arXiv:2304.03385v1 [cs.LG] 6 Apr 2023

宽神经网络：从初始化时的非高斯随机场到训练的神经切线核几何

LUIS CARVALHO、JOÃO LOPES COSTA、JOSÉ MOURÃO 和 GONÇALO OLIVEIRA

摘要。近年来，具有超过 $n = 10^{14}$ 个参数的人工神经网络在应用方面取得了进展，因此研究此类网络在大 n 时的行为变得极为重要。

大多数研究宽神经网络的工作都聚焦于此类网络在无限宽度 $n \to +\infty$ 极限下的情形，并表明它们在初始化时对应于高斯过程 [12, 20]。本文将研究它们在宽度较大但有限（n）时的行为。我们的主要贡献如下：

- 以关于 n^{-1} 的渐近级数形式计算对高斯性的修正。该展开式的系数由参数初始化的统计特性和激活函数决定。

- 在训练过程中，通过计算与无限宽度极限情形的偏差（在该极限下，网络沿线性流演化），控制有限宽度 n 网络输出的演化。这改进了先前的估计 [1, 16, 10, 18]；此外，我们还得到了更精确的有限宽度神经正切核衰减率，它们以 n 表示，并在整个训练过程中均成立。作为推论，我们还证明，以任意高的概率，宽度足够大的神经网络的训练会收敛到相应二次损失函数的全局最小值。

- 以n为尺度估计偏离高斯性的程度如何随训练演变。具体而言，利用测度空间中的某种度量，我们发现，在训练过程中，所得测度与对应于无限宽度网络的时变高斯过程相差不超过$n^{-\frac{1}{2}}(\log n)^{1+}$（该时变高斯过程可通过将初始高斯过程与对应于无限宽度极限下训练的线性流进行复合而明确给出）。

目录

1. 引言 2 1.1. 背景 2 1.2. 概要 2 1.3. 主要结果 3 1.4. 相关工作 7 2. 初始化时偏离高斯性 8 2.1. 所考虑的宽神经网络 8 2.2. 初始化时函数空间上诱导的概率测度 9 2.3. 收敛至高斯随机场 10 2.4. 偏离高斯性的定量分析 11 3. 函数空间中的训练几何 12 3.1. 训练与神经切线核 12 3.2. 作为将梯度流“推前”至函数空间的训练 14

3.3. 无限网络输出的演化与收敛 15 3.4. 宽但有限网络输出的演化与收敛 16 4. 训练期间演化的近似 18 4.1. 通过向量场的流实现演化 18 4.2. 无限宽度情形下的演化 19 4.3. 大但有限宽度下的演化 24 附录 A. 埃奇沃思展开 27 A.1. 特征函数 28 A.2. 累积量生成函数 28 附录 B. 中心极限定理的渐近改进（埃奇沃思展开） 30 B.1. 中心极限定理 30 B.2. $n \gg 1$ 中的渐近行为（定理 1 的证明） 31 B.3. $k = 1$ 情形（例1的细节） 34 附录 C. 训练期间输出的演化与收敛 36 C.1. 一个小幅改进 42 参考文献 43

1. 引言

1.1. 背景。近年来，具有超过 $n = 10^{14}$ 个参数的人工神经网络在应用方面取得了进展，因此研究此类网络的大n 行为变得极为重要。

大多数研究都聚焦于此类网络的无限宽度$n \to +\infty$ 极限，并表明，在初始化时，该极限对应于高斯随机场（高斯随机函数） [12, 20]；而对于二次损失，其训练对应于相对于神经正切核（神经正切核$^\infty$）的线性演化 [16, 1]。

1.2. 概要。本文的目标是研究具有单个隐藏层的人工神经网络在初始化和训练时，当n较大但有限的情况下的行为。

1.2.1. 初始化。 一个具有固定参数（权重和偏置，$\theta \in \mathbb{R}^{P_n}$）的人工神经网络，是一个从 $\mathbb{R}^{n_{in}}$ 到 $\mathbb{R}^{n_{out}}$ 的函数（为简化起见，我们将重点讨论 $n_{in} = n_{out} = 1$ 的情形，尽管我们的大多数结果都可以自然地推广，但代价是符号过于繁复，且一些证明会变得晦涩）。改变参数后，我们得到一个映射 F_n，它从超参数空间 $\mathbb{R}^{P_n}$ 映射到函数空间。在初始化时，构成此类网络的参数通常从某个预先设定的分布中随机生成；在 F_n 下进行推前后，这会在函数空间中诱导出一列分布 $\mathrm{P}^{(n)}$，或一列随机场。

众所周知，在$n \to +\infty$ 极限下，该序列会收敛到一个高斯随机函数，文献中已经确立了许多此类结果的形式，参见例如 [12, 20] 及其中的参考文献。尽管理解这种极限很重要，但现实中的网络只有有限数量的神经元n，因此必须计算（有限维）分布 $\mathrm{P}^{(n)}$ 与极限高斯分布之间的偏差。一些近期工作 [13] 提出了处理此类问题的方法

借助有效量子场论中的技术。相反，在本工作中，我们采用一种更简单的方法，使用埃奇沃思级数技术，以便将高斯性的修正显式计算为$n^{-\frac{1}{2}}$ 的渐近级数，其系数由参数初始化的统计特性决定。

1.2.2. 训练。 这类网络在训练过程中的演化，可以用有限的n神经切线核（$\mathrm{NTK}^{(n)}$）来建模。众所周知，当$n \to +\infty$时，它几乎必然收敛到一个常数。因此，在此极限下，这类网络的训练等价于一个线性流（在函数空间中）。在我们的工作中，我们改进了先前的估计 [16, 1, 10, 18]，并计算出更精确且关于时间一致的 $\mathrm{NTK}^{(n)}$ 衰减率，其表达式以n为变量。由此产生的偏差，反映了实际情况与刻画无限n 极限的线性流之间的差异；这使我们能够同时控制训练过程中输出随时间$t \geq 0$ 和宽度n的演化。事实上，正如非线性问题中常见的那样，我们需要“同时控制一切”，因为要仔细控制 $\mathrm{NTK}^{(n)}$，就需要详细控制网络的输出，反之亦然。这通过连续性／自举论证实现——这是一种非线性分析中的标准技术，据我们所知，它尚未广泛应用于人工神经网络研究。最后，我们的结果有一个具体推论：在适当条件下，通过梯度下降训练足够宽的神经网络，将以关于参数初始化的任意高概率收敛到相应二次损失函数的全局最小值。

这些结果使我们能够用n来估计我们在初始化时推导出的偏离高斯性的程度在训练过程中如何演变。我们首先证明，无限宽度网络生成的概率测度是一个时变高斯过程；该过程可通过将初始高斯过程与对应于无限宽度极限下训练的线性流进行复合，明确地表示出来。随后，利用测度空间中的 Prokhorov 度量，我们发现，在训练过程中，与有限宽度网络相关联的测度与对应于无限宽度网络的时变高斯过程之间的距离为$n^{-\frac{1}{2}}(\log n)^{+}$。

1.3. 主要结果。在陈述本文的主要结果之前，我们需要做一些准备。我们将这种网络的权重和偏置称为参数，并将它们记为

$$\theta := ((W_1^{(2)},\dots,W_n^{(2)},W_1^{(1)},\dots,W_n^{(1)}),(b_1^{(2)},\dots b_n^{(2)},b_1^{(1)},\dots,b_n^{(1)})) \in (\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^n \times \mathbb{R}^n) \cong \mathbb{R}^{4n} =: \mathbb{R}^{P_n}.$$

这些参数与非线性激活函数 $\sigma : \mathrm{R} \to \mathrm{R}$ 一起，确定了对该网络进行编码的标量值函数

$$f_\theta^{(n)}(x) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(W_j^{(2)}\, \sigma\left(W_j^{(1)} x + b_j^{(1)} \right) + b_j^{(2)} \right). \tag{1.1}$$

因此，我们得到一个映射 $\mathrm{F}_{(n)}$，它将一组参数 θ 映射为函数$f_{(n)\,\theta} \in \mathrm{Map}([a,b],\mathrm{R})$。

1.3.1. 初始化。 通常在应用中，参数 θ 会从 R^{4n} 上预先确定的概率测度中随机初始化。对于每个 n，通过经 F_n 推前，这会在无限维函数空间 $\mathrm{Map}([a,b],\mathrm{R})$ 上诱导出概率测度 $\mathrm{P}^{(n)}$。在某些

在 2.3 中回顾的非常自然的假设下，并且我们也将在本引言下文中再次回顾这些假设，根据中心极限定理，$\mathrm{P}^{(n)}$ 收敛于 Map(R,R) 上的高斯测度，称为高斯随机函数。本文第一部分的主要目标是计算 $\mathrm{P}^{(n)}$ 偏离高斯性的程度。为了研究此类偏离，与其直接研究函数，不如研究通过在有限个点上取值而获得的有限维分布。注意，正如我们在 2.2 中回顾的，根据柯尔莫哥洛夫扩张定理，这样做不失一般性。

现在，我们考虑通过固定$k \in \mathrm{N}$ 和$x_1,\ldots,x_k \in [a,b],$，考察这样一个映射所得到的有限维分布：该映射将一组参数映射为网络在这些点上的取值

$$\begin{aligned} \mathbb{R}^{4n} &\longrightarrow \mathrm{Map}([a,b],\mathbb{R}) \to \mathbb{R}^k, \\ \theta &\longmapsto (f_\theta(x_1),\cdots,f_\theta(x_k)), \end{aligned}$$

并将最初固定在 R^{4n} 上的分布推前至 R^k 上的一个分布，我们将其记为$Q^{(n)}_{x_1,\ldots,x_k}$。

借助这类有限维分布，可以轻松理解收敛到高斯过程的过程，具体如下。从现在起，假设所有元组（$W(2)_j, W(1)_j, b^{(2)}{}_j, b^{(1)}{}_j$）$\sim$（$W(2), W(1), b^{(2)}, b^{(1)}$）均为独立同分布，且$W(2)$ 和$b^{(2)}$均为零均值。于是，中心极限定理适用于 1.1，并得到

$$\mathbb{P}^{(n)}_{x_1,\ldots,x_k} \to \mathbb{P}^{C}_{x_1,\ldots,x_k}, \tag{1.2}$$

其中 $\mathrm{P}^{C}_{x_1,\ldots,x_k}$是 R^k 上的高斯分布，其协方差为 $C(x_i,xj) = \mathrm{E}[p(x_i)p(xj)]$,其中 $p(x) = W(2)\sigma(W(1)x + b^{(1)}) + b^{(2)}$而 $\mathrm{E}[\cdot]$ 表示相对于参数 $(W(2), W(1), b^{(2)}, b^{(1)})$ 上预先固定的分布所取的期望。更明确地说

$$d\mathbb{P}^{C}_{x_1,\ldots,x_k}(y_1,\ldots,y_k) = \frac{e^{-\frac{1}{2}C^{-1}_{ab}y_a y_b}}{(2\pi\det(C_{ij}))^{k/2}} dy^1\ldots dy^k, \tag{1.3}$$

其中 $Cij := C(x_i, xj)$，这里我们使用了对重复指标求和的爱因斯坦求和约定。特别地，若 $k = 1$，则 $C = \mathrm{E}[p(x)^2]$ 是随机变量 $p(x)$ 的二阶矩，我们将其记为 μ_2。那么，在这种 $k = 1$ 情形下，有

$$d\mathbb{P}^{C}_{x}(y) = \frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} dy. \tag{1.4}$$

结合上述中心极限定理与柯尔莫哥洛夫扩张定理，我们得出，有限维分布 (1.3) 在 Map($[a,b]$,R) 上定义了一个协方差为 C 的高斯测度。此外，若非线性激活函数 σ 是利普希茨函数，我们将在定理 ?? 中证明，高斯测度 P^{C} 实际上支撑在更小且性质更好的连续函数空间 $\mathrm{C} = C([a.b],\mathrm{R})$ 上。

本文的第一个主要结果是定理 1（见图 1.3.1 和图 1.3.1），它对于有限的 n，明确计算了初始化时分布 $\mathrm{P}^{(n)}_{x_1,\ldots,x_k}$ 中偏离高斯性的程度。这里，在引言中，我们仅陈述该结果针对 $k = 1$ 的一个特殊版本

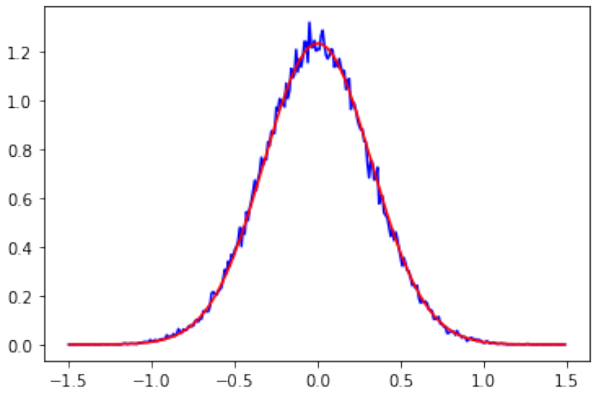

图1. 定理 A 的实验验证，揭示了高斯行为（特别是 $\mu_3 = 0$）。蓝色曲线对应于通过对一个具有 2.000 个神经元/感知机、采用 Sigmoid 非线性且参数从区间 $[-1/2, 1/2]$ 中均匀采样的神经网络，在 $x = 1$ 处的取值进行 80.000 次采样而得到的密度函数。红色曲线是定理的预测结果，不包括 $O(n^{-3/2})$ 项。

这应与公式 1.4 进行比较，并视为该公式在 n 较大但有限时的一个改进版本。

定理 A (定理 1 的特例)。 设 $x \in \mathbb{R}$，且对于所有 $i \in \{1, \dots, n\}$，元组 $(W^{(2)}{}_i, b^{(2)}{}_i, W^{(1)}{}_i, b^{(1)}{}_i)$ 均独立同分布，并且 $W^{(2)}{}_i, b^{(2)}{}_i$ 的均值为零。那么，

$$d\mathbb{P}_x^{(n)}(y) = \nu_x^{(n)}(y)dy$$

满足以下条件的密度 $\nu_x^{(n)}$

$$\nu_x^{(n)}(y) = \frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + \frac{1}{\sqrt{n}}\frac{\mu_3}{12\sqrt{2}\mu_2^{3/2}}H_3\left(\frac{y}{\sqrt{2\mu_2}}\right)\frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}}$$
$$+\frac{1}{n}\left(\frac{\mu_4 - 3\mu_2^2}{96\mu_2^2}H_4\left(\frac{y}{\sqrt{2\mu_2}}\right) + \frac{\mu_3^2}{576\mu_2^3}H_6\left(\frac{y}{\sqrt{2\mu_2}}\right)\right)\frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + O(n^{-3/2}),$$

其中，$H_n(\cdot)$ 表示第 n 阶厄米多项式，而 $\mu_r = \mathbb{E}[p(x)^r]$。

注释 1. 我们再次指出，这一结果是我们关于 $\mathbb{P}^{(n)}_{x_1,\dots,x_k}$ 的主要结果的一个示例，该结果将在 *2.4* 节中作更详细的介绍，并在 *B.2* 节中证明，如下所示：

1.3.2. 训练。 现在假设我们给定了一个训练集 $\{X_i, Y_i\}_{i=1}^N$，并希望通过梯度下降最小化二次损失函数来训练我们的网络 (1.1)。为此，我们像之前一样初始化参数 $\theta_0 = (W(2)\ 0,j, b_0^{(2)},j, W(1)\ 0,j, b_0^{(1)},j)_{j\in\mathbb{N}}$（但现在还假设它们具有有限三阶矩）。于是，我们在参数空间中得到一个定义良好的动力系统 $t \mapsto \theta(t)$，以及与之对应的、适定的输出演化

$$y(t) = f_{\theta(t)}^{(n)}(x)\,,$$

的有限 n 宽度为 n 的网络（再次回顾 (1.1)）。

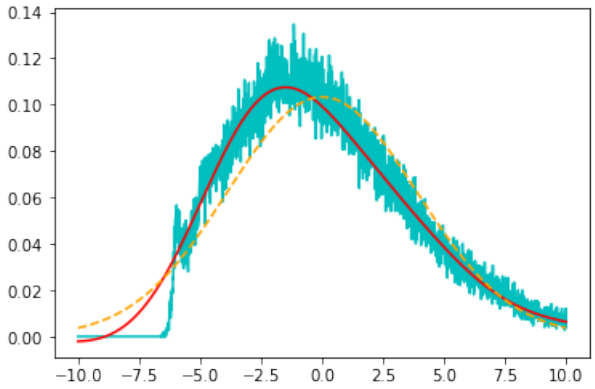

图2. 定理 A 的一次实验实现，揭示了“显著”偏离高斯性（尤其是 $\mu_3 \neq 0$）。青色曲线对应于通过对一个神经网络在 $x = 1$ 处的求值采样 150.000 次得到的密度函数；该网络包含 300 个神经元/感知机，采用 Sigmoid 非线性，参数取自一个矩为 $m_2 \approx 11{,}6$、$m_3 \approx 607{,}6$ 和 $m_4 \approx 34.131{,}4$ 的分布。红色曲线是定理的预测结果，其中不包含$O(n^{-3/2})$ 项；橙色虚线曲线则是具有相同 μ_2，的高斯分布，作为比较。

在本文中，我们还建立了各种估计（相应证明见附录 C），这些估计在第3.4节详述的适当条件下，对足够宽但有限的网络的输出的完整动态提供了清晰描述。特别地，我们将证明（该结果的完整版本见定理 2）

定理 B(定理 2 的简化版本)。 增大宽度 n 后，对于（参数）初始化，以任意高的概率：

- 该有限宽度神经切线核与其无限宽度极限之间的差异满足如下形式的估计

$$|\mathrm{NTK}^{(n)}_{zw}(\theta(t)) - \mathrm{NTK}^{\infty}_{zw}| \leq C\frac{(\log n)^{1/2+}}{\sqrt{n}}, \tag{1.5}$$

对于所有 $t \geq 0$，所有 $|z|$、$|w| \leq 1$。

- 训练输出（有限宽度神经网络的）以指数速度收敛到其标签：

$$|f^{(n)}_{\theta(t)}(X_l) - Y_l| \leq Ce^{-(\lambda_\infty - \varepsilon)t}, \tag{1.6}$$

其中 $\lambda_\infty > 0$ 是神经正切核$\Theta^{\infty}_{X_iX_j}$ [7]的最小特征值。

- 训练是稳定的，也就是说，只要由神经正切核Θ^{∞}给出的线性（无限宽度）动力学保持不变，输出动力学就会与其接近。更准确地说，我们证明，对于所有 $|x| \leq 1$,

$$|f^{(n)}_{\theta(t)}(x) - f^{\infty}_t(x)| \leq C\frac{(\log(n))^{1/2+}}{\sqrt{n}}. \tag{1.7}$$

最后，固定 $(x_1, \ldots, x_k) \in \mathbb{R}^k$，并考察训练过程中输出空间中各测度的演化。在每个训练时刻 $t \geq 0$，这给出两个测度 $\mathrm{P}^{(n)}_{x_1,\ldots,x_k}(t)$ 和 $\mathrm{P}^{\infty}_{x_1,\ldots,x_k}(t)$,

它们分别由使用有限宽度为 n 的网络进行训练，以及使用无限网络进行训练得到（注意，根据定理 3，对于所有 t ≥ 0，P∞(t) 都是高斯分布）。我们的目标是估计随着 t 演化，这两个测度之间的差异如何随 n 变化。为此，我们将使用普罗霍罗夫度量来衡量这两个测度之间的距离。对于两个测度 μ1,μ2，记作 $\pi(\mu_1,\mu_2)$，其定义为满足以下条件的正数 $\rho > 0$ 的下确界：$\mu_1(A) \le \mu_2(A^\rho) + \rho$，且 $\mu_2(A) \le \mu_1(A^\rho) + \rho$，其中 $A^\rho = \bigcup_{a\in A} B_\rho(a)$。接下来，我们证明如下结论。

定理 C。 设 $(x_1,\dots,x_k) \in \mathbb{R}^k$。那么，存在常数 $C > 0$，与 n 无关，使得对所有 $t \ge 0$

$$\pi\left(\mathbb{P}^{(n)}_{x_1,\dots,x_k}(t), \mathbb{P}^{\infty}_{x_1,\dots,x_k}(t)\right) \le C\frac{(\log(n))^{1+}}{\sqrt{n}}.$$

1.4. 相关工作。关于宽神经网络及其在无限宽度极限下的渐近行为，已有大量研究。这里我们重点介绍其中一些据我们所知与本文所述进展关系最为密切的结果。

如今，在机器学习领域，$n \to +\infty$ 宽度极限收敛到高斯过程（其适当含义已在前文说明）已是一个经典结果。该结论最初针对浅层网络情形在 [12] 中得到证明，最近又在 [20] 中推广到深层网络情形。另请参见 [21, 17]，其中对其性能与核方法进行了比较。

如前所述，其他近期研究已探讨了大但有限宽度情况下出现的偏离高斯性现象。例如，[13] 研究了深层网络，并采用了有效量子场论中的技术。此外，他们所偏好的参数初始化类型（零均值的标准高斯初始化）会导致以 n^{-1} 的幂次为形式的偏离，其中 n 表示网络宽度。相比之下，本文采用更简单的方法，使用埃奇沃思级数技术，以渐近级数的形式显式计算偏离高斯性的修正项，级数以 $n^{-\frac{1}{2}}$ 为变量，其系数由参数初始化的统计特性决定。

–

论文 [1] 也研究了深度神经网络。在该设定下，研究特别表明，在初始化时，通过增大所有隐藏层的宽度，可以使无限极限神经正切核∞ 与有限宽度神经切线核(n)之间的差异任意小。他们还表明，在训练之后，通过增大宽度，可以使有限宽度网络的输出与相应无限宽度网络的输出之间的差异任意小。与本文展示的结果相比，[1] 的明显优势在于其研究的是深层网络，而我们只考虑单层网络。尽管如此，他们给出的界并未明确量化地依赖于各层的宽度，而且其结果仅适用于 ReLU 非线性；而本文在很大程度上弥补了这一局限。

[18] 中的结果给出了关于宽度 n 的各种估计，用于控制神经正切核（NTK）的行为、该对象的高阶泛化能力，以及训练期间的输出。再一次，他们的结果适用于深度神经网络，但遗憾的是，他们提供的估计依赖于训练时间 t 的递增函数；尤其是，这一缺陷使得无法控制 $t \to \infty$ 极限。尽管如此，[18] 给出的估计足以表明，通过适当增加宽度 n，训练数据的输出可以任意接近其标签。

在前面提到的广泛工作中，[13] 还可以找到关于宽神经网络输出行为的结果，其中表明，训练后，输出与标签之间的差异可以控制在n^{-2} 阶的误差范围内，其中n 决定了网络的宽度。

关于训练期间输出的行为，据我们所知，[10, 9] 中的结果与本文所呈现的结果重合度最高。该研究同样考虑浅层网络，并且也得到了关于训练时间的一致估计，以宽度参数控制输出动力学。我们认为，与 [10]相比，本文的结果有以下优点：i) [10] 中的结果仅限于 ReLU 非线性；ii) 要求输出层参数的初始化服从均匀分布；iii) 本文在很大程度上消除了这两项限制。为此，我们需要克服一些新的挑战：由于无法再依赖 ReLU 非线性所允许的简化显式计算，我们必须处理不一定具有紧支撑的输出；而在 [10] 中，通过上述对输出参数采用均匀分布的严格初始化，输出具有紧支撑。最后，我们还提供了一种基于连续性/自助法方法的构造性证明。关于这一背景的其他相关结果可见 [8]。

其他探讨宽神经网络、且可能适用于本文所发展的技术的有趣相关工作，例如 [6]。

运用本文所发展的技术可以继续探索一个重要且有趣的方向：沿着[3, 23, 24, 25, 28, 29, 2]的思路，进一步研究过参数化的后果。

Chen 等人的论文 [14] 在初始化时和训练过程中，针对经验测度空间上的测度涨落，得出了均场理论结果，其阶为$m^{-1/2}$，其中m 是单层神经网络的神经元数量。若能深入理解这种针对测度空间上的概率测度的均场理论方法与我们采用的、针对函数空间上的测度的随机场理论方法之间的关系，将会非常有趣；后者涉及 R^4 上的$C([a,b],\mathrm{R})$。

沿着大致相似的思路，研究如何将我们的方法改编用于 [30],等其他参数化方式，并进一步探讨特征学习，也可能很有意义。

未来工作中，尝试理解如何运用我们的技术来更好地理解物理信息神经网络及其中出现的一些有趣现象，也将是一项有意思的研究。比如参见 [19, 27]。

2. 初始化时偏离高斯性

2.1. 要考虑的宽神经网络。我们考虑一个具有一个隐藏层且宽度为 $n \in \mathbb{N}$、输入和输出取值为实数的神经网络。为方便起见，我们进一步假设输入位于一个固定的非空紧区间 $[a,b] \subset \mathrm{R}$ 中。这意味着，对于每个参数选择 $\theta \in \mathrm{R}^{P_n}$，其中 $P_n = 4n$，这是以下内容的简写形式

$$\theta := ((W_1^{(2)},\dots,W_n^{(2)},W_1^{(1)},\dots,W_n^{(1)}),(b_1^{(2)},\dots b_n^{(2)},b_1^{(1)},\dots,b_n^{(1)})) \in (\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^n \times \mathbb{R}^n),$$

分别由权重和偏置组成，我们构造一个函数

$$f_\theta^{(n)} : [a,b] \to \mathbb{R},$$

给定为

$$f_\theta^{(n)}(x) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(W_j^{(2)} \, \sigma\left(W_j^{(1)} x + b_j^{(1)} \right) + b_j^{(2)} \right), \tag{2.1}$$

其中 σ： $\mathrm{R} \to \mathrm{R}$ 是一个称为激活函数的非线性连续标量函数。因此，对于每个 $n \in \mathbb{N}$，该神经网络是如下映射:

$$\begin{aligned} \mathcal{F}^{(n)} : \mathbb{R}^{4n} &\longrightarrow C([a,b],\mathbb{R}) \subset \mathrm{Map}([a,b],\mathbb{R}) \\ \theta &\longmapsto f_\theta^{(n)}. \end{aligned} \tag{2.2}$$

2.2. 初始化时函数空间上诱导的概率测度。在初始化时，参数 θ 从 R^{4n}上的一个概率测度中随机选取，我们将其记为 $Q^{(n)}$，其密度为 $\nu_n(\theta)$，满足 $dQ^{(n)} = \nu_n(\theta) d\theta$，其中 $d\theta := d\theta^1 \ldots d\theta^{4n}$。通过在 F (n) 下取推前，我们得到无限维空间上的一列概率测度 $\mathrm{P}^{(n)}$，该空间由 $[a,b]$,上的连续函数构成

$$\mathcal{C} := C([a,b],\mathbb{R}), \qquad \mathbb{P}^{(n)} := \left(\mathcal{F}^{(n)}\right)_* (Q^{(n)}), \tag{2.3}$$

其支撑于锥上,

$$\mathcal{C}^{(n)} := \mathcal{F}^{(n)}(\mathbb{R}^{4n}) \subset \mathcal{C} \subset \mathrm{Map}([a,b],\mathbb{R}), \tag{2.4}$$

其最大维数为 $3n+1$。在自然假设下（我们将在下文回顾这些假设），中心极限定理意味着序列 $\mathrm{P}^{(n)}$收敛于 $\mathrm{C} = C([a,b],\mathrm{R})$ 上的高斯测度,

$$\mathbb{P}_C^{(\infty)} = \lim_{n\to\infty} \mathbb{P}^{(n)}, \tag{2.5}$$

其中 C 表示协方差。粗略来说，分布 $\mathrm{P}^{(n)}$ 定义了一个（通常为）非高斯随机场序列，该序列收敛于一个高斯随机场。

本工作的一个主要目标将是研究$\mathrm{P}^{(n)}$ 偏离高斯性的情况。根据柯尔莫哥洛夫扩张定理，我们知道，$\mathrm{Map}([a,b],\mathrm{R})$上的概率测度由在有限个点处对场进行求值得到的所有可能的有限维分布定义，并且也由这些分布确定。更确切地说，对于 $k \in \mathbb{N}$ 和 $x_1, \ldots, x_k \in [a,b]$，记 $p_{x_1,\ldots,x_k}$为求值映射

$$\begin{aligned} p_{x_1,\ldots,x_k} : \mathrm{Map}([a,b],\mathbb{R}) &\longrightarrow \mathbb{R}^k \\ f &\longmapsto (f(x_1)\ldots, f(x_k)). \end{aligned}$$

因此，我们可以考虑复合映射

$$\begin{aligned} \mathcal{F}_{x_1,\ldots,x_k}^{(n)} = p_{x_1,\ldots,x_k} \circ \mathcal{F}^{(n)} : \mathbb{R}^{4n} &\longrightarrow \mathrm{Map}([a,b],\mathbb{R}) \to \mathbb{R}^k, \\ \theta &\longmapsto (f_\theta(x_1), \cdots, f_\theta(x_k)), \end{aligned} \tag{2.6}$$

以及测度$Q^{(n)}$相应的推前，

$$\mathbb{P}^{(n)}_{x_1,\dots,x_k} := \left(\mathcal{F}^{(n)}_{x_1,\dots,x_k}\right)_* (Q^{(n)}). \tag{2.7}$$

根据构造，测度 $\mathbf{P}^{(n)}_{x_1,\dots,xk}$ 满足以下两个相容条件

$$\int_{I_1\times\dots\times I_k} d\mathbb{P}^{(n)}_{x_{\sigma(1)},\dots,x_{\sigma(k)}} = \int_{I_{\sigma^{-1}(1)}\times\dots\times I_{\sigma^{-1}(k)}} d\mathbb{P}^{(n)}_{x_1,\dots,x_k}, \tag{2.8}$$

对于 $I_i \subset [a,b]$ 以及 $\sigma\{1,\dots,k\}$ 的一个置换，并且

$$\int_{I_1\times\dots\times I_k\times\mathbb{R}\times\dots\times\mathbb{R}} d\mathbb{P}^{(n)}_{x_1,\dots,x_k,x_{k+1},\dots,x_{k+m}} = \int_{I_1\times\dots\times I_k} d\mathbb{P}^{(n)}_{x_1,\dots,x_k}, \tag{2.9}$$

对于 $m\in\mathbb{N}$。反过来，根据柯尔莫哥洛夫扩张定理（ [31]第 37 页定理 9.2），任意一族有限维分布，

$$\left\{\mathbb{P}^{(n)}_{x_1,\dots,x_k}\right\}_{k\in\mathbb{N},x_1,\dots,x_k\in[a,b]}, \tag{2.10}$$

满足一致性条件 (2.8)、(2.9)，即可定义 $\mathrm{Map}([a,b],\mathbb{R})$ 上唯一的测度 $\mathbf{P}^{(n)}$。

2.3. 收敛到高斯随机场。回忆一下，根据定义，若函数空间 $\mathrm{Map}([a,b],\mathbb{R})$（或 $\mathbf{C}$）上的测度 $\mathbf{P}$ 满足：对于所有$k\in\mathbb{N},x_1,\dots,x_k\in[a,b],$，其在 $\mathbb{R}^k$ 上的所有测度 $\mathbf{P}_{x_1,\dots,xk}$ 均为高斯分布，则称 $\mathbf{P}$ 为高斯测度。

现在考虑如下标准情形：隐藏层中的所有感知机彼此独立，且其参数服从相同分布。在这种情况下，正如前文所述，当人工神经元的数量 $n\to\infty$ 时，测度 $\mathbf{P}^{(n)}$ 收敛到 $\mathrm{Map}([a,b],\mathbb{R})$ 上的高斯测度。

事实上，只需考察函数 $f^{(n)}_\theta$ 的表达式 (2.1)，便可轻易看出这一点。该表达式可写为

$$f^{(n)}_\theta(x) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} p_{\hat\theta_j}(x), \tag{2.11}$$

其中 $\hat\theta = (W(2),W(1),b^{(2)},b^{(1)}) \in \mathbb{R}^4$，

$$p_{\hat\theta_j}(x) = W^{(2)}_j\,\sigma\left(W^{(1)}_j x + b^{(1)}_j\right) + b^{(2)}_j,$$

表示第 j 个感知机。从现在起，我们假设所有元组 $(W(2)_j, W(1)_j, b^{(2)}{}_j, b^{(1)}{}_j) \sim (W(2),W(1),b^{(2)},b^{(1)})$ 均为独立同分布，因此这些感知机也独立同分布。此外，我们假设 $W^{(2)}$ 和 $b^{(2)}$ 均为零均值。于是，中心极限定理适用于 (2.11)，并得到

$$\lim_{n\to+\infty} \mathbb{P}^{(n)} = \mathbb{P}^{\mathbf{C}}, \tag{2.12}$$

其中，Pc表示 Map([a,b],R) 上的高斯测度，其协方差 C 由以下作用于 $[a,b]\times[a,b]\subset\mathbb{R}^2$, 的函数给出

$$\mathbf{C}(x,x')=\mathbb{E}_f[f(x)f(x')]:=\mathbb{E}_{\hat{\theta}}[p(x)p(x')]. \tag{2.13}$$

等价地，$\mathbb{R}^k$ 上的有限维分布 $\mathbb{P}^{\mathbf{C}}_{x_1,\dots,x_k}$ 是高斯分布，其协方差 $k\times k$矩阵由 $C_{ab}:=\mathbf{C}(x_a,x_b)$ 给出。

$$d\mathbb{P}^{\mathbf{C}}_{x_1,\dots,x_k}(y_1,\dots,y_k)=\frac{e^{-\frac{1}{2}C^{-1}_{ab}y_ay_b}}{\sqrt{(2\pi)^k\det(C)}}dy^1\dots dy^k, \tag{2.14}$$

其中我们对重复指标采用了爱因斯坦求和约定。我们将计算一个对大n值有效的渐近展开，从而细化这一结果。精确表述请参见定理 1。

注释 2.

- 我们假设 C_{ab} 是非退化的。如果 x_a's 彼此不同，并且激活函数 σ 足够正规且为非多项式函数，事实证明情况确实如此（更多详情请参见 [7]）。

- 由 [26] 中的定理 *1.1.1* 可知，若 $x\mapsto\mathbf{C}(x,x')$属于 C^{k+1}类，对所有 $x'\in[a,b]$,均成立，则测度 $\mathbb{P}$c 的支撑集是从 $[a,b]$ 到 $\mathbb{R}$ 的 C^k类函数集合。这显然在 σ 本身属于 C^{k+1}类时成立。

2.4. 偏离高斯性的定量分析。本节陈述我们的主要结果之一，即定理 1，以及它的一个简单推论，该推论以例 1 的形式给出。对于较大但有限的 n，我们估计有限宽度分布 $\mathbb{P}^{(n)}$ 与高斯分布 $\mathbb{P}^{\mathbf{C}}$之间的偏差。这些结果的证明见附录 B。

我们将陈　述由单个感知机生成的随机函数的结果　　n

$$p(x):=p_{\hat{\theta}}(x)=W^{(2)}\sigma(W^{(1)}x+b^{(1)})+b^{(2)}. \tag{2.15}$$

接下来，对于 $x_1,\dots,x_k\in\mathbb{R}$ 和 $r_1,\dots,r_k\in\mathbb{N}_0$，我们记作

$$\mu_{r_1,\dots,r_k}:=\mu_{r_1,\dots,r_k}(x_1,\cdots,x_k)=\mathbb{E}\left[p(x_1)^{r_1}\dots p(x_k)^{r_k}\right],$$

随机变量（$p(x_1)$，..，$p(x_k)$）的矩，以及

$$C(x_a,x_b)=\mathbb{E}[p(x_a)p(x_b)],$$

其协方差矩阵。进一步地，对于点 $x_1,\dots,x_k$ 的每一种取值，我们定义一组实数$\lambda_{r_1,\dots,r_k}=\lambda_{r_1,\dots,r_k}(x_1,\dots,x_k)$，称为（ $p(x_1),\dots,p(x_k)$ ）的累积量。这些累积量由 $\mu_{r_1,\dots,r_k}$ 确定，具体要求关于变量 $q_1,\dots,q_k$ 的以下形式幂级数满足等式

$$\sum_{r_1+\dots+r_k=1}^{\infty}\lambda_{r_1,\dots r_k}\frac{(q_1)^{r_1}\dots(q_k)^{r_k}}{r_1!\dots r_k!}=\sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{j}\left(\sum_{s_1+\dots+s_k=1}^{\infty}\mu_{s_1,\dots s_k}\frac{(q_1)^{s_1}\dots(q_k)^{s_k}}{s_1!\dots s_k!}\right)^j, \tag{2.16}$$

现在我们已经可以陈述主要结果了，其证明将在 B.2 中给出。

定理 1。 设元组 $\hat{}\theta^{(i)}=(W^{(2)i}、b^{(2)i}、W^{(1)i}、b^{(1)i})\sim\hat{}\theta$对于所有 $i\in\{1、\dots、n\}$均为独立同分布。另设 $W^{(2)i}、b^{(2)i}$的均值为零，且 $p_{\hat{\theta}}(x)$的所有矩对于

任意 $x \in \mathbb{R}$均有限，并令

$$\mathcal{F}^{(n)}(\theta)(x) = f_\theta^{(n)}(x) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(W_j^{(2)}\, \sigma\left(W_j^{(1)} x + b^{(1)} \right) + b_j^{(2)} \right). \tag{2.17}$$

那么，对于所有 $x_1, \ldots, x_k \in [a,b]$ 以及 $l \in \mathbb{N}$，这些测度

$$d\mathbb{P}^{(n)}_{x_1,\ldots,x_k}(y_1,\ldots,y_k) = \nu_{x_1,\ldots,x_k}(y_1,\ldots,y_k)dy^1 \cdots dy^k\,,$$

在 $\mathbb{R}^k$ 上具有密度 $\nu_{x_1,\ldots,x_k}(y_1,\ldots,y_k)$，其表达式为

$$\exp\left(\sum_{r_1+\ldots+r_k>2}^{l} \frac{1}{n^{\frac{r_1+\ldots+r_k}{2}-1}} \frac{\lambda_{r_1,\ldots r_k}}{r_1!\ldots r_k!} (-\partial_{y_1})^{r_1} \ldots (-\partial_{y_k})^{r_k} \right) \frac{e^{-\frac{1}{2}C^{-1}_{ij} y_i y_j}}{\sqrt{(2\pi)^k \det(C)}} + o(n^{-l/2+1})\,. \tag{2.18}$$

注释3。 值得探讨的是，是否可以对 $W^{(2)}{}_i$、$b^{(2)}{}_i$ 和 $W^{(1)}{}_i$、$b^{(1)}{}_i$ 的分布，也就是对感知机施加某些条件，从而加快收敛到高斯过程。

如果能使 $n^{-1/2}$ 阶项消失，就能实现这一点；而当 $r_1 + \ldots + r_k = 3$ 时，令 $\mu_{r_1,\ldots,r_k} = 0$ 即可做到。

对 $W^{(2)}$、$b^{(2)}$ 的分布施加一些合理条件，即可保证三阶矩消失。例如，除了 $W^{(2)}$、$b^{(2)}$ 的均值为零外，它们的三阶矩也为零。若它们关于0 （均值）对称分布，这一条件便成立；正态分布就是如此。

为了说明如何在实践中使用该定理，我们给出以下直接推论，证明见附录B.3。

例1 （情形 $k = 1$）。 设 $x \in \mathbb{R}$。那么，对于 （$W^{(2)}{}_i$、$b^{(2)}{}_i$、$W^{(1)}{}_i$、$b^{(1)}{}_i$）以及定理 *1* 中的 $f_\theta(\cdot)$，我们有

$$\begin{aligned} \nu_x(y) = {} & \frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + \frac{1}{\sqrt{n}} \frac{\mu_3}{12\sqrt{2}\mu_2^{3/2}} H_3\left(\frac{y}{\sqrt{2\mu_2}} \right) \frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} \\ & + \frac{1}{n}\left(\frac{\mu_4 - 3\mu_2^2}{96\mu_2^2} H_4\left(\frac{y}{\sqrt{2\mu_2}} \right) + \frac{\mu_3^2}{576\mu_2^3} H_6\left(\frac{y}{\sqrt{2\mu_2}} \right) \right) \frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + O(n^{-3/2}), \end{aligned}$$

其中 $H_n(\cdot)$ 表示第 n 阶厄米多项式，且 $\mu_r = \mathbb{E}[p(x)^r]$。

3. 函数空间中的训练几何

3.1. 训练与神经切线核。在训练过程中，参数 $\{\theta(t)\}_{t\geq 0}$ 按照代价函数 $\mathrm{L}^{(n)}\colon \mathrm{R}^{4n} \to \mathrm{R}$ 的负梯度流演化，即

$$\dot{\theta} = -\nabla \mathcal{L}^{(n)}, \tag{3.1}$$

其中，梯度是相对于 R^{4n} 上的平坦对角度量计算的，

$$\gamma^{(n)} = \sum_{j=1}^{4n} d\theta_j^2.$$

实际上，正是逆度量参与了函数的梯度向量场的定义，

$$\nabla\mathcal{L}^{(n)}: \qquad \gamma^{(n)}\left(\nabla\mathcal{L}^{(n)},\cdot\right)=d\,\mathcal{L}^{(n)},$$

或者，在局部坐标中，

$$\nabla\mathcal{L}^{(n)}=\sum_{j=1}^{4n}\frac{\partial\mathcal{L}^{(n)}}{\partial\theta_j}\frac{\partial}{\partial\theta_j}.$$

在实践中，人们将损失函数与训练数据 $Z=\{(X_a,Y_a)\}_{a=1}^{N}$相关联，其中我们假设$X_a\neq X_b$ 对于$a\neq b$ 成立，并且每个 (X_a,Y_a) 表示一个输入输出对，也就是说，我们期望近似的函数 f 满足$f(X_a)=Y_a$对所有$a=1,\ldots,N$均成立。接着，在 $\mathrm{Map}([a,b],\mathrm{R})$ 上考虑一个损失函数 $\mathrm{L}_Z\colon \mathrm{Map}([a,b],\mathrm{R})\longrightarrow\mathrm{R}$，例如二次损失

$$\mathcal{L}_Z(f):=\frac{1}{2}\sum_{i=1}^{N}(Y_i-f(X_i))^2. \tag{3.2}$$

那么，对于每个n，我们在 R^{4n} 上都有一个损失函数，

$$\mathcal{L}_Z^{(n)}=\mathcal{L}_Z\circ\mathcal{F}^{(n)},$$

以及相应的梯度流。令 $\theta(t)$ 为该流的一条积分曲线，对于 $x\in\mathrm{R}$，考虑曲线 $y(t):=f^{(n)}_{\theta(t)}(x)$。由 $\dot{y}=\sum_{j=1}^{4n}\frac{\partial f^{(n)}_{\theta}}{\partial\theta_j}(x)\,\dot{\theta}_j$，我们得到，

$$\begin{aligned}\dot{y}&=\sum_{j=1}^{4n}\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(x)\sum_{l=1}^{n}\left(Y_l-f_\theta^{(n)}(X_l)\right)\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(X_l)\\&=\sum_{l=1}^{N}\left(Y_l-f_\theta^{(n)}(X_l)\right)\sum_{j=1}^{4n}\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(x)\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(X_l).\end{aligned}$$

引入神经切线核

$$\mathrm{NTK}^{(n)}_{z_1z_2}(\theta):=\sum_{j=1}^{4n}\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(z_1)\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(z_2)$$

我们可以将之前的计算改写为

$$\dot{y}=\sum_{l=1}^{N}\mathrm{NTK}^{(n)}_{xX_l}(\theta)\left(Y_l-f_\theta^{(n)}(X_l)\right). \tag{3.3}$$

在我们所考虑的设定中，即当 $f_\theta^{(n)}$ 由公式 2.11 给出时，神经切线核为

$$\mathrm{NTK}^{(n)}_{x_1x_2}(\theta)=\sum_{j=1}^{4n}\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(x_1)\frac{\partial f_\theta^{(n)}}{\partial\theta_j}(x_2)=\frac{1}{n}\sum_{l,m=1}^{n}\sum_{j=1}^{4n}\frac{\partial p_l}{\partial\theta_j}(x_1)\frac{\partial p_m}{\partial\theta_j}(x_2)=\frac{1}{n}\left(\sum_{m=1}^{n}\sum_{j=1}^{4n}\frac{\partial p_m}{\partial\theta_j}(x_1)\frac{\partial p_m}{\partial\theta_j}(x_2)\right), \tag{3.4}$$

因为 $\frac{\partial p_l}{\partial\theta_j}(x_1)\frac{\partial p_m}{\partial\theta_j}(x_2)=0$ 对于 $l\neq m$。因此，大数定律保证，当$n\to+\infty$

$$\mathrm{NTK}^{(n)}_{x_1x_2}(\theta)\to\mathrm{NTK}^{\infty}_{x_1x_2}:=\mathbb{E}\left[K(x_1,x_2)\right]\text{ , a.s. in }\theta\text{ ,} \tag{3.5}$$

其中 $K(x_1,x_2)$ 是随机变量

$$K(x_1,x_2) := \sum_j \frac{\partial p}{\partial \theta_j}(x_1)\frac{\partial p}{\partial \theta_j}(x_2)\,.$$

事实上，众所周知，大数定律有如下定量版本 [11,定理 2.5。11]

$$\mathrm{NTK}^{(n)}_{x_1x_2}(\theta) = \mathrm{NTK}^{\infty}_{x_1x_2} + o(n^{-1/2}(\log n)^{1/2+}) \quad \text{a.s.}\,, \tag{3.6}$$

其中，和通常一样，$a+$ 表示任何严格大于a的数。

3.2. 将梯度流“推前”至函数空间来描述训练。与上一节一样，设 $\{\theta(t)\}_{t\in I\subset\mathbf{R}}$ 为 $\mathrm{L}^{(n)}{}_Z = \mathrm{L}_Z \circ \mathrm{F}^{(n)}$ 的负梯度的一条积分曲线，即 $\dot\theta = -\nabla \mathrm{L}^{(n)}{}_Z$，其中 $\nabla \mathrm{L}^{(n)}{}_Z$ 是 R^{4n} 上由下式确定的向量场

$$\nabla \mathcal{L}^{(n)}_Z(\alpha) = (\gamma^{(n)})^{-1}(d\mathcal{L}^{(n)}_Z, \alpha),$$

对所有 1-形式 α成立。这里我们使用了

$$(\gamma^{(n)})^{-1} = \sum_{j=1}^{4n} \frac{\partial}{\partial\theta_j}\otimes\frac{\partial}{\partial\theta_j}$$

表示由欧几里得度量 $\gamma^{(n)}$诱导的余向量上的度量。此外，由于$d\mathrm{L}^{(n)}{}_Z = d\mathrm{L}_Z \circ d\mathrm{F}^{(n)}$，我们实际上有

$$\nabla \mathcal{L}^{(n)}_Z(\cdot) = (\gamma^{(n)})^{-1}(d\mathcal{L}_Z \circ d\mathcal{F}^{(n)}, \cdot). \tag{3.7}$$

先记住这个表达式，接下来用 $\mathrm{F}^{(n)}$ 沿 $\{\mathrm{F}^{(n)}(\theta(t))\}_{t\in I\subset\mathbf{R}} \subset \mathrm{Map}([a,b],\mathrm{R})$ 推前逆度量，也就是我们所考虑的积分线的像。在 $\mathrm{F}^{(n)}(\theta(t))$ 的任意一点，我们可以将 $\gamma^{(n)}$ 定义为

$$\left(\mathcal{F}_*(\gamma^{(n)})^{-1}\right)_{\mathcal{F}^{(n)}(\theta(t))}(\alpha,\beta) = (\gamma^{(n)})^{-1}_{\theta(t)}\left(\alpha\circ d\mathcal{F}^{(n)}_{\theta(t)}, \beta\circ d\mathcal{F}^{(n)}_{\theta(t)}\right),$$

对于 $\mathrm{F}^{(n)}(\theta(t))$ 处的任意余向量 α,β。然后，利用这个度量，我们可以沿 $\{\mathrm{F}^{(n)}(\theta(t))\}_{t\in I\subset\mathbf{R}}$ 定义 L_Z 的梯度：

$$\nabla\mathcal{L}_Z(\cdot) = \left(\mathcal{F}^{(n)}_*(\gamma^{(n)})^{-1}\right)(d\mathcal{L}_Z,\cdot)$$

展开 $\mathrm{F}^{(n)}{}_*(\gamma^{(n)})^{-1}$ 的定义，并利用方程 3.7 中 $\nabla \mathrm{L}^{(n)}{}_Z$ 的定义，可得

$$(\nabla\mathcal{L}_Z)(\cdot) = (\gamma^{(n)})^{-1}\left(d\mathcal{L}_Z\circ d\mathcal{F}^{(n)}, (\cdot)\circ d\mathcal{F}^{(n)}\right) = d\mathcal{F}^{(n)}(\nabla\mathcal{L}^{(n)}_Z(\cdot)),$$

从而得出结论：

$$\nabla\mathcal{L}_Z = d\mathcal{F}^{(n)}(\nabla\mathcal{L}^{(n)}_Z).$$

[1]这里，给定 $\mathrm{F}^{(n)}(\theta(t))$ 处的一个余向量 α，$\alpha\circ d\mathrm{F}^{(n)}$ 表示 $\theta(t)$ 处的余向量；在切向量 V 上取值时，其值为 $\alpha(d\mathrm{F}^{(n)}\,\theta(t)V)$。

特别地，$F^{(n)}(\theta(t))$ 的演化速度向量为 $-\nabla L_Z$，即 L_Z 关于逆度量 $(\gamma^{(n)})^{-1}$ 的负梯度。确实，

$$\frac{d}{dt}(\mathcal{F}^{(n)}(\theta(t))) = d\mathcal{F}(\dot{\theta}(t)) = -(d\mathcal{F})_{\theta(t)}(\nabla\mathcal{L}_Z^{(n)})_{\theta(t)} = -(\nabla\mathcal{L}_Z)_{\mathcal{F}(\theta(t))},$$

因此，我们得到了神经切线核流的一种不同解释：它是相对于逆度量的推前，对 L_Z 的负梯度进行流动的一种特殊情形。

备注4。 注意，我们只沿梯度流线的像定义 $\nabla\mathcal{L}_Z$。虽然理想情况下我们希望在 $\mathcal{F}$ 的整个像上定义它，但一般而言这是不可能的，因为这个映射可能严重地不具有单射性。

3.3. 无限网络的演化与输出的收敛。本节回顾关于无限宽神经网络在训练过程中演化的一些众所周知的事实。总而言之，网络的输出以线性方式演化，其依据是神经切线核，[20]。

给定训练集 $\{X_i, Y_i\}_{i=1}^N$，我们将与核为神经正切核$^\infty$的“无限宽度”网络（其初始化与有限宽度网络相同）相关联的函数定义为函数 $f_\infty(\cdot, t) : R \to R$，它定义为下列初值问题的解。

$$\begin{cases} \partial_t f_\infty(x,t) = -\sum_{j=1}^N \mathrm{NTK}^\infty_{xX_j}(f_\infty(X_j,t) - Y_j) \\ f_\infty(x,0) = f^{(n)}_{\theta(0)}(x)\,, \end{cases} \tag{3.8}$$

其中 $f^{(n)}_{\theta_0}(\,)(x)$表示在$x$ 处取值的初始化（有限宽度）网络。这是一个常微分方程组，对于所有 $t \in R$都可以通过闭式形式求解。该结果在附录 C 中推导得出，现为完整起见将其重新表述如下：首先得到 $E_i(t) = f_\infty(X_i,t) - Y_i$ 的线性系统，然后演化方程 3.8 得出

$$\dot{E}_i = -\sum_{l=1}^N \mathrm{NTK}^\infty_{x_i X_l} E_l$$

其解为 $E(t) = e^{-\mathrm{NTK}^\infty t}E(0)$，因此

$$f_\infty(X_i,t) = Y_i - \sum_{j=1}^N \left(e^{-\mathrm{NTK}^\infty t}\right)_{ij}(f_\infty(X_j,0) - Y_j), \tag{3.9}$$

其中 $\left(e^{-\mathrm{NTK}^\infty t}\right)_{ij}$ 表示矩阵 $e^{-\mathrm{NTK}^\infty t}$ 的第 (i,j) 个元素。

现在，我们利用这一点来计算在一般的 R 上对 f_∞ 在 $x \in$ 处的求值的演化。正如我们将看到的，而且这在 3.8 中已隐含体现，其演化完全由 f_∞ 在标签 $\{X_i\}_{i=1}^N$ 处求值后的演化决定。事实上，利用 $\{E_i(t)\}_{i=1}^N$，演化方程 3.8 可改写为

$$\partial_t f_\infty(x,t) = -\sum_{l=1}^N \mathrm{NTK}^\infty_{x_{N+1}X_l} E_l(t).$$

利用 (神经正切核$^{\infty}$)$_{ij}$:= 神经正切核$^{\infty}_{X_iX_j}$ 可逆这一事实 [7], , 我们可以写成 $E_l(t) = -\sum_{m=1}^{N}$(神经正切核$^{\infty}$)$^{-1}_{lm}$ $\dot{E}_m$, 代入上式可得

$$\partial_t f_\infty(x,t) = \sum_{l,m=1}^{N} \mathrm{NTK}^{\infty}_{x_{N+1}X_l}(\mathrm{NTK}^{\infty})^{-1}_{lm}\dot{E}_m(t).$$

然后, 我们可以利用微积分基本定理得到

$$\begin{aligned} f_\infty(x,t) &= f_\infty(x,0) + \int_0^t \sum_{l,m=1}^{N} \mathrm{NTK}^{\infty}_{xX_l}(\mathrm{NTK}^{\infty})^{-1}_{lm}\dot{E}_m(s)ds \\ &= f_\infty(x,0) + \sum_{l,m=1}^{N} \mathrm{NTK}^{\infty}_{xX_l}(\mathrm{NTK}^{\infty})^{-1}_{lm}(E_m(t) - E_m(0)) \\ &= f_\infty(x,0) + \sum_{l,m=1}^{N} \mathrm{NTK}^{\infty}_{xX_l}(\mathrm{NTK}^{\infty})^{-1}_{lm}(f_\infty(X_m,t) - f_\infty(X_m,0)), \end{aligned}$$

再利用 3.9, 可得

$$f_\infty(x,t) = f_\infty(x,0) + \sum_{l,m,p=1}^{N} \mathrm{NTK}^{\infty}_{xX_l}(\mathrm{NTK}^{\infty})^{-1}_{lm}(e^{-\mathrm{NTK}^{\infty}t} - 1)_{mp}(f_\infty(X_p,0) - Y_p), \tag{3.10}$$

这就是一个以$f_\infty(x,0)$ 初始化的无限宽度网络在训练过程中的演化。

注 5. 特别地, 我们得到

$$\lim_{t\to+\infty} f_\infty(x,t) = f_\infty(x,0) - \sum_{l,m=1}^{N} \mathrm{NTK}^{\infty}_{x_{N+1}X_l}(\mathrm{NTK}^{\infty})^{-1}_{lm}(f_\infty(X_m,0) - Y_m),$$

这应与核回归进行比较。

3.4. 宽但有限的网络输出的演化与收敛。本节给出一些关于宽度为 n 的网络输出演化的结果。正如前文所述, 这种演化由 (3.3) 决定, 因此与神经正切核$^{(n)}$($\theta(t)$ 的演化密切相关。此外, 当 $n \to \infty$ 时, 这个有限宽度神经切线核几乎必然收敛到神经正切核$^{\infty}$, 后者在参数空间中为常数。因此, 在无限宽度极限下, 输出的演化由线性流决定。我们将利用这一事实（证明见附录 C）, 通过自助法／连续性论证, 以线性无限宽度动力学来控制有限宽度网络的非线性学习动力学。

现在我们可以陈述关于输出演化的主要结果:

定理 2. 考虑序列 $\theta_0 = (W^{(2)}_{0,j}, b^{(2)}_0, \mu, W^{(1)}_{0,\mu}, b^{(1)}_0, \mu)_{\mu\in\mathbb{N}}$, 其中的元组由具有有限三阶矩的随机变量组成, 且独立同分布。另考虑满足以下条件的非多项式激活函数。

$$|\sigma(z)| \le C_\sigma(|z|+1) \quad and \quad |\sigma'(z)| + |\sigma''(z)| \le C_\sigma\,, \tag{3.11}$$

对某个 $C_\sigma > 0$，且所有 $z \in \mathbb{R}$ 均成立。对于任意一组 $4n$ 个参数 $\theta = (W^{(2)\,i},\ b^{(2)\,i},\ W^{(1)\,i},\ b^{(1)\,i})_{i=1,\dots,n}$，令

$$f_\theta^{(n)}(x) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(W_j^{(2)}\, \sigma\left(W_j^{(1)} x + b_j^{(1)} \right) + b_j^{(2)} \right)$$

对应的宽度为 n 的神经网络。

另假设给定训练集 $\{X_i,\ Y_i\}_{i=1}^N$，其中 $X_i \neq X_j$，对于 $i \neq j$，并令 $\lambda_\infty > 0$ 为矩阵 神经正切核$\Theta^\infty_{X_i X_j}$ 的最小特征值。[2]

令 $t \mapsto \theta(t)$ 表示训练期间参数的演化，其初始化为 $\theta(0) = \theta_0$，训练采用梯度下降法 (3.1)，并使用均方损失 (3.2)。

对于任意给定的 $\delta > 0$，存在 $m > 0$ 和 $C_\delta > 0$，使得若 $n \geq m$，则以概率 $1 - \delta$（在参数初始化时），对所有 $t \geq 0$，均有：

(i) 有限宽度神经切线核与其无限宽度极限之间的差异：

$$|\mathrm{NTK}_{zw}^{(n)}(\theta(t)) - \mathrm{NTK}_{zw}^{\infty}| \leq \frac{C_\delta\, (\log n)^{1/2+}}{\sqrt{n}} + \sup_{k=1,\dots,N} |f_{\theta(0)}^{(n)}(X_k) - Y_k| \frac{NC}{\sqrt{n}(\lambda_\infty - \varepsilon)}, \tag{3.12}$$

对于所有 $z, w \in [a, b]$，且其中 $\varepsilon > 0$ 可通过增大 n 任意变小。

(ii) 训练输出（有限宽度神经网络的训练输出）指数收敛至其标签：

$$|f_{\theta(t)}^{(n)}(X_i) - Y_i| \leq C \sup_{l=1,\dots,N} |f_{\theta(0)}^{(n)}(X_l) - Y_l| e^{-(\lambda_\infty - \varepsilon)t}, \tag{3.13}$$

其中，通过增大 n，可以使 $\varepsilon > 0$ 任意小。

(iii) 训练的稳定性：设f_∞ 为与无限宽度核相关联的函数（如（3.10）所定义），则，对于所有 $x \in [a,b]$,

$$|f_{\theta(t)}^{(n)}(x) - f_\infty(t,x)| \quad \leq \quad \left(1 + \sup_{i=1,\dots,N} |\mathrm{NTK}_{xX_i}^{\infty}|\right) \frac{CE_0}{\lambda_\infty - \varepsilon} \frac{(\log(n))^{1/2+}}{\sqrt{n}}, \tag{3.14}$$

where $E_0 := \sup_{k=1,\dots,N} |f_{\theta(0)}^{(n)}(X_k) - Y_k|$ *and* $\varepsilon > 0$, *as before* .

这一结果的一个直接推论如下。

推论 1(网络向其极限的指数收敛)。 在定理 *2* 的条件下，存在一个正常数 C，使得对于所有 $x \in \mathbb{R}$，都有

$$|\lim_{s \to +\infty} f_{\theta(s)}^{(n)}(x) - f_{\theta(t)}^{(n)}(x)| \leq \frac{C \sup_{l=1,\dots,N} |f_{\theta(0)}^{(n)}(X_l) - Y_l|}{\lambda_\infty - \varepsilon} e^{-(\lambda_\infty - \varepsilon)t}, \tag{3.15}$$

证明。 方程 3.13 表明，对于充分大的 n，且 $i = 1,\dots N$，有

$$|f_{\theta(t)}^{(n)}(X_l) - Y_l| \leq C \sup_{k=1,\dots,N} |f_{\theta(0)}^{(n)}(X_k) - Y_k| e^{-(\lambda_\infty - \varepsilon)t}$$

[2]由 [7] 中的结果可知， λ_∞ 为正。

我们可以利用这个估计进一步计算 $f_\theta^{(n)}(t)(x)$ 与其极限之间的差。这需要使用在方程 3.3 中推导出的 $f_\theta^{(n)}(t)(x)$ 的演化方程，其形式为

$$\frac{\partial}{\partial t}\left(f_{\theta(t)}^{(n)}(x)\right) = -\sum_{i=1}^{N} \mathrm{NTK}_{xX_i}(f_{\theta(t)}^{(n)}(X_i) - Y_i).$$

然后，通过可能改变常数C，得到

$$\begin{aligned}
|\lim_{s\to+\infty} f_{\theta(s)}^{(n)}(x) - f_{\theta(t)}^{(n)}(x)| &= |\int_t^{+\infty} \partial_t f_{\theta(s)}^{(n)}(x) ds| \\
&\leq C \sup_{l=1,\dots,N} |f_{\theta(0)}^{(n)}(X_l) - Y_l| \int_t^{+\infty} e^{-(\lambda_\infty - \varepsilon)s} ds \\
&\leq \frac{C \sup_{l=1,\dots,N} |f_{\theta(0)}^{(n)}(X_l) - Y_l|}{\lambda_\infty - \varepsilon} e^{-(\lambda_\infty - \varepsilon)t},
\end{aligned}$$

其中 sho ws that $f_\theta^{(n)}(t)(x)$ 按照陈述所言以指数速度收敛到其极限。 t. □

此外，我们指出，在深度神经网络领域，有一些有趣的结果保证损失函数在训练过程中收敛至零。例如，参见 [10]以及 [8]。

4. 近似训练期间的演化

4.1. 向量场流所描述的演化。和之前一样，给定 $x_1, \dots, x_k \in \mathbb{R}$，我们考察网络在这些点上的取值 $(f_\theta(x_1), \dots, f_\theta(x_k)) \in \mathbb{R}^k$。为简单明了起见，我们用坐标 $(y_1, \dots, y_k)$ 来参数化输出空间 $\mathbb{R}^k$。此外，在适当的时候，我们会用 $x = (x_1, \dots, x_k)$ 和 $y = (y_1, \dots, y_k)$ 来简化记号。

严格来说，描述 $y(t)$ 在训练过程中如何演化并不难。事实上，记 $\{\phi_t^{(n)}\}_{t\geq 0}$ 为 $-\nabla_\theta C\theta$ 在 $\mathbb{R}^{4n}$ 中的流，即 $\{\theta(t) = \phi_t^{(n)}(\theta)\}_{t\geq 0}$ 满足 $\dot{\theta} = -\nabla_\theta C\theta$，那么 $y(t) = f_{\phi_t^{(n)}(\theta)}$ 将按照 3.3 节所述演化。考虑初始化时的诱导测度 $\mathrm{P}_{x_1,\dots,x_k}^{(n)}$，其定义见 2.7 节。遗憾的是，尽管这种说法是正确的，却无助于明确理解概率分布如何

$$\mathbb{P}_{x_1,\dots,x_k}^{(n)}(t) := \left(\mathcal{F}_{x_1,\dots,x_k}^{(n)} \circ \phi_t\right)_* (Q_n), \tag{4.1}$$

随时间演化。为此，如果 $\{y(t)\}_{t\geq 0}$ 沿着由 $-\nabla_\theta C\theta$ 经 $\mathrm{F}_{(n)x_1,\dots,x_k} : \mathbb{R}^{4n} \to \mathbb{R}^k$推前得到的向量场的流演化，那将非常方便。然而，由于 $\mathrm{F}_{(n)x_1,\dots,x_k}$ 通常并非单射，因此无法相容地定义这样的推前向量场。

克服这一困难的一种可能方法（至少在形式上）是，转而考虑 $\mathrm{F}_{(n)x_1,\dots,x_k}$ 的图像，并在其上定义一个向量场

$$\mathrm{graph}(\mathcal{F}_{x_1,\dots,x_k}^{(n)}) = \{(\theta, y) = ((\theta_1, \dots, \theta_P), (y_1, \dots, y_k)) \in \mathbb{R}^P \times \mathbb{R}^k \mid y_i = f_\theta(x_i) \text{ for } i = 1, \dots, k\}.$$

那么，该映射

$$\varphi_{x_1,\dots,x_k} : \mathbb{R}^{4n} \to \mathrm{graph}(\mathcal{F}_{x_1,\dots,x_k}^{(n)})$$

由下式给出

$$\varphi_{x_1,\dots,x_k}(\theta_1,\dots,\theta_{4n}) = ((\theta_1,\dots,\theta_{4n}),(f_\theta(x_1),\dots,f_\theta(x_k)))$$

是一个微分同胚，因此我们可以定义向量场 $(\varphi_{x_1,\dots,x_k})_*(-\nabla_\theta C_\theta)$。通过写作 $-\nabla_\theta C_\theta = -\sum_j \frac{\partial C_\theta}{\partial \theta_j}\frac{\partial}{\partial \theta_j}$，由方程 3.3 可得

$$(\varphi_{x_1,\dots,x_k})_*(-\nabla_\theta C_\theta) = -\sum_{j=1}^{4n}\frac{\partial C_\theta}{\partial \theta_j}\frac{\partial}{\partial \theta_j} + \sum_{i=1}^{k}\left(\sum_{l=1}^{N}\mathrm{NTK}^{(n)}_{x_i X_l}(\theta)(Y_l - f_\theta(X_l))\right)\frac{\partial}{\partial y_i}.$$

于是，图 $\mathrm{graph}(F^{(n)}_{x_1,\dots,x_k})$ 上的测度 $(\varphi_{x_1,\dots,x_k} \circ \phi_t)_*(Q_n)$ 沿向量场 V 的流演化，我们可以通过沿第二个因子上的投影 π 将其向前推送来简化 $Q^{(n)}_{x_1,\dots,x_k}(t)$：$\mathrm{graph}(\psi_{x_1,\dots,x_k}) \subset \mathbb{R}^{4n}\times\mathbb{R}^k \to \mathbb{R}^k$。

于是，$f^{(n)}_\theta(x_i) = y_i$，因此

$$(\varphi_{x_1,\dots,x_k})_*(-\nabla_\theta C_\theta) = -\sum_{j=1}^{P}\frac{\partial C_\theta}{\partial \theta_j}\frac{\partial}{\partial \theta_j} + \sum_{i=1}^{k}\left(\sum_{l=1}^{N}\mathrm{NTK}^{(n)}_{x_i X_l}(\theta)(Y_l - y_l)\right)\frac{\partial}{\partial y_i}.$$

注 6。如前所述，在这种情况下，当 $n \to +\infty$ 时，神经正切核 $\mathrm{NTK}^{(n)}_{z_1 z_2}(\theta) \to$ 神经正切核 $\mathrm{NTK}^{\infty}_{z_1 z_2}$ 的极限几乎必然与 θ 无关，因此我们确实可以在 $\mathbb{R}^k$ 上定义一个向量场，定义如下：

$$V = \sum_{i=1}^{k}\left(\sum_{l=1}^{N}\mathrm{NTK}^{\infty}_{x_i X_l}(Y_l - y_l)\right)\frac{\partial}{\partial y_i}.$$

实践中，考虑到其中涉及的不确定性，对于较大的 n，使用 V 来计算演化中的一阶偏差或许是个不错的选择。下一节将对此进行讨论。

4.2. 无限宽度情形下的演化。如注 6 所述，在无限宽度极限下，我们可以在输出空间 $\mathbb{R}^k$ 上定义一个向量场 V，其坐标为 $(y^1,\dots,y^k)$，其第 i 个分量为

$$V_i := \sum_{l=1}^{N}\mathrm{NTK}^{\infty}_{x_i X_l}(Y_l - f_\theta(X_l)).$$

在这个无限宽度极限下，训练对应于 V 的流。与之前一样，我们将其流记为 $\{\Phi^\infty_t\}_{t\ge 0}$，并定义测度的密度 $\nu^\infty_{x_1,\dots,x_k}(t,y_1,\dots,y_k)$

(4.2) $$\mathbb{P}^\infty_{x_1,\dots,x_k}(t) := \lim_{n\to\infty}\mathbb{P}^{(n)}_{x_1,\dots,x_k}(t),$$

by

$$\nu^\infty_{x_1,\dots,x_k}(t,y_1,\dots,y_k)dy^1\dots dy^k = d\mathbb{P}^\infty_{x_1,\dots,x_k}(t).$$

至于演化中的密度，也可以通过 Φ^∞_{-t} 的拉回等价地定义，即

$$\nu^\infty_{x_1,\dots,x_k}(t,y_1,\dots,y_k)dy^1\dots dy^k = (\Phi^\infty_{-t})^*(\nu^\infty_{x_1,\dots,x_k}(0,y_1,\dots,y_k)dy^1\dots dy^k).$$

使用 th 简写记法 $x = (x_1,\dots,x_k)$ 和 $y = (y_1,\dots,y_k)$ 下，我们可以明确写出

(4.3) $$\nu^\infty_x(t,y) = \nu^\infty_x(0,\Phi^\infty_{-t}(y))\,|\det(d\Phi^\infty_{-t})|,$$

由此我们发现， $\nu_x^\infty(t,y)$ 满足连续性方程

$$\partial_t \nu_x^\infty + \mathrm{div}_y(\nu_x^\infty V) = 0,$$

初始条件 $\nu_x^\infty(0,y)$ 由方程 2.14 给出。

注 7。 有时，使用傅里叶变换（即与 $\nu_x^\infty(t,y)$ 相关联的特征函数）会更方便。我们将其记为

$$Z_{\nu_x}^\infty(t,p) = \int_{\mathbb{R}^k} e^{i\sum_{j=1}^k p_j y_j} \nu_x^\infty(t,y)dy.$$

计算其随时间的演化可得

$$\partial_t Z_{\nu_x}^\infty(t,p) = \int_{\mathbb{R}^k} e^{i\sum_{j=1}^k p_j y_j} \partial_t \nu_x^\infty(t,y)dy = -\int_{\mathbb{R}^k} e^{i\sum_{j=1}^k p_j y_j} \mathrm{div}_y(\nu_x^\infty V)dy,$$

分部积分后得到

$$\begin{aligned}\partial_t Z_{\nu_x}^\infty(t,p) &= -\int_{\mathbb{R}^k}\left[\mathrm{div}_y(e^{i\sum_{j=1}^k p_j y_j}\nu_x^\infty V) - i\sum_{l=1}^k p_l V_l e^{i\sum_{j=1}^k p_j y_j}\nu_x^\infty\right]dy \\ &= i\sum_{l=1}^k p_l \int_{\mathbb{R}^k} V_l e^{i\sum_{j=1}^k p_j y_j}\nu_x^\infty dy. \end{aligned} \tag{4.4}$$

在许多情况下，这可用于有效计算 $Z_{\nu_x}^\infty(t,p)$ 的演化。为举例说明，简单起见，假设 V 不依赖于 y。那么，上述演化方程变为

$$\partial_t Z_{\nu_x}(t,p) = i\sum_{l=1}^k p_l V_l Z_{\nu_x}(t,p),$$

我们可以对其积分得到

$$Z_{\nu_x}(t,p) = Z_{\nu_x}(0,p)\exp\left(i\sum_{l=1}^k p_l \int_0^t V_l(s)ds\right).$$

因此，为了求出 $\nu_x(t,y)$，我们只需应用傅里叶逆变换，得到

$$\begin{aligned}\nu_x(t,y) &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} Z_{\nu_x}(t,p)e^{-i\sum_{j=1}^k y_j p_j}dp \\ &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} Z_{\nu_x}(0,p)e^{-i\sum_{j=1}^k p_j\left(y_j - \int_0^t V_j(s)ds\right)}dp \\ &= \nu_x\left(0, y - \int_0^t V(s)ds\right).\end{aligned}$$

我们首先考虑一个重要情形：$k = N$且$x_i = X_i$，其中$i = 1,\ldots,N$。在这种情况下，我们可以将 V写为

$$V = \sum_{i=1}^N\left(\sum_{l=1}^N \mathrm{NTK}_{X_i X_l}^\infty(Y_l - y_l)\right)\frac{\partial}{\partial y_i}.$$

在这种情况下，我们可以显式地计算流 $\{\Phi_t^\infty\}_{t\geq 0}$。事实上，这已经在公式 3.9 中完成，其中$y_i$ 被替换为$f_\infty(X_i,t)$。结果为

$$\Phi_t^\infty(y) = Y + e^{-\mathrm{NTK}^\infty t}(y - Y),$$

并使用密度演化的一般公式（方程 4.3），我们得到

$$\begin{aligned} \nu_x^\infty(t,y) &= \nu_x^\infty(0,\Phi_{-t}^\infty(y))|\det(d\Phi_{-t}^\infty)| \\ &= \nu_x^\infty(0, Y + e^{\mathrm{NTK}^\infty t}(y-Y))\det(e^{\mathrm{NTK}^\infty t}) \end{aligned} \tag{4.5}$$

注释 8. 我们可以通过求解 *4.4* 来验证这一点。事实上，我们有

$$\begin{aligned} \partial_t Z_{\nu_x}^\infty(t,p) &= i\sum_{i=1}^N p_i \int_{\mathbb{R}^k} V_i e^{i\sum_{j=1}^k p_j y_j}\nu_x^\infty dy \\ &= i\sum_{i=1}^N p_i \int_{\mathbb{R}^k}\left(\sum_{l=1}^N \mathrm{NTK}_{X_iX_l}^\infty (Y_l - y_l)\right) e^{i\sum_{j=1}^k p_j y_j}\nu_x^\infty dy \\ &= i\sum_{i,l=1}^N p_i\ \mathrm{NTK}_{X_iX_l}^\infty\left(Y_l Z_{\nu_x}^\infty(t,p) - \int_{\mathbb{R}^k} y_l \nu_x^\infty\ e^{i\sum_{j=1}^k p_j y_j} dy\right) \\ &= i\sum_{i,l=1}^N p_i\ \mathrm{NTK}_{X_iX_l}^\infty\left(Y_l Z_{\nu_x}^\infty(t,p) + i\partial_{p_l} Z_{\nu_x}^\infty(t,p)\right). \end{aligned}$$

我们可以将其重写为 $\left(\partial_t - i\sum_{i,l=1}^N p_i \mathrm{NTK}_{X_iX_l}^\infty\left(Y_l + i\partial_{p_l}\right)\right) Z_{\nu_x}^\infty(t,p) = 0$，或者

$$\left(\partial_t + \sum_{i,l=1}^N p_i\ \mathrm{NTK}_{X_iX_l}^\infty\ \partial_{p_l} - i\sum_{i,l=1}^N p_i\ \mathrm{NTK}_{X_iX_l}^\infty\ Y_l\right) Z_{\nu_x}^\infty(t,p) = 0.$$

该方程可以用特征线法求解。事实上，令 $T\in\mathbb{R}\mapsto P(T)\in\mathbb{R}^N$ 为一条满足该方程的路径

$$\dot{P}_j(T) = \sum_{i=1}^N \mathrm{NTK}_{X_iX_j}^\infty P_i(T),$$

i.e. $P(T) = \exp(\text{神经正切核}^\infty T)P(0)$。然后，函数 $U(T) := Z^{\nu_x}(T, P(T))$ 不断演化彩票计划h

$$\dot{U}(T) = \partial_t Z_{\nu_x}^\infty(T,P(T)) + \sum_{j=1}^N \partial_{p_j} Z_{\nu_x}^\infty(T,P(T))\dot{P}_j(T)$$

然后，

$$\begin{aligned}\dot{U}(T) &= -\sum_{i,l=1}^{N} P_i(T)\,\mathrm{NTK}^{\infty}_{X_iX_l}\,\partial_{p_l} Z^{\infty}_{\nu_x}(T,P(T)) + i\sum_{i,l=1}^{N} P_i(T)\,\mathrm{NTK}^{\infty}_{X_iX_l} Y_l Z^{\infty}_{\nu_x}(T,P(T)) + \\ &+ \sum_{j=1}^{N} \partial_{p_j} Z^{\infty}_{\nu_x}(T,P(T)) \sum_{i=1}^{N} \mathrm{NTK}^{\infty}_{X_iX_j} P_i(T) \\ &= -i\sum_{i,l=1}^{N} P_i(T)\,\mathrm{NTK}^{\infty}_{X_iX_l} Y_l Z^{\infty}_{\nu_x}(T,P(T)),\end{aligned}$$

并使用 $Z^{\nu_x}(T, P(T)) = U(T)$ 以及 $P(T) = e^{-\mathrm{NTK}^{\infty} T} P(0)$，上述常微分方程可积分为

$$U(T) = U(0)\exp\left(i\int_0^T \sum_{i,j=1}^{N} (e^{\mathrm{NTK}^{\infty} s} P(0))_i \mathrm{NTK}^{\infty}_{X_iX_j} Y_j\, ds \right).$$

不过，请注意，

$$\sum_{i,j=1}^{N} (e^{\mathrm{NTK}^{\infty} s} P(0))_i \mathrm{NTK}^{\infty}_{X_iX_j} Y_j = \frac{d}{ds}\langle e^{\mathrm{NTK}^{\infty} s} P(0), Y\rangle_{\mathbb{R}^N},$$

代入以上结果即可得到

$$U(T) = U(0)\exp\left[i\left(\langle e^{\mathrm{NTK}^{\infty} T} P(0), Y\rangle_{\mathbb{R}^N} - \langle P(0), Y\rangle_{\mathbb{R}^N} \right)\right],$$

换句话说

$$Z^{\infty}_{\nu_x}(T, e^{\mathrm{NTK}^{\infty} T} P(0)) = Z^{\infty}_{\nu_x}(0, P(0))\exp\left[i\left(\langle e^{\mathrm{NTK}^{\infty} T} P(0), Y\rangle_{\mathbb{R}^N} - \langle P(0), Y\rangle_{\mathbb{R}^N} \right)\right].$$

最后可以将其重新整理，得到 $Z^{\infty}_{\nu_x}(t,p)$ 的最终表达式：

$$Z^{\infty}_{\nu_x}(t,p) = Z^{\infty}_{\nu_x}(0, e^{-\mathrm{NTK}^{\infty} t} p)\exp\left[i\left(\langle p, Y\rangle_{\mathbb{R}^N} - \langle e^{-\mathrm{NTK}^{\infty} t} p, Y\rangle_{\mathbb{R}^N} \right)\right].$$

然后，我们可以通过傅里叶逆变换回过头来计算密度测度 ν_x

$$\begin{aligned}\nu^{\infty}_x(t,y) &= \frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} Z^{\infty}_{\nu_x}(t,p) e^{-i\langle y,p\rangle} dp \\ &= \frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} Z^{\infty}_{\nu_x}(0, e^{-\mathrm{NTK}^{\infty} t} p)\exp\left[-i\left(\langle e^{-\mathrm{NTK}^{\infty} t} p, Y\rangle - \langle p, Y\rangle + \langle p, y\rangle \right)\right] dp \\ &= \frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} Z^{\infty}_{\nu_x}(0,q)\exp\left[-i\langle q, Y - e^{\mathrm{NTK}^{\infty} t}(Y-y)\rangle\right]\det(e^{\mathrm{NTK}^{\infty} t})\, dq \\ &= \det(e^{\mathrm{NTK}^{\infty} t})\ \nu_x(0, Y - e^{\mathrm{NTK}^{\infty} t}(Y-y)),\end{aligned}$$

正如预期，这与 *4.5* 一致。

在这种情况下，当 $k = N$ 且 $x_i = X_i$ 对于 $i = 1,\ldots,N$ 时，为了确定演化密度 $\nu^{\infty}_x(t,y)$，我们必须将方程 2.14 中给出的初始密度公式代入 4.5。这是通过将 2.14 与 Φ^{∞}_{-t} 进行复合得出的，这需要计算

$$(C^{-1}_{ij} y_i y_j)\circ\Phi^{\infty}_{-t} = A(t)_{ij}(y_i + (e^{-\mathrm{NTK}^{\infty} t} - 1)_{im} Y_m)(y_j + (e^{-\mathrm{NTK}^{\infty} t} - 1)_{jp} Y_p),$$

其中

$$A(t)_{ij} = \left(e^{\mathrm{NTK}^{\infty} t}\right)_{il} C^{-1}_{lp} \left(e^{\mathrm{NTK}^{\infty} t}\right)_{pj},$$

and

$$\frac{\det(e^{\mathrm{NTK}^{\infty} t})}{\sqrt{\det(C)}} = \frac{1}{\sqrt{\det(A(t)^{-1})}}.$$

最后，将这些代入 4.5，得到

$$\nu^{\infty}_{x}(t,y) = \frac{e^{-A(t)_{ij}(y_i + (e^{-\mathrm{NTK}^{\infty} t} - 1)_{im} Y_m)(y_j + (e^{-\mathrm{NTK}^{\infty} t} - 1)_{jp} Y_p)}}{\sqrt{(2\pi)^k \det(A(t)^{-1})}}. \tag{4.6}$$

注 9. *F* 根据表达式 *4.6*，以及神经正切核$^{\infty}$ 是正定的 [7], 这一事实，我们得到

$$\lim_{t \to +\infty} \nu^{\infty}_{x}(t,y)dy = \prod_{i=1}^{N} \delta(y_i - Y_i)dy,$$

弱地。

现在我们考虑下一种情况，其中 $k = N + 1$ 以及 $x_i = X_i$对于 $i = 1,\dots N$，并且 $x_{N+1} \in \mathrm{R}$ 是一个额外的任意点。y_{N+1} 的演化，即 $(\Phi^{\infty}_t(y))_{N+1} = f_{\infty}(x,t)$，已经在 3.10 节中计算过。我们当时得到的结果是

$$(\Phi^{\infty}_t(y))_{N+1} = y_{N+1} + \sum_{l,m,p=1}^{N} \mathrm{NTK}^{\infty}_{x_{N+1} X_l} (\mathrm{NTK}^{\infty})^{-1}_{lm} (e^{-\mathrm{NTK}^{\infty} t} - 1)_{mp} (y_p - Y_p).$$

特别地，请注意 $\partial_{y^{N+1}} (\Phi^{\infty}_t(y))_{N+1} = 1$而 $\partial_{y^{N+1}} (\Phi^{\infty}_t(y))_i = 0$对于 $i \in 1,\dots,N$，从而表明 $\det(d\Phi^{\infty}_t) = \det(e^{\mathrm{NTK}^{\infty} t})$，与之前一样。我们发现，对于 $y = (y_1,\dots,y_N)$ 和 $Y = (Y_1,\dots,Y_N)$

$$\nu^{\infty}_{x}(t,y,y_{N+1}) = \nu^{\infty}_{x}\left(0, Y + e^{\mathrm{NTK}^{\infty} t}(y - Y), y_{N+1} + \mathrm{NTK}^{\infty}_{x_{N+1} X_l} (\mathrm{NTK}^{\infty})^{-1}_{lm} (e^{\mathrm{NTK}^{\infty} t} - 1)_{mp} (y_p - Y_p)\right) \times$$
$$\times \det(e^{\mathrm{NTK}^{\infty} t}),$$

对重复指标 l,m,p 从 1 到 N求和。

最后，我们考虑具有 $k > N$ 的一般情形。鉴于在前一种情形中，x_{N+1} 是任意的，导出 3.10 的计算对任意$x \in \mathrm{R}$ 都成立。从现在起，我们令$x_1 = X_1,\dots,x_N = X_N$，并令$x_{N+1},\dots x_{N+k} \in \mathrm{R}$ 任意。为简化记号，我们令$X = (X_1,\dots,X_N)$ 和$x = (x_{N+1},\dots,x_{N+k})$。于是，对于$i = 1,\dots,N$，我们有

$$(\Phi^{\infty}_t(y))_i = Y_i + \sum_{m=1}^{N} (e^{-\mathrm{NTK}^{\infty} t})_{im} (y_m - Y_m),$$

而对于$i = N + 1,\dots N + k$

$$(\Phi^{\infty}_t(y))_i = y_i + \sum_{l,m,p=1}^{N} \mathrm{NTK}^{\infty}_{x_i X_l} (\mathrm{NTK}^{\infty})^{-1}_{lm} (e^{-\mathrm{NTK}^{\infty} t} - 1)_{mp} (y_p - Y_p).$$

事实上，很容易验证，第一个方程实际上是第二个方程的一个特例，因此我们不妨只研究第二个方程。为此，我们将$y \in \mathrm{R}^k$ 视为一个列向量，将 $\pi_N(y) = (y_1,\dots,y_N) \in \mathrm{R}^N$视为它在前$N$个坐标上的投影，将神经正切核$^{\infty}_x$视为一个$k \times N$矩阵，其元素为神经正切核$^{\infty}_{x_i X_j}$，其中$i \in \{1,\dots,k\}$且$j \in \{1,\dots, N\}$；并将神经正切核$^{\infty}$ 视为一个$N \times N$矩阵，其元素为

神经正切核$\overset{\infty}{X_iX_j}$，其中$i,j \in \{1,\dots,N\}$，正如我们此前一直采用的记法。由此得到一般的演化方程

$$\Phi_t^{\infty}(y) = y + \mathrm{NTK}_x^{\infty}(\mathrm{NTK}^{\infty})^{-1}(e^{-\mathrm{NTK}^{\infty}t} - 1)(\pi_N(y) - Y). \tag{4.7}$$

使用与之前类似的论证，我们有 $\det(d\Phi_t^{\infty}) = \det(e^{\mathrm{NTK}^{\infty}t})$，并且可以用相同的方法计算 $\nu_x^{\infty}(t,y) = \nu_x^{\infty}(0,\Phi_{-t}^{\infty}(y))|\det(d\Phi_{-t}^{\infty})|$。这引出了以下结果。

定理 3。设 $k \geq N$，且 $x = (x_1,\dots,x_k)$，其中 $x_1 = X_1$，…、$x_N = X_N$ 为训练点。那么，测度 $\mathbb{P}_x^{\infty}(t)$ 在 $\mathbb{R}^k$ 上诱导的密度 $\nu_x^{\infty}(t,y)$ 满足

$$\nu_x^{\infty}(t,y) = \frac{e^{-\langle \Phi_{-t}^{\infty}(y), C^{-1}\Phi_{-t}^{\infty}(y)\rangle}}{\sqrt{(2\pi)^k \det(A(t)^{-1})}},$$

其中

$$(\Phi_{-t}^{\infty}(y))_i = y_i + \sum_{l,m,p=1}^{N} \mathrm{NTK}_{x_i X_l}^{\infty}(\mathrm{NTK}^{\infty})_{lm}^{-1}(e^{\mathrm{NTK}^{\infty}t} - 1)_{mp}(y_p - Y_p),$$

对于 $i = 1,\dots,k$，以及

$$A(t)_{ij} = \left(e^{\mathrm{NTK}^{\infty}t}\right)_{il} C_{lp}^{-1} \left(e^{\mathrm{NTK}^{\infty}t}\right)_{pj},$$

对于 $i,j = 1,\dots,k$。

注 10. 如果我们只关心$(y_{N+1},\dots,y_{N+k})$的分布，则可以通过对$(y_1,\dots,y_N,y_{N+1},\dots,y_{N+k})$中的前 N 个变量积分，从其分布中计算得到，即

$$\nu_x(t,y_{N+1},\dots,y_{N+k}) = \int_{\mathbb{R}^N} \nu_{X,x}(t,y_1,\dots y_N,y_{N+1},\dots,y_{N+k})dy_1\dots dy_N$$

4.3. 大宽度但有限宽度下的演化。本节将估计在有限宽度$n \in \mathbb{N}$ 网络下，诱导测度 $\mathrm{P}_n = (\mathrm{F}_{(n)})_*(Q^{(n)})$ 在 $C([a,b],\mathbb{R})$ 中的演化。我们将关注 $n \gg 1$ 的情形，并将 P_n 的演化与 P^{∞} 的演化进行比较；P^{∞} 的有限维分布具有定理 3 中计算得到的密度。

令$x \in [a,b]^k$ 固定，并考虑 4.1 和 4.2 中定义的有限维分布。为完整起见，我们在此回顾如下：

$$\mathbb{P}_x^{(n)}(t) = (\mathcal{F}_x^{(n)} \circ \phi_t^{(n)})_* Q^{(n)}$$
$$\mathbb{P}_x^{\infty}(t) = (\Phi_t^{\infty})_* \mathbb{P}_x^{\infty},$$

其中：(i) 函数 $\mathrm{F}_{(n)x}: \mathbb{R}^{4n} \to \mathbb{R}^k$ 表示 $\mathrm{F}_{(n)}$ 与在固定的 x 处进行求值的复合，如 2.6 节所示，即神经网络在x处的求值结果；(ii) $\phi^{(n)}{}_t: \mathbb{R}^{4n} \to \mathbb{R}^{4n}$ 描述了有限神经网络在通过梯度下降训练时参数的演化；(iii)Φ_t^{∞}描述了使用无限网络极限进行训练时输出的演化，如公式 4.7 和 3.10 所示。

为了在 $\mathrm{P}_x^{(n)}(t)$ 和 $\mathrm{P}_x^{(n)}(t)$ 之间进行“插值”，引入辅助测度将很有用

$$\mathbb{P}^{(n),\infty}(t) = \left(\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}\right)_* Q^{(n)}$$

在 $\mathbb{R}^k$ 上。这是先计算有限宽度网络，再根据无限宽度网络诱导的流演化其输出而得到的。

为了更直观地展示 $\mathrm{P}^{(n)}(t)$ 和 $\mathrm{P}^{(n),\ \infty}(t)$，我们注意到，它们是沿着下方非交换图 4.8 中两条不同的对角线进行推前得到的。

(4.8)
$$\begin{array}{ccc} \mathbb{R}^{4n} & \xrightarrow{\phi_t^{(n)}} & \mathbb{R}^{4n} \\ {\scriptstyle \mathcal{F}_x^{(n)}}\downarrow & & \downarrow{\scriptstyle \mathcal{F}_x^{(n)}} \\ \mathbb{R}^{k} & \xrightarrow[\Phi_t^{\infty}]{} & \mathbb{R}^{k} \end{array}$$

回顾一下，我们的目的是衡量 $\mathrm{P}_x^{(n)}(\mathrm{t})$ 与 $\mathrm{P}_x^{\infty}(\mathrm{t})$ 之间的距离，尤其是考察随着 t 的演化，该距离如何随 n 变化。为此，给定两个测度 μ_1,μ_2，我们将考虑它们之间的普罗霍罗夫度量，$\pi(\mu_1,\mu_2)$，其定义为所有能同时满足 $\mu_1(\mathrm{A}) \le \mu_2(\mathrm{A}^{\rho}) + \rho$ 和 $\mu_2(\mathrm{A}) \le \mu_1(\mathrm{A}^{\rho}) + \rho$的正 $\rho > 0$ 的下确界，其中 $\mathrm{A}^{\rho} = \bigcup_{\mathrm{a}\in\mathrm{A}} \mathrm{B}_{\rho}(\mathrm{a})$。用公式表示为，

$$\pi(\mu_1,\mu_2) = \inf\left\{\rho > 0 : \forall \text{ measurable } A,\ \mu_1(A) \le \mu_2(A^{\rho}) + \rho \text{ and } \mu_2(A) \le \mu_1(A^{\rho}) + \rho\right\}.$$

本节的主要结果如下。

定理 4。　设 $x \in \mathbb{R}^k$。那么，存在一个常数 $C > 0$，与 n 无关，使得对所有 $t \ge 0$

$$\pi\left(\mathbb{P}_x^{(n)}(t), \mathbb{P}_x^{\infty}(t)\right) \le C\frac{(\log(n))^{+}}{\sqrt{n}}.$$

为了证明这个定理，我们先证明三个中间结果。在这些结果中，我们将使用如下记号

$$\varepsilon_n := C\frac{(\log(n))^{+}}{\sqrt{n}}.$$

引理1. 存在 $m \in \mathbb{N}$ 和 $C' > 0$，具有如下含义。对于 $n > m$，存在集合 $\mathcal{A}_{\delta_n}^n \subset \mathbb{R}^{4n}$，使得 $Q^{(n)}(\mathcal{A}_{\delta_n}^n) \ge 1 - C'(\log(n))^{1/2+}/\sqrt{n}$，从而对于任意 $A \subseteq \mathbb{R}^{4n}$，

(4.9)
$$\left(\left(\Phi_t^{\infty}\circ\mathcal{F}_x^{(n)}\right)^{-1}(A)\right)\cap\mathcal{A}_{\delta_n}^n \subseteq \left(\left(\mathcal{F}_x^{(n)}\circ\phi_t^{(n)}\right)^{-1}(A^{\varepsilon_n})\right)\cap\mathcal{A}_{\delta_n}^n$$

(4.10)
$$\left(\left(\mathcal{F}_x^{(n)}\circ\phi_t^{(n)}\right)^{-1}(A)\right)\cap\mathcal{A}_{\delta_n}^n \subseteq \left(\left(\Phi_t^{\infty}\circ\mathcal{F}_x^{(n)}\right)^{-1}(A^{\varepsilon_n})\right)\cap\mathcal{A}_{\delta_n}^n$$

证明。 我们将使用定理 2 的结果 (3.13)，该结果指出，对于任意 $\delta \in (0,1)$，存在 $C_\delta > 0$和集合 $\mathrm{A}_{\delta}^n \subset \mathrm{R}^{4n}$，使得 $Q^{(n)}(\mathrm{A}_{\delta}^n) \ge 1 - \delta$，并且对所有 $\theta_0 \in \mathrm{A}_{\delta}^n$

$$\left|\mathcal{F}_x^{(n)}\circ\phi_t^{(n)}(\theta_0) - \Phi_t^{\infty}\circ\mathcal{F}_x^{(n)}(\theta_0)\right| \le C_{\delta}\frac{(\log(n))^{1/2+}}{\sqrt{n}}.$$

另一方面，我们可以令 $\delta_n \sim (\log(n))^{1/2+}/\sqrt{n} \to 0$，如例3中所示。然后，正如其中所示

$$C_{\delta_n}\frac{(\log(n))^{1/2+}}{\sqrt{n}} = \varepsilon_n$$

并且我们发现存在 $\mathrm{A}_{\delta_n}^n$，满足 $Q^{(n)}(\mathrm{A}_{\delta_n}^n) \geq 1-\delta_n$，使得对于所有 $\theta_0 \in \mathrm{A}_{\delta_n}^n$

$$\left|\mathcal{F}_x^{(n)} \circ \phi_t^{(n)}(\theta_0) - \Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}(\theta_0)\right| \leq \varepsilon_n. \tag{4.11}$$

接下来，我们证明对于任意 $y \in \mathrm{R}^k$，都有

$$\left(\left(\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}\right)^{-1}(y)\right) \cap \mathcal{A}_{\delta_n}^n \subseteq \left(\left(\mathcal{F}_x^{(n)} \circ \phi_t^{(n)}\right)^{-1}(B_{\varepsilon_n}(y))\right) \cap \mathcal{A}_{\delta_n}^n$$

$$\left(\left(\mathcal{F}_x^{(n)} \circ \phi_t^{(n)}\right)^{-1}(y)\right) \cap \mathcal{A}_{\delta_n}^n \subseteq \left(\left(\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}\right)^{-1}(B_{\varepsilon_n}(y))\right) \cap \mathcal{A}_{\delta_n}^n$$

为证明第一个包含关系，取一个 $\theta_0 \in \mathrm{A}_{\delta_n}^n$，使得 $\mathrm{F}(_n)_x \circ \phi^{(n)}{}_t(\theta_0) = y$。由 (4.11)，可得

$$\left|y - (\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)})(\theta_0)\right| \leq \varepsilon_n.$$

也就是说，$\Phi_t^{\infty} \circ \mathrm{F}(_n)_x(\theta_0) \in B_{\varepsilon_n}(y)$，从而得到该包含关系。第二个包含关系的证明方法相同。最后，对所有$y \in A$取这些包含关系的并集，即得所述结果。□

接下来，我们利用这一结果计算 $\mathrm{P}_x^{(n)}(t)$ 与辅助测度 $\mathrm{P}^{(n)}{}_{,x}{}^{\infty}(t)$ 之间 Prokhorov 距离的一个与 t无关的上界

引理 2。存在 $m > 0$，使得当 $n > m$ 时

$$\pi\left(\mathbb{P}_x^{(n),\infty}(t), \mathbb{P}_x^{(n)}(t)\right) \leq \varepsilon_n.$$

证明。由（4.9）可得

$$Q^{(n)}\left(\left(\mathcal{F}_x^{(n)} \circ \phi_t^{(n)}\right)^{-1}(A) \cap \mathcal{A}_{\delta_n}^n\right) \leq Q^{(n)}\left(\left(\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}\right)^{-1}(A^{\varepsilon_n}) \cap \mathcal{A}_{\delta_n}^n\right),$$

因为

$$\begin{aligned}
\left(\mathbb{P}_x^{(n)}(t)\right)(A) &= Q^{(n)}\left(\left(\mathcal{F}_x^{(n)} \circ \phi_t^{(n)}\right)^{-1}(A) \cap \mathcal{A}_{\delta_n}^n\right) + Q^{(n)}\left(\left(\mathcal{F}_x^{(n)} \circ \phi_t^{(n)}\right)^{-1}(A) \cap \left(\mathcal{A}_{\delta_n}^n\right)^c\right) \\
&\leq Q^{(n)}\left(\left(\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}\right)^{-1}(A^{\varepsilon_n}) \cap \mathcal{A}_{\delta_n}^n\right) + Q^{(n)}\left(\left(\mathcal{A}_{\delta_n}^n\right)^c\right) \\
&\leq Q^{(n)}\left(\left(\Phi_t^{\infty} \circ \mathcal{F}_x^{(n)}\right)^{-1}(A^{\varepsilon_n})\right) + \delta_n \\
&= \left(\mathbb{P}_x^{(n),\infty}(t)\right) + \delta_n \\
&\leq \left(\mathbb{P}_x^{(n),\infty}(t)\right)(A^{\varepsilon_n}) + \varepsilon_n,
\end{aligned}$$

由于 $\delta > 0$ 是任意的，我们实际上得到

$$\left(\mathbb{P}_x^{(n)}(t)\right)(A) \leq \left(\mathbb{P}_x^{(n),\infty}(t)\right)(A^{\varepsilon_n}) + \varepsilon_n.$$

此外，通过采用类似的分析，只是现在使用 4.10 而非 4.9，我们可得 $\left(\mathrm{P}^{(n)}{}_{,x}{}^{\infty}(t)\right)(A) \leq \left(\mathrm{P}_x^{(n)}(t)\right)(A^{\varepsilon_n}) + \varepsilon_n$，从而证明引理中所述的结论。□

此外，我们还可以比较 $\mathrm{P}_x^{\infty}(t)$ 和 $\mathrm{P}^{(n)}{}_{,\ x}{}^{\infty}(t)$。

引理 3. 存在 $m > 0$ 和 $C'' > 0$，使得对于 $n > m$

$$\pi\Big(\mathbb{P}_x^{(n),\infty}(t), \mathbb{P}_x^{\infty}(t)\Big) \le \frac{C''}{\sqrt{n}}$$

证明。利用定理 1 的表述：$\mathrm{P}_x^{(n)}(0) = \mathrm{P}_x^{(n)}$ 的密度 ν_x^n 接近 $\mathrm{P}_x^{\infty}(0) = \mathrm{P}_x^{\infty}$ 的密度 ν_x^{∞}，特别地，它说明

$$\left|\nu^n(y) - \nu^{\infty}(y)\right| \le \frac{\bar{C}}{\sqrt{n}}\lambda(y),$$

对于某个定义在 R^k 上且积分有限的正函数 λ。因此，当 $n \gg 1$ 足够大时，我们有

$$\begin{aligned}\Big(\mathbb{P}_x^{(n),\infty}(t)\Big)(A) &= \mathbb{P}_x^{(n)}\Big(\Phi_{-t}^{\infty}(A)\Big) \\ &\le \mathbb{P}_x^{\infty}\Big(\Phi_{-t}^{\infty}(A)\Big) + \frac{\bar{C}}{\sqrt{n}}\Lambda\Big(\Phi_{-t}^{\infty}(A)\Big),\end{aligned}$$

其中，Λ 表示 R^k 上密度为 ν 的有限测度。由此可得

$$\Big(\mathbb{P}_x^{(n),\infty}(t)\Big)(A) \le \Big(\mathbb{P}_x^{\infty}(t)\Big)(A^{\frac{C''}{\sqrt{n}}}) + \frac{C''}{\sqrt{n}}.$$

同样的推理可得不等式

$$\begin{aligned}\Big(\mathbb{P}_x^{\infty}(t)\Big)(A) &= \mathbb{P}_x^{\infty}\Big(\Phi_{-t}^{\infty}(A)\Big) \\ &= \mathbb{P}_x^{(n)}\Big(\Phi_{-t}^{\infty}(A)\Big) + \frac{\bar{C}\Lambda\Big(\Phi_{-t}^{\infty}(A)\Big)}{\sqrt{n}} \\ &\le \Big(\mathbb{P}_x^{(n),\infty}(t)\Big)(A) + \frac{C''}{\sqrt{n}} \\ &\le \Big(\mathbb{P}_x^{(n),\infty}(t)\Big)(A^{\frac{C''}{\sqrt{n}}}) + \frac{C''}{\sqrt{n}},\end{aligned}$$

从而完成了所述结果的证明。 □

定理 4 的证明。由引理 2 和 3 可得

$$\pi\Big(\mathbb{P}_x^{(n)}(t), \mathbb{P}_x^{\infty}(t)\Big) \le \pi\Big(\mathbb{P}_x^{(n),\infty}(t), \mathbb{P}_x^{(n)}(t)\Big) + \pi\Big(\mathbb{P}_x^{(n),\infty}(t), \mathbb{P}_x^{\infty}(t)\Big) \le 2\varepsilon_n.$$

□

附录 A。埃奇沃思展开

在本附录中，我们回顾概率论中一种广为人知的方法，即埃奇沃斯级数展开。关于全面介绍，请参见 [15]及其中的参考文献。

A.1. 特征函数。令 $X = (X_1, \ldots, X_k)$为一个取值于$\mathbb{R}^k$ 的随机变量，且其所有矩均有限。我们将其密度函数记为 $f_X(y) = f_X(y_1, \ldots, y_k)$，并将

$$Z_X(p_1, \ldots, p_k) = \mathbb{E}(e^{i \sum_{i=1}^k p_i X_i}),$$

X 的特征函数。特别地，将其明确写为 $\int_{\mathbb{R}^k} f_X(y_1, \ldots, y_k) e^{i \sum_{i=1}^k p_i y_i} dy_1 \ldots dy_k$，即可立即看出它就是$f_X$ 的傅里叶逆变换。因此，我们可以通过傅里叶变换从Z_X 中恢复f_X

$$f_X(y_1, \ldots, y_k) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} Z_X(p_1, \ldots, p_k) e^{-i \sum_{i=1}^k p_i y_i} dp_1 \ldots dp_k.$$

特征函数的一个重要性质是以下弱收敛结果。若 $\{Y_n\}_{n=1}^{+\infty}$ 是一族随机变量，其特征函数Z_{Y_n} 逐点收敛到一个在原点连续且满足$Z_Y(0) = 1$ 的函数Z_Y，则存在一个随机变量Y，其特征函数为Z_Y，并且

$$f_{Y_n}(y)\, dy_1 \ldots dy_k \rightharpoonup f_Y(y)\, dy_1 \ldots dy_k,$$

也就是说，Y_n 弱收敛到Y。

现在可以直接从定义轻松证明特征函数的若干其他有用性质。

- 对于 $r \in \mathbb{N}_k$，我们有

$$\begin{aligned} Z_X^{(r_1, \ldots, r_k)}(0) &:= \frac{d^{r_1} \ldots d^{r_k}}{dp_1^{r_1} \ldots dp_k^{r_k}} \Big|_{p=0} Z_X = \mathbb{E}((iX_1)^{r_1} \ldots (iX_k)^{r_k} e^{i \sum_{i=1}^k p_i X_i}|_{p=0}) \\ &= i^{r_1 + \ldots + r_k} \mathbb{E}(X_1^{r_1} \ldots X_k^{r_k}) = i^{r_1 + \ldots + r_k} \mu_{r_1, \ldots r_k}, \end{aligned}$$

  其中，若该矩有限，则 $\mu(r_1, \ldots, r_k)$ 是 $r_1, \ldots, r_k$-阶矩。作为这一结果和泰勒公式的一个特殊应用，我们发现，若 X 的前 l 阶矩均有限，则

$$Z_X(p) = 1 + \sum_{r_1 + \ldots + r_k = 1}^{l} \mu_{r_1, \ldots r_k} \frac{(ip_1)^{r_1} \ldots (ip_k)^{r_k}}{r_1! \ldots r_k!} + R_l(p), \tag{A.1}$$

  其中余项 R_l 满足 $\lim_{q \to 0} R_l(q) / |q|^l = 0$。

- 如果X，$\tilde{X}$ 是如前所述的独立随机变量，且$a, b \in \mathbb{R}$，那么

$$Z_{aX + b\tilde{X}}(p) = Z_X(ap) Z_{\tilde{X}}(bp).$$

A.2. 累积量生成函数。由方程 A.1 可知，特征函数Z_X 可解释为矩生成函数。同理，函数$F_X := \log Z_X$可展开为

$$F_X(p) = \sum_{r_1 + \ldots + r_k = 1}^{l} \lambda_{r_1, \ldots r_k} \frac{(ip_1)^{r_1} \ldots (ip_k)^{r_k}}{r_1! \ldots r_k!} + \tilde{R}_l(p), \tag{A.2}$$

其中余项 $\tilde{R}_l$ 满足 $\lim_{q \to 0} \tilde{R}_l(q) / |q|^l = 0$。这里，量 $\lambda_{r_1, \ldots r_k}$ 称为累积量。它们与矩的关系由方程 A.2 给出，其中 A.1 的对数为

如下所示

$$\sum_{r_1+\ldots+r_k=1}^{\infty}\lambda_{r_1,\ldots r_k}\frac{(ip_1)^{r_1}\ldots(ip_k)^{r_k}}{r_1!\ldots r_k!}=\sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{j}\left(\sum_{s_1+\ldots+s_k=1}^{\infty}\mu_{s_1,\ldots s_k}\frac{(ip_1)^{s_1}\ldots(ip_k)^{s_k}}{s_1!\ldots s_k!}\right)^j, \tag{A.3}$$

从而可以用矩来表示累积量。

注释 11. 展开 *A.3* 的前几项，我们发现

$$\frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}=\frac{\mu_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}-\frac{1}{2}\sum_{s_i+s_i'=r_i}\frac{\mu_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}\frac{\mu_{r_1',\ldots,r_k'}}{r_1'!\ldots r_k'!}+\frac{1}{3}\sum_{s_i+s_i'+s_i''=r_i}\frac{\mu_{s_1,\ldots,s_k}}{s_1!\ldots s_k!}\frac{\mu_{s_1',\ldots,s_k'}}{s_1'!\ldots s_k'!}\frac{\mu_{s_1'',\ldots,s_k''}}{s_1''!\ldots s_k''!}$$
$$+\ldots,$$

其中 .. . 表示包含矩的四次方或更高次幂的项 .

或者，也可以改为令 A.2 的指数等于 A.1。由此得到

$$\begin{aligned}\sum_{r_1+\ldots+r_k=0}^{\infty}\mu_{r_1,\ldots r_k}\frac{(ip_1)^{r_1}\ldots(ip_k)^{r_k}}{r_1!\ldots r_k!}&=\exp\left(\sum_{s_1+\ldots+s_k=1}^{\infty}\lambda_{s_1,\ldots,s_k}\frac{(ip_1)^{s_1}\ldots(ip_k)^{s_k}}{s_1!\ldots s_k!}\right)\\&=\prod_{s_1+\ldots+s_k=1}^{\infty}\exp\left(\lambda_{s_1,\ldots,s_k}\frac{(ip_1)^{s_1}\ldots(ip_k)^{s_k}}{s_1!\ldots s_k!}\right)\\&=\prod_{s_1+\ldots+s_k=1}^{\infty}\sum_{m_s=0}^{\infty}\frac{1}{m_s!}\left(\lambda_{s_1,\ldots,s_k}\frac{(ip_1)^{s_1}\ldots(ip_k)^{s_k}}{s_1!\ldots s_k!}\right)^{m_s}\\&=\sum_{r_1+\ldots+r_k=0}^{\infty}\sum_{m_s}\prod_{s_1,\ldots,s_k}\frac{\lambda_{s_1,\ldots,s_k}^{m_s}}{m_s!}\frac{(ip_1)^{r_1}\ldots(ip_k)^{r_k}}{(s_1!)^{m_s}\ldots(s_k!)^{m_s}},\end{aligned}$$

其中，$\sum_{m_{s_1}}\prod_{s1,\ldots,sk}$ 中的乘积与求和的组合受到限制，使得 $\sum_s m_s s_i=r_i$ 对所有 $i=1,\ldots,k$ 均成立。由此我们发现

$$\mu_{r_1,\ldots r_k}=\sum_{m_s}\prod_{s_1,\ldots,s_k}\frac{1}{m_s!}\frac{(r_1!)\ldots(r_k)!}{(s_1!)^{m_s}\ldots(s_k!)^{m_s}}\lambda_{s_1,\ldots,s_k}^{m_s}. \tag{A.4}$$

注释 12. 这个公式有一个直观的图示解释。设有 r_1 个颜色为 1 的球、r_2 个颜色为 2 的球，依此类推，直到有 r_k 个颜色为 k 的球。那么，上述组合因子表示将这些球分到 m_s 个盒子中的方法数，每个盒子里对每种颜色 $i=1,\ldots,k$ 都有 s_i 个球。于是，只需以所有可能的方式将球分到盒子中，即可由累积量计算矩；每一种分法所对应的贡献，都是各盒子相关累积量的乘积。

注释 13 ($r_1+\ldots+r_k\le 2$). 当 $r_1+\ldots+r_k=1$ 中只有一个 r_i 可以非零时，该值为 *1*，因此 $m_s=1$ 和 $s_i=1$ 也始终为 *1*。结合公式 *A.4* 可得

$$\mu_{r_1,\ldots,r_k}=\lambda_{r_1,\ldots,r_k},$$

只要 $r_1+\ldots+r_k=1$.

此外，当 $r_1+\ldots+r_k=2$ 时，$\mu_{r_1,\ldots,r_k}$ 的值编码了协方差矩阵。事实上，如果 r_i 和 r_j 是仅有的非零项，且 $r_i+r_j=2$，则有

$$\mu_{r_1,\ldots r_k}=\mathbb{E}[X_iX_j]=:C_{ij},$$

因此

$$\sum_{r_1+\ldots+r_k=2} \mu_{r_1,\ldots,r_k}(ip_1)^{r_1}\ldots(ip_k)^{r_k} = -\frac{1}{2}\sum_{i,j=1}^{k} C_{ij}\, p_i p_j.$$

一个特殊情况是对于 $r_1 + \ldots + r_k = 1$，所有 $\mu_{r_1,\ldots,r_k} = 0 = \lambda_{r_1,\ldots,r_k}$ 都成立。在这种情况下，只要 $r_1 + \ldots + r_k = 2$，我们也有 $\mu_{r_1,\ldots,r_k} = \lambda_{r_1,\ldots,r_k}$，因此

$$\sum_{r_1+\ldots+r_k=2} \lambda_{r_1,\ldots,r_k}(ip_1)^{r_1}\ldots(ip_k)^{r_k} = -\frac{1}{2}\sum_{i,j=1}^{k} C_{ij}\, p_i p_j,$$

也一样。

现在来看一个具体例子。

例 2 ($k = 1$). 在这种特殊情况下，我们用于根据累积量表示矩的公式 *A.4* 表明，

$$\mu_r = \sum_{m_s} \prod_s \frac{1}{m_s!}\frac{r!}{(s!)^{m_s}}\lambda_s^{m_s}, \tag{A.5}$$

在满足约束 $\sum_s s m_s = r$ 的条件下。特别是在某些情况下，这给出

$$\begin{aligned}
\mu_1 &= \lambda_1 \\
\mu_2 &= \lambda_2 + \lambda_1^2 \\
\mu_3 &= \lambda_3 + 3\lambda_1\lambda_2 + \lambda_1^3 \\
\mu_4 &= \lambda_4 + 4\lambda_3\lambda_1 + 3\lambda_2^2 + 6\lambda_2\lambda_1^2 + \lambda_1^4,
\end{aligned}$$

依此类推。反过来，利用方程 *A.3* 进行较为繁琐但直接的计算即可得到

$$\begin{aligned}
\lambda_1 &= \mu_1 \\
\lambda_2 &= \mu_2 - \mu_1^2 \\
\lambda_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 \\
\lambda_4 &= \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4,
\end{aligned}$$

等等。我们将遇到一种特别相关的情况，即 $\mu_1 = 0 = \lambda_1$，此时这些最后的关系可简化为

$$\begin{aligned}
\lambda_2 &= \mu_2 \\
\lambda_3 &= \mu_3 \\
\lambda_4 &= \mu_4 - 3\mu_2^2.
\end{aligned}$$

附录 B. 中心极限定理的渐近修正（埃奇沃思展开）

B.1. 中心极限定理。现在，假设 $\{X^{(n)}\}_{n\in\mathbb{N}}$ 是一族独立同分布的随机变量，其前 l-阶矩有限（对于 $l > 2$），且 $\mu_{r_1,\ldots,r_k} = 0$ 对任意 $r_1 + \ldots + r_k = 1$ 均成立，即所有前

阶矩均为零。定义：

$$Y_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X^{(i)}.$$

那么，

(B.1)
$$\begin{aligned} Z_{Y_n}(p) &= Z_X\left(p/\sqrt{n}\right)^n \\ &= \left(1 + \sum_{r_1+\ldots+r_k=1}^{l} \frac{\mu_{r_1,\ldots r_k}}{n^{\frac{r_1+\ldots+r_k}{2}}} \frac{(ip_1)^{r_1}\ldots(ip_k)^{r_k}}{r_1!\ldots r_k!} + R_l(p/\sqrt{n})\right)^n \\ &= \left(1 + \frac{1}{n}\left(\sum_{r_1+\ldots+r_k=2}^{l} \frac{\mu_{r_1,\ldots r_k}}{n^{\frac{r_1+\ldots+r_k}{2}-1}} \frac{(ip_1)^{r_1}\ldots(ip_k)^{r_k}}{r_1!\ldots r_k!} + nR_l(p/\sqrt{n})\right)\right)^n. \end{aligned}$$

由此公式可清楚看出，对于任意有限的$p \in \mathbb{R}$

$$\lim_{n\to+\infty} Z_{Y_n}(p) = \exp\left(-\sum_{r_1+\ldots+r_k=2} \mu_{r_1,\ldots r_k} \frac{p_1^{r_1}\ldots p_k^{r_k}}{r_1!\ldots r_k!}\right),$$

从而证明了中心极限定理。

B.2. $n \gg 1$ 中的渐近行为（定理 1 的证明）。然而，如果我们不只是试图计算 $\lim_{n\to+\infty} Z^{Y_n}(p)$，而是想评估$Z^{Y_n}(p)$收敛到其极限的速率，那么使用累积量生成函数会更方便。其主要原因从公式 B.1 中便可看出：取其对数会使指数中的n 降下来。事实上，我们发现

$$\begin{aligned} F_{Y_n}(p) &:= \log Z_{Y_n}(p) \\ &= nF_X(p/\sqrt{n}) \\ &= n \sum_{r_1+\ldots+r_k=1}^{l} \frac{\lambda_{r_1,\ldots r_k}}{n^{\frac{r_1+\ldots+r_k}{2}}} \frac{(ip_1)^{r_1}\ldots(ip_k)^{r_k}}{r_1!\ldots r_k!} + n\tilde{R}_l(p/\sqrt{n}) \end{aligned}$$

现在，我们先将 $\rho \gg 1$ 设为稍后确定的值，并考虑满足 $|p| < \rho$ 的p 。那么，存在正常数$n\rho, C_\rho > 0$，使得对于所有$n > n\rho$ p 和 $|p| \le \rho$，均有

$$n|\tilde{R}_l(p/\sqrt{n})| \sim o\left(\frac{C_\rho}{n^{l/2-1}}\right).$$

此时，我们利用假设 $\lambda_{r_1,\ldots,r_k} = 0 = \mu_{r_1,\ldots,r_k}$当 $r_1 + \ldots + r_k = 1$。代入上式可得出结论：对于 $|p| \le \rho$

$$\begin{aligned} F_{Y_n}(p) = &-\sum_{r_1+\ldots+r_k=2} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}(ip_1)^{r_1}\ldots(ip_k)^{r_k} \\ &+ \sum_{r_1+\ldots+r_k>2}^{l} \frac{1}{n^{\frac{r_1+\ldots+r_k}{2}-1}} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}(ip_1)^{r_1}\ldots(ip_k)^{r_k} + o(n^{-l/2+1}) \end{aligned}$$

其中 $o(n^{-l/2+1})$ 仅取决于 ρ。因此，对于满足 $|p| \leq \rho$ 的此类 p

$$\begin{aligned} Z_{Y_n}(p) &= \exp F_{Y_n}(p) \\ &= \exp\left(- \sum_{r_1+\ldots+r_k=2} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!} (ip_1)^{r_1}\ldots(ip_k)^{r_k} \right) \times \\ &\quad \times \exp\left(\sum_{r_1+\ldots+r_k>2}^{l} \frac{1}{n^{\frac{r_1+\ldots+r_k}{2}-1}} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!} (ip_1)^{r_1}\ldots(ip_k)^{r_k} + o(n^{-l/2+1}) \right). \end{aligned}$$

考虑到注释 13 中的讨论，第一个指数项由

$$\exp\left(-\frac{1}{2} \sum_{i,j=1}^{k} C_{ij}\, p_i p_j \right),$$

我们将其写作 $e^{-\frac{1}{2}C^{ij}p_i p^j}$，并采用爱因斯坦求和约定。此外，展开第二个指数项后，我们发现它可以写成

$$1 + \sum_{m=1}^{l} \frac{1}{m!} \left(\sum_{r_1+\ldots+r_k>2}^{l} \frac{1}{n^{\frac{r_1+\ldots+r_k}{2}-1}} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!} (ip_1)^{r_1}\ldots(ip_k)^{r_k} \right)^m + o(n^{-l/2+1}).$$

接下来我们定义

$$Z^l_{Y_n}(p) = \left(1 + \sum_{m=1}^{l} \frac{1}{m!} \left(\sum_{r_1+\ldots+r_k>2}^{l} \frac{1}{n^{\frac{r_1+\ldots+r_k}{2}-1}} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!} (ip_1)^{r_1}\ldots(ip_k)^{r_k} \right)^m \right) e^{-\frac{1}{2}\sum_{i,j=1}^{k} C_{ij}p_i p_j},$$

其性质为：对于 $|p| \leq \rho$

$$|Z_{Y_n}(p) - Z^l_{Y_n}(p)| \lesssim o\left(n^{-l/2+1} |p|^l e^{-\frac{1}{2}\sum_{i,j=1}^{k} C_{ij}p_i p_j} \right). \tag{B.2}$$

回顾一下，我们的最终目标是计算 Y_n 的密度函数，该函数由 $Z^{Y_n}(p)$ 的傅里叶变换给出。正如我们现在将要看到的，除 $o(n^{-l/2+1})$ 阶的项之外，它可由 $Z^l_{Y_n}(p)$ 的傅里叶变换给出，具体如下

$$\begin{aligned} f_{Y_n}(Y) &= \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} Z_{Y_n}(p_1,\ldots,p_k) e^{-ip_i y_i} dp_1\ldots dp_k \\ &= \frac{1}{(2\pi)^k} \int_{|p|\leq\rho} Z_{Y_n}(p_1,\ldots,p_k) e^{-ip_i y_i} dp_1\ldots dp_k + \frac{1}{(2\pi)^k} \int_{|p|\geq\rho} Z_{Y_n}(p_1,\ldots,p_k) e^{-ip_i y_i} dp_1\ldots dp_k \\ &= \frac{1}{(2\pi)^k} \int_{|p|\leq\rho} Z_{Y_n}(p_1,\ldots,p_k) e^{-ip_i y_i} dp_1\ldots dp_k + \varepsilon_\rho, \end{aligned}$$

其中 $\lim_{\rho\to+\infty}\varepsilon_\rho = 0$。然后，利用公式 B.2

$$\begin{aligned} f_{Y_n}(y) &= \frac{1}{(2\pi)^k}\int_{|p|\le\rho} Z^l_{Y_n}(p_1,\dots,p_k)e^{-ip_iy_i}dp_1\dots dp_k \\ &\quad + \frac{1}{(2\pi)^k}\int_{|p|\le\rho} o\left(n^{-l/2+1}|p|^l e^{-\frac{1}{2}\sum_{i,j=1}^k C_{ij}p_ip_j}\right)e^{-ip_iy_i}dp_1\dots dp_k + \varepsilon(\rho) \\ &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} Z^l_{Y_n}(p_1,\dots,p_k)e^{-ip_iy_i}dp_1\dots dp_k \\ &\quad + \delta_\rho + \frac{1}{(2\pi)^k}\int_{|p|\le\rho} o\left(n^{-l/2+1}|p|^l e^{-\frac{1}{2}\sum_{i,j=1}^k C_{ij}p_ip_j}\right)e^{-ip_iy_i}dp_1\dots dp_k + \varepsilon_\rho, \end{aligned}$$

其中

$$\delta_\rho = -\frac{1}{(2\pi)^k}\int_{|p|>\rho} Z^l_{Y_n}(p_1,\dots,p_k)e^{-ip_ix_i}dp_1\dots dp_k,$$

它也满足 $\lim_{\rho\to+\infty}\delta_\rho = 0$。因此，我们可以选取一个序列 $\{\rho_n\}_{n\gg1}$，使其以足够快的速度趋于无穷，从而使 $\varepsilon_{\rho_n}+\delta_{\rho_n} = o(n^{-l/2+1})$

$$\begin{aligned} f_{Y_n}(y) &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k}\left(1+\sum_{m=1}^l\frac{1}{m!}\left(\sum_{r_1+\dots+r_k>2}^l \frac{1}{n^{\frac{r_1+\dots+r_k}{2}-1}}\frac{\lambda_{r_1,\dots,r_k}}{r_1!\dots r_k!}(ip_1)^{r_1}\dots(ip_k)^{r_k}\right)^m\right)\times \\ &\quad \times e^{-\frac{1}{2}C_{ij}p_ip_j}e^{-ip_iy_i}dp_1\dots dp_k + o(n^{-l/2+1}) \\ &= \left(1+\sum_{m=1}^l\frac{1}{m!}\left(\sum_{r_1+\dots+r_k>2}^l \frac{1}{n^{\frac{r_1+\dots+r_k}{2}-1}}\frac{\lambda_{r_1,\dots,r_k}}{r_1!\dots r_k!}(-\partial_{y_1})^{r_1}\dots(-\partial_{y_k})^{r_k}\right)^m\right)\frac{e^{-\frac{1}{2}C^{-1}_{ij}y_iy_j}}{\sqrt{(2\pi)^k\det(C)}} \\ &\quad + o(n^{-l/2+1}), \end{aligned} \tag{B.3}$$

其中 C^{-1}_{ij}表示协方差矩阵 C^{ij}的逆矩阵的i,j项。注意，从形式上看，我们实际上可以将其改写为

$$f_{Y_n}(y) = \exp\left(\sum_{r_1+\dots+r_k>2}^l \frac{1}{n^{\frac{r_1+\dots+r_k}{2}-1}}\frac{\lambda_{r_1,\dots,r_k}}{r_1!\dots r_k!}(-\partial_{y_1})^{r_1}\dots(-\partial_{y_k})^{r_k}\right)\frac{e^{-\frac{1}{2}C^{-1}_{ij}y_iy_j}}{\sqrt{(2\pi)^k\det(C)}}, \tag{B.4}$$

忽略阶为 $o(n^{-l/2+1})$ 的项。用 $\mu_{r_1,\dots,r_k}$ 表示时，也可以利用关系式 A.3 来写出累积量与矩之间的关系。

注释 14. 为了说明这一结果的用法，我们现在将按 n 的幂展开

$$I := \sum_{m=1}^l\frac{1}{m!}\left(\sum_{r_1+\dots+r_k>2}^l \frac{1}{n^{\frac{r_1+\dots+r_k}{2}-1}}\frac{\lambda_{r_1,\dots,r_k}}{r_1!\dots r_k!}(-\partial_{y_1})^{r_1}\dots(-\partial_{y_k})^{r_k}\right)^m,$$

$n^{-1/2}$。展开前几项，我们得到

$$I = \frac{1}{\sqrt{n}} \sum_{r_1+\ldots+r_k=3} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}(-\partial_{y_1})^{r_1}\ldots(-\partial_{y_k})^{r_k} + \\ + \frac{1}{n} \sum_{r_1+\ldots+r_k=4} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!}(-\partial_{y_1})^{r_1}\ldots(-\partial_{y_k})^{r_k} \\ + \frac{1}{2n} \sum_{r_1+\ldots+r_k=3} \sum_{r_1'+\ldots+r_k'=3} \frac{\lambda_{r_1,\ldots,r_k}}{r_1!\ldots r_k!} \frac{\lambda_{r_1',\ldots,r_k'}}{r_1'!\ldots r_k'!}(-\partial_{y_1})^{r_1+r_1'}\ldots(-\partial_{y_k})^{r_k+r_k'} + \ldots,$$

其中，.....表示阶为 $O(n^{-3/2})$ 的项。此外，请记住，该表达式中的累积量可通过公式 *A.3* 由矩表示。

我们的主要结果 1 可以立即得出：将附录本节的结果应用于随机变量 $X^{(i)} = (p_i(x_1), \ldots, p_i(x_k))$，此时 $Y^n = \frac{1}{\sqrt{n}} X^{(i)}$ 与 $(f_\theta(x_1), \ldots, f_\theta(x_k))$ 重合。 —=

B.3. $k = 1$ 的情形（例1的详细说明）。为了说明如何使用我们的主定理，我们现在讨论最简单且相关的特例，即 $k = 1$ 且 $\mu_1 = 0 = \lambda_1$。此时，方程 B.3 给出

$$f_{Y_n}(y) = \left(1 + \sum_{m=1}^{l} \frac{1}{m!} \left(\sum_{r>2} \frac{1}{n^{\frac{r}{2}-1}} \frac{\lambda_r}{r!}(-\partial_y)^r \right)^m \right) \frac{e^{-\frac{y^2}{2\lambda_2}}}{(2\pi\lambda_2)^{1/2}}.$$

为计算这一项，我们将展开 $I := \sum_{m=1}^{l} \frac{1}{m!} \left(\sum_{r\geq 2} \frac{1}{n^{\frac{r}{2}-1}} \frac{\lambda_r}{r!}(-\partial_y)^r \right)^m$，如下所示

$$I = \sum_{r>2}^{l} \frac{1}{n^{\frac{r}{2}-1}} \frac{\lambda_r}{r!}(-\partial_y)^r + \frac{1}{2} \sum_{s,s'>2} \frac{1}{n^{\frac{s+s'}{2}-2}} \frac{\lambda_s}{s!} \frac{\lambda_{s'}}{s'!}(-\partial_y)^{s+s'} \\ + \frac{1}{6} \sum_{s,s',s''>2} \frac{1}{n^{\frac{s+s'+s''}{2}-3}} \frac{\lambda_s}{s!} \frac{\lambda_{s'}}{s'!} \frac{\lambda_{s''}}{s''!}(-\partial_y)^{s+s'+s''} + \ldots \\ = \sum_{r>2}^{l} \frac{1}{n^{\frac{r}{2}}} \left(n\frac{\lambda_r}{r!} + \frac{n^2}{2} \sum_{s+s'=r}{}' \frac{\lambda_s}{s!} \frac{\lambda_{s'}}{s'!} + \frac{n^3}{6} \sum_{s+s'+s''=r}{}' \frac{\lambda_s}{s!} \frac{\lambda_{s'}}{s'!} \frac{\lambda_{s''}}{s''!} + \ldots \right) \partial_y^r,$$

其中的求和 $\sum'$ 表示求和中的每一项都必须大于 2。

注释 15. 最后一个公式可以用更系统的方式写成

$$I = \sum_{r>2}^{l} \frac{1}{n^{\frac{r}{2}}} \left(\sum_{m\geq 1} \frac{n^m}{m!} \sum_{s_1+\ldots+s_m=r} \prod_{i=1}^{m} \frac{\lambda_{s_i}}{s_i!} \right) (-\partial_y)^r.$$

展开前几项可得

$$I = -\frac{\lambda_3}{3!n^{1/2}}\partial_y^3 + \frac{1}{n}\frac{\lambda_4}{4!}\partial_y^4 - \frac{1}{n^{3/2}}\frac{\lambda_5}{5!}\partial_y^5 + \frac{1}{n^3}\left(n\frac{\lambda_6}{6!} + \frac{n^2}{2}\frac{\lambda_3^2}{(3!)^2}\right)\partial_y^6 +$$
$$-\frac{1}{n^{7/2}}\left(n\frac{\lambda_7}{7!} + 2\frac{n^2}{2}\frac{\lambda_3}{3!}\frac{\lambda_4}{4!}\right)\partial_y^7 + \frac{1}{n^4}\left(n\frac{\lambda_8}{8!} + \frac{n^2}{2}\frac{\lambda_4^2}{(4!)^2}\right)\partial_y^8$$
$$-\frac{1}{n^{9/2}}\left(n\frac{\lambda_9}{9!} + 2\frac{n^2}{2}\frac{\lambda_5}{5!}\frac{\lambda_4}{4!} + \frac{n^3}{3}\frac{\lambda_3^3}{(3!)^3}\right)\partial_y^9 + \dots$$

并按$n^{-1/2}$的幂次整理后，得到

$$I = -\frac{\lambda_3}{3!n^{1/2}}\partial_y^3 + \frac{1}{n}\left(\frac{\lambda_4}{4!}\partial_y^4 + \frac{1}{2}\frac{\lambda_3^2}{(3!)^2}\partial_y^6\right) - \frac{1}{n^{3/2}}\left(\frac{\lambda_5}{5!}\partial_y^5 + \frac{\lambda_3}{3!}\frac{\lambda_4}{4!}\partial_y^7 + \frac{1}{3}\frac{\lambda_3^3}{(3!)^3}\partial_y^9\right) + \dots$$

其中$\dots$表示阶为$O(n^{-3/2})$的项。由例 2 可知，若 $\lambda_1 = 0 = \mu_1$，如我们当前情形所示，则 $\lambda_2 = \mu_2$，$\lambda_3 = \mu_3$ 和 $\lambda_4 = \mu_4 - 3\mu_2^2$。在这种情况下，我们有

$$I = \frac{1}{\sqrt{n}}\frac{\mu_3}{3!}\frac{d^3}{dy^3} + \frac{1}{n}\left[\frac{\mu_4 - 3\mu_2^2}{4!}\frac{d^4}{dy^4} + \frac{1}{2}\frac{\mu_3^2}{(3!)^2}\frac{d^6}{dy^6}\right] + \dots,$$

并回顾一下，$\dots$表示阶为$O(n^{-3/2})$的项。此外，将协方差矩阵C写作μ_2^2,，我们有$C^{-1} = {}^1\mu_2^2$；利用高斯函数的导数与厄米多项式$H_n(x)$之间的关系，可得

$$\frac{d^n}{dy^n}\frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} = \frac{(-1)^n}{(\sqrt{2\mu_2})^n}H_n\left(\frac{y}{\sqrt{2\mu_2}}\right)\frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}}.$$

将这两个方程代入 B.3 或 B.4，即得

$$f_{Y_n}(x) = \frac{e^{-\frac{x^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + \frac{1}{\sqrt{n}}\frac{\mu_3}{3!}\frac{1}{(\sqrt{2\mu_2})^3}H_3\left(\frac{x}{\sqrt{2\mu_2}}\right)\frac{e^{-\frac{x^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}}$$
$$+\frac{1}{n}\left(\frac{\mu_4 - 3\mu_2^2}{4!}\frac{1}{(\sqrt{2\mu_2})^4}H_4\left(\frac{x}{\sqrt{2\mu_2}}\right) + \frac{1}{2}\frac{\mu_3^2}{(3!)^2}\frac{1}{(\sqrt{2\mu_2})^6}H_6\left(\frac{x}{\sqrt{2\mu_2}}\right)\right)\frac{e^{-\frac{x^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + \dots,$$

我们将其更简洁地写作

$$f_{Y_n}(y) = \frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + \frac{1}{\sqrt{n}}\frac{\mu_3}{12\sqrt{2}\mu_2^{3/2}}H_3\left(\frac{y}{\sqrt{2\mu_2}}\right)\frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}}$$
$$+\frac{1}{n}\left(\frac{\mu_4 - 3\mu_2^2}{96\mu_2^2}H_4\left(\frac{y}{\sqrt{2\mu_2}}\right) + \frac{\mu_3^2}{576\mu_2^3}H_6\left(\frac{y}{\sqrt{2\mu_2}}\right)\right)\frac{e^{-\frac{y^2}{2\mu_2}}}{\sqrt{2\pi\mu_2}} + \dots,$$

至此完成计算，并得出例1中的结论。

附录 C. 训练期间输出的演化与收敛

考虑一个有限宽度单隐藏层神经网络。

$$f_\theta(\cdot) = f_\theta^{(n)}(\cdot) = \frac{1}{\sqrt{n}} \sum_{\mu=1}^{n} p_\mu(\cdot)\,, \tag{C.1}$$

由感知机构建

$$p_\mu(x) := \theta_\mu^{(3)} \sigma(\theta_\mu^{(1)} x + \theta_\mu^{(2)}) + \theta_\mu^{(4)}, \tag{C.2}$$

其中，转化为前述记号的对应关系为

$$W_\mu^{(k)} = \theta_\mu^{(2k-1)} \quad \text{and} \quad b_\mu^{(k)} = \theta_\mu^{(2k)}\,.$$

当需要明确表示对参数的依赖时，我们也将写作 $p(\theta,x)$。

与 f_θ 相关联的神经切线核定义为

$$\mathrm{NTK}_{zw}^{(n)}(\theta) := \sum_{i=1}^{4n} \frac{\partial f_\theta(z)}{\partial \theta_i} \frac{\partial f_\theta(w)}{\partial \theta_i}\,,$$

其中求和遍历参数 θ_i 的整个范围，即遍历 $i = 1, \cdots, P = 4n$。

接下来我们计算

$$\partial_t \mathrm{NTK}_{zw}^{(n)}(\theta(t)) = \sum_{i=1}^{4n} \left(\frac{d}{dt}\left(\frac{\partial f_\theta(z)}{\partial \theta_i}\right) \frac{\partial f_\theta(w)}{\partial \theta_i} + \frac{\partial f_\theta(z)}{\partial \theta_i} \frac{d}{dt}\left(\frac{\partial f_\theta(w)}{\partial \theta_i}\right) \right).$$

假设训练采用梯度下降法 (3.1)，并使用均方损失 (3.2)，则有

$$\begin{aligned}
\frac{d}{dt}\left(\frac{\partial f_\theta(z)}{\partial \theta_i}\right) &= \sum_{j=1}^{P} \frac{\partial^2 f_\theta(z)}{\partial \theta_j \partial \theta_i} \dot{\theta}_j \\
&= -\sum_{j=1}^{P} \frac{\partial^2 f_\theta}{\partial \theta_j \partial \theta_i}(z) \sum_{l=1}^{N} \frac{\partial f_\theta(X_l)}{\partial \theta_j} (f_\theta(X_l) - Y_l) \\
&= -\sum_{l=1}^{N} \left(\sum_{j=1}^{P} \frac{\partial^2 f_\theta}{\partial \theta_j \partial \theta_i}(z) \frac{\partial f_\theta(X_l)}{\partial \theta_j} \right) (f_\theta(X_l) - Y_l),
\end{aligned}$$

对于 $\frac{d}{dt}\left(\frac{\partial f_\theta}{\partial \theta_i}(w)\right)$，同理。将其代入神经切线核的演化方程可得

$$\partial_t \mathrm{NTK}_{zw} = -\sum_{l=1}^{N} \left(\sum_{i,j=1}^{P} \left(\frac{\partial^2 f_\theta}{\partial \theta_j \partial \theta_i}(z) \frac{\partial f_\theta(w)}{\partial \theta_i} \frac{\partial f_\theta(X_l)}{\partial \theta_j} + \frac{\partial^2 f_\theta}{\partial \theta_j \partial \theta_i}(w) \frac{\partial f_\theta(z)}{\partial \theta_i} \frac{\partial f_\theta(X_l)}{\partial \theta_j} \right) (f_\theta(X_l) - Y_l) \right)$$

现在，利用（C.1）和（C.2）给出的f_θ 的具体形式，我们发现

$$\frac{\partial^2 f_\theta}{\partial\theta_j\partial\theta_i}(z)\frac{\partial f_\theta(w)}{\partial\theta_i}\frac{\partial f_\theta(X_l)}{\partial\theta_j}=\frac{1}{n^{3/2}}\sum_{\mu,\nu,\lambda=1}^{n}\frac{\partial^2 p_\mu}{\partial\theta_j\partial\theta_i}(z)\frac{\partial p_\nu(w)}{\partial\theta_i}\frac{\partial p_\lambda(X_l)}{\partial\theta_j}$$
$$=\frac{1}{n^{3/2}}\sum_{\mu=1}^{n}\frac{\partial^2 p_\mu}{\partial\theta_j\partial\theta_i}(z)\frac{\partial p_\mu(w)}{\partial\theta_i}\frac{\partial p_\mu(X_l)}{\partial\theta_j}.$$

那么，

$$\begin{aligned}\partial_t\mathrm{NTK}^{(n)}_{zw}(\theta(t))&=-\frac{1}{n^{3/2}}\sum_{l=1}^{N}\left(\sum_{i,j=1}^{P}\sum_{\mu=1}^{n}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(w)}{\partial\theta_i}+\frac{\partial^2 p_\mu(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(z)}{\partial\theta_i}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_j}(f_{\theta(t)}(X_l)-Y_l)\right)\\&=-\frac{1}{n^{3/2}}\sum_{\mu=1}^{n}\left(\sum_{i,j=1}^{P}\sum_{l=1}^{N}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(w)}{\partial\theta_i}+\frac{\partial^2 p_\mu(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(z)}{\partial\theta_i}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_j}(f_{\theta(t)}(X_l)-Y_l)\right)\\&=-\frac{1}{\sqrt{n}}\left[\frac{1}{n}\sum_{\mu=1}^{n}\left(\sum_{i,j=1}^{P}\sum_{l=1}^{N}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(w)}{\partial\theta_i}+\frac{\partial^2 p_\mu(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(z)}{\partial\theta_i}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_j}(f_{\theta(t)}(X_l)-Y_l)\right)\right]\\&=-\frac{1}{\sqrt{n}}\sum_{l=1}^{N}(f_{\theta(t)}(X_l)-Y_l)\left[\frac{1}{n}\sum_{\mu=1}^{n}\left(\sum_{i,j=1}^{P}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(w)}{\partial\theta_i}+\frac{\partial^2 p_\mu(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(z)}{\partial\theta_i}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_j}\right)\right].\end{aligned}$$

方括号内的量是三阶核

$$K^{(3)}_n(z,w,X_l):=\frac{1}{n}\sum_{\mu=1}^{n}\left(\sum_{i,j=1}^{P}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(w)}{\partial\theta_i}+\frac{\partial^2 p_\mu(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(z)}{\partial\theta_i}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_j}\right).$$

注释 16. 顺便指出，根据初始化时的大数定律（即当 $t=0$ 时），$K^{(3)}$ 几乎必然收敛（当 $n\to+\infty$ 时）至

$$K^{(3)}_\infty(z,w,v):=\mathbb{E}\left[\sum_{i,j}\left(\frac{\partial^2 p(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p(w)}{\partial\theta_i}+\frac{\partial^2 p(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p(z)}{\partial\theta_i}\right)\frac{\partial p(v)}{\partial\theta_j}\right],$$

其中，$\mathbb{E}[\cdot]$ 表示关于编码感知机 $p(\cdot)$的参数$\{\theta^{(i)}\}_{i=1}^4$ 的期望值。

从现在起，固定一个足够大的n，其取值依据下文将予以说明的规则确定。如果权重以具有有限二阶矩的独立同分布方式初始化，那么根据上文所述的中心极限定理以及$f_{\theta(0)}$ 的连续性，对于所有 $\delta>0$，都存在$C_\delta>0$，与所有足够大的n无关，使得

$$\mathbb{P}\left(|f_{\theta(0)}(x)|\le C_\delta\right)\ge 1-\delta\ ,\ \forall|x|\le 1\ . \tag{C.3}$$

本节的结果仅适用于非多项式激活函数 σ，这类函数至多线性增长，且一阶和二阶导数有界；也就是说，我们假设存在$C_\sigma>0$，使得

$$|\sigma(z)|\le C_\sigma(|z|+1)\quad\text{and}\quad|\sigma'(z)|+|\sigma''(z)|\le C_\sigma\ ,\forall z\in\mathbb{R}\,. \tag{C.4}$$

引入范数

$$\|\theta_\mu(t)\|_\infty = \sup_{j=1,\dots,4} |\theta_\mu^{(j)}(t)|\,, \tag{C.5}$$

我们看到，对于 $|x| \le 1$,

$$\left|\frac{\partial p_\mu}{\partial \theta_\mu^{(i)}}(\theta_\mu(t),x)\right| \le 2C_\sigma\left(\|\theta_\mu(t)\|_\infty + 1\right) \tag{C.6}$$

以及】【。

$$\left|\frac{\partial^2 p_\mu}{\partial \theta_\mu^{(i)}\partial \theta_\mu^{(j)}}(\theta_\mu(t),x)\right| \le C_\sigma\left(\|\theta_\mu(t)\|_\infty + 1\right)^2\,. \tag{C.7}$$

因此

$$\begin{aligned}\left|\sum_{i,j=1}^{P}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(w)}{\partial\theta_i} + \frac{\partial^2 p_\mu(w)}{\partial\theta_j\partial\theta_i}\frac{\partial p_\mu(z)}{\partial\theta_i}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_j}\right| &= \left|\sum_{i,j=1}^{4}\left(\frac{\partial^2 p_\mu(z)}{\partial\theta_\mu^{(j)}\partial\theta_\mu^{(i)}}\frac{\partial p_\mu(w)}{\partial\theta_\mu^{(i)}} + \frac{\partial^2 p_\mu(w)}{\partial\theta_\mu^{(j)}\partial\theta_\mu^{(i)}}\frac{\partial p_\mu(z)}{\partial\theta_\mu^{(i)}}\right)\frac{\partial p_\mu(X_l)}{\partial\theta_\mu^{(j)}}\right| \\ &\le CC_\sigma^3\left(\|\theta_\mu(t)\|_\infty + 1\right)^3\,,\end{aligned}$$

由此我们得出结论：

$$\left|\partial_t \mathrm{NTK}_{zw}^{(n)}(\theta(t))\right| \le \frac{1}{n}\sum_{\mu=1}^{n} CC_\sigma^3\left(\|\theta_\mu(t)\|_\infty + 1\right)^3 \frac{1}{\sqrt{n}}\sum_{l=1}^{N}\left|f_{\theta(t)}(X_l) - Y_l\right|\,. \tag{C.8}$$

为对所有相关量（核、输出和参数）在所有 $t \ge 0$ 上统一获得所需的控制，我们通过自助法/连续性论证来推进：为此我们考虑集合

$$\mathcal{T} = \mathcal{T}_\varepsilon := \{t \ge 0 : |\mathrm{NTK}_{zw}^{(n)}(\theta(s))| \le \varepsilon\,, \text{ for all } s \in [0,t] \text{ and all } |z|\,,|w| \le 1\}\,. \tag{C.9}$$

现在，鉴于 (3.6) 以及核关于输入（z,w）的连续性，我们看到，对于任意$\varepsilon > 0$，当固定 $n \ge m(\varepsilon^{-1})$ 时，以概率 $1-\delta$ 有 $0 \in \mathrm{T}_\varepsilon$，其中 m 是一个充分大的递增函数。由连续性可知，T_ε 也是闭集。我们的最终目标是证明，对于适当选择的 n 和 ε，相应的 T 也是开集，因此等于 $[0,+\infty[$。

对于 $t \in \mathrm{T}_\varepsilon$，我们可以利用 (C.8) 和 (3.6) 证明，以概率 $1-\delta$,

$$\begin{aligned}|\text{神经正切核}_{zw}^{(n)}(\theta(t)) - \text{神经正切核}_{zw}^{\infty}| &\le |\text{神经正切核}_{zw}^{(n)}(\theta(0)) - \text{神经正切核}_{zw}^{\infty}| + |\text{神经正切核}_{zw}^{(n)}(\theta(t)) - \text{神经正切核}_{zw}^{(n)}(\theta(0))| \\ &\le \frac{C\,(\log n)^{1/2+}}{\sqrt{n}} + \int_0^t \left|\partial_t \mathrm{NTK}_{zw}^{(n)}(\theta(s))\right| ds \\ &\le \frac{C\,(\log n)^{1/2+}}{\sqrt{n}} + \int_0^t \left[\frac{1}{n}\sum_{\mu=1}^{n} CC_\sigma^3\left(\|\theta_\mu(s)\|_\infty + 1\right)^3\right]\frac{1}{\sqrt{n}}\sum_{l=1}^{N}\left|f_{\theta(s)}(X_l) - Y_l\right| ds\,.\end{aligned} \tag{C.10}$$

我们通过估计积分中的项来继续。为此，令 $K_{ij}^{\infty} =$ 神经正切核$_{X_iX_j}^{\infty}$，$K^{ij} = K^{ij}(t) =$ 神经正切核$^{(n)}{}_{X_iX_j}(\theta(t))$，$y_i = f^{\theta(t)}(X_i)$ 以及 $E_i = y_i - Y_i$。此时，重要的是回顾一下，K_{ij}^{∞}几乎处处与时间无关（3.5），并且还是严格正定的 [7]；因此，通过选择

ε 足够小时，我们看到 K 的最小特征值有一个与时间无关的下界 $\lambda > 0$。因此，当$t \in \mathrm{T}\varepsilon$ 时，我们有（回顾 (3.3)）

$$\partial_t E_j = -K_{ji} E_i \, .$$

乘以 $2E^j$ 并对j 求和可得

$$\partial_t \|E\|_2^2 = -2E^\intercal K E \le -2\lambda E^\intercal E \le 2\lambda \|E\|_2^2 \, .$$

由此可得 $\|E(t)\|_2^2 \le C\|E(0)\|_2^2 e^{-2\lambda t}$。利用 $\|g\|_\infty^2 \le \|g\|_2^2 \le N\|g\|_\infty^2$ 可得

$$\|E(t)\|_\infty^2 \le C\|E(0)\|_2^2 e^{-2\lambda t} \le CN\|E(0)\|_\infty^2 e^{-2\lambda t} \, .$$

因此 $\|E(t)\|_\infty \le C\|E(0)\|_\infty e^{-\lambda t}$。特别地，我们有

$$|f_{\theta(t)}(X_i) - Y_i| = |E_i(t)| = |y_i(t) - Y_i| \le C\|E(0)\|_\infty e^{-\lambda t} \, , \tag{C.11}$$

对于所有 $t \in \mathrm{T}$，且其中 $\|E(0)\|_\infty := \sup_{i=1,\dots,N} |E_i(0)|$。请注意，当我们最终完成自助法论证并证明 $\mathrm{T} = [0, \infty[$, 时，上述估计还将意味着有限宽度网络的训练收敛于全局最小值。

继续之前，回顾一下（有限宽度网络的）参数在训练期间的演化由下式支配

$$\partial_t \theta_\nu^{(j)}(t) = -\sum_{l=1}^{N} \frac{\partial f_\theta(X_l)}{\partial \theta_\nu^{(j)}} (f_\theta(X_l) - Y_l) \, .$$

But

$$\frac{\partial f_\theta(x)}{\partial \theta_\nu^{(j)}} = \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta_\nu^{(j)}} \sum_{\mu=1}^{n} p_\mu(x) = \frac{1}{\sqrt{n}} \frac{\partial p_{\nu_j}(x)}{\partial \theta_\nu^{(j)}}$$

因此，鉴于 (C.6)，我们有

$$\left| \frac{\partial f_\theta(x)}{\partial \theta_\nu^{(j)}} \right| \le 2 \frac{C_\sigma}{\sqrt{n}} (\|\theta_\nu\|_\infty + 1) \, .$$

回顾 (C.11)，可知，对于 $t \in \mathrm{T}$,

$$\partial_t \theta_\mu^{(j)}(t) \le \left| \partial_t \theta_\mu^{(j)}(t) \right| \le \frac{NCC_\sigma}{\sqrt{n}} \left(\|\theta_\mu(t)\|_\infty + 1 \right) \|E(0)\|_\infty e^{-\lambda t} \, .$$

$$\begin{aligned} \left| \theta_\mu^{(j)}(t) \right| &\le \left| \theta_\mu^{(j)}(0) \right| + \int_0^t \left| \partial_t \theta_\mu(s) \right| ds \\ &\le \left| \theta_\mu^{(j)}(0) \right| + \int_0^t \frac{NCC_\sigma}{\sqrt{n}} \left(\|\theta_\mu(s)\|_\infty + 1 \right) \|E(0)\|_\infty e^{-\lambda s} ds \, . \end{aligned}$$

对最后一个估计式两边分别取上确界，得到

$$\left\| \theta_\mu(t) \right\|_\infty + 1 \le \left\| \theta_\mu(0) \right\|_\infty + 1 + \int_0^t \frac{NCC_\sigma}{\sqrt{n}} \left(\|\theta_\mu(s)\|_\infty + 1 \right) \|E(0)\|_\infty e^{-\lambda s} ds \, ,$$

并应用 Gronwäll不等式，我们得到

$$\begin{aligned}\|\theta_\mu(t)\|_\infty + 1 &\leq \left(\|\theta_\mu(0)\|_\infty + 1\right)\exp\left(\int_0^t \frac{NCC_\sigma}{\sqrt{n}}\|E(0)\|_\infty e^{-\lambda s} ds\right) \\ &\leq \left(\|\theta_\mu(0)\|_\infty + 1\right)\exp\left(\frac{NCC_\sigma}{\sqrt{n}\lambda}\|E(0)\|_\infty\right).\end{aligned}$$

由 (C.3) 可知，以概率 $1-\delta$ 有

$$\|E(0)\|_\infty \leq C_\delta + C_Y\,,$$

对于某个常数的取值，使得 $|Y_l| \leq C_Y$，对所有 $l = 1\ldots, N$ 均成立；特别地， $\|E(0)\|_\infty$ 有界，且其上界为与n无关的常数。

最后这些估计的一个重要推论是，存在一个$C > 0$，它与n无关，使得

$$\frac{1}{n}\sum_{\mu=1}^{n}\left(\|\theta_\mu(t)\|_\infty + 1\right)^3 \leq \frac{1}{n}\sum_{\mu=1}^{n} e^C\left(\|\theta_\mu(0)\|_\infty + 1\right)^3\,. \tag{C.12}$$

由于我们假设权重以独立同分布方式初始化，且具有有限三阶矩，因此根据大数定律，右侧几乎必然收敛到一个常数。特别地，存在一个常数$C > 0$，与n无关，使得

$$\frac{1}{n}\sum_{\mu=1}^{n}\left(\|\theta_\mu(t)\|_\infty + 1\right)^3 \leq C \quad,\quad \text{a.s.} \tag{C.13}$$

因此

$$\begin{aligned}\left|\partial_t \mathrm{NTK}_{zw}^{(n)}(\theta(t))\right| &\leq \frac{1}{n}\sum_{\mu=1}^{n} CC_\sigma^3\left(\|\theta_\mu(t)\|_\infty + 1\right)^3 \frac{1}{\sqrt{n}}\sum_{l=1}^{N}\left|f_{\theta(t)}(X_l) - Y_l\right| \\ &\leq \frac{NC\|E(0)\|_\infty}{\sqrt{n}} e^{-\lambda t}\,,\end{aligned} \tag{C.14}$$

对所有 $t \in \mathrm{T}$。

最后，我们回到 (C.10)，并利用 (C.11) 证明

$$|\mathrm{NTK}_{zw}^{(n)}(\theta(t)) - \mathrm{NTK}_{zw}^{\infty}| \leq \frac{C\,(\log n)^{1/2+}}{\sqrt{n}} + \frac{NC\|E(0)\|_\infty}{\sqrt{n}\lambda}\,, \tag{C.15}$$

通过适当选择 n，可以使其小于 $\varepsilon/2$。这表明 T 是开集；又因为它也是闭集且非空，我们最终得出 $\mathrm{T} = [0, +\infty[$。特别要注意的是，由此可知，所有推导出的估计对所有 $t \geq 0$ 均成立。

现在，我们借此机会研究并比较有限宽度和无限宽度网络的通用输出在训练过程中的演化。

回想一下，与有限宽度网络$f^{\theta(t)}$相关联、核为神经正切核$^\infty$的“无限宽度”网络，是函数$f_\infty(\cdot, t) : \mathrm{R} \to \mathrm{R}$，定义如下：首先，我们在训练集 $\{X_i\}$ 上定义它为

初值问题的解

$$\begin{cases} \partial_t f_\infty(X_i,t) = -\mathrm{NTK}^\infty_{X_iX_j}(f_\infty(X_j,t)-Y_j) \\ f_\infty(X_i,0) = f_{\theta(0)}(X_i)\,, \end{cases} \tag{C.16}$$

对重复指标求和。对于一般的 x，我们将 $f_\infty(x,t)$ 定义为以下方程的解：

$$\begin{cases} \partial_t f_\infty(x,t) = -\mathrm{NTK}^\infty_{xX_j}(f_\infty(X_j,t)-Y_j) \\ f_\infty(x,0) = f_{\theta(0)}(x)\,. \end{cases} \tag{C.17}$$

我们再次强调，在我们的框架中，无限核 NTK^∞ 几乎必然与时间无关；特别地，这意味着前述常微分方程组几乎必然是线性的，因此对所有 $t\in\mathbb{R}$ 都适定。

为简化记号，令 $K^\infty_{ij}=\mathrm{NTK}^\infty_{X_iX_j}$，并设 $y^\infty_i = f_\infty(X_i,\theta(t))$ 和 $E^\infty_i = y^\infty_i - Y_i$。于是，定义常微分方程组 (C.16) 可写为：

$$\partial_t E^\infty_j = \partial_t y^\infty_j = -K^\infty_{ji}E^\infty_i\,,\ j=1,\cdots,N\,. \tag{C.18}$$

由于 $K^\infty=(K^\infty_{ij})$ 是正定的，其最小特征值（记为 λ_∞）为正。由此，根据 (C.18)，我们得出：

$$|E^\infty_i| = |y^\infty_i - Y_i| \le C\|E^\infty(0)\|_\infty e^{-\lambda_\infty t}\,, \tag{C.19}$$

其中

$$\|E^\infty(0)\|_\infty := \sup_{i=1,\cdots,N}|E^\infty_i(0)|\,.$$

令K^∞_{xi}:= 神经正切核$^\infty_{xX_i}$ 并引入记号 $y^\infty = y^\infty(t) = f_\infty(x,t)$，则

$$\partial_t y^\infty = -K^\infty_{xi}E^\infty_i \tag{C.20}$$

并且由于 (C.18) 意味着

$$E^\infty_i = -(K^\infty)^{-1}_{ij}\partial_t E^\infty_j\,, \tag{C.21}$$

利用 (C.20)，我们得出

$$y^\infty(t) = y^\infty(0) - \int_0^t K^\infty_{xi}E^\infty_i(s)ds \tag{C.22}$$

$$= y^\infty(0) + K^\infty_{xi}(K^\infty)^{-1}_{ij}\int_0^t \partial_t E^\infty_j(s)ds \tag{C.23}$$

$$= y^\infty(0) + K^\infty_{xi}(K^\infty)^{-1}_{ij}\left(E^\infty_j(t) - E^\infty_j(0)\right)\,. \tag{C.24}$$

取极限并使用（C.19），我们得到

$$\lim_{t\to\infty} y^\infty(t) = y^\infty(0) - K^\infty_{xi}(K^\infty)^{-1}_{ij}\left(y^\infty_j(0) - Y_j\right)\,, \tag{C.25}$$

这应与 [1,式 5]进行比较。

现在，我们可以通过注意到有限宽度网络与无限宽度网络之间的差异来进行估计，即

$$\begin{aligned}\partial_t(y_j^\infty - y_j) &= -K_{ji}^\infty E_i^\infty + K_{ji}E_i \\ &= -K_{ji}^\infty(y_i^\infty - y_i) + (K_{ji} - K_{ji}^\infty)E_i\,.\end{aligned}$$

将上式作为关于 $d = (d_1, \ldots, d_N)$、$d_i = y_i^\infty - y_i$ 的一阶线性方程积分，并回想根据构造，$d(0) = 0$，我们得到

$$d(t) = e^{-K^\infty t}\int_0^t (K(s) - K^\infty)E(s)e^{K^\infty s}ds\,.$$

利用 (C.11) 和 (C.15)，我们由此得出

$$|y_i^\infty(t) - y_i(t)| \le \|d(t)\|_\infty \le e^{-\lambda_\infty t}\frac{C\|E(0)\|}{\lambda}\frac{(\log(n))^{1/2+}}{\sqrt{n}}\int_0^t e^{(\lambda_\infty - \lambda)s}ds \tag{C.26}$$

$$\le \frac{C\|E(0)\|}{\lambda}\frac{(\log(n))^{1/2+}}{\sqrt{n}}e^{-\min\{\lambda_\infty,\lambda\}t}\,. \tag{C.27}$$

对于一般的输出，我们有

$$\begin{aligned}\partial_t(y^\infty - y) &= -K_{xi}^\infty E_i^\infty + K_{xi}E_i \\ &= -K_{xi}^\infty(y_i^\infty - y_i) + (K_{xi} - K_{xi}^\infty)(y_i - Y_i)\,,\end{aligned}$$

因此，由于 $y^\infty(t) = y(0)$，我们可以对其积分得到（前提是 $|x| \le 1$）

$$\begin{aligned}|y(t) - y^\infty(t)| &\le \int_0^t |K_{xi}^\infty||y_i(s) - y_i^\infty(s)|ds + \int_0^t \left|K_{xi}(s) - K_{xi}^\infty\right||y_i(s) - Y_i|ds \\ &\le \left(\max_{i=1,\cdots,N}|K_{xi}^\infty| + 1\right)\frac{C\|E(0)\|}{\min\{\lambda_\infty,\lambda\}}\frac{(\log(n))^{1/2+}}{\sqrt{n}}\,.\end{aligned}$$

这应与 [1, 定理 3.2]进行比较。

C.1. 一个小幅改进。本节刚刚证明的结果表明，对于任意 $\delta > 0$，存在：$C > 0$ 以及一个集合 $\mathrm{A}_\delta \subset \mathrm{R}^{4n}$，其测度为 $1 - \delta$；使得对于任意满足 $|x| \le 1$ 且 $\theta(0) \in \mathrm{A}\delta$ 的 x

$$|f_{\theta(t)}(x) - f_\infty(x,t)| \le \left(\max_{i=1,\cdots,N}|K_{xi}^\infty| + 1\right)\frac{C\|E(0)\|}{\min\{\lambda_\infty,\lambda\}}\frac{(\log(n))^{1/2+}}{\sqrt{n}}\,,$$

对于任意 $n \in \mathrm{N}$。然而，在某些应用中，让 δ 以一种使其随n 变化的方式依赖于 $n\,\delta \to 0$ as$n \to +\infty$是有用的。不幸的是，这样做时，量 $\|E(0)\|$ 无法得到一致有界。事实上，$\|E(0)\| \le C_\delta + \max_{i=1,\cdots,N}|Y_i|$，其中$C_\delta > 0$ 是方程 C.3 中引入的常数，而该方程又隐式定义了 A_δ。现在，根据 Berry-Esseen 定理可知

$$-\frac{1}{\sqrt{n}} + \mathbb{P}(|G| \le C_\delta) \lesssim \mathbb{P}(|f_{\theta(t)}(x)| \le C_\delta) \lesssim \frac{1}{\sqrt{n}} + \mathbb{P}(|G| \le C_\delta),$$

与 G 高斯分布。因此，$\mathrm{P}(|G| \le C\delta) \sim 1 - e^{-C_\delta^2}$。于是，只要满足以下条件，公式 C.3 就能保证成立：

$$-\frac{1}{\sqrt{n}} + (1 - e^{-C_\delta^2}) \gtrsim 1 - \delta,$$

我们可以将其改写为

$$\delta \gtrsim e^{-C_\delta^2} + \frac{1}{\sqrt{n}}.$$

我们得出以下结论。

命题 1.存在正常数$C_1, C_2 > 0$，具有以下含义。令 δ_n 和 C^{δ_n}为满足

$$C_1 \delta_n > e^{-C_{\delta_n}^2} + \frac{1}{\sqrt{n}}. \tag{C.28}$$

那么，存在一个集合 $\mathcal{A}^{\delta_m} \subset \mathbb{R}^{4n}$，其测度为 $1 - \delta_n$；使得对于任意满足 $|x| \le 1$ 且 $\theta(0) \in \mathcal{A}^{\delta_n}$ 的 x

$$|f_{\theta(t)}(x) - f_\infty(x,t)| \quad \le \quad \left(\max_{i=1,\cdots,N} |K_{xi}^\infty| + 1 \right) \frac{C_2 \|E(0)\|}{\min\{\lambda_\infty, \lambda\}} \frac{(\log(n))^{1/2+}}{\sqrt{n}},$$

对于任意$n \in \mathbb{N}$。此外，

$$\|E(0)\| \le C_{\delta_n} + \max_{i=1,\cdots,N} |Y_i|.$$

例3。考虑 $\delta_n = C^{\sqrt{(\log n)^{1/2+}}} n$ 对于一个常数 $C > 0$ 且

$$C_{\delta_n}^2 = \frac{1}{2} \log n - \left(\frac{1}{2} \right)^{1+} \log\log n.$$

我们可以检查 这些满足 $C.28$，因此前一个命题的结论有效，且re热 h

$$|f_{\theta(t)}(x) - f_\infty(x,t)| \quad \lesssim \quad \frac{(\log(n))^{1+}}{\sqrt{n}},$$

参考文献

[1] S. Arora, S. S. Du, W.Hu, Z. Li, R. Salakhutdinov, and R. Wang. 关于无限宽神经网络精确计算的研究。发表于神经信息处理系统进展，32，2018。[2] S. Arora, S. S. Du, W. Hu, Z. Li, and Ruosong Wang. 过参数化双层神经网络优化与泛化的细粒度分析。发表于第*36*届国际机器学习大会，ICML 2019。[3] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. 线性回归中的良性过拟合。发表于美国国家科学院院刊，117.48: 30063-30070，2020。[4] P. Billingsley.概率测度的收敛。约翰·威立父子公司，2013。[5] V. Bogachev.测度的弱收敛。美国数学会，2018。[6] Y. Cao, and and G. Quanquan. 宽而深的神经网络中随机梯度下降的泛化界。发表于神经信息处理系统进展，32，2019。[7] L. Carvalho, J. L. Costa, J. Mourão, and G. Oliveira. 关于宽而深的神经网络的神经切线核谱隙的研究，准备中。[8] L. Chizat, E. Oyallon, and F. Bach. 关于可微编程中的惰性训练。发表于神经信息处理系统进展，32，2019。[9] S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. 梯度下降可找到深度神经网络的全局极小值。发表于国际机器学习大会，2019。[10] S. Du, X. Zhai, B. Poczos, A. Singh. 梯度下降可证明能够优化过参数化神经网络。发表于国际学习表征大会，2018。[11] R. Durrett.概率：理论与例题，剑桥大学出版社，2019。[12] N. Radford.无限网络的先验，（技术报告编号 crg-tr-94-1），多伦多大学，1994。

[13] D. A. Roberts, 和 S. Yaida。《深度学习理论原理：理解神经网络的有效理论方法》*ff*，剑桥大学出版社，2022。[14] Z. Chen、G. Rotskoff、J. Bruna 和 E. Vanden-Eijnden。《浅层神经网络的动力学中心极限定理》，载于神经信息处理系统大会论文集，2020。[15] Hall, Peter。《自助法与埃奇沃思展开》。施普林格科学与商业媒体，2013。[16] A. Jacot、G. Franck 和 C. Hongler。《神经切线核：神经网络中的收敛与泛化》。载于神经信息处理系统进展，2018。[17] B. Ghorbani、S. Mei、T. Misiakiewicz 和 A. Montanari。《神经网络何时优于核方法？》。载于神经信息处理系统进展，33，2020。[18] J. Huang、H.-T. Yau。《深度神经网络动力学与神经切线层级》。载于第*37*届国际机器学习大会论文集，机器学习研究论文集 119:4542-4551，2020。[19] G. Karniadakis、I. G. Kevrekidis、L. Lu、P. Perdikaris、S. Wang 和 L. Yang。《物理信息机器学习》。载于《自然综述：物理》，3.6: 422-440，2021。[20] J. Lee、Y. Bahri、R. Novak、S. Schoenholz、J. Pennington、J. Sohl-Dickstein。《深度神经网络作为高斯过程》。载于国际学习表征大会，2018。[21] J. Lee、S. Schoenholz、J. Pennington、B. Adlam、L. Xiao、R. Novak 和 J. Sohl-Dickstein。《有限与无限神经网络：一项实证研究》。载于神经信息处理系统进展，33: 15156-15172，2020。[22] J. Lee、L. Xiao、S. Schoenholz、Y. Bahri、R. Novak、J. Sohl-Dickstein 和 J. Pennington。《任意深度的宽神经网络在梯度下降下演化为线性模型》。载于神经信息处理系统进展，32，2019。[23] S. Mei 和 A. Montanari。《随机特征回归的泛化误差：精确渐近与双下降曲线》。《纯粹与应用数学通讯》，75.4: 667-766，2022。[24] B. Neyshabur、Z. Li、S. Bhojanapalli、Y. LeCun 和 N. Srebro。《探究过参数化在神经网络泛化中的作用》。载于国际学习表征大会，2019。[25] S. Oymak 和 M. Soltanolkotabi。《迈向适度过参数化：训练浅层神经网络的全局收敛保证》。载于*IEEE*信息论精选领域期刊，1.1: 84-105，2020。[26] Rivera, Alejandro。《高斯场的统计力学》。博士论文，格勒诺布尔阿尔卑斯大学，2018。[27] S. Wang、X. Yu 和 P. Perdikaris。《物理信息神经网络何时以及为何训练失败：神经切线核视角》。载于《计算物理学杂志》，449: 110768，2022。[28] K. Xu、M. Zhang、J. Li、S. Du、K. Kawarabayashi 和 S. Jegelka。《神经网络如何外推：从前馈网络到图神经网络》。载于国际学习表征大会，2021。[29] G. Yang。《权重共享的宽神经网络的尺度极限：高斯过程行为、梯度独立性与神经切线核推导》。预印本 arXiv:1902.04760，2019。[30] G. Yang 和 E. J. Hu。《张量程序 IV：无限宽神经网络中的特征学习》。载于国际机器学习大会，2021。[31] Y. Yamasaki。《无限维空间上的测度》，世界科学出版社，1985。

电子邮件地址: galato97@gmail.com