多层平均场网络的深度分离

Yunwei Ren卡内基梅隆大学
yunweir@andrew.cmu.edu

Mo Zhou杜克大学
mozhou@cs.duke.edu

Rong Ge杜克大学
rongge@cs.duke.edu

arXiv:2304.01063v1 [cs.LG] 3 Apr 2023

摘要

深度分离——为什么更深的网络比更浅的网络更强大——一直是深度学习理论中的一个重大问题。以往的研究往往聚焦于表征能力。例如，Safran 等人(2019年) 构造了一个函数，用三层网络可以轻松近似，但任何两层网络都无法近似。本文表明，这种分离实际上具有算法意义：可以使用一个含有多项式数量神经元的过参数化网络，高效地学习 Safran 等人(2019年) 构造的函数。我们的结果依赖于一种将平均场极限推广到多层网络的新方法，以及一种对损失进行分解的方法，该分解将无限宽平均场网络离散化所引入的误差分离出来。

1 引言

深度学习理论中的一个谜题是，为什么我们需要更深的网络。早期研究中，研究人员证明了深层网络能够表示浅层网络难以近似的函数（Eldan 和 Shamir（2016年）；Telgarsky，2016年；Poole 等人，2016年；Daniely，2017年；Yarotsky，2017年；Liang 和 Srikant，2017年；Safran 和 Shamir，2017年；Poggio 等人，2017年；Safran 等人（2019年）；Malach 和 Shalev-Shwartz，2019年；Vardi 和 Shamir，2020年；Venturi 等人，2022年；Malach 等人，2021年）。特别是，Eldan 和 Shamir（2016年）以及 Safran 等人（2019年）的开创性研究构造了一个简单函数（$f_*(\boldsymbol{x}) = \mathrm{ReLU}(1 - \|\boldsymbol{x}\|)$），它可以由一个三层神经网络计算，但无法由两层网络近似。

然而，这些结果只涉及神经网络的表征能力，并不能保证从合理的初始化出发训练深度神经网络就一定能学会这类函数。本文证明，可以训练一个神经网络，以任意精度逼近$f_*(\boldsymbol{x}) = \mathrm{ReLU}(1 - \|\boldsymbol{x}\|)$——这在两层网络与三层网络的能力之间实现了算法方面的分离。

为了分析训练动态，我们开发了一个新框架，将神经网络的平均场分析推广到多层（Chizat & Bach，2018年；Mei 等人，2018年）。因此，在训练过程中，所有层的权重都可能发生显著变化（这与许多先前关于神经切线核或固定低层表征的研究不同）。我们的分析还给出了损失的分解形式，使我们能够将多层的训练解耦。

在本文余下部分，我们首先介绍新的多层平均场分析框架，然后给出主要结果和技术。我们将在第1.3节讨论深度分离算法方面的若干相关工作。与标准平均场分析类似，我们先在第3节考察无限宽度动力学，然后讨论如何将结果离散化为多项式规模网络的新思路（见第4节）。

1.1 多层平均场框架

我们提出了一种将平均场分析扩展到多层网络的新方法。为简便起见，这里以三层网络为例进行说明。一般框架请参见附录A 。简而言之，我们将中间层拆分为两个线性层，并限制中间层的规模。更准确地说，我们

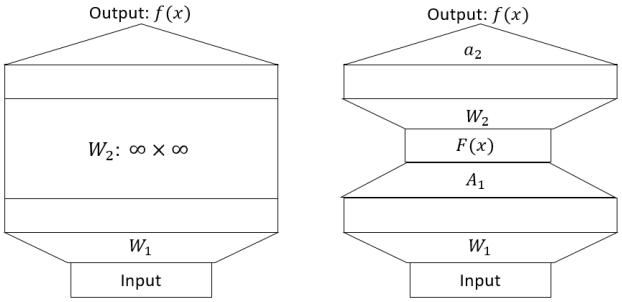


图 1：先前方法的区别Nguyen 与 Pham (2020)（左）与我们的框架（右）。

定义

$$f(\boldsymbol{x}) = \frac{1}{m_2}\boldsymbol{a}_2^\top \sigma(\boldsymbol{W}_2\boldsymbol{F}(\boldsymbol{x})), \quad \boldsymbol{F}(\boldsymbol{x}) = \frac{1}{m_1}\boldsymbol{A}_1\sigma(\boldsymbol{W}_1\boldsymbol{x}),$$

其中 $\boldsymbol{W}_1 \in \mathbb{R}^{m_1\times d}$， $\boldsymbol{A}_1 \in \mathbb{R}^{D\times m_1}$， $\boldsymbol{W}_2 \in \mathbb{R}^{m_2\times D}$ $\boldsymbol{a}_2 \in \mathbb{R}^{m_2}$ 是参数，而 $\boldsymbol{F}(\boldsymbol{x}) \in \mathbb{R}^D$表示隐藏特征。图1 展示了示意图。后文将把$\boldsymbol{x} \mapsto \boldsymbol{F}(\boldsymbol{x})$ 这一步称为第一层，把 $\boldsymbol{F}(\boldsymbol{x}) \mapsto f(\boldsymbol{x})$ 称为第二层，尽管它们实际上都是双层网络。

在无限宽极限下，我们固定隐藏特征维度 D，并令神经元数量m_1， m_2 趋于无穷。于是，我们得到无限宽网络

$$f(\boldsymbol{x}) = \mathop{\mathbb{E}}_{(a_2,\boldsymbol{w}_2)\sim\mu_2} a_2\sigma(\boldsymbol{w}_2\cdot\boldsymbol{F}(\boldsymbol{x})), \quad F_i(\boldsymbol{x}) = \mathop{\mathbb{E}}_{(a_1,\boldsymbol{w}_1)\sim\mu_{1,i}} a_1\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x}), \quad \forall i\in[D],$$

其中 $(\mu_{1,\ i})^{i\in[D]}$ 是定义在 $\mathbb{R}^{1+d}$ 上的分布，且在 $\boldsymbol{w}_1$上具有共享边缘分布；而 μ_2是定义在 $\mathbb{R}^{1+D}$上的分布。注意，与 Nguyen 与 Pham (2020年) 中的表述不同，这里使用神经元分布来描述隐藏层，因此它们在神经元置换下自动保持不变；这是均场网络最重要的性质之一。可以选取 μ_1， μ_2 作为有限多个神经元上的经验分布，从而还原有限宽网络。实际上，本文大部分内容都会采用这种做法，使我们的结果适用于神经元数量为多项式级别的有限宽网络。该网络可视为一个三层网络，其中间层为 $\boldsymbol{W}_2\boldsymbol{A}$，且为低秩结构。这让人联想到残差网络中使用的瓶颈结构（He 等人(2016年)），Allen-Zhu 与 Li (2020年) 等人也曾出于其他目的，在先前的理论分析中使用过这种结构。

学习器网络 现在我们准备介绍用于学习目标函数的具体网络。我们设定 $D = 1$，并将 a_1 与 $\boldsymbol{w}_1$ 耦合。

$$\begin{cases} F(\boldsymbol{x}) = F(\boldsymbol{x};\mu_1) := \mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu_1}\{\|\boldsymbol{w}\|\,\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\}, \\ f(\boldsymbol{x}) = f(\boldsymbol{x};\mu_2,\mu_1) := \mathop{\mathbb{E}}_{(w_2,b_2)\sim\mu_2}\sigma(w_2F(\boldsymbol{x};\mu_1)+b_2). \end{cases} \tag{1}$$

这里， σ 是 ReLU 激活函数，而 $\mu_1 \in \mathscr{P}(\mathbb{R}^d)$ 和 $\mu_2 \in \mathscr{P}(\mathbb{R}^2)$ 分别是编码第一层和第二隐藏层权重的分布。我们将每个第一层神经元乘以 $\|\boldsymbol{w}\|$，使 F 更加规整。这种2-齐次参数化也用于 Li 等人(2020年) 和Wang 等人 (2020年)。在论文的大部分内容中， μ_1 和 μ_2 是定义在多项式数量神经元上的经验分布。我们使用 μ_1， μ_2 来统一讨论无限宽和有限宽网络时的记号。

将中间层限制为只有一个维度（$D = 1$）就足够了，因为可以使用第一层 F（$\boldsymbol{x}$）和第二层的 $\alpha\|\boldsymbol{x}\| \mapsto \sigma$（$1 - \|\boldsymbol{x}\|$）来学习某些 $\alpha \in \mathbb{R}$ 的 $\boldsymbol{x} \mapsto \alpha\|\boldsymbol{x}\|$。对于计算 F（$\boldsymbol{x}$）的网络，我们不需要偏置项，因为目标函数关于 $\boldsymbol{x}$是齐次的。尽管我们将第一层限制为正值，但这不会限制网络的表征能力，因为第二层可以是正值，也可以是负值。对于第二层，尽管一个神经元就足够了，我们仍遵循该框架，对网络进行过参数化。

1.2 主要结果与我们的方法

我们的主要结果将上一节中的框架应用于Safran 等人(2019年)构造的函数（详见第2节）。简而言之，我们证明：[1]

定理 *1.1*（主要结果，非正式表述）。对于式 *(1)* 中定义的输入维度为 d 的学习器网络，给定任意 $\epsilon > 0$，我们可以选择层宽为 $m_1 = \textit{poly}(d, 1/\epsilon)$、$m_2 = \Theta(1)$，使得以至少 $1 - 1/\textit{poly}(d, 1/\varepsilon)$ 的概率，在随机初始化下运行一种简单的梯度流变体[2]，便可在 $T = \textit{poly}(d, 1/\epsilon)$ 时间内将损失 $\mathcal{L} := \mathbb{E}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \right\} /2$ 降至 ε。

这一结果表明，可以训练一个多层神经网络来学习函数 $\mathrm{ReLU}(1 - \|x\|)$，而任何两层网络都无法对其进行近似。由 Safran et al.（2019年）中重尾输入分布的选择所引发的一些技术细节，我们将在第 2 节讨论。

为了证明这一结果，我们首先刻画无限宽度动力学（见第3节）。具体而言，我们证明，在无限宽度动力学中，第一层始终会计算出 $\|\boldsymbol{x}\|$的一个倍数，而第二层的行为则类似于单个神经元。

然而，要将这种无限宽度分析离散化为多项式宽度网络，往往十分困难。主要难点在于网络中的误差可能会被放大：如果在训练开始时，第一层仅以 δ的误差接近于计算 $\|x\|$的一个倍数，那么这个 δ 值在训练过程中可能会呈指数增长（Mei 等人（2018年））。考虑到我们的动力学需要较长的多项式训练时间，这种指数增长是不可接受的。

为了解决这一问题，我们将分析划分为两个阶段；对于耗时较长的第二阶段，我们依赖于对损失函数的分解：

$$\mathcal{L} := \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \right\} \approx \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))^2 \right\} + \frac{\bar{w}_2^2}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2 \right\}. \quad (2)$$

这里，$\tilde{F}(\boldsymbol{x})$ 是 $\|\boldsymbol{x}\|$ 的一个倍数，且接近实际的第一层输出 $F(\boldsymbol{x})$；$\tilde{f}(\boldsymbol{x})$ 是在用 $\tilde{F}(x)$ 替代第一层时网络的输出——也就是说，第一层实际计算 $\|\boldsymbol{x}\|$ 的一个倍数时的输出（精确定义见（5））。因此，第一项刻画的是在第一层完美的条件下的损失；第二项则刻画第一层输出与 $\|\boldsymbol{x}\|$的一个倍数之间的差异。我们证明，这两项的梯度至少近似地互不影响。因此，我们可以将训练过程视为同时进行两件事：在第一层表示良好的条件下最小化损失（减小第一项），并使第一层输出更接近 $\|x\|$ 的一个倍数（减小第二项）。我们认为，这种分解揭示了神经网络中的下层如何获得有用的梯度信息，从而针对这一特定目标学习良好的表示。

1.3 相关工作

深度分离的算法方面 还有其他研究从算法角度深入探讨了深度分离。 Allen-Zhu &Li (2020年) 证明，多层二次网络能够以层次化的方式学习某些目标函数，而任何核方法或浅层神经网络都无法学习这些目标函数。我们的工作研究的是更常见的神经网络架构和目标函数。一项同期研究 Safran & Lee (2021年) 研究了与我们相似的问题，证明梯度下降结合特定的三层网络能够学习球指示函数，而任何两层网络都无法逼近该函数。从概念上说，我们的结果与其主要区别在于训练动态——Safran&Lee (2021年) 固定了第一层，而我们同时训练两层。这导致训练动态和证明技术都截然不同。

过参数化神经网络 一类研究探讨了过参数化神经网络的优化，这类网络将训练动态与神经切线核（NTK）核回归联系起来（例如，Jacot 等人，2018年；Allen-Zhu 等人，2018b；Du 等人，2018年）。然而，研究表明

[1]我们称某个量 a 为 $\mathrm{poly}(d, 1/\varepsilon)$，若其上界为 $C(d/\varepsilon)^C$，其中 $C > 0$ 是某个可能随行变化的普适常数。[2]严格来说，梯度流并不是一种真正的算法，但在理论分析中，通常会用它作为梯度下降的替代方法。关于如何将论证转化为梯度下降论证，请参阅附录 E 。

神经网络在 NTK 机制下的行为类似于核方法，并且已有多项下界研究（Yehudai & Shamir, 2019；Wei et al., 2019；Ghorbani et al., 2019；2020）。我们的训练动态不处于 NTK 机制，因为所有权重都发生了显著变化。另一类研究探讨了平均场极限下过参数化神经网络的优化（Mei et al., 2018；Chizat & Bach, 2018；Nitanda & Suzuki, 2017；Wei et al., 2019；Rotskoff & Vanden-Eijnden, 2018；Sirignano & Spiliopoulos, 2020）。Chizat 等人（2019）表明，在平均场机制下，参数可以偏离其初始化并学习有用特征，这与 NTK 机制不同。然而，现有的大多数研究都需要指数级/无限多个神经元，且未能提供多项式收敛速率。更多讨论见附录 A。

多层平均场 尽管平均场分析在双层过参数化网络的优化中已取得成功，但由于中间层的宽度趋于无穷，将其扩展到多层网络并不容易。许多研究尝试解决这一问题，以便将平均场分析推广到深度网络。例如，Nguyen 与 Pham (2020年); Pham 与 Nguyen (2021年);Araújo 等人 (2019年); Sirignano 与 Spiliopoulos (2021年); Fang 等人(2021年); Lu 等人 (2020年); Ding 等人 (2021年) 及其中引用的文献。与大多数现有研究不同，我们的多层平均场框架仍保持有限的隐藏特征维度，同时神经元数量可以趋于无穷，从而形成神经元分布。更多讨论见第 1.1 节和附录 A 。

轻度过参数化神经网络 近年来，许多研究探讨了如何使用轻度过参数化（多项式规模）网络学习特定目标函数（Allen-Zhu 等人，2018a；Allen-Zhu 与 Li，2019；Bai 与 Lee，2019；Dyer 与 Gur-Ari，2019；Woodworth 等人，2020；Bai 等人，2020；Huang 与 Yau，2020；Chen 等人，2020；Li 等人，2020；Wang 等人，2020；Zhou 等人，2021）。具体而言，这些研究不同于典型的平均场分析（通常考虑无限宽网络），也不同于典型的 NTK 分析（其中神经网络的行为类似于核方法）。我们的研究方向与之相近，但要将离散化扩展到新的多层框架，还需要新的见解。

2 预备知识

本节讨论 Safran 等人 （2019年）针对输入分布提出的额外技术条件，以及我们如何在训练过程中处理这些条件。

记号 对于向量 $\boldsymbol{x}$，我们令 $\|\boldsymbol{x}\|$ 表示其欧几里得范数。我们使用 $a = b \pm c$ 作为条件 $a \in [b-|c|, b+|c|]$的简写。对于分布 μ，我们用 $\boldsymbol{v} \in \mu$ 表示条件 $\boldsymbol{v}$属于 μ的支集。我们使用的其他记号大多是标准的。我们通常使用 $\boldsymbol{v}_1$ 和 $\boldsymbol{w}_1$ 表示第一层神经元，并使用（v_2，r_2）和（w_2，b_2）表示第二层神经元。对于神经元保留两套记号是有意为之的。当我们对神经元取期望时，使用$\boldsymbol{w}_1$和（w_2，b_2）。当考虑单个神经元时，使用 $\boldsymbol{v}_1$和（v_2，r_2）。对于向量，我们写作$\bar{\boldsymbol{v}} := \boldsymbol{v}/\|\boldsymbol{v}\|$。当上下文明确时，我们将使用 $\mathbb{E}_{\boldsymbol{x}}$ 作为 $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}$ 的简写。我们还使用 $\boldsymbol{v} \in \mu$作为 $\boldsymbol{v} \in \mathrm{supp}(\mu)$ 的简写。

目标函数与输入分布 我们考虑的目标函数为 $f_*(\boldsymbol{x}) = \sigma(1-\|\boldsymbol{x}\|)$，其中 $\sigma: \mathbb{R} \to \mathbb{R}$ 是 ReLU 激活函数。为描述输入分布，首先定义$\varphi(\boldsymbol{x}) := \left(\frac{R_d}{\|\boldsymbol{x}\|}\right)^{d/2} J_{d/2}(2\pi R_d \|\boldsymbol{x}\|)$，其中 $R_d = \frac{1}{\sqrt{\pi}}(\Gamma(d/2+1))^{1/d}$，且 J_ν是第一类 ν 阶贝塞尔函数。令 $\alpha, \beta > 0$为 Safran 等人（2019）中的通用常数（参见定理 5 的证明）。我们假设输入 $\boldsymbol{x} \in \mathbb{R}^d$ 从分布 $\mathcal{D}$ 中采样，其密度由 $\boldsymbol{x} \mapsto (\sqrt{d\beta}\alpha)^d \varphi^2(\sqrt{d\beta}\alpha\boldsymbol{x})$给出。Eldan 和 Shamir（2016）以及 Safran 等人（2019）已验证，这确实是一个有效的概率分布。另请注意，$\mathcal{D}$ 是一个球对称分布。关于 $\mathcal{D}$的更多性质，请参见附录 B.2。根据 Safran 等人（2019）的定理 5，宽度为 $\mathrm{poly}(d, 1/\varepsilon)$ 的双层网络无法在 $L^2(\mathcal{D})$中以 ε 的精度逼近 f_*。[3] 这种分布是重尾的，因为 $\mathbb{E}_{x\sim\mathcal{D}}[\|x\|^2]$ 未定义。选择这种重尾分布主要是为了证明下界。我们的训练结果适用于大多数合理的球对称分布。

[3]严格来说，Safran 等人 (2019年) 的结果要求 $\varepsilon = O(1/d^6)$。即使在该条件下，我们的算法仍能用多项式数量的(d) 个神经元学习该函数，而任何两层网络都无法做到这一点，因此这仍然构成有效的分离。

训练算法与主要结果 我们使用带裁剪的梯度流，通过均方误差损失训练一个多项式规模网络。我们将损失写作

$$\mathcal{L} = \mathcal{L}(\mu_1, \mu_2) = \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \right\} =: \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}), \tag{3}$$

定义 $S(\boldsymbol{x}) = (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \mathbb{E}_{w_2,\, b_2} \{\sigma'(w_2 F(\boldsymbol{x}) + b_2) w_2\}$。可以验证，神经元的动力学由下式给出

$$\begin{cases} \dot{\boldsymbol{v}}_1 = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}} \left\{ \Pi_{R_{\boldsymbol{v}_1}} \left[S(\boldsymbol{x}) \left(\bar{\boldsymbol{v}}_1 \sigma(\boldsymbol{v}_1 \cdot \boldsymbol{x}) + \|\boldsymbol{v}_1\| \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \boldsymbol{x} \right) \right] \right\}, \\ \dot{v}_2 = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}} \left\{ \Pi_{R_{v_2}} \left[(f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma'(v_2 F(\boldsymbol{x}) + r_2) F(\boldsymbol{x}) \right] \right\}, \\ \dot{r}_2 = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}} \left\{ \Pi_{R_{r_2}} \left[(f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma'(v_2 F(\boldsymbol{x}) + r_2) \right] \right\}, \end{cases} \tag{4}$$

其中 Π_R 表示投影到半径为 R的球上，而 $R_{\boldsymbol{v}_1} = \Theta(d)$、 $R_{v_2} = \Theta(d^3)$和$R_{r_2} = \Theta(1)$ 是投影阈值。我们加入这些额外的梯度裁剪，因为若不进行裁剪，由于分布 $\mathcal{D}$具有重尾性质，梯度将无法得到良好定义。值得注意的是，实践中确实广泛使用梯度裁剪来避免梯度爆炸（Pascanu 等人，2013；Zhang 等人，2020）。事实上，我们认为，只要一般的球对称分布 $\mathcal{D}$ 具有更好的正则性，即使不使用梯度裁剪，我们的优化结果仍然成立。

为了初始化学习器网络，我们使用 Unif($\sigma_1 \mathbb{S}^{d-1}$) 初始化第一层权重 $\boldsymbol{w}_1$，
$\mathcal{N}\ 0\ \sigma_2^2$ w_2 b_2 σ_r (,) 作为第二层权重，并将所有第二层偏置设为 ，其中σ_1，σ_2，σ_r 是一些较小的正实数。出于技术上的便利，我们使用高斯分布来初始化球面上的 $\boldsymbol{w}_1$。我们将偏置项初始化为一个较小的正值，以确保所有第二层神经元在初始化时都被激活，从而避免出现零梯度。

现在我们可以给出主要结果了。该结果表明，在我们的平均场框架中定义的、具有多项式规模学习器网络（1）的梯度流能够高效地学习 $f_*(\boldsymbol{x}) = \sigma(1 - \|\boldsymbol{x}\|)$，而这无法由任何两层网络近似（Safran 等人，2019 年）。

定理 *2.1*（主要结果）。给定式 *(1)* 中定义、并采用上述初始化的学习器网络，假设我们在该有限宽网络上对损失 *(3)* 使用裁剪 *(4)* 运行梯度流（假设其存在）。那么，对于任意 $\epsilon > 0$，我们可以选择 $m_1 = \mathit{poly}_{m_1}(d, 1/\epsilon)$、$m_2 = \Theta(1)$、 $\sigma_1 = 1/\Theta(1)$ $_{\sigma_1}\ d\ 1/\epsilon$ $\sigma_2 = 1/$ $_{\sigma_2}\ d\ 1/\epsilon$ $\sigma_r = \Theta\ 1$ $R_{\boldsymbol{v}_1} = \Theta\ d$ $R_{v_2} = \Theta\ d^3$ $R_{r_2} =$ *poly (,)*、*poly (,)*、*( )*、*( )*、*( )* 和 *( )*，使得在随机初始化下，以至少 $1 - 1/\mathit{poly}(, 1/\varepsilon)$ 的概率，经过 $T = \mathit{poly}(d, 1/\epsilon)$ 时间后，损失 $\mathcal{L} \le \varepsilon$ *within*。

3 无限宽度动力学

我们的证明包括分析无限宽平均场网络的动力学，并控制离散化误差。本节中，我们刻画了无限宽度动力学。为便于阐述，在本小节中，我们假设不存在投影且梯度定义良好，并将如何处理投影的讨论留到第 4 节。

首先请注意，输入分布 $\mathcal{D}$ 和无限宽网络都是球对称的。也就是说，对于任意 $\boldsymbol{x}$，$\boldsymbol{x}' \in \mathbb{R}^d$，满足 $\|\boldsymbol{x}\| = \|\boldsymbol{x}'\|$，其密度/函数值都相同。任何球对称的 $g\colon \mathbb{R}^d \to \mathbb{R}$ 都可以由函数 $h\colon [0, \infty) \to \mathbb{R}$ 表征，该函数满足 $h(\|\boldsymbol{x}\|) = g(\boldsymbol{x})$。为方便起见，我们也将滥用记号，用 $g\colon \mathbb{R} \to \mathbb{R}$ 表示该函数 h。

假设 μ_1 第一层神经元的分布是球对称的（这在至少初始化时成立），我们可以利用以下引理，通过一个简单的函数来近似第一层。其证明可见附录B.3。

引理 3.1。设 μ为球对称分布。我们有

$$\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu} \|\boldsymbol{w}\| \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) = C_\Gamma \frac{\mathbb{E}_{\boldsymbol{w}\sim\mu} \|\boldsymbol{w}\|^2}{\sqrt{d}} \|\boldsymbol{x}\| \quad \textit{where} \quad C_\Gamma := \frac{\Gamma(d/2)\sqrt{d}}{2\sqrt{\pi}\Gamma((d+1)/2)}.$$

注意，由于 $d \to \infty$，我们有 $C_\Gamma \to 1/\sqrt{2\pi}$，因此对于所有 d，C_Γ 均有统一上界。

该引理意味着，在无限宽极限下，我们有 $F(\boldsymbol{x}) = \alpha \|\boldsymbol{x}\|$，其中某个实数 $\alpha > 0$ 至少在初始化时成立。这启发我们将无限宽近似定义为：

$$\alpha := \frac{C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{w}_1 \sim \mu_1} \|\boldsymbol{w}_1\|^2, \quad \tilde{F}(\boldsymbol{x}) := \alpha \|\boldsymbol{x}\|, \quad \tilde{f}(\boldsymbol{x}) := \mathop{\mathbb{E}}_{(w_2, b_2) \sim \mu_2} \sigma(w_2 \tilde{F}(\boldsymbol{x}) + r_2). \tag{5}$$

注意，(5) 定义良好，无论 μ_1 是否为无限宽，尽管只有在无限宽的情况下才有 $\tilde{F} = F$。稍后在第 4 节中，我们将证明，在离散化部分的整个过程中，$F \approx \tilde{F}$。

对于无限宽网络，可以设想，多亏了对称性，只要 μ_1 在时刻 t 呈球对称，那么第一层神经元的方向都不会改变，且范数的变化也一致，即不依赖于方向 $\bar{\boldsymbol{v}}_1$。（证明见附录 B.4 。）因此，μ_1 将保持球对称。形式化地，可以证明，对于任意球对称的 g：$\mathbb{R}^d \to \mathbb{R}$，有

$$\mathop{\mathbb{E}}_{\boldsymbol{x}} \{g(\boldsymbol{x}) \sigma(\boldsymbol{v} \cdot \boldsymbol{x})\} = \frac{C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \{g(\boldsymbol{x}) \|\boldsymbol{x}\|\} \|\boldsymbol{v}\| \text{ 以及 } \mathop{\mathbb{E}}_{\boldsymbol{x}} \{g(\boldsymbol{x}) \sigma'(\boldsymbol{v} \cdot \boldsymbol{x}) \boldsymbol{x}\} = \frac{C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \{g(\boldsymbol{x}) \|\boldsymbol{x}\|\} \bar{\boldsymbol{v}}, \text{ 其}$$

中 $\bar{\boldsymbol{v}} = \boldsymbol{v} / \|\boldsymbol{v}\|$。同样，这两个恒等式的证明见附录 B.3。将这些恒等式应用于 $\dot{\boldsymbol{v}}_1$，并令 $g \equiv S$，即可得到

$$\dot{\boldsymbol{v}}_1 = \frac{2 C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \{S(\boldsymbol{x}) \|\boldsymbol{x}\|\} \boldsymbol{v}_1.$$

因此，μ_1 始终是某个球面上的均匀分布。此外，我们有[4]

$$\dot{\alpha} = \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \frac{\partial \alpha}{\partial \boldsymbol{w}_1} \frac{\mathrm{d} \boldsymbol{w}_1}{\mathrm{d} t} = \frac{4 C_\Gamma^2}{d} \mathop{\mathbb{E}}_{\boldsymbol{x}} \{S(\boldsymbol{x}) \|\boldsymbol{x}\|\} \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \|\boldsymbol{w}_1\|^2 = \frac{4 C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \{S(\boldsymbol{x}) \|\boldsymbol{x}\|\} \alpha.$$

这意味着，第一层的动力学也可以仅由 α 刻画。这使得第一层的动力学可以简化为一个实数 α。也就是说，第一层的输出仅取决于 α 和 $\boldsymbol{x}$，而 α 的动力学也仅取决于 α，而非每个神经元 $\boldsymbol{w}_1$。换句话说，在这种无限宽情形下，我们无需考察 $\boldsymbol{w}_1$ 的实际动力学。我们将在后文证明，第二层的分散度始终很小，因此第二层可以用 $\alpha \|\boldsymbol{x}\| \mapsto \sigma(\bar{w}_2 \alpha \|\boldsymbol{x}\| + b_2)$ 近似，其中 $(\bar{w}_2, b_2) = \mathbb{E}(w_2, b_2)$。综合这些观察，可以用三个量刻画整个网络的动力学：α、$\bar{w}_2$ 和 b_2。

我们用对 F 的另一种解释来结束本节，这种解释在第4.2节中会很有用。由于我们知道，在理想化情形下，F 应当是球对称的。因此，将“理想化”的 $\tilde{F}$ 定义为球面上的平均值是合理的，也就是 F，$F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}' \in \|\boldsymbol{x}\| \mathbb{S}^{d-1}} F(\boldsymbol{x}')$。注意，在引理3.1中，期望是对神经元取的，而这里则是对输入取的。不过，与引理3.1的证明类似，仍然可以证明

$$\mathop{\mathbb{E}}_{\boldsymbol{x}' \in \|\boldsymbol{x}\| \mathbb{S}^{d-1}} F(\boldsymbol{x}') = \mathop{\mathbb{E}}_{\boldsymbol{w} \sim \mu_1} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in \|\boldsymbol{x}\| \mathbb{S}^{d-1}} \|\boldsymbol{w}\|^2 \sigma(\bar{\boldsymbol{w}} \cdot \boldsymbol{x}) = \frac{C_\Gamma \mathbb{E}_{\boldsymbol{w} \sim \mu_1} \|\boldsymbol{w}\|^2}{\sqrt{d}} \|\boldsymbol{x}\| = \alpha \|\boldsymbol{x}\|.$$

换言之，这两种推导是等价的。从某种意义上说，这意味着无限宽网络可以解释为对实际有限宽网络进行的对称化。

4 使用多项式规模网络离散化动力学

本节将展示如何对无限宽度动力学进行离散化，从而得到我们的主要结果。仿真结果见图 2 。可以看到，尽管网络宽度有限，但在任意时刻步，函数 $f(x)$ 都接近于形如 $\boldsymbol{x} \mapsto \sigma(b_2 - \bar{w}_2 \alpha \|\boldsymbol{x}\|)$的函数，并且在整个训练过程中，第二层权重都高度集中。

令 $\delta_2 := \max_{(v_2, r_2), (v_2', r_2')} \|(v_2, r_2) - (v_2', r_2')\|$ 为第二层的分散度，我们将训练过程分为两个阶段。回顾一下，$(\bar{w}_2, b_2) := \mathbb{E}_{(w_2, b_2) \sim \mu_2} (w_2, b_2)$。在阶段 1 中，$\bar{w}_2$ 将降至 $-\mathrm{poly}(d) \delta_2$。我们证明，一旦满足这一条件，就可以忽略 (4) 中的投影算子（也就是说，相应的项永远不会超过阈值，见引理4.1）。在阶段 2 中，我们证明该网络可以在多项式时间内拟合目标函数。

[4] 与标准的平均场论证一样，我们将梯度按 m 重新缩放，使其不会随着 $m \to \infty$ 而趋于 0。在大多数梯度计算中，$\partial_{\boldsymbol{v}} g \cdot \boldsymbol{v} = \boldsymbol{w}$ 这等价于使用形式规则 $\partial_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{w}} g(\boldsymbol{v}, \boldsymbol{w}) = \mathbb{E}_{\boldsymbol{w}} \partial_{\boldsymbol{v}} g(\boldsymbol{v}, \boldsymbol{w})$。

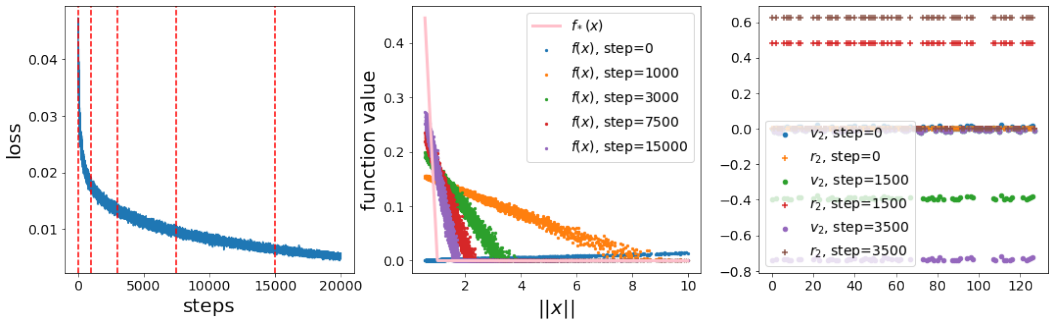


图 2：仿真结果。左图展示了训练过程中的损失。每条竖直虚线对应于另外两幅图中绘制的一个时间点。中间图展示了 f 在某些步骤中的形状。右图展示了某些步骤中第二层神经元的值。可以观察到，$f \approx f$ 确实成立，并且第二层神经元集中在 $(\bar{w}_2,\ b_2)$ 附近，这与我们的理论分析相符。仿真是在一个有限宽网络上进行的，其宽度为 $m_1 = 512$、$m_2 = 128$，输入维度为 $d = 100$。

4.1 阶段 1：移除投影

我们的第一步证明，经过一小段时间的训练后，就可以忽略 (4) 中的投影算子。为了说明在某些情况下为什么可以忽略投影，首先注意到，if $f \approx f$，第二层神经元会集中在其均值附近，$b_2 = \Theta(1)$ and $\bar{w}_2 < 0$，那么 $f \approx \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\| + b_2)$ 在 $\{\|\boldsymbol{x}\| \le \Theta(1/|\bar{w}_2\alpha|)\}$之外会消失，因此对于这些较大的 $\boldsymbol{x}$，梯度也会消失。同时，通过对梯度范数取上界，可以证明，若要触发投影，$\|\boldsymbol{x}\|$ 必须足够大。因此，当 f 下降得足够快时，$f(\boldsymbol{x})$ 会在 $\|\boldsymbol{x}\|$ 变得过大之前达到 0。形式上，我们有以下引理，其证明见附录C。

引理4.1。选择投影阈值 $R_{\boldsymbol{v}_1} = \Theta(d)$，$R_{v_2} = \Theta(d^3)$ 和 $R_{r_2} = \Theta(1)$ (4) $\alpha = \Theta(1/\sqrt{d})$。那么，如果所有第二层权重均为非正数，则 $\dot{r}_2$、$\dot{\boldsymbol{v}}_1$ 和 $\dot{v}_2$ 中的投影算子将不再被激活；对于某个足够大的常数，$-\bar{w}_2 > \Theta(1)\delta_2$，并且对于某个足够大的常数，$-\bar{w}_2 \ge \Theta(1)/R_{v_2}$。

基于这个引理，我们进一步将阶段 1 分为三个子阶段。定义 $T_{1.1}$ 为所有第二层权重均变为负值的最早时刻，$T_{1.2}$ 和 $T_{1.3}$ 分别为 $|\bar{w}_2|$ 变为 $\Theta(d)\delta_2$和 $\Theta(1/R_{v_2})$的最早时刻。它们分别表示阶段 1.1、1.2 和 1.3 的结束时刻。我们要求 $|\bar{w}_2|$ 在 $\Theta(d)\delta_2$，而不是 $\Theta(1)\delta_2$，以便阶段 1.3 的初始状态更加规整。根据定义和引理4.1，每个子阶段结束后，都可以忽略一个额外的投影；阶段 1 结束后，所有投影都可以忽略。

T阶段 1 的主引理如下。回忆一下，$R_{\boldsymbol{v}_1}$，R_{v_2}，R_{r_2} 是裁剪阈值 .

引理 4.2 (阶段 1，非正式表述)。将阶段 *1* 的结束时间定义为 $T_1 := \inf\{t \ge 0 : -\bar{w}_2(t) = C_1/R_{v_2}\}$ 对于某个较大的常数 C_1。在定理 *2.1* 的假设下，我们有 $T_1 \le \mathrm{poly}(d, 1/\varepsilon)$，且在整个阶段 1 期间，以下条件均成立。

(a) **第一层的近似误差**。对于每个 $\boldsymbol{v}_1 \in \mu_1$，切向移动和径向扩散度均可控制为 $\|\bar{\boldsymbol{v}}_1(t) - \bar{\boldsymbol{v}}_1(0)\| \le \delta_{1\,T,}^{(1)}(t)$ 以及 $\|\boldsymbol{v}_1\|^2 = (1 \pm \delta_{1\,R,}^{(1)}(t))\mathbb{E}\|\boldsymbol{w}_1\|^2$，其中 $\delta_{1\,T,}^{(1)}$ 和 $\delta_{1\,,\,R}^{(1)}$ 是两个始终较小的过程。

(b) 第二层的扩散。对于任意 (v_2, r_2)、(v_2', r_2')，$\in \mu_2$, $\|(v_2, r_2) - (v_2', r_2')\|$ 都很小。(c) 正则性条件。 $r_2 = \Theta(1)$ **对所有** (v_2, r_2)，$\in \mu_2$, $|\bar{w}_2| = O(1/R_{v_2}) = O(1/d^3)$ **且** $\alpha = \Theta(\sqrt{d}/R_{\boldsymbol{v}_1}) = \Theta(1/d^{1.5})$。

前两个条件意味着近似 $f(\boldsymbol{x}) \approx \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\| + b_2)$ 在整个阶段 1 中都成立，而第三个条件描述了 f 在阶段 1 中的形状。为保持这些条件，我们采用所谓的连续性论证，它可以看作数学归纳法的连续版本。关于这一技巧的说明，请参见附录 B.1 。

根据近似 $F(\boldsymbol{x}) \approx \alpha\|\boldsymbol{x}\|$ 以及事实 $f(\boldsymbol{x})\sigma'(v_2 F(\boldsymbol{x}) + r_2) = f(\boldsymbol{x})$ 对大多数 $\boldsymbol{x}$成立，我们可以将 v_2 的动力学改写为

$$\dot{v}_2 \approx \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{\Pi_{R_{v_1}}\left[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\alpha\|\boldsymbol{x}\|\right]\right\}.$$

由于 f 比 f_*平坦得多，当 f_* 趋于零时， f 仍为 $\Omega(1)$，这是因为 $\|\boldsymbol{x}\| \geq 1$。因此，右端始终为负。事实上，我们证明它是 $-\Theta(\alpha\log d)$。回想一下， $T_{1.2}$ 是 $|\bar{w}_2|$ 达到 $\Theta(d\delta_2)$ 的时间。如果 δ_2 大致保持不变，那么阶段 1.1 和 1.2 所需的时间与初始 δ_2 成正比。于是，我们可以通过选择足够小的 σ_2来减小初始 δ_2 。这也有助于控制阶段 1.1 和 1.2 中 $\boldsymbol{v}_1$ 和 r_2 的变化，因为它们的动力学取决于 $|w_2|$。

还需要证明，在阶段 1.1 和 1.2 期间， δ_2 不会增加太多，以维持 $\bar{}$近似 $f(\boldsymbol{x}) \approx \sigma(\bar{w}_2 F(\boldsymbol{x}) + b_2)$。直观地说，这是因为对于范数较小的输入，梯度 $\nabla_{v_2}\mathcal{L}(\boldsymbol{x})$ 不依赖于 (v_2, r_2) 本身；对于范数较大的输入，由于梯度裁剪，它们对梯度的贡献不会太大。因此，在阶段 1.1 和 1.2 中， v_2 的动力学近似均匀，于是不同 (v_2, r_2)、(v_2', r_2') 之间的距离保持较小。

在阶段 1.3 中，同样的方法行不通，因为此时 $\bar{w}_2$ 的目标值不再依赖于 δ_2，而我们需要对第一层进行更精细的分析。回想一下，在阶段 1.2 之后，可以忽略 $\dot{\boldsymbol{v}}_1$ 中的投影。因此，我们可以将 $\dot{\boldsymbol{v}}_1$ 分解为径向方向和切向方向上的分量：

$$\begin{aligned}\dot{\boldsymbol{v}}_1 = \mathrm{Rad}(\dot{\boldsymbol{v}}_1) + \mathrm{Tan}(\dot{\boldsymbol{v}}_1) &= \langle\dot{\boldsymbol{v}}_1, \bar{\boldsymbol{v}}_1\rangle\bar{\boldsymbol{v}}_1 + (\boldsymbol{I} - \bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\dot{\boldsymbol{v}}_1 \\ &= 2\mathop{\mathbb{E}}_{\boldsymbol{x}}\{S(\boldsymbol{x})\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\} + \|\boldsymbol{v}_1\|\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{S(\boldsymbol{x})\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{I} - \bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\boldsymbol{x}\right\}.\end{aligned}$$

˜ ˜接着，我们将 $S(\boldsymbol{x}) \approx (f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\,\bar{w}_2 = (f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\,\bar{w}_2 + (f(\boldsymbol{x}) - f(\boldsymbol{x}))\,\bar{w}_2$ 写成.....。与 $f_* - f$ 相关的˜项，基本上就是无限宽度动力学中预期会出现的项。对于这些项，径向移动是均匀的，而切向移动为 0。接着，我们利用第一层的径向扩散和切向移动，对与 $f - f$ 相关的˜项进行界定，得到 $\mathrm{d}\,\mathrm{d}_t\left(\delta_{1,R}^{(1)} + \delta_{1,T}^{(1)}\right) \lesssim^{O}\left(\frac{1}{d^{2.5}}\left(\delta_{1,R}^{(1)} + \delta_{1,T}^{(1)}\right)\right)$（参见引理 C.16）。尽管有了这个界，误差仍可能以指数速度增长（$\exp(t/d^{2.5})$），但这已足够，因为阶段 1.3 只需 $O(d^{1.5})$ 的时间。

4.2 阶段 2：拟合目标函数

阶段 2 的目标是使梯度流在多项式时间内收敛到一个损失至多为 ε 的点。本阶段的主要难点在于，我们需要更仔细地界定第一层的近似误差，因为阶段 2 可能持续较长时间，而阶段 1 中使用的蛮力估计在训练后期过于宽松。我们记 $F := F/\alpha$，并使用 $\left\|\bar{F}|_{\mathbb{S}^{d-1}} - 1\right\|$ 和 $\left\|\bar{F} - \|\cdot\|_2\right\|_{L^2}$ 来衡量近似误差。严格来说，对于 L^2 误差，我们只考虑满足 $\|\boldsymbol{x}\| \leq \Theta(1/|\bar{w}\alpha|) = \mathrm{poly}(d)$ 的那些$\boldsymbol{x}$，否则它可能没有定义。这样做是合理的，因为正如我们前面讨论过的，当 $\boldsymbol{x}$ 较大时， f 会趋于零。在阶段 2 中， $\mathbb{E}_{\boldsymbol{x}}$ 始终表示 $\mathbb{E}_{\|\boldsymbol{x}\|\leq\Theta(1/|\bar{w}_2\alpha|)}$；为简化表述，我们通常不明确说明这一点。阶段 2 的主要结果如下。

引理 4.3 (阶段 2（非正式）)。 将阶段 *2* 的结束时间定义为 $T_2 := \text{下确界}\{t \geq T_1 : \mathcal{L} = \varepsilon\}$。在定理 *2.1*的假设下，我们有 $T_2 - T_1 \leq \text{多项式}(d, 1/\varepsilon)$，且在整个阶段 2 期间满足以下条件：

(a) **第一层的近似误差**。两者 $\left\|\bar{F} - \|\cdot\|\right\|_{L^2}$ 和 $\left\|\bar{F}|_{\mathbb{S}^{d-1}} - 1\right\|_{L^\infty}$ 均很小。 *e*

(b) **第二层的分散度**。最大值$_{(v_2,r_2),\ (v_2',\ r_2')}$ $\|(v_2,\ r_2) - (v_2',\ r_2')\|$ 不会增长。

(c) **正则性条件**。 形状 f 与图 *2* 中所示的形状相似。

如前所述，主要的技术挑战是界定第一层的近似误差。整体策略是先证明，在阶段 2 中，L^2 误差几乎不会增大；然后证明，只要 L^2 误差足够小，就也能控制 L^∞ 误差。与阶段 1 不同，阶段 2 中 $|\bar{w}_2\alpha|$ 相当大，因此第一层可以从损失函数中接收到一些信号。

直观来说，这一信号应促使第一层更接近 $\|\boldsymbol{x}\|$ 的某个倍数，因为全局最优解正是如此。形式上，我们先证明以下近似关系：

$$\mathcal{L} \approx \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))^2\right\}+\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\tilde{F}(\boldsymbol{x})-F(\boldsymbol{x}))^2\right\}, \tag{6}$$

从这个意义上说，两侧的梯度 $\nabla_{\boldsymbol{v}_1}$ 近似相同，其中 $f(x)$ 定义为 $\mathbb{E}_{(w_2,b_2)\sim\mu_2}\sigma(w_2F(x)+b_2)$。式（6）的第一项衡量目标函数与无限宽网络之间的距离，第二项衡量第一层的近似误差。从某种意义上说，可以将这个公式视为对均场网络进行离散化时的偏差-方差分解。

有了这一近似结果，我们接下来证明，得益于 F 的二次齐次性，经过一定归一化后，第一项不会影响第一层的近似误差。与此同时，由于我们采用梯度流，第二项只会使近似误差减小。

为建立式（6），我们首先将损失函数分解为

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))^2\right\}+\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\tilde{f}(\boldsymbol{x})-f(\boldsymbol{x}))^2\right\}+\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))(\tilde{f}(\boldsymbol{x})-f(\boldsymbol{x}))\right\}\\ &=: \mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3.\end{aligned}$$

我们声称，$\mathcal{L}_2$ 近似等于式（6）的第二项，而第三项近似等于 0[5]。设 X_1 为一个最大的球对称集合，在该集合上 $v_2F(\boldsymbol{x})+r_2>0$ 对所有 $(v_2,r_2)\in\mu_2$ 都成立。我们证明，这些 $\boldsymbol{x}$ 位于 X_1 之外的部分贡献很小。因此，我们可以将 $\mathcal{L}_2$ 改写为

$$\begin{aligned}\mathcal{L}_2 &\approx \frac{1}{2}\mathop{\mathbb{E}}_{X_1}\left\{\left(\mathop{\mathbb{E}}_{w_2,b_2}(w_2\tilde{F}(\boldsymbol{x})+b_2)-\mathop{\mathbb{E}}_{w_2,b_2}(w_2F(\boldsymbol{x})+b_2)\right)^2\right\}\\ &= \frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{X_1}\left\{(\tilde{F}(\boldsymbol{x})-F(\boldsymbol{x}))^2\right\}\approx\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\tilde{F}(\boldsymbol{x})-F(\boldsymbol{x}))^2\right\}.\end{aligned}$$

类似地，我们可以将 $\mathcal{L}_3$ 改写为 $\mathcal{L}_3\approx\bar{w}_2\mathbb{E}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))(\tilde{F}(\boldsymbol{x})-F(\boldsymbol{x}))\right\}$。回顾第3节，其中 $F(\boldsymbol{x})=\mathbb{E}_{\boldsymbol{x}'\in\|\boldsymbol{x}\|\mathbb{S}^{d-1}}F(\boldsymbol{x})$。由此可知，对于任意球对称函数 g，很容易验证：$\mathbb{R}^d\to\mathbb{R}$, $\mathbb{E}_{\boldsymbol{x}}\{g(\boldsymbol{x})F(\boldsymbol{x})\}=\mathbb{E}_{\boldsymbol{x}}\left\{g(\boldsymbol{x})\tilde{F}(\boldsymbol{x})\right\}$。令 $g=f_*(x)-f(x)$，得到 $\mathcal{L}_3\approx 0$。将这两个估计结合起来，即得式（6）。

在 L^2 误差始终较小的前提下，我们证明，忽略一些高阶项后，

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\bar{\boldsymbol{x}})\right|\lesssim O(d^3)\left\|\bar{F}-\|\cdot\|_2\right\|_{L^2},\quad\forall\bar{\boldsymbol{x}}\in\mathbb{S}^{d-1}.$$

换言之，$\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x})$ 的变化可以由 L^2 误差界定。因此，$\left\|\bar{F}|_{\mathbb{S}^{d-1}}-1\right\|_{L^\infty}$ 始终较小，只要我们选择足够大的 m_1，使得 $F(x)|_{x\in\mathbb{S}^{d-1}}$ 在初始化时接近1 即可。这并不令人意外，毕竟在无限宽度动力学中 $F(x)|_{x\in\mathbb{S}^{d-1}}=1$。上述论证的正式证明见第D.2节。

鉴于可以控制近似误差，随后便可利用无限宽度动力学推导出收敛速率。详见第 D.3 节。

5 结论

本文提出了一个将平均场极限推广到多层网络的新框架，并利用该框架证明，三层网络能够学习一种无法由双层网络近似的函数。仍有许多开放问题：对于当前的目标函数，损失具有球对称性，因此第一层神经元几乎不会沿切向移动；如果函数改为 $\sigma(1-\|P_S\boldsymbol{x}\|)$，其中 P_S 是投影到某个未知子空间的投影，又会怎样？对于需要大小大于 1 的中间层的函数，又该如何处理？能否将鞍点分析推广到更深的网络？我们希望这项工作能成为理解深度神经网络如何学习有用特征的起点。

[5]为便于表述，这里讨论的是函数值，而不是梯度。严格来说，这种说法并不准确，因为函数值较小并不一定意味着梯度较小。不过，两者的思路本质上相同。请参阅第D.2 节，了解实际证明。

致谢

本研究获得美国国家科学基金会资助，项目编号为 DMS-2031849、CCF-1845171（CAREER）和 CCF-1934964（Tripods），并获得斯隆研究奖学金支持。

参考文献

Zeyuan Allen-Zhu 和 Yuanzhi Li。残差网络能够高效学习什么，超越核方法？*arXiv* 预印本 *arXiv:1905.10337*，2019年。

Zeyuan Allen-Zhu 和 Yuanzhi Li。反向特征校正：深度学习如何执行深度学习。*arXiv* 预印本 *arXiv:2001.04413*，2020年。

Zeyuan Allen-Zhu、Yuanzhi Li 和 Yingyu Liang。超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a。

Zeyuan Allen-Zhu、Yuanzhi Li 和 Zhao Song。通过过度参数化建立深度学习的收敛理论。*arXiv* 预印本 *arXiv:1811.03962*，2018年b。

Dyego Ara´ujo、Roberto I Oliveira 和 Daniel Yukimura。某些深度神经网络的平均场极限。*arXiv* 预印本 *arXiv:1906.00193*，2019年。

Yu Bai 和 Jason D Lee。超越线性化：宽神经网络的二次及高阶近似。*arXiv* 预印本 *arXiv:1910.01619*，2019年。

YuBai、Ben Krause、Huan Wang、Caiming Xiong 和 Richard Socher。泰勒化训练：旨在更好地近似有限宽度下的神经网络训练。*arXiv* 预印本 *arXiv:2002.04010*，2020年。

Minshuo Chen、Yu Bai、Jason D Lee、Tuo Zhao、Huan Wang、Caiming Xiong 和 Richard Socher。迈向理解层次化学习：神经表征的优势。*arXiv* 预印本 *arXiv:2006.13436*，2020年。

Lenaic Chizat 和 Francis Bach。利用最优传输研究过参数化模型中梯度下降的全局收敛。在神经信息处理系统进展，第3036–3046页，2018年。

Lenaic Chizat、Edouard Oyallon 和 Francis Bach。关于可微编程中的惰性训练。在神经信息处理系统进展，第2933–2943页，2019年。Amit Daniely。神经网络的深度分离。在学习理论会议，第690–696页。PMLR，2017年。Zhiyan Ding、Shi Chen、Qin Li 和 Stephen Wright。深度残差网络的过参数化：零损失与平均场分析。*arXiv* 预印本 *arXiv:2105.14417*，2021年。Simon S Du、Xiyu Zhai、Barnabas Poczos 和 Aarti Singh。梯度下降可证明能够优化过参数化神经网络。*arXiv* 预印本 *arXiv:1810.02054*，2018年。Ethan Dyer 和 Guy Gur-Ari。基于费曼图的宽神经网络渐近行为。*arXiv* 预印本 *arXiv:1909.11304*，2019年。

Ronen Eldan 和 Ohad Shamir。前馈神经网络的深度力量。载于 Vitaly Feldman、Alexander Rakhlin 和 Ohad Shamir（编），第*29*届学习理论会议，机器学习研究论文集第49卷，第907–940页，美国纽约州纽约市哥伦比亚大学，2016年6月。PMLR。网址：http://proceedings.mlr.press/v49/eldan16.html。

Cong Fang、Jason Lee、Pengkun Yang 和 Tong Zhang。基于特征建模：过参数化深度神经网络的平均场框架。载于学习理论会议，第 1887–1936 页。PMLR，2021年。

Behrooz Ghorbani、Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。双层神经网络惰性训练的局限性。载于神经信息处理系统大会，2019年。

Behrooz Ghorbani、Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。神经网络何时优于核方法？arXiv 预印本 arXiv:2006.13409，2020年。何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于 2016年 IEEE计算机视觉与模式识别会议（CVPR），第 770–778 页，2016年。doi：10.1109/CVPR.2016.90。黄嘉阳和 Horng-Tzer Yau。深度神经网络的动力学与神经切线层级。载于国际机器学习会议，第 4542–4551 页。PMLR，2020年。Arthur Jacot、Franck Gabriel 和 Cl´ement 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年。I. Krasikov。贝塞尔函数的一致界。《应用分析杂志》，12(1):83–91，2006年。doi：doi:10.1515/JAA.2006.83。URL https://doi.org/10.1515/JAA.2006.83。Yuanzhi Li、Tengyu Ma 和 Hongyang R Zhang。超越 NTK 的过参数化双层神经网络学习。载于学习理论会议，第 2613–2682 页。PMLR，2020年。梁世宇和 R Srikant。为什么深度神经网络适用于函数逼近？载于第 5 届国际学习表征会议（ICLR 2017），2017年。陆逸平、马超、陆玉龙、陆剑峰和应乐行。深度残差网络及其拓展的平均场分析：通过深度过参数化实现可证明优化。载于国际机器学习会议，第 6426–6436 页。PMLR，2020年。Eran Malach 和 Shai Shalev-Shwartz。只有当浅层网络表现良好时，更深的网络才更好吗？《神经信息处理系统进展》，32，2019年。Eran Malach、Gilad Yehudai、Shai Shalev-Schwartz 和 Ohad Shamir。神经网络中的逼近、深度分离与可学习性之间的联系。载于学习理论会议，第 3265–3295 页。PMLR，2021年。Song Mei、Andrea Montanari 和 Phan-Minh Nguyen。双层神经网络损失景观的平均场视角。《美国国家科学院院刊》，115(33):E7665–E7671，2018年。Phan-Minh Nguyen 和 Huy Tuan Pham。多层神经网络平均场极限的严格框架。arXiv 预印本 arXiv:2001.11443，2020年。Atsushi Nitanda 和 Taiji Suzuki。无限集成的随机粒子梯度下降。arXiv 预印本 arXiv:1712.05438，2017年。Razvan Pascanu、Tomas Mikolov 和 Yoshua Bengio。循环神经网络训练的困难。载于第 30 届国际机器学习会议论文集，第 28 卷，ICML 2013，第 III–1310–III–1318 页。JMLR.org，2013年。Huy Tuan Pham 和 Phan-Minh Nguyen。平均场机制下三层神经网络的全局收敛。载于国际学习表征会议，2021年。URL https://openreview.net/forum?id=KvyxFqZS_D。Tomaso Poggio、Hrushikesh Mhaskar、Lorenzo Rosasco、Brando Miranda 和 Qianli Liao。深层网络何以以及何时能够避免维数灾难，而浅层网络不能：综述。《国际自动化与计算杂志》，14(5):503–519，2017年。Ben Poole、Subhaneil Lahiri、Maithra Raghu、Jascha Sohl-Dickstein 和 Surya Ganguli。瞬态混沌带来的深度神经网络指数级表达能力。《神经信息处理系统进展》，29，2016年。Grant M Rotskoff 和 Eric Vanden-Eijnden。神经网络的可训练性与准确性：一种相互作用粒子系统方法。arXiv 预印本 arXiv:1805.00915，2018年。Itay Safran 和 Jason D Lee。神经网络的基于优化的分离。arXiv 预印本 arXiv:2112.02393，2021年。

Itay Safran 和 Ohad Shamir。《利用神经网络逼近自然函数时的深度-宽度权衡》。载于 Doina Precup 和 Yee Whye Teh（编），第 *34* 届国际机器学习会议论文集，第 70 卷，机器学习研究论文集，第 2979–2987 页。PMLR，2017年8月。网址：https://proceedings.mlr.press/v70/safran17a.html。

Itay Safran、Ronen Eldan 和 Ohad Shamir。《神经网络中的深度分离：究竟分离的是什么？》。载于 Alina Beygelzimer 和 Daniel Hsu（编），第三十二届学习理论会议论文集，第 99 卷，机器学习研究论文集，第 2664–2666 页，美国凤凰城，2019年6月。PMLR。网址 http://proceedings.mlr.press/v99/safran19a.html。

Justin Sirignano 和 Konstantinos Spiliopoulos。神经网络的平均场分析：中心极限定理。《随机过程及其应用》，130(3):1820–1852，2020年。

Justin Sirignano 和 Konstantinos Spiliopoulos。深度神经网络的平均场分析。《运筹学数学》，2021年。

陶哲轩。《非线性色散方程：局部与整体分析》。第106号，数学科学委员会区域会议数学系列。美国数学会，2006年。ISBN 978-0-8218-4143-3。OCLC：ocm65165502。

Matus Telgarsky。深度对神经网络的益处。载于学习理论会议，第1517–1539页。PMLR，2016年。

Gal Vardi 和 Ohad Shamir。具有小权重的神经网络与深度分离障碍。神经信息处理系统进展，33:19433–19442，2020年。

Luca Venturi、Samy Jelassi、Tristan Ozuch 和 Joan Bruna。《超越径向函数的深度分离》。《机器学习研究杂志》，23(122)：1–56，2022年。

Xiang Wang、Chenwei Wu、Jason D Lee、Tengyu Ma 和 Rong Ge。《超越过参数化张量分解中的惰性训练》。*arXiv*预印本 *arXiv:2010.11356*，2020年。

Colin Wei、Jason D Lee、Qiang Liu 和 Tengyu Ma。《正则化很重要：神经网络与其诱导核的泛化和优化》。载于神经信息处理系统进展，第9712–9724页，2019年。

Blake Woodworth、SuriyaGunasekar、Jason D Lee、Edward Moroshko、Pedro Savarese、Itay Golan、Daniel Soudry 和 Nathan Srebro。《过参数化模型中的核机制与丰富机制》。载于学习理论会议，第3635–3673页。PMLR，2020年。

Dmitry Yarotsky。深度 ReLU 神经网络近似的误差界。神经网络，94：103–114，2017年。

Gilad Yehudai 和 Ohad Shamir。随机特征在理解神经网络方面的能力与局限。*arXiv* 预印本 *arXiv:1904.00687*，2019年。

Jingzhao Zhang、Tianxing He、Suvrit Sra 和 Ali Jadbabaie。为什么梯度裁剪能加速训练：对自适应性的理论论证。发表于 国际学习表征会议，2020年。网址https://openreview.net/forum?id=BJgnXpVYwS。

Mo Zhou、Rong Ge 和 Chi Jin。《轻度过参数化双层神经网络的局部收敛理论》。载于学习理论会议，第 4577–4632 页。PMLR，2021年。

多层平均场网络

本节首先简要回顾双层平均场网络的现有理论，然后介绍我们针对多层平均场网络提出的框架。

A.1 双层网络与置换不变性

一个宽度为 m 的两层网络 f 通常可以表示为[6]

$$f(\boldsymbol{x};\boldsymbol{W},\boldsymbol{a})=\frac{1}{m}\boldsymbol{a}^{\top}\sigma(\boldsymbol{W}\boldsymbol{x})=\frac{1}{m}\sum_{i=1}^{m}a_i\sigma(\boldsymbol{w}_i\cdot\boldsymbol{x}). \tag{7}$$

其中，$\boldsymbol{W}\in\mathbb{R}^{m\times d}$ 是隐藏层的权重矩阵，$\boldsymbol{a}\in\mathbb{R}^{m}$ 是输出权重。令 μ 为 $\{(a_i,\ \boldsymbol{w}_i)\}_{i=1}^{m}\subset\mathbb{R}^{d+1}$ 的经验分布。于是，我们可以写成

$$f(\boldsymbol{x};\mu)=\mathop{\mathbb{E}}_{(a,\boldsymbol{w})\sim\mu}\left\{a\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\right\}. \tag{8}$$

允许 μ 成为定义在 $\mathbb{R}^d$ 上任意充分正则的分布，我们得到一个由 (8) 表示的神经网络，其中可以包含无限多个神经元。

要描述这个无限宽网络的梯度流，只需为$\mathbb{R}^{d+1}$ 指定一个向量场，描述每个神经元$(a,\boldsymbol{w})\in\mathbb{R}^{d+1}$ 在时间 t应如何移动。一种简单的启发式方法是，先计算有限宽情形下的梯度，然后按照（8）将所有求和替换为期望，并将梯度视为向量场。现在我们在可实现设定和均方误差损失下说明这一思路

$$\mathcal{L}=\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))^2\right\}.$$

该理论可以推广到更一般的设定，并且可以借助 Wasserstein梯度流理论进行严格论证。详情请参阅例如Chizat 与 Bach (2018年) 以及 Mei 等人 (2018年)。对于有限宽网络（7），$\mathcal{L}$ 关于神经元$(a_k,\boldsymbol{w}_k)$ 的梯度为

$$\begin{aligned}-m\nabla_{a_k}\mathcal{L}&=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x};\boldsymbol{W},\boldsymbol{a}))\sigma(\boldsymbol{w}_k\cdot\boldsymbol{x})\right\},\\-m\nabla_{\boldsymbol{w}_k}\mathcal{L}&=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x};\boldsymbol{W},\boldsymbol{a}))a_k\sigma'(\boldsymbol{w}_k\cdot\boldsymbol{x})\boldsymbol{x}\right\}.\end{aligned}$$

将 $f(\boldsymbol{x};\ \boldsymbol{W},\boldsymbol{a})$ 替换为 $f(\boldsymbol{x};\ \mu)$，将 $(a_k,\boldsymbol{w}_k)$ 视为一个通用神经元，便可得到一个向量场$\tilde{\nabla}:\mathbb{R}^{d+1}\to\mathbb{R}^{d+1}$

$$-\tilde{\nabla}(a,\boldsymbol{w}):=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x};\mu))\begin{bmatrix}\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\\a\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\end{bmatrix}\right\}.$$

˜在每个时间 t，我们根据 $-\nabla$更新 μ 中的神经元。

这种平均场表述最重要的性质之一是**它消除了神经元的置换不变性**。也就是说，我们可以对 $(a_1,\boldsymbol{w}_1)\ldots,(a_m,\boldsymbol{w}_m)$ 进行置换，而不改变网络的输出。然而，当我们将训练视为在 $(\boldsymbol{a},\boldsymbol{W})$ 空间上进行的优化问题，即 $\mathbb{R}^m\times\mathbb{R}^{m\times d}$ 时，对 $(a_i,\boldsymbol{w}_i)$ 进行置换会彻底改变 $(\boldsymbol{a},\boldsymbol{W})$。另一方面，如果我们用一个关于 $\mathbb{R}^{d+1}$ 的分布 μ 来描述网络，那么它天然具有置换不变性。请注意，这并非仅适用于无限宽网络。当我们选择 μ 为有限多个神经元上的经验分布时，就能在不破坏置换不变性的情况下得到有限宽网络。

[6]这里，$\boldsymbol{w}_i\in\mathbb{R}^d$ 表示 $\boldsymbol{W}$的第 i行。之后，我们将使用记号 $\boldsymbol{v}_i$、$\boldsymbol{a}_i$ 表示相应矩阵的第 i行或列。根据维度可以很容易地判断它是行还是列。一般规则是，如果 $\boldsymbol{V}\in\mathbb{R}^{D\times m}$，其中 m 表示神经元的数量，那么 $\boldsymbol{v}_i\in\mathbb{R}^D$ 是第 i列；如果 $\boldsymbol{W}\in\mathbb{R}^{m\times D}$，那么 $\boldsymbol{w}_i\in\mathbb{R}^D$ 是第 i行。

A.2 多层平均场网络

遗憾的是，上述策略无法直接推广到多层网络。考虑如下三层网络

$$f(\boldsymbol{x};\boldsymbol{a},\boldsymbol{W}_2,\boldsymbol{W}_1)=\frac{1}{m_2}\boldsymbol{a}^\top\sigma\left(\boldsymbol{W}_2\boldsymbol{h}(\boldsymbol{x};\boldsymbol{W}_1)\right),\quad \boldsymbol{h}(\boldsymbol{x};\boldsymbol{W}_1)=\frac{1}{m_1}\sigma(\boldsymbol{W}_1\boldsymbol{x}),$$

其中 $\boldsymbol{a}\in\mathbb{R}^{m_2}$, $\boldsymbol{W}_2\in\mathbb{R}^{m_2\times m_1}$, $\boldsymbol{W}_1\in\mathbb{R}^{m_1\times d}$。仍然可以写成

$$f(\boldsymbol{x};\boldsymbol{a},\boldsymbol{W}_2,\boldsymbol{W}_1)=\frac{1}{m_2}\sum_{i=1}^{m_2}a_i\sigma(\boldsymbol{w}_{2,i}\cdot\boldsymbol{h}(\boldsymbol{x};\boldsymbol{W}_1))=\mathop{\mathbb{E}}_{(a_i,\boldsymbol{w}_2)\sim\mu_2}\left\{a\sigma(\boldsymbol{w}_2\cdot\boldsymbol{h}(\boldsymbol{x};\boldsymbol{W}_1))\right\}.$$

然而，现在 μ_2 是一个 $\mathbb{R}^{m_1}$上的分布；如果 $m_1\to\infty$，它就会变成一个$\mathbb{R}^\infty$上的分布，而这并没有明确的定义。一种解决这一问题的方法是将 $\boldsymbol{W}_2$ 视为从$[m_2]\times[m_1]$到 $\mathbb{R}$ 的函数，然后通过将指标集$[m_2]$，$[m_1]$ 替换为两个可能不可数的一般指标集 I_2，I_1 来推广它，以处理无限宽情形。例如，我们可以选择 $I_1=I_2=\mathbb{R}$。Nguyen & Pham (2020年)采用的正是这一策略。（参见 Pham& Nguyen (2021年)，这篇论文的版本更易于理解。）这种形式的缺点在于，引入指标集后，置换不变性不再被消去。尽管采用这种形式仍可得到无限宽网络的全局收敛结果，但在分析有限宽网络时，它就不那么有用了，因为它本质上与常见的矩阵形式无异。

现在，我们提出一种能够分离神经元置换不变性的表述，它通过复合一系列向量值两层网络构建。首先，我们考虑一个输出维度为 D的两层网络：

$$\boldsymbol{f}(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W})=\frac{1}{m}\boldsymbol{A}\sigma(\boldsymbol{W}\boldsymbol{x}),\tag{9}$$

其中 $\boldsymbol{A}\in\mathbb{R}^{D\times m}$ 且 $\boldsymbol{W}\in\mathbb{R}^{m\times d}$。对于每个索引 $i\in[D]$，我们仍有

$$f_i(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W})=\frac{1}{m}\sum_{j=1}^{m}a_{i,j}\sigma(\boldsymbol{w}_j\cdot\boldsymbol{x})=\mathop{\mathbb{E}}_{(a,\boldsymbol{w})\sim\mu_i}\left\{a\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\right\},$$

其中 μ_i 是 $\{(a_{i,\,j},\ \boldsymbol{w}_j)\}_{j\in[m]}\subset\mathbb{R}^{d+1}$的经验分布。遍历 i 即可得到该网络的输出向量。对于标量输出的两层网络，要得到其平均场对应形式，只需允许 μ 取关于 $\mathbb{R}\times\mathbb{R}^d$的一般分布。然而，对于向量输出的网络，情况并非如此，因为 μ_i 的 $\boldsymbol{W}$ 部分彼此耦合。因此，我们还需要施加约束：所有 $(\mu_i)^{i\in[D]}$ 具有相同的第二边缘分布，即对于某个关于 $\mathbb{R}^d$ 的分布 $\mu_{\boldsymbol{W}}$ 以及所有 $\in[D]$，均有 $\pi_2\#\mu_i=\mu_{\boldsymbol{W}}$，其中 $\pi_2\colon\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}^d$ 是将 $(a,\ \boldsymbol{w})$ 投影为 $\boldsymbol{w}$的投影。直观地说，这一条件表示它们共享相同的第一层权重 $\boldsymbol{W}$。我们将在下述定义中将这一想法形式化。

定义 A.1. 设 $(\mu_i)^{i\in[D]}$ 是 D 足够正则的[7] 分布，定义在 $\mathbb{R}\times\mathbb{R}^d$上。若存在一个定义在 $\mathbb{R}^d$ 上的测度 $\mu_{\boldsymbol{W}}$，使得$\pi_2\#\mu_i=\mu_{\boldsymbol{W}}$ 对所有 $i\in[D]$都成立，则称 $(\mu_i)_{i=1}^D$为一个**维度为** (D,d) 的可容许配置。

注。 请注意，这里所说的神经元是一个 $(D+d)$维向量 $(a_1,\ldots,a_D,\boldsymbol{w})$。在有限宽网络（9）中，这对应于 $\boldsymbol{W}$ 中的一行以及 $\boldsymbol{A}$中对应的一列。推导无限宽度梯度流时，这一视角十分重要，因为与两层情形一样，某个神经元所在位置的向量场只能依赖于其余神经元的整体。♣

为补充上述讨论，我们在此考虑这样一个问题：给定一个可容许的无限宽度配置 $(\mu_i)^{i\in[D]}$，如何获得一个包含 m 个神经元的有限宽网络。对于标量值

[7]我们的重点是剥离置换不变性；在本文中，几乎所有分布本质上都是定义在有限多个神经元上的经验分布，对于这类分布，积分就是求和，并且总是良定义的。具体正则性条件的确定留待未来工作。

平均场网络，其特征由 μ 描述，只需从 μ 中生成 m 个样本即可。对于向量值网络，步骤略有不同。我们首先从共享边缘分布 $\mu_{\boldsymbol{W}}$ 中采样一个权重向量 $\boldsymbol{w}$。然后，对每个 $i \in [D]$，在给定 $\boldsymbol{w}$ 的条件下生成一个实数 a_i。由此得到一个神经元 $(a_1, \ldots, a_D, \boldsymbol{w}) \in \mathbb{R}^D \times \mathbb{R}^d$。重复此过程 m 次，即可得到一个包含 m 个神经元的有限宽网络。

我们正式定义两层向量值平均场网络如下。

定义 A.2. 给定一个可容许$(\mu_i)^{i\in[D]}$，它所定义的两层向量值网络为

$$\boldsymbol{F}(\boldsymbol{x}; \mu_1, \ldots, \mu_D) = (F_1(\boldsymbol{x}; \mu_1), \ldots, F_D(\boldsymbol{x}; \mu_D)), \tag{10}$$

其中

$$F_i(\boldsymbol{x}; \mu_i) = \mathop{\mathbb{E}}_{(a,\boldsymbol{w})\sim\mu_i} \{a\sigma(\boldsymbol{w} \cdot \boldsymbol{x})\}, \quad \forall i \in [D].$$

现在，我们可以定义多层平均场网络了。基本上，多层平均场网络是由一系列两层向量值网络（10）复合而成的。

定义 A.3. 设 $L \geq 1$ 为一个整数。设 $D^{(1)}, \ldots, D^{(L)}$ 为一列正整数，并令 $D^{(0)} = d$。对于每个 $l \in [L]$，令 $(\mu_i^{(l)})^{i\in[D_l]}$ 为维度为 $(D^{(l)}, D^{(l-1)})$ 的可容许配置。由配置$\Theta := ((\mu_i^{(l)})^{i\in[D_l]})^{l\in[L]}$ 定义的 L 层平均场网络 $\boldsymbol{f}$ 递归定义为

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}; \Theta) &= \boldsymbol{F}^{(L)}(\boldsymbol{x}; \Theta), \\ \boldsymbol{F}^{(l)}(\boldsymbol{x}; \Theta) &:= \boldsymbol{F}\left(\boldsymbol{F}^{(l-1)}(\boldsymbol{x}; \Theta); \mu_1^{(l)}, \ldots, \mu_{D_l}^{(l)}\right), \quad \forall l \geq 1, \\ \boldsymbol{F}^{(0)}(\boldsymbol{x}; \Theta) &:= \boldsymbol{x}, \end{aligned} \tag{11}$$

其中 $\boldsymbol{F}$ 是由 （*10*）给出的双层平均场网络。

示例 例如，我们考虑这里的情形 $L = 3$。在这种情况下，与（11）对应的有限宽网络为

$$\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{A}_2, \boldsymbol{W}_2, \boldsymbol{A}_1, \boldsymbol{W}_1) = \frac{1}{m_2}\boldsymbol{A}_2\sigma\left(\boldsymbol{W}_2\frac{1}{m_1}\boldsymbol{A}_1\sigma(\boldsymbol{W}_1\boldsymbol{x})\right),$$

这与实践中使用的常见多层网络完全相同，除了归一化项 $1/m_2$、$1/m_1$ 以及一个额外的矩阵 $\boldsymbol{A}_1 \in \mathbb{R}^{D_1\times m_1}$。该矩阵将一个 m_1 维特征向量压缩为一个 D_1 维向量，其中 D_1 是一个不会趋于 ∞的整数。这让人想起残差网络中使用的瓶颈结构（He 等人（2016年））。

注。 请注意，这种表述确实在每一层神经元的置换下保持不变。不过，它并未消除深度网络的所有置换不变性。例如，可以置换 $\boldsymbol{W}_1$ 的列，并相应地调整 $\boldsymbol{A}_1$、$\boldsymbol{W}_2$、$\boldsymbol{A}_2$，而不改变网络的输出。从某种意义上说，这相当于置换隐藏特征 $\boldsymbol{F}^{(1)}$ 的元素。我们认为没有必要也没有益处消除这种对称性，因为毕竟，即使在双层情形下，我们也不会置换输入 $\boldsymbol{x}$ 的元素。♣

最后，我们考虑如何构造平均场梯度流，使其与通常的梯度流相匹配。思路很简单：我们先在有限宽度设定下计算梯度，然后用积分替换求和。为便于说明，我们考虑一个三层网络和均方误差损失。同样，这一框架可以轻松推广到更深的网络和其他损失函数。我们记

$$\begin{aligned} f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}_2, \boldsymbol{V}_1, \boldsymbol{W}_1) &= \frac{1}{m_2}\boldsymbol{a}^\top\sigma\left(\boldsymbol{W}_2\boldsymbol{F}(\boldsymbol{x}; \boldsymbol{V}_1, \boldsymbol{W}_1)\right), \\ \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x}; \boldsymbol{V}_1, \boldsymbol{W}_1) &= \frac{1}{m_1}\boldsymbol{V}_1\sigma(\boldsymbol{W}_1\boldsymbol{x}), \\ \mathcal{L} = \mathcal{L}(\boldsymbol{a}, \boldsymbol{W}_2, \boldsymbol{V}, \boldsymbol{W}_1) &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}_2, \boldsymbol{V}, \boldsymbol{W}_1))^2\right\}, \end{aligned}$$

其中 $\boldsymbol{a} \in \mathbb{R}^{m_2}, \boldsymbol{W}_2 \in \mathbb{R}^{m_2 \times D}, \boldsymbol{V}_1 \in \mathbb{R}^{D \times m_1}, \boldsymbol{W}_1 \in \mathbb{R}^{m_1 \times d}$。我们有

$$
\begin{aligned}
-m_2 \nabla_{a_i} \mathcal{L} &= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma(\boldsymbol{w}_{2,i} \cdot \boldsymbol{F}(\boldsymbol{x})) \right\}, && \forall i \in [m_2], \\
-m_2 \nabla_{\boldsymbol{w}_{2,i}} \mathcal{L} &= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) a_i \sigma'(\boldsymbol{w}_{2,i} \cdot \boldsymbol{F}(\boldsymbol{x})) \boldsymbol{F}(\boldsymbol{x}) \right\}, && \forall i \in [m_2], \\
-m_1 \nabla_{\boldsymbol{v}_{1,i}} \mathcal{L} &= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \frac{1}{m_2} \sum_{j=1}^{m_2} a_j \sigma'(\boldsymbol{w}_{2,j} \cdot \boldsymbol{F}(\boldsymbol{x})) \boldsymbol{w}_{2,j} \sigma(\boldsymbol{w}_{1,i} \cdot \boldsymbol{x}) \right\}, && \forall i \in [m_1], \\
-m_1 \nabla_{\boldsymbol{w}_{1,i}} \mathcal{L} &= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \frac{1}{m_2} \sum_{j=1}^{m_2} a_j \sigma'(\boldsymbol{w}_{2,j} \cdot \boldsymbol{F}(\boldsymbol{x})) \left\langle \boldsymbol{w}_{2,j}, \boldsymbol{v}_{1,i} \right\rangle \sigma'(\boldsymbol{w}_{1,i} \cdot \boldsymbol{x}) \boldsymbol{x} \right\}, && \forall i \in [m_1].
\end{aligned}
$$

将求和替换为积分，我们得到

$$
\begin{aligned}
-\tilde{\nabla}_{(a, \boldsymbol{w}_2)} &= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \begin{bmatrix} \sigma(\boldsymbol{w}_2 \cdot \boldsymbol{F}(\boldsymbol{x})) \\ a \sigma'(\boldsymbol{w}_2 \cdot \boldsymbol{F}(\boldsymbol{x})) \boldsymbol{F}(\boldsymbol{x}) \end{bmatrix} \right\}, \\
-\tilde{\nabla}_{(\boldsymbol{v}_1, \boldsymbol{w}_1)} &= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \mathop{\mathbb{E}}_{(a, \boldsymbol{w}_2) \sim \mu_2} \left\{ a \sigma'(\boldsymbol{w}_2 \cdot \boldsymbol{F}(\boldsymbol{x})) \begin{bmatrix} \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \boldsymbol{w}_2 \\ \langle \boldsymbol{w}_2, \boldsymbol{v}_1 \rangle \sigma'(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \boldsymbol{x} \end{bmatrix} \right\} \right\}. \quad (12)
\end{aligned}
$$

也就是说，在每一步 t，我们都会更新第二层神经元$(a, \boldsymbol{w}_2)$，使其达到$-\tilde{\nabla}_{(a, \boldsymbol{w}_2)}$；并更新第一层神经元$(\boldsymbol{v}_1, \boldsymbol{w}_1)$，使其达到$-\tilde{\nabla}_{(\boldsymbol{v}_1, \boldsymbol{w}_1)}$。请注意，与许多其他多层平均场框架不同，我们并未引入路径的概念。由于我们对 μ_2 在式（12）中取期望，每个第一层神经元的动力学都依赖于整体的第二层。第二层神经元也是如此。从某种意义上说，额外的矩阵 V_1 将第一层和第二层神经元的动力学解耦。

B 预备知识

B.1 归纳假设与连续性论证

在证明中，我们广泛使用数学归纳法的连续时间版本，也称为连续性论证。本小节将简要讨论这一技巧，并说明我们在撰写证明时采用的一些约定。例如，可参阅Tao (2006年)了解详情。

与离散时间归纳论证类似，目标是在一段时间内维持一组条件，我们称之为归纳假设（参见归纳假设 C.2和归纳假设 D.1）。条件主要有两类。

第一类形式为“某个过程 A_t 受另一个过程 B_t的上界约束”。在证明中，A_t 通常是我们要控制的误差，而 B_t 是表示相应上界的非递减过程。要维持这类条件，只需证明在初始化时 $A_t \leq B_t$ 成立，并且只要归纳假设成立，A_t 就成立。

对于这类条件，通常我们也有 B_t的一个上界，例如 $B_t \leq B_\infty$。维持这些界限最严谨的方法是采用反证法。令 T 为过程结束时刻 T_1 与该上界首次被突破时刻 T_2 中的较早者。根据定义，对于任意 $t \leq T$，归纳假设都成立。利用归纳假设，可以推导出 T_1的一个上界 T'，进而得到 T的一个上界。接下来，只需证明 $B_{T'}$ 小于 B_∞，这样 T 便由 T_1 而非 T_2达到。为便于表述，对于这类条件，我们不再显式采用反证法，而只需证明：若归纳假设在 $[0, T_1]$ 上成立，则 $B_{T_1} \leq B_\infty$ 成立。

第二种类型的形式为“某个过程 C_t 被某个值 D所界定”。这里，C_t 通常是与学习函数形状相关的某个量，例如 $\bar{w}_2$ 和 α。为了保持，比如说 $C_t \leq D$，我们证明当 $C_t \in [D - \varepsilon, D]$ 时，有 $C_t < 0$。这意味着，只要C_t 是连续的，C_t 就永远不会达到 D。

B.2 输入分布的性质

在本小节中，我们推导输入分布的一些基本性质，这些性质将在后续分析中发挥作用。

下面的引理给出了 $\|\boldsymbol{x}\|$ 的分布及其尾界。

引理 B.1. 设 $\boldsymbol{x} \sim \mathcal{D}$，并令 $\|\mathcal{D}\|$ 表示 $\|\boldsymbol{x}\|$ 的分布。我们有

$$\|\mathcal{D}\|(r) = \frac{d}{r} J_{d/2}^2(2\pi R_d \beta\alpha\sqrt{d}r) = O\left(\frac{1}{r^2}\right), \quad \forall r > 0.$$

因此，我们得到如下尾界：对于所有 $R > 0$，$\mathbb{P}[\|\boldsymbol{x}\| \geq R] \leq O(1/R)$。

现在我们给出输入分布的一些正则性条件，这些条件将在证明中使用。粗略地说，这表明该分布具有重尾特征，同时对于$\|\boldsymbol{x}\| \in [0, 1]$仍有足够大的质量

引理 B.2 (输入分布的正则性条件)。 对于输入分布 $\mathcal{D}$，我们有

(a) $\mathbb{E}_{\|\boldsymbol{x}\| \leq 0.99} \|\boldsymbol{x}\| = \Theta(1)$。 *(b)* $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} f_*(\boldsymbol{x}) = \Omega(1)$。 *(c)* $\mathbb{E}_{\|\boldsymbol{x}\| \leq \Omega(d)} \|\boldsymbol{x}\| \geq \Theta(\log d)$ 且 $\mathbb{E}_{\|\boldsymbol{x}\| \leq \mathrm{poly}(d)} \|\boldsymbol{x}\| \leq \Theta(\log(d))$。

引理 *B.1* 的证明。 回顾一下，$\boldsymbol{x}$ 的输入分布为

$$\left(\beta\alpha\sqrt{d}\right)^d \varphi^2(\beta\alpha\sqrt{d}\boldsymbol{x}),$$

where α，$\beta > 0$ 是来自 Safran 等人 (2019年)（参见定理 5 的证明），

$$\varphi(\boldsymbol{x}) = \left(\frac{R_d}{\|\boldsymbol{x}\|}\right)^{d/2} J_{d/2}(2\pi R_d \|\boldsymbol{x}\|), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

$R_d = \frac{1}{\sqrt{\pi}}(\Gamma(d/2+1))^{1/d} = \Theta(\sqrt{d})$ (引理 5，出自 Eldan 和 Shamir (2016年))，且 J_ν 是第一类贝塞尔函数，阶数为。注意，由于 φ 仅依赖于 $\|\boldsymbol{x}\|$，我们可以滥用记号，用 $\varphi(r)$ 表示 $\varphi(\boldsymbol{x})$，其中 $\|\boldsymbol{x}\| = r$。

对于任意测试函数 g： $\mathbb{R} \mapsto \mathbb{R}$，我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[g(\|\boldsymbol{x}\|)] &= \int_{\mathbb{R}^d} g(\|\boldsymbol{x}\|) \left(\beta\alpha\sqrt{d}\right)^d \varphi^2(\beta\alpha\sqrt{d}\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \left(\beta\alpha\sqrt{d}\right)^d S_{d-1} \int_0^\infty g(r)\varphi^2(\beta\alpha\sqrt{d}r) r^{d-1}\mathrm{d}r,\end{aligned}$$

其中 $S_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$是单位球面 $\mathbb{S}^{d-1}$。因此，我们得到$\|\boldsymbol{x}\|$ 在 $\|\boldsymbol{x}\| = r$ 下的密度为

$$\begin{aligned}\left(\beta\alpha\sqrt{d}\right)^d S_{d-1}\varphi^2(\beta\alpha\sqrt{d}r) r^{d-1} =& \frac{2\pi^{d/2}\left(\beta\alpha\sqrt{d}\right)^d}{\Gamma(d/2)} \frac{R_d^d}{\left(\beta\alpha\sqrt{d}r\right)^d} J_{d/2}^2(2\pi R_d\beta\alpha\sqrt{d}r) r^{d-1} \\ =& \frac{d}{r} J_{d/2}^2(2\pi R_d\beta\alpha\sqrt{d}r) \\ =& O\left(\frac{1}{r^2}\right),\end{aligned}$$

其中我们利用了以下事实： $J_\nu(z) = O(1/\sqrt{z})$ (Krasikov (2006年))。接着，很容易看出$\mathbb{P}(\|\boldsymbol{x}\| \geq R) = O(1/R)$。

□

引理的证明*B.2*。

(a) 很容易看出上界

$$\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 0.99}\|\boldsymbol{x}\| \le 0.99.$$

对于下界，注意到 $\mathbb{E}_{\|\boldsymbol{x}\|\le 0.99}\|\boldsymbol{x}\| \ge 0.1\,\mathbb{P}(0.1 \le \|\boldsymbol{x}\| \le 0.99)$。因此，只需对 $\mathbb{P}(0.1 \le \|\boldsymbol{x}\| \le 0.99)$求下界。我们有

$$\begin{aligned}\mathbb{P}(0.1 \le \|\boldsymbol{x}\| \le 0.99) =& \int_{0.1}^{0.99} \frac{d}{r} J_{d/2}^2(2\pi R_d \beta\alpha\sqrt{d}r)\mathrm{d}r \\ \ge& \Omega(1)\int_{0.2\pi R_d\beta\alpha\sqrt{d}}^{1.98\pi R_d\beta\alpha\sqrt{d}} J_{d/2}^2(r)\mathrm{d}r \\ =& \Omega(1),\end{aligned}$$

其中，最后一行用到了 Eldan & Shamir (2016年 年的引理 23。这意味着 $\mathbb{E}_{\|\boldsymbol{x}\|\le 0.99}\|\boldsymbol{x}\| = \Omega(1)$。结合上界，可得 $\mathbb{E}_{\|\boldsymbol{x}\|\le 0.99}\|\boldsymbol{x}\| = \Theta(1)$。

(b) 我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}} f_*(\boldsymbol{x}) =& \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}[1-\|\boldsymbol{x}\|] \ge \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 0.99}[1-\|\boldsymbol{x}\|] \\ \ge& 0.01\,\mathbb{P}(\|\boldsymbol{x}\| \le 0.99) \ge 0.01\,\mathbb{P}(0.1 \le \|\boldsymbol{x}\| \le 0.99) = \Omega(1),\end{aligned}$$

其中，最后一个不等式用到了 (a) 中的计算。

(c) 上界可直接由尾界 $\|\mathcal{D}\|(r) \le O(1/r^2)$ 得出。对于下界，回忆一下当 $\|\boldsymbol{x}\| = r$ 为 $\frac{d}{r}J_{d/2}^2(2\pi R_d\beta\alpha\sqrt{d}r)$ 时，$\|\boldsymbol{x}\|$ 的密度。为简化记号，令 $R_{\mathcal{D}} = \Theta(d)$。我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le R_{\mathcal{D}}}\|\boldsymbol{x}\| =& \int_0^{R_{\mathcal{D}}} dJ_{d/2}^2(2\pi R_d\beta\alpha\sqrt{d}r)\mathrm{d}r \\ =& \frac{d}{2\pi R_d\beta\alpha\sqrt{d}}\int_0^{2\pi R_d R_{\mathcal{D}}\beta\alpha\sqrt{d}} J_{d/2}^2(r)\mathrm{d}r \\ \ge& \Omega(1)\int_{cd}^{cd^2} J_{d/2}^2(r)\mathrm{d}r,\end{aligned}$$

其中 c 是一个足够大的常数。

要得到下界 $\mathbb{E}\|\boldsymbol{x}\|$，只需对 $\int_{cd}^{cd^2} J_{d/2}^2(r)\mathrm{d}r$ 求下界。接下来，我们将参照 Eldan& Shamir(2016年) 中引理 23 的证明，采用类似的计算来求其下界。根据 Eldan & Shamir (2016年) 中引理 23 的证明，对于 $x \ge d \ge 2$

$$J_{d/2}^2(x) \ge \frac{2}{\pi x}\cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x}x\right) - 3x^{-2},$$

其中 $f_{d,\,x}$ 是一个依赖于 d 和 x 的量，并满足 $1.3 \ge f_{d,\,x} \ge 0.85$。

于是，我们有

$$\begin{aligned}\int_{cd}^{cd^2} J_{d/2}^2(x)\mathrm{d}x \ge& \int_{cd}^{cd^2}\frac{2}{\pi x}\cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x}x\right)\mathrm{d}x - \int_{cd}^{cd^2} 3x^{-2}\mathrm{d}x \\ =& \frac{2}{\pi}\int_{cd}^{cd^2}\frac{1}{x}\cos^2\left(-\frac{(d+1)\pi}{4} + f_{d,x}x\right)\mathrm{d}x - \frac{3(d-1)}{cd^2}\end{aligned}$$

注意，在Eldan 和 Shamir（2016年）中引理 23 的证明表明

$$\frac{\partial}{\partial x}(f_{d,x}x) = \sqrt{1-\frac{d^2-1}{4x^2}} \le 1.$$

于是，由于 $1.3 \geq f_{d,\ x} \geq 0.85$我们有

$$\begin{aligned}
&\frac{2}{\pi}\int_{cd}^{cd^2}\frac{1}{x}\cos^2\left(-\frac{(d+1)\pi}{4}+f_{d,x}x\right)\mathrm{d}x\\
\geq&\frac{2}{\pi}\int_{cd}^{cd^2}\frac{0.85}{f_{d,x}x}\cos^2\left(-\frac{(d+1)\pi}{4}+f_{d,x}x\right)\frac{\partial}{\partial x}(f_{d,x}x)\mathrm{d}x\\
=&\frac{2}{\pi}\int_{f_{d,cd^2}cd}^{f_{d,cd^2}cd^2}\frac{0.85}{z}\cos^2\left(-\frac{(d+1)\pi}{4}+z\right)\mathrm{d}z\\
\geq&\frac{1.7}{\pi}\int_{1.3cd}^{0.85cd^2}\frac{1}{z}\cos^2\left(-\frac{(d+1)\pi}{4}+z\right)\mathrm{d}z.
\end{aligned}$$

然后，利用分部积分以及 $\cos^2(z-(d+1)\pi/4) = \frac{\partial}{\partial z}(z/2+\sin(2z-(d+1)\pi/2)/4)$ 这一事实，我们得到

$$\begin{aligned}
&\int_{1.3cd}^{0.85cd^2}\frac{1}{z}\cos^2\left(-\frac{(d+1)\pi}{4}+z\right)\mathrm{d}z\\
=&\frac{(\frac{z}{2}+\frac{1}{4}\sin(2z-\frac{(d+1)\pi}{2}))}{z}\Big|_{1.3cd}^{0.85cd^2}+\int_{1.3cd}^{0.85cd^2}\frac{(\frac{z}{2}+\frac{1}{4}\sin(2z-\frac{(d+1)\pi}{2}))}{z^2}\mathrm{d}z\\
\geq&-\frac{1}{4}\left(\frac{1}{0.85cd^2}+\frac{1}{1.3cd}\right)+\int_{1.3cd}^{0.85cd^2}\frac{1}{4z}\mathrm{d}z\\
=&-\frac{1}{4}\left(\frac{1}{0.85cd^2}+\frac{1}{1.3cd}\right)+\frac{1}{4}\ln\frac{0.85cd^2}{1.3cd}=\Omega(\log d).
\end{aligned}$$

因此，我们得到

$$\int_{cd}^{cd^2}J_{d/2}^2(x)\mathrm{d}x=\Omega(\log d),$$

这意味着 $\mathbb{E}_{\|\boldsymbol{x}\|\leq\Theta(d)}\|\boldsymbol{x}\|=\Omega(\log d)$。

□

B.3 球对称函数和分布的性质

在本小节中，我们给出一些关于球对称函数和分布的有用性质。这些性质将成为我们后续分析中的有用工具。基本而言，这些引理使我们在考虑对球对称函数进行积分时，能够将输入 $\boldsymbol{x}$ 和神经元 $\boldsymbol{v}$ 分离开来。

引理 3.1. 设 μ是一个球对称分布。我们有

$$\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu}\|\boldsymbol{w}\|\sigma(\boldsymbol{w}\cdot\boldsymbol{x})=C_\Gamma\frac{\mathbb{E}_{\boldsymbol{w}\sim\mu}\|\boldsymbol{w}\|^2}{\sqrt{d}}\|\boldsymbol{x}\|\quad \textit{where}\quad C_\Gamma:=\frac{\Gamma(d/2)\sqrt{d}}{2\sqrt{\pi}\Gamma((d+1)/2)}.$$

注意，由于 $d\to\infty$，我们有 $C_\Gamma\to 1/\sqrt{2\pi}$，因此对于所有 d，C_Γ 都一致有界。

引理 B.3. 对于任意球对称的 g： $\mathbb{R}^d\to\mathbb{R}$ 和 $\boldsymbol{v}\in\mathbb{R}^d$，我们有

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})\sigma(\boldsymbol{v}\cdot\boldsymbol{x})\}=\frac{C_\Gamma}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}\|\boldsymbol{v}\|.$$

推论 B.4. 设 g： $\mathbb{R}^d\to\mathbb{R}$ 是一个球对称函数。我们有

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})F(\boldsymbol{x})\}=\alpha\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}.$$

引理 B.5. 设 $g\colon\mathbb{R}^d\to\mathbb{R}$ 是一个球对称函数。那么，对于任意 $\boldsymbol{v}\in\mathbb{R}^d$，我们有

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x}\}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}\frac{C_\Gamma}{\sqrt{d}}\bar{\boldsymbol{v}}.$$

引理 3.1 的证明。为简便起见，令 $g(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{w}\sim\mu} \|\boldsymbol{w}\| \sigma(\boldsymbol{w}\cdot\boldsymbol{x})$。由于 σ 是 1-齐次的，且 μ是球对称的，因此我们有

$$\begin{aligned} g(\boldsymbol{x}) &= \int_{\mathbb{R}^d} \|\boldsymbol{w}\|^2 \sigma(\bar{\boldsymbol{w}}\cdot\boldsymbol{x})\mu(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \\ &= \int_0^\infty \int_{\mathbb{S}^{d-1}} r^2\sigma(\bar{\boldsymbol{w}}\cdot\boldsymbol{x})\mu(r\bar{\boldsymbol{w}})r^{d-1}\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}})\mathrm{d}r \\ &= \int_0^\infty r^{d+1}\mu(r)\,\mathrm{d}r \int_{\mathbb{S}^{d-1}} \sigma(\bar{\boldsymbol{w}}\cdot\boldsymbol{x})\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}). \end{aligned}$$

对于第一项，注意到[8]

$$\int_{\mathbb{R}^d} \|\boldsymbol{w}\|^2 \mu(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} = \int_0^\infty \int_{\mathbb{S}^{d-1}} r^2\mu(r\bar{\boldsymbol{w}})\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}})\mathrm{d}r = \frac{2\pi^{d/2}}{\Gamma(d/2)}\int_0^\infty r^{d+1}\mu(r)\,\mathrm{d}r.$$

因此，

$$\int_0^\infty r^{d+1}\mu(r)\,\mathrm{d}r = \frac{\Gamma(d/2)}{2\pi^{d/2}}\int_{\mathbb{R}^d}\|\boldsymbol{w}\|^2\mu(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} = \frac{\Gamma(d/2)}{2\pi^{d/2}}\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu}\|\boldsymbol{w}\|^2.$$

接下来按如下方式计算第二项。由于它也是球对称的，我们有

$$\int_{\mathbb{S}^{d-1}} \sigma(\bar{\boldsymbol{w}}\cdot\boldsymbol{x})\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}) = \|\boldsymbol{x}\|\int_{\mathbb{S}^{d-1}}\sigma(\bar{w}_1)\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}) = \frac{\|\boldsymbol{x}\|}{2}\int_{\mathbb{S}^{d-1}}|\bar{w}_1|\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}).$$

定义 $I = \int_{\mathbb{R}^d}|w_1|e^{-\|\boldsymbol{w}\|^2}\,\mathrm{d}\boldsymbol{w}$。我们有

$$I = \int_{\mathbb{R}^d}|w_1|\prod_{i=1}^d e^{-w_i^2}\,\mathrm{d}\boldsymbol{w} = \left(\int_{-\infty}^\infty |w_1|e^{-w_1^2}\,\mathrm{d}w_1\right)\prod_{i=2}^d\int_{-\infty}^\infty e^{-w_i^2}\,\mathrm{d}w_i = \pi^{(d-1)/2}.$$

我们还有

$$\begin{aligned} I = \int_{\mathbb{S}^{d-1}}\int_0^\infty r|\bar{w}_1|e^{-r^2}r^{d-1}\,\mathrm{d}r\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}) &= \int_0^\infty e^{-r^2}r^d\,\mathrm{d}r\int_{\mathbb{S}^{d-1}}|\bar{w}_1|\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}) \\ &= \frac{\Gamma((d+1)/2)}{2}\int_{\mathbb{S}^{d-1}}|\bar{w}_1|\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}). \end{aligned}$$

因此，

$$\int_{\mathbb{S}^{d-1}}\sigma(\bar{\boldsymbol{w}}\cdot\boldsymbol{x})\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}) = \frac{\|\boldsymbol{x}\|}{2}\int_{\mathbb{S}^{d-1}}|\bar{w}_1|\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{w}}) = \frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)}\|\boldsymbol{x}\|. \tag{13}$$

于是，

$$g(\boldsymbol{x}) = \frac{\Gamma(d/2)}{2\pi^{d/2}}\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu}\|\boldsymbol{w}\|^2\frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)}\|\boldsymbol{x}\| = C_\Gamma\frac{\mathbb{E}_{\boldsymbol{w}\sim\mu}\|\boldsymbol{w}\|^2}{\sqrt{d}}\|\boldsymbol{x}\|.$$

□

引理的证明*B.3*. 我们计算

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma(\boldsymbol{v}\cdot\boldsymbol{x})\} &= \int_{\mathbb{R}^d} g(\boldsymbol{x})\sigma(\boldsymbol{v}\cdot\boldsymbol{x})\mathcal{D}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \int_0^\infty\int_{\mathbb{S}^{d-1}} g(r\bar{\boldsymbol{x}})\sigma(\boldsymbol{v}\cdot(r\bar{\boldsymbol{x}}))\mathcal{D}(r\bar{\boldsymbol{x}})r^{d-1}\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{x}})\mathrm{d}r \\ &= \int_0^\infty\int_{\mathbb{S}^{d-1}} g(r)\sigma(\boldsymbol{v}\cdot\bar{\boldsymbol{x}})\mathcal{D}(r)r^d\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{x}})\mathrm{d}r \\ &= \int_0^\infty g(r)\mathcal{D}(r)r^d\,\mathrm{d}r\int_{\mathbb{S}^{d-1}}\sigma(\boldsymbol{v}\cdot\bar{\boldsymbol{x}})\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{x}}) \\ &= \int_0^\infty g(r)\mathcal{D}(r)r^d\,\mathrm{d}r\frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)}\|\boldsymbol{v}\|, \end{aligned}$$

[8]回顾一下，d维单位球面的表面积为 $\int\mathrm{d}\sigma^{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$。

其中最后一行由（13）得出。（注意，此处积分是关于 $\bar{\boldsymbol{x}}$ 而不是关于 $\bar{\boldsymbol{w}}$ 进行的。对于第一项，注意

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\} &= \int_{\mathbb{R}^d} g(\boldsymbol{x})\|\boldsymbol{x}\|\mathcal{D}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= \int_0^\infty\int_{\mathbb{S}^{d-1}} g(r)\mathcal{D}(\boldsymbol{x})r^d\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{x}})\mathrm{d}r\\ &= \int_0^\infty\int_{\mathbb{S}^{d-1}} g(r)\mathcal{D}(\boldsymbol{x})r^d\,\mathrm{d}\sigma^{d-1}(\bar{\boldsymbol{x}})\mathrm{d}r\\ &= \frac{2\pi^{d/2}}{\Gamma(d/2)}\int_0^\infty g(r)\mathcal{D}(\boldsymbol{x})r^d\,\mathrm{d}r.\end{aligned}$$

因此，

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma(\boldsymbol{v}\cdot\boldsymbol{x})\} &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}\left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^{-1}\frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)}\|\boldsymbol{w}\|\\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}\frac{C_\Gamma}{\sqrt{d}}\|\boldsymbol{v}\|.\end{aligned}$$

□

推论的证明 *B.4*。 根据前面的引理，我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})F(\boldsymbol{x})\} &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{g(\boldsymbol{x})\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu_1}\{\|\boldsymbol{w}\|\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\}\right\}\\ &= \mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu_1}\left\{\|\boldsymbol{w}\|\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\}\right\}\\ &= \mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mu_1}\left\{\|\boldsymbol{w}\|^2\frac{C_\Gamma}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}\right\} = \alpha\mathop{\mathbb{E}}_{\boldsymbol{x}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}.\end{aligned}$$

□

引理的证明 *B.5*。 定义 $\boldsymbol{R} = \bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^\top - (\boldsymbol{I}_d - \bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^\top) = 2\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^\top - \boldsymbol{I}_d$。也就是说， $\boldsymbol{R}$ 是与 $\bar{\boldsymbol{v}}$相关联的反射矩阵。由于 $\mathcal{D}$ 是球对称的，因此我们有 $\boldsymbol{R}\#\mathcal{D} = \mathcal{D}$。出于同样的原因， $g\circ\boldsymbol{R} = g$。此外，根据构造， $\boldsymbol{R}\boldsymbol{v} = \boldsymbol{v}$。因此，

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x}\} &= \frac{1}{2}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x}\} + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{R}\#\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x}\}\right)\\ &= \frac{1}{2}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x} + g(\boldsymbol{R}\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{R}\boldsymbol{x})\boldsymbol{R}\boldsymbol{x}\}\right)\\ &= \frac{1}{2}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x} + g(\boldsymbol{R}\boldsymbol{x})\sigma'(\boldsymbol{R}\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{R}\boldsymbol{x}\}\right)\\ &= \frac{1}{2}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})(\boldsymbol{x}+\boldsymbol{R}\boldsymbol{x})\}\right).\end{aligned}$$

注意， $\boldsymbol{x} + \boldsymbol{R}\boldsymbol{x} = 2\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^\top\boldsymbol{x} = 2\langle\bar{\boldsymbol{v}},\boldsymbol{x}\rangle\bar{\boldsymbol{v}}$。因此，

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma'(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{x}\} = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\sigma(\bar{\boldsymbol{v}}\cdot\boldsymbol{x})\}\bar{\boldsymbol{v}} = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\{g(\boldsymbol{x})\|\boldsymbol{x}\|\}\frac{C_\Gamma}{\sqrt{d}}\bar{\boldsymbol{v}},$$

其中第二个等式来自引理 B.3。 □

B.4 无限宽网络保持球对称

在本小节中，我们证明无限宽网络在整个过程中始终保持球对称。显然， μ_1 在初始化时具有球对称性。现在，假设它在时间 t具有球对称性。我们断言， $\boldsymbol{v}_1$ 不会沿切向移动，且其径向速度与方向 $\bar{\boldsymbol{v}}_1$无关。也就是说，对于某个函数 h，有 $\dot{\boldsymbol{v}}_1 = h(\|\boldsymbol{v}_1\|)\bar{\boldsymbol{v}}_1$。

根据归纳假设， S在时间 t也具有球对称性。令 $\boldsymbol{T}:=2\bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top-\boldsymbol{I}_d$为关于 $\boldsymbol{v}_1$的反射。显然， $\boldsymbol{T}\boldsymbol{v}_1=\boldsymbol{v}_1$。此外，它不会改变范数，因此，$S(T\boldsymbol{x})=S(\boldsymbol{x})$， $\boldsymbol{T}\#\mathcal{D}=\mathcal{D}$且 $\Pi\circ\boldsymbol{T}=\boldsymbol{T}\circ\Pi$。因此，我们有

$$\begin{aligned}\dot{\boldsymbol{v}}_1&=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\\&=\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\\&\quad+\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{T}\#\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}.\end{aligned}$$

对于第二项，我们有

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{T}\#\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\\=&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{T}\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{T}\boldsymbol{x})\boldsymbol{T}\boldsymbol{x}\right)\right]\right\}\\=&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{T}\boldsymbol{x}\right)\right]\right\}\\=&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\boldsymbol{T}\left(\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\bar{\boldsymbol{v}}_1+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\\=&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\boldsymbol{T}\,\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\bar{\boldsymbol{v}}_1+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}.\end{aligned}$$

因此，

$$\begin{aligned}\dot{\boldsymbol{v}}_1&=\frac{1}{2}(\boldsymbol{I}+\boldsymbol{T})\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\\&=2\left\langle\bar{\boldsymbol{v}}_1,\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\right\rangle\bar{\boldsymbol{v}}_1.\end{aligned}$$

即， $\dot{\boldsymbol{v}}_1=h(\boldsymbol{v}_1)\bar{\boldsymbol{v}}_1$ 其中

$$h(\boldsymbol{v}_1)=2\left\langle\bar{\boldsymbol{v}}_1,\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\right\rangle.$$

现在，我们证明 h 是球对称的，以完成证明。令 $\boldsymbol{R}$ 为任意旋转矩阵。我们有

$$\begin{aligned}h(\boldsymbol{R}\boldsymbol{v}_1)&=2\left\langle\boldsymbol{R}\bar{\boldsymbol{v}}_1,\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\boldsymbol{R}\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{R}\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{R}\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\right\rangle\\&=2\left\langle\boldsymbol{R}\bar{\boldsymbol{v}}_1,\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\boldsymbol{R}\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{R}^\top\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{R}^\top\boldsymbol{x})\boldsymbol{R}\boldsymbol{R}^\top\boldsymbol{x}\right)\right]\right\}\right\rangle\\&=2\left\langle\boldsymbol{R}\bar{\boldsymbol{v}}_1,\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{R}^\top\#\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\boldsymbol{R}\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{R}\boldsymbol{x}\right)\right]\right\}\right\rangle\\&=2\left\langle\boldsymbol{R}\bar{\boldsymbol{v}}_1,\boldsymbol{R}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\right\rangle\\&=2\left\langle\bar{\boldsymbol{v}}_1,\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left\{\Pi_{R_{\boldsymbol{v}_1}}\left[S(\boldsymbol{x})\left(\bar{\boldsymbol{v}}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})+\|\boldsymbol{v}_1\|\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right]\right\}\right\rangle\\&=h(\boldsymbol{v}_1).\end{aligned}$$

因此， h 是球对称的。

C 阶段 1

阶段 1 的目标是让所有 v_2 降低到 $-\Theta(1/R_{\boldsymbol{v}_2})$ ，这样我们就可以忽略 $\dot{r}_2$、 $\dot{\boldsymbol{v}}_1$ 和 $\dot{v}_2$中的所有投影算子。我们将阶段 1 分为三个子阶段，在每个子阶段中， v_2 都会降低到0、$-\mathrm{poly}(d)$ δ_2 和 $-\Theta(1/R_{\boldsymbol{v}_2})$。根据引理C.3，在每个子阶段结束时，都可以再忽略一个投影算子。我们还证明，在阶段 1 中，第一层的近似误差和第二层的分散度不会增长太多。

First，对于初始化，根据一些标准的集中性论证，我们有如下引理 .

引理 C.1（初始化）。我们选择 $1/\quad d\ 1/\varepsilon\quad \sigma_r\quad m_1 =\quad \boldsymbol{w}_1 \sim\quad \sigma_1\mathbb{S}^{d-1}\quad \sigma_2 = \mu_1^-$ $\mathrm{poly}(d, 1/\varepsilon)$、$m_2 = \Theta(1)$、$\sigma_1 = 1/\sqrt{d}$、$\mathrm{poly}(,)$，并将其设为一个较小的常数。对于，我们初始化为 $\mathrm{Unif}()$，并且 $w_2 \sim \mathcal{N}$ $(0、\sigma_2^2)$ 以及 $b_2 = \sigma_r$ 适用于 μ_2。

给定 $\delta_{1,\ I} = 1/$ 多项式$1(d, 1/\varepsilon)$，我们选择一个足够大的 m_1，使得在初始化时，以至少 $1 - 1/$ 多项式(d) 的概率，$\left\|\bar{F}|_{\mathbb{S}^{d-1}} - 1\right\|_{L^\infty} \le \delta_{1,\ I}$。我们还选择 $\sigma_2 = \delta_{1,\ I}/d^7$。以至少 $1 - 1/$多项式(d) 的概率，有 最大$_{w_2}|w_2| \le O$(对数 $d)\sigma_2$。

接下来，我们正式陈述在阶段 1 中将要维持的归纳假设。

归纳假设 C.2 (阶段 1)。我们定义 $T_1 :=$ 无穷 $\{t \ge 0 :\quad -\bar{w}_2(t) = \Theta(1)/R_{v_2}\}$ 其中某个常数足够大。定义 $\delta_{1\ T}^{(1)}$，$\delta_{1\ R}^{(1)}$，$\delta_2^{(1)}$ 为[9]

$$\begin{cases} \delta_{1,T}^{(1)} = \max\left\{\delta_{1,T}^{(1)}(0). \max\limits_{\boldsymbol{v}_1\in\mu_1} \|\bar{\boldsymbol{v}}_1(t) - \bar{\boldsymbol{v}}_1(0)\|\right\}, \\ \delta_{1,R}^{(1)} = \max\left\{\delta_{1,R}^{(1)}(0), \max\limits_{\boldsymbol{v}_1\in\mu_1} \left|\dfrac{\|\boldsymbol{v}_1\|^2 - \mathbb{E}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2}{\mathbb{E}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2}\right|\right\}, \qquad \textit{in Stage 1.1 and Stage 1.2,} \\ \delta_2^{(1)} = \max\left\{\delta_2^{(1)}(0), \max\limits_{(v_2,r_2),(v_2',r_2')} \|(v_2, r_2) - (v_2', r_2')\|\right\}, \end{cases}$$

and

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\delta_{1,T}^{(1)} = \mathrm{ReLU}\left(\dfrac{\mathrm{d}}{\mathrm{d}t}\max\limits_{\boldsymbol{v}_1\in\mu_1} \|\bar{\boldsymbol{v}}_1(t) - \bar{\boldsymbol{v}}_1(0)\|\right), \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\delta_{1,R}^{(1)} = \mathrm{ReLU}\left(\dfrac{\mathrm{d}}{\mathrm{d}t}\max\limits_{\boldsymbol{v}_1\in\mu_1} \left|\dfrac{\|\boldsymbol{v}_1\|^2 - \mathbb{E}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2}{\mathbb{E}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2}\right|\right), \qquad \textit{in Stage 1.3,} \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\delta_2^{(1)} = \mathrm{ReLU}\left(\dfrac{\mathrm{d}}{\mathrm{d}t}\max\limits_{(v_2,r_2),(v_2',r_2')} \|(v_2, r_2) - (v_2', r_2')\|\right), \end{cases}$$

初始值为 $\delta_{1\ T}^{(1)}(0) = \delta_{1\ R}^{(1)}(0) = 0$ 以及 $\delta_2^{(1)}(0) = \Theta(\sigma_2 \log d)$。

如果以下条件成立，我们就说该归纳假设在时刻 $t \in [0, T_1]$ 成立。[10]

(a) **第一层的近似误差**。对于每个 $\boldsymbol{v}_1 \in \mu_1$，$\|\bar{\boldsymbol{v}}_1(t) - \bar{\boldsymbol{v}}_1(0)\| \le \delta$ $^{(1)}_{1,}{}_T$ 和 $\|\boldsymbol{v}_1\|^2 = \left(1 \pm \delta_{1,R}^{(2)}\right)\mathbb{E}_{\boldsymbol{w}_1\sim\mu_1}\|\boldsymbol{w}_1\|^2$.

(b) **第二层的扩散**。对于任意 $(v_2,\ r_2)$，$(v_2',\ r_2') \in \mu_2$，$\|(v_2,\ r_2) - (v_2',\ r_2')\| \le \delta_2^{(1)}$。

(c) **偏置项**。对于任意 $(v_2,\ r_2) \in \mu_2$，$r_2 = \Theta(1)$。

(d) **大小为** f**.** $|\bar{w}_2| = O(1/R_{v_2}) = O(1/d^3)$ 和 $\alpha = \Theta(\sqrt{d}/R_{\boldsymbol{v}_1}) = \Theta(1/d^{1.5})$.

(e) **误差的上界**。$\delta_2^{(1)} \le O(d^{1.5}(\log\ d)\sigma_2)$ 和 $\delta_{1\ ,R}^{(1)} + \delta_{1\ ,T}^{(1)} \le O(d^7(\log\ d)\sigma_2 + \delta_{1,I})$

下一个引理说明何时可以忽略投影算子。粗略地说，我们首先对梯度进行界定，以表明要触发投影算子，$\|\boldsymbol{x}\|$ 必须大于某个量。同时，注意到 f，因此，对于满足 $\|\boldsymbol{x}\| \ge \Theta(1/|\bar{w}_2\alpha|)$ 的那些 $\boldsymbol{x}$，梯度会消失。因此，只要 $\Theta(1/|\bar{w}_2\alpha|)$小于该量，我们就可以忽略投影。

[9]请注意，我们将这些 δ定义为相应值的上界，而不是值本身。我们以这种曲折的方式定义这些 δ 的唯一原因，是为了让证明更容易严谨地书写。关于进一步讨论，请参见归纳假设 D.1中的脚注，其中这类定义在技术上发挥着更重要的作用。[10]前两个条件实际上可直接由 δ的定义推出。我们将其重复列出，仅为便于查阅。我们需要证明这些 δ的实际结果是条件 (e)，即这些δ始终很小。

引理 C.3. 假设归纳假设 C.2 成立。如果所有第二层权重均为非正数，那么当 $\dot{r}_2$, $\dot{\boldsymbol{v}}_1$ 和$\dot{v}_2$ 分别满足 $-\bar{w}_2 > \Theta(1)\delta_2^{(1)}$ （其中某个常数足够大），以及 $-\bar{w}_2 \geq \Theta(1)/R_{v_2}$ （其中某个常数足够大）时，其中的投影算子将不再被激活。

注。 尽管我们只需要 $-\bar{w}_2$ 为 $\Theta(1)\delta_2^{(1)}$，以便忽略 $\dot{\boldsymbol{v}}_1$中的投影算子，但实际上，我们将阶段 1.2 的结束定义为 $-\bar{w}_2$ 变为 $\text{poly}(d)\delta_2^{(1)}$ 的时刻，从而为阶段 1.3 获得一个更加规整的起点。♣

现在，我们给出阶段 1 的主引理。可以看出，通过适当选择参数，可以使误差任意小，同时不影响 α 和 $\bar{w}_2$的最终值。要证明主要引理，只需将引理C.6、引理C.9 和引理 C.10 结合起来即可。

引理 C.4(阶段 1 的主引理)。 归纳假设 C.2 在整个阶段 *1* 中均成立。阶段 *1* 至多需要 $O(d^4\sigma_2 + 1/d^{1.5})$ 的时间。在阶段 *1* 结束时，我们有 $\alpha = \Theta(1/d^{1.5})$ 以及$-\bar{w}_2 = \Theta(1/d^3)$。对于误差，我们有 $\delta_2^{(2)} \leq O(d^{1.5}\log d\sigma_2)$ 以及 $\delta_{1,R}^{(1)} + \delta_{1,T}^{(1)} \leq O(\delta_{1,I})$。

引理 C.3 的证明。首先，注意到当所有 v_2 均为非正时，我们有 $f = O(1)$。由于我们选择 R_{r_2} 为一个较大的常数，这意味着 $\dot{r}_2$ 中的投影算子不会被激活。当 $-\bar{w}_2 > \Theta(1)\delta_2^{(1)}$ 时，对于某个较小的常数 $c > 0$，我们有 $f(\boldsymbol{x}) \leq \sigma(c\bar{w}_2\alpha\|\boldsymbol{x}\| + O(1))$。因此，$f$在 $\{\|\boldsymbol{x}\| \geq (-c\bar{w}_2\alpha)^{-1}\}$上消失。接着，对于满足 $\|\boldsymbol{x}\| \leq (-c\bar{w}_2\alpha)^{-1}$ 的那些 $\boldsymbol{x}$，相对于 $\boldsymbol{v}_1$ 的梯度可界定为

$$\|\nabla_{\boldsymbol{v}_1}\mathcal{L}(\boldsymbol{x})\| \leq O(1)|\bar{w}_2|\,\|\boldsymbol{x}\|\,\|\boldsymbol{v}_1\| \leq O(1)|\bar{w}_2|\,\|\boldsymbol{v}_1\|\,\frac{1}{|\bar{w}_2|\alpha} \leq O(d).$$

由于我们选择 $R_{\boldsymbol{v}_1} = \Theta(d)$ 为一个较大的常数，这意味着 $\dot{\boldsymbol{v}}_1$ 中的投影算子不会被触发。最后，对于 $\dot{v}_2$，对于满足 $\|\boldsymbol{x}\| \leq (-c\bar{w}_2\alpha)^{-1}$ 的那些 $\boldsymbol{x}$，我们有

$$|\nabla_{v_2}\mathcal{L}(\boldsymbol{x})| \leq O(1)\alpha\|\boldsymbol{x}\| \leq \frac{O(1)}{|\bar{w}_2|}.$$

根据假设，$|\bar{w}_2| = \Theta(1)/R_{v_2}$ 对于某个足够大的常数成立。因此，该不等式意味着 $\dot{v}_2$ 中的投影算子将不会被触发。 □

C.1 阶段 1.1

阶段 1.1 的目标是确保所有第二层权重 v_2 都变为非正，即

$$T_{1.1} := \inf\{t \geq 0 : \forall(v_2, r_2) \in \mu_2,\, v_2 \leq 0\}.$$

因此，在阶段 1.1 结束时，f 是 $O(1)$ 并且，根据引理 C.3，$\dot{r}_2$ 中的投影算子可以忽略。由于该阶段只需耗费极少的时间，我们将直接通过界定 $\boldsymbol{v}_1$的移动来控制第一层误差。对于第二层，我们以同样的直接方式界定偏置项的移动。对于第二层权重，我们证明正的v_2会比负的 v_2下降得更快，因此分散度不会增加。

引理 C.5。 假设归纳假设 C.2 在时刻 t 和 $t \leq T_{1.1}$成立。那么以下结论成立。

(a) $\|\dot{\boldsymbol{v}}_1\| \leq R_{\boldsymbol{v}_1}$ 和 $|\dot{r}_2| \leq R_{r_2}$。*(b)* 最大值$_{w_2}\, w_2$ –最小值$_{w_2}\, w_2$ 是非递增的。*(c)* 对任意正的第二层权重 v_2，我们有 $\dot{v}_2 \leq -\Theta(\log d/d^{1.5})$。

注。 事实上，只要对于任意 $(v_2,\ r_2) \in \mu_2$和 $\boldsymbol{x} \in \{\|\boldsymbol{x}\| \leq d^{1.5}\}$，$\alpha = \Omega(1/d^{1.5})$ 且 $v_2F(\boldsymbol{x}) + r_2 \geq \Theta(1)$，(c) 就成立；这在整个阶段 1 中始终成立。此估计也将在阶段 1.2 和阶段 1.3 中使用。♣

引理 C.6 (阶段 1.1 的主引理)。 阶段 *1.1* 最多需要 $O(d^{1.5}\delta_2^{(1)}(0))$ 的时间。在阶段 *1.1* 结束时，所有第二层权重 v_2 均为非正。因此， $f = O(1)$ 并且根据引理 *C.3*， $\dot{r}_2$ 中的投影算子不再可能被激活。

对于误差，我们有 $\delta_2^{(1)}(T_{1.1}) \le O(d^{1.5}\delta_2^{(1)}(0))$，且 $\delta_{1R}^{(1)}(T_{1.1})$ 和 $\delta_{1T}^{(1)}(T_{1.1})$ 均可由 $O(d^3\delta_2^{(1)}(0))$界定。

引理*C.5*的证明。

(a) 这是显而易见的。

(b) 首先，我们将 v_2 分解为

$$\dot{v}_2 = \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))F(\boldsymbol{x})\} - \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\Pi_{R_{v_2}}\left[f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x})+r_2)F(\boldsymbol{x})\right]\right\}.$$

注意，第一项不依赖于 v_2，并且对于第二项，当 $v_2 \ge 0$ 时， $\sigma'(v_2F(\boldsymbol{x})+r_2) = 1$。因此，正 v_2 的速度是一致的，并且比那些 $v_2 < 0$ 更负。因此，$\max_{w_2} w_2 - \min_{w_2} w_2$ 是非递增的。

(c) 显然， $\mathbb{E}_{\|\boldsymbol{x}\|\le 1}\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))F(\boldsymbol{x})\} = O(\alpha)$。对于第二项，首先注意到，对于任意满足 $\|\boldsymbol{x}\| \le d^{1.5}$ 的 $\boldsymbol{x}$，我们有

$$f(\boldsymbol{x})F(\boldsymbol{x}) \le O\left(1 + \max_{w_2} w_2\alpha\|\boldsymbol{x}\|\right)\alpha\|\boldsymbol{x}\| \le R_{v_2} \quad \text{and} \quad f(\boldsymbol{x}) \ge \Theta(1) - \max_{w_2}|w_2|\alpha\|\boldsymbol{x}\| = \Theta(1).$$

因此，

$$\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\Pi_{R_{v_2}}\left[f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x})+r_2)F(\boldsymbol{x})\right]\right\} \ge \Theta(\alpha)\mathop{\mathbb{E}}_{1\le\|\boldsymbol{x}\|\le d^{1.5}}\|\boldsymbol{x}\| = \Theta\left((\log d)\alpha\right).$$

于是， $\dot{v}_2 \le -\Theta(\log d/d^{1.5})$。

□

引理的证明*C.6*。 由引理 C.5 可知，所有 v_2 变为非正值所需的时间至多为 $O(d^{1.5}\delta_2^{(1)}(0))$。在这段时间内， r_2 的变化至多为 $O(d^{1.5}\delta_2^{(1)}(0))$。由于 w_2 的分散度不会增加，这意味着 $\delta_2^{(1)}(T_{1.1}) \le O(d^{1.5}\delta_2^{(1)}(0))$。最后， $\boldsymbol{v}_1$ 的变化可由 $O(d^{2.5}\delta_2^{(1)}(0))$ 上界。因此， $\delta_{1R}^{(1)}(T_{1.1})$ 和 $\delta_{1T}^{(1)}(T_{1.1})$ 均可由 $O(d^3\delta_2^{(1)}(0))$ 上界。

□

C.2 阶段 1.2

阶段 1.2 的目标是确保 $-\bar{w}_2 \ge d\delta_2^{(1)}(T_{1.1})$。即，

$$T_{1.2} := \inf\left\{t \ge T_{1.1} : -\bar{w}_2 = d\delta_2^{(1)}(T_{1.1})\right\}.$$

我们还将证明 $\delta_2^{(1)}(T_{1.2}) = O(\delta_2^{(1)}(T_{1.1}))$ 因此 $\delta_2^{(1)}(T_{1.2})/|\bar{w}_2| = O(1/d)$ 在阶段 1.2 结束时成立。此外，根据引理 C.3，在阶段 1.2 结束时， $\dot{\boldsymbol{v}}_1$ 中的投影算子将不再激活。我们还将证明 r_2 在本小节的整个阶段 1 期间保持 $\Theta(1)$。

第一层误差仍以蛮力方式控制。对于第二层分散度，我们证明，由于 $|v_2|$ 很小，对于大多数 $\boldsymbol{x}$， $\sigma'(v_2F(\boldsymbol{x}) + r_2) = 1$，因此 $(v_2,\ r_2)$ 的变化近似均匀。

引理 C.7. 假设归纳假设 *C.2* 在时刻 t成立。那么，对于任意 $(v_2, r_2) \in \mu_2$，$\dot{r}_2 > 0$ 当 $r \le \mathbb{E} f_*/2$ 且 $\dot{r}_2 < 0$ 当 $r \ge 2\mathbb{E} f_*$时。因此， $r_2 = \Theta(1)$ 在整个阶段 *1* 期间成立。

引理 C.8 (第二层的扩散)。 假设归纳假设 C.2 在时刻 t和 $t \le T_{1.2}$成立。那么，对于任意 (v_2, r_2),$(v_2', r_2') \in \mu_2$，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t} \|(v_2, r_2) - (v_2', r_2')\|^2 \le O(d^{2.5}) \left(\delta_2^{(1)}\right)^2.$$

不过，根据这个引理，误差 $\delta_2^{(1)}$ 可以指数级快速增长，且增长率相当大，但不会爆炸到 $\dot{v}_2 \le -\Theta(\log d/d^{1.5})$，因此阶段 1.2 所需的时间远短于 $1/d^{2.5}$。

引理 C.9 (阶段 1.2 的主引理)。 阶段 *1.2* 至多需要 $O(d^{2.5}\delta_2^{(1)}(T_{1.1}))$ 的时间。阶段 *1.2* 结束时，对于任意$(v_2, r_2) \in \mu_2$, $-v_2 \ge \Theta(d)\delta_2^{(1)}(T_{1.1})$。

对于误差，第二层的分散度为$(1 + o(1))\delta_2^{(1)}(T_{1.1})$，且 $\delta_{1R}^{(1)}(T_{1.2})$ 和$\delta_{1T}^{(1)}(T_{1.2})$ 都可以被界定为 $O(d^4\delta_2^{(1)}(T_{1.1}))$。

引理的证明C.7. 我们写作

$$\dot{r}_2 = \mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\sigma'(v_2F(\boldsymbol{x}) + r_2)\} = \mathop{\mathbb{E}}_{\boldsymbol{x}} f_*(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{x}}\{f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2)\}$$

由于 b_2 的分散度为 $o(1)$，当 $r_2 \le \mathbb{E}_{\boldsymbol{x}} f_*(\boldsymbol{x})/2 = \Theta(1)$ 时，右端是一个正的常数。换句话说， r_2 将持续增长。同时，由于第二项可以界定为 $\mathbb{E}_{\boldsymbol{x}}\{f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2)\} \ge \mathbb{E}_{\|\boldsymbol{x}\|\le d^2}\{f(\boldsymbol{x})\} \ge (1 - o(1))b_2$，当 $r_2 \ge 2\,\mathbb{E} f_*(\boldsymbol{x})$时， $\dot{r}_2$将变为负常数，而 r_2 将减小。综合这两种情况，即完成证明。 □

引理的证明 C.8. 由于 $|v_2| \le d\delta_2^{(1)}(T_{1.1})$， $F(\boldsymbol{x}) = \Theta(\alpha)\|\boldsymbol{x}\|$ 且 $r_2 = \Theta(1)$， $v_2F(\boldsymbol{x}) + r_2 > 0$ 对所有 $\boldsymbol{x}$ 满足 $\|\boldsymbol{x}\| \le \Theta(\sqrt{d}/\delta_2^{(1)}(T_{1.1}))$。因此，我们可以将 $\dot{v}_2$ 改写为

$$\dot{v}_2 = \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le\Theta(\sqrt{d}/\delta_2^{(1)}(T_{1.1}))}\left\{\Pi_{R_{v_2}}[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))F(\boldsymbol{x})]\right\} - \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge\Theta(\sqrt{d}/\delta_2^{(1)}(T_{1.1}))}\left\{\Pi_{R_{v_2}}[f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2)F(\boldsymbol{x})]\right\}.$$

第一项不依赖于 v_2，并且根据尾界，第二项可以界定为$O(R_{v_2}\delta_2^{(1)}(T_{1.1})/\sqrt{d})$。类似地，对于 $\dot{r}_2$，我们有

$$\dot{r}_2 = \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le\Theta(\sqrt{d}/\delta_2^{(1)}(T_{1.1}))}\{f_*(\boldsymbol{x}) - f(\boldsymbol{x})\} \pm O\left(\delta_2^{(1)}(T_{1.1})/\sqrt{d}\right).$$

因此，对于任意(v_2, r_2)，$(v_2', r_2') \in \mu_2$，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}\|(v_2, r_2) - (v_2', r_2')\|^2 \le (v_2 - v_2')O\left(\frac{R_{v_2}\delta_2^{(1)}(T_{1.1})}{\sqrt{d}}\right) + (r_2 - r_2')O\left(\frac{\delta_2^{(1)}(T_{1.1})}{\sqrt{d}}\right) \le O(d^{2.5})\left(\delta_2^{(1)}\right)^2.$$

□

引理 C.9 的证明。回顾引理 C.5，有 $\dot{v}_2 = -\Theta(\log d/d^{1.5})$，因此阶段 1.2 至多需要 $O(d^{2.5}\delta_2^{(1)}(T_{1.1}))$的时间。根据引理 C.8，我们有

$$\left(\delta_2^{(1)}(T_{1.2})\right)^2 \le \left(\delta_2^{(1)}(T_{1.1})\right)^2 \exp\left(O(d^5)\delta_2^{(1)}(T_{1.1})\right) \le (1 + o(1))\left(\delta_2^{(1)}(T_{1.1})\right)^2.$$

对于 $\boldsymbol{v}_1$，与引理 C.6 的证明类似， $\delta_{1R}^{(1)}(T_{1.2})$ 和 $\delta_{1T}^{(1)}(T_{1.2})$ 均可由$O(d^4\delta_2^{(1)}(T_{1.1}))$ 界定。 □

C.3 阶段 1.3

阶段 1.3 的目标是确保 $-\bar{w}_2 = \Theta(1/R_{v_2})$ 对于某个较大的常数成立，这样根据引理C.3，就可以忽略 $\dot{v}_2$ 中的投影算子。也就是说，我们定义

$$T_{1.3} := \inf\{t \geq T_{1.2} : -\bar{w}_2(t) = \Theta(1/R_{v_2})\}.$$

这一阶段所需的时间比前几个阶段更长，因此我们需要用不那么粗暴的方法来控制误差。对于第一层，我们证明切向移动几乎为零，而径向移动近似均匀。对于第二层，我们证明分散度 $\delta_2^{(1)}$ 不会增长得太快。

引理 C.10 (阶段 1.3 的主引理)。 阶段 *1.3* 至多需要 $O(1/d^{1.5})$ 的时间。在阶段 *1.3* 结束时，我们有 $-\bar{w}_2 = \Theta(1/R_{v_2})$ 以及 $\alpha = \Theta(\sqrt{d}/R_{v_1})$。
对于误差，第二层的分散度为 $O\left(\delta_2^{(1)}(T_{1.2})\right)$，而第一层误差为
$O\left(\delta_{1,R}^{(1)}(T_{1.2}) + \delta_{1,T}^{(1)}(T_{1.2}) + \delta_{1,I} + \log(d)\delta_2^{(1)}(T_{1.2})\right)$.

证明。 由于 $\dot{v}_2 = -\Omega(\log d/d^{1.5})$ 且 $R_{v_2} = \Theta(d^3)$，阶段 1.3 至多需要 $O(1/d^{1.5})$ 的时间。在这段时间内，根据引理C.18，我们有

$$(\delta_2^{(1)}(T_{1.3}))^2 \leq (\delta_2^{(1)}(T_{1.2}))^2 \exp\left(\frac{O(1)}{d^{2.5}}\frac{1}{d^{1.5}}\right) = (1+o(1))(\delta_2^{(1)}(T_{1.2}))^2.$$

对于第一层，根据引理 C.16，我们有

$$\begin{aligned}\delta_{1,R}^{(1)}(T_{1.3}) + \delta_{1,T}^{(1)}(T_{1.3}) &\leq \left(\delta_{1,R}^{(1)}(T_{1.2}) + \delta_{1,T}^{(1)}(T_{1.2}) + \frac{O(1)}{d^3}\delta_{1,I} + O\left(\log(d)\delta_2^{(1)}\right)\right)\exp\left(\frac{O(1)}{d^{2.5}}\frac{1}{d^{1.5}}\right)\\ &= O\left(\delta_{1,R}^{(1)}(T_{1.2}) + \delta_{1,T}^{(1)}(T_{1.2}) + \frac{\delta_{1,I}}{d^3} + \log(d)\delta_2^{(1)}(T_{1.2})\right).\end{aligned}$$

最后，根据引理 C.17，我们有 $\alpha(T_{1.3}) = (1+o(1))\alpha(T_{1.2})$。 □

C.3.1 与 $\sigma'(v_2F(\mathbf{x}) + r_2)$ 相关的估计

首先，我们需要一些辅助结果来处理 $\sigma'(v_2F(\boldsymbol{x}) + r_2)$。这些结果成立的条件较为宽松，并且在整个训练过程中始终成立；后续阶段中我们也会用到这些结果。

首先，我们证明，当 $\sigma'(v_2F(\boldsymbol{x}) + r_2)$ 的值在不同的 (v_2, r_2) 之间可能发生变化时，函数值必定较小。请注意，此处的误差取决于比值 $\delta_2/|\bar{w}_2|$，这就是为什么在阶段 1.2 结束时，我们需要让 $|\bar{w}_2|$ 为 $\Theta(d)\delta_2$，而不能仅仅令其为 $\Theta(1)\delta_2$。

引理 C.11. 假设对于任意 $(v_2, r_2) \in \mu_2$，$r_2 = \Theta(1)$，$-v_2 \geq \Omega(\delta_2)$ 成立，其中 δ_2 是第二层的分散度。如果对于某个 $(,)$，$v_2F(\boldsymbol{x}) + r_2 = 0$ $v_2\ r_2 \in \mu_2$ $v_2'F(\boldsymbol{x}) + r_2' \leq ()$ 成立，那么对于所有 $(,)$，$()\,(())$ 均成立 $O\ |\bar{w}_2|^{-1} + 1\ \delta_2$ $v_2'\ r_2' \in \mu_2$。

注. 并不一定真的存在满足 $v_2F(\boldsymbol{x}) + r_2 = 0$ 的 $(v_2, r_2) \in \mu_2$。只要对于某个 $(v_2', r_2'), (v_2'', r_2'') \in \mu_2$，$v_2'F(\boldsymbol{x}) + r_2' \leq 0$ 且 $v_2''F(\boldsymbol{x}) + r_2'' \geq 0$，根据连续性，在 (v_2', r_2') 与 (v_2'', r_2'') 之间总存在某个点 (v_2, r_2)，使得 $v_2F(\boldsymbol{x}) + r_2 = 0$。此外，该点位于第二层的分散度范围内，因此此引理仍然适用。♣

然后，我们证明可以将 σ' 吸收到 f_* 和 f 中。

引理 C.12。 假设引理 *C.11* 的假设成立，且所有第二层神经元都在 $\{\|\boldsymbol{x}\| \leq 1\}$ 上激活。那么，对于任意 $(v_2,\ r_2) \in \mu_2$ 和 $\boldsymbol{x} \in \mathbb{R}^d$，有

$$f_*(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2) = f_*(\boldsymbol{x}) \quad \textit{and} \quad f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2) = f(\boldsymbol{x}) \pm O\left((|\bar{w}_2|^{-1} + 1)\delta_2\right).$$

作为一个推论，我们有

$$\begin{aligned}f(\boldsymbol{x}) &= \sigma(v_2F(\boldsymbol{x}) + r_2) \pm O\left((|\bar{w}_2|^{-1} + 1)\delta_2\right),\\ f(\boldsymbol{x}) &= \sigma(\bar{w}_2F(\boldsymbol{x}) + \bar{b}_2) \pm O\left((|\bar{w}_2|^{-1} + 1)\delta_2\right).\end{aligned}$$

作为引理 C.11的推论， $\sigma'(v_2F(\boldsymbol{x}) + r_2)$ 对不同的(v_2, r_2)可能出现差异的测度也很小。这里我们还利用了这样一个事实：这些 $\boldsymbol{x}$ 位于 $\Theta(1/|\bar{w}_2\alpha|)$ 尾界 $\|\mathcal{D}\|(r) \leq O(1/r^2)$附近。

引理 C.13。 假设归纳假设 *C.2* 在时刻 t成立。对于任意 $(v_2, r_2),(v_2', r_2') \in \mu_2$，有

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{|\sigma'(v_2F(\boldsymbol{x}) + r_2) - \sigma'(v_2'F(\boldsymbol{x}) + r_2')|\right\} \leq O\left(\alpha\delta_2^{(1)}\right).$$

引理 *C.11* 的证明。 对于任意 $(v_2', r_2') \in \mu_2$，我们可以写成

$$\begin{aligned} v_2'F(\boldsymbol{x}) + r_2' &= \underbrace{v_2F(\boldsymbol{x}) + r_2}_{=0} + (v_2' - v_2)F(\boldsymbol{x}) + (r_2' - r_2) \\ &= \frac{v_2' - v_2}{v_2}(\underbrace{v_2F(\boldsymbol{x}) + r_2}_{=0} - r_2) + (r_2' - r_2) = r_2\frac{v_2 - v_2'}{v_2} + (r_2' - r_2). \end{aligned}$$

最后一项可界定为 $O((|\bar{w}_2|^{-1} + 1)\delta_2)$。 □

引理 C.12 的证明。由于所有第二层神经元在 $\{\|\boldsymbol{x}\| \leq 1\}$ 上都被激活，因此我们总有$f_*(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2) = f_*(\boldsymbol{x})$。现在考虑 $f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2)$。如果 $v_2F(\boldsymbol{x}) + r_2 > 0$，则得证。如果对所有 $(v_2', r_2') \in \mu_2$，都有 $v_2'F(\boldsymbol{x}) + r_2' < 0$，那么 $f(\boldsymbol{x})\sigma'(v_2F(\boldsymbol{x}) + r_2)$ 和$f(\boldsymbol{x})$ 均为 0。因此，只需考虑 $v_2F(\boldsymbol{x}) + r_2 \leq 0$而 $f(\boldsymbol{x}) > 0$ 的情况。根据引理 D.6，在这种情况下，我们有 $f(\boldsymbol{x}) \leq O\left((|\bar{w}_2|^{-1} + 1)\delta_2\right)$。 □

引理*C.13*. 由于 $\boldsymbol{x}$ 的范数和方向相互独立，只需固定一个方向 $\bar{\boldsymbol{x}}$ 并考虑

$$\mathop{\mathbb{E}}_{r\sim\|\mathcal{D}\|}\left\{|\sigma'(v_2rF(\bar{\boldsymbol{x}}) + r_2) - \sigma'(v_2'rF(\bar{\boldsymbol{x}}) + r_2')|\right\}.$$

为简化记号，定义 $h(v_2, r_2, r) = v_2rF(\bar{\boldsymbol{x}}) + r_2$。被积函数非零，当且仅当 $h(v_2, r_2, r)$ 和 $h(v_2', r_2', r)$ 的符号不同。为了界定符号可能不同的区间长度，我们写作

$$\begin{aligned} h(v_2, r_2, r) &= \bar{w}_2rF(\bar{\boldsymbol{x}}) + \bar{b}_2 + (v_2 - \bar{w}_2)rF(\bar{\boldsymbol{x}}) + (r_2 - \bar{b}_2) \\ &= \left(\bar{w}_2 \pm O\left(\delta_2^{(1)}\right)\right)rF(\bar{\boldsymbol{x}}) + \bar{b}_2 \pm O\left(\delta_2^{(1)}\right). \end{aligned}$$

因此，该区间的长度可由 $O(\delta_2^{(1)}/(\bar{w}_2^2\alpha))$ 界定。此外，注意该区间位于 $\Theta(1/|\bar{w}_2\alpha|)$，故其上的密度为 $O(\bar{w}_2^2\alpha^2)$。因此，该区间的测度为 $O(\alpha\delta_2^{(1)})$。

□

C.3.2 第一层的估计

在控制误差增长之前，我们需要一个引理，用来关联近似误差与第一层的切向移动和径向扩散。

引理 C.14. 假设第一层神经元的切向移动和径向扩散可以界定为 $\|\bar{\boldsymbol{v}}_1(t) - \bar{\boldsymbol{v}}_1(0)\| \leq \delta_{1,T}$ 以及 $\|\boldsymbol{v}_1\|^2 = (1 \pm \delta_{1,R})\,\mathbb{E}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2$。那么

$$F(\boldsymbol{x}; \mu_1) = \left(1 + \delta_{1,I} + \sqrt{d}\delta_{1,R} + \sqrt{d}\delta_{1,T}\right)\alpha\|\boldsymbol{x}\|.$$

作为一个简单的推论，我们有如下结果。

推论 C.15. 假设归纳假设 *C.2* 在时刻成立 t。那么，我们有

$$|f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| = \left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\alpha\|\boldsymbol{x}\|.$$

因此，我们有

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\|\boldsymbol{x}\|\right\} &\leq \left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\alpha\,\mathbb{E}\|\boldsymbol{x}\|^2 \\ &\leq \delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}. \end{aligned}$$

现在，我们可以控制第一层的误差了。

引理 C.16. 假设归纳假设 C.2 在时刻 t 和 $t \in [T_{1.2}, T_{1.3}]$成立。那么我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\delta_{1,R}^{(1)}+\delta_{1,T}^{(1)}\right) &\le O\left(\left(\delta_{1,I}+\sqrt{d}\delta_{1,R}^{(1)}+\sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\right)+O\left(\log(d)\delta_2^{(1)}\right)\\ &\le \frac{O(1)}{d^{2.5}}\left(\delta_{1,R}^{(1)}+\delta_{1,T}^{(1)}\right)+\frac{O(1)}{d^3}\delta_{1,I}+O\left(\log(d)\delta_2^{(1)}\right).\end{aligned}$$

最后，我们估计 $\boldsymbol{v}_1$ 的径向速度，以估计阶段 1 结束时 α 的大小。

引理 C.17. 假设归纳假设 C.2 在时刻 t 和 $t \in [T_{1.2}, T_{1.3}]$成立。那么我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}_1\|^2 = \Theta\left(\frac{\log d}{\sqrt{d}}\right)\bar{w}_2\|\boldsymbol{v}_1\|^2.$$

引理 C.14的证明。 定义 $N^2 = \mathbb{E}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2$。令 μ_1' 为通过将 μ_1 中神经元的范数设为 N而得到的分布。我们有

$$F(\boldsymbol{x};\mu_1) = \mathop{\mathbb{E}}_{\boldsymbol{w}_1\sim\mu_1}\left\{(1\pm\delta_{1,R})N^2\sigma(\bar{\boldsymbol{w}}_1\cdot\boldsymbol{x})\right\} = F(\boldsymbol{x};\mu_1')\pm O(\delta_{1,R}N^2\|\boldsymbol{x}\|).$$

令 μ_1'' 为通过将 $\bar{\boldsymbol{v}}_1(t)$ 在 μ_1'中从 $\bar{\boldsymbol{v}}_1$移动到 (0) 而得到的分布。那么我们有

$$F(\boldsymbol{x};\mu_1') = N^2\mathop{\mathbb{E}}_{\boldsymbol{w}_1\sim\mu_1(0)}\{\sigma(\bar{\boldsymbol{w}}_1\cdot\boldsymbol{x})\}\pm O\left(\delta_{1,T}N^2\|\boldsymbol{x}\|\right) = F(\boldsymbol{x};\mu_1'')\pm O\left(\delta_{1,T}N^2\|\boldsymbol{x}\|\right).$$

最后，注意到

$$F(\boldsymbol{x};\mu_1'') = \frac{N_t^2}{N_0^2}F(\boldsymbol{x};\mu_1(0)) = \frac{N_t^2}{N_0^2}(1\pm\delta_{1,I})\alpha_0\|\boldsymbol{x}\| = (1\pm\delta_{1,I})\alpha_t\|\boldsymbol{x}\|.$$

将这些结果合并，即完成证明。 □

*Pro*引理 C.16的证明。 首先，我们将 $\dot{\boldsymbol{v}}_1$ 沿切向方向和径向方向分解如下：

$$\begin{aligned}\mathrm{Rad}(\dot{\boldsymbol{v}}_1) &:= \langle\dot{\boldsymbol{v}}_1,\bar{\boldsymbol{v}}_1\rangle\bar{\boldsymbol{v}}_1 = 2\mathop{\mathbb{E}}_{\boldsymbol{x}}\{S(\boldsymbol{x})\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\}\bar{\boldsymbol{v}}_1,\\ \mathrm{Tan}(\dot{\boldsymbol{v}}_1) &:= (\boldsymbol{I}-\bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\dot{\boldsymbol{v}}_1 = \|\boldsymbol{v}_1\|\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{S(\boldsymbol{x})\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{I}-\bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\boldsymbol{x}\right\}.\end{aligned}$$

注意， $\dot{\boldsymbol{v}}_1 = \mathrm{Rad}(\dot{\boldsymbol{v}}_1) + \mathrm{Tan}(\dot{\boldsymbol{v}}_1)$。根据引理 C.12，可得

$$\begin{aligned}\mathrm{Rad}(\dot{\boldsymbol{v}}_1) &= 2\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\}\bar{\boldsymbol{v}}_1\pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right),\\ \mathrm{Tan}(\dot{\boldsymbol{v}}_1) &= \|\boldsymbol{v}_1\|\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{I}-\bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\boldsymbol{x}\right\}\pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right).\end{aligned}$$

对于径向项，根据引理 B.3 和引理C.15，可得

$$\begin{aligned}\mathrm{Rad}(\dot{\boldsymbol{v}}_1) &= 2\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\right\}\bar{\boldsymbol{v}}_1+2\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\tilde{f}(\boldsymbol{x})-f(\boldsymbol{x}))\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x})\right\}\bar{\boldsymbol{v}}_1\pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right)\\ &= \frac{2C_\Gamma\bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))\|\boldsymbol{x}\|\right\}\boldsymbol{v}_1\\ &\quad\pm O\left(\left(\delta_{1,I}+\sqrt{d}\delta_{1,R}^{(1)}+\sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\|\boldsymbol{v}_1\|\right)\pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right).\end{aligned}$$

因此，

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}_1\|^2 &= 2\langle\boldsymbol{v}_1,\mathrm{Rad}(\dot{\boldsymbol{v}}_1)\rangle\\ &= \frac{4C_\Gamma\bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))\|\boldsymbol{x}\|\right\}\|\boldsymbol{v}_1\|^2\\ &\quad\pm O\left(\left(\delta_{1,I}+\sqrt{d}\delta_{1,R}^{(1)}+\sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\|\boldsymbol{v}_1\|^2\right)\pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|^2\right).\end{aligned}$$

对于任意 $\boldsymbol{v}_1, \boldsymbol{v}_1' \in \mu_1$，满足 $\|\boldsymbol{v}_1\| \geq \|\boldsymbol{v}_1'\|$，我们有

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left(\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2\right)}{\|\boldsymbol{v}_1'\|^2} - \frac{\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2} \\
&= \frac{4C_\Gamma \bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\|\boldsymbol{x}\|\right\}\frac{\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2} \\
&\quad \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\right) \pm O\left(\log(d)\delta_2^{(1)}\right) \\
&\quad - \frac{\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2}\frac{4C_\Gamma \bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\|\boldsymbol{x}\|\right\} \\
&\quad \pm \frac{\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2}O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\right) \pm \frac{\|\boldsymbol{v}_1\|^2 - \|\boldsymbol{v}_1'\|^2}{\|\boldsymbol{v}_1'\|^2}O\left(\log(d)\delta_2^{(1)}\right) \\
&= \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\right) \pm O\left(\log(d)\delta_2^{(1)}\right).
\end{aligned}$$

现在我们考虑切向移动。根据引理B.5 和引理 C.15，我们有

$$\begin{aligned}
\mathrm{Tan}(\dot{\boldsymbol{v}}_1) &= \|\boldsymbol{v}_1\|\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{I} - \bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\boldsymbol{x}\right\} \\
&\quad + \|\boldsymbol{v}_1\|\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{I} - \bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\boldsymbol{x}\right\} \pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right) \\
&= \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\|\boldsymbol{v}_1\|\right) \pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right).
\end{aligned}$$

因此，

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{v}}_1 = \frac{\mathrm{Tan}(\dot{\boldsymbol{v}}_1)}{\|\boldsymbol{v}_1\|} = \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\right) \pm O\left(\log(d)\delta_2^{(1)}\right).$$

将这两个界合并起来，即完成证明。 □

引理 *C.17* 的证明。 根据引理 C.16的证明，我们有

$$\begin{aligned}
\mathrm{Rad}(\dot{\boldsymbol{v}}_1) &= \frac{2C_\Gamma\bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\|\boldsymbol{x}\|\right\}\boldsymbol{v}_1 \\
&\quad \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\|\boldsymbol{v}_1\|\right) \pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right) \\
&= \Theta\left(\frac{\log d}{\sqrt{d}}\right)\bar{w}_2\boldsymbol{v}_1 \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\|\boldsymbol{v}_1\|\right) \pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|\right).
\end{aligned}$$

回想一下， $\delta_2^{(1)} \leq |\bar{w}_2|/d$。因此，

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}_1\|^2 &= \Theta\left(\frac{\log d}{\sqrt{d}}\right)\bar{w}_2\|\boldsymbol{v}_1\|^2 \pm O\left(\left(\delta_{1,I} + \sqrt{d}\delta_{1,R}^{(1)} + \sqrt{d}\delta_{1,R}^{(1)}\right)\bar{w}_2\|\boldsymbol{v}_1\|^2\right) \pm O\left(\log(d)\delta_2^{(1)}\|\boldsymbol{v}_1\|^2\right) \\
&= \Theta\left(\frac{\log d}{\sqrt{d}}\right)\bar{w}_2\|\boldsymbol{v}_1\|^2,
\end{aligned}$$

□

C.3.3 第二层的估计

现在，我们来界定第二层分散度的增长。读者可以先查看引理D.14的证明；该结果本质上是一个更简单的情形，因为其中无需处理投影。在引理D.14中，我们证明了分散度永远不会增长。这里的误差来自投影。

引理 C.18。 假设归纳假设 *C.2* 在时刻 t成立。那么我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}(\delta_2^{(1)})^2 \leq \frac{O(1)}{d^{2.5}}(\delta_2^{(1)})^2.$$

证明。令 $(v_2,\ r_2)$，$(v_2',\ r_2') \in \mu_2$ 并定义 $h_2(\boldsymbol{x}) = v_2 F(\boldsymbol{x}) + r_2$ 和 $h_2'(\boldsymbol{x}) = v_2' F(\boldsymbol{x}) + r_2'$。记作

$$\dot{v}_2 = \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))F(\boldsymbol{x})\} - \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\Pi_{R_{v_2}}\left[f(\boldsymbol{x})\sigma'(h_2(\boldsymbol{x}))F(\boldsymbol{x})\right]\right\} =: \mathrm{T}_1(\dot{v}_2) + \mathrm{T}_2(\dot{v}_2).$$

T_1 不依赖于 v_2。对于 T_2，，注意到

$$\Pi_{R_{v_2}}\left[f(\boldsymbol{x})\sigma'(h_2(\boldsymbol{x}))F(\boldsymbol{x})\right] = \Pi_{R_{v_2}/F(\boldsymbol{x})}\left[f(\boldsymbol{x})\right]\sigma'(h_2(\boldsymbol{x}))F(\boldsymbol{x}).$$

同样地，对于 $\dot{r}_2$，我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(r_2 - r_2')^2 &= -2\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\{f(\boldsymbol{x})(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x})))(r_2 - r_2')\}\\ &= -2\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\Pi_{R_{v_2}/F(\boldsymbol{x})}[f(\boldsymbol{x})](\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x})))(r_2 - r_2')\right\}\\ &\quad - 2\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\left(f(\boldsymbol{x}) - \Pi_{R_{v_2}/F(\boldsymbol{x})}[f(\boldsymbol{x})]\right)(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x})))(r_2 - r_2')\right\}.\end{aligned}$$

将这两个方程结合起来可得

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left((v_2 - v_2')^2 + (r_2 - r_2')^2\right)\\ =\ &-2\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\Pi_{R_{v_2}/F(\boldsymbol{x})}\left[f(\boldsymbol{x})\right](\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x})))\,(h_2(\boldsymbol{x}) - h_2'(\boldsymbol{x}))\right\}\\ &\quad - 2\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{\left(f(\boldsymbol{x}) - \Pi_{R_{v_2}/F(\boldsymbol{x})}[f(\boldsymbol{x})]\right)(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x})))(r_2 - r_2')\right\}.\end{aligned}$$

由于 σ' 是非递减的，第一项为非正。对于第二项，根据引理 C.11和引理 C.13，可将其界定为

$$\max_{\boldsymbol{x}:\mathrm{sgn}(h_2(\boldsymbol{x}))\ne\mathrm{sgn}(h_2'(\boldsymbol{x}))} f(\boldsymbol{x})\times\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|}\{|\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))|\}\times|r_2 - r_2'| \le O\left(\frac{\alpha(\delta_2^{(1)})^3}{|\bar{w}_2|}\right) \le \frac{O(1)}{d^{2.5}}(\delta_2^{(1)})^2.$$

□

D 阶段 2

阶段 2 的目标是使梯度流收敛到一个损失为 ε 的点。与阶段 1 类似，我们维持一组归纳假设。

归纳假设 D.1。定义 $T_2 :=$ 下确界$\{t \ge T_1 :\ \mathcal{L} = \varepsilon\}$。定义 $\delta_1^{(2)}$，L^2，δ_1，L^∞，$\delta_2^{(2)}$ a s

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta_{1,L^2}^{(2)} = \mathrm{ReLU}\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|\bar{F} - \|\cdot\|\right\|_{L^2}\right),\quad \frac{\mathrm{d}}{\mathrm{d}t}\delta_{1,L^\infty}^{(2)} = \mathrm{ReLU}\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|\bar{F}|_{\mathbb{S}^{d-1}} - 1\right\|_{L^\infty}\right),\quad \frac{\mathrm{d}}{\mathrm{d}t}\delta_2^{(2)} = 0,$$

其初始值满足[11]

$$\begin{aligned}\Theta\left(\frac{d^{17}}{\varepsilon}(\delta_{1,L^\infty}^{(2)})^2\right) \le \delta_{1,L^2}^{(2)} \le \Theta\left(\frac{\varepsilon}{d^6}\delta_{1,L^\infty}^{(2)}\right),\\ \delta_{1,L^2}^{(2)} \le O\left(\frac{\varepsilon^2}{d^7}\right),\quad \delta_{1,L^\infty}^{(2)}(T_1) \le O\left(\frac{\varepsilon}{d^{14}}\right),\quad \delta_2^{(2)} \le O\left(\frac{\varepsilon^2}{d^{10}}\right).\end{aligned}$$

对于任意 $t \in [T_1, T_2]$，如果满足以下条件，我们就称该归纳假设成立。

[11]正如我们在归纳假设脚注中提到的C.2，这些 δ被定义为相应误差的上界。这使得我们在选择其初始值时拥有一定的自由度。根据引理C.4，我们可以选择参数，使得阶段 2 开始时的误差任意小，从而确实满足这些条件。第一个条件要求 L^2 误差受到 L^∞ 误差的左右控制，乍看之下似乎有些奇怪。我们需要这一条件的唯一原因，是将一些二阶误差项并入一阶误差项。

(a) 第一层的误差。 $\left\|\bar{F}-\|\cdot\|\right\|_{L^2} \leq \delta_1^{(2)}$,$_{L^2}$ **和** $\left\|\bar{F}|_{\mathbb{S}^{d-1}}-1\right\|_{L^\infty} \leq \delta_1^{(2)}$,$_{L^\infty}$。 **(b) 第二层的扩散。** $\|(v_2, r_2) - (v_2', r_2')\| \leq \delta_2^{(2)}$ **对所有** (v_2, r_2),$(v_2', r_2') \in \mu_{2^0}$ **(c) 正则性条件。** $\bar{b}_2 \leq 1-\Theta(\sqrt{\varepsilon})$。 $\bar{w}_2\alpha \geq -1+\Theta(\sqrt{\varepsilon})$。 $|\bar{w}_2| \leq d$。 $|\bar{w}_2| \geq \Theta(1/d^3)$。 $\alpha \geq \Theta(1/d^{1.5})$。 **(d) 误差的界。** $\delta_1^{(2)}$,$_{L^\infty} = O(\delta_1^{(2)}$,$_{L^\infty}(T_1))$ **和** $\delta_1^{(2)}$,$_{L^2} = O(\delta_1^{(2)}$,$_{L^2}(T_1))$。

阶段 2 的主要引理如下。

引理 D.2 （阶段 2）。 归纳假设 *D.1* 在整个阶段 *2* 中均成立，且阶段 *2* 至多耗时 $O(d^3/\varepsilon)$。

本节其余部分安排如下。在第 D.1 节中，我们汇总了一些后续将用到的辅助结果。在第 D.2 节中，我们证明归纳假设 D.1 在整个阶段 2 中始终成立。（有关所用技术和一些约定的讨论，也请参见第 B.1 节。）随后，我们在第 D.3 节中推导收敛速率的下界。最后，我们在第 D.2 节中证明引理D.4。

D.1 辅助引理

D.1.1 F、 f 和 $\mathcal{L}$的动力学

回想一下，在阶段 2 中，我们可以忽略投影算子，因此神经元的动力学由下式给出

$$\begin{aligned}
\dot{\boldsymbol{v}}_1 &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{S(\boldsymbol{x})\left(\bar{v}_1\sigma(\boldsymbol{v}_1\cdot\boldsymbol{x}) + \|\boldsymbol{v}_1\|\,\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\boldsymbol{x}\right)\right\},\\
\dot{v}_2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\sigma'(v_2F(\boldsymbol{x}) + r_2)F(\boldsymbol{x})\right\},\\
\dot{r}_2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\sigma'(v_2F(\boldsymbol{x}) + r_2)\right\}.
\end{aligned}$$

现在，我们推导描述 α、 F 的动力学以及损失 $\mathcal{L}$ 的方程。

引理 D.3（α的动力学）。在阶段 *2* 中，我们有

$$\dot{\alpha} = \frac{4C_\Gamma}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}'}\left\{S(\boldsymbol{x}')F(\boldsymbol{x}')\right\}.$$

引理 D.4 （F的动力学）。在阶段 *2* 中，对每个固定的 $\boldsymbol{x}$，我们有

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}F(\boldsymbol{x}) = 4\mathop{\mathbb{E}}_{\boldsymbol{x}'}\left\{S(\boldsymbol{x}')\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x}')\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\right\}\right\}\\
+\mathop{\mathbb{E}}_{\boldsymbol{x}'}\left\{S(\boldsymbol{x}')\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\|\boldsymbol{w}_1\|^2\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x}')\sigma'(\boldsymbol{v}_1\cdot\boldsymbol{x})\left\langle(\boldsymbol{I}-\bar{\boldsymbol{v}}_1\bar{\boldsymbol{v}}_1^\top)\boldsymbol{x}', \boldsymbol{x}\right\rangle\right\}\right\}.
\end{aligned}$$

注意，在上述引理中，我们将 $\frac{\mathrm{d}}{\mathrm{d}t}F(\boldsymbol{x})$ 分解为两项，其中第一项对应于 $\boldsymbol{v}_1$ 的径向移动，第二项对应于切向移动。

引理 D.5 （$\mathcal{L}$的动力学）。定义 $W_2(\boldsymbol{x}) = \mathbb{E}_{w_2,\, b_2}\{\sigma'(w_2F(\boldsymbol{x}) + b_2)w_2\}$。在阶段 *2* 中，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L} = -\mathop{\mathbb{E}}_{w_2,b_2,\boldsymbol{w}_1}\|\nabla_{w_2,b_2,\boldsymbol{w}_1}\|^2,$$

其中

$$\nabla_{w_2,b_2,\boldsymbol{w}_1} := \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\begin{bmatrix}\sigma'(w_2F(\boldsymbol{x}) + b_2)F(\boldsymbol{x})\\ \sigma'(w_2F(\boldsymbol{x}) + b_2)\\ 2\bar{W}_2(\boldsymbol{x})\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\\ \|\boldsymbol{w}_1\|\,\bar{W}_2(\boldsymbol{x})\sigma'(\boldsymbol{w}_1\cdot\boldsymbol{x})(\boldsymbol{I}-\bar{\boldsymbol{w}}_1\bar{\boldsymbol{w}}_1^\top)\boldsymbol{x}\end{bmatrix}\right\}.$$

$\nabla_{w_2、\, b_2、\, \boldsymbol{w}_1}$ 分别对应 v_2、 r_2的移动、 $\boldsymbol{v}_1$ 的径向移动以及 $\boldsymbol{v}_1$的切向移动。

这三个引理的证明如下。

引理的证明*D.3*. 回顾 $\alpha := \frac{C_\Gamma}{\sqrt{d}} \mathbb{E}_{\boldsymbol{w}_1} \|\boldsymbol{w}_1\|^2$。因此，$\dot{\alpha} = \frac{2C_\Gamma}{\sqrt{d}} \mathbb{E}_{\boldsymbol{w}_1} \langle \boldsymbol{w}_1, \dot{\boldsymbol{w}}_1 \rangle$。我们 compute

$$\langle \dot{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle = \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ S(\boldsymbol{x}) \left(\sigma(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \ \langle \bar{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle + \|\boldsymbol{v}_1\| \, \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \ \langle \boldsymbol{x}, \boldsymbol{v}_1 \rangle \right) \} = 2 \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ S(\boldsymbol{x}) \|\boldsymbol{v}_1\| \, \sigma(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \} .$$

因此，

$$\dot{\alpha} = \frac{4C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ S(\boldsymbol{x}) \|\boldsymbol{w}_1\| \, \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \} \right\} = \frac{4C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ S(\boldsymbol{x}) \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \{ \|\boldsymbol{w}_1\| \, \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \} \right\} = \frac{4C_\Gamma}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ S(\boldsymbol{x}) F(\boldsymbol{x}) \} .$$

□

引理的证明*D.4*。首先，我们写出

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\boldsymbol{x}) = \frac{\mathrm{d}}{\mathrm{d}t} \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \|\boldsymbol{w}_1\|^2 \sigma(\bar{\boldsymbol{w}}_1 \cdot \boldsymbol{x}) \right\} = \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \left(\frac{\mathrm{d}}{\mathrm{d}t} \|\boldsymbol{w}_1\|^2 \right) \sigma(\bar{\boldsymbol{w}}_1 \cdot \boldsymbol{x}) \right\} + \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \|\boldsymbol{w}_1\|^2 \frac{\mathrm{d}}{\mathrm{d}t} \sigma(\bar{\boldsymbol{w}}_1 \cdot \boldsymbol{x}) \right\} .$$

根据引理 D.3的证明，第一项为 $4 \mathbb{E}_{\boldsymbol{x}'} \{ S(\boldsymbol{x}') \mathbb{E}_{\boldsymbol{w}_1} \{ \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}') \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \} \}$. 对于第二项，我们计算

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \sigma(\bar{\boldsymbol{v}}_1 \cdot \boldsymbol{x}) &= \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \left\langle (\boldsymbol{I} - \bar{\boldsymbol{v}}_1 \bar{\boldsymbol{v}}_1^\top) \frac{\dot{\boldsymbol{v}}_1}{\|\boldsymbol{v}_1\|}, \boldsymbol{x} \right\rangle \\
&= \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \left\langle \mathop{\mathbb{E}}_{\boldsymbol{x}'} \{ S(\boldsymbol{x}') \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}') (\boldsymbol{I} - \bar{\boldsymbol{v}}_1 \bar{\boldsymbol{v}}_1^\top) \boldsymbol{x}' \}, \boldsymbol{x} \right\rangle \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}'} \{ S(\boldsymbol{x}') \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}') \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \langle (\boldsymbol{I} - \bar{\boldsymbol{v}}_1 \bar{\boldsymbol{v}}_1^\top) \boldsymbol{x}', \boldsymbol{x} \rangle \} .
\end{aligned}$$

因此，第二项为

$$\mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \|\boldsymbol{w}_1\|^2 \frac{\mathrm{d}}{\mathrm{d}t} \sigma(\bar{\boldsymbol{w}}_1 \cdot \boldsymbol{x}) \right\} = \mathop{\mathbb{E}}_{\boldsymbol{x}'} \left\{ S(\boldsymbol{x}') \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \|\boldsymbol{w}_1\|^2 \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}') \sigma'(\boldsymbol{v}_1 \cdot \boldsymbol{x}) \langle (\boldsymbol{I} - \bar{\boldsymbol{v}}_1 \bar{\boldsymbol{v}}_1^\top) \boldsymbol{x}', \boldsymbol{x} \rangle \right\} \right\} .$$

将这些结果合并，即完成证明。 □

引理的证明*D.5*。首先，我们写出

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) &= \mathop{\mathbb{E}}_{w_2, b_2} \{ \sigma'(w_2 F(\boldsymbol{x}) + b_2) \dot{w}_2 F(\boldsymbol{x}) \} + \mathop{\mathbb{E}}_{w_2, b_2} \left\{ \sigma'(w_2 F(\boldsymbol{x}) + b_2) \dot{b}_2 \right\} + \bar{W}_2(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}t} F(\boldsymbol{x}) \\
&=: \mathrm{T}_1 \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) + \mathrm{T}_2 \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) + \mathrm{T}_3 \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) .
\end{aligned}$$

注意，$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{L} = -\sum_{i=1}^3 \mathbb{E}_{\boldsymbol{x}} \{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \mathrm{T}_i \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) \}$. 现在我们分别计算这三项。我们有

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \mathrm{T}_1 \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) \right\} &= \mathop{\mathbb{E}}_{w_2, b_2} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma'(w_2 F(\boldsymbol{x}) + b_2) F(\boldsymbol{x}) \} \dot{w}_2 \right\} \\
&= \mathop{\mathbb{E}}_{w_2, b_2} \left\{ \left(\mathop{\mathbb{E}}_{\boldsymbol{x}} \{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma'(w_2 F(\boldsymbol{x}) + b_2) F(\boldsymbol{x}) \} \right)^2 \right\} , \\
\mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \mathrm{T}_2 \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) \right\} &= \mathop{\mathbb{E}}_{w_2, b_2} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma'(w_2 F(\boldsymbol{x}) + b_2) \} \dot{b}_2 \right\} \\
&= \mathop{\mathbb{E}}_{w_2, b_2} \left\{ \left(\mathop{\mathbb{E}}_{\boldsymbol{x}} \{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \sigma'(w_2 F(\boldsymbol{x}) + b_2) \} \right)^2 \right\} .
\end{aligned}$$

同时，对于 T_3, 根据引理D.4，我们有

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \mathrm{T}_3 \left(\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}) \right) \right\} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \bar{W}_2(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}t} F(\boldsymbol{x}) \right\} \\
&= 4 \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \bar{W}_2(\boldsymbol{x}) \mathop{\mathbb{E}}_{\boldsymbol{x}'} \left\{ S(\boldsymbol{x}') \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \{ \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}') \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \} \right\} \right\} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - f(\boldsymbol{x})) \bar{W}_2(\boldsymbol{x}) \mathop{\mathbb{E}}_{\boldsymbol{x}'} \left\{ S(\boldsymbol{x}') \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \|\boldsymbol{w}_1\|^2 \sigma'(\boldsymbol{w}_1 \cdot \boldsymbol{x}') \sigma'(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \langle (\boldsymbol{I} - \bar{\boldsymbol{w}}_1 \bar{\boldsymbol{w}}_1^\top) \boldsymbol{x}', \boldsymbol{x} \rangle \right\} \right\} \right\} \\
&= 4 \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \left(\mathop{\mathbb{E}}_{\boldsymbol{x}} \{ S(\boldsymbol{x}) \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \} \right)^2 \right\} + \mathop{\mathbb{E}}_{\boldsymbol{w}_1} \left\{ \left\| \mathop{\mathbb{E}}_{\boldsymbol{x}} \{ S(\boldsymbol{x}) \|\boldsymbol{w}_1\| \, \sigma'(\boldsymbol{w}_1 \cdot \boldsymbol{x}) (\boldsymbol{I} - \bar{\boldsymbol{w}}_1 \bar{\boldsymbol{w}}_1^\top) \boldsymbol{x} \} \right\|^2 \right\} .
\end{aligned}$$

将这些结果结合起来，我们便完成了证明。 □

D.1.2 误差相关估计

我们在此汇总一些误差相关估计。其中大多数已在阶段 1 中证明，除此之外，这里我们使用 $|\bar{w}_2| \geq \Theta(1/d^3)$ 将 $(|\bar{w}_2|^{-1}+1)$ 替换为 $O(d^3)$。为便于查阅，我们在此重复该陈述。

引理 D.6。 假设归纳假设 *D.1* 在时刻 t 成立。如果 $v_2 F(\boldsymbol{x}) + r_2 = 0$ 对于某个$(v_2,\ r_2) \in \mu_2$ 成立，那么 $v_2' F(\boldsymbol{x}) + r_2' \leq O\left(d^3 \delta_2^{(2)}\right)$ 对于所有 $(v_2',\ r_2') \in \mu_2$均成立。

证明。 参见引理C.11。 □

引理 D.7。 假设归纳假设 *D.1* 在时刻 t成立。那么，对于任意 $(v_2, r_2) \in \mu_2$ 和 $\boldsymbol{x} \in \mathbb{R}^d$，都有

$$f_*(\boldsymbol{x})\sigma'(v_2 F(\boldsymbol{x}) + r_2) = f_*(\boldsymbol{x}) \quad \textit{and} \quad f(\boldsymbol{x})\sigma'(v_2 F(\boldsymbol{x}) + r_2) = f(\boldsymbol{x}) \pm O\left(d^3 \delta_2^{(2)}\right).$$

作为推论，我们有

$$f(\boldsymbol{x}) = \sigma(v_2 F(\boldsymbol{x}) + r_2) \pm O\left(d^3 \delta_2^{(2)}\right),$$
$$f(\boldsymbol{x}) = \sigma(\bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2) \pm O\left(d^3 \delta_2^{(2)}\right).$$

证明。 参见引理C.12。 □

引理 D.8。 假设归纳假设 *D.1* 在时刻 t成立。那么有 $\left\|f - \tilde{f}\right\|_{L^2} \leq O\left(|\bar{w}_2 \alpha| \delta_{1,L^2}^{(2)}\right)$.

证明。 由于 σ 是 1-Lipschitz 的，我们有

$$|f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| = \left| \mathop{\mathbb{E}}_{w_2,b_2} \left\{ \sigma(w_2 F(\boldsymbol{x}) + b_2) - \sigma(w_2 \tilde{F}(\boldsymbol{x}) + b_2) \right\} \right| \leq O\left(|\bar{w}_2| |F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x})|\right).$$

因此，

$$\left\|f - \tilde{f}\right\|_{L^2}^2 \leq O\left(\bar{w}_2^2 \alpha^2 \left\|\bar{F} - \|\cdot\|\right\|_{L^2}^2\right) \leq O\left(\bar{w}_2^2 \alpha^2 (\delta_{1,L^2}^{(2)})^2\right).$$

□

D.2 维持归纳假设

在本节中，我们将证明归纳假设 D.1 在阶段 2 的整个过程中均成立。关于本文所用技术的讨论和惯例，请参见第 B.1节。

D.2.1 第一层的误差

回顾一下，我们可以将损失分解为

$$\mathcal{L} = \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))^2 \right\} + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \right\} + \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\{ (f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))(\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x})) \right\}$$
$$=: \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3.$$

正如正文中所讨论的，目标是证明 $\mathcal{L}_2 \approx \frac{\bar{w}_2^2}{2} \mathbb{E}\left\{ (\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2 \right\}$ 和$\mathcal{L}_3 \approx 0$，从而 $\mathcal{L}$ 可以分解为两项，其中第一项刻画目标函数 f_* 与无限宽网络 f之间的差异，第二项则衡量 $\tilde{F}$ 与 F 之间的近似误差。我们将在引理 D.中证明，正如预期的那样，不会影响 F。估计 $\mathcal{L}_2$ 和 $\mathcal{L}_3$ 的梯度稍微复杂一些。首先，我们需要引入输入空间上的如下划分。

引理 D.9. 定义

$$R_1 := \left\{R > 0 \ : \ \forall (v_2, r_2) \in \mu_2, \boldsymbol{x} \in R\mathbb{S}^{d-1}, \ v_2 F(\boldsymbol{x}) + r_2 > 0\right\},$$
$$R_2 := \left\{R > 0 \ : \ \exists (v_2, r_2) \in \mu_2, \boldsymbol{x} \in R\mathbb{S}^{d-1}, \ v_2 F(\boldsymbol{x}) + r_2 > 0\right\}.$$

然后，我们将输入空间划分为

$$X_1 := \{\|\boldsymbol{x}\| \le R_1\}, \qquad X_2 := \{R_1 \le \|\boldsymbol{x}\| \le R_2\}, \qquad X_3 := \{R_2 \le \|\boldsymbol{x}\|\}.$$

换言之， X_1 是所有第二层神经元均被激活的最大球对称集合，而 $X_1 \cup X_2$ 是至少有一个第二层神经元被激活的最大球对称集合。假设归纳假设*D.1* 在时刻 t成立。那么以下结论成立。

(a) f_* *vanishes on* $X_2 \cup X_3$*, i.e.,* $R_1 \ge 1$*,* f *vanishes on* X_3*, and* $R_3 \le O(1/|\bar{w}_2|/\alpha)$*.*

(b) $R_2 - R_1 \le O\left(\delta_{1,L^\infty}^{(2)} + d\delta_2^{(2)}\right) \frac{1}{|\bar{w}_2|\alpha} =: \delta_{X_2}^{(2)}$*. As a corollary, we have* $\mathbb{P}[X_2] \le O\left(\delta_{X_2}^{(2)}\right)$*.*

(c) $f \le O\left(\delta_{X_2}^{(2)}\right)$ *on* X_2*.*

上述引理意味着 $\mathcal{L}_2 \approx \frac{1}{2}\mathbb{E}_{X_1}\left\{(\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right\} = \frac{\bar{w}_2^2}{2}\mathbb{E}_{X_1}\left\{(\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2\right\}$ 和 $\mathcal{L}_3 \approx \mathbb{E}_{X_1}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))(\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))\right\} = \bar{w}_2\mathbb{E}_{X_1}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))(\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))\right\} = 0$。我们将在下面的引理中正式证明这一近似。

引理 D.10 ($\mathcal{L}_2$ 和 $\mathcal{L}_3$的梯度)。 假设归纳假设 *D.1* 在时刻 t成立。那么，对于每个 $\boldsymbol{v}_1 \in \mu_1$，我们有

$$\nabla_{\boldsymbol{v}_1}\mathcal{L}_2 = \nabla_{\boldsymbol{v}_1}\left(\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{X_1}\left\{(\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2\right\}\right) \pm O\left(\left(\delta_{X_2}^{(2)}\right)^2 \frac{1}{\alpha}\right)\|\boldsymbol{v}_1\|,$$
$$\nabla_{\boldsymbol{v}_1}\mathcal{L}_3 = \pm O\left(\left(\delta_{X_2}^{(2)}\right)^2 \frac{1}{\alpha}\right)\|\boldsymbol{v}_1\|.$$

现在，我们准备推导描述 F动力学的方程。注意，这个引理意味着，至少近似而言， F的动力学只取决于 $\mathcal{L}_2$。

引理 D.11 (F的动力学)。 假设归纳假设 *D.1* 在时刻 t成立。然后，对于每个固定的 $\boldsymbol{x}$，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x}) = -\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\bar{F}(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\tilde{F}(\boldsymbol{x}') - F(\boldsymbol{x}'))^2\right\}\right\rangle\right\} \pm O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2 \frac{1}{\alpha}\right)\|\boldsymbol{x}\|.$$

接下来，我们证明， $\mathrm{d}\,\mathrm{d}t\,\bar{F}(\boldsymbol{x})$ 中的信号项只能降低 L^2 误差。直观上这是成立的，毕竟， $\mathcal{L}_2$ 就是（经过重缩放的） L^2 误差。因此， L^2 误差几乎不会增长。

引理 D.12 (L^2 近似误差)。 假设归纳假设 *D.1* 在时刻 t成立，那么我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\bar{F} - \|\cdot\|\right\|_{L^2}^2 \le O\left(d^5\delta_{1,L^2}^{(2)}\left(\delta_{X_2}^{(2)}\right)^2\right).$$

最后，我们证明这种变化 $\bar{F}|_{\mathbb{S}^{d-1}}$ 取决于 L^2 误差。因此，只要 L^2 误差足够小， L^∞ 误差就不会增长得太快。

引理 D.13 (L^∞ 近似误差)。 假设归纳假设 *D.1* 在时刻 t成立。那么，对任意 $\bar{\boldsymbol{x}} \in \mathbb{S}^{d-1}$，有

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\bar{\boldsymbol{x}})\right| \le O\left(d^3\delta_{1,L^2}^{(2)} + d^2\left(\delta_{X_2}^{(2)}\right)^2\right).$$

这些引理的证明如下。

引理的证明*D.9*。

（a）这可直接由划分的构造和归纳假设推出 D.1。

(b) 首先，我们写出

$$F(\boldsymbol{x}) = \alpha \|\boldsymbol{x}\| + \alpha \|\boldsymbol{x}\| (\bar{F}(\bar{\boldsymbol{x}}) - 1) = \alpha \|\boldsymbol{x}\| \pm \alpha \|\boldsymbol{x}\| \delta_{1,L^\infty}^{(2)} = \alpha \|\boldsymbol{x}\| \pm O\left(\frac{\delta_{1,L^\infty}^{(2)}}{|\bar{w}_2|}\right),$$

其中，最后一个等号是因为 f 在 $\{\|\boldsymbol{x}\| \geq \Omega(-b_2/(\alpha|\bar{w}_2|))\}$ 上消失。类似地，对于任意 $(v_2, r_2) \in \mu_2$，我们有

$$\begin{aligned} v_2 F(\boldsymbol{x}) + r_2 = \left\langle \begin{bmatrix} v_2 \\ r_2 \end{bmatrix}, \begin{bmatrix} F(\boldsymbol{x}) \\ 1 \end{bmatrix} \right\rangle &= \left\langle \begin{bmatrix} \bar{w}_2 \\ \bar{b}_2 \end{bmatrix}, \begin{bmatrix} F(\boldsymbol{x}) \\ 1 \end{bmatrix} \right\rangle + \left\langle \begin{bmatrix} v_2 \\ r_2 \end{bmatrix} - \begin{bmatrix} \bar{w}_2 \\ \bar{b}_2 \end{bmatrix}, \begin{bmatrix} F(\boldsymbol{x}) \\ 1 \end{bmatrix} \right\rangle \\ &= \bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2 \pm O\left(\delta_2^{(2)} \sqrt{\alpha^2 \|\boldsymbol{x}\|^2 + 1}\right) \\ &= \bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2 \pm O\left(d^3 \delta_2^{(2)}\right). \end{aligned}$$

因此，对于任意 $R > 0$ 和 $\boldsymbol{x} \in R\mathbb{S}^{d-1}$，我们有

$$v_2 F(\boldsymbol{x}) + r_2 = \bar{w}_2 \alpha \|\boldsymbol{x}\| + \bar{b}_2 \underbrace{\pm O\left(\delta_{1,L^\infty}^{(2)}\right) \pm O\left(d^3 \delta_2^{(2)}\right)}_{=: \delta_{\text{Tmp}}}.$$

因此，

$$\begin{aligned} v_2 F(\boldsymbol{x}) + r_2 > 0, \quad \text{if } \|\boldsymbol{x}\| < \frac{\bar{b}_2 - \delta_{\text{Tmp}}}{-\bar{w}_2 \alpha} = \tilde{R} - \frac{\delta_{\text{Tmp}}}{-\bar{w}_2 \alpha}, \\ v_2 F(\boldsymbol{x}) + r_2 < 0, \quad \text{if } \|\boldsymbol{x}\| > \frac{\bar{b}_2 + \delta_{\text{Tmp}}}{-\bar{w}_2 \alpha} = \tilde{R} + \frac{\delta_{\text{Tmp}}}{-\bar{w}_2 \alpha}. \end{aligned}$$

换言之，$R_1 \geq \tilde{R} - \frac{\delta_{\text{Tmp}}}{-\bar{w}_2 \alpha}$ 和 $R_2 \leq \tilde{R} + \frac{\delta_{\text{Tmp}}}{-\bar{w}_2 \alpha}$。因此，

$$R_2 - R_1 \leq \frac{\delta_{\text{Tmp}}}{-\bar{w}_2 \alpha} \leq O\left(\delta_{1,L^\infty}^{(2)} + O\left(d^3 \delta_2^{(2)}\right)\right) d^{4.5} = \delta_{X_2}.$$

要完成证明，只需调用引理B.1。

(c) 注意，根据 R_2 的定义，对于任意 $\boldsymbol{x}_0 \in R_2 \mathbb{S}^{d-1}$，我们有 $f(\boldsymbol{x}_0) = 0$。因此，对于任意 $\boldsymbol{x} \in X_2$，存在某个 $\boldsymbol{x}_0$，使得 $f(\boldsymbol{x}_0) = 0$ 且 $\|\boldsymbol{x} - \boldsymbol{x}_0\| \leq R_2 - R_1 = \delta_{X_2}^{(2)}$。由于 f 是 $O(1)$-利普希茨连续的，对于任意 $\boldsymbol{x} \in X_2$，我们有 $f(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}_0) \leq O(\delta_{X_2}^{(2)})$。

□

引理*D.10*的证明。由于 f_* 和 f 在 X_3上均为零，因此只需考虑 X_1 和 X_2。回顾一下，所有第二层神经元在 X_1上均被激活。因此，

$$\mathcal{L}_2\Big|_{X_1} := \frac{1}{2} \mathop{\mathbb{E}}_{X_1} \left\{(\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right\} = \frac{\bar{w}_2^2}{2} \mathop{\mathbb{E}}_{X_1} \left\{(\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2\right\},$$

$$\mathcal{L}_3\Big|_{X_1} := \mathop{\mathbb{E}}_{X_1} \left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))(\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))\right\} = \bar{w}_2 \mathop{\mathbb{E}}_{X_1} \left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))(\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x}))\right\} = 0,$$

其中，最后一个等式由推论 B.4 得出。现在，我们来界定 X_2的影响。注意~，$\nabla_{\boldsymbol{v}_1} f(\boldsymbol{x})$和 $\nabla_{\boldsymbol{v}_1} f(\boldsymbol{x})$均由 $O(|\bar{w}_2| \|\boldsymbol{v}_1\| \|\boldsymbol{x}\|)$界定。回顾引理 D.9，可知$f \leq O(\delta_{X_2}^{(2)})$在 X_2上成立，且 $\mathbb{P}[X_2] \leq O(\delta_{X_2}^{(2)})$。因此，

$$\left\|\nabla_{\boldsymbol{v}_1} \mathcal{L}_2\Big|_{X_2}\right\| \leq O(\delta_{X_2}^{(2)}) \times O\left(\delta_{X_2}^{(2)}\right) \times O\left(|\bar{w}_2| \|\boldsymbol{v}_1\| \frac{1}{|\bar{w}_2|\alpha}\right) \leq O\left(\left(\delta_{X_2}^{(2)}\right)^2 \frac{1}{\alpha}\right) \|\boldsymbol{v}_1\|.$$

对于 $\nabla_{\boldsymbol{v}_1} \mathcal{L}_3|_{X_2}$ 的证明是相同的。 □

引理的证明 *D.11*。 对于固定的 $\boldsymbol{x} \in \mathbb{R}^d$，我们写作

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x}) = \frac{\frac{\mathrm{d}}{\mathrm{d}t}F(\boldsymbol{x})}{\alpha} - \bar{F}(\boldsymbol{x})\frac{\dot{\alpha}}{\alpha} = -\frac{1}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\langle\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathcal{L}\rangle + \bar{F}(\boldsymbol{x})\frac{1}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\langle\nabla_{\boldsymbol{w}_1}\alpha, \nabla_{\boldsymbol{w}_1}\mathcal{L}\rangle.$$

首先，我们考虑 $\mathcal{L}_1$。对于每个 $\boldsymbol{v}_1 \in \mu_1$，我们有

$$\begin{aligned}\nabla_{\boldsymbol{v}_1}\mathcal{L}_1 &= -\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\nabla_{\boldsymbol{v}_1}\tilde{f}(\boldsymbol{x})\right\}\\ &= -\frac{2C_\Gamma}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))\mathop{\mathbb{E}}_{w_2,b_2}\{\sigma(w_2\alpha\|\boldsymbol{x}\| + b_2)w_2\}\right\}\boldsymbol{v}_1 =: C_{\mathtt{Tmp},1}\boldsymbol{v}_1.\end{aligned}$$

同时，注意到

$$\langle\nabla_{\boldsymbol{v}_1}F(\boldsymbol{x}), \boldsymbol{v}_1\rangle = \left\langle\nabla_{\boldsymbol{v}_1}(\|\boldsymbol{v}_1\|^2\sigma(\bar{\boldsymbol{v}}_1\cdot\boldsymbol{x})), \boldsymbol{v}_1\right\rangle = \left\langle\nabla_{\boldsymbol{v}_1}(\|\boldsymbol{v}_1\|^2)\sigma(\bar{\boldsymbol{v}}_1\cdot\boldsymbol{x}), \boldsymbol{v}_1\right\rangle = 2\|\boldsymbol{v}_1\|^2\sigma(\bar{\boldsymbol{v}}_1\cdot\boldsymbol{x}),$$

$$\langle\nabla_{\boldsymbol{v}_1}\alpha, \boldsymbol{v}_1\rangle = \frac{C_\Gamma}{\sqrt{d}}\left\langle\nabla_{\boldsymbol{v}_1}\|\boldsymbol{v}_1\|^2, \boldsymbol{v}_1\right\rangle = \frac{2C_\Gamma}{\sqrt{d}}\|\boldsymbol{v}_1\|^2.$$

因此，

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x})\Big|_{\mathcal{L}_1} &:= -\frac{1}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\langle\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathcal{L}_1\rangle + \bar{F}(\boldsymbol{x})\frac{1}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\langle\nabla_{\boldsymbol{w}_1}\alpha, \nabla_{\boldsymbol{w}_1}\mathcal{L}_1\rangle\\ &= -C_{\mathtt{Tmp},1}\frac{2}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\|\boldsymbol{w}_1\|^2\sigma(\bar{\boldsymbol{w}}_1\cdot\boldsymbol{x})\right\} + C_{\mathtt{Tmp},1}\bar{F}(\boldsymbol{x})\frac{1}{\alpha}\frac{2C_\Gamma}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\|\boldsymbol{w}_1\|^2\\ &= -C_{\mathtt{Tmp},1}\frac{2}{\alpha}F(\boldsymbol{x}) + 2C_{\mathtt{Tmp},1}\bar{F}(\boldsymbol{x})\\ &= 0.\end{aligned}$$

也就是说， $\mathcal{L}_1$ 不会影响 F。现在我们考虑 $\mathcal{L}_2$。根据引理D.10，我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x})\Big|_{\mathcal{L}_2} &:= -\frac{1}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\langle\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathcal{L}_2\rangle + \bar{F}(\boldsymbol{x})\frac{1}{\alpha}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\langle\nabla_{\boldsymbol{w}_1}\alpha, \nabla_{\boldsymbol{w}_1}\mathcal{L}_2\rangle\\ &= -\frac{1}{\alpha}\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\tilde{F}(\boldsymbol{x}') - F(\boldsymbol{x}'))^2\right\}\right\rangle\right\}\\ &\quad + \frac{1}{\alpha}\frac{\bar{w}_2^2}{2}\bar{F}(\boldsymbol{x})\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\alpha, \mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\tilde{F}(\boldsymbol{x}') - F(\boldsymbol{x}'))^2\right\}\right\rangle\right\}\\ &\quad \pm O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2\frac{1}{\alpha}\right)\|\boldsymbol{x}\|.\end{aligned}$$

注意，我们可以将第一项中的 $\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x})$ 重写为 $(\nabla_{\boldsymbol{w}_1}\alpha)F(\boldsymbol{x}) + \alpha\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x})$，这样其中一部分就会与第二项相消。然后，我们得到

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x})\Big|_{\mathcal{L}_2} = -\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\bar{F}(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\tilde{F}(\boldsymbol{x}') - F(\boldsymbol{x}'))^2\right\}\right\rangle\right\} \pm O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2\frac{1}{\alpha}\right)\|\boldsymbol{x}\|.$$

对于 $\mathcal{L}_3$，我们可以直接将其并入 $\mathrm{d}\,\mathrm{d}t\,\bar{F}(\boldsymbol{x})|_{\mathcal{L}_2}$ 的误差项中。 □

引理证明*D.12*。 根据引理D.11，我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\bar{F} - \|\cdot\|\right\|_{L^2}^2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\bar{F}(\boldsymbol{x}) - \|\boldsymbol{x}\|)\frac{\mathrm{d}}{\mathrm{d}t}F(\boldsymbol{x})\right\}\\ &= -\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\bar{F}(\boldsymbol{x}) - \|\boldsymbol{x}\|)\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\bar{F}(\boldsymbol{x}), \nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\tilde{F}(\boldsymbol{x}') - F(\boldsymbol{x}'))^2\right\}\right\rangle\right\}\right\}\\ &\quad \pm \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\bar{F}(\boldsymbol{x}) - \|\boldsymbol{x}\|)O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2\frac{1}{\alpha}\right)\|\boldsymbol{x}\|\right\}.\end{aligned}$$

第二项可以界定为 $O\left(\delta_1^{(2)}\cdot L^2\left(\delta_{X_2}^{(2)}\right)^2 d^5\right)$。第一项等于

$$\mathtt{Tmp} := -\frac{\bar{w}_2^2}{4}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\bar{F}(\boldsymbol{x}) - \|\boldsymbol{x}\|)^2\right\}, \nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\alpha\|\boldsymbol{x}'\| - F(\boldsymbol{x}'))^2\right\}\right\rangle\right\}.$$

为完成证明，只需证明该值为负。对于每个 $\boldsymbol{w}_1$，我们有

$$\nabla_{\boldsymbol{w}_1} \mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\alpha\|\boldsymbol{x}'\|-F(\boldsymbol{x}'))^2\right\}=\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\bar{F}(\boldsymbol{x}')-\|\boldsymbol{x}'\|)^2\right\}\nabla_{\boldsymbol{w}_1}\alpha^2+\alpha^2\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{\nabla_{\boldsymbol{w}_1}(\bar{F}(\boldsymbol{x}')-\|\boldsymbol{x}'\|)^2\right\}.$$

由于 $\boldsymbol{x}$ 的分布是球对称的，$\mathbb{E}_{\boldsymbol{x}'\in X_1}\left\{\nabla_{\boldsymbol{w}_1}(\bar{F}(\boldsymbol{x}')-\|\boldsymbol{x}'\|)^2\right\}$ 和 $\mathbb{E}_{\boldsymbol{x}}\left\{\nabla_{\boldsymbol{w}_1}(\bar{F}(\boldsymbol{x})-\|\boldsymbol{x}\|)^2\right\}$ 具有相同的方向。因此，

$$\begin{aligned}\texttt{Tmp}&\le-\frac{\bar{w}_2^2}{4}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\bar{F}(\boldsymbol{x})-\|\boldsymbol{x}\|)^2\right\},\nabla_{\boldsymbol{w}_1}\alpha^2\right\rangle\right\}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\bar{F}(\boldsymbol{x}')-\|\boldsymbol{x}'\|)^2\right\}\\&=-\frac{C_\Gamma}{\sqrt{d}}\bar{w}_2^2\alpha\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\bar{F}(\boldsymbol{x}')-\|\boldsymbol{x}'\|)^2\right\}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}(\bar{F}(\boldsymbol{x})-\|\boldsymbol{x}\|)^2,\boldsymbol{w}_1\right\rangle\right\}\right\}.\end{aligned}$$

接下来，我们计算

$$\begin{aligned}\left\langle\nabla_{\boldsymbol{w}_1}(\bar{F}(\boldsymbol{x})-\|\boldsymbol{x}\|)^2,\boldsymbol{w}_1\right\rangle&=2(\bar{F}(\boldsymbol{x})-\|\boldsymbol{x}\|)\left\langle\frac{\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x})}{\alpha}-\bar{F}(\boldsymbol{x})\frac{\nabla_{\boldsymbol{w}_1}\alpha}{\alpha},\boldsymbol{w}_1\right\rangle\\&=2(\bar{F}(\boldsymbol{x})-\|\boldsymbol{x}\|)\left(\frac{2\|\boldsymbol{w}_1\|^2\sigma(\bar{\boldsymbol{w}}_1\cdot\boldsymbol{x})}{\alpha}-\bar{F}(\boldsymbol{x})\frac{1}{\alpha}\frac{2C_\Gamma}{\sqrt{d}}\|\boldsymbol{w}_1\|^2\right).\end{aligned}$$

对 $\boldsymbol{w}_1$ 取期望，可知其值为 0。因此，$\texttt{Tmp}\le 0$。 □

引理的证明 *D.13*。 回顾引理D.11 可知

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x})=-\frac{\bar{w}_2^2}{2}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\left\langle\nabla_{\boldsymbol{w}_1}\bar{F}(\boldsymbol{x}),\nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{(\tilde{F}(\boldsymbol{x}')-F(\boldsymbol{x}'))^2\right\}\right\rangle\right\}\pm O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2\frac{1}{\alpha}\right)\|\boldsymbol{x}\|.$$

对于第一项，我们有

$$\begin{aligned}\left\|\nabla_{\boldsymbol{w}_1}\bar{F}(\boldsymbol{x})\right\|&\le\left\|\frac{\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x})}{\alpha}\right\|+\left\|\bar{F}(\boldsymbol{x})\frac{\dot{\alpha}}{\alpha}\right\|\le O\left(\frac{\|\boldsymbol{w}_1\|\|\boldsymbol{x}\|}{\alpha}\right),\\\left\|\nabla_{\boldsymbol{w}_1}\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{\left(\tilde{F}(\boldsymbol{x}')-F(\boldsymbol{x}')\right)^2\right\}\right\|&\le\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{\left|\tilde{F}(\boldsymbol{x}')-F(\boldsymbol{x}')\right|\left(\left\|\nabla_{\boldsymbol{w}_1}\tilde{F}(\boldsymbol{x}')\right\|+\|\nabla_{\boldsymbol{w}_1}F(\boldsymbol{x}')\|\right)\right\}\\&\le O(1)\mathop{\mathbb{E}}_{\boldsymbol{x}'\in X_1}\left\{\left|\tilde{F}(\boldsymbol{x}')-F(\boldsymbol{x}')\right|\|\boldsymbol{x}'\|\right\}\|\boldsymbol{w}_1\|\\&\le O\left(\delta_{1,L^2}^{(2)}\frac{1}{\sqrt{\alpha|\bar{w}_2|}}\|\boldsymbol{w}_1\|\right).\end{aligned}$$

因此，

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(\boldsymbol{x})\right|&\le O\left(\bar{w}_2^2\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\{\frac{\|\boldsymbol{w}_1\|\|\boldsymbol{x}\|}{\alpha}\delta_{1,L^2}^{(2)}\frac{1}{\sqrt{\alpha|\bar{w}_2|}}\|\boldsymbol{w}_1\|\right\}\right)+O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2\frac{1}{\alpha}\right)\|\boldsymbol{x}\|\\&\le O\left(\frac{\sqrt{d}|\bar{w}_2|^{1.5}}{\sqrt{\alpha}}\delta_{1,L^2}^{(2)}\right)\|\boldsymbol{x}\|+O\left(\sqrt{d}\left(\delta_{X_2}^{(2)}\right)^2\frac{1}{\alpha}\right)\|\boldsymbol{x}\|\\&\le O\left(d^3\delta_{1,L^2}^{(2)}+d^2\left(\delta_{X_2}^{(2)}\right)^2\right)\|\boldsymbol{x}\|.\end{aligned}$$

□

D.2.2 第二层的分散度

引理 D.14。 假设归纳假设 *D.1* 在时刻 t 成立。那么对于任意$(v_2,r_2),(v_2',r_2')\in\mu_2$，$\frac{\mathrm{d}}{\mathrm{d}t}\|(v_2,r_2)-(v_2',r_2')\|^2\le 0$。换言之，第二层的分散度永远不会增大。

证明。 设 (v_2, r_2),$(v_2', r_2') \in \mu_2$ 是两个第二层神经元。为方便记号，我们定义 $h_2(\boldsymbol{x}) = v_2 F(\boldsymbol{x}) + r_2$ 以及 $h_2'(\boldsymbol{x}) = v_2' F(\boldsymbol{x}) + r_2'$。我们有

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left((v_2 - v_2')^2 + (r_2 - r_2')^2\right) \\
&= (v_2 - v_2')\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))F(\boldsymbol{x})\left(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))\right)\right\} \\
&\qquad + (r_2 - r_2')\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\left(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))\right)\right\} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\left(h_2(\boldsymbol{x}) - h_2'(\boldsymbol{x})\right)\left(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))\right)\right\}.
\end{aligned}$$

根据引理 D.9，$\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x})) = 0$ 对所有 $\boldsymbol{x}$ 满足 $\|\boldsymbol{x}\| \le 1$的情况均成立。因此，

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left((v_2 - v_2')^2 + (r_2 - r_2')^2\right) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}:\|\boldsymbol{x}\|>1}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\left(h_2(\boldsymbol{x}) - h_2'(\boldsymbol{x})\right)\left(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))\right)\right\} \\
&= -\mathop{\mathbb{E}}_{\boldsymbol{x}:\|\boldsymbol{x}\|>1}\left\{f(\boldsymbol{x})\left(h_2(\boldsymbol{x}) - h_2'(\boldsymbol{x})\right)\left(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))\right)\right\}.
\end{aligned}$$

注意，$f \ge 0$ 并且，由于 σ' 是非递减的，$(h_2(\boldsymbol{x}) - h_2'(\boldsymbol{x}))(\sigma'(h_2(\boldsymbol{x})) - \sigma'(h_2'(\boldsymbol{x}))) \ge 0$。因此，$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left((v_2 - v_2')^2 + (r_2 - r_2')^2\right) \le 0$。 □

D.2.3 正则性条件

如前所述，我们主要通过连续性论证来维持正则性条件，因此问题可以归结为估计边界上的导数。举例来说，假设 $b_2 = 1 - \delta$ 对于某个较小的 $\delta > 0$。那么根据引理 D.15，该引理使用 和 $1 - b_2$ 对损失给出 $-1 - \bar{w}_2\alpha$上界，我们知道 $|-1 - \bar{w}_2\alpha|$ 必须足够大，否则就会有 $\mathcal{L} < \varepsilon$。接着，我们可以利用 $|-1 - \bar{w}_2\alpha|$ 较大这一事实来估计导数。其他正则性条件的证明与此类似，唯一的例外是 $|\bar{w}_2|$，其证明思路与第一层误差的证明相同。

引理 D.15. 假设归纳假设 *D.1* 在时刻成立 t。那么我们有

$$\mathcal{L} \le O\left((1 - \bar{b}_2)^2 + \frac{(-1 - \bar{w}_2\alpha)^2}{\bar{w}_2^2\alpha^2} + \left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right)^2\right).$$

引理 D.16. 假设归纳假设 *D.1* 在时刻 t 和 $b_2 = 1 - \Theta(\sqrt{\varepsilon})$成立。那么，$\frac{\mathrm{d}}{\mathrm{d}t}\bar{b}_2 < 0$。

引理 D.17. 假设归纳假设 *D.1* 在时刻 t 和 $\bar{w}_2\alpha = -1 + \Theta(\sqrt{\varepsilon})$成立。那么我们有 $\frac{\mathrm{d}}{\mathrm{d}t}(\bar{w}_2\bar{\alpha}) > 0$。

引理 D.18. 假设归纳假设 *D.1* 在阶段 *2* 的整个过程中成立。那么 $|\bar{w}_2| \le d$。**引理 D.** **19** *1* *2* $|\bar{w}_2| \ge$ **.** 假设归纳假设 *D.*在阶段 **.** 的整个过程中成立。() 且 ()。 $\Theta(1/d^3)$ $\alpha \ge \Theta(1/d^{1.5})$

本小节各部分的证明汇总如下。

引理 *D.15* 的证明。 对于任意 $\boldsymbol{x} \in \mathbb{R}^d$，根据引理 D.7 以及 σ 的 Lipschitz 性，对于任意 $\boldsymbol{x} \in X_1 \cup X_2$,

$$\begin{aligned}
f(\boldsymbol{x}) &= \sigma(\bar{w}_2\alpha\bar{F}(\boldsymbol{x}) + \bar{b}_2) \pm O\left(d^3\delta_2^{(2)}\right) \\
&= \sigma(1 - \|\boldsymbol{x}\|) \pm |1 - \bar{b}_2| \pm \left|-\|\boldsymbol{x}\| - \bar{w}_2\alpha\bar{F}(\boldsymbol{x})\right| \pm O\left(d^3\delta_2^{(2)}\right).
\end{aligned}$$

根据归纳假设D.1，对于任意 $\boldsymbol{x} \in X_1 \cup X_2$，我们有

$$\begin{aligned}
\left|-\|\boldsymbol{x}\| - \bar{w}_2\alpha\bar{F}(\boldsymbol{x})\right| &= \left|-1 - \bar{w}_2\alpha\bar{F}(\bar{\boldsymbol{x}})\right|\|\boldsymbol{x}\| \\
&\le |-1 - \bar{w}_2\alpha|\,\|\boldsymbol{x}\| + \left|1 - \bar{F}(\bar{\boldsymbol{x}})\right||\bar{w}_2|\alpha\,\|\boldsymbol{x}\| \le O\left(\frac{|-1 - \bar{w}_2\alpha|}{|\bar{w}_2\alpha|}\right) + O\left(\delta_{1,L^\infty}^{(2)}\right).
\end{aligned}$$

因此，

$$f(\boldsymbol{x}) = f_*(\boldsymbol{x}) \pm |1 - \bar{b}_2| \pm O\left(\frac{|-1-\bar{w}_2\alpha|}{|\bar{w}_2\alpha|}\right) \pm O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right).$$

因此，

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right\} &\le \frac{1}{2}\left(|1-\bar{b}_2| + O\left(\frac{|-1-\bar{w}_2\alpha|}{|\bar{w}_2\alpha|}\right) + O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right)\right)^2 \\ &\le O\left((1-\bar{b}_2)^2 + \frac{(-1-\bar{w}_2\alpha)^2}{\bar{w}_2^2\alpha^2} + \left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right)^2\right).\end{aligned}$$

□

引理 *D.16* 的证明。 根据引理 D.7，对于任意 $(v_2, r_2) \in \mu_2$，我们有

$$\dot{r}_2 = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{f_*(\boldsymbol{x}) - \sigma(\bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2)\right\} \pm O\left(d^3\delta_2^{(2)}\right).$$

然后，根据归纳假设 D.1 以及 σ 的 Lipschitz 性，我们有

$$\sigma(\bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2) = \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\|\bar{F}(\bar{\boldsymbol{x}}) + \bar{b}_2) = \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\| + \bar{b}_2) \pm O\left(\delta_{1,L^\infty}^{(2)}\right).$$

因此，

$$\dot{\bar{b}}_2 = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{f_*(\boldsymbol{x}) - \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\| + \bar{b}_2)\right\} \pm O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right).$$

由于 $\mathcal{L} \ge \varepsilon$，根据引理 D.15，我们有

$$\frac{(-1-\bar{w}_2\alpha)^2}{\bar{w}_2^2\alpha^2} \ge \Omega(\varepsilon) - O(\delta^2) - O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right)^2 \ge \Omega(\varepsilon).$$

由于 $\bar{w}_2\alpha \ge -1$，这意味着 $\bar{w}_2\alpha \ge -1 + \Omega(|\bar{w}_2|\alpha\sqrt{\varepsilon})$。事实上，即使当 $|\bar{w}_2|\alpha$ 为 $o(1)$ 时，这也意味着 $\bar{w}_2\alpha \ge -1 + \Omega(\sqrt{\varepsilon})$，因为在这种情况下，$\bar{w}_2\alpha \ge -1 + \Omega(\sqrt{\varepsilon})$ 直接成立。因此，

$$\sigma(\bar{w}_2\alpha\|\boldsymbol{x}\| + \bar{b}_2) \ge \sigma\left(\left(-1 + \Omega\left(\sqrt{\varepsilon}\right)\right)\|\boldsymbol{x}\| + 1 - \delta\right) = \sigma\left(1 - \|\boldsymbol{x}\| + \Omega\left(\sqrt{\varepsilon}\|\boldsymbol{x}\|\right) - \delta\right).$$

Thus,

$$\begin{aligned}\dot{\bar{b}}_2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{f_*(\boldsymbol{x}) - \sigma\left(1 - \|\boldsymbol{x}\| + \Omega\left(\sqrt{\varepsilon}\|\boldsymbol{x}\|\right) - \delta\right)\right\} \pm O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right) \\ &\le \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le}\left\{1 - \|\boldsymbol{x}\| - \left(1 - \|\boldsymbol{x}\| + \Omega\left(\sqrt{\varepsilon}\|\boldsymbol{x}\|\right) - \delta\right)\right\} + O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right) \\ &= -\Omega\left(\sqrt{\varepsilon}\right) + \delta + O\left(\delta_{1,L^\infty}^{(2)} + d^3\delta_2^{(2)}\right).\end{aligned}$$

As l 只要 $\delta = \Theta(\sqrt{\varepsilon})$ 中的常数足够小，当 $b_2 = 1 - \delta$ 时，这就意味着 $b_2 < 0$。 □

引理的证明 *D.17*。 由引理 D.3 和引理 D.7，我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\bar{w}_2\alpha) = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{\left(f_*(\boldsymbol{x}) - \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\|\bar{F}(\bar{\boldsymbol{x}}) + \bar{b}_2)\right)F(\boldsymbol{x})\right\}\left(\alpha + \frac{4C_\Gamma\bar{w}_2^2}{\sqrt{d}}\right) \\ \pm O\left(d^3\log(d)\delta_2^{(2)}\right)\alpha\left(\alpha + \frac{4C_\Gamma\bar{w}_2^2}{\sqrt{d}}\right).\end{aligned}$$

现在我们估计第一项的系数。假设 $\bar{w}_2\alpha = -1 + \delta$ 存在某个 $\delta \le \Theta(\sqrt{\varepsilon})$ 满足某个充分小的常数。于是，根据引理 D.15，我们有 $(1-\bar{b}_2)^2 \ge \Omega(\varepsilon) - O(\delta^2) \ge \Omega(\varepsilon)$。因此，$b_2 \le 1 - \Theta(\sqrt{\varepsilon})$。还需注意，$\bar{w}_2\alpha = \Theta(1)$ 表明只需考虑 $\boldsymbol{x}$ 满足 $\|\boldsymbol{x}\| = \Theta(1)$ 的情形。因此，我们有

$$\begin{aligned}\sigma(\bar{w}_2\alpha\|\boldsymbol{x}\|\bar{F}(\bar{\boldsymbol{x}}) + \bar{b}_2) &= \sigma(\bar{w}_2\alpha\|\boldsymbol{x}\| + \bar{b}_2) \pm O\left(\delta_{1,L^\infty}^{(2)}\right) \\ &\le \sigma\left(1 - \|\boldsymbol{x}\| - \Theta(\sqrt{\varepsilon})\right) + O\left(\delta_{1,L^\infty}^{(2)}\right).\end{aligned}$$

然后，我们将该系数分解为

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{\left(f_*(\boldsymbol{x})-\sigma(\bar{w}_2\alpha\left\|\boldsymbol{x}\right\|\bar{F}(\bar{\boldsymbol{x}})+\bar{b}_2)\right)F(\boldsymbol{x})\right\} &= \mathbb{E}\left\{\left(f_*(\boldsymbol{x})-\sigma(\bar{w}_2\alpha\left\|\boldsymbol{x}\right\|\bar{F}(\bar{\boldsymbol{x}})+\bar{b}_2)\right)F(\boldsymbol{x})\right\}\\ &\geq \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\leq 1}\left\{\left(\Theta(\sqrt{\varepsilon})-O\left(\delta_{1,L^\infty}^{(2)}\right)\right)F(\boldsymbol{x})\right\}\\ &\geq \Omega\left(\alpha\sqrt{\varepsilon}\right).\end{aligned}$$

因此，

$$\frac{\mathrm{d}}{\mathrm{d}t}(\bar{w}_2\alpha)\geq\left(\Omega(\sqrt{\varepsilon})-O\left(d^3\delta_2^{(2)}\right)\log(d)\right)\alpha\left(\alpha+\frac{4C_\Gamma\bar{w}_2^2}{\sqrt{d}}\right)>0.$$

□

引理的证明 *D.18*。 根据引理 D.3 和引理 D.7，我们有

$$\begin{aligned}\dot{\bar{w}}_2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))F(\boldsymbol{x})\right\}\pm\left(d^3\log d\delta_2^{(2)}\right)\\ \dot{\alpha} &= \frac{4C_\Gamma}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))F(\boldsymbol{x})\right\}\bar{w}_2\pm O\left(d^{2.5}(\log d)\delta_2^{(2)}\right)\end{aligned}$$

因此，

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left(\alpha-\frac{2C_\Gamma}{\sqrt{d}}\bar{w}_2^2\right)\right|\leq O\left(d^4\delta_2^{(2)}\right).$$

另请回忆， $\bar{w}_2^2\ll\alpha$ 在 T_1处。因此，在整个阶段 2 中，我们始终有 $\left|\alpha-\frac{2C_\Gamma}{\sqrt{d}}\bar{w}_2^2\right|\ll 1/d$。由于 $|\bar{w}_2\alpha|\leq 1$，这意味着 $|\bar{w}_2|\leq O(d^{1/6})\leq d$。 □

引理的证明 *D.19*。 回顾引理 D.18 的证明可知 $|\alpha-\frac{2C_\Gamma}{\sqrt{d}}\bar{w}_2^2|\ll 1/d$。因此，当 $\alpha=\Theta(1/d^{1.5})$时，我们有 $|\bar{w}_2|\leq O(1/d)$。阶段 1 中的估计，作相应调整，表明在这种情况下， α 和 $|\bar{w}_2|$ 都会增长。

□

D.3 收敛速率

回顾引理 D.5，有 $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}=-\mathbb{E}_{w_2,\ b_2,\ \boldsymbol{w}_1}\left\|\nabla_{w_2,\ b_2,\ \boldsymbol{w}_1}\right\|^2$，其中

$$\nabla_{w_2,b_2,\boldsymbol{w}_1}:=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\begin{bmatrix}\sigma'(w_2F(\boldsymbol{x})+b_2)F(\boldsymbol{x})\\ \sigma'(w_2F(\boldsymbol{x})+b_2)\\ 2\bar{W}_2(\boldsymbol{x})\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\\ \|\boldsymbol{w}_1\|\bar{W}_2(\boldsymbol{x})\sigma'(\boldsymbol{w}_1\cdot\boldsymbol{x})(\boldsymbol{I}-\bar{\boldsymbol{w}}_1\bar{\boldsymbol{w}}_1^\top)\boldsymbol{x}\end{bmatrix}\right\}.$$

引理 D.20。 假设归纳假设 *D.1* 在时刻 t成立。那么有

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}\leq-\left\|\tilde{\nabla}\right\|^2+O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)d^4\right),$$

其中

$$\tilde{\nabla}:=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\begin{bmatrix}\|\boldsymbol{x}\|\sqrt{\alpha^2+\frac{4C_\Gamma}{\sqrt{d}}\bar{w}_2^2\alpha}\\ 1\end{bmatrix}\right\}.$$

引理 D.21。 假设归纳假设 *D.1* 在时刻 t成立。那么有

$$\left\|\tilde{\nabla}\right\|\geq\Omega(\alpha\mathcal{L})-O\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right).$$

引理 D.22 （阶段 2）。 假设归纳假设 *D.1* 在整个阶段 *2* 内均成立。那么$T_2-T_1\leq O(d^3/\varepsilon)$。

引理 *D.20* 的证明。由于它是 $\nabla_{w_2,\ b_2,\ \boldsymbol{w}_1}$的范数，我们可以放心地忽略最后一项，只考虑前三项。根据引理 D.7，我们有

$$[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_{1:3}=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\begin{bmatrix}F(\boldsymbol{x})\\ 1\\ 2\bar{w}_2\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\end{bmatrix}\right\}\pm O\left(d^3\delta_2^{(2)}\begin{bmatrix}\alpha\log(d)\\ 1\\ \bar{w}_2\|\boldsymbol{w}_1\|\log(d)\end{bmatrix}\right).$$

此外，我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))F(\boldsymbol{x})\}&=\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\alpha\left\|\boldsymbol{x}\right\|\}+\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\alpha(\bar{F}(\boldsymbol{x})-\left\|\boldsymbol{x}\right\|)\right\}\\&=\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\alpha\left\|\boldsymbol{x}\right\|\}+O\left(\alpha\delta_{1,L^2}^{(2)}\right).\end{aligned}$$

同时，对于 $[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_3$，根据引理 B.3 和引理 D.8，我们有

$$\begin{aligned}&2\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\}\\&=2\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\right\}+2\bar{w}_2\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(\tilde{f}(\boldsymbol{x})-f(\boldsymbol{x}))\sigma(\boldsymbol{w}_1\cdot\boldsymbol{x})\right\}\\&=\frac{2C_\Gamma\bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))\left\|\boldsymbol{x}\right\|\right\}\left\|\boldsymbol{w}_1\right\|\pm2\bar{w}_2\left\|\boldsymbol{w}_1\right\|\left\|f-\tilde{f}\right\|_{L^2}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{x}\in X_2}\left\|\boldsymbol{x}\right\|^2}\\&=\frac{2C_\Gamma\bar{w}_2}{\sqrt{d}}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-\tilde{f}(\boldsymbol{x}))\left\|\boldsymbol{x}\right\|\right\}\left\|\boldsymbol{w}_1\right\|\pm O\left(|\bar{w}_2|^{1.5}\alpha^{0.5}\left\|\boldsymbol{w}_1\right\|\delta_{1,L^2}^{(2)}\right).\end{aligned}$$

~重复上述过程，我们可以将第一项中的 f 替换为 f_*。因此，

$$\begin{aligned}[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_{1:3}&=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\begin{bmatrix}\alpha\left\|\boldsymbol{x}\right\|\\1\\\frac{2C_\Gamma\bar{w}_2}{\sqrt{d}}\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{w}_1\right\|\end{bmatrix}\right\}\\&\quad\pm O\left(\delta_{1,L^2}^{(2)}\begin{bmatrix}\alpha\\0\\|\bar{w}_2|^{1.5}\alpha^{0.5}\left\|\boldsymbol{w}_1\right\|\end{bmatrix}\right)\pm O\left(d^3\delta_2^{(2)}\begin{bmatrix}\alpha\log(d)\\1\\\bar{w}_2\left\|\boldsymbol{w}_1\right\|\log(d)\end{bmatrix}\right)\\&=\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\begin{bmatrix}\alpha\left\|\boldsymbol{x}\right\|\\1\\\frac{2C_\Gamma\bar{w}_2}{\sqrt{d}}\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{w}_1\right\|\end{bmatrix}\right\}\\&\quad\pm O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)\begin{bmatrix}\alpha\log(d)\\1\\|\bar{w}_2|\left\|\boldsymbol{w}_1\right\|\log(d)\end{bmatrix}\right).\end{aligned}$$

现在，我们估计 $[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_{1:3}$的期望范数。首先，我们有

$$\begin{aligned}[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_1^2&=\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\left\|\boldsymbol{x}\right\|\}\right)^2\alpha^2\pm O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)\alpha^2\log(d)\right),\\[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_2^2&=\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\{f_*(\boldsymbol{x})-f(\boldsymbol{x})\}\right)^2\pm O\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right).\end{aligned}$$

对于 $[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_3$，我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{w}_1}[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_2^3&=\frac{4C_\Gamma^2\bar{w}_2^2}{d}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\left\|\boldsymbol{x}\right\|\}\right)^2\mathop{\mathbb{E}}_{\boldsymbol{w}_1}\left\|\boldsymbol{w}_1\right\|^2\\&\quad\pm O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)\bar{w}_2^2\frac{\mathbb{E}\left\|\boldsymbol{w}\right\|_1^2}{\sqrt{d}}\log(d)\right)\\&=\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\left\|\boldsymbol{x}\right\|\}\right)^2\frac{4C_\Gamma\bar{w}_2^2}{\sqrt{d}}\alpha\\&\quad\pm O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)\bar{w}_2^2\alpha\log(d)\right).\end{aligned}$$

因此，

$$\begin{aligned}\left\|[\nabla_{w_2,b_2,\boldsymbol{w}_1}]_{1:3}\right\|^2&=\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\{(f_*(\boldsymbol{x})-f(\boldsymbol{x}))\left\|\boldsymbol{x}\right\|\}\right)^2\left(\alpha^2+\frac{4C_\Gamma\bar{w}_2^2}{\sqrt{d}}\alpha\right)+\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\{f_*(\boldsymbol{x})-f(\boldsymbol{x})\}\right)^2\\&\quad\pm O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)d^4\right)\\&=\left\|\tilde{\nabla}\right\|^2\pm O\left(\left(\delta_{1,L^2}^{(2)}+d^3\delta_2^{(2)}\right)d^4\right).\end{aligned}$$

□

引理的证明 *D.21*。 为简化记号，令 $A := \sqrt{\alpha^2 + \frac{4C_\Gamma}{\sqrt{d}}\bar{w}_2^2}\alpha$。于是我们可以写作

$$\tilde{\nabla} = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\begin{bmatrix} A\|\boldsymbol{x}\| \\ 1\end{bmatrix}\right\}.$$

定义

$$\hat{\nabla} = \begin{bmatrix} -1 - \alpha\bar{w}_2 \\ A(1-\bar{b}_2)\end{bmatrix}.$$

根据归纳假设 D.1， $\left\|\hat{\nabla}\right\| \le O(1)$。因此，为了给出 $\left\|\tilde{\nabla}\right\|$的下界，只需给出 $\left\langle\tilde{\nabla}, \hat{\nabla}\right\rangle$ 的下界。我们有

$$\left\langle\tilde{\nabla}, \hat{\nabla}\right\rangle = A\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\left(-\|\boldsymbol{x}\| + 1 - (\alpha\bar{w}_2\|\boldsymbol{x}\| + \bar{b}_2)\right)\right\}.$$

首先，对于这些 $\boldsymbol{x} \in \{\|\boldsymbol{x}\| \le 1\}$，我们有 $f_*(\boldsymbol{x}) = -\|\boldsymbol{x}\| + 1$ 和

$$f(\boldsymbol{x}) = \bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2 = \bar{w}_2\alpha\|\boldsymbol{x}\| + \bar{b}_2 + \bar{w}_2\alpha(\bar{F}(\boldsymbol{x}) - \|\boldsymbol{x}\|).$$

因此，我们有

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\left(-\|\boldsymbol{x}\| + 1 - (\alpha\bar{w}_2\|\boldsymbol{x}\| + \bar{b}_2)\right)\right\} \\
=& \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right\} + \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\bar{w}_2\alpha(\bar{F}(\boldsymbol{x}) - \|\boldsymbol{x}\|)\right\} \\
=& \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\le 1}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right\} \pm O\left(|\bar{w}_2\alpha|\delta^{(2)}_{1,L^2}\right).
\end{aligned}$$

然后，对于 $\boldsymbol{x} \in \{\|\boldsymbol{x}\| \ge 1\}$，注意 $-\|\boldsymbol{x}\| + 1 \le 0$ 和 $f_*(\boldsymbol{x}) = 0$。因此，我们有

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))\left(-\|\boldsymbol{x}\| + 1 - (\alpha\bar{w}_2\|\boldsymbol{x}\| + \bar{b}_2)\right)\right\} \\
=& -\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{f(\boldsymbol{x})\left(-\|\boldsymbol{x}\| + 1 - (\alpha\bar{w}_2\|\boldsymbol{x}\| + \bar{b}_2)\right)\right\} \ge \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{f(\boldsymbol{x})(\alpha\bar{w}_2\|\boldsymbol{x}\| + \bar{b}_2)\right\}.
\end{aligned}$$

然后，我们计算

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{f(\boldsymbol{x})(\alpha\bar{w}_2\|\boldsymbol{x}\| + \bar{b}_2)\right\} \\
=& \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{f(\boldsymbol{x})(\bar{w}_2 F(\boldsymbol{x}) + \bar{b}_2)\right\} + \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{f(\boldsymbol{x})(\alpha\bar{w}_2(\|\boldsymbol{x}\| - \bar{F}(\boldsymbol{x})))\right\} \\
=& \mathop{\mathbb{E}}_{\|\boldsymbol{x}\|\ge 1}\left\{f^2(\boldsymbol{x})\right\} \pm O\left(d^3\delta^{(2)}_2\right) \pm O\left(\alpha\bar{w}_2\delta^{(2)}_{1,L^2}\right).
\end{aligned}$$

其中，第二个等式由引理D.7和归纳假设D.1得出。将这两种情况合并，可得

$$\left\langle\tilde{\nabla}, \hat{\nabla}\right\rangle \ge A\mathop{\mathbb{E}}_{\boldsymbol{x}}\left\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right\} - O\left(|\bar{w}_2\alpha|\delta^{(2)}_{1,L^2}\right) - O\left(d^3\delta^{(2)}_2\right).$$

最后，注意到 $A \ge \alpha$。因此，

$$\left\|\tilde{\nabla}\right\| \ge \Omega(\alpha\mathcal{L}) - O\left(\delta^{(2)}_{1,L^2} + d^3\delta^{(2)}_2\right).$$

□

引理*D.22*的证明。根据引理D.20 和引理D.21，

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L} \le -\Omega(\alpha^2\mathcal{L}^2) + O\left(\left(\delta^{(2)}_{1,L^2} + d^3\delta^{(2)}_2\right)d^4\right) \le -\Omega\left(\frac{\mathcal{L}^2}{d^3}\right)$$

因此，对于任意 $T \in [T_1, T_2]$，

$$\mathcal{L}(T) \le \left(\Omega\left(d^{-3}\right)(T - T_1) + \frac{1}{\mathcal{L}(T_1)}\right)^{-1} \le O\left(\frac{d^3}{T - T_1}\right).$$

因此， $\mathcal{L}$ 到达 ε最多需要 $O(d^3/\varepsilon)$的时间。 □

D.4 主要引理的证明

引理证明 *D.2*。归纳假设在第 D.2 节中得到维持，并且根据引理 D.22，我们有 $T_2 - T_1 \le O(d^3/\varepsilon)$。现在我们考虑第一层误差。回顾一下，

$$\frac{\mathrm{d}}{\mathrm{d}t}(\delta_{1,L^2}^{(2)})^2 = O\left(d^5\delta_{1,L^2}^{(2)}\left(\delta_{X_2}^{(2)}\right)^2\right),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\delta_{1,L^\infty}^{(2)} = O\left(d^3\delta_{1,L^2}^{(2)} + d^2\left(\delta_{X_2}^{(2)}\right)^2\right).$$

回顾一下，$\delta_{X_2} := O(1)d^{4.5}(\delta_1^{(2)}, {}_{L^\infty} + d^3\delta_2^{(2)})$。为简单起见，我们选择 $\delta_1^{(2)}, {}_{L^\infty} \ge d^3\delta_2^{(2)}$ 使得$\delta_{X_2} = O(d^{4.5}\delta_1^{(2)}, {}_{L^\infty})$。于是，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}(\delta_{1,L^2}^{(2)})^2 = O\left(d^{14}\delta_{1,L^2}^{(2)}(\delta_{1,L^\infty}^{(2)})^2\right),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\delta_{1,L^\infty}^{(2)} = O\left(d^3\delta_{1,L^2}^{(2)} + d^{11}(\delta_{1,L^\infty}^{(2)})^2\right).$$

我们选择 $\delta_1^{(2)}, {}_{L^2}(T_1)$和 $\delta_1^{(2)}, {}_{L^\infty}(T_1)$，使得

$$\Theta\left(\frac{d^{17}}{\varepsilon}(\delta_{1,L^\infty}^{(2)})^2\right) \le \delta_{1,L^2}^{(2)} \le \Theta\left(\frac{\varepsilon}{d^6}\delta_{1,L^\infty}^{(2)}\right) \quad \text{and} \quad \delta_{1,L^\infty}^{(2)}(T_1) \le \Theta\left(\frac{\varepsilon}{d^{14}}\right). \tag{14}$$

注意，这是可行的，因为 $\delta_{1\ L^2}^{(2)}, (T_1)$和 $\delta_1^{(2)}, {}_{L^\infty}(T_1)$可以选取得任意小，且其大小关于多项式倒数而言。若此条件成立，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}(\delta_{1,L^2}^{(2)})^2 \le O\left(\frac{\varepsilon}{d^3}(\delta_{1,L^2}^{(2)})^2\right) \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}\delta_{1,L^\infty}^{(2)} = O\left(\frac{\varepsilon}{d^3}\delta_{1,L^\infty}^{(2)}\right).$$

因此，根据归纳法，在 $O(d^3/\varepsilon)$的时间内，这两个误差至多分别为 $O(\delta_1^{(2)}, {}_{L^2}(T_1))$和$O(\delta_1^{(2)}, {}_{L^\infty}(T_1))$。 □

E 从梯度流到梯度下降

只要我们能在每次迭代时生成新样本，就可以用标准方法将上述梯度流论证转化为梯度下降论证。首先，通过选择足够小的步长，可以确保每一步中梯度下降与梯度流之间的差异是逆多项式小。请注意，我们的论证建立在归纳假设之上。因此，我们无需担心误差的累积。此外，我们的估计可以容忍逆多项式大的误差。然后，在梯度下降的每一步中，我们生成足够多（但仍为多项式数量的）样本，以确保总体梯度与有限样本梯度之间的差异以高概率足够小。由于完成该过程只需多项式次迭代，因此所需样本总数为多项式数量。