价。

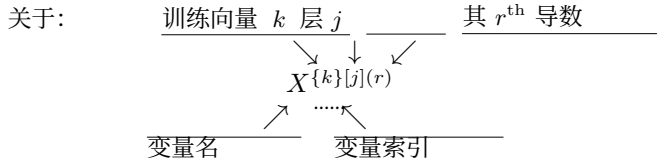
3 符号说明

令 $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{N}_n = \{0, 1, \ldots, n-1\}$。向量和矩阵分别记为$\mathbf{a}$和$\mathbf{A}$，其元素为 a_i,$A_{i,\ i'}$，列$\mathbf{A}_{:i}$ 由 $i,\ i' \in \mathbb{N}$索引。定义如下:

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

$\forall p, q \in [-\infty, 0) \cup (0, \infty]$，当 $p, q \in [1, \infty]$ 时它们是范数。弗罗贝尼乌斯范数和内积分别为 $\|\cdot\|_F = \|\cdot\|_2$,2,$\langle$A, B$\rangle_F$ = Tr(A^{T}B)。克罗内克积和哈达玛积分别为 a $\otimes$ b、a $\odot$ b。克罗内克幂和哈达玛幂分别为 a$^{\otimes c}$ = a $\otimes$ a $\otimes^c$次... $\otimes$ a、a$^{\odot c}$ = a $\odot$ a $\odot^c$次... $\odot$ a。逐元素绝对值和符号函数分别为 |a|、sgn(a)。最后，对于向量 a、b，令 $\boldsymbol{\varrho}$(a, b) = [a_0 (b$^{\otimes 1}$)^{T}, a_1(b$^{\otimes 2}$)^{T}, ...]^{T}。diag(A) 是一个包含 A 的对角元素的向量；反之，diag(a) 是一个对角元素取自 a 的对角矩阵。

我们研究全连接的 D 层神经网络 $\mathbf{f} : (\mathbb{X} \subset \mathbb{R}^n) \to (\mathbb{Y} \subset \mathbb{R}^m)$，其层宽为 $H^{[j]}$ （以及 $H^{[-1]} = n$），并在训练集 $\{(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}) \in \mathbb{X} \times \mathbb{Y} :\ k \in \mathbb{N}_N\}$ 上进行训练。我们采用索引范围约定 $k \in \mathbb{N}_N$、 $j \in \mathbb{N}_D$ 和 $i_{j+1} \in \mathbb{N}_{H^{[j]}}$，为清晰起见，记作:



例如，$\mathbf{W}^{[j]}$ 是第 j层的权重矩阵，$\mathbf{x}^{\{k\}[j]}$ 是网络在给定输入 $\mathbf{x}^{\{k\}}$时，第 j 层的输入（图像），而 $f^{(2)}(z)$ 是 $f(z)$ 的 2^{nd}导数。关于训练，$X_{\mathcal{O}}$表示“迭代前变量 X 的值”，而 X_Δ 表示“由于迭代导致的 X 的变化”。最后，在适用的情况下，我们使用上标 $X^{\boxtimes}$ 表示 X 与梯度下降（反向传播）有关，使用 $X^\bullet$ 表示 X 与 RKBS 正则化风险最小化有关。

4 背景

4.1 再生核巴拿赫空间

再生核希尔伯特空间（RKHS） [Aronszajn， 1950] 是一个函数希尔伯特空间 $\mathcal{H}$ $f : \mathbb{X} \to \mathbb{Y}$，其中点求值泛函 $\delta_{\mathbf{x}}(f) = f$(x)是连续的。因此，应用 Riesz表示定理,

存在一个核 K 使得:

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

随后， $K(\mathbf{x}, \mathbf{x}') = \langle K(\mathbf{x} \cdot) K(\mathbf{x}', \cdot)\rangle$ 并且，根据 Moore-Aronszajn定理， K 由 $\mathcal{H}$ 唯一确定，反之亦然。 K 被称为再生核，相应的再生核希尔伯特空间记作 $\mathcal{H}_K$。基于再生核希尔伯特空间的方法日益受到欢迎，因为它们非常适用于机器学习的诸多方面 [Steinwart and Christman, 2008, Shawe-Taylor and Cristianini, 2004]。内积结构使核技巧成为可能，并且核可以很容易地理解为一种相似性度量。此外，再生核希尔伯特空间的结构带来了丰富的复杂度分析和泛化界框架 [Steinwart and Christman, 2008, Shawe-Taylor and Cristianini, 2004]。近年来，神经切线核被提出 [Jacot等人, 2018], ，从而使再生核希尔伯特空间理论能够应用于过参数化情形下的神经网络训练。

为在再生核希尔伯特空间理论中引入更丰富的几何结构，再生核巴拿赫空间（RKBS）从函数巴拿赫空间出发，对再生核希尔伯特空间进行推广 [Der and Lee, 2007, Zhang et al., 2009, Song et al., 2013, Xu and Ye, 2014, Lin et al., 2022] 等。具体而言:

定义1 （再生核巴拿赫空间（RKBS - [Lin et al., 2022]））。 定义在集合 $\mathbb{X}$ 上的再生核巴拿赫空间$\mathcal{B}$ 是一个函数巴拿赫空间 $f: \mathbb{X} \to \mathbb{Y}$ ，其中每个点求值映射 $\delta_{\mathbf{x}}: \mathcal{B} \to \mathbb{Y}$， $\mathbf{x} \in \mathbb{X}$，在 $\mathcal{B}$ 上都是连续的（因此 $\forall \mathbf{x} \in \mathbb{X}\ \exists C_{\mathbf{x}} \in \mathbb{R}_+$满足 $|\delta_{\mathbf{x}}(f)| = |f(\mathbf{x})| \leq C_{\mathbf{x}}\|f\|_{\mathcal{B}}\ \forall f \in \mathcal{B}$）。

再生核巴拿赫空间构造有几种不同的方法。不过，在当前语境下，我们发现 [Lin 等，2022，定理2.1] 的方法最为方便。给定图1中所概述的组成部分，并假设$\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})$在 $\mathcal{X}_{\mathcal{O}}$ 中稠密，且 $\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})$在 $\mathcal{W}_{\mathcal{O}}$中稠密，我们在 $\mathbb{X}$ 上定义再生核巴拿赫空间 $\mathcal{B}_{\mathcal{O}}$ 如下:

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

其再生巴拿赫核为:

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

5 设置与假设

我们假设一个全连接的 D层前馈神经网络 $\mathbf{f}: (\mathbb{X} \subseteq \mathbb{R}^n) \to (\mathbb{Y} \subseteq \mathbb{R}^m)$，其各层宽度为$H^{[0]}$， $H^{[1]}$， ....., $H^{[D-1]}$，其中 $H^{[D-1]} = m$，并定义 $H^{[-1]} = n$。我们假设第 $j \in \mathbb{N}_D$层（始终为$j \in \mathbb{N}_D$）仅包含使用激活函数 $\tau^{[j]}$ 的神经元： $\mathbb{R} \to \mathbb{R}$。该网络递归定义为 $\forall j \in \mathbb{N}_D$:

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

其中 $\mathbf{W}^{[j]} \in \mathbb{R}^{H^{[j-1]} \times H^{[j]}}$ 和 $\mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}}$ 是权重和偏置，我们将其概括记为 $\mathbf{W} \in \mathbb{W}$，而 $\alpha^{[j]} \in \mathbb{R}_+$ 是固定的。这种形式的函数集记为 $\mathcal{F}$。

我们假设目标　　是给定一个训练集，找到使经验风险最小化的权重和偏置　　sk:

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

其中 $\mathbf{f}_{\mathbf{W}}$ 是一个具有式（3）形式的网络，其权重和偏置为 $\mathbf{W}$，$\mathbb{D} = \{(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}) \in \mathbb{X} \times \mathbb{Y}: k \in \mathbb{N}_N\}$ 是一个训练集（全文均如此），而 $E: \mathbb{X} \times \mathbb{Y} \times \mathbb{R}^m \to \mathbb{R}$ 是定义网络用途的误差函数。

我们作出以下技术性假设:

1. 输入空间： $\mathbb{X} = [-1, 1]^n$。

2. 误差函数： E： $\mathbb{X} \times \mathbb{Y} \times \mathbb{R}^m \to \mathbb{R}$ 在其第三个参数上是 $\mathcal{C}^1$ 且满足 L_E-利普希茨条件。

3. 激活函数：对于所有 $j \in \mathbb{N}_D$， $\tau^{[j]}$： $\mathbb{R} \to [-1, 1]^{H^{[j]}}$有界、 $\mathcal{C}^{\infty}$，且具有幂级数表示，其在所有 $z \in \mathbb{R}$周围的收敛域（ROC）至少为 $\rho^{[j]} \in \mathbb{R}_+$。

4. 权重非平凡性：对于所有 $j \in \mathbb{N}_D$， $\mathbf{W}^{[j]} \neq \mathbf{0}$ 在训练期间始终成立。[1]

5. 权重初始化：我们假设采用 LeCun 初始化，因此对于所有 $j \in \mathbb{N}_D$， $W_{i_j}^{[j]}{}_{,\ i_{j+1}}$， $b_{i_{j+1}}^{[j]} \sim \mathcal{N}(0, 1)$。

6. 训练：我们假设训练采用学习率为 $\eta \in \mathbb{R}_+$的梯度下降（反向传播）。

[1]不满足这一要求的网络，其输出为常数，且与输入$\mathbf{x}$无关。我们认为这并不是一个苛刻的假设，因为随机初始化的网络经过典型训练集训练后，几乎不可能达到这种状态。

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

图 2：神经网络在一次迭代前后的术语定义。

5.1 反向传播训练

如上所述，我们假设网络使用反向传播（梯度下降） [Goodfellow et al., 2016]进行训练。这是一种迭代方法。一次迭代从初始权重和偏置$\mathbf{W}_{\mathcal{O}}\in\mathbb{W}$开始。一个权重步长：

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

计算得出后，权重和偏置更新为 $\mathbf{W}\ =\mathbf{W}_{\mathcal{O}}+\mathbf{W}_{\Delta}^{\circledast}$。关于迭代前的激活以及权重步长导致的后续变化，我们采用的记号约定见图 2。权重步长为 [Goodfellow 等人， 2016] （推导见附录 B）：

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

对所有 $j\in\mathbb{N}_{D}$， i_{j+1} 其中，递归地定义为 $\forall j\in\mathbb{N}_{D-1}$:

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

注意，偏置 $\mathbf{b}_{\Delta}^{[j]}$ 的变化与 $\alpha^{[j]}$成正比。

6 单次迭代分析

在分析的第一阶段，我们考察权重和偏置发生微小变化（权重步长）所引起的神经网络行为变化。根据第 7 节的内容，可以据此直接推演出整体训练序列。我们的首要目标是将经过一次训练迭代后的神经网络改写为:

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

其中，$\mathbf{f}_{\mathcal{O}}=\mathbf{f}\mathbf{w}_{\mathcal{O}}:(\mathbb{X}\subset\mathbb{R}^{n})\to(\mathbb{Y}\subset\mathbb{R}^{m})$ 表示迭代前的神经网络，而 $\mathbf{f}_{\Delta}:(\mathbb{X}\subset\mathbb{R}^{n})\ \to\mathbb{R}^{m}$ 表示本次迭代中权重和偏置发生变化 $\mathbf{W}\ =\mathbf{W}_{\mathcal{O}}\to\mathbf{W}\ =\mathbf{W}_{\mathcal{O}}+\mathbf{W}_{\Delta}$ 所导致的网络行为变化，详见图 2，因此:

$$\begin{aligned}&\mathbf{f}_{\Delta}\left(\mathbf{x}\right)=\left\langle\boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right),\boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}\\ \text{where:}\quad&\boldsymbol{\Phi}_{\mathcal{O}}:\mathbb{X}\to\mathcal{X}_{\mathcal{O}}=\operatorname{span}\left(\boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbb{X}\right)\right)\subset\mathbb{R}^{\infty\times m}\\ &\boldsymbol{\Psi}_{\mathcal{O}}:\mathbb{W}_{\mathcal{O}}\to\mathcal{W}_{\mathcal{O}}=\operatorname{span}\left(\boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbb{W}_{\mathcal{O}}\right)\right)\subset\mathbb{R}^{\infty\times m}\end{aligned}\tag{7}$$

它们是完全由网络结构（层数及宽度、激活函数）和初始权重和偏置 $\mathbf{W}_{\mathcal{O}}$决定的特征映射；而 $\left\langle\boldsymbol{\Xi},\boldsymbol{\Omega}\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}=\operatorname{diag}\left(\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Omega}\right)$ 是一种双线性形式。随后，我们的第二个目标是根据这些特征映射推导出核和范数，以便研究它们的收敛性质。最后，我们将证明，由单次反向传播产生的权重步长 $\mathbf{W}_{\Delta}^{\circledast}$ 与使（再生核巴拿赫空间）正则化风险最小化的对应权重步长 $\mathbf{W}_{\Delta}^{\bullet}$ 等价。

6.1 贡献一：特征映射展开

在本节中，我们推导出适当的特征映射，以表示有限权重步长下神经网络行为的变化。原则上，我们的方法很简单，但涉及一些技术细节，因此详见附录 B。不过，概括而言，对于光滑激活函数 $\tau^{[j]}:\ \mathbb{R}\to\mathbb{R},z\in\mathbb{R}$ 以及有限维向量 $\mathbf{c},\ \mathbf{c}'$，若其内积落在收敛半径 $\rho^{[j]}$ 内（因此 $|\langle\mathbf{c},\ \ \mathbf{c}\rangle|<\rho^{[j]}$），则 $\tau^{[j]}$关于 z的幂级数表示可写为 $\tau^{[j]}(z+\langle\mathbf{c},\ \ \mathbf{c}\rangle)$ $=\tau^{[j]}(z)\ +\left\langle\boldsymbol{\varrho}(\mathbf{g}^{[j]}(z),\ \ \mathbf{c})\ \boldsymbol{\varrho}(\mathbf{1}_{\infty},\ \ \mathbf{c}')\right\rangle$，其中:

$$\begin{aligned}\boldsymbol{\varrho}\left(\mathbf{a},\mathbf{d}\right)&=\left[\begin{array}{cccc}a_{0}\mathbf{d}^{\otimes1\mathrm{T}}&a_{1}\mathbf{d}^{\otimes2\mathrm{T}}&a_{2}\mathbf{d}^{\otimes3\mathrm{T}}&\ldots\end{array}\right]^{\mathrm{T}}\\ \mathbf{g}^{[j]}\left(z\right)&=\left[\begin{array}{cccc}\frac{1}{1!}\tau^{[j](1)}\left(z\right)&\frac{1}{2!}\tau^{[j](2)}\left(z\right)&\frac{1}{3!}\tau^{[j](3)}\left(z\right)&\ldots\end{array}\right]^{\mathrm{T}}\end{aligned}$$

给定输入 $\mathbf{x}$，从第 0 层开始向前计算，并参照图 2，我们可以将权重步长 $\mathbf{W}_\Delta$ 导致第 0 层输出的变化 $\mathbf{x}_\Delta^{[1]}$ 写为：

$$\begin{aligned}\mathbf{x}_\Delta^{[1]} &= \left[\left\langle \boldsymbol{\Phi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{x}), \boldsymbol{\Psi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{W}_\Delta)\right\rangle\right]_{i_1} \\ \text{where: } \boldsymbol{\Phi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{x}) &= \boldsymbol{\varrho}\left(\mathbf{g}^{[0]}\left(\tilde{x}_{\mathcal{O}i_1}^{[0]}\right), \mu_{i_1}^{[0]}\begin{bmatrix}\frac{1}{\sqrt{2}}\alpha^{[0]} \\ \frac{1}{\sqrt{2H^{[0]}}}\mathbf{x}_\mathcal{O}\end{bmatrix}\right) \\ \boldsymbol{\Psi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{W}_\Delta) &= \boldsymbol{\varrho}\left(\mathbf{1}_\infty, \frac{1}{\mu_{i_1}^{[0]}}\begin{bmatrix}\sqrt{2}b_{\Delta i_1}^{[0]} \\ \sqrt{2}\mathbf{W}_{\Delta:i_1}^{[0]}\end{bmatrix}\right)\end{aligned} \tag{8}$$

这里需要注意的是，两个特征映射的收敛半径都是有限的。这些特征映射由缩放因子 $\mu_{i_1}^{[0]} \in \mathbb{R}_+$ 参数化，其作用主要是技术性的，因为它们将使我们能够证明 RKBS 正则化风险最小化与反向传播之间的等价性。[2] 在此处，除了其存在性之外，它们的确切取值并不重要。

该过程会 在后续层中重复（详见附录 B）。完成所有层的处理后：

$$\mathbf{f}_\Delta(\mathbf{x}) = \langle \boldsymbol{\Phi}_\mathcal{O}(\mathbf{x}), \boldsymbol{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}} \tag{9}$$

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

$$\begin{aligned}\boldsymbol{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x}) &= \boldsymbol{\varrho}\left(\mathbf{g}^{[j]}\left(\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}\right), \mu_{i_{j+1}}^{[j]}\begin{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}}\alpha^{[j]} \\ \frac{1}{\sqrt{H^{[j]}}}\mathbf{x}_\mathcal{O}^{[j]}\end{bmatrix} \\ \left[\frac{\omega_{i_j}^{[j]} W_{\mathcal{O}i_j,i_{j+1}}^{[j]}}{\widetilde{\omega}_{i_j i_{j+1}}^{[j]}\sqrt{H^{[j]}}}\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right]_{i_j} \\ \left[\frac{\omega_{i_j}^{[j]}}{\sqrt{H^{[j]}}}\boldsymbol{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right]_{i_j}\end{bmatrix}\right) \\ \boldsymbol{\Psi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{W}_\Delta) &= \boldsymbol{\varrho}\left(\mathbf{1}_\infty, \frac{1}{\mu_{i_{j+1}}^{[j]}}\begin{bmatrix}\begin{bmatrix}\sqrt{2}b_{\Delta i_{j+1}}^{[j]} \\ \mathbf{W}_{\Delta:i_{j+1}}^{[j]}\end{bmatrix} \\ \left[\frac{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]}}{\omega_{i_j}^{[j]}}\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right]_{i_j} \\ \left[\frac{W_{\Delta i_j,i_{j+1}}^{[j]}}{\omega_{i_j}^{[j]}}\boldsymbol{\Psi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{W}_\Delta)\right]_{i_j}\end{bmatrix}\right)\end{aligned} \tag{10}$$

递归地 $\forall j \in \mathbb{N}_D\backslash\{0\}$，它们由缩放因子 $\mu_{i_{j+1}}^{[j]} \in \mathbb{R}_+$ 和影子权重 $\omega_{i_j}^{[j]}$，$\widetilde{\omega}_{i_j,\,i_{j+1}}^{[j]} \in \mathbb{R}_+$ 参数化，并在 RKBS 正则化风险最小化与反向传播之间的等价性中发挥作用（否则，它们的确切取值并不重要）。

6.2 贡献二：诱导核与范数

在 前一节中，我们证明了，经过一次权重步长 $\mathbf{W}_\Delta$ 后，可以写成 ：

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

其中 $\mathbf{f}_\mathcal{O} : \mathbb{X} \to \mathbb{Y}$ 是迭代前的神经网络，而特征映射 $\boldsymbol{\Phi}_\mathcal{O}, \boldsymbol{\Psi}_\mathcal{O}$ 是特征映射（8–10）。利用这些，我们通过核技巧在 $\mathbb{X}$，$\mathbb{W}_\mathcal{O}$ 上诱导出核：

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

我们称这些核为 神经邻域核 （NNK），因为它们描述了$\mathbf{W}_\mathcal{O}$ 的有限邻域内的相似性结构（与神经切线核不同，后者描述的是与$\mathbf{W}_\mathcal{O}$相切的行为，或其无穷小邻域内的行为）。这些矩阵值核按构造是对称且正定的，可能可用于支持向量机（SVM）或类似的基于核的方法（进行迁移），分别衡量 $\mathbb{X}$ 和 $\mathbb{W}_\mathcal{O}$上的相似性。同样，我们诱导出一个巴拿赫核：

$$\begin{aligned}\mathbf{K}_\mathcal{O}(\mathbf{x},\mathbf{W}'_\Delta) &= \langle \boldsymbol{\Phi}_\mathcal{O}(\mathbf{x}), \boldsymbol{\Psi}_\mathcal{O}(\mathbf{W}'_\Delta)\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}} \\ &= \mathrm{diag}(\mathbf{f}_\Delta(\mathbf{x}))\end{aligned} \tag{11}$$

这显然就是对输入向量 $\mathbf{x}$施加权重步长 $\mathbf{W}_\Delta$ 后，网络输出 $\mathbf{f}_\Delta(\mathbf{x})$ 的变化，对其进行对角化。

[2]更多讨论请参见附录。

神经邻域核的具体形式相当复杂（推导过程很直接，但得到的递归方程非常冗长）。NNK 是 NTK 未经近似的对应物。然而，尽管神经网络在一阶近似下沿着 NTK 所定义的 RKHS 演化，但 NNK 并非如此。事实上，不难看出，使用 *NNK* 进行再生核希尔伯特空间正则化，得到的权重向量将始终不是权重步长在 $\mathbf{\Psi}_{\mathcal{O}}$ 下的像（也就是说，再生核希尔伯特空间理论不足以解决问题，我们需要再生核巴拿赫空间理论才能继续）。因此，由于 NNK 并非本文的重点，我们不会在论文正文中重现这些核；感兴趣的读者可参阅附录 C。利用这些诱导核，我们可以得到 $\mathbb{X}$和 $\mathbb{W}_{\mathcal{O}}$ 在特征空间中的像的范数表达式：

$$\begin{aligned}\left\|\mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right)\right\|_{\mathcal{X}_{\mathcal{O}}}^{2} &= \left\|\mathbf{\Phi}_{\mathcal{O}}^{[D-1]}\left(\mathbf{x}\right)\right\|_{F}^{2} \\ &= \mathrm{Tr}\left(\mathbf{K}_{\mathcal{X}_{\mathcal{O}}}\left(\mathbf{x},\mathbf{x}\right)\right) \\ \left\|\mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\|_{\mathcal{W}_{\mathcal{O}}}^{2} &= \left\|\mathbf{\Psi}_{\mathcal{O}}^{[D-1]}\left(\mathbf{W}_{\Delta}\right)\right\|_{F}^{2} \\ &= \mathrm{Tr}\left(\mathbf{K}_{\mathcal{W}_{\mathcal{O}}}\left(\mathbf{W}_{\Delta},\mathbf{W}_{\Delta}\right)\right)\end{aligned} \tag{12}$$

再一次，这些表达式的精确形式较为复杂（推导过程直接明了，但答案很长——详情见附录 C）。这些范数的重要性在于可用于推导特征映射收敛的条件。定义辅助函数：

$$\overline{\sigma}^{[j]}\left(\zeta\right) = \max_{z\in\mathbb{R}_{+}\cup\{0\}}\left\{\sum_{l=1}^{\infty}\left(\tfrac{1}{l!}\right)^{2}\tau^{[j](l)}\left(z\right)\tau^{[j](l)}\left(z\right)\zeta^{l}\right\} \tag{13}$$

以及常数：

$$\begin{aligned}s^{[j]2} &= \begin{cases}\frac{1}{2}\alpha^{[j]2}+\frac{1}{2}\frac{H^{[-1]}}{H^{[0]}} & \text{if } j=0\\ \frac{1}{2}\alpha^{[j]2}+\frac{H^{[j-1]}}{H^{[j]}} & \text{otherwise}\end{cases} \\ t_{\Delta i_{j+1}}^{[j]2} &= \begin{cases}\left[2b_{\Delta i_1}^{[0]2}+2\left\|\mathbf{W}_{\Delta:i_1}^{[0]}\right\|_2^2\right]_{i_1} & \text{if } j=0\\ \left[2b_{\Delta i_{j+1}}^{[j]2}+\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2\right]_{i_{j+1}} & \text{otherwise}\end{cases}\end{aligned} \tag{14}$$

粗略而言，它们分别是第 j 层到第 $j-1$ 层宽度扩张/收缩（扇出）的替代量，以及第 j层权重步长大小的替代量；在附录 C.6 中，我们推导出以下引理（参见附录中定理 5 和定理 6 的相应证明）：

引理 1. 令 $\epsilon_{\phi}^{[j]}\in(0,1)\ \ \forall j\in\mathbb{N}_{D}$。对于给定的神经网络和初始权重 $\mathbf{W}_{\mathcal{O}}$，定义 $\underset{H^{[j]}}{\phi^{[j]}}=\sigma^{[j]}((1-\epsilon_{\phi}^{[j]})\sqrt{\rho^{[j]}})\ \ \forall j\in\mathbb{N}_{D}$。若缩放因子满足：

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

$\forall j\in\mathbb{N}_{D}, i_{j+1}$ 则 $\left\|\mathbf{\Phi}_{\mathcal{O}}^{[j]}\left(\mathbf{x}\right)\right\|_{F}^{2}\leq\phi^{[j]}\ \forall\mathbf{x}\in\mathbb{X}$。

引理 2。令 $\epsilon_{\psi}^{[j]}\in(0,1)\ \ \forall j\in\mathbb{N}_{D}$。对于给定的神经网络、初始权重 $\mathbf{W}_{\mathcal{O}}$ 和权重步长 $\mathbf{W}_{\Delta}$，若：

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

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

引理 1 使我们能够对缩放因子设定界限，以确保对于所有 $\mathbf{x}\in\mathbb{X}$，特征映射 $\mathbf{\Phi}_{\mathcal{O}}^{[j]}:\mathbb{X}\to\mathcal{X}_{\mathcal{O}}$ 是有限的（定义良好）。引理 2 与此类似，但它并非对缩放因子设定界限，而是将其视为已知（由引理 1 给出），并对权重步长 $\mathbf{W}_{\Delta}$ 的大小设定界限，使得特征映射 $\mathbf{\Psi}_{\mathcal{O}}^{[j]}:\mathbb{W}\to\mathcal{W}_{\mathcal{O}}$ 是有限的（定义良好）。因此，综合这两个引理，我们可以得到权重步长大小的某种界限，该步长可针对给定的神经网络结构和初始权重 $\mathbf{W}_{\mathcal{O}}$进行建模。不过，它们并未直接说明影子权重。在下一节中，我们将利用这些引理，建立梯度下降产生的权重步长与再生核巴拿赫空间中的学习之间的联系，并进一步阐明使用我们的构造进行建模时，权重步长的大小范围。

6.3 贡献 3：再生核巴拿赫空间中梯度下降与正则化风险最小化的等价性

在我们此前的研究中，我们证明，神经网络行为的变化$\mathbf{f}_\Delta$ 可以表示为形式 (9)，其中特征映射为 (8,10)；我们还根据这些特征映射导出了核，并给出了该表示成立时权重步长大小的界限。在本节中，我们利用这些结果，建立神经网络中的梯度下降学习与再生核巴拿赫空间（再生核巴拿赫空间）中的正则化风险最小化之间的联系。

首先，不难看出，特征映射 $\mathbf{\Phi}_\mathcal{O}\colon \mathbb{X}\to\mathcal{X}_\mathcal{O}$ 和 $\mathbf{\Psi}_\mathcal{O}\colon \mathbb{W}_\mathcal{O}\to\mathcal{W}_\mathcal{O}$ 利用式 (1) 定义了一个再生核巴拿赫空间 $\mathcal{B}_\mathcal{O}$，其赋有 $\|\cdot\|_{\mathcal{B}_\mathcal{O}}=\|\cdot\|_F$，其中使用的是由式 (2) 导出的再生巴拿赫核 (11)：[3]

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

据此，反向传播迭代引起的网络行为变化 $\mathbf{f}_\Delta^{\boxtimes}$ 可写为：

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

对于一个初始权重为 $\mathbf{W}_\mathcal{O}$ 的给定神经网络，我们假设权重步长通过梯度下降来选取。[4] 另一种方法是选择权重步长，使再生核巴拿赫空间中的正则化风险最小，具体而言：

$$\begin{aligned}\mathbf{W}_\Delta^{\bullet}&=\operatorname*{argmin}_{\mathbf{W}_\Delta\in\mathbb{W}_\mathcal{O}}R_\lambda\left(\mathbf{W}_\Delta\right)\\R_\lambda\left(\mathbf{W}_\Delta\right)&=\lambda\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta\right)\right\|_{\mathcal{W}_\mathcal{O}}+R_E\left(\mathbf{W}_\Delta+\mathbf{W}_\mathcal{O},\mathbb{D}\right)\end{aligned}\tag{15}$$

其中，我们将 λ 称为权衡系数。权衡系数越大，权重步长越小，反之亦然。与通过反向传播推导出的权重步长相比，这种形式的优点在于，我们可以直接应用再生核巴拿赫空间理论中的复杂度界等结果，并将其扩展到完整的训练过程。这促使我们提出以下问题：

> 对于给定的神经网络，其初始权重和偏置为 $\mathbf{W}_\mathcal{O}$，令 $\mathbf{W}_\Delta^{\boxtimes}$ 为由式（*5*）定义的反向传播权重步（学习率为 η 的梯度下降），并令 $\mathbf{W}_\Delta^{\bullet}$ 为求解正则化风险最小化问题（*15*）的权重步。给定由梯度下降得到的权重步$\mathbf{W}_\Delta^{\boxtimes}$，我们能否选择缩放因子、影子权重和权衡参数 λ（作为 $\mathbf{W}_\Delta^{\boxtimes}$ 的函数），以保证 $\mathbf{W}_\Delta^{\bullet}=\mathbf{W}_\Delta^{\boxtimes}$？

如果答案是肯定的（我们将对此加以证明），那么我们就可以通过分析（15）来加深对反向传播的理解。现在，式（15）的解必须满足一阶最优性条件（为简便起见，假设其可微），因此：

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

注意，如果 正则化项的梯度满足：

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

对于某个 $\nu\in\mathbb{R}_+$ 和 $\frac{1}{\lambda}=\eta\nu$，则 $\mathbf{W}_\Delta^{\bullet}=\mathbf{W}_\Delta^{\boxtimes}$。因此，对于给定的学习率 η，正则化风险最小化权重步是否对应于梯度下降权重步这一问题，可以通过证明规范缩放的存在性得到肯定答案。规范缩放定义如下：

定义 2（规范缩放）。对于给定的神经网络、由反向传播生成的初始权重 $\mathbf{W}_\mathcal{O}$ 和权重步 $\mathbf{W}_\Delta^{\boxtimes}$，我们将一组缩放因子和影子权重称为规范缩放，其中$\partial\,{}^{\partial}\mathbf{w}_\Delta\left\|\mathbf{\Psi}_\mathcal{O}\left(\mathbf{W}_\Delta^{\boxtimes}\right)\right\|_{\mathcal{W}_\mathcal{O}}=\nu\mathbf{W}_\Delta^{\boxtimes}$ 对于 $\nu\in\mathbb{R}_+$成立，且 $\left\|\mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta^{\boxtimes})\right\|_{\mathcal{W}_\mathcal{O}}$，$\|\mathbf{\Phi}_\mathcal{O}(\mathbf{x})\|_{\mathcal{X}_\mathcal{O}}<\infty\ \forall\mathbf{x}\in\mathbb{X}$。

为了证明规范缩放的存在性，从而揭示梯度下降（反向传播）与再生核巴拿赫空间中的学习之间的关键联系，利用（14），定义：

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

$\forall j\in\mathbb{N}_D$，i_{j+1}，在附录中，我们基于引理1和引理2证明以下关键结果：

定理1. 设 $\epsilon,\chi\in(0,1)$。对于一个初始权重为 $\mathbf{W}_\mathcal{O}$ 的给定神经网络，设 $\mathbf{W}_\Delta^{\boxtimes}$ 为通过反向传播得到的权重步长，并不失一般性地假设 $\alpha^{[j]}\in\mathbb{R}_+$ 的选取满足 $\forall j\in\mathbb{N}_{D-1}$：[5]

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

[3]我们在附录中证明，这些映射满足相关的稠密性要求。[4]请注意，$\mathbb{D}$ 此处的训练集[5]仅用于本次迭代，也可能是较大训练集的随机子集。$b_{\Delta i_{j+1}}^{[j]\boxtimes}$ 请注意，$\alpha^{[j]}$与 $t_{\Delta i_{j+1}}^{[j]\boxtimes 2}$ 成正比，因此我们总能通过调整 $\alpha^{[j]}$来增大，确保条件成立。

令 $\epsilon_\psi = 1 - \frac{1}{1-\chi}\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}}\left\|\mathbf{t}_\Delta^{[D-1]\boxplus}\right\|_\infty^2$ 且：[6]

$$\epsilon_\phi^{[j]} = 1 - \frac{1}{\sqrt{\rho^{[j]}}}\overline{\sigma}^{[j]-1}\left(\frac{\frac{1}{1-\chi}\frac{\chi^{D-1-j}\epsilon_\psi\left(1-\epsilon_\phi^{[j+1]}\right)\sqrt{\rho^{[j+1]}}}{\left\|\mathbf{t}_\Delta^{[j+1]\boxplus}\right\|_\infty^2} - \frac{\chi^{D-1-j}\epsilon_\psi}{1-\chi^{D-1-j}\epsilon_\psi}s^{[j+1]2}}{\max_{i_{j+2}}\left\{\frac{\frac{1}{H^{[j+1]}}\left\|\mathbf{w}_{\mathcal{O}:i_{j+2}}^{[j+1]}\right\|_2^2}{\left\|\mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxplus}\right\|_2^2 - \left\|\mathbf{w}_{\Delta:i_{j+2}}^{[j+1]\boxplus}\right\|_\infty^2} + \frac{H^{[j]}}{H^{[j+1]}}\right\}}\right)$$

$\forall j \in \mathbb{N}_{D-1}$，其中 $\epsilon_\phi^{[D-1]} = \epsilon$。如果权重步长满足：

$$\left\|\mathbf{t}_\Delta^{[j]\boxplus}\right\|_\infty^2 \quad < B^{[j]2} = \begin{cases} \frac{(1-\chi)^2\left(1-\chi^{D-j-1}\epsilon_\psi\right)}{\left(\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}\right)} & if\ j < D-1 \\ \frac{(1-\chi)^2}{\left(\frac{s^{[D-1]2}}{(1-\epsilon)\sqrt{\rho^{[D-1]}}}\right)} & otherwise \end{cases}$$

$\forall j \in \mathbb{N}_D$ 那么存在一种规范缩放：

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

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

这在附录中作为推论 10 得到了证明。该定理告诉我们，对于任意一组初始权重 $\mathbf{W}_\mathcal{O}$，由反向传播生成的足够小的权重步长 $\mathbf{W}_\Delta^{\boxplus}$，都存在一种规范缩放——即一组缩放因子和影子权重，使得梯度下降权重步与使用适当权衡参数 λ 的正则化 RKBS 学习所生成的步长完全等价。请注意：

- 第 j 层中的最大步长 $B^{[j]}$（忽略接近恒等变换的缩放项）由扇出 $1/s^{[j]}$ 的倒数决定（其大致按 $H^{[j]}/H^{[j-1]}$ 缩放），并受缩放后的收敛半径（$1-\epsilon_\phi^{[j]}$）$\sqrt{\rho^{[j]}}$ 影响（其大致与后续层中的权重步长成反比）。从输出层向输入层移动时，这个上界往往会逐渐减小，但在许多情况下，权重步长也会随之减小，这是由梯度消失问题导致的。

- 这种规范缩放所需的权衡系数（正则化程度）为：

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

如图3所示。请注意：(a) 为生成等效的再生核巴拿赫空间权重步长，所需的正则化程度与生成原始反向传播权重步长时所用的学习率成反比；(b) 较大的梯度下降权重步长等效于正则化程度较低的再生核巴拿赫空间权重步长，这与预期相符。

7 应用：拉德马赫复杂度

在建立了权重和偏置有限邻域内神经网络的精确表示、确立了其与再生核巴拿赫空间理论的联系，并证明通过适当的事后缩放因子和影子权重选择（规范缩放），梯度下降步等价于再生核巴拿赫空间中的正则化步之后，我们现在考虑将这一框架应用于使用拉德马赫复杂度的一致收敛分析。实值函数集合 $\mathcal{G}$ 的拉德马赫复杂度是衡量其容量的一种指标。假设训练向量 $\mathbf{x}_i \sim \nu$ 和拉德马赫随机变量 $\epsilon_i \in \{-1, 1\}$，则 $\mathcal{G}$ 的拉德马赫复杂度为 [Mendelson, 2003]:

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

这可用于一致收敛分析，以界定经验风险收敛到期望风险的速度，通常形式如下：

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

下面的定理展示了如何运用我们的框架来界定标量输出神经网络的拉德马赫复杂度：

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

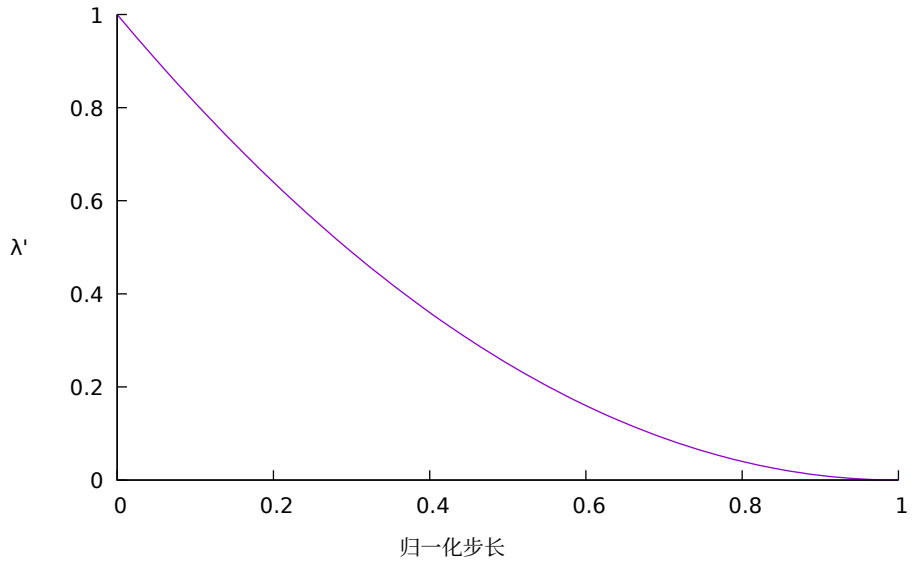


图3：归一化规范权衡系数 $\lambda' = \frac{4\eta}{B^{[D-1]2}}\lambda$ 与归一化梯度下降步长 $\frac{\|\mathbf{t}_{\Delta}^{[j]\boxplus}\|_{\infty}}{B^{[D-1]}}$。

定理2。设 $\epsilon, \chi \in (0,1)$，对于一个初始权重为 $\mathbf{W}_{\mathcal{O}}$ 的给定神经网络，令 $\mathbf{W}_{\Delta}^{\boxplus}$ 为通过反向传播得到的权重步长，并满足推论 *10* 中给出的条件。则 $f_{\Delta}^{\boxplus} \in \mathcal{F}^{\bullet}$，其中 $\mathcal{F}^{\bullet}$ 的拉德马赫复杂度满足以下界：

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

该定理的证明是定理1的直接应用（详情请参见附录中定理 11 的证明）。除通常的 $\frac{1}{N}$ 缩放外，该定理与典型的拉德马赫复杂度界非常不同，因为它直接利用梯度下降单次迭代的权重步长大小来界定复杂度。注意：

- 根据拉德马赫复杂度的性质 [Bartlett 和 Mendelson，2002，定理 12]训练后的神经网络经过 T 次迭代，其复杂度可由每一步界的简单求和来界定。这支持了通过提前停止来防止神经网络过拟合的做法。[7]

- 反向传播权重步长的大小与学习率成正比，因此对于足够小的学习率，我们得到：

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

  然而，随着学习率增大，这个界会以加速的速度增大。[8] 这支持了较低的学习率在神经网络训练中起到正则化作用的观点。

最后，我们考虑一个具体例子。假设有一个标量输出的两层网络（$D = 2$，$n = H^{[1]} = 1$），其激活函数为 $\tau^{[0]} = \tau^{[1]} =$ 双曲正切（因此 $\rho^{[0]} = \rho^{[1]} = \frac{\pi}{2}$ 且 $\tau_{(1)}{}^{[0]}(\zeta)\ \tau^{(1)[1]}(\zeta) \in [-1,1]$。$\sigma_z^{[j]}{}_{,z'}$ 的形式并不平凡（详情见附录 F），但不难看出 $\sigma^{[j]} = \sigma_{0\,0}^{[j]}{}_{,}$，因此，利用关于 0 的幂级数：

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

此外，$s^{[0]2} = \alpha^{[0]2}/2 + m/2H^{[0]}$ 和 $s^{[1]2} = \alpha^{[1]2}/2 + H^{[0]}$。利用我们的假设和反向传播方程，不难看出：

$$\begin{aligned}\left\|\mathbf{t}_{\Delta}^{[0]\boxplus}\right\|_{\infty}^2 &\leq 2\eta^2 N^2 L_E^2 \left\|\mathbf{W}_{\mathcal{O}}^{[1]}\right\|_F^2 \left(1+\alpha^{[0]2}\right)\\ \left\|\mathbf{t}_{\Delta}^{[1]\boxplus}\right\|_{\infty}^2 &\leq 2\eta^2 N^2 L_E^2 \left(1+\alpha^{[1]2}\right)\end{aligned}$$

[7]我们推测，若更仔细地分析各步骤所对应的再生核 Banach 空间之间的重叠，可能会得到对 T 的次线性依赖，但这超出了本文的研究范围。[8]最终，当步长超过我们能够用再生核巴拿赫空间建模的范围时，该界将失去意义。

以及 $B^{[1]} = \frac{(1-\chi)^2 \, 1-\epsilon \sqrt{\frac{\pi}{2}}}{\frac{1}{2}\alpha^{[1]2}+H^{[0]}}$ 。假设学习率为：

$$\eta = s\frac{(1-\chi)}{NL_E\sqrt{2\left(1+\alpha^{[1]2}\right)}}\sqrt{\frac{(1-\epsilon)\sqrt{\frac{\pi}{2}}}{\frac{1}{2}\alpha^{[1]2}+H^{[0]}}}$$

其中 $0 < s \ll 1$ ——遵循学习率与批量大小及网络宽度成反比的缩放惯例——我们得到训练 T 次迭代后的神经网络的以下拉德马赫复杂度上界：

$$\frac{\mathcal{R}_N(\mathcal{F}^\bullet)}{H^{[D-1]}} \lessapprox sT\sqrt{2\frac{\bar{\sigma}^{[D-1]}\left((1-\epsilon)\sqrt{\frac{\pi}{2}}\right)}{N}}\frac{(1-\chi)^2(1-\epsilon)\sqrt{\frac{\pi}{2}}}{\frac{1}{2}\alpha^{[1]2}+H^{[0]}}$$

我们注意到，该上界与训练向量的数量 N 和 网络宽度 $H^{[0]}$均成反比。

8 结论与未来方向

在本文中，我们建立了使用梯度下降进行神经网络训练与再生核巴拿赫空间中的正则化学习之间的联系。我们通过两个特征映射的内积，给出了神经网络行为的精确表示：当权重和偏置在某些初始权重和偏置的有限邻域内变化时，其中一个特征映射从数据空间映射到特征空间，另一个则从权重步长空间映射到特征空间。基于此，我们证明了反向传播单次迭代导致的神经网络行为变化属于再生核巴拿赫空间，并且反向传播得到的权重步长可以通过再生核巴拿赫空间中的正则化风险最小化精确复现。随后，我们给出了适用于过参数化和欠参数化情形的神经网络拉德马赫复杂度上界，并讨论了该上界如何取决于学习率、数据集大小、网络宽度以及所用的训练迭代次数。

关于未来的工作，我们预计有若干值得探索的方向。首先，应将分析扩展到 ReLU 等非光滑激活函数，或许可以通过使用幂级数展开以外的表示来修改特征映射。其次，需要进一步阐明学习率的具体影响，以及定理1的其他细节。对于神经邻域核本身，如果能将其化为闭式，将有助于在实践中应用。[9] 最后，还需要进一步研究网络深度对这一理论的影响。

参考文献

Z. Allen-Zhu、Y. Li 和 Y. Liang. 过参数化神经网络中的学习与泛化：超越两层网络。*arXiv*预印本 *arXiv:1811.04918*，2018年。Z. Allen-Zhu、Y. Li 和 Z. Song. 通过过参数化建立深度学习的收敛理论。载于国际机器学习会议，第 242–252 页。PMLR，2019年。N. Aronszajn. 再生核理论。《美国数学学会会刊》，68:337–404，1950 年 1 月–6 月。S. Arora、R. Ge、B. Neyshabur 和 Y. Zhang. 通过压缩方法为深度网络建立更强的泛化界。载于*ICML* 会议论文集，2018年。S. Arora、S. Du、W. Hu、Z. Li 和 R. Wang. 对过参数化双层神经网络的优化与泛化进行细粒度分析。载于国际机器学习会议，第 322–332 页。PMLR，2019年a。S. Arora、S. S. Du、W. Hu、Z. Li、R. R. Salakhutdinov 和 R. Wang. 关于无限宽神经网络的精确计算。载于神经信息处理系统进展，第 8139–8148 页，2019年b。F. Bach. 关于核求积规则与随机特征展开的等价性。《机器学习研究杂志》，18(1):714–751，2017年。F. R. Bach. 用凸神经网络打破维数灾难。*CoRR*，abs/1412.8690，2014年。网址：`http://arxiv。org/abs/1412。8690`。Y. Bai 和 J. D. Lee. 超越线性化：宽神经网络的二次及高阶逼近。*arXiv*预印本 *arXiv:1910.01619*，2019年。P. L. Bartlett 和 S. Mendelson. 拉德马赫复杂度与高斯复杂度：风险界与结构性结果。《机器学习研究杂志》，3:463–482，2002年。P. L. Bartlett、D. J. Foster 和 M. J. Telgarsky. 神经网络的谱归一化间隔界。载于神经信息处理系统进展，第 6240–6249 页，2017年。

[9]神经邻域核的泰勒近似很容易得到，但如果不作进一步分析，就很难判断其精确度。

F. Bartolucci、E. De Vito、L. Rosasco 和 S. Vigogna. 通过再生核巴拿赫空间理解神经网络。*arXiv*预印本 *arXiv:2109.09710*，2021年。

Y. Cao 和 Q. Gu. 宽而深的神经网络随机梯度下降泛化界。载于神经信息处理系统进展，第 32 卷，2019年。Y. Cho 和 L. K. Saul. 深度学习的核方法。载于 B. Y.、S. D.、L. J. D.、C. K. I. Williams 和 A. Culotta 编，神经信息处理系统进展第 *22* 卷，第 342–350 页。Curran Associates, Inc.，2009。网址：http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf。S. R. Chowdhury 和 A. Gopalan. 关于核化多臂老虎机。载于 D. Precup 和 Y. W. Teh 编，第 *34* 届国际机器学习会议论文集，《机器学习研究论文集》第 70 卷，第 844–853 页，澳大利亚悉尼国际会议中心，2017 年 8 月。PMLR。C. Cortes 和 V. Vapnik. 支持向量网络。机器学习，20(3):273–297，1995。N. Cristianini 和 J. Shawe-Taylor. 支持向量机及其他基于核的学习方法导论。剑桥大学出版社，英国剑桥，2005。A. Daniely. SGD 学会网络的共轭核类。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展第 *30* 卷，第 2422–2430 页。Curran Associates, Inc.，2017。网址：http://papers.nips.cc/paper/6836-sgd-learns-the-conjugate- kernel-newlineclass-of-the-network.pdf。A. Daniely、R. Frostig 和 Y. Singer. 深入理解神经网络：初始化的力量及对表达能力的对偶视角。载于 D. D. Lee、M. Sugiyama、U. V. Luxburg、I. Guyon 和 R. Garnett 编，神经信息处理系统进展第 *29* 卷，第 2253–2261 页。Curran Associates, Inc.，2016。网址：http://papers.nips.cc/paper/6427-toward-deeper-understanding-of-neural-networks-the-power-of-initialization-and-a-dual-view-on-expressivity.pdf。J. D'Aurizio. tanh x 的泰勒级数展开。数学 Stack Exchange https://math.stackexchange.com/q/ 1052926（版本：2014-12-05），2014。R. Der 和 D. Lee. 巴拿赫空间中的大间隔分类。载于*JMLR* 研讨会与会议论文集第 *2* 卷：*AISTATS2007*，第 91–98 页，2007。F. Dr¨axler、K. Veschgini、M. Salmhofer 和 F. A. Hamprecht. 神经网络能量景观中几乎不存在障碍。载于第 *35* 届国际机器学习会议论文集（*ICML 2018*），2018。S. Du、J. Lee、H. Li、L. Wang 和 X. Zhai. 梯度下降找到深度神经网络的全局最小值。载于国际机器学习会议，第 1675–1685 页。PMLR，2019年a。S. S. Du、X. Zhai、B. Poczos 和 A. Singh. 梯度下降可证明地优化过参数化神经网络。载于学习表征会议，2019年b。M. G. Genton. 机器学习的核类：统计学视角。《机器学习研究杂志》，2:299–312，2001。N. Golowich、A. Rakhlin 和 O. Shamir. 神经网络的样本复杂度与规模无关。载于$COLT$，2018。M. G¨onen 和 E. Alpaydin. 多核学习算法。《机器学习研究杂志》，12:2211–2268，2011。I. Goodfellow、Y. Bengio 和 A. Courville. 深度学习。麻省理工学院出版社，2016。http://www.deeplearningbook.org。I. S. Gradshteyn 和 I. M. Ryzhik. 积分、级数与乘积表。学术出版社，伦敦，2000。N. Harvey、C. Liaw 和 A. Mehrabian. 分段线性神经网络近乎紧致的 VC 维数界。载于第 *30* 届学习理论会议论文集（*COLT 2017*），2017。R. Herbrich. 核分类器学习：理论与算法。麻省理工学院出版社，2002。A. Jacot、F. Gabriel 和 C. Hongler. 神经切线核：神经网络中的收敛与泛化。载于神经信息处理系统进展，第 8571–8580 页，2018。M. P. Knapp. 等差数列中角的正弦与余弦。《数学杂志》，82(5):371，2009。J. Lee、J. Sohl-dickstein、J. Pennington、R. Novak、S. Schoenholz 和 Y. Bahri. 深度神经网络即高斯过程。载于国际学习表征会议，2018。C. Li、L. Venturi 和 R. Xu. 学习用于分类与回归的核。*arXiv*预印本 *arXiv:1712.08597*，2017。

Y. Li 和 Y. Liang. 在结构化数据上通过随机梯度下降学习过参数化神经网络。载于神经信息处理系统进展第 *31* 卷：神经信息处理系统年度会议，2018年。R. Lin、H. Zhang 和 J. Zhang. 关于再生核巴拿赫空间：一般性定义与构造的统一框架。《数学学报（英文版）》，2022。A. G. d. G. Matthews、M. Rowland、J. Hron、R. E. Turner 和 Z. Ghahramani. 宽深神经网络中的高斯过程行为。*arXiv* 电子版，2018年。S. Mendelson. 关于统计学习理论的几点札记。载于 S. Mendelson 和 A. J. Smola 编，《机器学习高级讲座：*2002* 年 *2* 月 *11–22* 日在澳大利亚堪培拉举行的机器学习暑期学校修订讲义》，第 1–40 页。Springer Berlin Heidelberg，柏林、海德堡，2003。K.-R. M¨uller、S. Mika、G. R¨atsch、K. Tsuda 和 B. Sch¨olkopf. 基于核的学习算法导论。《*IEEE*神经网络汇刊》，12(2):181–198，2001 年 3 月。V. Nagarajan 和 J. Z. Kolter. 一致收敛可能无法解释深度学习中的泛化。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d´e Buc、E. Fox 和 R. Garnett 编，神经信息处理系统进展第 *32* 卷，第 11615–11626 页。Curran Associates, Inc.，2019年a。网址：http://papers.nips.cc/paper/9336-uniform-convergence-may-be-unable-to-explain-generalization-in-deep-learning.pdf。V. Nagarajan 和 Z. Kolter. 通过泛化抗噪性得到深度网络的确定性 PAC- 贝叶斯泛化界。载于国际学习表征会议（*ICLR*），2019 年b。R. M. Neal. 无限网络的先验，第 29–53 页。Springer，1996。B. Neyshabur、R. Tomioka 和 N. Srebro. 神经网络中基于范数的容量控制。载于学习理论会议论文集，第 1376–1401 页，2015。B. Neyshabur、S. Bhojanapalli、D. McAllester 和 N. Srebro. 探索深度学习中的泛化。载于第 *31* 届神经信息处理系统国际会议论文集，第 5949–5958 页，2017。B. Neyshabur、S. Bhojanapalli 和 N. Srebro. 用于神经网络谱归一化间隔界的PAC-贝叶斯方法。载于*ICLR* 论文集，2018。B. Neyshabur、Z. Li、S. Bhojanapalli、Y. LeCun 和 N. Srebro. 过参数化在神经网络泛化中的作用。载于*ICLR* 论文集，2019。R. Parhi 和 R. D. Nowak. 神经网络与脊样条的巴拿赫空间表示定理。《机器学习研究杂志》，22(43):1–40，2021。A. Rahimi 和 R. Benjamin. 随机厨房水槽的加权和：以随机化取代学习中的最小化。载于 D. Koller、D. Schuurmans、Y. Bengio 和 L. Bottou 编，神经信息处理系统进展第 *21* 卷，第 1313–1320 页。Curran Associates, Inc.，2009。K. Sanders. 以测度参数化的函数形式表示的神经网络：表示定理与逼近优势。硕士论文，埃因霍温理工大学，2020年。J. Shawe-Taylor 和 N. Cristianini. 模式分析的核方法。剑桥大学出版社，2004。A. J. Smola 和 B. Sch¨olkopf. 关于一种用于模式识别、回归、逼近和算子求逆的基于核的方法。《算法学》，22:211–231，1998。技术报告 1064，GMD First，1997 年 4 月。G. Song、H. Zhang 和 F. J. Hickernell. 带有 ℓ^1 范数的再生核巴拿赫空间。《应用与计算调和分析》，34(1):96–116，2013 年 1 月。B. K. Sriperumbudur、K. Fukumizu 和 G. R. Lanckriet. 希尔伯特空间与巴拿赫空间中的学习：一种测度嵌入视角。载于神经信息处理系统进展，第 1773–1781 页，2011。I. Steinwart 和 A. Christman. 支持向量机。Springer，2008。M. Unser. 用于反问题与机器学习的统一表示定理。《计算数学基础》，21(4):941–960，2021。Y. Xu 和 Q. Ye. 广义 Mercer 核与再生核巴拿赫空间。*arXiv*预印本 *arXiv:1412.8663*，2014。H. Zhang 和 J. Zhang. 将巴拿赫空间中的正则化学习视作优化问题：表示定理。《全球优化杂志》，54(2):235–250，2012 年 10 月。

H. Zhang、Y. Xu 和 J. Zhang。用于机器学习的再生核巴拿赫空间。《机器学习研究杂志》，10：2741–2775，2009。W. Zhou、V. Veitch、M. Austern、R. P. Adams 和 P. Orbanz。ImageNet 规模下非空泛化界：一种 PAC-贝叶斯压缩方法。载于国际学习表征会议（*ICLR*），2019。D. Zou 和 Q. Gu。对过参数化深度神经网络训练的改进分析。载于神经信息处理系统进展，第 32 卷，2019。D. Zou、Y. Cao、D. Zhou 和 Q. Gu。梯度下降优化过参数化深度 ReLU 网络。机器学习，109(3)：467–492，2020。

推导与证明。

本补充材料旨在呈现与本文相关的所有推导与证明，以及任何可能有用的补充材料和说明。

我们假设一个全连接的、 D 层前馈神经网络 $\mathbf{f}:(\mathbb{X}\subseteq\mathbb{R}^n)\to(\mathbb{Y}\subseteq\mathbb{R}^m)$，其各层宽度为$H^{[0]},H^{[1]},\ldots,H^{[D-1]}$，其中 $H^{[D-1]}=m$，并定义 $H^{[-1]}=n$。我们假设第 $j\in\mathbb{N}_D$ 层（全文采用约定 $j\in\mathbb{N}_D$）由使用相同激活函数 $\tau^{[j]}$ 的神经元组成： $\mathbb{R}\to\mathbb{R}$。该网络递归定义如下:

$$\begin{aligned}&\mathbf{f}(\mathbf{x})=\mathbf{x}^{[D]}\in\mathbb{R}^{H^{[D-1]}}\\&\mathbf{x}^{[j+1]}=\tau^{[j]}\left(\tilde{\mathbf{x}}^{[j]}\right)\in\mathbb{R}^{H^{[j]}}\\&\quad\tilde{\mathbf{x}}^{[j]}=\tfrac{1}{\sqrt{H^{[j]}}}\mathbf{W}^{[j]\mathrm{T}}\mathbf{x}^{[j]}+\alpha^{[j]}\mathbf{b}^{[j]}\in\mathbb{R}^{H^{[j]}}\qquad\forall j\in\mathbb{N}_D\\&\mathbf{x}^{[0]}=\mathbf{x}\in\mathbb{X}\subset\mathbb{R}^{H^{[-1]}}\ (H^{[-1]}=n)\end{aligned}\tag{16}$$

其中 $\mathbf{W}^{[j]}\in\mathbb{R}^{H^{[j-1]}\times H^{[j]}}$ 和 $\mathbf{b}^{[j]}\in\mathbb{R}^{H^{[j]}}$ 分别为权重和偏置，而 $\alpha^{[j]}\in\mathbb{R}_+$ 是我们稍后将使用的常数。我们将具有形式 (16) 的神经网络集合定义如下:

$$\mathcal{F}=\{\mathbf{f}:\mathbb{R}^n\to\mathbb{R}^m|\,\mathbf{f}\text{ has form (16) given }\mathbf{W}\in\mathbb{W}\}\tag{17}$$

其中 $\mathbf{W}$ 概括了 (16) 中的权重和偏置:

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

通常，神经网络训练的目标是给定一个训练集，找到权重和偏置，使真实训练标签与网络预测之间的某种误差度量最小化。为简单起见，假设我们希望最小化:

$$\mathbf{f}^\star=\operatorname*{argmin}_{\mathbf{f}\in\mathcal{F}}\textstyle\sum_k E\left(\mathbf{x}^{\{k\}},\mathbf{y}^{\{k\}},\mathbf{f}\left(\mathbf{x}^{\{k\}}\right)\right)\tag{18}$$

或者，等价地:

$$\mathbf{W}^\star=\operatorname*{argmin}_{\mathbf{W}\in\mathbb{W}}\textstyle\sum_k E\left(\mathbf{x}^{\{k\}},\mathbf{y}^{\{k\}},\mathbf{f}_{\mathbf{W}}\left(\mathbf{x}^{\{k\}}\right)\right)\tag{19}$$

其中 $\mathbf{f}_{\mathbf{W}}$ 是形式如 (16) 的网络，其权重和偏置为 $\mathbf{W}$，$\mathbb{D}=\{(\mathbf{x}^{\{k\}},\mathbf{y}^{\{k\}})\in\mathbb{X}\times\mathbb{Y}:\ k\in\mathbb{N}_N\}$ 是一个训练集（我们始终采用约定 $k\in\mathbb{N}_N$），而 $E:\mathbb{X}\times\mathbb{Y}\times\mathbb{R}^m\to\mathbb{R}$ 是定义网络用途的误差函数。

我们作出以下假设:

1. 输入空间：我们假设 $\mathbb{X}=[-M^{[-1]},M^{[-1]}]^n$。
2. 误差函数：我们假设误差函数 $E:\mathbb{X}\times\mathbb{Y}\times\mathbb{R}^m\to\mathbb{R}$ 是 $\mathcal{C}^1$ 且关于其第三个参数满足 L_E-利普希茨条件。
3. 激活函数：我们假设激活函数 $\tau^{[j]}:\mathbb{R}\to[-M^{[j]},M^{[j]}]^{H^{[j]}}$有界且 $\mathcal{C}^\infty$，并且 $\tau^{[j]}$ 具有幂级数表示，其收敛域（ROC）至少在 z 周围包含 $\rho^{[j]}\in\mathbb{R}_+$，对所有$z\in\mathbb{R}$, $j\in\mathbb{N}_D$均成立。
4. 权重非平凡性：我们假设在训练期间的任何时刻，对于所有 $j\in\mathbb{N}_D$，$\mathbf{W}^{[j]}\neq\mathbf{0}$。[10]
5. 权重初始化：我们假设初始时 $W^{[j]}_{i_j,\,i_{j+1}}$，$b^{[j]}_{i_{j+1}}\sim\mathcal{N}(0,1)$（LeCun 初始化）。
6. 训练：我们假设网络使用学习率为 $\eta\in\mathbb{R}_+$ 的反向传播进行训练。

A.1 反向传播回顾

现在简要回顾反向传播训练，它是对权重和偏置 $\mathbf{W}$ 执行梯度下降以求解 (19) 的系统化实现。考虑一次训练迭代：我们从初始权重和偏置 $\mathbf{W}_{\mathcal{O}}\in\mathbb{W}$ 开始，并计算权重步长 $\mathbf{W}^{\boxplus}_{\Delta}$，使得此次迭代后，$\mathbf{W}=\mathbf{W}_{\mathcal{O}}+\mathbf{W}^{\boxplus}_{\Delta}$，其中:

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

将迭代前网络对输入$\mathbf{x}$ 的激活记为:

$$\begin{aligned}&\mathbf{f}_{\mathcal{O}}(\mathbf{x})=\mathbf{x}^{[D]}_{\mathcal{O}}\\&\mathbf{x}^{[j+1]}_{\mathcal{O}}=\tau^{[j]}\left(\tilde{\mathbf{x}}^{[j]}_{\mathcal{O}}\right)\\&\quad\tilde{\mathbf{x}}^{[j]}_{\mathcal{O}}=\tfrac{1}{\sqrt{H^{[j]}}}\mathbf{W}^{[j]\mathrm{T}}_{\mathcal{O}}\mathbf{x}^{[j]}_{\mathcal{O}}+\alpha^{[j]}\mathbf{b}^{[j]}_{\mathcal{O}}\\&\mathbf{x}^{[0]}_{\mathcal{O}}=\mathbf{x}\in\mathbb{X}\end{aligned}\tag{20}$$

[10]请注意，不满足这一要求的网络，其输出为常数且与输入 $\mathbf{x}$ 无关。我们认为这并非一个限制性假设，因为随机初始化的网络在使用典型训练集进行训练时，几乎不可能达到这种状态。

对于所有层 $j \in \mathbb{N}_{D^\circ}$ 层 $D-1$ 的反向传播迭代是一个学习率为 η 的梯度下降步，即:

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

其中 $\tau^{(r)}(z) = \frac{\partial^r}{\partial^r z}\tau(z)$，并且:

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

随后:

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

其中:

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

依此类推，遍历所有层。综上，对于所有 $j \in \mathbb{N}_D$ （并采用 MATLAB 记号 $\mathbf{A}_{:i}$ 表示矩阵$\mathbf{A}$的第 i 列）:

$$\begin{aligned}
\mathbf{W}_{\Delta:i_{j+1}}^{[j]\boxplus} &= -\eta \frac{1}{\sqrt{H^{[D-1]}H^{[D-2]}\ldots H^{[j+1]}}} \sum_k \gamma_{\mathcal{O}i_{j+1}}^{\{k\}[j]} \frac{1}{\sqrt{H^{[j]}}} \mathbf{x}_{\mathcal{O}}^{\{k\}[j]} \\
b_{\Delta:i_{j+1}}^{[j]\boxplus} &= -\eta \frac{1}{\sqrt{H^{[D-1]}H^{[D-2]}\ldots H^{[j+1]}}} \sum_k \gamma_{\mathcal{O}i_{j+1}}^{\{k\}[j]} \alpha^{[j]}
\end{aligned} \tag{21}$$

其中，递归地:

$$\begin{aligned}
\gamma_{\mathcal{O}i_j}^{\{k\}[j-1]} &= \sum_{i_{j+1}} \gamma_{\mathcal{O}i_{j+1}}^{\{k\}[j]} W_{\mathcal{O}i_j,i'_{j+1}}^{[j]} \tau^{[j-1](1)}\left(\tilde{x}_{\mathcal{O}i_j}^{\{k\}[j-1]}\right) \\
\gamma_{\mathcal{O}i_D}^{\{k\}[D-1]} &= \nabla_{i_D} E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}\left(\mathbf{x}^{\{k\}}\right)\right) \tau^{[D-1](1)}\left(\tilde{x}_{\mathcal{O}i_D}^{\{k\}[D-1]}\right)
\end{aligned} \tag{22}$$

B 神经网络的对偶形式步骤

我们的首要目标是将训练迭代后的神经网络改写为:

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

其中，$\mathbf{f}_{\mathcal{O}} : (\mathbb{X} \subset \mathbb{R}^n) \to (\mathbb{Y} \subset \mathbb{R}^m)$ 表示迭代前的神经网络，而 $\mathbf{f}_{\Delta} : (\mathbb{X} \subset \mathbb{R}^n) \to \mathbb{R}^m$ 表示由于本次迭代中权重和偏置的变化 $\mathbf{W}_{\Delta} \in \mathbb{W}_{\mathcal{O}}$ 所导致的网络行为变化，因此:

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

其中:

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

是特征映射;

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

且 $\langle\cdot,\cdot\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}}$： $\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}\to\mathbb{R}^m$ 是双线性形式：

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

此外，我们旨在证明特征映射完全由以下因素决定：

1. 网络的结构，即层数、各层的宽度和激活函数 s.

2. 迭代之前的权重和偏置 $\mathbf{W}_\mathcal{O}\in\mathbb{W}$ 。

随后，我们将利用这一点为权重特征空间 $\mathcal{W}_\mathcal{O}$ 和数据特征空间 $\mathcal{X}_\mathcal{O}$构造合适的范数，从而证明$\mathbf{f}_\Delta$ 属于再生核巴拿赫空间，并分析神经网络的复杂度和收敛性。

B.1 预备知识：幂级数记号

这里采用的方法很直接——我们构造网络的幂级数展开，不进行截断，并重新排列各项，使所得求和分离成所需的形式。为了简化这一过程，我们定义：

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

注意到这关于逐元素（Hadamard）乘积是线性的，因此：

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

利用此记号，可以很容易地得到以下多维泰勒展开的简洁表达式（使用光滑性假设），该展开针对 $\tau^{[j]}$： $\mathbb{R}\to\mathbb{R}$ 关于 $z\in\mathbb{R}$，其中 $\mathbf{c},\mathbf{c}'\in\mathbb{R}^p$, $|\langle\mathbf{c},\mathbf{c}'\rangle|\le\rho^{[j]}$:

$$\tau^{[j]}\left(z+\langle\mathbf{c},\mathbf{c}'\rangle\right)=\tau^{[j]}(z)+\left\langle\boldsymbol{\varrho}\left(\mathbf{g}^{[j]}(z),\mathbf{c}\right),\boldsymbol{\varrho}\left(\mathbf{1}_\infty,\mathbf{c}'\right),\right\rangle \tag{23}$$

其中， $\tau^{[j]}$ 在 z 处的导数表示为：

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

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

B.2 神经网络的特征映射

我们将迭代前后的运算以及迭代带来的变化表示为：

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

$$\Uparrow$$

$$\begin{pmatrix}\mathbf{x}_\Delta^{[j+1]}=\tau^{[j]}\left(\tilde{\mathbf{x}}_\mathcal{O}^{[j]}+\tilde{\mathbf{x}}_\Delta^{[j]}\right)-\tau^{[j]}\left(\tilde{\mathbf{x}}_\mathcal{O}^{[j]}\right)\\ \tilde{\mathbf{x}}_\Delta^{[j]}=\frac{1}{\sqrt{H^{[j]}}}\mathbf{W}_\Delta^{[j]\mathrm{T}}\left(\mathbf{x}_\mathcal{O}^{[j]}+\mathbf{x}_\Delta^{[j]}\right)+\frac{1}{\sqrt{H^{[j]}}}\mathbf{W}_\mathcal{O}^{[j]\mathrm{T}}\mathbf{x}_\Delta^{[j]}+\alpha^{[j]}\mathbf{b}_\Delta^{[j]}\end{pmatrix} \tag{24}$$

我们的目标是将 $\mathbf{f}_\Delta$： $\mathbb{X}\to\mathbb{R}^m$ 写为：

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

其中 $\boldsymbol{\Phi}_\mathcal{O}$： $\mathbb{X}\to\mathcal{X}_\mathcal{O}$ 和 $\boldsymbol{\Psi}_\mathcal{O}$： $\mathbb{W}_\mathcal{O}\to\mathcal{W}_\mathcal{O}$ 是特征映射，且 $\langle\boldsymbol{\Xi},\boldsymbol{\Omega}\rangle_{\mathcal{X}_\mathcal{O}\times\mathcal{W}_\mathcal{O}}=\operatorname{diag}(\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Omega})$ 是一个双线性形式。从网络的第一层开始，不难看出：

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

其中 $\mu_{i_1}^{[0]} \in \mathbb{R}_+$ 是我们称为缩放因子的参数，而 $\mathbf{A}_{:l}$ 表示矩阵 $\mathbf{A}$ 的第 l 列。缩放因子在我们对 $\mathbf{f}$ 的表示中总会相互抵消，但稍后我们在特征空间上定义范数时会很重要。按照式 (23) 使用幂级数展开，如果 $\|\tilde{}\mathbf{x}_\Delta^{[0]}\|_\infty \leq \rho^{[0]}$:

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

因此:

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

其中:

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

转到第二层，我们看到:

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

其中 $\mu_{i_2}^{[1]} \in \mathbb{R}_+$ 为缩放因子， $\omega_{i_1}^{[1]}, \widetilde{\omega}_{i_1}^{[1]}{}_{,i_2} \in \mathbb{R}^+$ 为影子权重（注意，影子权重和缩放因子在内积中都会相互抵消）。进行泰勒展开并利用 $\boldsymbol{\varrho}$ 的性质，我们看到，只要 $\|\tilde{}\mathbf{x}_\Delta^{[1]}\|_\infty \leq \rho^{[1]}$:

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

其中:

$$\begin{aligned}\boldsymbol{\Phi}_{\mathcal{O}:i_2}^{[1]}(\mathbf{x}) &= \boldsymbol{\varrho}\left(\mathbf{g}^{[1]}\left(\tilde{x}_{\mathcal{O}i_2}^{[1]}\right), \mu_{i_2}^{[1]}\begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[1]} \\ \frac{1}{\sqrt{H^{[1]}}}\mathbf{x}_{\mathcal{O}}^{[1]} \end{bmatrix} \\ \left[\frac{1}{\widetilde{\omega}_{i_1,i_2}^{[1]}}\frac{\omega_{i_1}^{[1]}W_{\mathcal{O}i_1,i_2}^{[1]}}{\sqrt{H^{[1]}}}\boldsymbol{\Phi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{x})\right]_{i_1} \\ \left[\frac{\omega_{i_1}^{[1]}}{\sqrt{H^{[1]}}}\boldsymbol{\Phi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{x})\right]_{i_1} \end{bmatrix}\right) \\ \boldsymbol{\Psi}_{\mathcal{O}:i_2}^{[1]}(\mathbf{W}_\Delta) &= \boldsymbol{\varrho}\left(\mathbf{1}_\infty, \frac{1}{\mu_{i_2}^{[1]}}\begin{bmatrix} \begin{bmatrix} \sqrt{2}b_{\Delta i_2}^{[1]} \\ \mathbf{W}_{\Delta:i_2}^{[1]} \end{bmatrix} \\ \left[\widetilde{\omega}_{i_1,i_2}^{[1]}\frac{1}{\omega_{i_1}^{[1]}}\boldsymbol{\Psi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{W}_\Delta)\right]_{i_1} \\ \left[\frac{W_{\Delta i_1,i_2}^{[1]}}{\omega_{i_1}^{[1]}}\boldsymbol{\Psi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{W}_\Delta)\right]_{i_1} \end{bmatrix}\right)\end{aligned}$$

继续来看，不难发现，对于任意 $j \in \mathbb{N}_D$，如果:

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

那么，取缩放因子 $\mu_{i_{j+1}}^{[j]} \in \mathbb{R}_+$ 和影子权重 $\omega_{i_j}^{[j]}$，$\widetilde{\omega}_{i_j,\ i_{j+1}}^{[j]} \in \mathbb{R}_+$：

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

因此，若 $\|\tilde{\mathbf{x}}_{\Delta}^{[j]}\|_\infty \leq \rho^{[j]}$，则有：

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

其中：

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

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

因此，通过归纳法，只要 $\|\tilde{\mathbf{x}}_{\Delta}^{[j]}\|_\infty \leq \rho^{[j]}\ \forall j$：

$$\begin{aligned}
\mathbf{f}(\mathbf{x}) &= \mathbf{f}_{\mathcal{O}}(\mathbf{x}) + \mathbf{f}_{\Delta}(\mathbf{x}) \\
\mathbf{f}_{\Delta}(\mathbf{x}) &= \langle \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}), \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}
\end{aligned} \tag{25}$$

其中：

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

并递归地进行：

$$\mathbf{\Phi}_{\mathcal{O}:i_{j+1}}^{[j]}(\mathbf{x}) = \begin{cases} \boldsymbol{\varrho} \left(\mathbf{g}^{[j]} \left(\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]} \right), \mu_{i_{j+1}}^{[j]} \begin{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[j]} \\ \frac{1}{\sqrt{H^{[j]}}}\mathbf{x}_{\mathcal{O}}^{[j]} \end{bmatrix} \\ \left[\frac{1}{\widetilde{\omega}_{i_j i_{j+1}}^{[j]}} \frac{\omega_{i_j}^{[j]} W_{\mathcal{O}i_j,i_{j+1}}^{[j]}}{\sqrt{H^{[j]}}} \mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x}) \right]_{i_j} \\ \left[\frac{\omega_{i_j}^{[j]}}{\sqrt{H^{[j]}}} \mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x}) \right]_{i_j} \end{bmatrix} \right) & \text{if } j > 0 \\ \boldsymbol{\varrho} \left(\mathbf{g}^{[0]} \left(\tilde{x}_{\mathcal{O}i_1}^{[0]} \right), \mu_{i_1}^{[0]} \begin{bmatrix} \frac{1}{\sqrt{2}}\alpha^{[0]} \\ \frac{1}{\sqrt{H^{[0]}}} \frac{1}{\sqrt{2}} \mathbf{x}_{\mathcal{O}} \end{bmatrix} \right) & \text{if } j = 0 \end{cases} \quad \forall j \in \mathbb{N}_D$$

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

(26)

其中缩放因子为 $\mu_{i_{j+1}}^{[j]} \in \mathbb{R}_+$，影子权重为 $\omega_{i_j}^{[j]}, \widetilde{\omega}_{i_j}^{[j]}, i_{j+1} \in \mathbb{R}_+, \forall j \in \mathbb{N}_D$。

总而言之，我们已将神经网络分解为所需的形式 (25)，其包含以下组成部分：

- 一项 $\mathbf{f}_{\mathcal{O}}(\mathbf{x})$，即迭代之前的网络在 $\mathbf{x}$ 上的取值。
- 一个特征映射 $\mathbf{\Phi}_{\mathcal{O}} : \mathbb{X} \to \mathcal{X}_{\mathcal{O}} = \text{span}(\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})) \subset \mathbb{R}^{\infty \times m}$，从数据输入空间映射到数据特征空间，且仅取决于网络的结构（层数、各层宽度和神经元类型），以及迭代之前的权重和偏置 $\mathbf{W}_{\mathcal{O}}$。
- 一个特征映射 $\mathbf{\Psi}_{\mathcal{O}} : \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} = \text{span}(\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})) \subset \mathbb{R}^{\infty \times m}$，从权重步输入空间映射到权重步特征空间，且仅取决于网络的结构（层数及各层宽度）。

B.3 特征映射的稠密性

本节将证明特征映射的一项关键性质，这项性质对于证明 $\mathbf{f}_\Delta$ 属于再生核巴拿赫空间是必需的。作为预备，我们先展示非平凡性假设的一个推论，即：

引理3。 设$\mathbf{f}_{\mathcal{O}} \in \mathcal{F}$ 是满足非平凡性假设的神经网络。则$\mathbf{f}_\Delta(\mathbf{x})$ 随$\mathbf{x}$非平凡地变化（即不是常数）。

证明。 回顾如下：

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

我们想知道在什么条件下，$\tilde{}\mathbf{x}_\Delta^{[D-1]}$（从而$\mathbf{f}_\Delta(\mathbf{x})$是一个与$\mathbf{x}$无关的常数。转而考虑 $\tilde{}\mathbf{x}_\Delta^{[j]}$，上式中唯一依赖$\mathbf{x}$ 的部分是$\mathbf{x}_\Delta^{[j]}$，因此，$\tilde{}\mathbf{x}_\Delta^{[j]}$ 为常数，当且仅当$\mathbf{W}_\Delta^{[j]} = -\mathbf{W}_{\mathcal{O}}^{[j]}$ 或$\mathbf{x}_\Delta^{[j]}$ 是与$\mathbf{x}$无关的常数。由此递归可知，$\tilde{}\mathbf{x}_\Delta^{[D-1]}$ 为常数（与$\mathbf{x}$无关），当且仅当$\mathbf{W}_\Delta^{[j]} = -\mathbf{W}_{\mathcal{O}}^{[j]}$ 对某个 $j \in \mathbb{N}_D$成立。然而，请回想我们的非平凡性假设明确排除了这种情况（这对应于迭代后的平凡神经网络），因此对所有 $j \in \mathbb{N}_D$，$\mathbf{W}_\Delta^{[j]} \neq -\mathbf{W}_{\mathcal{O}}^{[j]}$，这足以证明所需结论。

□

完成这项准备工作后，我们现在转向本节的主要结果：[11]

引理4 (特征映射的稠密性)。 $\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})$ 的线性张成在 $\mathcal{X}_{\mathcal{O}}$ 中关于 $\langle\cdot,\cdot\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$稠密；而 $\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})$ 的线性张成在 $\mathcal{W}_{\mathcal{O}}$ 中关于 $\langle\cdot,\cdot\rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}}$稠密，其中 $\langle \mathbf{\Omega}\ \mathbf{\Xi} \rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}} = \langle \mathbf{\Xi}, \mathbf{\Omega} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}^{\mathrm{T}}$。

证明。 根据定义 [Lin等人， 2022]，$\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})$的线性张成span($\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})$)在 $\mathcal{X}_{\mathcal{O}}$ 中关于 $\langle\cdot,\cdot\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}$稠密，当且仅当对于任意$\mathbf{\Omega}_\Delta \in \mathcal{W}_{\mathcal{O}}$，以下陈述成立：

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

这意味着$\mathbf{\Omega}_\Delta = \mathbf{0}$。根据定义 $\mathcal{W}_{\mathcal{O}} = \text{span}(\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}}))$，因此对于某些 $\tilde{\alpha}^{\{0\}}, \tilde{\alpha}^{\{1\}}, \ldots \in \mathbb{R}$，有$\mathbf{\Omega}_\Delta = \sum_l \tilde{\alpha}^{\{l\}} \mathbf{W}_\Delta^{\{l\}}$，其中 $\tilde{\alpha}^{\{0\}}, \tilde{\alpha}^{\{1\}}, \ldots \in \mathbb{R}$，$\mathbf{W}_\Delta^{\{0\}}, \mathbf{W}_\Delta^{\{1\}}, \ldots \in \mathbb{W}_{\mathcal{O}}$，因此：

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

其中，我们用$\mathbf{f}_\Delta^{\{l\}}$表示$\mathbf{f}$因权重步长$\mathbf{W}_\Delta^{\{l\}}$而产生的变化。根据前面的结论，对于所有$\mathbf{x} \in \mathbb{X}$，当且仅当 $\tilde{\alpha}^{\{0\}} = \tilde{\alpha}^{\{l\}} = \ldots = 0$时，$\mathbf{0}$，因此$\mathbf{\Omega}_\Delta = \mathbf{0}$，这就证明了所需的稠密性。类似地，如果对于任意 $\mathbf{\Xi}_\Delta \in \mathcal{X}_{\mathcal{O}}$，以下观察成立：那么 $\text{span}(\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}}))$ 关于 $\langle\cdot,\cdot\rangle_{\mathcal{W}_{\mathcal{O}} \times \mathcal{X}_{\mathcal{O}}}$ 在$\mathcal{W}_{\mathcal{O}}$ 中稠密：

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

这意味着 $\mathbf{\Xi}_\Delta = \mathbf{0}$。根据定义，$\mathcal{X}_{\mathcal{O}} = \text{span}(\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X}))$，因此对于某个 $\tilde{\beta}^{\{0\}}, \tilde{\beta}^{\{1\}}, \ldots \in \mathbb{R}$，有 $\mathbf{\Xi}_\Delta = \sum_l \tilde{\beta}^{\{l\}} \mathbf{x}^{\{l\}}$，其中$\mathbf{x}^{\{0\}}, \mathbf{x}^{\{1\}}, \ldots \in \mathbb{X}$，因此：

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

但同样，根据我们的初步结果，对于所有 $\mathbf{W}_\Delta \in \mathbb{W}_{\mathcal{O}}$，当且仅当 $\tilde{\beta}^{\{0\}} = \tilde{\beta}^{\{1\}} = \ldots = 0$ 时，$\mathbf{0}$ 成立，因此 $\mathbf{\Xi}_\Delta = \mathbf{0}$，证明完毕。 □

[11] 参见 [Lin 等人 2022]，了解此处所用稠密性的讨论。

注释 1。 有人可能会提出异议，认为非平凡性假设是任意的，而且无论如何都无法保证成立，因此上述结果值得怀疑。然而，考虑到神经网络通常以非平凡的权重和偏置进行初始化（更准确地说，从例如均匀分布或正态分布中抽取权重和偏置时，任意一层的权重和偏置全为零的概率为 0），并且训练数据通常本身也是非平凡的（也就是说，目标值会变化，而不仅仅是输入会变化，因此就固定输出所带来的潜在最优性而言，平凡情形并无特殊意义），权重和偏置在有限次反向传播迭代中回到平凡状态（至少一层中的所有权重和偏置均为 0）的可能性似乎极低。事实上，我们推测应该存在这样一个结果：在随机权重初始化，且噪声影响的数据取自某个分布的条件下，训练期间遇到权重和偏置构成的平凡网络的概率恰为 0。但遗憾的是，我们未能得到相应的证明，因此目前这仍只是推测。

C 诱导核与诱导范数

在上一节中，我们证明了，神经网络经过单次迭代（权重和偏置发生阶跃变化）后，其运算可以写成：

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

其中，$\mathbf{f}_{\mathcal{O}}:\mathbb{X}\to\mathbb{Y}$ 是迭代前的神经网络，而$\mathbf{f}_{\Delta}:\mathbb{X}\to\mathbb{Y}_{\mathcal{O}}$ 是由于本次迭代中权重和偏置发生变化$\mathbf{W}_{\Delta}\in\mathbb{W}_{\mathcal{O}}$ 而导致的网络行为变化。此外，我们还证明：

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

其中：

$$\begin{array}{c}\mathbf{\Phi}_{\mathcal{O}}:\mathbb{X}\quad\to\mathcal{X}_{\mathcal{O}}\;=\operatorname{span}\left(\mathbf{\Phi}_{\mathcal{O}}\left(\mathbb{X}\right)\right)\quad\subset\mathbb{R}^{\infty\times m}\\ \mathbf{\Psi}_{\mathcal{O}}:\mathbb{W}_{\mathcal{O}}\to\mathcal{W}_{\mathcal{O}}=\operatorname{span}\left(\mathbf{\Psi}_{\mathcal{O}}\left(\mathbb{W}_{\mathcal{O}}\right)\right)\subset\mathbb{R}^{\infty\times m}\end{array}$$

是特征映射，而 $\left\langle \cdot,\cdot\right\rangle _{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}:\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}\to\mathbb{R}^{m}$ 是双线性形式：

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

在本节中，我们将为 $\mathbb{X}_{\mathcal{O}}$ 和 $\mathbb{W}_{\mathcal{O}}$赋予（诱导）核和范数，从而能够约束神经网络行为变化的空间：

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

使用 H¨older 不等式：

$$\left\|\mathbf{f}_{\Delta}\left(\mathbf{x}\right)\right\|_{2}\leq\left\|\mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right)\right\|_{\mathcal{X}_{\mathcal{O}}}\left\|\mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\|_{\mathcal{W}_{\mathcal{O}}}$$

C.1 诱导核

我们开 始利用特征映射和核技巧，在 $\mathbb{X}$ 和 $\mathbb{W}_{\mathcal{O}}$ 上诱导核，具体而言 ly:

$$\begin{array}{c}\mathbf{K}_{\mathcal{X}_{\mathcal{O}}}\left(\mathbf{x},\mathbf{x}'\right)=\left\langle \mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}\right),\mathbf{\Phi}_{\mathcal{O}}\left(\mathbf{x}'\right)\right\rangle _{\mathcal{X}_{\mathcal{O}}\times\mathcal{X}_{\mathcal{O}}}\\ \mathbf{K}_{\mathcal{W}_{\mathcal{O}}}\left(\mathbf{W}_{\Delta},\mathbf{W}'_{\Delta}\right)=\left\langle \mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right),\mathbf{\Psi}_{\mathcal{O}}\left(\mathbf{W}'_{\Delta}\right)\right\rangle _{\mathcal{W}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}\end{array}$$

其中我们定义双线性形式：

$$\begin{array}{c}\left\langle \mathbf{\Xi},\mathbf{\Xi}'\right\rangle _{\mathcal{X}_{\mathcal{O}}\times\mathcal{X}_{\mathcal{O}}}=\mathbf{\Xi}^{\mathrm{T}}\mathbf{\Xi}'\\ \left\langle \mathbf{\Omega},\mathbf{\Omega}'\right\rangle _{\mathcal{W}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}=\mathbf{\Omega}^{\mathrm{T}}\mathbf{\Omega}'\end{array}$$

矩阵值核 $\mathbf{K}_{\mathcal{X}_{\mathcal{O}}}$ 和 $\mathbf{K}_{\mathcal{W}_{\mathcal{O}}}$ 根据构造均为正定（Mercer）核。除了理论用途外，这些核还可能用于支持向量机（SVM）、高斯过程（GP）或类似的基于核的方法，分别衡量 $\mathbb{X}$ 和 $\mathbb{W}_{\mathcal{O}}$ 上的相似度。

对于所有层 $j\in\mathbb{N}_{D}$，我们定义：

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

回顾式 (26)，我们看到，对于所有 $j \in \mathbb{N}_D \backslash \{0\}$:

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

$$K^{[j]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) = \sum_l \varrho_l \left(\mathbf{1}_\infty, \frac{1}{\mu^{[j]}_{i_{j+1}} \mu^{[j]}_{i'_{j+1}}} \begin{bmatrix} \begin{bmatrix} 2 b^{[j]}_{\Delta i_{j+1}} b'^{[j]}_{\Delta i'_{j+1}} \\ \mathbf{W}^{[j]}_{\Delta: i_{j+1}} \odot \mathbf{W}'^{[j]}_{\Delta: i'_{j+1}} \end{bmatrix} \\ \left[\frac{\widetilde{\omega}^{[j]}_{ij, i_{j+1}} \widetilde{\omega}^{[j]}_{ij, i'_{j+1}}}{\omega^{[j]2}_{ij}} \mathbf{\Psi}^{[j-1]}_{\mathcal{O}: i_j}(\mathbf{W}_\Delta) \odot \mathbf{\Psi}^{[j-1]}_{\mathcal{O}: i_j}(\mathbf{W}'_\Delta) \right]_{i_j} \\ \left[\frac{W^{[j]}_{\Delta ij, i_{j+1}} W'^{[j]}_{\Delta ij, i'_{j+1}}}{\omega^{[j]2}_{ij}} \mathbf{\Psi}^{[j-1]}_{\mathcal{O}: i_j}(\mathbf{W}_\Delta) \odot \mathbf{\Psi}^{[j-1]}_{\mathcal{O}: i_j}(\mathbf{W}'_\Delta) \right]_{i_j} \end{bmatrix} \right)$$

and:

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

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

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

and:

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

我们定义:

$$\begin{aligned} \theta(\zeta) &= \sum_{l=1}^{\infty} \zeta^l = \tfrac{\zeta}{1-\zeta} \quad \forall \zeta \in (-1, 1) \\ \sigma^{[j]}_{z, z'}(\zeta) &= \sum_{l=1}^{\infty} \left(\tfrac{1}{l!}\right)^2 \tau^{[j](l)}(z) \tau^{[j](l)}(z') \zeta^l \end{aligned} \tag{27}$$

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

$$\theta^{-1}(\zeta) = \tfrac{\zeta}{1+\zeta} \quad \forall \zeta \in (1, \infty) \tag{28}$$

注意，$\sigma^{[j]}_{z}{}_{,\ z'}$ 是由激活函数在 z 附近的幂级数表示推导而来，推导时去掉了常数项；若 $z = z'$，符号函数也一并去掉，从而使 $\sigma^{[j]}_{z'},{}_{z}$ 成为递增函数。[12] 根据此定义和观察，只要自变量仍处于所有相关级数的收敛域内，我们就有:

$$\begin{aligned} \mathbf{K}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') &= \mathbf{K}^{[D-1]}_{\mathcal{X}_{\mathcal{O}}}(\mathbf{x}, \mathbf{x}') \\ \mathbf{K}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) &= \mathbf{K}^{[D-1]}_{\mathcal{W}_{\mathcal{O}}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) \end{aligned} \tag{29}$$

[12]顺便提一下，激活函数 $\tau^{[j]}$ 在任意 z 处的幂级数表示都代表同一个底层函数，只是减去了 $\tau^{[j]}(0)$，因此原则上我们可以从其中一个表示出发，通过解析延拓在各处重构 $\tau^{[j]}$。遗憾的是，对于 $\sigma^{[j]}_{z\ z'}$，情况并非如此，这在此处或许令人遗憾。

其中，递归地定义为 $\forall j \in \mathbb{N}_D$:

$$K^{[j]}_{\mathcal{X}\mathcal{O} i_{j+1},i'_{j+1}}(\mathbf{x},\mathbf{x}') = \begin{cases} \sigma^{[j]}_{\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}},\tilde{x}'^{[j]}_{\mathcal{O}i'_{j+1}}}\left(\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}\left(\begin{array}{l}\left(\frac{1}{2}\alpha^{[j]2}+\frac{1}{H^{[j]}}\left\langle \mathbf{x}^{[j]}_{\mathcal{O}},\mathbf{x}'^{[j]}_{\mathcal{O}}\right\rangle\right)+\dots\\ \dots\sum\limits_{i_j}\left(\frac{W^{[j]}_{\mathcal{O}i_j,i_{j+1}}W^{[j]}_{\mathcal{O}i_j,i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j,i_{j+1}}\widetilde{\omega}^{[j]}_{i_j,i'_{j+1}}}+1\right)\frac{\omega^{[j]2}_{i_j}}{H^{[j]}}K^{[j-1]}_{\mathcal{X}\mathcal{O}i_j,i_j}(\mathbf{x},\mathbf{x}')\end{array}\right)\right) & \text{if } j>0\\ \sigma^{[0]}_{\tilde{x}^{[0]}_{\mathcal{O}i_1},\tilde{x}'^{[0]}_{\mathcal{O}i'_1}}\left(\mu^{[0]}_{i_1}\mu^{[0]}_{i'_1}\left(\frac{1}{2}\alpha^{[0]2}+\frac{1}{2}\frac{1}{H^{[0]}}\left\langle \mathbf{x}^{[0]}_{\mathcal{O}},\mathbf{x}'^{[0]}_{\mathcal{O}}\right\rangle\right)\right) & \text{if } j=0\end{cases}$$

$$K^{[j]}_{\mathcal{W}\mathcal{O} i_{j+1},i'_{j+1}}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta) = \begin{cases} \theta\left(\frac{1}{\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}}\left(\begin{array}{l}\left(2b^{[j]}_{\Delta i_{j+1}}b'^{[j]}_{\Delta i'_{j+1}}+\left\langle \mathbf{W}^{[j]}_{\Delta:i_{j+1}},\mathbf{W}'^{[j]}_{\Delta:i'_{j+1}}\right\rangle\right)+\dots\\ \dots\sum\limits_{i_j}\left(\widetilde{\omega}^{[j]}_{i_j,i_{j+1}}\widetilde{\omega}^{[j]}_{i_j,i'_{j+1}}+W^{[j]}_{\Delta i_j,i_{j+1}}W'^{[j]}_{\Delta i_j,i'_{j+1}}\right)\frac{K^{[j-1]}_{\mathcal{W}\mathcal{O}i_j,i_j}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}}\end{array}\right)\right) & \text{if } j>0\\ \theta\left(\frac{1}{\mu^{[0]}_{i_1}\mu^{[0]}_{i'_1}}\left(2b^{[0]}_{\Delta i_1}b'^{[0]}_{\Delta i'_1}+2\left\langle \mathbf{W}^{[0]}_{\Delta:i_1},\mathbf{W}'^{[0]}_{\Delta:i'_1}\right\rangle\right)\right) & \text{if } j=0\end{cases} \tag{30}$$

C.2 诱导核梯度

在论文后文中，我们将需要诱导核的梯度。回顾式（29）、（30），并应用链式法则，可得:

$$\begin{aligned}\nabla_{\mathbf{x}}\mathbf{K}_{\mathcal{X}_\mathcal{O}}(\mathbf{x},\mathbf{x}') &= \left[\left[\frac{\partial}{\partial x_{i''_1}}\nabla_{\mathbf{x}}K^{[D-1]}_{\mathcal{X}\mathcal{O}i_{j+1},i'_{j+1}}(\mathbf{x},\mathbf{x}')\right]_{i''_1}\right]_{i_{j+1},i'_{j+1}}\\ \nabla_{\mathbf{W}^{[j']}_\Delta}\mathbf{K}_{\mathcal{W}_\mathcal{O}}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta) &= \left[\left[\frac{\partial}{\partial W^{[j']}_{\Delta i''_{j'},i''_{j'+1}}}K^{[D-1]}_{\mathcal{W}\mathcal{O}i_{j+1},i'_{j+1}}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta)\right]_{i''_{j'},i''_{j'+1}}\right]_{i_{j+1},i'_{j+1}}\\ \nabla_{\mathbf{b}^{[j']}_\Delta}\mathbf{K}_{\mathcal{W}_\mathcal{O}}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta) &= \left[\left[\frac{\partial}{\partial b^{[j']}_{\Delta i''_{j'+1}}}K^{[D-1]}_{\mathcal{W}\mathcal{O}i_{j+1},i'_{j+1}}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta)\right]_{i''_{j'+1}}\right]_{i_{j+1},i'_{j+1}}\end{aligned} \tag{31}$$

其中，递归地定义 $\forall j \in \mathbb{N}_D$:

$$\frac{\partial}{\partial x_{i''_1}}K^{[j]}_{\mathcal{X}\mathcal{O} i_{j+1},i'_{j+1}}(\mathbf{x},\mathbf{x}') = \begin{cases} \sigma^{[j](1)}_{\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}},\tilde{x}'^{[j]}_{\mathcal{O}i'_{j+1}}}\left(\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}\left(\begin{array}{l}\left(\frac{1}{2}\alpha^{[j]2}+\frac{1}{H^{[j]}}\left\langle \mathbf{x}^{[j]}_{\mathcal{O}},\mathbf{x}'^{[j]}_{\mathcal{O}}\right\rangle\right)+\dots\\ \dots\sum\limits_{i_j}\left(\frac{W^{[j]}_{\mathcal{O}i_j,i_{j+1}}W^{[j]}_{\mathcal{O}i_j,i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j,i_{j+1}}\widetilde{\omega}^{[j]}_{i_j,i'_{j+1}}}+1\right)\frac{\omega^{[j]2}_{i_j}}{H^{[j]}}K^{[j-1]}_{\mathcal{X}\mathcal{O}i_j,i_j}(\mathbf{x},\mathbf{x}')\end{array}\right)\right)\dots & \text{if } j>0\\ \qquad\dots\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}\sum\limits_{i_j}\left(\frac{W^{[j]}_{\mathcal{O}i_j,i_{j+1}}W^{[j]}_{\mathcal{O}i_j,i'_{j+1}}}{\widetilde{\omega}^{[j]}_{i_j,i_{j+1}}\widetilde{\omega}^{[j]}_{i_j,i'_{j+1}}}+1\right)\frac{\omega^{[j]2}_{i_j}}{H^{[j]}}\frac{\partial}{\partial x_{i''_1}}K^{[j-1]}_{\mathcal{X}\mathcal{O}i_j,i_j}(\mathbf{x},\mathbf{x}') & \\ \sigma^{[0](1)}_{\tilde{x}^{[0]}_{\mathcal{O}i_1},\tilde{x}'^{[0]}_{\mathcal{O}i'_1}}\left(\mu^{[0]}_{i_1}\mu^{[0]}_{i'_1}\left(\frac{1}{2}\alpha^{[0]2}+\frac{1}{2}\frac{1}{H^{[0]}}\langle\mathbf{x},\mathbf{x}'\rangle\right)\right)\mu^{[0]}_{i_1}\mu^{[0]}_{i'_1}\frac{1}{H^{[0]}}\delta_{i_1,i''_1}\frac{1}{2}x'_{i'_1} & \text{if } j=0\end{cases} \tag{32}$$

$$\begin{aligned}&\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)=\dots\\ &\dots= \begin{cases} \theta^{(1)}\left(\frac{1}{\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}}\left(\begin{array}{l}\left(2b^{[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)+\dots\\ \dots\sum\limits_{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{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)\dots & \text{if } j>j'\\ \quad\dots\frac{1}{\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}}\sum\limits_{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}}\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}{l}\left(2b^{[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)+\dots\\ \dots\sum\limits_{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{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)\dots & \text{if } j=j'>0\\ \qquad\dots\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(2b^{[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}2W'^{[0]}_{\Delta i'_0,i'_1} & \text{if } j=j'=0\end{cases}\end{aligned} \tag{33}$$

$$\frac{\partial}{\partial b^{[j']}_{\Delta i''_{j'+1}}} K^{[j]}_{\mathcal{W}_{\mathcal{O}} i_{j+1}, i'_{j+1}}(\mathbf{W}_\Delta, \mathbf{W}'_\Delta) = \dots$$

$$\dots = \begin{cases} \theta^{(1)}\left(\frac{1}{\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}}\left(\begin{pmatrix}2b^{[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\end{pmatrix} + \dots \right.\right. \\ \left.\left. \dots \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{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j,i_j}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}}\right)\right)\dots \\ \dots \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}}\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) & \text{if } j > j' \\ \theta^{(1)}\left(\frac{1}{\mu^{[j]}_{i_{j+1}}\mu^{[j]}_{i'_{j+1}}}\left(\begin{pmatrix}2b^{[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\end{pmatrix} + \dots \right.\right. \\ \left.\left. \dots \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{K^{[j-1]}_{\mathcal{W}_{\mathcal{O}} i_j,i_j}(\mathbf{W}_\Delta,\mathbf{W}'_\Delta)}{\omega^{[j]2}_{i_j}}\right)\right)\dots \\ \dots \frac{1}{\mu^{[j]}_{i''_{j+1}}\mu^{[j]}_{i'_{j+1}}}\delta_{i_{j+1},i''_{j+1}} 2b'^{[j]}_{\Delta i'_{j+1}} & \text{if } j = j' > 0 \\ \theta^{(1)}\left(\frac{1}{\mu^{[0]}_{i_1}\mu^{[0]}_{i'_1}}\left(2b^{[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} 2b'^{[0]}_{\Delta i'_1} & \text{if } j = j' = 0 \end{cases} \tag{34}$$

C.3 诱导范数

现在我们考虑由特征映射诱导的范数。具体而言，我们希望在 $\mathcal{X}_{\mathcal{O}} \subset \mathbb{R}^{\infty\times m}$ and$\mathcal{W}_{\mathcal{O}} \subset \mathbb{R}^{\infty\times m}$ 上施加范数，并据此计算 $\|\mathbf{\Phi_O(x)}\|_{\mathcal{X}_\mathcal{O}}$、 $\|\mathbf{\Psi_O(W_\Delta)}\|_{\mathcal{W}_\mathcal{O}}$，从而得到构造$\mathbf{f}_\Delta$的再生核巴拿赫空间时所需的诱导范数。原则上，我们可以在这些空间上施加任意范数。然而，请回想:

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

因此:

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

因此，一个自然的选择是一对对偶范数——也就是说，一对范数$\|\cdot\|_{\mathcal{X}_\mathcal{O}}$和$\|\cdot\|_{\mathcal{W}_\mathcal{O}}$，满足:

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

因此，根据 H¨older 不等式:

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

In t在当前语境下，使用 Frobenius 范数（逐元素欧几里得范数）会更方便 s:

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

因此:

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

在继续之前，我们先定义:

$$s^{[j]2} = \begin{cases} \frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{H^{[-1]}}{H^{[0]}}M^{[-1]2} & \text{if } j = 0 \\ \frac{1}{2}\alpha^{[j]2} + \frac{H^{[j-1]}}{H^{[j]}}M^{[j-1]2} & \text{otherwise} \end{cases}$$

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

其中，粗略地说，$t^{[j-1]2}$ 表示第 j 层的权重步长大小，$s^{[j]2}$ 表示向量 $\mathbf{x} \in \mathbb{X}$ 在第 j层输入处的图像上界。使用此记号:

$$\begin{aligned} \|\mathbf{\Phi}_\mathcal{O}(\mathbf{x})\|^2_{\mathcal{X}_\mathcal{O}} &= \textstyle\sum_{i_D} \left\|\mathbf{\Phi}^{[D-1]}_{\mathcal{O}:i_D}(\mathbf{x})\right\|^2_2 \\ \|\mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\|^2_{\mathcal{W}_\mathcal{O}} &= \textstyle\sum_{i_D} \left\|\mathbf{\Psi}^{[D-1]}_{\mathcal{O}:i_D}(\mathbf{W}_\Delta)\right\|^2_2 \end{aligned} \tag{35}$$

其中，递归地定义 $\forall j \in \mathbb{N}_D$:

$$\left\|\boldsymbol{\Phi}^{[j]}_{\mathcal{O}:i_{j+1}}(\mathbf{x})\right\|_2^2 = \begin{cases} \sigma^{[j]}_{\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}},\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}}} \left(\begin{array}{l} \mu^{[j]2}_{i_{j+1}} \left(\left(\frac{1}{2}\alpha^{[j]2} + \frac{1}{H^{[j]}} \left\|\mathbf{x}^{[j]}_{\mathcal{O}}\right\|_2^2 \right) + \dots \right. \\ \left. \dots \sum\limits_{i_j} \left(1 + \frac{W^{[j]2}_{\mathcal{O}i_j,i_{j+1}}}{\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}} \right) \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} \left\|\boldsymbol{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x})\right\|_2^2 \right) \end{array} \right) & \text{if } j > 0 \\ \sigma^{[0]}_{\tilde{x}^{[0]}_{\mathcal{O}i_1},\tilde{x}^{[0]}_{\mathcal{O}i_1}} \left(\mu^{[0]2}_{i_1} \left(\frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{1}{H^{[0]}} \left\|\mathbf{x}^{[0]}_{\mathcal{O}}\right\|_2^2 \right) \right) & \text{if } j = 0 \end{cases} \tag{36}$$

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

C.4 诱导范数梯度

在论文后文中，我们将需要诱导范数的梯度。回顾式 (35)、(36)，并应用链式法则可得:

$$\begin{aligned} \nabla_{\mathbf{x}} \left\|\boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x})\right\|^2_{\mathcal{X}_{\mathcal{O}}} &= \sum\nolimits_{i_D} \left[\frac{\partial}{\partial x_{i''_1}} \left\|\boldsymbol{\Phi}^{[D-1]}_{\mathcal{O}:i_D}(\mathbf{x})\right\|_2^2 \right]_{i''_1} \\ \nabla_{\mathbf{W}^{[j']}_\Delta} \left\|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\right\|^2_{\mathcal{W}_{\mathcal{O}}} &= \sum\nolimits_{i_D} \left[\frac{\partial}{\partial W^{[j']}_{\Delta i''_{j'},i''_{j'+1}}} \left\|\boldsymbol{\Psi}^{[D-1]}_{\mathcal{O}:i_D}(\mathbf{W}_\Delta)\right\|_2^2 \right]_{i''_{j'},i''_{j'+1}} \\ \nabla_{\mathbf{b}^{[j']}_\Delta} \left\|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta)\right\|^2_{\mathcal{W}_{\mathcal{O}}} &= \sum\nolimits_{i_D} \left[\frac{\partial}{\partial b^{[j']}_{\Delta i''_{j'}}} \left\|\boldsymbol{\Psi}^{[D-1]}_{\mathcal{O}:i_D}(\mathbf{W}_\Delta)\right\|_2^2 \right]_{i''_{j'+1}} \end{aligned} \tag{37}$$

其中，递归定义为 $\forall j \in \mathbb{N}_D$:

$$\frac{\partial}{\partial x_{i'_1}} \left\|\boldsymbol{\Phi}^{[j]}_{\mathcal{O}:i_{j+1}}(\mathbf{x})\right\|_2^2 = \dots$$

$$\dots = \begin{cases} \sigma^{[j](1)}_{\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}},\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}}} \left(\mu^{[j]2}_{i_{j+1}} \left(\left(\frac{1}{2}\alpha^{[j]2} + \frac{1}{H^{[j]}} \left\|\mathbf{x}^{[j]}_{\mathcal{O}}\right\|_2^2 \right) + \sum\limits_{i_j} \left(\frac{W^{[j]2}_{\mathcal{O}i_j,i_{j+1}}}{\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}} + 1 \right) \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} \left\|\boldsymbol{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x})\right\|_2^2 \right) \dots \right. \\ \left. \dots \mu^{[j]2}_{i_{j+1}} \sum\limits_{i_j} \left(\frac{W^{[j]2}_{\mathcal{O}i_j,i_{j+1}}}{\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}} + 1 \right) \frac{\omega^{[j]2}_{i_j}}{H^{[j]}} \frac{\partial}{\partial x_{i'_1}} \left\|\boldsymbol{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x})\right\|_2^2 \right) & \text{if } j > 0 \\ \sigma^{[0](1)}_{\tilde{x}^{[0]}_{\mathcal{O}i_1},\tilde{x}^{[0]}_{\mathcal{O}i_1}} \left(\mu^{[0]2}_{i_1} \left(\frac{1}{2}\alpha^{[0]2} + \frac{1}{2}\frac{1}{H^{[0]}} \left\|\mathbf{x}\right\|_2^2 \right) \right) \mu^{[0]2}_{i_1} \frac{1}{H^{[0]}} \delta_{i_1,i'_1} x_{i'_1} & \text{if } j = 0 \end{cases} \tag{38}$$

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

$$\dots = \begin{cases} \theta^{(1)} \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} \left(t^{[j]}_{\Delta i_{j+1}} + \sum\limits_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}} \right) \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i_j}} \right) \right) \dots \\ \dots \frac{1}{\mu^{[j]2}_{i_{j+1}}} \sum\limits_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}} \right) \frac{1}{\omega^{[j]2}_{i_j}} \frac{\partial}{\partial W^{[j']}_{\Delta i'_{j'},i'_{j'+1}}} \left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2 & \text{if } j > j' \\ \theta^{(1)} \left(\frac{1}{\mu^{[j]2}_{i_{j+1}}} \left(t^{[j]}_{\Delta i_{j+1}} + \sum\limits_{i_j} \left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}} + W^{[j]2}_{\Delta i_j,i_{j+1}} \right) \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i_j}} \right) \right) \dots \\ \dots \frac{1}{\mu^{[j]2}_{i'_{j+1}}} \delta_{i_{j+1},i'_{j+1}} 2 \left(1 + \frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i'_j}(\mathbf{W}_\Delta)\right\|_2^2}{\omega^{[j]2}_{i'_j}} \right) W^{[j]}_{\Delta i'_j,i_{j+1}} & \text{if } j = j' > 0 \\ \theta^{(1)} \left(\frac{1}{\mu^{[0]2}_{i_1}} t^{[0]}_{\Delta i_1} \right) \frac{1}{\mu^{[0]2}_{i'_1}} \delta_{i_1,i'_1} 4 W^{[0]}_{\Delta i'_0,i_1} & \text{if } j = j' = 0 \end{cases} \tag{39}$$

$$\frac{\partial}{\partial b^{[j']}_{\Delta i'_{j'+1}}}\left\|\boldsymbol{\Psi}^{[j]}_{\mathcal{O}:i_{j+1}}\left(\mathbf{W}_{\Delta}\right)\right\|_2^2=\begin{cases}\theta^{(1)}\left(\frac{1}{\mu^{[j]2}_{i_{j+1}}}\left(t^{[j]}_{\Delta i_{j+1}}+\sum_{i_j}\left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}+W^{[j]2}_{\Delta i_j,i_{j+1}}\right)\frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_{\Delta})\right\|_2^2}{\omega^{[j]2}_{i_j}}\right)\right)\cdots & \\ \quad\cdots\frac{1}{\mu^{[j]2}_{i_{j+1}}}\sum_{i_j}\left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}+W^{[j]2}_{\Delta i_j,i_{j+1}}\right)\frac{1}{\omega^{[j]2}_{i_j}}\frac{\partial}{\partial b^{[j']}_{\Delta i'_{j'+1}}}\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{W}_{\Delta}\right)\right\|_2^2 & \text{if } j>j' \\ \theta^{(1)}\left(\frac{1}{\mu^{[j]2}_{i_{j+1}}}\left(t^{[j]}_{\Delta i_{j+1}}+\sum_{i_j}\left(\widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}+W^{[j]2}_{\Delta i_j,i_{j+1}}\right)\frac{\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{W}_{\Delta})\right\|_2^2}{\omega^{[j]2}_{i_j}}\right)\right)\cdots & \\ \quad\cdots\frac{1}{\mu^{[j]2}_{i'_{j+1}}}\delta_{i_{j+1},i'_{j+1}}4b^{[j]}_{\Delta i_{j+1}} & \text{if } j=j'>0 \\ \theta^{(1)}\left(\frac{1}{\mu^{[0]2}_{i_1}}t^{[0]}_{\Delta i_1}\right)\frac{1}{\mu^{[0]2}_{i'_1}}\delta_{i_1,i'_1}4b^{[0]}_{\Delta i_1} & \text{if } j=j'=0\end{cases} \tag{40}$$

θ 和 σ的性质

Here 我们介绍 θ and 和 $\sigma_{z,z'}$ 函数的一项重要预备性质。回顾一下, :

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

其中:

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

我们注意到，对于 $z\in\mathbb{R}$、 $j\in\mathbb{N}_D$:

- $\sigma^{[j]}_{z,\ z}$： $(-\overline{\sqrt{\rho}}^{[j]},\overline{\sqrt{\rho}}^{[j]})\to\mathbb{R}$ 通常是无界、光滑且非利普希茨的。
- $\sigma^{[j]}_{z,\ z}$： $[0,\overline{\sqrt{\rho}}^{[j]})\to\mathbb{R}_+$ 通常是无界、光滑、严格递增且非利普希茨的。
- θ： $(-1,\ 1)\to\mathbb{R}$ 无界、光滑且非利普希茨。
- θ: $[0,1)\to[0,\infty)$ 无界、光滑、严格递增且为非利普希茨函数。
- θ^{-1}: $[0,\infty)\to[0,1)$ 有界、光滑、严格递增且为利普希茨函数。
- $\theta_{(1)}$: $[0,1)\to[1,\infty)$ 无界、光滑、严格递增且为非利普希茨函数。
- $\theta_{(1)-1}$: $[1,\infty)\to[0,1)$ 有界、光滑、严格递增且为利普希茨函数。

在分析我们的特征映射的收敛性质时， $\sigma^{[j]}_{z,z}$ 和 θ 函数无界且为非利普希茨函数这一特性会带来问题。不过，如果我们通过定义 $\epsilon^{[j]}_{\psi},\epsilon^{[j]}_{\phi}\in(0,1)\ \forall j\in\mathbb{N}_D$ 并且：来限制 $\sigma^{[j]}_{z,z}$ 和 θ 的定义域

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

那么:

- $\overline{\sigma}^{[j]}$： $[0,(1-\epsilon^{[j]}_{\phi})\sqrt{\rho^{[j]}}]\to[0,\frac{\phi^{[j]}}{H^{[j]}}]\subset\mathbb{R}_+$ 有界、光滑、严格递增且满足 $L^{[j]}_{\phi}$-利普希茨条件，其中:

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

  并且我们注意到，当 $\epsilon^{[j]}_{\phi}\to 0$ 时， $\frac{\phi^{[j]}}{H^{[j]}}\to\infty$。

- $\overline{\sigma}^{[j]-1}$： $[0,\frac{\phi^{[j]}}{H^{[j]}}]\to[0,(1-\epsilon^{[j]}_{\phi})\sqrt{\rho^{[j]}}]$ 有界、光滑且严格递增。

- θ： $[0,1-\epsilon^{[j]}_{\psi}]\to[0,\frac{\psi^{[j]}}{H^{[j]}}]\subset\mathbb{R}_+$ 有界、光滑、严格递增且满足 $L^{[j]}_{\psi}$-利普希茨条件，其中:

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

  并且我们注意到，当 $\epsilon^{[j]}_{\psi}\to 0$ 时， $\frac{\psi^{[j]}}{H^{[j]}}\to\infty$。

- θ^{-1}： $[0, \frac{\psi^{[j]}}{H^{[j]}}] \to [0, 1-\epsilon_\psi^{[j]}]$ 有界、光滑且严格递增。

在下一节分析诱导范数的收敛性和有限性时，我们将用到这一对定义域的限制。我们还发现，定义相关函数很有用：

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

其中，我们注意到：

- κ： $[0,1) \to [0,\infty)$ 严格递增。
- κ^{-1}： $[0,\infty) \to [0,1)$ 严格递增。
- λ： $(0,\ 1] \to [0,\infty)$ 严格递减。
- λ^{-1}： $[0,\infty) \to (0,1]$ 严格递减。

C.6 收敛条件

本节最后，我们考虑诱导范数的收敛问题，并指出这些结果也适用于诱导核：

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

并利用诱导核的正定性。对于$\mathbf{f}_\Delta$ 的特征表示，只需确保 $\|\tilde{\mathbf{x}}_\Delta^{[j]}\|_\infty \le \rho^{[j]}$ 对所有 i_{j+1}, j都成立；但在此处无法直接采用这一方法，因为特征映射 $\mathbf{\Phi}$ 无法访问 $\mathbf{W}_\Delta$ （无论如何，在此情境下它是不确定的），同样，特征映射 $\mathbf{\Psi}$ 也无法访问$\mathbf{x}$。回顾我们的定义：

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

使用这一记号，以下定理给出了对所有$\mathbf{x} \in \mathbb{X}$ 以及给定权重步长$\mathbf{W}_\Delta$，诱导范数收敛（有限）的条件：

定理 5. 令 $\epsilon_\phi^{[j]} \in (0,1)\ \forall j \in \mathbb{N}_D$，对于给定的神经网络和初始权重$\mathbf{W}_\mathcal{O}$，定义：

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

如果缩放因子满足：

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

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

证明。 回顾（35）、（36），我们有：

$$\begin{aligned}&\left\|\mathbf{\Phi}^{[j]}_{\mathcal{O}:i_{j+1}}(\mathbf{x})\right\|_2^2 = \dots\\ &\dots = \begin{cases}\sigma^{[j]}_{\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}},\tilde{x}^{[j]}_{\mathcal{O}i_{j+1}}}\left(\mu^{[j]2}_{i_{j+1}}\left(\left(\alpha^{[j]2} + \frac{2}{H^{[j]}}\left\|\mathbf{x}^{[j]}_\mathcal{O}\right\|_2^2\right) + \sum_{i_j}\frac{\omega^{[j]2}_{ij}}{H^{[j]}}\left(\frac{W^{[j]2}_{\mathcal{O}i_j i_{j+1}}}{\tilde{\omega}^{[j]2}_{i_j,i_{j+1}}}+1\right)\left\|\mathbf{\Phi}^{[j-1]}_{\mathcal{O}:i_j}(\mathbf{x})\right\|_2^2\right)\right) & \text{if } j>0\\ \sigma^{[0]}_{\tilde{x}^{[0]}_{\mathcal{O}i_1},\tilde{x}^{[0]}_{\mathcal{O}i_1}}\left(\mu^{[0]2}_{i_1}\left(\alpha^{[0]2} + \frac{1}{H^{[0]}}\left\|\mathbf{x}^{[0]}_\mathcal{O}\right\|_2^2\right)\right) & \text{if } j=0\end{cases}\end{aligned}$$

其中，正如 C.5 节所讨论的，在受限值域 $\sigma_{z,\ z}^{[j]}$ 上：$[0,(1-\epsilon_\phi^{[j]})\sqrt{\rho[j]}] \to [\overline{0,\phi^{[j]}}]$ 对于所有 $z \in \mathbb{R}$ 都是有界、递增且满足利普希茨条件的。根据我们的假设，我们有：

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

因此：

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

我们将通过对这个界进行界定，构造收敛的充分条件。从输入层开始，如果 $\sigma_{\tilde{x}_{\mathcal{O}i_1}^{[0]}, \tilde{x}_{\mathcal{O}i_1}^{[0]}}^{[0]}$ 的自变量位于受限 ROC 内，即满足以下条件，则可保证 $\|\mathbf{\Phi}_{\mathcal{O}\ :i_1}^{[0]}(\mathbf{x})\|_2^2$ 收敛：

$$\mu_{i_1}^{[0]2} s^{[0]2} \leq \left(1 - \epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}} \quad \forall i_1 \tag{41}$$

此外，若满足此条件，则：

$$\left\|\mathbf{\Phi}_{\mathcal{O}:i_1}^{[0]}(\mathbf{x})\right\|_2^2 \leq \tfrac{\phi^{[0]}}{H^{[0]}} \quad \forall i_1 \tag{42}$$

对于层 $j \in \mathbb{N}_D \backslash \{0\}$，如果 $\sigma_{\tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}, \tilde{x}_{\mathcal{O}i_{j+1}}^{[j]}}^{[j]}$ 的自变量位于受限 ROC 内，则可保证 $\|\mathbf{\Phi}_{\mathcal{O}\ :i_{j+1}}^{[j]}(\mathbf{x})\|_2^2$ 收敛：

$$\mu_{i_{j+1}}^{[j]2}\left(s^{[j]2} + \sum_{i_j} \frac{\omega_{i_j}^{[j]2}}{H^{[j]}}\left(\frac{W_{\mathcal{O}i_j i_{j+1}}^{[j]2}}{\widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}} + 1\right)\left\|\mathbf{\Phi}_{\mathcal{O}:i_j}^{[j-1]}(\mathbf{x})\right\|_2^2\right) \leq \left(1 - \epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}} \quad \forall i_{j+1} \tag{43}$$

此外，若满足此条件，则：

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

假设层 $j' < j$ 满足 (41)、(43)，则满足以下条件即可保证 $\|\mathbf{\Phi}_{\mathcal{O}\ :i_{j+1}}^{[j]}(\mathbf{x})\|_2^2$ 收敛：

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

或者，充分地：

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

这就完成了证明。 □

定理 6. 令 $\epsilon_\psi^{[j]} \in (0,1)\ \forall j \in \mathbb{N}_D$，对于给定的神经网络和初始权重 $\mathbf{W}_\mathcal{O}$，定义：

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

对于权重步长$\mathbf{W}_\Delta$，若：

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

那么 $\left\|\mathbf{\Psi}_\mathcal{O}^{[j]}(\mathbf{W}_\Delta)\right\|_F^2 \leq \psi^{[j]}$ 对所有 $j \in \mathbb{N}_D$均成立。

证明。 回顾 (35)、(36)，我们有：

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

如 C.5 节关于受限值域 θ 的讨论所述：$[0,(1-\epsilon_\psi^{[j]})] \to [0,\psi^{[j]}]$ 有界、递增且是 $\frac{1}{\epsilon_\psi^{[j]2}}$-利普希茨函数。从第 0 层开始，我们看到，如果：

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

那么：

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

此外，对于层 j，如果：

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

那么：

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

由此得到一般情况下的充分条件：

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

至此，证明完成。 □

注意，这个定理也可以用缩放因子的界来等价表述：

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

而不是作为对权重步长的界。因此，定理 5 和定理 6 揭示了缩放因子在我们方案中的作用。粗略地说，缩放因子的选择必须在特征映射 $\mathbf{\Phi}_\mathcal{O}$ 和 $\mathbf{\Psi}_\mathcal{O}$的收敛之间权衡；这些定理说明了这种权衡可以如何调整，以及调整的幅度。具体而言：

- 定理 5 给出的缩放因子上界会随着 $\mathbb{X}$ 的“大小”以及网络所有层的输出范围（以 $s^{[j]}$衡量）变小而增大（限制变宽松），并在极限 $s^{[j]} \to 0^+$下趋于 ∞ （无界）。当然，因子 $s^{[j]}$ 由网络结构、数据集 $\mathbb{X}$ 的范围以及我们对 α 的选择决定，因此在实践中，缩放因子的这一上界是固定的。

- 定理 6 给出的缩放因子下界会随着权重步长以及依赖于 $\widetilde{\omega}_{i_j}^{[j]}$ 的因子，i_{j+1}变小而减小，并在它们趋于零时降至 0（无界）。因此，缩放因子实际上受特征映射 $\mathbf{\Psi}_\mathcal{O}$的收敛条件约束，下界由权重步长的大小决定。与定理 5 给出的上界不同，只要要求权重步长和偏移量足够小，我们就可以使下界任意小。

这两项观察——缩放因子的上界和下界——共同决定了我们能够在所有输入集合 $\mathbb{X}$ 上建模哪些权重步长。此外，定理 6 下界中缩放因子的存在，表明了这些因子可以如何选取。具体而言，要使这些界限有意义，我们需要满足以下条件：

- 影子权重 $\widetilde{\omega}_{i_j}^{[j]}$，$i_{j+1}$ 必须与权重步长处于相同（或更低）的阶（尺度），以确保当权重步长足够小时，缩放因子的下界会低于其上界。

- 影子权重 $\omega_{i_j}^{[j]}$ 必须与范数 $\|\mathbf{\Psi}_\mathcal{O}^{[j-1]:i_j}(\mathbf{W}_\Delta)\|_2^2$ 处于相同（或更高）的阶（尺度），从而抵消其对缩放因子下界的影响。

我们通过结合这些定理，将这一直觉具体化，得到关于权重步长、影子权重和缩放因子大小的充分条件，以确保对于所有 $\mathbf{x}\in\mathbb{X}$，$\mathbf{\Phi}_\mathcal{O}$ 和 $\mathbf{\Psi}_\mathcal{O}$ 均收敛：

定理 7。 设 $\widetilde{\epsilon},\epsilon_\phi^{[j]},\epsilon_\psi \in (0,1)$ $\forall j\in\mathbb{N}_D$，对于给定的神经网络和初始权重 $\mathbf{W}_\mathcal{O}$，定义：

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

递归地定义，随后：

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

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

对于权重步长 $\mathbf{W}_{\Delta}$，如果影子权重满足：

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

并且权重步长满足：

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

并且缩放因子满足：

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

那么 $\left\|\boldsymbol{\Phi}_{\mathcal{O}}^{[j]}(\mathbf{x})\right\|_F^2 \leq \phi^{[j]}\ \forall \mathbf{x} \in \mathbb{X}, j \in \mathbb{N}_D$ 并且 $\left\|\boldsymbol{\Psi}_{\mathcal{O}}^{[j]}\left(\mathbf{W}_{\Delta}\right)\right\|_F^2 \leq \psi^{[j]}$ 对所有 $j \in \mathbb{N}_D$均成立。

证明。 结合定理 5 和定理 6，我们可以看出，要使 $\boldsymbol{\Phi}_{\mathcal{O}}, \boldsymbol{\Psi}_{\mathcal{O}}$ 均收敛，需要满足：

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

使用我们的条件 $\omega_{i_j}^{[j]2} \in \left[\|\boldsymbol{\Psi}_{\mathcal{O}\,i_j}^{[j-1]}:(\mathbf{W}_{\Delta})\|_2^2,\ \underline{\frac{\psi^{[j-1]}}{H^{[j-1]}}}\right]$ 那么只需满足：

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

We r要求只要权重步足够小，就能满足这一条件，因此 $\forall j \in \mathbb{N}_D \setminus \{0\},\ i_{j+1}$:

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

其中:

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

为使其定义良好，我们要求右侧为正；满足以下条件即可:

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

我们可以用 $\epsilon^{[j-1]}_\psi$ 将其改写为:

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

在这种情况下，我们的充分条件变为:

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

and:

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

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

就缩放因子而言，回顾定理 5 和定理 6，我们要求（充分地）:

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

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

注意到 $\chi_{i_{j+1}}^{[j]} \leq 1$，我们可以进一步收紧约束，以得到充分条件：

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

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

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

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

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

由此完成证明。 □

最后，我们通过考察一个稍后将发挥重要作用的特殊情况来结束本节。

定理 8. 令 $\epsilon_{\phi}^{[j]}, \epsilon_{\psi}^{[j]}, \delta \in (0,1)\ \forall j \in \mathbb{N}_D$，对于一个具有初始权重 $\mathbf{W}_{\mathcal{O}}$ 和权重步长 $\mathbf{W}_{\Delta}$ 的给定神经网络，使用影子权重：

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

其中：

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

如果权重步长满足：

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

$$t_{\Delta i_{j+1}}^{[j]2}+\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 \leq \frac{\left(1-\epsilon_{\psi}^{[j]}\right)}{\frac{s^{[j]2}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}+\frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{w}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_{\infty}^2}+H^{[j-1]}\right)\frac{\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}{\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}}} \quad \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$$

并且缩放因子满足：

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

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

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* $j \in \mathbb{N}_D$.

证明。 我们的证明与定理 refth:convergephipsi 的证明简单类似，只是针对我们预先定义的缩放因子做了调整。结合定理 5 和定理 6可知，要使 $\boldsymbol{\Phi}_{\mathcal{O}}, \boldsymbol{\Psi}_{\mathcal{O}}$ 都收敛，按照我们对缩放因子的定义，需要满足：

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

因此，只需满足：

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

根据缩放因子，并回顾定理 5 和定理 6，我们要求（充分地）：

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

$\forall i_1, \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1}$，由此完成证明。 □

D 规范缩放

In 在下一节中，我们将考虑如下形式的正则化风险最小化问题：

$$\mathbf{W}_{\Delta}^{\bullet} = \underset{\mathbf{W}_{\Delta}\in\mathbb{W}_{\mathcal{O}}}{\operatorname{argmin}} R_{\lambda}\left(\mathbf{W}_{\Delta}\right) = \lambda h\left(\left\|\boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\|_{\mathcal{W}_{\mathcal{O}}}\right) + \sum_k E\left(\mathbf{x}^{\{k\}}, \mathbf{y}^{\{k\}}, \mathbf{f}_{\mathcal{O}}\left(\mathbf{x}^{\{k\}}\right) + \left\langle \boldsymbol{\Phi}_{\mathcal{O}}\left(\mathbf{x}^{\{k\}}\right), \boldsymbol{\Psi}_{\mathcal{O}}\left(\mathbf{W}_{\Delta}\right)\right\rangle_{\mathcal{X}_{\mathcal{O}}\times\mathcal{W}_{\mathcal{O}}}\right) \tag{44}$$

我们将探讨的问题是：

> 给定一个具有初始权重和偏置 $\mathbf{W}_{\mathcal{O}}$ 的神经网络，设 $\mathbf{W}_{\Delta}^{\circledast}$ 为由式 *(21)* 定义的反向传播权重步（学习率为 η 的梯度下降），并设 $\mathbf{W}_{\Delta}^{\bullet}$ 为求解正则化风险最小化问题 *(44)* 的权重步。给定由梯度下降得到的权重步 $\mathbf{W}_{\Delta}^{\circledast}$，是否存在一种影子权重和缩放因子的选取方式（可能依赖于 $\mathbf{W}_{\Delta}^{\circledast}$），以及正则化参数 λ，使得 $\mathbf{W}_{\Delta}^{\bullet} = \mathbf{W}_{\Delta}^{\circledast}$？

如果答案是肯定的（我们将证明这一点），那么通过分析 (44)，我们就能加深对反向传播的理解。现在，式 (44) 的解必须满足一阶最优性条件（为简洁起见，此处假设其可微）：

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

现在注意到，式 (44) 中第二项的导数对应于反向传播中的梯度；若第一项（正则化项）的梯度满足：

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

对于某个 $\nu \in \mathbb{R}_+$，并且：

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

那么：

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

因此，对于是否存在缩放因子、影子权重和 λ，使正则化风险最小化的权重步与指定学习率 η 下的梯度下降权重步相对应这一问题，我们可以通过证明规范缩放的存在来肯定地回答，定义如下：

定义3（规范缩放）。对于给定的神经网络、初始权重 $\mathbf{W}_{\mathcal{O}}$ 以及由反向传播生成的权重步 $\mathbf{W}_{\Delta}^{\circledast}$，我们将满足以下条件的一组影子权重和缩放因子定义为规范缩放：

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

对于某个 $\nu\in\mathbb{R}_{+}$，且 $\left\|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta}^{\circledast})\right\|_{\mathcal{W}_{\mathcal{O}}}<\infty$，$\left\|\boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x})\right\|_{\mathcal{X}_{\mathcal{O}}}<\infty\ \forall\mathbf{x}\in\mathbb{X}$。我们分别将利用规范缩放诱导出的核和范数称为规范诱导核 和规范诱导范数。

关于其存在性，以下定理给出了规范缩放对应的权重步存在的充分条件。定义 $\delta\in(0,1)$ 以及：

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

我们的核心结果如下：

定理 9. 令 $\epsilon_{\phi},\epsilon_{\psi},\delta,\chi\in(0,1)$，对于一个初始权重为 $\mathbf{W}_{\mathcal{O}}$ 的给定神经网络，令 $\mathbf{W}_{\Delta}^{\circledast}$ 为通过反向传播得到的权重步长。令：

$$0\leq\epsilon_{\psi}^{[j]}\leq\begin{cases}1-\frac{s^{[D-1]2}}{(1-\epsilon_{\phi})\sqrt{\rho^{[D-1]}}}\left\|\mathbf{t}_{\Delta}^{[D-1]\circledast}\right\|_{\infty}^{2} & \textit{if } j=D-1\\ \frac{1}{1-\delta}\frac{\left\|\mathbf{W}_{\Delta}^{[j+1]\circledast}\right\|_F^2}{\left\|\mathbf{t}_{\Delta}^{[j]\circledast}\right\|_{\infty}^{2}} & \textit{otherwise}\end{cases}\qquad\forall j\in\mathbb{N}_D$$

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

and:

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

$$\frac{\phi^{[j]}}{H^{[j]}}=\overline{\sigma}^{[j]}\left(\left(1-\epsilon_{\phi}^{[j]}\right)\sqrt{\rho^{[j]}}\right)$$

$\forall j\in\mathbb{N}_D$。对于某个 $\alpha^{[j]}\in\mathbb{R}_{+}$，存在 $\varpi^{[j]}\in(0,1)\ \ \forall j\in\mathbb{N}_D\backslash\{0\}$，使得：[13]

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

如果权重步长满足：

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

[13]注意，$b_{\Delta i_{j+1}}^{[j]\circledast}$ 与 $\alpha^{[j]}$成正比，因此我们总能通过充分增大 $\alpha^{[j]}$ 来增大 $t_{\Delta i_{j+1}}^{[j]\circledast 2}$，从而确保条件成立。

那么就存在一种规范缩放：

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

其中：

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

且 $\left\| \mathbf{\Phi}_{\mathcal{O}}^{[j]} \left(\mathbf{x} \right) \right\|_F^2 \leq \phi^{[j]} \ \forall \mathbf{x} \in \mathbb{X}, \ \ j \in \mathbb{N}_D$ 且 $\left\| \mathbf{\Psi}_{\mathcal{O}}^{[j]} \left(\mathbf{W}_\Delta \right) \right\|_F^2 \leq \psi^{[j]}$ 对所有 $j \in \mathbb{N}_D$均成立。

证明。 我们的目标是为一般情形推导规范缩放，其中权重步长 $\mathbf{W}_\Delta^{\circledast}$ 是先验给定的。由 (37)、(39) 和 (40) 可知：

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

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

其中，递归地定义 $\forall j \in \mathbb{N}_D$：

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

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

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

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

并且我们定义：

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

因此：

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

请注意，关于特定权重 $W_{\Delta i'_{j'}, \ i'_{j'+1}}^{[j']}$ 或偏置 $b_{\Delta i'_{j'+1}}^{[j']}$ 的范数梯度递归公式，精准地反映了网络的结构：计算从输出层开始，并沿着所有可能的路径反向递归至第 j' 层的神经元 $i'_{j'+1}$。实际上，如果我们将从输出层中的任意神经元到第 j' 层神经元 $i'_{j'+1}$ 的所有路径集合定义为：

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

那么，我们可以将范数梯度重写为路径和：

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

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

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

$$\ldots=\left(\sum_{\mathtt{i}\in\mathcal{P}^{[j']}_{i'_{j'+1}}} g^{[j']\circledast}_{i_{j'+1}}\prod_{j''=j'+1}^{D-1} g^{[j'']\circledast}_{i_{j''+1}}h^{[j''-1,j'']\circledast}_{i_{j''},i_{j''+1}}\right)4b^{[j]\circledast}_{\Delta i_{j+1}}$$

其中，路径求和中的每一项都是神经元代价的乘积：

$$g^{[j]\circledast}_{i_{j+1}}=\begin{cases}\frac{1}{\mu^{[D-1]2}_{i_D}}\theta^{(1)}\left(\frac{t^{[D-1]\circledast 2}_{\Delta i_D}}{\mu^{[D-1]2}_{i_D}}\right) & \text{if } j=D-1\\ \frac{1}{\omega^{[j+1]2}_{i_{j+1}}}\frac{1}{\mu^{[j]2}_{i_{j+1}}}\theta^{(1)}\left(\frac{t^{[j]\circledast 2}_{\Delta i_{j+1}}}{\mu^{[j]2}_{i_{j+1}}}\right) & \text{otherwise}\end{cases}$$

以及连接代价：

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

对于所有神经元 a 以及该路径上从输出层回溯到第 j 层神经元 $i'_{j'+1}$ 的神经元之间的连接 $'$..

暂且不考虑收敛问题，我们可以看出，要获得规范缩放，必须使用缩放因子和影子权重来“均衡”神经元代价和连接代价对路径的依赖性，并消除终止神经元 i_{j+1}（位于第 j 层）的特征图范数对结果的影响。为此，很容易看出，若我们选择：

$$\begin{aligned}\omega^{[j]2}_{i_j}&=\left\|\boldsymbol{\Psi}^{[j-1]}_{\mathcal{O}:i_j}\left(\mathbf{W}^{\circledast}_{\Delta}\right)\right\|^2_2\\ \widetilde{\omega}^{[j]2}_{i_j,i_{j+1}}&=\frac{1}{1-\delta}\left\|\mathbf{W}^{[j]\circledast}_{\Delta:i_{j+1}}\right\|^2_2-W^{[j]\circledast 2}_{\Delta i_j,i_{j+1}}\end{aligned}\qquad\forall j\in\mathbb{N}_D,i_j,i_{j+1}\tag{45}$$

其中 $\delta\in(0,\ 1)$（请记住，我们假设反向传播权重步 $\mathbf{W}^{\circledast}_{\Delta}$ 是预先计算的，然后尝试使基于正则化的权重步长 $\mathbf{W}^{\bullet}_{\Delta}$ 与之对应），则：

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

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

and:

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

$$h^{[j-1,j]\circledast}_{i_j,i_{j+1}}=\frac{1}{1-\delta}\left\|\mathbf{W}^{[j]\circledast}_{\Delta:i_{j+1}}\right\|^2_2$$

利用（36），我们得到

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

$$g^{[j]\circledast}_{i_{j+1}}=\begin{cases}\frac{1}{\mu^{[D-1]2}_{i_D}}\theta^{(1)}\left(\frac{t^{[D-1]\circledast 2}_{\Delta i_D}}{\mu^{[D-1]2}_{i_D}}\right) & \text{if } j=D-1\\ \frac{\mu^{[j]2}_{i_{j+1}}-t^{[j]\circledast 2}_{\Delta i_{j+1}}}{t^{[j]\circledast 2}_{\Delta i_{j+1}}}\frac{1}{\mu^{[j]2}_{i_{j+1}}}\theta^{(1)}\left(\frac{t^{[j]\circledast 2}_{\Delta i_{j+1}}}{\mu^{[j]2}_{i_{j+1}}}\right) & \text{otherwise}\end{cases}$$

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

从现在起，假设采用缩放因子（45）。利用缩放因子均衡范数梯度中的连接代价后，接下来要做的是利用缩放因子均衡神经元代价，并得到规范缩放。同样，暂且不考虑收敛问题，考察输出层权重和偏置的梯度，它们具有特别简单的形式：

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

对于规范缩放，我们必须满足：

$$\frac{4}{\mu_{i_D}^{[D-1]2}}\theta^{(1)}\left(\frac{t_{\Delta i_D}^{[D-1]\boxtimes 2}}{\mu_{i_D}^{[D-1]2}}\right)=\frac{4}{t_{\Delta i_D}^{[D-1]\boxtimes 2}}\kappa\left(\frac{t_{\Delta i_D}^{[D-1]\boxtimes 2}}{\mu_{i_D}^{[D-1]2}}\right)=\nu$$

回顾：

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

因此：

$$\mu_{i_D}^{[D-1]2}=\frac{t_{\Delta i_D}^{[D-1]\boxtimes 2}}{\kappa^{-1}\left(\frac{t_{\Delta i_D}^{[D-1]\boxtimes 2}\nu}{4}\right)}=\frac{2\frac{\nu t_{\Delta i_D}^{[D-1]\boxtimes 2}}{4}}{2\frac{\nu t_{\Delta i_D}^{[D-1]\boxtimes 2}}{4}+1-\sqrt{4\frac{\nu t_{\Delta i_D}^{[D-1]\boxtimes 2}}{4}+1}}t_{\Delta i_D}^{[D-1]\boxtimes 2}$$

令 $\mu_{i_D}^{[D-1]2}$ 取 th　　其值，考虑第 $D-2$ 层权重和偏置的梯度，它们可简化为：

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

由此可见，对于规范缩放，我们必须满足：

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

对于层 $D-3$，使用目前推导出的缩放因子：

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

依此类推。因此，逐层向后推导，我们发现，选择以下影子权重即可实现规范缩放：

$$\begin{aligned}\omega_{i_j}^{[j]2}&=\left\|\mathbf{\Psi}_{\mathcal{O}:i_j}^{[j-1]}\left(\mathbf{W}_{\Delta}^{\boxtimes}\right)\right\|_2^2 &&\forall j\in\mathbb{N}_D, i_j\\ \widetilde{\omega}_{i_j,i_{j+1}}^{[j]2}&=\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]\boxtimes}\right\|_F^2-W_{\Delta i_j,i_{j+1}}^{[j]\boxtimes 2} &&\forall j\in\mathbb{N}_D\backslash\{0\}, i_j, i_{j+1}\end{aligned}\tag{46}$$

以及以下缩放因子：

$$\begin{aligned}\mu_{i_D}^{[D-1]2} &= \frac{t_{\Delta i_D}^{[D-1]\boxtimes 2}}{\kappa^{-1}\left(\frac{t_{\Delta i_D}^{[D-1]\boxtimes 2\,\nu}}{4}\right)} = \frac{2\frac{\nu t_{\Delta i_D}^{[D-1]\boxtimes 2}}{4}}{2\frac{\nu t_{\Delta i_D}^{[D-1]\boxtimes 2}}{4}+1-\sqrt{4\frac{\nu t_{\Delta i_D}^{[D-1]\boxtimes 2}}{4}+1}} t_{\Delta i_D}^{[D-1]\boxtimes 2} \quad \forall i_D \\ \mu_{i_{j+1}}^{[j]2} &= \frac{t_{\Delta i_{j+1}}^{[j]\boxtimes 2}}{t_{\Delta i_{j+1}}^{[j]\boxtimes 2}-\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta}^{[j+1]\boxtimes}\right\|_F^2} t_{\Delta i_{j+1}}^{[j]\boxtimes 2} \qquad \forall j \in \mathbb{N}^{D-1}, i_{j+1}\end{aligned} \tag{47}$$

我们还得到关于权重步长的第一个附加约束，即利用 $\mu_{i_{j+1}}^{[j]2}$ 必须为正这一要求：

$$\tfrac{1}{1-\delta}\left\|\mathbf{W}_{\Delta}^{[j+1]\boxtimes}\right\|_F^2 \leq \left(1-\varpi^{[j+1]}\right)\left\|\mathbf{t}_{\Delta}^{[j]\boxtimes}\right\|_\infty^2 \tag{48}$$

最后，我们必须考虑给定缩放因子和影子权重 (46)、(47) 时的收敛问题。回顾定理 8，给定缩放因子 (46)，并使用其中的定义，若：

$$\begin{aligned} t_{\Delta i_1}^{[0]2} &\leq \frac{\left(1-\epsilon_\psi^{[0]}\right)}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)} \qquad \forall i_1 \\ t_{\Delta i_{j+1}}^{[j]2} + \tfrac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2 &\leq \frac{\left(1-\epsilon_\psi^{[j]}\right)}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}+\frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{w}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_\infty^2}+H^{[j-1]}\right)\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}} \quad \forall j \in \mathbb{N}_D \backslash \{0\}, i_{j+1} \end{aligned} \tag{49}$$

我们将其视为已知，且：

$$\begin{aligned} \mu_{i_1}^{[0]2} &\in \left[\frac{1}{1-\epsilon_\psi^{[0]}} t_{\Delta i_1}^{[0]2}, \frac{1}{\left(\frac{s^{[0]2}}{\left(1-\epsilon_\phi^{[0]}\right)\sqrt{\rho^{[0]}}}\right)}\right] \\ \mu_{i_{j+1}}^{[j]2} &\in \left[\frac{1}{1-\epsilon_\psi^{[j]}}\left(t_{\Delta i_{j+1}}^{[j]2}+\tfrac{1}{1-\delta}\left\|\mathbf{W}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2\right), \frac{1}{\frac{s^{[j]2}}{\left(1-\epsilon_\phi^{[j]}\right)\sqrt{\rho^{[j]}}}+\frac{1}{H^{[j]}}\left(\frac{\left\|\mathbf{w}_{\mathcal{O}:i_{j+1}}^{[j]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]}\right\|_\infty^2}+H^{[j-1]}\right)\frac{\psi^{[j-1]}}{H^{[j-1]}}\frac{\phi^{[j-1]}}{H^{[j-1]}}}\right] \end{aligned} \tag{50}$$

$\forall i_1, \forall j \in \mathbb{N}_D \backslash \{0\}$，$i_{j+1}$，那么范数将收敛。我们必须证明，我们的规范缩放所用的缩放因子满足收敛所需的界限。首先考虑 $\mu_{i_D}^{[D-1]2}$。在此情况下，根据 (47)，我们要求：

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

以及 $\forall i_D$：

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

我们通过以下选择，在满足第一个界的同时最大化该界的右侧：

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

因此条件变为（第一个条件是上述右侧为正，第二个条件则保证不等式成立）：

$$\begin{aligned} \epsilon_\psi^{[D-1]} &< 1-(1-\chi)\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}\left\|t_{\Delta}^{[D-1]\boxtimes}\right\|_\infty^2 \\ \frac{\phi^{[D-2]}}{H^{[D-2]}} &\leq \frac{\frac{1}{1-\chi}\frac{1-\epsilon_\psi^{[D-1]}}{\left\|\mathbf{t}_{\Delta}^{[D-1]\boxtimes}\right\|_\infty^2}-\frac{s^{[D-1]2}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}}{\frac{1}{H^{[D-1]}}\max_{i_D}\left\{\frac{\left\|\mathbf{w}_{\mathcal{O}:i_D}^{[D-1]}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_D}^{[D-1]\boxtimes}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_D}^{[D-1]\boxtimes}\right\|_\infty^2}+H^{[D-2]}\right\}\frac{\frac{\psi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_\phi^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}} \end{aligned}$$

或者，等价地：

$$
\begin{aligned}
\nu &= \frac{4}{\left\|\mathbf{t}_{\Delta}^{[D-1]\boxplus}\right\|_{\infty}^{2}}\kappa\left(\frac{1-\epsilon_{\psi}^{[D-1]}}{1-\chi}\right)\\
\epsilon_{\psi}^{[D-1]} &\in \left(0, 1-(1-\chi)\frac{s^{[D-1]2}}{\left(1-\epsilon_{\phi}^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}\left\|t_{\Delta}^{[D-1]\boxplus}\right\|_{\infty}^{2}\right)\\
\epsilon_{\phi}^{[D-2]} &\in \left(1-\frac{1}{\sqrt{\rho^{[D-2]}}}\overline{\sigma}^{[D-2]-1}\left(\frac{\frac{1}{1-\chi}\frac{1-\epsilon_{\psi}^{[D-1]}}{\left\|\mathbf{t}_{\Delta}^{[D-1]\boxplus}\right\|_{\infty}^{2}}-\frac{s^{[D-1]2}}{\left(1-\epsilon_{\phi}^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}}{\frac{1}{H^{[D-1]}}\max\limits_{i_D}\left\{\frac{\left\|\mathbf{w}_{\mathcal{O}:i_D}^{[D-1]\boxplus}\right\|_2^2}{\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta:i_D}^{[D-1]\boxplus}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_D}^{[D-1]\boxplus}\right\|_{\infty}^2}+H^{[D-2]}\right\}\frac{\frac{\psi^{[D-2]}}{H^{[D-2]}}}{\left(1-\epsilon_{\phi}^{[D-1]}\right)\sqrt{\rho^{[D-1]}}}}\right),1\right)
\end{aligned}
\tag{51}
$$

由此我们得到第二个附加约束：

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

接下来我们考虑第 $0<j<D-1$ 层。为满足（50），根据（47），我们要求：

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

因此只需满足：

$$
\begin{aligned}
&\frac{t_{\Delta i_{j+1}}^{[j]\boxplus 2}}{t_{\Delta i_{j+1}}^{[j]\boxplus 2}-\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta}^{[j+1]\boxplus}\right\|_F^2}t_{\Delta i_{j+1}}^{[j]\boxplus 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]\boxplus}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]\boxplus}\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]\boxplus}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]\boxplus}\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]\boxplus 2}-\frac{1}{1-\delta}\left\|\mathbf{W}_{\Delta}^{[j+1]\boxplus}\right\|_F^2}{t_{\Delta i_{j+1}}^{[j]\boxplus 4}} \quad \forall i_{j+1}\\
\Rightarrow\quad & \frac{\phi^{[j-1]}}{H^{[j-1]}} \leq \frac{\frac{t_{\Delta i_{j+1}}^{[j]\boxplus 2}-\frac{1}{1-\delta}\left\|\mathbf{w}_{\Delta}^{[j+1]\boxplus}\right\|_F^2}{t_{\Delta i_{j+1}}^{[j]\boxplus 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]\boxplus}\right\|_2^2-\left\|\mathbf{w}_{\Delta:i_{j+1}}^{[j]\boxplus}\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}
$$

因此只需令：[14]

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

或者，等价地：

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

其中，正性要求我们进一步约束权重步长，以确保该值为正。根据（48），正性要求第三个附加约束：

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

[14]注意，若 $0<a<x_{\min}\leq x$，则通过设置 $x=x_{\min}$可使 $\frac{x-a}{x^2}$ 最小化。

最后，我们考虑 $j=0$的情况。我们要求：

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

and:

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

因此，只需令：

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

或者，等价地：

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

以及我们的第四个附加约束：

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

最后还有一个技术细节。选择

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

将得到：

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

这是不允许的，因为 $\epsilon_{\phi}^{[D-2]}<1$ （也就是说， $\epsilon_{\phi}^{[D-2]}$ 的值域在极限情况下会变成空集）。因此，我们进一步收紧对 $\epsilon_{\psi}^{[D-1]}$ 的约束：

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

经过一些小幅调整后，证明即告完成。 □

由此可得论文正文中使用的推论：

推论 10. 设 $\epsilon,\chi\in(0,1)$，对于给定的初始权重为 $\mathbf{W}_{\mathcal{O}}$ 的神经网络，令$\mathbf{W}_{\Delta}^{\boxplus}$为由反向传播得到的权重步长。令：

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

对于某个 $\alpha^{[j]}\in\mathbb{R}_+$：

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

如果权重步长满足：

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

那么存在一种规范缩放：

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

其中：

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

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

sa

证明。 我们从定理 9 开始。设 $\varpi^{[j]}=1-\chi$ 和 $\epsilon_{\phi}^{[D-1]}=\epsilon_{\phi\circ}$ 定理 9 中的定义变为：

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

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

and:

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

随后，对于某个 $\alpha^{[j]}\in\mathbb{R}_{+}$：

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

权重步长的条件满足：

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

这意味着存在一种规范缩放：

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

其中：

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

满足 $\|\quad\boldsymbol{\Phi}_{\mathcal{O}}(\mathbf{x})\|_F^2 \leq \phi^{[D-1]},\ \|\boldsymbol{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta})\|_F^2 \leq \psi^{[D-1]}$。为进一步简化约束，注意：

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

因此，我们可以将 $\epsilon_{\phi}^{[j]}$ 和 $\|\mathbf{t}_{\Delta}^{[j]\circledast}\|_{\infty}^{2}$ 的约束收紧为：

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

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

随后，定理 9 中的定义可以收紧为以下形式，同时注意确保 $\left\|\mathbf{t}_{\Delta}^{[j]\circledast}\right\|_{\infty}^{2}$ 不会受到过度约束：

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

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

随后，对于某个 $\alpha^{[j]}\in\mathbb{R}_{+}$：

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

权重步长需满足以下条件（第一个约束是确保 $\epsilon_{\psi}>0$ 所必需的）：

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

这意味着规范缩放的存在：

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

其中：

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

满足 $\|\mathbf{\Phi}_{\mathcal{O}}\mathbf{x}\|_{F}^{2}\leq H^{[D-1]}\sigma^{[D-1]}\underline{\left(\left(1-\epsilon_{\phi}\right)\sqrt{\rho^{[D-1]}}\right)}$，$\underline{\|\mathbf{\Psi}_{\mathcal{O}}\mathbf{W}_{\Delta}\|_{F}^{2}}\leq H^{[D-1]}\frac{1-\epsilon_{\psi}}{\epsilon_{\psi}}$。

此时，我们的构造还有足够的灵活性，可以令 $\delta\to 0$，并通过 $\epsilon_{\phi}=\epsilon$ 加以整理，因此约束变为：

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

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

随后对某个 $\alpha^{[j]}\in\mathbb{R}_{+}$：

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

The 权重步长满足的条件（需要第一个约束以确保 $\epsilon_{\psi}>0$）：

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

这意味着规范缩放的存在：

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

其中：

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

满足 $\|\mathbf{\Phi}_{\mathcal{O}}(\mathbf{x})\|_{F}^{2}\leq H^{[D-1]}\sigma^{[D-1]}\underline{\left(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right)}$，$\underline{\|\mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_{\Delta})\|_{F}^{2}}\leq H^{[D-1]}\frac{1-\epsilon_{\psi}}{\epsilon_{\psi}}$。

最后，我们需要确保 ν 定义良好，为此要求 $1-\epsilon_\psi < 1-\chi$，或者等价地， $\epsilon_\psi > \chi$。对此，只需满足：

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

因此，稍微收紧界：

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

由此完成证明。 □

随后我们得到拉德马赫复杂度的一个界：

定理 11. 令 $\epsilon, \chi \in (0,1)$，对于一个初始权重为$\mathbf{W}_\mathcal{O}$的给定神经网络，令$\mathbf{W}_\Delta^{\circledast}$为由反向传播得到的权重步长，并满足推论 *10* 中给出的条件。则 $f_\Delta^{\circledast} \in \mathcal{F}^\bullet$，其中 $\mathcal{F}^\bullet$ 的拉德马赫复杂度满足以下界：

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

证明。根据推论 10，我们知道，在满足关于权重步长大小的假设时，必定存在缩放因子、影子权重和正则化参数 λ，使得反向传播导致的神经网络运作变化 $f_\Delta^{\circledast}$ 与最小化正则化风险 R_λ 所导致的神经网络运作变化 $f_\Delta^\bullet$ 相一致，因此 $f_\Delta^{\circledast} = f_\Delta^\bullet \in \mathcal{B}_\mathcal{O}$。此外，固定这些参数后，根据推论 10，我们知道$\|\mathbf{\Phi}_\mathcal{O}(\mathbf{x})\|_F^2 \leq H^{[D-1]}\sigma^{[D-1]}\left(\left(1-\epsilon\right)\sqrt{\rho^{[D-1]}}\right)$ $\forall \mathbf{x} \in \mathbb{X}$, $\|\mathbf{\Psi}_\mathcal{O}(\mathbf{W}_\Delta)\|_F^2 \preceq H^{[D-1]}\frac{1-\epsilon_\psi}{\epsilon_\psi}$，因此：

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

因此，对于拉德马赫随机变量 ϵ，拉德马赫复杂度有如下界：

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

与数据分布 ν 无关。

接下来回顾推论 10 中的定义：

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

因此：

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

因此：

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

□

	本文中的记号	Lin 等人 [使用的记号， 2022]
数据空间:	$\mathbb{X} \subset \mathbb{R}^n$	Ω_1 （输入空间）
权重步长空间:	$\mathbb{W}_{\mathcal{O}} \subset \prod_{j \in \mathbb{N}_D} \mathbb{R}^{H^{[j-1]} \times H^{[j]}} \times \mathbb{R}^{H^{[j]}}$	Ω_2 （权重空间）
数据特征映射:	$\mathbf{\Phi}_{\mathcal{O}}: \mathbb{X} \to \mathcal{X}_{\mathcal{O}} \subset \mathbb{R}^{\infty \times m}$	$\Phi_1: \Omega_1 \to \mathcal{W}_1$
权重步长特征映射:	$\mathbf{\Psi}_{\mathcal{O}}: \mathbb{W}_{\mathcal{O}} \to \mathcal{W}_{\mathcal{O}} \subset \mathbb{R}^{\infty \times m}$	$\Phi_2: \Omega_2 \to \mathcal{W}_2$
数据巴拿赫空间:	$\mathcal{X}_{\mathcal{O}} = \operatorname{span}(\mathbf{\Phi}_{\mathcal{O}}(\mathbb{X})) \subset \mathbb{R}^{\infty \times m}$，其中 $\|\cdot\|_{\mathcal{X}_{\mathcal{O}}} = \|\cdot\|_F$	$\mathcal{W}_1$ 范数为 $\|\cdot\|_{\mathcal{W}_1}$
权重步巴拿赫空间:	$\mathcal{W}_{\mathcal{O}} = \operatorname{span}(\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}})) \subset \mathbb{R}^{\infty \times m}$，其中 $\|\cdot\|_{\mathcal{W}_{\mathcal{O}}} = \|\cdot\|_F$	$\mathcal{W}_2$ 范数为 $\|\cdot\|_{\mathcal{W}_2}$
双线性形式:	$\langle\cdot,\cdot\rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}}: \mathbb{R}^{\infty \times m} \times \mathbb{R}^{\infty \times m} \to \mathbb{R}^m$, $\langle \mathbf{\Omega}\ \mathbf{\Xi} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} = \text{对角}\ (\mathbf{\Omega}^{\mathrm{T}} \mathbf{\Xi})$	$\langle\cdot,\cdot\rangle_{\mathcal{W}_1 \times \mathcal{W}_2}: \mathcal{W}_1 \times \mathcal{W}_2 \to \mathbb{Y}$

图 4: 根据 [Lin等人， 2022]对再生核巴拿赫空间构造的总结。

E 神经网络与再生核巴拿赫空间

再生核希尔伯特空间（RKHS） [Aronszajn， 1950] 是一个函数希尔伯特空间 $\mathcal{H}$ $f: \mathbb{X} \to \mathbb{Y}$ 其中点求值函数是连续的。因此，应用 Riesz表示定理，存在一个核 K，使得:

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

对所有 $f \in \mathcal{H}$。随后， $K(\mathbf{x}, \mathbf{x}') = \langle K(\mathbf{x} \cdot) K(\mathbf{x}', \cdot) \rangle$ 并且，根据 Moore-Aronszajn 定理， K 由 $\mathcal{H}$ 唯一确定，反之亦然。 K 称为再生核，对应的再生核希尔伯特空间记作 $\mathcal{H}_K$。再生核希尔伯特空间在机器学习中日益普及，因为它们非常适用于机器学习的诸多方面 [Steinwart 和 Christman，2008；Shawe-Taylor 和 Cristianini，2004；Cortes 和 Vapnik，1995；Chowdhury 和 Gopalan，2017；Cristianini 和 Shawe-Taylor，2005；Genton，2001；G¨onen 和 Alpaydin，2011；Herbrich，2002；Li 等人，2017；M¨uller 等人，2001；Smola 和 Sch¨olkopf， 1998]。内积结构使核技巧成为可能，这在实践中非常有用，而且核本身很容易被理解为一种相似性度量。更重要的是，再生核希尔伯特空间的结构催生了丰富的复杂度分析和泛化界框架 [Steinwart 和 Christman，2008；Shawe-Taylor 和 Cristianini， 2004] ，为这一机器学习分支奠定了基础。近来，借助神经切线核理论，这一框架已扩展至神经网络 [Jacot等人， 2018]。

再生核巴拿赫空间（RKBS）是再生核希尔伯特空间的一种推广，它以函数巴拿赫空间而非希尔伯特空间为基础 [Der and Lee, 2007, Lin 等，2022, Zhang et al., 2009, Zhang and Zhang, 2012, Song et al., 2013, Sriperumbudur et al., 2011, Xu and Ye, 2014]。具体而言:

定义4 （再生核巴拿赫空间（RKBS， [Lin 等， 2022]））。 定义在集合 $\mathbb{X}$ 上的再生核巴拿赫空间 $\mathcal{B}$是函数空间 $f: \mathbb{X} \to \mathbb{Y}$ ，且其中每个点求值映射 $\delta_{\mathbf{x}}: \mathcal{B} \to \mathbb{Y}$， $\mathbf{x} \in \mathbb{X}$，在 $\mathcal{B}$ 上都是连续的。也就是说， $\exists C_{\mathbf{x}} \in \mathbb{R}_+$ 使得:

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

对所有 $f \in \mathcal{B}$均成立。

这引入了更丰富的几何结构，并为常见的核框架带来了令人振奋的新拓展，例如非对称核、特征空间中的稀疏学习、统计学中的套索回归以及 m-核。在当前语境下，它使我们能够将神经切线核从过参数化状态下的一阶近似扩展为精确表示，而无需进行无限宽度近似。文献中有许多再生核巴拿赫空间理论的研究方法，但在当前语境下，我们发现 [Lin 等， 2022] 的方法最为有用。

在这种表述中，再生核可以由一组基本要素构造而成，这些要素与我们在前几节中构造$\mathbf{f}_\Delta$ 的方式十分相似。具体而言，根据 [Lin等人 2022], 的结果，并结合引理4，我们可以利用图 4 中定义的要素构造一个包含$\mathbf{f}_\Delta$ 的再生核巴拿赫空间 $\mathcal{B}_{\mathcal{O}}$ 。由此，我们得到定义在 $\mathbb{X}$上的再生核巴拿赫空间 $\mathcal{B}_{\mathcal{O}}$:

$$\mathcal{B}_{\mathcal{O}} = \left\{ \langle \mathbf{\Phi}_{\mathcal{O}}(\cdot), \mathbf{\Omega} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \middle| \mathbf{\Omega} \in \mathcal{W}_{\mathcal{O}} \right\}, \text{ where } \left\| \langle \mathbf{\Phi}_{\mathcal{O}}(\cdot), \mathbf{\Omega} \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \right\|_{\mathcal{B}_{\mathcal{O}}} = \|\mathbf{\Omega}\|_{\mathcal{W}_{\mathcal{O}}} \tag{52}$$

$\mathcal{B}_{\mathcal{O}}$ 中的元素包括函数$\mathbf{f}_\Delta : \mathbb{R}^n \to \mathbb{R}^m$ - 即在满足收敛条件的情况下，单次迭代中权重和偏置发生变化所引起的神经网络行为变化$\mathbf{W}_\Delta \in \mathbb{W}_{\mathcal{O}}$ ——以及它们的所有线性组合（回顾一下， $\mathcal{W}_{\mathcal{O}} = \operatorname{span}(\mathbf{\Psi}_{\mathcal{O}}(\mathbb{W}_{\mathcal{O}}))$）。其再生核为$\mathbf{K}_{\mathcal{O}}$: $\mathbb{X} \times \mathbb{W}_{\mathcal{O}} \to \mathbb{R}^{m \times m}$:

$$\mathbf{K}_{\mathcal{O}}(\mathbf{x}, \mathbf{W}_\Delta) = \operatorname{diag}\left(\langle \mathbf{\Phi}_{\mathcal{O}}(\mathbf{x}), \mathbf{\Psi}_{\mathcal{O}}(\mathbf{W}_\Delta) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \right) = \operatorname{diag}(\mathbf{f}_\Delta(\mathbf{x})) \tag{53}$$

我们还得到了定义在 $\mathbb{W}_{\mathcal{O}}$上的相应再生核巴拿赫空间 $\mathcal{B}_{\mathcal{O}}^*$:

$$\mathcal{B}_{\mathcal{O}}^* = \left\{ \langle \mathbf{\Xi}, \mathbf{\Psi}_{\mathcal{O}}(\cdot) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \middle| \mathbf{\Xi} \in \mathcal{X}_{\mathcal{O}} \right\}, \text{ where } \left\| \langle \mathbf{\Xi}, \mathbf{\Psi}_{\mathcal{O}}(\cdot) \rangle_{\mathcal{X}_{\mathcal{O}} \times \mathcal{W}_{\mathcal{O}}} \right\|_{\mathcal{B}_{\mathcal{O}}^*} = \|\mathbf{\Xi}\|_{\mathcal{X}_{\mathcal{O}}} \tag{54}$$

其元素是 $\mathcal{B}_{\mathcal{O}}$ 的求值泛函，它具有如下再生核 $\mathbf{K}_{\mathcal{O}}^*$: $\mathbb{W}_{\mathcal{O}} \times \mathbb{X} \to \mathbb{R}^{m \times m}$ 其定义为:

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

F 案例研究——双曲正切激活函数

考虑如下激活函数：

$$\tau^{[j]}(\zeta) = \tanh(\zeta) \tag{55}$$

我们希望为任意 $\sigma_z^{[j]}{}_{,z'}$ 构造相应的 $z, z' \in \mathbb{R}$。正如我们在 F.1 节所示，双曲正切在任意 $z \in \mathbb{R}$ 处的泰勒展开为：

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

其中：

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

该展开对所有满足 $|\zeta| \leq \frac{\pi}{2}$ 的 $z, \zeta \in \mathbb{R}$ 均成立。因此，根据定义 (27)，利用这一展开：

$$\sigma_{(z,z')}^{[j]}(\zeta) = \sum_{m=1}^{\infty} a_{(z)m}^{\tanh} a_{(z')m}^{\tanh} \zeta^m \tag{56}$$

在 $z = z' = 0$ 的情况下，这可简化为：

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

其中 B_{2m} 为伯努利数。因此：

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

F.1 双曲正切在任意点处的泰勒展开

双曲正切在 0 处的泰勒展开广为人知，具体如下：

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

其中，B_{2n} 是伯努利数，该展开在 $|\xi| \leq \frac{\pi}{2}$ 时成立。然而，本文需要的是双曲正切在任意点 z 处的泰勒展开；但我们未能在文献中找到这样的展开。本节将采用一种受帖子 [D'Aurizio, 2014] 中所述方法启发的方法，推导出这样的展开。

我们知道，双曲余弦的魏尔斯特拉斯乘积展开为：

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

对于所有 z，$\xi \in \mathbb{R}$（这里的"符号吸收"将在后续推导中发挥重要作用）。因此：

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

于是，对两边求导数 $\frac{\partial}{\partial \xi}$：

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

其中 $\mathrm{i} = \sqrt{-1}$。因此，对所有 $\xi^2 \leq z^2 + (\frac{\pi}{2})^2$：

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

因此，重新排列我们的索引：

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

利用 Knapp 的结果 [, 2009]:

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

以及和与差的正弦（余弦）的通常结果，我们有：

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

and:

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

因此:

$$\begin{aligned}
&\operatorname{sgn}(z)\tanh(z+\xi)=\ldots\\
&=2(|z|+\operatorname{sgn}(z)\xi)\sum_{m=0}^{\infty}(-\operatorname{sgn}(z))^m\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(|z|+\operatorname{sgn}(z)\xi)\sum_{m=0}^{\infty}(-\operatorname{sgn}(z))^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}(-\operatorname{sgn}(z))^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\sum_{m=0}^{\infty}(-\operatorname{sgn}(z))^{m+1}\ldots\\
&\ldots\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}}\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)+\ldots\\
&\ldots2|z|\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^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\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\ldots\\
&\ldots\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}}\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)+\ldots\\
&\ldots2|z|\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^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\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\ldots\\
&\ldots\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)\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+\ldots\\
&\ldots2\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\ldots\\
&\ldots\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}}\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\\
&=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)+\ldots\\
&\ldots2|z|\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^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\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\ldots\\
&\ldots\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)\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{-\frac{1}{2}}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{1}{2}}|z|\right)\xi^m+\ldots\\
&\ldots2\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\ldots\\
&\ldots\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}}\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}}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{1}{2}}\left(n+\frac{1}{2}\right)\pi\right)\xi^m\\
&=2|z|\sum_{n=0}^{\infty}\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}+2|z|\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^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-\ldots\\
&\ldots2|z|\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^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+\ldots\\
&\ldots2\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\left(\sum_{n=0}^{\infty}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\cos\left((m+1)\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}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}+2\sum_{m=1}^{\infty}(-\operatorname{sgn}(z))^m\left(\sum_{n=0}^{\infty}\left(\frac{1}{z^2+\left(n+\frac{1}{2}\right)^2\pi^2}\right)^{\frac{m+1}{2}}\cos\left((m+1)\operatorname{atan}\left(\frac{\left(n+\frac{1}{2}\right)\pi}{|z|}\right)\right)\right)\xi^m
\end{aligned}$$

and so:

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

(作为快速合理性检查，令 $z \to 0^{\pm}$，则反正切函数($\frac{(n+\frac{1}{2})\pi}{|z|} \to \frac{\pi}{2}$且符号函数($z$) $= \pm 1$:

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

其中 ζ 是黎曼ζ函数， B_{2n} 是伯努利数，这就是双曲正切在 0 处的通常泰勒展开）。继续利用 [Gradshteyn 和 Ryzhik, 2000, （1.421.2）]:

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

因此，利用双曲正切是奇函数这一事实:

$$\tanh(z+\xi) = \tanh(z) + \textstyle\sum_{m=1}^{\infty} a_{(z)m}^{\tanh} \xi^m \tag{57}$$

其中利用了余弦函数和反正切函数是奇函数这一点:

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