arXiv:2303.03950v1 [cs.LG] 6 Mar 2023

具有 ReLU 激活函数的最优浅层前馈网络的存在性

STEFFEN DEREICH 和 SEBASTIAN KASSING

ABSTRACT. 我们证明，使用具有 ReLU 激活函数的浅层前馈人工神经网络逼近连续目标函数时，损失景观中存在全局最小值。这一性质是 ReLU 与其他常用激活函数的根本区别之一。我们提出一种搜索空间的闭包，使得在扩展空间中极小化元得以存在。接下来，我们在温和的假设下证明，扩展中新加入的函数，其表现不如适当的、可由 ReLU 网络表示的函数。由此可知，扩展目标空间中的最优响应实际上就是某个 ReLU 网络的响应函数。

1. INTRODUCTION

现代机器学习算法通常通过基于梯度的算法优化人工神经网络（ANNs）。这些方法在实际应用中取得的巨大成功，促使许多科学家为机器学习奠定数学基础，尤其是试图找出训练动力学中的普遍结构，以解释实践者观察到的令人惊叹的现象。人工神经网络的一个关键组成部分是激活函数。在已提出的各种激活函数中，整流线性单元（ReLU）定义为零与输入值中的较大者，已成为使用最广泛、效果最好的激活函数。ReLU 如此受欢迎有几个原因，例如，它易于实现、计算效率高，并且能够克服梯度消失问题；在训练人工神经网络时，这一问题常见于其他激活函数。本文指出并证明了 ReLU 函数的一项更为微妙的特性，使其区别于其他几种常见激活函数；这或许也是它在实践中广受欢迎的关键原因之一：优化景观中存在全局最小值。

一条热门研究路线是使用梯度下降（GD）类方法研究人工神经网络（ANN）的优化过程（也称为训练）。由于典型机器学习优化任务中的损失/误差函数具有非线性、非凸性，甚至是非强制的，因此，即使是在优化浅层人工神经网络（即仅包含一个隐藏层的人工神经网络）这一最简单的情形下，要严格证明（或否证）GD 的收敛性仍然是一个悬而未决的问题。现有的理论收敛性结果通常假设该过程保持有界，即对于每个实现，都存在一个紧集，使得训练过程中该过程不会离开此集合，参见例如 [BDL07、EJRW21、JR22]中关于梯度流的结果， [AMA05、AB09] 中关于确定性梯度方法的结果， [Tad15、DDKL20、MHKC20、DK21] 中关于随机梯度方法的结果，以及 [DK22] 中关于基于梯度的扩散过程的结果。许多结果都可追溯至 Lojasiewicz 关于解析目标函数的梯度不等式的经典工作

2020 *Mathematics Subject Classification.* 主要分类号：68T07；次要分类号：68T05、41A50。
Key words and phrases. 神经网络、浅层网络、最佳逼近、ReLU 激活函数、相对紧性。

函数，以及在假设过程保持局部性的条件下，梯度流轨迹收敛性的直接推论 [Loj63、Loj65、Loj84]。

在这一背景下，自然会提出这样的问题：是否存在能够在搜索空间内解决最小化任务的人工神经网络。更具体地说，如果优化景观中不存在全局最小值，那么任何趋近于最小损失值的序列都会发散至无穷。这可能导致收敛速度缓慢，甚至使损失值无法收敛，而这正是实践者最为关注的性质。因此，选择一种网络架构、激活函数和损失函数，使得优化景观中存在全局最优解，似乎是合理的。

在经验风险最小化的设定下，过参数化人工神经网络（隐藏层中的神经元数量多于待拟合的数据点）通常能够对有限个数据点进行完美插值，从而存在一个实现零误差的网络配置，并因此在搜索空间中得到一个全局最小值。然而，对于一般测度 μ （不一定由有限个狄拉克测度组成），关于全局最小值存在性的文献非常有限。已有一些关于使用海维赛德激活函数 [KKV03], 的浅层前馈人工神经网络逼近空间 $L_p([0,1]^d, \|\cdot\|_p)$ 中函数的正面结果；关于使用 ReLU 激活函数和标准均方误差 [JR22], 的浅层前馈人工神经网络逼近一维利普希茨连续目标函数的结果；以及关于使用 ReLU 激活函数的浅层残差人工神经网络逼近多维连续目标函数的结果，参见 [DJK23]。相反，对于若干常见的（光滑）激活函数，例如标准逻辑斯谛激活函数、Softplus、反正切函数、双曲正切函数和 Softsign，对于光滑目标函数（甚至多项式），优化景观中通常不存在极小化元，参见 [PRV21] 和 [GJL22]。在双曲正切激活函数的经验风险最小化中也可以观察到这一现象。正如 [LMQ22], 所示，在欠参数化设定下，存在这样的输入数据：对于来自某个具有正勒贝格测度的集合的所有输出数据，优化景观中都不存在极小化元。对于 ReLU 激活函数，这一现象是否仍然存在，尚未得到解决。

本文首次证明了具有多维输入空间的浅层前馈 ReLU 人工神经网络的存在性结果。有趣的是，在对优化问题作出非常温和的假设下，极小化元依然存在。这一存在性表明 ReLU 激活函数具有稳健性，这或许也是其在实际应用中取得成功的原因之一。证明过程如下。首先，我们证明，在一个扩展目标空间中极小化元的存在性；该空间由人工神经网络的可表示响应以及额外的不连续广义响应组成。请注意，对于许多激活函数（包括 ReLU），实现函数/响应函数的集合在适当的度量下并不闭，另见 [GP90]。其次，我们证明，在对优化问题的温和条件下，额外的不连续响应的表现不如可表示响应。与 [DJK23], 中对残差网络的研究相比，本文所讨论的经典前馈 ReLU 网络的情况更加复杂。这是由于扩展搜索空间具有更为精细的结构，见定义 2.1。

在我们关注使用 ReLU 激活函数的浅层人工神经网络在 L_p-损失下对连续目标函数进行逼近的情形下，我们给出了主要结果的一个特例。

定理 1.1. 设 $d_{\text{in}}, d \in \mathbb{N}, p > 1$ 和 $\mathfrak{d} = (d_{\text{in}}+2)d+1$。设 $f\colon \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ 和 $h\colon \mathbb{R}^{d_{\text{in}}} \to [0,\infty)$ 为连续函数，并假设 $h^{-1}((0,\infty))$ 是一个有界凸集。对于每个 $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$，令 $\mathrm{err}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ 由 $\mathrm{err}(\theta) = \int_{\mathbb{R}^{d_{\text{in}}}} |f(x) - \mathfrak{N}_\theta(x)|^p h(x)\, dx$ 给出，其中

$$\mathfrak{N}_\theta(x) = \theta_{\mathfrak{d}} + \big(\textstyle\sum_{j=1}^{d} \theta_{(d_{\text{in}}+1)d+j} \max\{\theta_{d_{\text{in}}d+j} + \sum_{i=1}^{d_{\text{in}}} \theta_{(j-1)d_{\text{in}}+i} x_i, 0\}\big).$$

那么存在 $\theta \in \mathbb{R}^{\mathfrak{d}}$，使得 $\mathrm{err}(\theta) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathrm{err}(\vartheta)$。

定理 1.1 是更一般的定理 1.2 的特例，后者处理更广泛的一类损失函数和测度。

下面介绍本文的核心对象和记号。接下来，我们将以更加结构化的方式表示人工神经网络。我们考虑具有 d_{in} 维输入空间的网络，其中包含一个隐藏层，该层由 d 个应用 ReLU 激活函数的神经元组成，即 $(x)^+ = \max(x, 0)$。我们用矩阵 $W^1 = (w^1_{j,\,i})_{j=1,\,\ldots,\,d,\;i=1,\,\ldots,\,d_{\mathrm{in}}}$ 和行向量 $W^2 = (w^2_1, \ldots, w^2_d)$ 描述人工神经网络的权重，并用列向量 $b^1 = (b^1_i)_{i=1,\,\ldots,\,d}$ 和标量 b^2 描述偏置。此外，对于 $j = 1, \ldots, d$，我们记 $w^1_j = (w^1_{j,\,1}, \ldots, w^1_{j,\,d_{\mathrm{in}}})^\dagger$，其中 $a^\dagger$ 表示向量或矩阵 a 的转置。令

$$\mathbb{W} = (W^1, b^1, W^2, b^2) \in \mathbb{R}^{d \times d_{\mathrm{in}}} \times \mathbb{R}^d \times \mathbb{R}^{1 \times d} \times \mathbb{R} =: \mathcal{W}_d$$

并称 $\mathbb{W}$ 为网络配置，$\mathcal{W}_d$ 为参数化类。我们常将神经网络的一种配置称为（神经）网络 $\mathbb{W}$。配置 $\mathbb{W} \in \mathcal{W}_d$ 描述了一个函数 $\mathfrak{N}^{\mathbb{W}} \colon \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}$，其定义如下：

$$\mathfrak{N}^{\mathbb{W}}(x) = \sum_{j=1}^{d} w^2_j \big(w^1_j \cdot x + b^1_j\big)^+ + b^2, \tag{1}$$

其中 $\cdot$ 表示 $\mathbb{R}^{d_{\mathrm{in}}}$上的内积。我们将 $\mathfrak{N}^{\mathbb{W}}$ 称为网络 $\mathbb{W}$的实现函数或响应函数。我们允许参数 d取自 $\mathbb{N}_0$中的所有值：$= \{0, 1, \ldots\}$，其中，按定义，零神经元网络的响应函数是一个常函数。关于可能包含多个隐藏层的一般神经网络的介绍，参见例如 [PRV21]。

请注意，一般而言，网络的响应函数是一个从 $\mathbb{R}^{d_{\mathrm{in}}}$ 到 $\mathbb{R}$的连续分段仿射函数。我们将 $\mathbb{W} \mapsto \mathfrak{N}^{\mathbb{W}}$ 视为一类潜在响应函数 $\{\mathfrak{N}^{\mathbb{W}} \colon \mathbb{W} \in \mathcal{W}_d\}$ 的参数化，它们用于一个最小化问题。更明确地说，设 μ 是定义在 $\mathbb{R}^{d_{\mathrm{in}}}$的博雷尔集上的有限测度，设 $\mathbb{D} = \mathrm{supp}(\mu)$，且 $\mathcal{L} \colon \mathbb{D} \times \mathbb{R} \to \mathbb{R}_+$ 是一个乘积可测函数，即损失函数。我们的目标是最小化误差

$$\mathrm{err}^{\mathcal{L}}(\mathbb{W}) = \int_{\mathbb{D}} \mathcal{L}(x, \mathfrak{N}^{\mathbb{W}}(x)) \, \mathrm{d}\mu(x)$$

对所有 $\mathbb{W} \in \mathcal{W}_d$，给定 $d \in \mathbb{N}_0$，并令

$$\mathrm{err}^{\mathcal{L}}_d = \inf_{\mathbb{W} \in \mathcal{W}_d} \mathrm{err}^{\mathcal{L}}(\mathbb{W}) \tag{2}$$

为使用隐藏层含有 d 个神经元的神经网络时该优化任务的最小误差。

本文旨在给出关于损失函数 $\mathcal{L}$ 和测度 μ 的充分条件，以保证存在一个网络 $\mathbb{W} \in \mathcal{W}_d$，使得 $\mathrm{err}^{\mathcal{L}}(\mathbb{W}) = \mathrm{err}^{\mathcal{L}}_d$。我们强调，如果不存在满足 $\mathrm{err}^{\mathcal{L}}(\mathbb{W}) = \mathrm{err}^{\mathcal{L}}_d$ 的神经网络 $\mathbb{W} \in \mathcal{W}_d$，那么每个满足 $\lim_{n\to\infty} \mathrm{err}^{\mathcal{L}}(\mathbb{W}_n) = \mathrm{err}^{\mathcal{L}}_d$ 的网络序列 $(\mathbb{W}_n)_{n \in \mathbb{N}} \subset \mathcal{W}_d$ 都会发散至无穷大。我们的框架涵盖了使用 L_p-损失对连续函数进行逼近的情形，见定理 1.1。

关于赋范空间中最佳逼近元的一般介绍，请读者参阅 [Sin70]。关于神经网络训练中损失景观的优秀文献综述可见于 [EMWW20]。关于（浅层）网络训练中存在非最优局部极小值的相关论述，请读者参阅 [SCP16、SS18、VBB19] 和 [CK22]。

我们陈述本文的主要结果。

定理 1.2. 假设 $\mathbb{D} = \operatorname{supp}(\mu)$ 是紧集，且 μ 具有连续的勒贝格密度 $h : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}_+$。假设对于每个与 $\mathbb{D}$ 的凸包内部相交的超平面 H，都存在一个 $x \in H$，使得 $h(x) > 0$。此外，假设损失函数 $\mathcal{L} : \mathbb{D} \times \mathbb{R} \to \mathbb{R}_+$ 满足以下条件：

（i）（关于第一参数的连续性）对于每个 $y \in \mathbb{R}$，$\mathbb{D} \ni x \mapsto \mathcal{L}(x, y)$ 是连续的。（ii）（关于第二参数的严格凸性）对于所有 $x \in \mathbb{D}$，$y \mapsto \mathcal{L}(x, y)$ 是严格凸的，并且能取得其最小值。

那么，对于每个 $d \in \mathbb{N}_0$，存在一个最优网络 $\mathbb{W} \in \mathcal{W}_d$，使得 $\operatorname{err}^{\mathcal{L}}(\mathbb{W}) = \operatorname{err}_d^{\mathcal{L}}$.

定理 1.2 是下面命题 3.1 的直接推论。我们强调，命题 3.1 的表述更强，因为它甚至表明，在许多情况下，新加入扩展目标空间的函数表现得严格劣于可表示响应。

例 1.3（回归问题）。设 μ 如定理 1.2 中所述，并假设 $f : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ 是一个连续函数，且 $L : \mathbb{R} \to \mathbb{R}_+$ 是一个能取得其最小值的严格凸函数。那么，$\mathcal{L} : \mathbb{R}^{d_{\text{in}}} \times \mathbb{R} \to \mathbb{R}_+$ 定义为

$$\mathcal{L}(x, y) = L(y - f(x))$$

满足后一个定理的假设，因此下确界

$$\inf_{\mathbb{W} \in \mathcal{W}_d} \int L(\mathfrak{N}^{\mathbb{W}}(x) - f(x)) \, \mathrm{d}\mu(x)$$

在网络 $\mathbb{W} \in \mathcal{W}_d$ 上取到。

2. Generalized response of neural networks

我们将采用更直观的几何描述来刻画网络 $\mathbb{W} \in \mathcal{W}_d$ 的实现函数，正如 [DJK23]中所引入的那样。若对所有 $j = 1, \dots, d$ 都有 $w_j^1 \neq 0$，则称网络 $\mathbb{W} \in \mathcal{W}_d$ 为非退化网络。对于一个非退化网络 $\mathbb{W}$，我们称神经元 $j \in \{1, \dots, d\}$ 具有

- *normal* $\mathfrak{n}_j = \dfrac{1}{|w_j^1|} w_j^1 \in \mathcal{O} := \{x \in \mathbb{R}^{d_{\text{in}}} : |x| = 1\}$,
- *offset* $o_j = -\dfrac{1}{|w_j^1|} b_j^1 \in \mathbb{R}$ and
- *kink* $\Delta_j = |w_j^1| w_j^2 \in \mathbb{R}$.

此外，我们称 $\mathfrak{b} = b^2$ 为 $\mathbb{W}$ 的偏置。我们称满足 $\mathfrak{n} =$ 的 $(\mathfrak{n}, o, \Delta, \mathfrak{b})$、$\in \mathcal{O}^d(\mathfrak{n}_1, \dots, \mathfrak{n}_d)$、$o = (o_1, \dots, o_d)$、$\in \mathbb{R}^d(\Delta = , \dots, \Delta_1)$ 和 $\in \mathbb{R}^d$ $\mathfrak{b} \in \mathbb{R}$ 的集合为 $\mathbb{W}$ 的有效元组，并将使用 d 个 ReLU 神经元时所有有效元组的集合记作 $\mathcal{E}_d$。

首先注意到，非退化网络的响应函数 $\mathbb{W}$ 可以用其有效元组表示：对于 $x \in \mathbb{R}^{d_{\text{in}}}$，有

$$\begin{aligned}\mathfrak{N}^{\mathbb{W}}(x) &= \mathfrak{b} + \sum_{j=1}^{d} w_j^2 (w_j^1 \cdot x + b_j^1)^+ = \mathfrak{b} + \sum_{j=1}^{d} \Delta_j \Big(\frac{1}{|w_j^1|} w_j^1 \cdot x + \frac{1}{|w_j^1|} b_j^1 \Big)^+ \\ &= \mathfrak{b} + \sum_{j=1}^{d} \Delta_j \big(\mathfrak{n}_j \cdot x - o_j \big)^+ .\end{aligned}$$

稍微滥用一下记号，我们也写作

$$\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R},\ x \mapsto \mathfrak{b} + \sum_{j=1}^{d} \Delta_j \big(\mathfrak{n}_j \cdot x - o_j\big)^+$$

以及 $\mathrm{err}^{\mathcal{L}}(\mathfrak{n}, o, \Delta, \mathfrak{b}) = \int \mathcal{L}(x, \mathfrak{N}^{\mathfrak{n},\ o,\ \Delta,\ \mathfrak{b}}(x))\, \mathrm{d}\mu(x)$。尽管元组（$\mathfrak{n}, o, \Delta, \mathfrak{b}$）无法唯一描述一个神经网络，但它能唯一描述一个响应函数，因此我们将讨论具有有效元组（$\mathfrak{n}, o, \Delta, \mathfrak{b}$）的神经网络。

我们强调，退化网络 $\mathbb{W}$ 的响应函数也可以描述为与一个有效元组相关联的响应函数。事实上，对于每个 $j \in \{1, \dots, d\}$ 满足 $w_j^1 = 0$ 的情形，相应神经元的贡献 $w_j^2(b_j^1)^+$是一个常数。现在，可以任意选择一个法向量 $\mathfrak{n}_j$和偏移量 o_j，将折点设为零（$\Delta_j = 0$），并将常数 $w_j^2(b_j^1)^+$加到偏置 $\mathfrak{b}$上。对每个这样的神经元重复此过程，我们得到一个有效元组（$\mathfrak{n}, o, \Delta, \mathfrak{b}$）$\in \mathcal{E}_d$，其满足 $\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}} = \mathfrak{N}^{\mathbb{W}}$。反过来，对于每个有效元组（$\mathfrak{n}, o, \Delta, \mathfrak{b}$）$\in \mathcal{E}_d$, $\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}$ 都是某个适当网络 $\mathbb{W} \in \mathcal{W}_d$的响应函数。事实上，可以选择 $b^2 = \mathfrak{b}$，并且对于 $j = 1, \dots, d$，选择 $w_j^1 = \mathfrak{n}_j$ 和 $b_j^1 = -o_j$，使得对于所有 $x \in \mathbb{R}^{d_{\text{in}}}$

$$w_j^2(w_j^1 \cdot x + b_j^1)^+ = \Delta_j(\mathfrak{n}_j \cdot x - o_j)^+.$$

这意味着

$$\mathrm{err}_d^{\mathcal{L}} = \inf_{(\mathfrak{n},o,\Delta,\mathfrak{b})\in\mathcal{E}_d} \int \mathcal{L}(x, \mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}(x))\, \mathrm{d}\mu(x)$$

当且仅当存在一个网络 $\mathbb{W} \in \mathcal{W}_d$ 使得式 (2) 中的下确界能够取到时，下确界才能取到。对于一个有效元组 $(\mathfrak{n}, o, \Delta, \mathfrak{b}) \in \mathcal{E}_d$，我们称第 j 个 ReLU 神经元的折线为

$$H_j = \big\{x \in \mathbb{R}^{d_{\text{in}}} : \mathfrak{n}_j \cdot x = o_j\big\}$$

并将

$$A_j = \{x \in \mathbb{R}^{d_{\text{in}}} : \mathfrak{n}_j \cdot x > o_j\}$$

称为第 j 个 ReLU 神经元的活动域。根据构造，我们有

$$\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}(x) = \mathfrak{b} + \sum_{j=1}^{d} \mathbb{1}_{A_j}(x)\big(\Delta_j(\mathfrak{n}_j \cdot x - o_j)\big).$$

在折线之外，函数 $\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}$ 可微，且

$$D\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{a}}(x) = \sum_{j=1}^{d} \mathbb{1}_{A_j}(x)\Delta_j\mathfrak{n}_j.$$

注意，对于每个加项 $j = 1,\dots,d$，沿折线在 A_j 和 $\overline{A_j}^c$ 上微分之差等于 $\Delta_j\mathfrak{n}_j$（对于响应函数 $\mathfrak{N}^{\mathbb{W}}$，只要它在参考点处可微，且不存在具有相同折线 H_j 的第二个神经元，这一点也成立）。

接下来，我们介绍广义网络响应这一类函数。该类函数的优势在于，在相当宽松的假设下，可以通过应用紧致性论证找到极小化元。

定义 2.1。若函数 $\mathcal{R}$: $\mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ 具有以下表示，我们称其为广义响应函数：存在 $K \in \mathbb{N}_0$，一个由 $=($ $A_1, \dots, A_K)$ 个开半空间组成的元组 A $\mathbb{R}^{d_{\text{in}}}$，这些开半空间属于 ∂A_1，且其边界两两不同，分别为 ∂A_K, $\dots$, $=$；还存在 m $m_0(m_K, \dots, \in \{0,1\} \times \{1,2\}^K)$，一个仿射

映射 $\mathfrak{a}\colon \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$、向量 $\delta_1, \ldots, \delta_K \in \mathbb{R}^{d_{\text{in}}}$ 以及实数 $\mathfrak{b}_1, \ldots, \mathfrak{b}_K \in \mathbb{R}$，使得对所有$x \in \mathbb{R}^{d_{\text{in}}}$

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{k=1}^{K} \mathbb{1}_{A_k}(x)\big(\delta_k \cdot x + \mathfrak{b}_k\big), \tag{3}$$

对于所有 $k = 1,\ldots, K$，其中 $m_k = 1$,

$$\partial A_k \subset \{x \in \mathbb{R}^{d_{\text{in}}} : \delta_k \cdot x + \mathfrak{b}_k = 0\}$$

且以下性质至少有一项成立

(a) $m_0 = 1$，(b) ($\mathfrak{n}_k\colon k \in \{1, \ldots, K\}$ 与 $m_k = 2$) 线性相关，

或 (c) $\mathfrak{a}$ 是常函数。

可达到的最小 $m_0 + \ldots + m_K$ 称为广义响应的维数。

我们称 $A_1, \ldots, A_K$ 为响应函数的活跃半空间，称 $m_1, \ldots, m_K$ 为半空间 $A_1, \ldots, A_K$ 的重数，或称为加项。对于 $d \in \mathbb{N}_0$，我们用 $\mathfrak{R}_d$ 表示所有维数至多为 d 的广义响应所构成的空间。此外，若某个响应函数 $\mathcal{R} \in \mathfrak{R}_d$ 的维数为 d 或更低，或该响应函数不连续，则称其在维数 d 下是严格响应。用 $\mathfrak{R}_d^{\text{strict}}$ 表示维数为 d 时严格响应的集合。此外，若某个响应函数连续，或等价地，重数 $m_1, \ldots, m_K$ 均可取为 1，则称该响应函数可表示。

注 2.2. 严格来说，维数为 d 的可表示响应函数不一定是隐藏层中含有 d 个神经元的网络的响应函数，因为在 $m_0 = 1$ 的情形下，我们可能需要两个 ReLU 神经元（而不是一个）来生成 $\mathfrak{a}$ 的线性分量。然而，对于每个紧集 $K \subset \mathbb{R}^{d_{\text{in}}}$ 和每个可表示响应函数 $\mathcal{R} \in \mathfrak{R}_d$，我们都能找到一个隐藏层中含有$d$ 个神经元的网络，使其在 K 上与 $\mathcal{R}$一致。因此，对于紧支集测度 μ，可表示广义响应的子集都能通过适当的浅层网络在相关定义域上实现。

在分析广义响应函数类上的最小化问题时，我们可以采用比主定理（定理 1.2）更弱的假设。我们将使用以下概念。

定义 2.3. (i) 超平面 $H \subset \mathbb{R}^{d_{\text{in}}}$ 中的元素 x 若满足 $x \in \operatorname{supp} \mu|_A$ 且 $x \in \operatorname{supp} \mu|_{\overline{A}^c}$，则称其为 H-正则，其中 A 是满足 $\partial A = H$的开半空间。

(ii) 若测度 μ 定义在 $\mathbb{R}^{d_{\text{in}}}$ 上，则称其为良好测度，若所有超平面的 μ测度均为零，并且对于每个满足 $\mu(A)$ $\mu(\overline{A}^c)$ 的开半空间 A， ∂A-正则点的集合不能被有限个不同于 ∂A 的超平面覆盖。

命题 2.4. 假设 μ 是定义在 $\mathbb{R}^{d_{\text{in}}}$ 上的良好测度，且损失函数 $\mathcal{L}\colon \mathbb{D} \times \mathbb{R} \to \mathbb{R}_+$ 可测并满足以下假设:

(i)（损失关于第二个自变量下半连续）对于所有 $x \in \mathbb{D}$ 和 $y \in \mathbb{R}$，我们有

$$\liminf_{y' \to y} \mathcal{L}(x, y') \geq \mathcal{L}(x, y).$$

(ii)（在第二个参数上无界）对于所有 $x \in \mathbb{D}$，我们有

$$\lim_{|y| \to \infty} \mathcal{L}(x, y) = \infty.$$

设 $d \in \mathbb{N}_0$，误差为 $\overline{\mathrm{err}}_d^{\mathcal{L}} < \infty$。那么存在一个 $\mathcal{R} \in \mathfrak{R}_d$，满足

$$\int \mathcal{L}(x, \mathcal{R}(x))\,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_d^{\mathcal{L}} := \inf_{\tilde{\mathcal{R}} \in \mathfrak{R}_d} \int \mathcal{L}(x, \tilde{\mathcal{R}}(x))\,\mathrm{d}\mu(x).$$

此外，对于 $d \geq 1$，下确界

$$\inf_{\tilde{\mathcal{R}} \in \mathfrak{R}_d^{\mathrm{strict}}} \int \mathcal{L}(x, \tilde{\mathcal{R}}(x))\,\mathrm{d}\mu(x)$$

在 $\mathfrak{R}_d^{\mathrm{strict}}$上可达到。

证明。设 $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ 是维数至多为 d 的一列广义响应函数，满足

$$\lim_{n\to\infty} \int \mathcal{L}(x, \mathcal{R}^{(n)}(x))\,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_d^{\mathcal{L}}.$$

我们使用式 (3) 中的表示，并写作

$$\mathcal{R}^{(n)}(x) = \mathfrak{a}^{(n)}(x) + \sum_{k=1}^{K_n} \mathbb{1}_{A_k^{(n)}} \big(\delta_k^{(n)} \cdot x + \mathfrak{b}_k^{(n)}\big).$$

此外，用 $\mathfrak{n}_k^{(n)} \in \mathcal{O}$ 和 $o_k^{(n)} \in \mathbb{R}$ 表示满足 $A_k^{(n)} = \{x \in \mathbb{R}^{d_{\mathrm{in}}} \colon \mathfrak{n}_k^{(n)} \cdot x > o_k^{(n)}\}$ 的量。步骤 1：导出一个具有表示 (3) 的极限。

我们选取一个子序列 $(n_l)_{l\in\mathbb{N}}$，沿着该子序列，K-数和重数始终保持不变，并且始终属于同一种情形，即 (a)、(b) 或 (c)。此外，我们假设对于每个 $k = 1, \ldots..., K$，$(\mathfrak{n}^{(n_l)}{}_k)_{l\in\mathbb{N}}$ 在 $\mathcal{O}$ 中收敛到 $\mathfrak{n}_k$，且 $(o^{(n_l)}{}_k)_{l\in\mathbb{N}}$ 在 $\mathbb{R} \cup \{\pm\infty\}$ 中收敛到 o_k。为记号简便起见，我们假设对于整个序列这一结论都成立。

我们称

$$A_k = \{x \in \mathbb{R}^{d_{\mathrm{in}}} : \mathfrak{n}_k \cdot x > o_k\}$$

k项的渐近活跃区域，并令 $H_k = \partial A_k$。记 $\mathbb{J}$ 为所有满足下述集合条件的子集 $J \subset \{1, \ldots, K\}$ 的集族：

$$A_J = \bigcap_{j\in J} A_j \cap \bigcap_{j\in J^c} \overline{A}_j^c$$

满足 $\mu(A_J) > 0$。注意到，集合 $(A_J \colon J \in \mathbb{J})$ 非空、开且两两不相交，并且它们的并集具有满 μ-测度，这是因为

$$\mu\Big(\mathbb{R}^{d_{\mathrm{in}}} \setminus \bigcup_{J\subset\{1,\ldots,K\}} A_J\Big) \leq \sum_{j=1}^{K} \mu\big(H_j\big) = 0.$$

此外，对于每个 $J \in \mathbb{J}$ 以及每个满足 $B \subset A_J$ 的紧集 B，从某个依赖于 B 的 n 起，广义响应 $\mathcal{R}^{(n)}$ 对所有满足 $x \in B$ 的情况都满足

$$\mathcal{R}^{(n)}(x) = \mathcal{D}_J^{(n)} \cdot x + \beta_J^{(n)},$$

其中

$$\mathcal{D}_J^{(n)} := \mathfrak{a}'^{(n)} + \sum_{j\in J} \delta_j^{(n)} \quad \text{and} \quad \beta_J^{(n)} := \mathfrak{a}^{(n)}(0) + \sum_{j\in J} \mathfrak{b}_j^{(n)}.$$

令 $J \in \mathbb{J}$。接下来，我们证明沿着某个适当的子序列，$(\mathcal{D}_J^{(n)})_{n\in\mathbb{N}}$ 在 $\mathbb{R}^{d_{\mathrm{in}}}$中收敛。首先假设沿着某个子序列，$(|\mathcal{D}_J^{(n)}|)_{n\in\mathbb{N}}$ 收敛到∞。为使记号简便，不失一般性地假设 $|\mathcal{D}_J^{(n)}| \to \infty$成立。我们令

$$\mathcal{H}_J^{(n)} = \{x \in \mathbb{R}^{d_{\mathrm{in}}} : \mathcal{D}_J^{(n)} \cdot x + \beta_J = 0\}.$$

对于每个满足 $\mathcal{D}_J^{(n)} \neq 0$ 的 n， $\mathcal{H}_J^{(n)}$ 都是一个超平面，可以通过指定一个法向量及其相应的偏移量来参数化。如上所述，我们可以证明沿着某个适当的子序列（这里同样假设它就是整个序列），法向量在 $\mathcal{O}$ 中收敛，且偏移量在 $\mathbb{R}$中收敛。我们用 $\mathcal{H}_J$ 表示与极限法向量和偏移量对应的超平面（若偏移量在 $\mathbb{R}$中不收敛，则假设该超平面为空集）。由于梯度 $\mathcal{D}_J^{(n)}$ 的范数趋于无穷大，因此对于每个 $x \in A_J \backslash \mathcal{H}_J$，都有$|\mathcal{R}^{(n)}(x)| \to \infty$，从而 $\mathcal{L}(x, \mathcal{R}^{(n)}(x)) \to \infty$。因此，由 Fatou 引理可得

$$\liminf_{n\to\infty} \int_{A_J \backslash \mathcal{H}_J} \mathcal{L}(x, \mathcal{R}^{(n)}(x)) \,\mathrm{d}\mu(x) \geq \int_{A_J \backslash \mathcal{H}_J} \liminf_{n\to\infty} \mathcal{L}(x, \mathcal{R}^{(n)}(x)) \,\mathrm{d}\mu(x) = \infty$$

这与 $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ 的渐近最优性相矛盾。我们证明了序列 $(\mathcal{D}_J^{(n)})_{n\in\mathbb{N}}$ 是相对紧的，并且通过转而选取适当的子序列，可以保证极限 $\mathcal{D}_J = \lim_{n\to\infty} \mathcal{D}_J^{(n)}$ 存在。

类似地，我们证明沿着适当的子序列，$(\beta_J^{(n)})_{n\in\mathbb{N}}$ 收敛到某个值 $\beta_J \in \mathbb{R}$。假设情况并非如此，那么存在一个子序列，使得 $|\beta_J^{(n)}| \to \infty$。同样地，为了记号简便，我们假设整个序列都满足这一情况。于是，对于每个 $x \in A_J$，都有 $|\mathcal{R}^{(n)}(x)| \to \infty$，并且我们按照上述方法论证这将与 $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ 的最优性相矛盾。因此，在紧集上我们有 $B \subset A_J$ 一致收敛

$$\lim_{n\to\infty} \mathcal{R}^{(n)}(x) = \mathcal{D}_J \cdot x + \beta_J. \tag{4}$$

由于 $\bigcup_{J\in\mathbb{J}} A_J$ 具有满的 μ测度，利用 $\mathcal{L}$ 关于第二个自变量的下半连续性以及法图引理可知，对于每个满足以下条件的可测函数 $\mathcal{R}: \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}$，对每个 $J \in \mathbb{J}$ 和 $x \in A_J$，都有

$$\mathcal{R}(x) = \mathcal{D}_J \cdot x + \beta_J \tag{5}$$

我们有

$$\begin{aligned}\int \mathcal{L}(x, \mathcal{R}(x)) \,\mathrm{d}\mu(x) &\leq \int \liminf_{n\to\infty} \mathcal{L}(x, \mathcal{R}^{(n)}(x)) \,\mathrm{d}\mu(x) \\ &\leq \liminf_{n\to\infty} \int \mathcal{L}(x, \mathcal{R}^{(n)}(x)) \,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_d^{\mathcal{L}}.\end{aligned}$$

如果加项 $j \in \{1, \ldots, K\}$ 满足 A_j或 $\overline{A}_j^c$的 μ测度为零，我们就称其为退化项。现在，设 j是一个非退化项。由于 μ 性质良好，存在一个 ∂A_j-正则点 x，使其不属于$\bigcup_{A\in\mathbb{A}: \partial A \neq \partial A_j} \partial A$，其中 $\mathbb{A} := \{A_i : i \text{ 是非退化的}\}$。令

$$J_-^x = \{i : x \in A_i\} \cup \{i : \overline{A}_i^c = A_j\} \quad \text{and} \quad J_+^x = \{i : x \in A_i\} \cup \{i : A_i = A_j\}.$$

由于 $x \in \mathrm{supp}(\mu|_{\overline{A}_j^c})$，我们得到单元 $A_{J_-^x}$具有严格为正的 μ测度，因此 $J_-^x \in \mathbb{J}$。类似地，$x \in \mathrm{supp}(\mu|_{A_j})$意味着 $J_+^x \in \mathbb{J}$。（注意，J_+^x和 J_-^x正是位于超平面 ∂A_j 两侧、在 x处相对的单元。）因此我们得到

$$\delta_{A_j}^{(n)} := \sum_{i:A_i=A_j} \delta_i^{(n)} - \sum_{i:\overline{A}_i^c=A_j} \delta_i^{(n)} = \mathcal{D}_{J_+^x}^{(n)} - \mathcal{D}_{J_-^x}^{(n)} \to \mathcal{D}_{J_+^x} - \mathcal{D}_{J_-^x} =: \delta_{A_j},$$

其中，$\delta_{A_j}^{(n)}$ 和 δ_{A_j} 的定义不依赖于 x的选择。同理，

$$\mathfrak{b}_{A_j}^{(n)} := \sum_{i:A_i=A_j} \mathfrak{b}_i^{(n)} - \sum_{i:\overline{A}_i^c=A_j} \mathfrak{b}_i^{(n)} = \beta_{J_+^x}^{(n)} - \beta_{J_-^x}^{(n)} \to \beta_{J_+^x} - \beta_{J_-^x} =: \mathfrak{b}_{A_j}.$$

我们通过对 $\mathbb{A}$ 进行删减来构造集合 $\mathbb{A}_0$，使得对于边界位于同一超平面上的两个活跃区域，仅保留其中一个（也就是说，对于位于某个超平面两侧的两个活跃区域，仅保留其中一个）。于是，存在一个仿射函数 $\mathfrak{a}\colon \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$，使得对于 $x \in \bigcup_{J\in\mathbb{J}} A_J$

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{A\in\mathbb{A}_0} \mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A). \tag{6}$$

我们利用后一个恒等式在整个空间 $\mathbb{R}^{d_{\text{in}}}$ 上定义 $\mathcal{R}$。

步骤 2：分析表示（6）的重数。

我们为 $A \in \mathbb{A}_0$ 中的活跃区域赋予重数。如果

$$\partial A \subset \{x \in \mathbb{R}^{d_{\text{in}}} : \delta_A \cdot x + \mathfrak{b}_A = 0\},$$

或者，等价地，$x \mapsto 1\mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A)$ 连续，那么我们为 A 赋予重数一；否则赋予重数二。

接下来，我们证明：若某个活跃区域 $A \in \mathbb{A}_0$ 的折线仅由一个重数为一的加项 k 提供，那么该区域仍被赋予重数一。为此，只需证明对于这样的 A，$x \mapsto 1\mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A)$ 连续。假设第 k 个加项是对 A 有贡献的唯一加项。那么 $\delta_k^{(n)} = \delta_A^{(n)} \to \delta_A$ 和 $\mathfrak{b}_k^{(n)} = \mathfrak{b}_A^{(n)} \to \mathfrak{b}_A$。此外，有

$$\{x \in \mathbb{R}^{d_{\text{in}}} : \mathfrak{n}_k^{(n)} \cdot x - o_k^{(n)} = 0\} \subset \{x \in \mathbb{R}^{d_{\text{in}}} : \delta_k^{(n)} \cdot x + \mathfrak{b}_k^{(n)} = 0\}$$

这意味着，特别是，$\delta_k^{(n)}$ 是 $\mathfrak{n}_k^{(n)}$ 的倍数。后两个向量均收敛，且$|\mathfrak{n}_k| = 1$，这也意味着极限 δ_A 是 $\mathfrak{n}_k$的倍数。为了证明

$$\partial A \subset \{x \in \mathbb{R}^{d_{\text{in}}} : \delta_A \cdot x + \mathfrak{b}_A = 0\},$$

因此，只需验证左侧超平面上的某个点也属于右侧的集合。事实上，$x = o_k\mathfrak{n}_k$ 满足这一条件，因为对于 $n \to \infty$

$$\delta_A \cdot (o_k\mathfrak{n}_k) + \mathfrak{b}_A = \lim_{n\to\infty} \delta_k^{(n)} \cdot (o_k^{(n)}\mathfrak{n}_k^{(n)}) + \mathfrak{b}_k^{(n)} = 0.$$

这意味着包含连续性；我们还证明，对于这样的 k，$1\mathbb{1}_k^{A_k^{(n)}}(x)(\delta_k^{(n)} \cdot x + \mathfrak{b}_k^{(n)}) \to 1\mathbb{1}_A(\delta_A \cdot x + \mathfrak{b}_A)$ 在 x中逐点成立。

注意，分配给活跃区域 $A \in \mathbb{A}_0$ 的所有重数之和严格小于 d，如果

(1) 对所有 $n \in \mathbb{N}$、$m_0^{(n)} = 1$均成立；(2) 存在一个退化渐近活跃区域；或者 (3) 存在一个活跃区域 $A \in \mathbb{A}_0$，其贡献项的累计重数严格大于分配给 A的重数。

In t在后两种情形下，我们可以像（6）中那样选择 $m_0 = 1$ 和 $\mathcal{R}$，其维数至多为 d.

3. 步骤：分别处理 $\mathcal{R}^{(n)}$ 属于类型（a）、（b）或（c）的情形。

在情形（a）中，已满足上述性质（1），因此证明完成。

现在假设 $\mathcal{R}^{(n)}$ 都是类型（b），并令 $\mathbb{I} \subset \{1, \dots, K\}$ 表示重数为二的加项的指标。那么，$(\mathfrak{n}_j^{(n)} : j \in \mathbb{I})$ 线性相关意味着极限 $(\mathfrak{n}_j : j \in \mathbb{I})$ 线性相关。如果不存在退化渐近活跃区域，且 $(\partial A_j : j \in \mathbb{I})$ 的条目两两不同，那么表示（6）满足性质（b），并且我们已验证 $\mathcal{R}$ 属于 $\mathfrak{R}_d$。反过来，如果 $(\partial A_j : j \in \mathbb{I})$ 中的条目并非两两不同，那么就存在一个活跃区域 $A \in \mathbb{A}_0$，其贡献加项的重数总和至少为四，因此上述性质（3）成立，问题得以解决。

剩下需要考虑的是情形（c）。我们可以假设存在非退化活跃区域，并且每个 $A \in \mathbb{A}_0$ 都由总重数至多为二的项所服务；否则，上述性质（2）或（3）成立，问题便得以解决。那么每个 $A \in \mathbb{A}_0$ 都由

(i) 一个加项；(ii) 两个重数为一且具有相同渐近活跃区域的加项；或 (iii) 两个重数为一且渐近区域位于相关超平面两侧的加项。

对于类型 (i) 或 (ii) 的 $A \in \mathbb{A}_0$，相关加项的渐近贡献在超平面 ∂A之外满足

$$\lim_{n\to\infty} \sum_{j:A_j=A} \mathbb{1}_{A_j^{(n)}}(x)(\delta_j^{(n)} \cdot x + \mathfrak{b}_j^{(n)}) = \mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A).$$

如果有一个重数为一的加项对极限作出贡献，那么出于与上述相同的原因，极限将是连续的。因此，如果不存在类型 (iii) 的 $A \in \mathbb{A}_0$，我们利用 $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ 在 $\bigcup_{J\in\mathbb{J}} A_J$ 上收敛这一点，来推导出 $\lim_{n\to\infty} \mathfrak{b}^{(n)} =: \mathfrak{b}$ 存在，并得到一个满足 $\mathfrak{a} \equiv \mathfrak{b}$ 的表示 (3)，且属于情形 (c)。

现在令 $\mathbb{A}_0^*$ 表示所有类型 (iii) 的 $A \in \mathbb{A}_0$ 构成的子集，并假设 $\mathbb{A}_0^* \neq \emptyset$。如果 $A \in \mathbb{A}_0^*$ 由孪生点对 (i, j) 服务，我们称之为由该孪生点对服务，其中 $A_i = A$ 且 $A_j = \overline{A}^c$。此外，如果某个加项 k 对一个类型 (iii) 的活跃区域有贡献，我们称该加项为类型 (iii)。通过转到一个适当的子序列，我们可以确保存在一个类型 (iii) 的加项 k，使得对于所有 $n \in \mathbb{N}$，$|\delta_k^{(n)}|$ 在所有类型 (iii) 的加项中取最大值。

我们区分两种情况。若存在一个子序列，使得 $(|\delta_k^{(n)}|)_{n\in\mathbb{N}}$ 一致有界，那么同样通过转而取适当的子序列，我们可得，对于每个类型 (iii) 的加项j，$\lim_{n\to\infty} \delta_j^{(n)} =: \delta_j \in \mathbb{R}^{d_{\text{in}}}$ 存在。由于 $\mathbb{1}_{A_j^{(n)}}(x)(\delta_j^{(n)} \cdot x + \mathfrak{b}_j^{(n)})$ 是连续的，对于满足 $x_n \in \partial A_j^{(n)}$ 的序列 $(x_n)_{n\in\mathbb{N}}$，对所有 $n \in \mathbb{N}$，且收敛到一个 $x \in \partial A_j$

$$\mathfrak{b}_j^{(n)} = -\delta_j^{(n)} \cdot x_n \to -\delta_j \cdot x =: \mathfrak{b}_j,$$

其中左端不取决于 $(x_n)_{n\in\mathbb{N}}$ 或 x 的选择。此外，极限 $\mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)$ 是连续的。因此，对于一对孪生点 (i, j)，贡献项 i 和 j 对 A_i 的总重数为二，但相应的项

$$\mathbb{1}_{A_i}(x)(\delta_{A_i} \cdot x + \mathfrak{b}_{A_i}) = \mathbb{1}_{A_i}(x)((\delta_i - \delta_j) \cdot x + \mathfrak{b}_i - \mathfrak{b}_j)$$

是连续的 ，因此重数为一。因此，我们属于上述情形（3），证明完毕 .

剩下需要考虑的情形是 $(|\delta_k^{(n)}|)_{n\in\mathbb{N}}$ 趋于无穷大。对于每一对孪生点 (i, j)，以及$n \in \mathbb{N}$，我们可以选取 $\alpha_i^{(n)} \in [-1, 1]$ 和 $\alpha_j^{(n)} \in [-1, 1]$，使得

$$\frac{1}{|\delta_k^{(n)}|}\delta_i^{(n)} = \alpha_i^{(n)}\mathfrak{n}_i^{(n)} \quad \text{and} \quad \frac{1}{|\delta_k^{(n)}|}\delta_j^{(n)} = -\alpha_j^{(n)}\mathfrak{n}_j^{(n)}$$

并且沿着适当的子序列，所有 $(\alpha_i^{(n)})_{n\in\mathbb{N}}$ 和 $(\alpha_j^{(n)})_{n\in\mathbb{N}}$ 都收敛到极限 $\alpha_i \in [-1, 1]$ 和 $\alpha_j \in [-1, 1]$。由于 $(\delta_i^{(n)} - \delta_j^{(n)})_{n\in\mathbb{N}}$ 收敛到 δ_{A_i}，$|\delta_k^{(n)}| \to \infty$ 并且

$$\lim_{n\to\infty} \mathfrak{n}_i^{(n)} = \mathfrak{n}_i = -\mathfrak{n}_j = \lim_{n\to\infty} -\mathfrak{n}_j^{(n)}$$

我们得到 $\alpha_i = \alpha_{j_0}$。因此，对于 $x \in \bigcup_{J\in\mathbb{J}} A_J$，有

$$\lim_{n\to\infty} \frac{1}{|\delta_k^{(n)}|}(\mathcal{R}^{(n)}(x) - \mathfrak{b}^{(n)}) = \sum_{i:A_i\in\mathbb{A}_0^*} \alpha_i(\mathfrak{n}_i \cdot x - o_i).$$

如果 $\sum_{i:A_i\in\mathbb{A}_0^*} \alpha_i\mathfrak{n}_i$ 不等于零，那么右侧的线性项就不会消失。这与 $\mathcal{R}^{(n)}$ 在 $\bigcup_{J\in\mathbb{J}} A_J$上的收敛相矛盾。因此，我们得到 $\sum_{i:A_i\in\mathbb{A}_0^*} \alpha_i\mathfrak{n}_i = 0$。注意，并非所有 α_i都等于零，因为 $\alpha_k \in \{\pm 1\}$，且求和中出现了 k 或它的孪生点。因此，我们证明了属于 $\mathbb{A}_0^*$ 中活跃区域的法向量线性相关。由此可知，我们处于情形 (b)，证明完成。

4. step: 维数为 d时严格响应的最小化问题讨论

现在，我们从 $\mathfrak{R}_d^{\text{strict}}$中选取一个响应序列 $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$，使得

$$\lim_{n\to\infty} \int \mathcal{L}(x, \mathcal{R}^{(n)}(x))\,\mathrm{d}\mu(x) = \inf_{\tilde{\mathcal{R}}\in\mathfrak{R}_d^{\text{strict}}} \int \mathcal{L}(x, \tilde{\mathcal{R}}(x))\,\mathrm{d}\mu(x).$$

如果有无穷多个响应 $\mathcal{R}^{(n)}$ 属于 $\mathfrak{R}_{d-1}$，那么上述构造的响应就是一个最小化解，并且属于 $\mathfrak{R}_{d-1}$。反过来，如果除有限多个之外的所有响应都是不连续的，那么上述构造会得到一个极限 $\mathcal{R} \in \mathfrak{R}_d$，使误差达到最小。注意，至少有一个加项对极限 $\mathcal{R}$的某个活跃区域贡献了重数二。如果 $\mathcal{R}$ 沿相应的超平面连续，那么它的维数至多为 $d-1$；否则，该响应不连续。两种情况下都有 $\mathcal{R} \in \mathfrak{R}_d^{\text{strict}}$。□

3. Strct generalized responses are not better than representable ones

我们将证明，在定理 1.2 的设定下，对于每个 $d \in \{2, 3, \dots\}$ with $\text{err}_d^{\mathcal{L}} < \text{err}_{d-1}^{\mathcal{L}}$，对严格广义响应取下确界所得的误差大于最佳可表示响应的误差。这将推出定理 1.2。关于损失函数 $\mathcal{L}$ that implies $\text{err}_d^{\mathcal{L}} < \text{err}_{d-1}^{\mathcal{L}}$ 的充分且必要假设的讨论，请参见 [DJK23] 中的命题 3.5 和例 3.7。

命题 3.1。假设定理 1.2 的条件成立。令 $d \in \mathbb{N}_0$。则存在一个最优网络 $\mathbb{W} \in \mathcal{W}_d$ with

$$\text{err}^{\mathcal{L}}(\mathbb{W}) = \text{err}_d^{\mathcal{L}} = \overline{\text{err}}_d^{\mathcal{L}}.$$

此外，如果 $d \geq 2$ and $\text{err}_d^{\mathcal{L}} < \text{err}_{d-1}^{\mathcal{L}}$，则有

$$\inf_{\mathcal{R}\in\mathfrak{R}_d^{\text{strict}}} \int \mathcal{L}(x, \mathcal{R}(x))\,\mathrm{d}\mu(x) > \text{err}_d^{\mathcal{L}}. \tag{7}$$

证明。不失一般性，我们可以假设 $\mu \neq 0$。首先，我们验证命题 2.4 的条件，以便得出存在广义响应 $\mathcal{R} \in \mathfrak{R}_d$ with

$$\int \mathcal{L}(x, \mathcal{R}(x))\,\mathrm{d}\mu(x) = \overline{\text{err}}_d^{\mathcal{L}}.$$

我们验证 μ 是一个良好测度：事实上，由于 μ 具有勒贝格密度 h，对于所有超平面 $H \subset \mathbb{R}^{d_{\text{in}}}$，都有 $\mu(H) = 0$。此外，对于每个满足 $\mu(A)\ \mu(\overline{A}^c) > 0$ 的半空间 A，∂A 与 $\mathbb{D}$ 的凸包的内部相交，因此存在一点 $x \in \partial A$，使得$h(x) > 0$。由于 $\{x \in \mathbb{R}^{d_{\text{in}}} : h(x) > 0\}$ 是一个开集，$\{x \in \partial A : h(x) > 0\}$ 不能被与 ∂A不同的有限多个超平面覆盖。此外，由于对于所有 $x \in \mathbb{R}^{d_{\text{in}}}$，函数

$y \mapsto \mathcal{L}(x,y)$ 是严格凸的并且能达到其最小值，因此显然，对于固定的 $x \in \mathbb{R}^{d_{\text{in}}}$， $y \mapsto \mathcal{L}(x,y)$ 具有连续性，并且

$$\lim_{|y|\to\infty} \mathcal{L}(x,y) = \infty.$$

我们通过对维数 d进行归纳法来证明该结论。若 $d \le 1$，则所有维数为 d 的广义响应在紧集 $\mathbb{D}$上都可由神经网络表示，结论即得。现在令 $d \ge 2$，并假设 $\mathcal{R}$ 是维数为 d 且满足以下条件的严格广义响应

$$\int \mathcal{L}(x,\mathcal{R}(x))\,\mathrm{d}\mu(x) = \inf_{\tilde{\mathcal{R}}\in\mathfrak{N}_d^{\text{strict}}} \int \mathcal{L}(x,\tilde{\mathcal{R}}(x))\,\mathrm{d}\mu(x).$$

只需证明以下两种情况至少有一种成立：存在

$$\int \mathcal{L}(x,\mathcal{R}(x))\,\mathrm{d}\mu(x) \ge \overline{\mathrm{err}}_{d-1}^{\mathcal{L}} \tag{8}$$

or

$$\int \mathcal{L}(x,\mathcal{R}(x))\,\mathrm{d}\mu(x) > \overline{\mathrm{err}}_{d}^{\mathcal{L}}. \tag{9}$$

事实上，若 (9) 不成立，则由 (8) 可得

$$\overline{\mathrm{err}}_{d-1}^{\mathcal{L}} \le \int \mathcal{L}(x,\mathcal{R}(x))\,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_{d}^{\mathcal{L}}$$

且归纳假设意味着 $\mathrm{err}_{d-1}^{\mathcal{L}} = \mathrm{err}_{d-1}^{\mathcal{L}} \le \mathrm{err}_{d}^{\mathcal{L}} \le \mathrm{err}_{d}^{\mathcal{L}}$，因此 $\mathrm{err}_{d}^{\mathcal{L}} = \mathrm{err}_{d}^{\mathcal{L}}$ 且 $\mathrm{err}_{d}^{\mathcal{L}} = \mathrm{err}_{d-1}^{\mathcal{L}}$。因此，维数至多为$d-1$ 的最优可表示响应 $\mathcal{R}$ （由归纳假设可知其存在），在对维数为 d 或更小的所有广义响应取最小值时，也是最优的。反过来，若 (9) 成立，则一个最优广义响应（由命题 2.4 可知其存在）是可表示的，因此特别有 $\mathrm{err}_{d}^{\mathcal{L}} = \mathrm{err}_{d}^{\mathcal{L}}$。这表明最优可表示响应总是存在。此外还可推出，当 $\mathrm{err}_{d}^{\mathcal{L}} < \mathrm{err}_{d-1}^{\mathcal{L}}$时，性质 (8) 和 (9) 中任一者都蕴含性质 (7)。

设在维数为 d 时，最优严格广义响应 $\mathcal{R}$ 由标准表示 (3) 给出

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{j=1}^{K} \mathbb{1}_{A_j}(x)\big(\delta_j \cdot x + \mathfrak{b}_j\big),$$

其中 A_1、…、A_K 是边界两两不同的半空间，而 m_0、m_1、…、m_K 是相应的重数。设 $\mathcal{R}$ 是一个最优响应，并且其中重数为二的项数最少。

首先注意，对于任意 $k = 1$、…、K，若 ∂A_k 不与 $\mathbb{D}$、A_k 的凸包的内部相交，则 ∂A_k 要么具有零 μ 测度，要么具有满 μ 测度。在这两种情况下，我们都可以移除第 k 个加项，令 $m_0 = 1$，并适当地调整仿射函数 $\mathfrak{a}$，从而得到一个与原响应在 μ 几乎处处一致、且维数仍至多为 d 的响应。因此不失一般性，可以假设对于任意 $k = 1$、…、K，∂A_k 都与 $\mathbb{D}$ 的凸包的内部相交。

我们分情况讨论。若有 $m_0 = 1$ （情形 (a)）或 $\mathfrak{a} \equiv \mathfrak{b} \in \mathbb{R}$ （情形 (c)），则 [DJK23] 中命题 3.3 的证明表明，用两个重数为一的加项替换一个重数为二的加项，可以减小误差，这就证明了 (9)。

剩下需要处理情形 (b)。设 $\mathbb{I}$ 为重数为二的指标集合。现在，向量 $(\mathfrak{n}_j : j \in \mathbb{I})$ 线性相关。如果该集合不是极小的，也就是说，移除其中一个向量后仍能得到线性相关的集合，那么我们可以像上文那样论证，并将这个特定加项适当地替换为两个

重数为一且仍满足 (b)、误差严格更小的加项。因此，我们可以不失一般性地假设，集合 $(\mathfrak{n}_j : j \in \mathbb{I})$ 是极小的，即对于一个非平凡线性组合

$$\sum_{j\in\mathbb{I}} \alpha_j \mathfrak{n}_j = 0 \tag{10}$$

可知，对于所有 $j \in \mathbb{I}$， $\alpha_j \neq 0$。对于 $x \in \mathbb{D}$ 和 $y \in \mathbb{R}$，令

$$\tilde{\mathcal{L}}(x, y) = \mathcal{L}\Big(x, y + \sum_{j\in\mathbb{I}^c} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)\Big)$$

~，并注意到，由于 $x \mapsto \sum_{j\in\mathbb{I}^c} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)$ 的连续性，函数 $\mathcal{L}$ 满足命题中对 $\mathcal{L}$ 所施加的相同假设。若我们能找到一个维数为 $2\#\mathbb{I}$ 的可表示响应函数 $\mathcal{R}$，使得

$$\int \tilde{\mathcal{L}}(x, \hat{\mathcal{R}}(x))\,\mathrm{d}\mu(x) < \int \tilde{\mathcal{L}}(x, \tilde{\mathcal{R}}(x))\,\mathrm{d}\mu(x),$$

其中

$$\tilde{\mathcal{R}}(x) = \mathfrak{a}(x) + \sum_{j\in\mathbb{I}} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j),$$

then

$$\int \mathcal{L}\Big(x, \hat{\mathcal{R}}(x) + \sum_{j\in\mathbb{I}^c} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)\Big)\,\mathrm{d}\mu(x) < \int \mathcal{L}(x, \mathcal{R}(x))\,\mathrm{d}\mu(x)$$

由此可得，式 (9) 成立。

因此，不失一般性，我们可以假设 $\mathbb{I} = \{1, \ldots, K\}$，且最优严格响应函数为

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{j=1}^{K} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j).$$

我们固定一个满足 (10) 的非平凡向量 $(\alpha_j)_{j\in\mathbb{I}}$，并选取 $(\delta_j^+, \delta_j^-) \in \mathbb{R}^{d_{\mathrm{in}}} \times \mathbb{R}^{d_{\mathrm{in}}}$、$(\mathfrak{b}_j^+, \mathfrak{b}_j^-) \in \mathbb{R}^2$ 以及 $\mathfrak{b} \in \mathbb{R}$，其中 $\delta_j = \delta_j^+ - \delta_j^-$ 和 $\mathfrak{b}_j = \mathfrak{b}_j^+ - \mathfrak{b}_j^-$，使得对所有 $x \in \mathbb{R}^{d_{\mathrm{in}}}$

$$\mathcal{R}(x) = \mathfrak{b} + \sum_{j=1}^{K} \mathbb{1}_{A_j}(x)(\delta_j^+ \cdot x + \mathfrak{b}_j^+) + \mathbb{1}_{A_j^c}(x)(\delta_j^- \cdot x + \mathfrak{b}_j^-).$$

通过交换活跃区域的两侧，我们不失一般性地假设所有$\alpha_j > 0$ （事实上，这只会在相应的超平面上改变 $\mathcal{R}$ 的定义）。我们将用以下表达式替换 $\mathcal{N}_j(x) := \mathbb{1}_{A_j}(x)(\delta_j^+ \cdot x + \mathfrak{b}_j^+) + \mathbb{1}_{A_j^c}(x)(\delta_j^- \cdot x + \mathfrak{b}_j^-)$:

$$\mathcal{N}_j^\kappa(x) := (\delta_j^+ \cdot x + \mathfrak{b}_j^+ + \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j))^+ - (-\delta_j^- \cdot x - \mathfrak{b}_j^- - \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j))^+,$$

其中 $\kappa > 0$。令

$$Q_j^\kappa := \{x \in \mathbb{R}^{d_{\mathrm{in}}} : \mathcal{N}_j^\kappa(x) \neq \mathcal{N}_j(x) + \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j)\}$$

并将 $\mathcal{R}$与

$$\mathcal{R}^\kappa(x) := \mathfrak{b}^\kappa + \sum_{j=1}^{K} \mathcal{N}_j^\kappa(x),$$

其中 $\mathfrak{b}^\kappa := \mathfrak{b} + \sum_{j=1}^K \kappa\alpha_j o_j$。注意，由于 $\sum_{j=1}^K \alpha_j \mathfrak{n}_j = 0$，我们有 $\mathcal{R}^\kappa = \mathcal{R}$，在 $\mathbb{R}^{d_{\text{in}}} \setminus (\bigcap_{j=1,\dots,K} Q_j^\kappa)$ 上。此外，集合 $\{x \in \mathbb{R}^{d_{\text{in}}} : x \text{ 分为两个部分，大小为 } Q_j^\kappa\} \cap \mathbb{D}$，即 $\mathcal{O}(\kappa^{-2})$。因此，

(11)
$$\int \mathcal{L}(x, \mathcal{R}^\kappa(x)) - \mathcal{L}(x, \mathcal{R}(x))\,\mathrm{d}\mu(x) = \sum_{j=1}^K \int_{Q_j^\kappa} h(x)(\mathcal{L}(x, \mathcal{R}^\kappa(x)) - \mathcal{L}(x, \mathcal{R}(x)))\,\mathrm{d}x + \mathcal{O}(\kappa^{-2}).$$

此外，利用 h 在紧集 $\mathbb{D}$ 上的一致连续性，以及 $|\mathcal{L}(x, \mathcal{R}^\kappa(x)) - \mathcal{L}(x, \mathcal{R}(x))|$ 对所有 $x \in \mathbb{D}$ 和 $\kappa \geq 1$ 的一致有界性，我们得出，当 $\kappa \to \infty$

(12)
$$\begin{aligned}
&\int_{Q_j^\kappa} h(x)(\mathcal{L}(x, \mathcal{R}^\kappa(x)) - \mathcal{L}(x, \mathcal{R}(x)))\,\mathrm{d}x \\
&\quad = \int_{H_j} \int_{(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa} h(z)\big(\mathcal{L}(z, \mathcal{R}^\kappa(z)) - \mathcal{L}(z, \mathcal{R}(z))\big)\,\mathrm{d}z\,\mathrm{d}x' \\
&\quad = \int_{H_j} h(x') \int_{(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa} \big(\mathcal{L}(x', \mathcal{R}^\kappa(z)) - \mathcal{L}(x', \mathcal{R}(z))\big)\,\mathrm{d}z\,\mathrm{d}x' + o(\kappa^{-1}),
\end{aligned}$$

其中 $H_j = \partial A_j$。现在注意，对于固定的 $x' \in H_j$，若 $(x' + \mathfrak{n}_j\mathbb{R}) \cap Q_j^\kappa$ 不与某个满足 $i \neq j$ 的 Q_i^κ 相交，则有

$$\begin{aligned}
\int_{(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa} \mathcal{L}(x', \mathcal{R}(z))\,\mathrm{d}z =& |(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j|\,(L_j^+(x') + o(1)) \\
&+ |(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j^c|\,(L_j^-(x') + o(1)),
\end{aligned}$$

其中 $|\cdot|$ 表示一维豪斯多夫测度（即线段的长度），

$$L_j^+(x') = \mathcal{L}(x', \delta_j^+ \cdot x' + \mathfrak{b}_j^+ + \hat{\mathcal{R}}_j(x')) \quad \text{and} \quad L_j^-(x') = \mathcal{L}(x', \delta_j^- \cdot x' + \mathfrak{b}_j^- + \hat{\mathcal{R}}_j(x'))$$

with

$$\hat{\mathcal{R}}_j(x') = \mathfrak{b} + \sum_{i \neq j} \big(\mathbb{1}_{A_i}(x')(\delta_i^+ \cdot x' + \mathfrak{b}_i^+) + \mathbb{1}_{A_i^c}(x')(\delta_i^- \cdot x' + \mathfrak{b}_i^-)\big).$$

此外，对于相同的 x'

$$\int_{(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa} \mathcal{L}(x', \mathcal{R}^\kappa(z))\,\mathrm{d}z = |(x' + \mathfrak{n}_j\mathbb{R}) \cap Q_j^\kappa|\,(\overline{L}_j(x') + o(1)),$$

其中，$\overline{L}_j(x')$ 是线段 $[\delta_j^- \cdot x' + \mathfrak{b}_j^- + \hat{\mathcal{R}}_j(x')\ \ \delta_j^+ \cdot x' + \mathfrak{b}_j^+ + \hat{\mathcal{R}}_j(x')]$ 上 $\mathcal{L}(x', \cdot)$ 的平均值。我们计算线段 $(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j$ 和 $(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j^c$ 的豪斯多夫测度。注意，对于 $t \in \mathbb{R}$，如果 t 位于 $t_{+/-}^\kappa$ 的解之间，则 $x' + t\mathfrak{n}_j$ 位于 Q_j^κ 中

$$\delta_j^{+/-} \cdot (x' + t_{+/-}^\kappa \mathfrak{n}_j) + \mathfrak{b}_j^{+/-} + \kappa\alpha_j t_{+/-}^\kappa = 0$$

从而 $|(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j| = |[t_-^\kappa, t_+^\kappa] \cap [0, \infty)|$。由于

$$\lim_{\kappa \to \infty} \kappa\, t_{+/-}^\kappa = -\frac{1}{\alpha_j}(\delta_j^{+/-} \cdot x' + \mathfrak{b}_j^{+/-}) =: t_{+/-}(x')$$

我们得到

$$q_j^+(x') := \lim_{\kappa \to \infty} \kappa \left|(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j\right| = \left|[t_-(x'), t_+(x')] \cap [0, \infty)\right|$$

类似地，可得

$$q_j^-(x') := \lim_{\kappa \to \infty} \kappa \left|(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j^c\right| = \left|[t_-(x'), t_+(x')] \cap (-\infty, 0]\right|.$$

结合这些估计可得，对于每个 $x' \in H_j \backslash \bigcup_{i\neq j} H_i$，都有

$$\begin{aligned}\lim_{\kappa\to\infty} \kappa \int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^\kappa} \big(\mathcal{L}(x', \mathcal{R}^\kappa(z)) - \mathcal{L}(x', \mathcal{R}(z))\big)\,\mathrm{d}z \\ = (q_j^+(x') + q_j^-(x'))\overline{L}_j(x') - (q_j^-(x')\,L_j^+(x') + q_j^+(x')\,L_j^-(x'))\end{aligned}$$

结合 (11)、(12) 和控制收敛定理，我们得到

$$\begin{aligned}&\lim_{\kappa\to\infty} \kappa \int \big(\mathcal{L}(x, \mathcal{R}^\kappa(x)) - \mathcal{L}(x, \mathcal{R}(x))\big)\,\mathrm{d}\mu(x) \\ &\quad = \sum_{j=1}^K \int_{H_j} h(x')\big((q_j^+(x') + q_j^-(x'))\,\overline{L}_j(x') - (q_j^+(x')\,L_j^+(x') + q_j^-(x')\,L_j^-(x'))\big)\,\mathrm{d}x',\end{aligned}$$

其中我们利用了 $\kappa\int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^\kappa} \big(\mathcal{L}(x', \mathcal{R}^\kappa(z)) - \mathcal{L}(x', \mathcal{R}(z))\big)\,\mathrm{d}z$ 在所有 $j = 1, \ldots K$、$x' \in H_j \cap \mathbb{D}$ 和 $\kappa \geq 1$上均一致有界。

现在考虑由下式给出的 $\mathcal{R}^{-\kappa}$

$$\mathcal{R}^{-\kappa}(x) = \mathfrak{b}^{-\kappa} + \sum_{j=1}^K -(-\delta_j^+ \cdot x - \mathfrak{b}_j^+ + \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j))^+ + (\delta_j^- \cdot x + \mathfrak{b}_j^- - \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j))^+.$$

其中 $\mathfrak{b}^{-\kappa} := \mathfrak{b} - \sum_{j=1}^K \kappa\alpha_j o_j$。按照上述相同的论证，我们得到

$$\begin{aligned}&\lim_{\kappa\to\infty} \kappa \int \mathcal{L}(x, \mathcal{R}^{-\kappa}(x)) - \mathcal{L}(x, \mathcal{R}(x))\,\mathrm{d}\mu(x) \\ &\quad = \sum_{j=1}^K \int_{H_j} h(x')\big((q_j^+(x') + q_j^-(x'))\overline{L}_j(x') - (q_j^-(x')\,L_j^+(x') + q_j^+(x')\,L_j^-(x'))\big)\,\mathrm{d}x'.\end{aligned}$$

将利用 $q_j(x') = q_j^+(x') + q_j^-(x')$ 得到的估计相加，我们得到

$$\begin{aligned}\lim_{\kappa\to\infty} \kappa \int \big(\mathcal{L}(x, \mathcal{R}^{-\kappa}(x)) + \mathcal{L}(x, \mathcal{R}^\kappa(x)) - 2\mathcal{L}(x, \mathcal{R}(x))\big)\,\mathrm{d}\mu(x) \\ = \sum_{j=1}^K \int_{H_j} h(x')q_j(x')\big(2\overline{L}_j(x') - (L_j^+(x') + L_j^-(x'))\big)\,\mathrm{d}x'.\end{aligned} \tag{13}$$

由 $\mathcal{L}(x', \cdot)$的严格凸性可知，$2\overline{L}_j(x') \leq L_j^+(x') + L_j^-(x')$，且当$\delta_j^- \cdot x' + \mathfrak{b}_j^- \neq \delta_j^+ \cdot x' + \mathfrak{b}_j^+$时严格不等。由于 $\partial A_1 \not\subset \{x \in \mathbb{R}^{d_{\mathrm{in}}} \colon \delta_1 \cdot x + \mathfrak{b}_1 = 0\}$，我们知道，由所有满足 $h(x') > 0$、$q_1(x') > 0$ 且 $\delta_j^- \cdot x' + \mathfrak{b}_j^- \neq \delta_j^+ \cdot x' + \mathfrak{b}_j^+$的 $x' \in H_1$ 构成的集合 H_1'具有严格为正的（$d_{\mathrm{in}} - 1$）维豪斯多夫测度。因此，(13)中的极限严格为负，并且存在 $\kappa > 0$，使得$\mathcal{R}^\kappa$ 或 $\mathcal{R}^{-\kappa}$ 是比 $\mathcal{R}$更好的响应函数，这与 $\mathcal{R}$的最优性相矛盾。□

致谢。本研究由德国研究联合会（DFG）通过德国卓越战略项目 EXC 2044–390685587“明斯特数学：动力学—几何—结构”资助。本研究亦由德国研究联合会（DFG）通过 SFB1283/2 2021 项目（317210226）资助。

References

[AB09] H. Attouch 和 J. Bolte。《涉及解析特征的非光滑函数近端算法的收敛性》。 *Mathematical Programming*, 116(1)：5–16，2009。[AMA05] P. A. Absil、R. Mahony 和 B. Andrews。《解析代价函数下降法迭代序列的收敛性》。 *SIAM Journal on Optimization*，16(2)：531–547，2005。[BDL07] J. Bolte、A. Daniilidis 和 A. Lewis。《非光滑亚解析函数的 Łojasiewicz 不等式及其在次梯度动力系统中的应用》。 *SIAM Journal on Optimization*, 17(4)：1205–1223，2007。[CK22] C. Christof 和 J. Kowalczyk。《某些神经网络训练问题中伪局部极小值的普遍存在》。arXiv:2202.12262，2022。[DDKL20] D. Davis、D. Drusvyatskiy、S. Kakade 和 J. D. Lee。《随机次梯度法在温和函数上的收敛性》。 *Foundations of Computational Mathematics*，20(1)：119–154，2020。[DJK23] S. Dereich、A. Jentzen 和 S. Kassing。《浅层残差 ReLU 神经网络优化景观中极小化元的存在性》。arXiv:2302.14690，2023。[DK21] S. Dereich 和 S. Kassing。《Łojasiewicz 景观上的随机梯度下降方法的收敛性》。arXiv:2102.09385，2021。[DK22] S. Dereich 和 S. Kassing。《随机微分方程的冷却：几乎必然收敛》。 *Stochastic Processes and their Applications*，152：289–311，2022。[EJRW21] S. Eberle、A. Jentzen、A. Riekert 和 G. S. Weiss。《ReLU 激活函数人工神经网络训练中梯度流的存在性、唯一性及收敛速率》。arXiv:2108.08106，2021。[EMWW20] W. E、C. Ma、S. Wojtowytsch 和 L. Wu。《迈向对基于神经网络的机器学习的数学理解：我们已知与未知的内容》。arXiv:2009.10713，2020。[GJL22] D. Gallon、A. Jentzen 和 F. Lindner。《人工神经网络训练中梯度下降优化方法的爆破现象》。arXiv:2211.15641，2022。[GP90] F. Girosi 和 T. Poggio。《网络与最佳逼近性质》。 *Biological Cybernetics*，63 3 169( )：–176，1990。[JR22] A. Jentzen 和 A. Riekert。《深度神经网络训练中全局最小值的存在性及梯度下降方法的收敛性分析》。 *Journal of Machine Learning* 1 2 141 246, , ( )：–。[2022
KKV03] P. C. Kainen、V. Kurkov´a 和 A. Vogt。《半空间示性函数线性组合的最佳逼近》。 *Journal of Approximation Theory*，122(2)：151–159，2003。[LMQ22] L.-H. Lim、M. Micha lek 和 Y. Qi。《最佳 k 层神经网络逼近》。 *Constructive Approximation* 55 1 583 604, 2022 , ( )：–。[-Loj63] S. Lojasiewicz。《实解析子集的一种拓扑性质》。 , ：–。[-Loj65] S. Lojasiewicz。《半解析集合》。 *Lectures Notes IHES (Bures-sur-Yvette)*，1965。[-Loj84] S. Lojasiewicz。《解析函数梯度的轨迹》。 *Seminari di geometria*，1982 1983 115/：–117，1984。[MHKC20] P. Mertikopoulos、N. Hallak、A. Kavis 和 V. Cevher。《非凸问题中随机梯度下降几乎必然收敛》。载于 *Advances in Neural Information Processing Systems*，第 33 卷，第 1117–1128 页，2020。[PRV21] P. Petersen、M. Raslan 和 F. Voigtlaender。《固定规模神经网络所生成函数集合的拓扑性质》。 *Foundations of Computational Mathematics*，21(2)：375–444，2021。[SCP16] G. Swirszcz、W. M. Czarnecki 和 R. Pascanu。《神经网络训练中的局部极小值》。arXiv:1611.06310，2016。[Sin70] I. Singer。《 *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*》。Springer，1970。[SS18] I. Safran 和 O. Shamir。《伪局部极小值在两层 ReLU 神经网络中很常见》。载于 *International conference on machine learning*，第 4433–4441 页。PMLR，2018。[Tad15] V. B. Tadi´c。《多重且非孤立极值情形下随机梯度搜索的收敛性及收敛速率》。 *Stochastic Processes and their Applications*，125(5)：1715–1755，2015。[VBB19] L. Venturi、A. S. Bandeira 和 J. Bruna。《单隐层神经网络优化景观中的伪谷》。 *Journal of Machine Learning Research*，20(133)：1–34，2019。

Steffen Dereich, Institute for Mathematical Stochastics, Faculty of Mathematics and Computer Science, University of Münster, Germany*Email address*: steffen.dereich@uni-muenster.de

Sebastian Kassing, Faculty of Mathematics, University of Bielefeld, Germany
Email address: skassing@math.uni-bielefeld.de