arXiv:2302.00401v1 [stat.ML] 1 Feb 2023

深度随机特征学习的确定性等价物与误差普适性

Dominik Schr¨oder[1*]、Hugo Cui[2*]、Daniil Dmitriev[3], 和 Bruno Loureiro[4]

[1]苏黎世联邦理工学院数学系，瑞士苏黎世 8006 ¨[2]计算统计物理实验室，物理研究所， ´洛桑联邦理工学院（EPFL），瑞士洛桑 1015´ [3]苏黎世联邦理工学院数学系及 ETH AI 中心，瑞士苏黎世 8092¨[4]巴黎高等师范学院（ENS）计算机科学系，PSL 与法国国家科学研究中心（CNRS），法国巴黎 75230 邮政信箱 05*dschroeder@ethz.ch, hugo.cui@epfl.ch,daniil.dmitriev@ai.ethz.ch,bruno.loureiro@di.ens.fr*[*]主要贡献

2023年2月2日

摘要

本文研究了使用中间层冻结、读出层可训练的全连接网络学习随机高斯网络函数的问题。这个问题可以视为将广泛研究的随机特征模型自然推广到更深架构。首先，我们证明了在学习器网络与目标网络共享相同中间层的岭回归设置下，测试误差具有高斯普适性，并给出了其精确的渐近公式。要建立这一结果，需要证明深度随机特征样本协方差矩阵的迹具有确定性等价物，这一结果本身也可能具有独立的研究价值。其次，我们猜想，在采用任意凸损失以及一般学习器／目标架构的更一般设置下，测试误差渐近地具有高斯普适性。我们为这一猜想提供了大量数值证据。基于这些结果，我们探究了架构设计与隐式正则化之间的相互作用。

1 引言

尽管深度神经网络在几乎所有知识领域的应用都取得了令人瞩目的实际进展，但我们目前对其理论理解在很大程度上仍不完整。理论方面的近期进展源于对简化设置的研究；尽管存在局限，这些设置往往能够捕捉“现实世界”神经网络的一些关键性质。一个值得注意的例子是近期关于随机特征网络（RFs）的一系列研究。随机特征最初由 [1] 提出，作为核方法的一种计算高效近似技术；近年来，它们则被用作惰性区间中双层神经网络的替代模型 [2–5]。随机特征网络是随机神经网络的一种特殊形式，其统计性质已在大量研究中得到探讨[6–10]。关于这类网络读出层训练的问题，在浅层（一层隐藏层）情形下已有研究 [4, 5], ，并给出了测试误差的精确渐近刻画。然而，类似研究在一般深层情形下仍然缺失。本文通过研究深度全连接随机神经网络最后一层的学习问题来弥补这一空白；下文将此模型称为深度随机特征（dRF）模型。具体而言，本文的**主要贡献**如下：

- 在第 3节中，我们陈述定理 3.3，该定理证明了确定性矩阵与逐层后激活的共轭核和样本协方差矩阵的乘积的迹所满足的渐近确定性等价。

- 作为定理 3.3的推论，我们在第 4 节中，针对目标网络与学习器网络共享相同中间层，且读出层采用平方损失进行训练的特殊情形，推导出 dRF模型测试误差的精确渐近公式。该结果确立了这一设置下岭回归测试误差的高斯等价性。

- 最后，我们猜想，对于一般的凸损失以及一般的目标函数/学习器网络架构，dRF模型具有高斯普适性，并为此提供了有力的数值证据。更具体地说，我们提供了精确的

测试误差渐近公式，该公式利用了高维统计领域的最新进展 [11] ，以及 [12]中网络激活总体协方差的闭式公式。这些公式表明，就二阶统计量而言，dRF等价于一个具有噪声层的线性网络。我们将在第5节讨论这种有效噪声如何转化为深度诱导的隐式正则化。

本研究所使用代码的 GitHub 仓库可在此处找到。

相关工作

随机特征最初由 [1]提出。单层共轭核的渐近谱密度在 [3, 13, 14]中得到了刻画。随机特征模型测试误差的尖锐渐近结果见于 [4, 15] ，针对岭回归；[5, 16] ，针对一般凸损失；以及 [17, 18] ，针对其他惩罚项。随机特征网络的隐式正则化在 [19]中进行了讨论。随机特征网络模型在许多不同的场景中都被用作理解过参数化的替代模型，例如用于不确定性量化 [20], 、集成 [21, 22], 和训练动力学 [23, 24], ，同时也用于揭示惰性训练的局限性 [25–28];

深度随机网络已被证明会收敛到高斯过程 [6, 7]。此外，文献 [29, 30], 研究了其在推断中的应用，文献 [31–33]则研究了其作为逆问题生成先验的应用。文献 [9, 10]刻画了深度随机网络输出的分布。与我们的工作密切相关的是文献 [8], ，该文为比例极限下神经切线核和共轭核的渐近谱密度及斯蒂尔切斯变换给出了精确公式。我们关于样本协方差和总体协方差的公式与其互为补充。近期，文献 [34, 35] 从贝叶斯学习的角度研究了线性宽度深度网络的测试误差;

随机特征网络的高斯普适性 测试误差已在 [4], 中得到证明； [5]猜想其对一般损失函数成立，并在 [36, 37]中得到了证明。研究还表明，高斯普适性对其他类别的特征也成立，例如双层神经切线核 [38], 核特征 [19, 24, 39, 40]。[11]为更一般的特征映射（包括预训练深度特征）的高斯普适性提供了数值证据。

样本协方差矩阵的确定性等价物 最早在 [41, 42] 中针对可分离协方差建立，从而推广了关于各向异性意义下谱的自由卷积的开创性工作 [43] 。最近，这些结果已扩展到非可分离协方差，首先是在迹意义下 [44], ，随后也扩展到各向异性意义下 [45, 46]。

2 设置与预备知识

设 $(\mathrm{x}^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n] \coloneqq \{1, \cdots, n\}$ 表示一些训练数据，其中 $\mathbf{x}^\mu \sim \mathcal{N}(0_d, \Omega_0)$ 相互独立，且$y^\mu = f_\star(\mathbf{x}^\mu)$ 是一个（可能为随机的）目标函数。本文旨在刻画广义线性估计的学习性能：

$$\hat{y} = \sigma\left(\frac{\theta^\top \varphi(\mathrm{x})}{\sqrt{k}}\right), \tag{1}$$

采用深度随机特征（dRF）：

$$\varphi(\mathrm{x}) \coloneqq \underbrace{(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1)}_{L}(\mathrm{x}), \tag{2}$$

其中后激活定义如下:

$$\varphi_\ell(h) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot h\right), \quad \ell \in [L]. \tag{3}$$

权重 $\{W_\ell \in \mathbb{R}^{k_\ell \times k_{\ell-1}}\}_{\ell \in [L]}$ 假设由具有独立同分布元素的高斯矩阵独立抽取得到$(W_\ell)_{ij} \sim \mathcal{N}(0, \Delta_\ell)$ $\forall 1 \le i \le k_\ell$, $1 \le j \le k_{\ell-1}$。为简化记号，有时将$k_L = k$记为 $\theta \in \mathbb{R}^k$。仅训练1中的读出权重，采用常规的正则化经验风险最小化方法:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \left[\sum_{\mu=1}^{n} \ell(y^\mu, \theta^\top \varphi(\mathrm{x}^\mu)) + \frac{\lambda}{2}||\theta||^2\right], \tag{4}$$

其中 $\ell: \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ 是一个损失函数，假设其为凸函数，而 $\lambda > 0$ 设定正则化强度 .

为了评估经验风险最小化器（4）的训练和测试性能，我们令 $g: \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ 为任意性能指标（例如损失函数本身，或者在分类情形下的误分类概率），并定义测试误差：

$$\epsilon_g(\hat{\theta}) := \mathbf{E}\left[g(y, \hat{\theta}^\top \varphi(\mathrm{x}))\right] \tag{5}$$

本文的主要目标是在5比例渐近区间$n,\ d,\ k_\ell \to \infty$ 且 $\mathcal{O}(1)$ 长宽比 $\alpha := n/d$ 和 $\gamma_\ell := k_\ell/d$固定时，对所有层索引 $\ell \in [L]$，精确刻画样本协方差、总体协方差以及后激活的格拉姆矩阵。

2.1 样本协方差矩阵的背景

Marchenko-Pastur 与自由概率：我们简要介绍样本协方差矩阵的基本术语。对于一个均值为零 $\mathbf{E}\, x = 0$、协方差为 $\Sigma := \mathbf{E}\, xx^\top \in \mathbb{R}^{d\times d}$ 的随机向量 $x \in \mathbb{R}^d$，我们将由 n 个独立副本 $x_1, \ldots, x_n$ 组成的矩阵$\widehat{\Sigma}$：$= \mathcal{X}\mathcal{X}^\top/n \in \mathbb{R}^{d\times d}$，用矩阵形式写作 $\mathcal{X} := (x_1, \ldots, x_n)$，称为与总体协方差矩阵 Σ 对应的样本协方差矩阵。Gram矩阵 $\check{\Sigma} := \mathcal{X}^\top\mathcal{X}/n \in \mathbb{R}^{n\times n}$ 与样本协方差矩阵具有相同的非零特征值，但特征向量并不相关。对样本协方差矩阵和格拉姆矩阵进行系统的数学研究由来已久，可追溯至 [47]。在 d 固定的“经典”统计极限 $n \to \infty$ 下，样本协方差矩阵收敛到总体协方差矩阵

$\widehat{\Sigma} \to \Sigma$;

而在比例区间 $d \sim n \gg 1$ 下，Σ 与Σ 的谱之间非平凡的渐近关系最早由开创性论文 [43] 得出：$\widehat{\Sigma}$ 的经验谱密度 $\mu(\widehat{\Sigma}) := d^{-1}\sum_{\lambda\in\mathrm{Spec}(\widehat{\Sigma})}$，近似等于 $\mu(\Sigma)$ 与长宽比为 $c = d/n$的 Marchenko-Pastur 分布 μ_{MP}^c 的自由乘法卷积，

$$\mu(\widehat{\Sigma}) \approx \mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^{d/n}. \tag{6}$$

这里，自由乘法卷积 $\mu \boxtimes \mu_{\mathrm{MP}}^c$ 可定义为唯一的分布 ν，其斯蒂尔切斯变换$m = m_\nu(z) := \int (x - z)^{-1}\, \mathrm{d}\nu(x)$ 满足标量自洽方程

$$zm = \frac{z}{1 - c - czm} m_\mu\left(\frac{z}{1 - c - czm}\right). \tag{7}$$

谱渐近（6）最初是在高斯 $\mathcal{X}$ 情形下得到的，或者更一般地，对于可分离相关性 $\mathcal{X} = \sqrt{\Sigma}Y$以及某个独立同分布矩阵 $Y \in \mathbb{R}^{d\times n}$也得到了这些结果。后来，这些结果被推广到 [44] 一般情形，其假设关于二次型集中于其期望附近，且本质上是最优的 $x^\top Ax \ \mathrm{Tr}\, A\Sigma$。

确定性等价物： 直到很久之后，人们才认识到 [41, 42] 渐近谱之间的关系（6）不仅适用于 Σ 和 $\widehat{\Sigma}$，$\check{\Sigma}$ 实际上也适用于特征向量，而且预解式 $\widehat{G}(z) := (\widehat{\Sigma} - z)^{-1}$，$\check{G}(z) := (\check{\Sigma} - z)^{-1}$ 渐近等于 确定性等价物

$$\widehat{M}(z) := -\frac{(\Sigma\check{m}(z) + I_d)^{-1}}{z}, \quad \check{M}(z) := \check{m}(z) I_n, \tag{8}$$

也适用于各向异性意义，而不只是迹意义；这凸显出，尽管它们的平均迹之间关系简单，

$$\widehat{m}(z) := m_{\mu(\Sigma)\boxtimes\mu_{\mathrm{MP}}^c}(z), \quad \check{m}(z) = \frac{c-1}{z} + c\widehat{m}(z),$$

样本协方差矩阵和格拉姆矩阵所包含的非谱信息却大不相同。预解式的各向异性集中（或用物理学术语来说，自平均）最初也是在高斯情形或可分离情形中得到的 [41, 42]。在利普希茨集中假设下，将其推广到一般样本协方差矩阵则是最近才实现的 [45, 46] 。在这项工作中，我们具体采用了 [46] 中针对一般协方差的样本协方差矩阵确定性等价物，并将其推广至格拉姆矩阵。

深度随机特征模型的应用： 在这项工作中，我们将各向异性确定性等价物的一般理论应用于深度随机特征模型。如第4节所述，即使对于简单的岭回归情形，要证明误差普适性，仅考虑矩阵的谱收敛还不够，还需要更强的结果。非线性激活函数的应用使该模型既非高斯，也不可分离，因此我们的分析依赖于 [46] 中的确定性等价物，以及我们对格拉姆矩阵的拓展；格拉姆矩阵会自然地出现在显式误差推导中。

2.2 记号

我们采用以下记号:

- 对于 $A \in \mathbb{R}^{n\times n}$，我们记 $\langle A\rangle := {}^1\!/_n \operatorname{tr} A$。
- 对于矩阵 $A \in \mathbb{R}^{n\times m}$，我们将（关于 ℓ^2-向量范数的）算子范数记为 $\|A\|$，将最大范数记为 $\|A\|_{\max} := \max_{ij} |A_{ij}|$，并将弗罗贝尼乌斯范数记为 $\|A\|_{\mathrm{F}}^2 := \sum_{ij} |A_{ij}|^2$。
- 对于任意分布 μ，我们将映射 $\lambda \mapsto a\lambda + b$ 下的推前记为 $a \otimes \mu \oplus b$，以避免与例如测度 μ_1, μ_2的凸组合 $a\mu_1 + (1-a)\mu_2$ 混淆。
- 如果对于所有充分小的 $\epsilon > 0$ 和充分大的 $D < \infty$，当 n 足够大时都有 $P(X_n > n^\epsilon Y_n) \le n^{-D}$，我们就称随机变量序列 $(X_n)_n$ 被另一随机变量序列 $(Y_n)_n$ 随机支配，此时记作$X_n \prec Y_n$。

3 确定性等价物

考虑由以下递推关系定义的方差序列

$$r_{\ell+1} = \Delta_{\ell+1}\, \mathbf{E}_{\xi\sim\mathcal{N}(0,r_\ell)} \left[\sigma_\ell(\xi)^2\right] \tag{9}$$

初始条件为 $r_1 := \Delta_1 {}^{\langle\Omega_0\rangle}\!/_d$，系数为

$$\begin{aligned} \kappa_1^\ell &= \frac{1}{r_\ell}\, \mathbf{E}_{\xi\sim\mathcal{N}(0,r_\ell)} \left[\xi\sigma_\ell(\xi)\right], \\ \kappa_*^\ell &= \sqrt{\mathbf{E}_{\xi\sim\mathcal{N}(0,r_\ell)} \left[\sigma_\ell(\xi)^2\right] - r_\ell \left(\kappa_1^\ell\right)^2}. \end{aligned} \tag{10}$$

3.1 多层样本协方差和 Gram 矩阵的严格结果 s

关于 dRFs 的各向异性确定性等价的主要结果，源于反复应用以下命题。我们考虑一个数据矩阵 $X_0 \in \mathbb{R}^{d\times n}$，其 Gram矩阵按如下方式集中:

$$\left\| \frac{X_0^\top X_0}{d} - r_1 I \right\|_{\max} \prec \frac{1}{\sqrt{n}}, \quad \left\| \frac{X_0}{\sqrt{d}} \right\| \prec 1 \tag{11}$$

对于某个正常数 r_1。例如，如果 X_0 的列向量 x 相互独立，均值为 $\mathrm{E}\,\mathbf{x} = 0$，协方差为 $\mathrm{E}\,\mathbf{x}\,\mathbf{x}^\top = \Omega_0 \in \mathbb{R}^{d\times d}$（并满足关于四阶矩的一些温和假设），则假设 (11) 得到满足；在这种情况下，$r_1 = \langle\Omega_0\rangle$ 是该协方差的归一化迹。随后我们考虑 $X_1 := \sigma_1(W_1 X_0/\sqrt{d})$，并假设 $W_1 \in \mathbb{R}^{k_1\times d}$ 的元素是独立同分布的。$\mathcal{N}(0,1)$ 元素，且 σ_1 在成比例$n \sim d \sim k_1$ 情形下满足 $\mathrm{E}_{\xi\sim\mathcal{N}(0,1)}\, \sigma_1(\sqrt{r_1}\xi) = 0$。通过改变 σ_1，不失一般性地假设 $\Delta_1 = 1$，出于记号简便起见，我们将采用这一假设。

命题 3.1（随机特征的确定性等价物）。对于任意确定性的 A 和利普希茨连续激活函数σ_1，在上述假设下，对于任意 $z \in \mathbf{C} \setminus \mathbb{R}_+$

$$\left| \left\langle A \Big[\Big(\frac{X_1^\top X_1}{k_1} - z \Big)^{-1} - \widehat{M}(z) \Big] \right\rangle \right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

and

$$\left| \left\langle A \Big(\frac{X_1 X_1^\top}{k_1} - z \Big)^{-1} \right\rangle - \langle A\rangle \breve{m}(z) \right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

其中 $\delta := \operatorname{dist}(z, \mathbb{R}_+)$,

$$\begin{aligned} -z\widehat{M}(z) &:= \Big(\breve{m}(z)\Sigma_{\mathrm{lin}} + I \Big)^{-1}, \\ \Sigma_{\mathrm{lin}} &:= (\kappa_1^1)^2 \frac{X_0^\top X_0}{d} + (\kappa_*^1)^2 I, \end{aligned} \tag{12}$$

and

$$\widehat{m}(z) := m_{\mu(\Sigma_{\text{lin}})\boxtimes\mu_{\text{MP}}^{n/k_1}}(z), \quad \breve{m}(z) = \frac{n-k_1}{nz} + \frac{n}{k_1}\widehat{m}(z).$$

此外，假设 (11) 成立，其中 X_0, r_1 分别由 X_1, r_2替代，并且我们有 $\operatorname{dist}(-1/\breve{m}(z)\,\mathbb{R}_+) \geq \operatorname{dist}(z, \mathbb{R}_+)$。

注释 3.2。 命题 3.1 的迹版本已在文献中多次出现，例如 [44]。它意味着 $\widehat{\mu}_1$ 的 $X_1^\top X_1/k_1$ 谱近似由自由乘法卷积给出

$$\begin{aligned}\widehat{\mu}_1 &\approx \mu\Big((\kappa_1^1)^2\frac{X_0^\top X_0}{d} + (\kappa_*^1)^2 I\Big) \boxtimes \mu_{\text{MP}}^{n/k_1} \\ &= \Big(\mu\Big((\kappa_1^1)^2\frac{X_0^\top X_0}{d}\Big) \boxplus \delta_{(\kappa_*^1)^2}\Big) \boxtimes \mu_{\text{MP}}^{n/k_1}.\end{aligned} \tag{13}$$

在 $c \leq 1$ 的情况下，也就是说，当 μ_{MP}^c 没有零点原子时， [48] 证明了

$$\sqrt{\mu \boxtimes \mu_{\text{MP}}^c} \boxplus_c \sqrt{\mu' \boxtimes \mu_{\text{MP}}^c} = \sqrt{(\mu \boxplus \mu') \boxtimes \mu_{\text{MP}}^c} \tag{14}$$

这使得可以简化（13）。这里 $\boxplus_c$ 是矩形自由卷积，它刻画了两个自由矩形随机矩阵之和的奇异值分布，而平方根应理解为平方根映射的推前。将（14）应用于（13）可得

$$\sqrt{\widehat{\mu}_1} \approx \Big(\kappa_1^1 \otimes \sqrt{\widehat{\mu}_0 \boxtimes \mu_{\text{MP}}^{n/k_1}}\Big) \boxplus_{n/k_1} \kappa_*^1 \otimes \sqrt{\mu_{\text{MP}}^{n/k_1}}, \tag{15}$$

这表明， $X_1/\sqrt{k}$ 的非零奇异值可以用以下高斯等效模型的非零奇异值来建模:

$$c'W'X_0 + c''W'' \tag{16}$$

对于某些适当选取的常数 c'， c'' 以及独立的高斯矩阵 W， W'。

命题 3.1 的最后一条断言允许对任意（但有限）数量的层进行迭代。事实上，经过一层后，我们有

$$\begin{aligned}\Big(\frac{X_1^\top X_1}{k_1} - z_1\Big)^{-1} &\approx \Big(-\breve{m}(z_1)z_1\Sigma_{\text{lin}} - z_1\Big)^{-1} \\ &= c_1\Big(\frac{X_0^\top X_0}{k_0} - z_0\Big)^{-1},\end{aligned} \tag{17}$$

使用定理 3.3 中的定义，针对 c_1， z_0 如下。

定理 3.3 （dRF 的确定性等价物）。 对于任意确定性的 A 以及 *Lipschitz* 连续的激活函数σ_1，……， σ_ℓ 满足 $\mathbf{E}^{\xi\sim\mathcal{N}(0,1)}\sigma_m(\sqrt{r_m}\xi) = 0$，在上述假设 (11) 下，对于任意 $z_\ell \in \mathbf{C} \setminus \mathbb{R}_+$

$$\left|\left\langle A\Big(\frac{X_\ell^\top X_\ell}{k_\ell} - z_\ell\Big)^{-1}\right\rangle - c_1\cdots c_\ell \breve{m}_0\langle A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}$$

并且

$$\left|\left\langle A\Big(\frac{X_\ell X_\ell^\top}{k_\ell} - z_\ell\Big)^{-1}\right\rangle - \breve{m}_\ell\langle A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

其中 $\delta :=\operatorname{dist}(z_\ell, \mathbb{R}_+)$，并递归定义

$$\begin{aligned}\Sigma_{\text{lin}}^{\ell-1} &:= (\kappa_1^\ell)^2\frac{X_{\ell-1}^\top X_{\ell-1}}{k_{\ell-1}} + (\kappa_*^\ell)^2 I, \\ \breve{m}_\ell &:= \frac{n-k_\ell}{nz_\ell} + \frac{n}{k_\ell} m_{\mu(\Sigma_{\text{lin}}^{\ell-1})\boxtimes\mu_{\text{MP}}^{n/k_\ell}}(z_\ell) \\ -\frac{1}{c_\ell} &:= \breve{m}_\ell z_\ell (\kappa_1^\ell)^2, \quad z_{\ell-1} := c_\ell z_\ell - \Big(\frac{\kappa_*^\ell}{\kappa_1^\ell}\Big)^2\end{aligned} \tag{18}$$

对于 $\ell \geq 1$， 最后

$$\breve{m}_0 := \frac{d-n}{nz_0} + \frac{d^2}{n^2} m_{\mu(\Omega_0)\boxtimes\mu_{\text{MP}}^{d/n}}\Big(\frac{d}{n}z_0\Big). \tag{19}$$

命题 3.1 和定理 3.3 的证明见附录A。

注记 3.4。 同样的迭代论证此前已出现在 [8] 中。与我们当前工作的主要区别在于，我们的估计具有各向异性，因此能够针对任意确定性矩阵检验样本协方差和 Gram 矩阵预解式。正如我们将在下一节讨论的，这对于给出深度随机特征模型测试误差的闭式渐近表达式至关重要。

3.2 总体协方差的闭式公式

在命题 3.1 和定理 3.3中，我们通过迭代将 $X_\ell^\top X_\ell/k_\ell$ 视为总体协方差为

$$\mathbf{E}_{W_\ell}\frac{X_\ell^\top X_\ell}{k_\ell}=\mathbf{E}_w\,\sigma_\ell\Big(\frac{X_{\ell-1}^\top w}{\sqrt{k_{\ell-1}}}\Big)\sigma_\ell\Big(\frac{w^\top X_{\ell-1}}{\sqrt{k_{\ell-1}}}\Big)\approx\Sigma_{\rm lin}^\ell$$

的样本协方差矩阵，并由此得到 $X_\ell^\top X_\ell$ 和 $X_\ell X_\ell^\top$ 的确定性等价物公式。更自然的方法是将 $X_\ell X_\ell^\top/n$ 视为总体协方差为

$$\Omega_\ell:=\mathbf{E}_{X_0}\frac{X_\ell X_\ell^\top}{n},\tag{20}$$

注意到，矩阵 X_ℓ 在给定 W_1，……，W_ℓ 的条件下具有独立列。定理 A.3 和命题 A.4 也适用于这种情形，但由于缺乏 Ω_ℓ 的严格表达式，所得的确定性等价物不如定理 3.3中的结果直观。近期，文献 [12]推导出了一个关于总体协方差的启发式闭式公式，并猜想该公式是精确的。现在我们讨论这一结果，并为求完整，在附录 B中给出推导。考虑由递推关系定义的矩阵序列 $\{\Omega_\ell^{\rm lin}\}_\ell$：

$$\Omega_{\ell+1}^{\rm lin}=\kappa_1^{(\ell+1)2}\frac{W_{\ell+1}\Omega_\ell^{\rm lin}W_{\ell+1}^\top}{k_\ell}+\kappa_*^{(\ell+1)2}I_{k_{\ell+1}}.\tag{21}$$

其中 $\Omega_0^{\rm lin}:=\Omega_0$。非正式地说，$\Omega_\ell^{\rm lin}$提供了 Ω_ℓ 的渐近近似，其意义在于归一化距离$|\Omega_\ell^{\rm lin}-\Omega_\ell||_F/\sqrt{d}$ 的阶为 $\mathcal{O}(1/\sqrt{d})$。此外，递推关系（21）意味着 $\Omega_\ell^{\rm lin}$可以表示为高斯矩阵（及其转置）的乘积之和，并且为推导其渐近谱分布的解析表达式提供了一种直接方法。该推导见附录 B。

一个有趣的问题是，能否通过定理21间接得到类似方程（3.3）中的总体协方差矩阵近似公式。关于这一逆问题已有大量研究，即如何根据样本协方差谱推断总体协方差谱的谱性质，例如 [49] 但我们将这一途径留待未来工作。

3.3 定理3.3与近似总体协方差的一致性

然而，我们可以注意到，公式（21）与定理 一致 3.3。我们在维度相等的情况下展示这一点 $n=d=k_1=\cdots=k_\ell$，以避免零特征值带来的不必要技术问题。我们定义

$$\widehat{\mu}_\ell:=\mu\Big(\frac{X_\ell^\top X_\ell}{k_\ell}\Big)=\check{\mu}_\ell:=\mu\Big(\frac{X_\ell X_\ell^\top}{n}\Big)\tag{22}$$

并回顾命题 3.1 意味着

$$\widehat{\mu}_\ell\approx((\kappa_1^l)^2\otimes\widehat{\mu}_{l-1}\oplus(\kappa_*^l)^2)\boxtimes\mu_{\rm MP}.\tag{23}$$

另一方面，将（6）应用于样本协方差矩阵 $X_\ell X_\ell^\top/n$ 及总体协方差 $\Omega_\ell\approx\Omega_\ell^{\rm lin}$意味着

$$\begin{aligned}\check{\mu}_\ell&\approx\mu(\Omega_\ell^{\rm lin})\boxtimes\mu_{\rm MP}\\&=\mu\Big((\kappa_1^\ell)^2\frac{W_\ell\Omega_{\ell-1}^{\rm lin}W_\ell^\top}{k_{\ell-1}}+(\kappa_*^\ell)^2I_{k_\ell}\Big)\boxtimes\mu_{\rm MP}\\&\approx\Big((\kappa_1^\ell)\otimes\mu(\Omega_{\ell-1}^{\rm lin})\boxtimes\mu_{\rm MP}\oplus(\kappa_*^\ell)^2\Big)\boxtimes\mu_{\rm MP}\\&\approx\Big((\kappa_1^\ell)\otimes\check{\mu}_{\ell-1}\oplus(\kappa_*^\ell)^2\Big)\boxtimes\mu_{\rm MP},\end{aligned}\tag{24}$$

这表明两种方法都导出了相同的递推关系。在这里，第三步中我们将 (6) 应用于样本协方差矩阵 $\sqrt{\Omega_{\ell-1}^{\rm lin}}W_\ell^\top$，并在第四步中使用了将 ℓ 替换为 $\ell-1$ 的一阶近似。

4 测试误差的高斯普适性

在本文的第二部分，我们将讨论如何利用关于特征的经验协方差和总体协方差渐近谱的结果，在标签由深度随机神经网络生成时，为测试误差和训练误差（5）给出精确表达式：

$$f_\star(\mathrm{x}^\mu) = \sigma^\star\left(\frac{\theta_\star^\top \varphi^\star(\mathrm{x}^\mu)}{\sqrt{k^\star}}\right). \tag{25}$$

特征映射 $\varphi^\star$ 表示 $L^\star+1$ 层的复合映射 $\varphi_{L^\star}^\star \circ \ldots \circ \varphi_1^\star$：

$$\varphi_\ell^\star(\mathrm{x}) = \sigma_\ell^*\left(\frac{1}{\sqrt{k_{\ell-1}^\star}} W_\ell^\star \cdot \mathrm{x}\right),$$

以及 $\theta_\star \in \mathbb{R}^{k^\star}$ 是最后一层权重。为简化记号，我们记 $k^\star := k_L^\star$。权重矩阵 $\{W_\ell^\star\}_{\ell\in[L^\star]}$的元素为独立同分布高斯条目，采样自 $\mathcal{N}(0,\ \Delta_\ell^\star)$。注意，我们不要求激活序列 $\{\sigma_\ell^\star\}_\ell$和宽度 $\{\gamma_\ell := k_\ell^\star/d\}_\ell$ 与学习器 dRF 的序列相匹配（2）。我们依次讨论

- 设定正确情形下，目标网络与学习器网络共享相同的中间层（即架构、激活函数和权重均相同） $\varphi_\ell^\star = \varphi_\ell$, $\ell \in [L]$ 其中 $L^\star = L$，且 dRF 的读出层通过岭回归进行训练。这等价于一个有趣的设定：对线性目标进行岭回归，特征则来自非高斯分布，该分布由高斯数据经过若干非线性层传播而形成。

- 一般情形下，目标网络与学习器网络的架构、激活函数和权重通常各不相同 ，并采用一般的凸损失。

对于这两种情形，我们都给出了测试误差的精确渐近刻画。此外，我们证明，后者等于一个等价学习问题在具有匹配的总体协方差的高斯样本上的测试误差，从而表明测试误差具有高斯普适性。在设定正确情形下，我们的结果是严格的，并利用了定理 3.3给出的确定性等价物。在完全一般的情形下，我们提出了一个猜想，并通过有限规模数值实验为其提供了有力支持。

4.1 设定正确情形

我们首先证明，在匹配设定下，采用平方损失训练的读出层所对应的深度随机特征的测试误差具有高斯普适性 $\varphi = \varphi^\star$。这对应于 $\mathcal{Y} = \mathbb{R}$，$\ell(y,\ \hat{}y) = 1/2(y - \hat{}y)^2$。这种情形尤为简单，因为经验风险最小化问题(4)存在如下闭式解：

$$\hat{\theta} = 1/\sqrt{k}(\lambda I_k + 1/k X_L X_L^\top)^{-1} X_L y \tag{26}$$

这里提醒读者，$X_L \in \mathbb{R}^{k\times n}$ 是将最后一层特征按列堆叠得到的矩阵，而$y \in \mathbb{R}^n$ 是标签向量。对于给定的目标函数，计算测试误差归结为一个随机矩阵理论问题，该问题取决于确定性矩阵与特征样本协方差矩阵的预解式之积的迹的变化（推导见附录C ）：

$$\begin{aligned}\epsilon_g(\hat{\theta}) = \Delta\left(\left\langle \Omega_L\left(\lambda I_k + 1/k X_L X_L\right)^{-1}\right\rangle + 1\right)\\ - \lambda(\lambda-\Delta)\partial_\lambda\left\langle \Omega_L\left(\lambda I_k + 1/k X_L X_L\right)^{-1}\right\rangle\end{aligned} \tag{27}$$

应用定理3.3可得以下推论：

推论 4.1 (匹配目标的岭回归普适性)。 设 $\lambda > 0$。在渐近极限 $n,\ d,\ k_\ell \to \infty$ 且固定 $\mathcal{O}(1)$比率 $\alpha = n/d,\ \gamma_\ell := k_\ell/d$ 并满足定理3.3的假设时，岭估计量(26) 在目标函数(25) 上的渐近测试误差为 $L = L^*$，且 $\varphi_\ell^* = \varphi_\ell$ 并具有方差为 $\Delta > 0$ 的加性高斯噪声时，其表达式为：

$$\begin{aligned}\epsilon_g(\hat{\theta}) \xrightarrow{k\to\infty} \epsilon_g^\star = \Delta\left(\langle\Omega_L\rangle \breve{m}_L(-\lambda) + 1\right)\\ - \lambda(\lambda-\Delta)\langle\Omega_L\rangle\partial_\lambda \breve{m}_L(-\lambda)\end{aligned} \tag{28}$$

其中 $\breve{m}_L$ 可分别由 (18) 递归计算。特别地，这意味着该模型的渐近均方误差具有高斯普适性，因为 (28) 与文献 *[50]*中推导的高斯数据 $\mathrm{x} \sim \mathcal{N}(0_d, \Omega_L)$ 上的岭回归渐近测试误差完全一致。

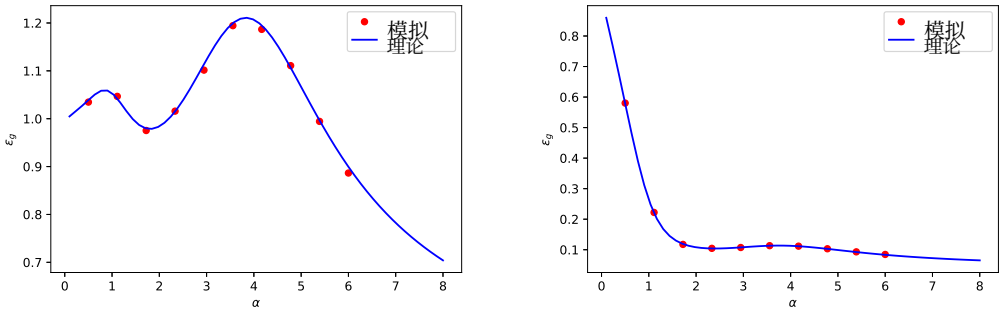

图1：岭回归的学习曲线 $\epsilon_g(\alpha)$（$\sigma_\star = id$, $\ell(y,z) = 1/2(y-z)^2$, 和 $g(y,\hat{}y) = (y-\hat{}y)^2$）。红点对应于对学习模型 (2) (25) 进行数值模拟的结果，取 20 次运行的平均值。实线对应于猜想4.3给出的尖锐渐近刻画，详见附录D。（左）双层目标（$L^\star = 1$,$\sigma_1^\star =$ 符号函数），（右）单层目标（$L^\star = 0$）。两者均使用具有 σ_1、$2(x)$ 双曲正切$(2x)$ 激活函数和正则化 $\lambda = 0.001$ 的 2–隐藏层随机特征 (2) 进行学习。

附录 C中给出了对（27）和推论 4.1 的详细推导，并讨论了确定性末层权重和一般目标函数的可能扩展。请注意，虽然建立设定正确情形下岭回归深层随机特征回归的高斯等价性并不需要总体协方差的迹 $\langle\Omega_L\rangle$，但它可以根据闭式公式（21）显式计算。

4.2 一般情形

尽管将随机矩阵理论工具箱应用于学习问题取得了重大进展，但其应用大多局限于二次问题，因为这类问题可以得到估计量的闭式表达式，例如（26）。要证明类似推论 4.1的普适性结果适用于二次问题以外的情形，是一项颇具挑战性的任务，近来已成为广受关注的研究课题。在广义线性估计（4）中，关于一般凸损失函数下 $L = 1$随机特征模型的测试误差普适性，文献 [5], 对其进行了启发式研究。作者表明，在高斯设计假设下得到的测试误差渐近公式，与采用真实特征进行的有限规模模拟结果完全吻合。此后，文献 [37] 通过结合林德伯格插值方案与广义中心极限定理，证明了测试误差的这种高斯普适性。接下来，我们旨在为多层随机特征情形作出与文献 [5] 类似的贡献。这一结果建立在本文所考虑的高维极限下，关于模型设定错误的广义线性估计渐近测试误差的严格闭式公式之上；该公式由文献 [11]推导得出。

我们证明，在高维极限下，第 2 节引入的模型的渐近测试误差属于高斯普适性类。更准确地说，该模型的测试误差渐近等价于一个等效高斯协变量模型（GCM）的测试误差；该模型是在数据集 $\check{\mathcal{D}} = \{v^\mu, \check{y}^\mu\}_{\mu\in[n]}$ 上进行广义线性估计，其标签为 $\check{y}^\mu = f_\star(1/\sqrt{k^\star}\theta_\star^\top u^\mu)$，协变量则服从联合高斯分布：

$$(u,v) \sim \mathcal{N}\begin{pmatrix} \Psi_{L^\star} & \Phi_{L^\star L} \\ \Phi_{L^\star L}^\top & \Omega_L \end{pmatrix} \tag{29}$$

这里回顾一下，Ω_L 是模型特征的方差（20），而 $\Phi \in \mathbb{R}^{k^\star\times k}$ 和 $\Psi \in \mathbb{R}^{k^\star\times k^\star}$ 分别是模型特征与目标特征之间的协方差，以及目标函数的方差：

$$\Phi_{L^\star L} := \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi(\mathrm{x})^\top\right], \ \Psi_{L^\star} := \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi^\star(\mathrm{x})^\top\right] \tag{30}$$

这一结果进一步丰富了近期高维线性估计普适性结果的研究 [11, 38, 51], , 并将 [15, 36, 37] 的随机特征普适性推广至 $L > 1$。该结果可概括为以下猜想：

猜想 *4.2*。在固定 $\mathcal{O}(1)$ 比率 $\alpha := n/d$ 和 $\gamma_\ell := k_\ell/d$ 时的高维极限 $n, d, k_\ell \to \infty$ 中，在 $\mathcal{D} = \{(x^\mu, y^\mu)\}_{\mu\in[n]}$ 上训练的经验风险最小化器 *(4)* 的测试误差，与具有匹配二阶矩 Ψ, Φ 的高斯协变量模型 *(29)* 的测试误差相同；其中协变量为 $x^\mu \sim \mathcal{N}(0_d, \Omega_0)$，标签来自 *(25)*，$\Omega$ 的定义见 *(20)* 和 *(30)*。

我们更进一步，给出测试误差的精确渐近表达式。递归构造以下矩阵序列

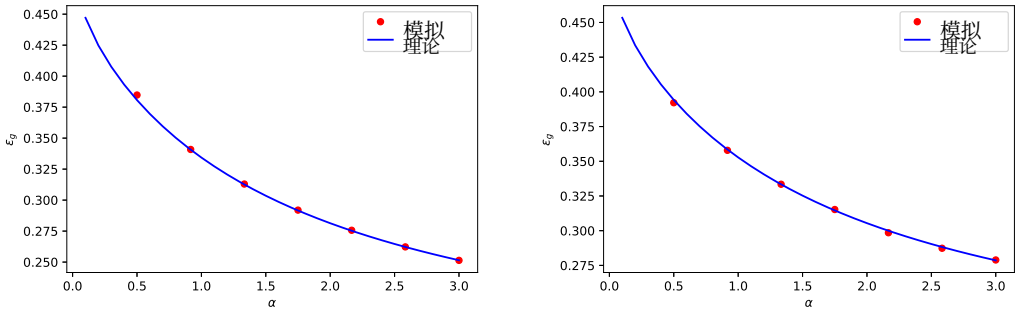

图2：逻辑回归的学习曲线 $\epsilon_g(\alpha)$（$\sigma_\star$ = 符号函数，$\ell(y,z) = \ln(1+e^{-yz})$，度量为 $g(y,\hat{y}) = 1-\Theta(y\hat{y})$）。红点表示对学习模型 (2) (25) 进行数值模拟的结果，取 20 次运行的平均值。实线表示猜想4.3给出的精确渐近刻画，详见附录D。（左）单层目标函数（$L^\star = 0$），（右）双层目标函数（$L^\star = 1, \sigma_1^\star = \mathrm{erf}$），带有（25）隐藏符号层。两者均采用深度为 $L = 2$ 的 dRF (2) 进行学习，激活函数为 σ_1,2(x) 双曲正切(2x)，正则化为 $\lambda = 0.05$（上）和 σ_1,2(x) = erf(x)，以及 $\lambda = 0.1$（下）。

的

$$\Psi_{\ell+1}^{\mathrm{lin}} = \left(\kappa_1^{\star(\ell+1)}\right)^2 \frac{W_{\ell+1}^\star \Psi_\ell^{\mathrm{lin}} W_{\ell+1}^{\star\top}}{k_\ell^\star} + \left(\kappa_*^{\star(\ell+1)}\right)^2 I_{k_{\ell+1}^\star} \tag{31}$$

在初始条件 $\Omega_0^{\mathrm{lin}} = \Psi_0^{\mathrm{lin}} := \Omega_0$ 下。进一步定义

$$\Phi_{L^\star L}^{\mathrm{lin}} = \left(\prod_{\ell=L^\star}^{1} \frac{\kappa_1^{\star\ell} W_\ell^\star}{\sqrt{k_\ell^\star}}\right) \cdot \Omega_0 \cdot \left(\prod_{\ell=1}^{L} \frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_\ell}}\right). \tag{32}$$

序列 $\{\kappa_1^{\star\ell}\kappa_*^{\star\ell}\}_{\ell=1}^{L^\star}$ 由（10）定义，其中 $\sigma_\ell^\star$，$\Delta_\ell^\star$。在特殊情形 $L^\star = 0$下，对应于单指标目标函数，$\Phi_{L^\star L}^{\mathrm{lin}}$ 中的第一个乘积应替换为 I_d。当 $L = 1$时，这种特定的目标架构也称为隐藏流形模型 [5, 52]，并为结构化数据提供了一种程式化模型。本文将这些研究推广到任意深度 L。由此可以提出以下更强的猜想：

猜想*4.3*。在与猜想 *4.2* 相同的极限下，使用协变量 $x^\mu \sim \mathcal{N}(0_d, \Omega_0)$ 和来自 *(25)* 的标签，在$\mathcal{D} = \{(x^\mu, y^\mu)\}_{\mu\in[n]}$ 上训练的经验风险最小化器 *(4)* 的测试误差，等于具有矩阵 $\Psi_{L^\star}^{\mathrm{lin}}, \Omega_L^{\mathrm{lin}}, \Phi_{L^\star L}^{\mathrm{lin}}$ *(21),(32)* 的高斯协变量模型 *(29)* 的测试误差。

猜想 4.3 使我们能够对测试误差进行完全解析的精确渐近刻画，详见附录 D。值得注意的是，它还为总体协方差提供了简洁的闭式公式Ω_L，$\Phi_{L^\star L}$，$\Psi_{L^\star}$。特别是，$\Psi_{L^\star}^{\mathrm{lin}}$、$\Omega_L^{\mathrm{lin}}$的谱可以通过解析方法计算，且与经验数值模拟的结果高度吻合。我们在附录 B中详细报告了这些结果。图 1 和 2 展示了由此得到的理论曲线，并将其与维度为 $d = 1000$时的数值模拟结果进行对比，二者吻合极佳。

5 深度诱导的隐式正则化

从猜想 4.3，尤其是闭式表达式（21）中，可以得到一个非正式但极具启发性的结论：深度非线性 dRF（2）中的激活，与具有以下各层的深度噪声线性网络中的激活具有相同的总体统计量

$$\varphi_\ell^{\mathrm{lin}}(\mathrm{x}) = \kappa_1^\ell \frac{W_\ell^\top \mathrm{x}}{\sqrt{k_{\ell-1}}} + \kappa_*^\ell \xi_\ell, \tag{33}$$

其中 $\xi_\ell \sim \mathcal{N}(0_{k_\ell}, I_{k_\ell})$ 是一个高斯噪声项。显然，(33) 得出的递推关系与 (21) 相同。这一观察来自相关工作 [12],，基本上表明，可以将使用 dRF (2) 进行学习的问题等效地视为使用线性含噪网络进行学习。事实上，猜想4.3 基本上指出，渐近测试误差取决于 dRF 与等效线性网络共有的最后一层激活的二阶统计量。最后值得强调的是，尽管学习器 dRF 在给定

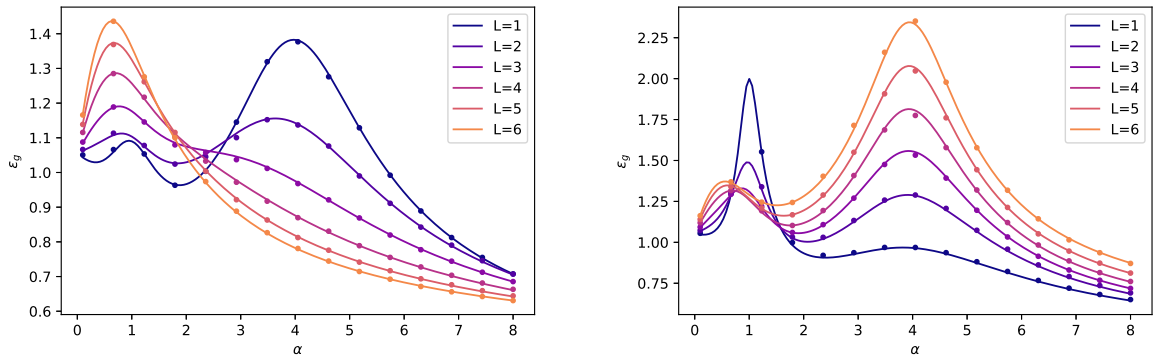

图3：使用宽度为 $\gamma_1 = \ldots = \gamma_L = 4$ 和$\sigma_1,\ldots,\sigma_L =$、采用tanh 激活函数的 L–隐藏层学习器（左），或采用$\sigma_1,\ldots,\sigma_L(x) = 1.1 \times$符号函数$(x) \times\min(2, |x|)$ 截断线性激活函数的学习器（右），对$\gamma_1^\star = 2$层目标函数（ $\sigma_1^\star =$ 符号函数）进行岭回归的学习曲线，深度为 $1 \le L \le 6$。正则化参数为 $\lambda = 0.001$。实线表示根据猜想4.3的精确刻画计算出的理论曲线，圆点表示对50次运行取平均后的数值模拟结果。在线性峰值可见于 $\alpha = 1$，而非线性峰值出现在 $\alpha = \gamma = 4$[53]。尽管架构相同，不同激活函数的使用会导致不同的隐式正则化；随着深度增加，截断线性激活函数下的线性峰值（分别是双曲正切激活函数下的非线性峰值）受到更强的抑制。

权重 $\{W_\ell\}$的条件下是确定性的，但等效线性网络 (33) 本质上具有随机性，这是由于每一层都会注入有效噪声 ξ_ℓ。统计学常识表明，这种有效噪声注入会通过在学习过程中引入一定的随机性而产生正则化效应，并有助于缓解过拟合。由于有效噪声是经过非线性层传播后产生的，这表明添加随机非线性层会产生隐式正则化。我们将在本节最后探讨这一观点。

首先注意，等效含噪线性神经网络（33）可简化为一个简单的浅层含噪线性模型

$$\hat{y}_\theta^{\text{lin}}(\mathrm{x}) = \sigma\left(\frac{1}{\sqrt{k}}\theta^\top (A_L \cdot \mathrm{x} + \xi_L)\right) \tag{34}$$

其中，有效权重矩阵 A为

$$A_L := \prod_{\ell=1}^{L}\left(\kappa_1^\ell \frac{W_\ell}{\sqrt{k_{\ell-1}}}\right)$$

且有效噪声 ξ_L 服从协方差为 C_ξ^L的高斯分布

$$C_\xi^L = \sum_{\ell_0=1}^{L-1}(\kappa_*^{\ell_0})^2\left(\prod_{\ell=\ell_0+1}^{L}\frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_{\ell-1}}}\right)^\top\left(\prod_{\ell=\ell_0+1}^{L}\frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_{\ell-1}}}\right) + (\kappa_*^L)^2 I_k.$$

等效线性特征（34）中的信号加噪声结构会对模型（2）学习曲线的水平产生深远影响：

- 当 $\alpha = 1$ 时，训练样本数与数据 $d-$ 维子流形 $A_L\mathbf{x}$ 的维度相同，从而出现标准的插值峰值。噪声部分 ξ_L 会产生隐式正则化，有助于缓解过拟合。

- 当 $\alpha = \gamma_L$ 时，训练样本数与噪声的维度 k_L 相匹配，噪声部分用于插值训练样本，从而产生另一个峰值。[53] 将这个第二个峰值称为非线性峰值。

因此，这两个峰值之间存在相互作用：较高的噪声 ξ_L 既有助于减弱线性峰值，也会加剧非线性峰值。神经网络的深度也发挥着作用，它通过递推关系（10）调节信号部分和噪声部分的幅度，具体取决于激活函数。

我们在图中给出深度的正则化效应的两个示例。 3。我们考察了两种激活函数： $\sigma_a =$ 双曲正切(其噪声水平以 $\operatorname{tr} C_\xi^L$ 衡量，并随深度增加而降低)，以及一种非线性非常弱的激活函数

$\sigma_b(x) = 1.1 \times$ 符号函数$(x) \times \min(2, |x|)$，对应于截断在 -2.2 和 2之间的线性函数。2 (其 $\operatorname{tr} C_\xi^L$随深度增加而增大)。请注意，对于 σ_a，有效噪声随深度增加而降低，因此更深的神经网络会加剧线性峰值，同时抑制非线性峰值。反之，对于 σ_b，额外的层会引入更多噪声并导致更高的非线性峰值，而由此产生的隐式正则化则会缓解线性峰值。关于架构设计对深度随机特征（2）泛化能力的影响，更多讨论见附录 E。

6 结论

我们研究了通过训练深度神经网络的读出层来学习深度随机网络目标函数的问题，其中隐藏层保持随机且固定（深度随机特征）。我们首先证明了深度高斯随机网络中激活的共轭核和样本协方差的渐近确定性等价物。借助这一结果，我们针对学习者与教师神经网络共享相同中间层、且通过岭损失学习读出层这一特定情形，建立了测试误差的精确渐近刻画。这证明了针对最后一层激活所对应的非线性特征进行岭回归时，测试误差具有高斯普适性。对于完全一般的情形，我们猜想，在目标函数和学习者架构完全一般且采用凸损失时，测试误差满足一个精确的渐近公式。这些公式表明，dRF 的行为类似于具有隐式正则化的线性含噪网络。我们探讨了这一等价性对 dRF 架构与其泛化能力之间相互作用的影响。

致谢

我们感谢 Gabriele Sicuro 在本项目期间与我们进行讨论。BL 感谢法国国家科学研究中心的*ChooseFrance- CNRS AIRising Talents* 计划提供支撑。DD 获得 ETH AI Center 博士奖学金的支撑。DS 获得 SNSF Ambizione Grant PZ00P2 209089 的支撑。HC 感谢欧洲研究委员会通过欧盟地平线 2020 研究与创新计划资助协议 714608-SMiLe 提供支撑。

参考文献

[1] Ali Rahimi 和 Benjamin Recht。用于大规模核机器的随机特征。载于 J. Platt、D. Koller、Y. Singer 和 S. Roweis 编，神经信息处理系统进展，第 20 卷。Curran Associates, Inc.，2007 年。[2] L´ena¨ıc Chizat、Edouard Oyallon 和 Francis Bach。关于可微编程中的惰性训练。载于 神经信息处理系统第 *33* 届国际会议论文集，美国纽约州雷德胡克，2019 年。Curran Associates Inc.[3] Jeffrey Pennington 和 Pratik Worah。用于深度学习的非线性随机矩阵理论。《统计力学：理论与实验》，2019(12)，2019 年。[4] Song Mei 和 Andrea Montanari。随机特征回归的泛化误差：精确渐近分析与双下降曲线。《纯粹与应用数学通讯》，75(4):667–766，2022 年。[5] Federica Gerace、Bruno Loureiro、Florent Krzakala、Marc Mezard 和 Lenka Zdeborova。使用随机特征和隐藏流形模型进行学习时的泛化误差。载于 Hal Daum´e III 和 Aarti Singh 编，第 *37* 届国际机器学习会议论文集，机器学习研究论文集第 119 卷，3452–3462 页。PMLR，2020年7月13日至18日。[6] Jaehoon Lee、Yasaman Bahri、Roman Novak、Samuel S. Schoenholz、Jeffrey Pennington 和 Jascha Sohl-Dickstein。作为高斯过程的深度神经网络。载于第六届国际学习表征会议（*ICLR 2018*）论文集，*2018* 年 *4* 月 *30* 日至 *5* 月 *3* 日，加拿大不列颠哥伦比亚省温哥华。OpenReview.net，2018 年。[7] Alexander G. De G. Matthews、Jiri Hron、Mark Rowland、Richard E. Turner 和 Zoubin Ghahramani。宽深度神经网络中的高斯过程行为。载于 国际学习表征会议，2018 年。[8] Zhou Fan 和 Zhichao Wang。线性宽度神经网络的共轭核与神经切线核的谱。载于 神经信息处理系统第 *34* 届国际会议论文集，NIPS'20，美国纽约州雷德胡克，2020 年。Curran Associates Inc.

[9] Jacob Zavatone-Veth 和 Cengiz Pehlevan。有限贝叶斯神经网络的精确边缘先验分布。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，第 34 卷，3364–3375 页。Curran Associates, Inc.，2021 年。[10] Lorenzo Noci、Gregor Bachmann、Kevin Roth、Sebastian Nowozin 和 Thomas Hofmann。深层 ReLU 网络先验预测分布的精确刻画。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，第 34 卷，20851–20862 页。Curran Associates, Inc.，2021 年。[11] Bruno Loureiro、Cedric Gerbelot、Hugo Cui、Sebastian Goldt、Florent Krzakala、Marc M´ezard 和 Lenka Zdeborov´a。使用教师-学生模型，在真实数据集上对通用特征映射进行学习曲线分析。《统计力学：理论与实验》，2022(11):论文编号 114001，共 78 页，2022 年。[12] Hugo Cui、Lenka Zdeborov´a 和 Florent Krzakala。广延宽度深度随机网络的最优学习，2023 年。私人通信。[13] Zhenyu Liao 和 Romain Couillet。关于高维数据随机特征映射的谱。载于 Jennifer Dy 和 Andreas Krause 编，《第 35 届国际机器学习会议论文集》，机器学习研究论文集第 80 卷，3063–3071 页。PMLR，2018年7月10日至15日。[14] Lucas Benigni 和 Sandrine P´ech´e。若干随机矩阵非线性模型的特征值分布。《电子概率学报》，26:论文编号 150，共 37 页，2021 年。[15] Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。随机特征与核方法的泛化误差：超收缩性与核矩阵集中性。《应用与计算调和分析》，59:3–84，2022 年。[16] Oussama Dhifallah 和 Yue M. Lu。随机特征学习的精确性能分析。arXiv:2008.11904，2020 年。[17] Tengyuan Liang 和 Pragya Sur。提升法与最小-ℓ_1-范数插值分类器的精确高维渐近理论。《统计年刊》，50(3):1669–1695，2022 年。.[18] David Bosch、Ashkan Panahi、Ayca Ozcelikkale 和 Devdatt Dubhash。随机特征模型中的双重下降：一般凸正则化的精确渐近分析。arXiv:2204.02678，2022 年。[19] Arthur Jacot、Berfin Simsek、Francesco Spadaro、Clement Hongler 和 Franck Gabriel。随机特征模型的隐式正则化。载于 Hal Daum´e III 和 Aarti Singh 编，《第 37 届国际机器学习会议论文集》，机器学习研究论文集第 119 卷，4631–4640 页。PMLR，2020年7月13日至18日。[20] Lucas Clart´e、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborov´a。过参数化高维模型中的不确定性量化研究。arXiv:2210.12760，2022 年。[21] St´ephane D'Ascoli、Maria Refinetti、Giulio Biroli 和 Florent Krzakala。双重下降中的双重难题：惰性区间中的偏差与方差。载于 Hal Daum´e III 和 Aarti Singh 编，《第 37 届国际机器学习会议论文集》，机器学习研究论文集第 119 卷，2280–2290 页。PMLR，2020年7月13日至18日。[22] Bruno Loureiro、Cedric Gerbelot、Maria Refinetti、Gabriele Sicuro 和 Florent Krzakala。波动、偏差、方差与学习器集成：高维凸损失的精确渐近分析。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，《第 39 届国际机器学习会议论文集》，机器学习研究论文集第 162 卷，14283–14314 页。PMLR，2022年7月17日至23日。[23] Antoine Bodin 和 Nicolas Macris。模型、样本与轮次层面的下降：随机特征模型中梯度流的精确解。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，第 34 卷，21605–21617 页。Curran Associates, Inc.，2021 年。[24] Blake Bordelon 和 Cengiz Pehlevan。随机梯度下降在结构化特征上的学习曲线。载于《国际学习表征会议》，2022 年。[25] Behrooz Ghorbani、Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。两层神经网络惰性训练的局限性。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，《神经信息处理系统进展》，第 32 卷。Curran Associates, Inc.，2019 年。

[26] Behrooz Ghorbani、Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。神经网络何时优于核方法? 《统计力学：理论与实验》，2021(12)：第124009号论文，110页，2021年。[27] Gilad Yehudai 和 Ohad Shamir。随机特征在理解神经网络方面的能力与局限性。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett（编），《神经信息处理系统进展》第32卷。Curran Associates, Inc.，2019年。[28] Maria Refinetti、Sebastian Goldt、Florent Krzakala 和 Lenka Zdeborova。对高维高斯混合进行分类：核方法失效而神经网络成功之处。载于 Marina Meila 和 Tong Zhang（编），《第38届国际机器学习会议论文集》，《机器学习研究论文集》第139卷，第8936–8947页。PMLR，2021年7月18日至24日。[29] Andre Manoel、Florent Krzakala、Marc M´ezard 和 Lenka Zdeborov´a。多层广义线性估计。载于《2017年IEEE国际信息论研讨会（ISIT）》，第2098–2102页，2017年。[30] Marylou Gabri´e、Andre Manoel、Cl´ement Luneau、jean barbier、Nicolas Macris、Florent Krzakala 和 Lenka Zdeborov´a。深度神经网络模型中的熵与互信息。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett（编），《神经信息处理系统进展》第31卷。Curran Associates, Inc.，2018年。[31] Benjamin Aubin、Bruno Loureiro、Antoine Maillard、Florent Krzakala 和 Lenka Zdeborov´a。带生成先验的尖峰矩阵模型。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett（编），《神经信息处理系统进展》第32卷。Curran Associates, Inc.，2019年。[32] Paul Hand、Oscar Leong 和 Vlad Voroninski。生成先验下的相位恢复。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett（编），《神经信息处理系统进展》第31卷。Curran Associates, Inc.，2018年。[33] Benjamin Aubin、Bruno Loureiro、Antoine Baker、Florent Krzakala 和 Lenka Zdeborov´a。采用随机生成先验进行相位恢复与压缩感知的精确渐近分析。载于 Jianfeng Lu 和 Rachel Ward（编），《第一届数学与科学机器学习会议论文集》，《机器学习研究论文集》第107卷，第55–73页。PMLR，2020年7月20日至24日。[34] Qianyi Li 和 Haim Sompolinsky。深度线性神经网络的统计力学：反向传播核重整化。《物理评论X》，11:031059，2021年。[35] S. Ariosto、R. Pacelli、M. Pastore、F. Ginelli、M. Gherardi 和 P. Rotondo。无限宽度极限之外的深度学习统计力学。arXiv:2209.04882，2022年。[36] Sebastian Goldt、Bruno Loureiro、Galen Reeves、Florent Krzakala、Marc Mezard 和 Lenka Zdeborov´a。浅层神经网络学习中生成模型的高斯等价性。载于《第二届数学与科学机器学习会议论文集》，《机器学习研究论文集》第145卷，第426–471页，2021年。[37] Hong Hu 和 Yue M. Lu。随机特征高维学习的普适性定律。《IEEE信息论汇刊》，2022年。[38] Andrea Montanari 和 Basil N. Saeed。经验风险最小化的普适性。载于 Po-Ling Loh 和 Maxim Raginsky（编），《第三十五届学习理论会议论文集》，《机器学习研究论文集》第178卷，第4310–4312页。PMLR，2022年7月2日至5日。[39] Hugo Cui、Luca Saglietti 和 Lenka Zdeborov'a。感知机模型的大偏差及其对主动学习的影响。《机器学习：科学与技术》，2:45001，2019年。[40] Hugo Cui、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborov´a。源条件与容量条件下核分类的误差率。arXiv，abs/2201.12655，2022年。[41] Z. Burda、A. G¨orlich、A. Jarosz 和 J. Jurkiewicz。相关矩阵中的信号与噪声。《物理学A辑》，343(1-4):295–310，2004年。[42] Antti Knowles 和 Jun Yin。随机矩阵的各向异性局部定律。《概率论及相关领域》，169(1-2):257–352，2017年。

[43] V A Marˇcenko 和 L A Pastur。若干随机矩阵集合的特征值分布。《苏联数学文集》，1(4):457–483，1967年。[44] Zhidong Bai 和 Wang Zhou。列间不存在独立性结构的大样本协方差矩阵。《统计新纪元》，18(2):425–442，2008年。[45] Cosme Louart 和 Romain Couillet。测度集中与大随机矩阵及其在样本协方差矩阵中的应用。arXiv:1805.08295，2018年。[46] Cl´ement Chouard。具有一般依赖结构的样本协方差矩阵的定量确定性等价物。arXiv:2211.13044，2022年。[47] John Wishart。来自正态多元总体的样本中的广义乘积矩分布。《生物计量学》，20A(1-2):32–52，1928年。[48] Florent Benaych-Georges。Marchenko–Pastur定律、矩形自由卷积与方形自由卷积之间一个令人惊讶的关系。《亨利·庞加莱研究所概率与统计年刊》，46(3):644–652，2010年。[49] Noureddine El Karoui。利用随机矩阵理论估计高维协方差矩阵的谱。《统计年刊》，36(6):2757–2790，2008年。[50] Edgar Dobriban 和 Stefan Wager。预测的高维渐近性：岭回归与分类。《统计年刊》，46(1):247–279，2018年。[51] Federica Gerace、Florent Krzakala、Bruno Loureiro、Ludovic Stephan 和 Lenka Zdeborov´a。高维中带随机标签的线性分类器的高斯普适性，2022年。[52] Sebastian Goldt、Marc M´ezard、Florent Krzakala 和 Lenka Zdeborov´a。对数据结构影响神经网络学习的建模：隐藏流形模型。《物理评论X》，10:041044，2020年。[53] St´ephane D’Ascoli、Levent Sagun 和 Giulio Biroli。三重下降与两类过拟合：它们何时何地出现，又为何出现？《统计力学：理论与实验》，2021(12)：第124002号论文，21页，2021年。[54] Radosław Adamczak。关于具有依赖关系的随机向量的Hanson–Wright不等式的注记。《电子概率通讯》，20，2015年。[55] Roman Vershynin。《高维概率》，剑桥统计与概率数学丛书第47卷。剑桥大学出版社，剑桥，2018年。一本附带数据科学应用的导论，序言作者为 Sara van de Geer。[56] St´ephane D’Ascoli、Marylou Gabri´e、Levent Sagun 和 Giulio Biroli。分类问题中数据结构与损失函数之间的相互作用。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），《神经信息处理系统进展》第34卷，第8506–8517页。Curran Associates, Inc.，2021年。[57] Hugo Cui、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborov´a。核回归的泛化误差率：从无噪声状态到有噪声状态的转变。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），《神经信息处理系统进展》第34卷，第10131–10143页。Curran Associates, Inc.，2021年。[58] Theodor Misiakiewicz。多项式状态下内积核矩阵的谱及核岭回归中的多重下降现象。arXiv:2204.10425，2022年。[59] Hong Hu 和 Yue M. Lu。线性状态之外核岭回归的尖锐渐近分析。arXiv:2205.06798，2022年。[60] Noureddine El Karoui、Derek Bean、Peter J. Bickel、Chinghway Lim 和 Bin Yu。关于高维预测变量下的稳健回归。《美国国家科学院院刊》，110(36):14557–14562，2013年。[61] Denny Wu 和 Ji Xu。过参数化线性回归中的最优加权正则化。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），《神经信息处理系统进展》第33卷，第10112–10123页。Curran Associates, Inc.，2020年。

[62] Trevor Hastie、Andrea Montanari、Saharon Rosset 和 Ryan J. Tibshirani。高维无岭最小二乘插值中的意外现象。《统计年刊》，50(2)：949–986，2022年。

[63] Alexander Wei、Wei Hu 和 Jacob Steinhardt。《不只是玩具：随机矩阵模型预测真实世界的神经表示如何泛化》。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），第*39*届国际机器学习会议论文集，第162卷，机器学习研究论文集，第23549–23588页。PMLR，2022年7月17日至23日。

A 各向异性确定性等价

A.1 样本协方差矩阵

考虑一个随机向量 $x \in \mathbb{R}^d$，其 $\mathbf{E}\, x = 0$ 和 $\mathbf{E}\, xx^\top = \Sigma$，对于 $n \in \mathbf{N}$，构造 $\mathcal{X} = (x_1, \dots, x_n) \in \mathbb{R}^{d\times n}$，使用 n 个 x 的独立副本 $x_1, \dots, x_n$。我们关注样本协方差和格拉姆矩阵

$$\widehat{\Sigma} := \frac{\mathcal{X}\mathcal{X}^\top}{n} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{d\times d} \qquad \text{and} \qquad \check{\Sigma} := \frac{\mathcal{X}^\top\mathcal{X}}{n} = \Big(\frac{x_i^\top x_j}{n}\Big)_{i,j=1}^n \in \mathbb{R}^{n\times n} \tag{35}$$

及其预解式

$$\widehat{G}(z) := (\widehat{\Sigma} - z)^{-1} \in \mathbb{C}^{d\times d} \qquad \text{and} \qquad \check{G}(z) := (\check{\Sigma} - z)^{-1} \in \mathbb{C}^{n\times n}. \tag{36}$$

样本协方差矩阵和格拉姆矩阵的期望为

$$\mathbf{E}\,\widehat{\Sigma} = \Sigma, \quad \mathbf{E}\,\check{\Sigma} = \frac{d}{n}\langle\Sigma\rangle I_n, \tag{37}$$

其中我们引入了平均迹 $\langle A\rangle := m^{-1}\operatorname{Tr} A$，适用于 $A \in \mathbb{R}^{m\times m}$。

请注意，虽然两个预解式作为矩阵的表现不同，但由于$\widehat{\ }$ $\check{\ }$非零特征值 Σ 和 Σ 相同，它们的迹之间存在关联，因此

$$\langle\widehat{G}(z)\rangle = \frac{n}{d}\langle\check{G}(z)\rangle + \frac{n-d}{pz}. \tag{38}$$

关于样本协方差预解式和 Gram 矩阵预解式的归一化迹，经典结果是在基本达到最优的条件下成立的以下方差估计。

定理 A.1 (一般总体下样本协方差矩阵的迹收敛 [44])。 假设 $\|\Sigma\| \lesssim 1$，$d/n \sim 1$ 并且

$$\mathbf{E}\left|\frac{x^\top Ax}{d} - \mathbf{E}\,\frac{x^\top Ax}{d}\right|^2 = \mathbf{E}\left|\frac{x^\top Ax}{d} - \langle\Sigma A\rangle\right|^2 = o(\|A\|) \tag{39}$$

对所有确定性矩阵 A 均成立。那么有

$$\mathbf{E}\left|\langle(\widehat{\Sigma} - z)^{-1}\rangle - \widehat{m}(z)\right|^2 = o(1), \quad \mathbf{E}\left|\langle(\check{\Sigma} - z)^{-1}\rangle - \check{m}(z)\right|^2 = o(1), \quad \textit{as} \quad n, d \to \infty, \tag{40}$$

对所有固定的 $z \in \mathbb{C}\setminus\mathbb{R}_+$ 均成立，其中 $\check{m} = \check{m}(z)$ 是标量方程的唯一解

$$1 - \frac{d}{n} + z\check{m} = -\frac{d}{n}\langle(\Sigma\check{m} + 1)^{-1}\rangle. \tag{41}$$

and

$$\widehat{m}(z) := \frac{n}{d}\check{m}(z) + \frac{d-n}{d}\frac{1}{-z} \tag{42}$$

这里，$\widehat{m}$ 是 Marchenko-Pastur 方程（6）的解，相应的测度是经验谱测度 $\mu(\Sigma)$ 与以下分布的自由乘法卷积：$= d^{-1}\sum_{\lambda\in\operatorname{Spec}(\Sigma)}\delta_\lambda$ 属于 Σ，以及长宽比为 μ_{MP}^c 的 Marchenko-Pastur 分布 $c = d/n$。因此，通过 Stieltjes 反演，定理 A.1 的结果可表述为

$$\mu\Big(\frac{\mathcal{X}^\top\mathcal{X}}{n}\Big) = \mu(\check{\Sigma}) \approx \frac{d}{n}\mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^{d/n} + \frac{n-d}{n}\delta_0, \quad \mu\Big(\frac{\mathcal{X}\mathcal{X}^\top}{n}\Big) = \mu(\widehat{\Sigma}) \approx \mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^{d/n} \tag{43}$$

在弱意义和整体意义下。注意，我们有以下极限

$$\lim_{c\to\infty} \mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^c = \delta_0, \quad \lim_{c\to 0} \mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^c = \mu(\Sigma) \tag{44}$$

这正是预期行为，因为当 $c = d/n$ 较大时，$\mathcal{X}\mathcal{X}^\top$ 的秩 n 增长得远比 d 慢，因此经验测度 $\mu(\Sigma)$集中在原点；而当 $c = d/n$ 较小时，根据大数定律 $\mathcal{X}\mathcal{X}^\top/n \approx \mathbf{E}\,\mathcal{X}\mathcal{X}^\top/n = \Sigma$。

A.2 各向异性确定性等价物

$\widehat{\ }$ $\check{\ }$定理 A.1 中的迹意义结果只能控制 Σ的特征值， Σ 却无法控制特征向量。关于 $\widehat{\ }$ $\check{\ }$的非迹确定性等价物已有大量研究， Σ， Σ形式包括逐项渐近$\check{G}_{ij} \approx \cdots$、针对确定性向量 $x^\top \widehat{G} y \approx \cdots$的各向同性渐近 x， y 以及针对确定性矩阵 $\langle A\widehat{G}\rangle \approx \cdots\widehat{\ }$ $\check{\ }$的函数型迹渐近 A。这些结果都包含关于 G、G 在渐近极限下作为矩阵的行为的非平凡信息，并可用于推断有关特征向量的信息。

对于可分离相关性，文献 [42] 已得到了各向同性形式和迹意义下的最优局部定律：

定理 A.2 ([42], 定理 3.6)。 若 $\mathcal{X} = \Sigma^{1/2} X$ 对某个矩阵 X 满足其元素独立同分布[1]， 均值为 0， 方差为 1，且谱密度 $\mu(\Sigma) \boxtimes \mu^{d/n}{}_{\mathrm{MP}}$ 是正则的[2],， 则有

$$\left|\langle(\widehat{\Sigma} - z)^{-1}\rangle - \widehat{m}(z)\right| + \left|\langle(\check{\Sigma} - z)^{-1}\rangle - \check{m}(z)\right| \prec \frac{1}{n \operatorname{Im} z}, \tag{45}$$

在迹意义下成立；对于任意确定性向量 x、 y

$$\left|x^\top\left[(\widehat{\Sigma} - z)^{-1} - (-\Sigma\check{m}(z)z - z)^{-1}\right]y\right| + \left|x^\top(\check{\Sigma} - z)^{-1}y - \check{m}(z)x^\top y\right| \prec \frac{\|x\|\|y\|}{\sqrt{n \operatorname{Im} z}} \tag{46}$$

在各向同性意义下成立。

$\widehat{\ }$请注意，特别是，矩阵 $G(z)$ 渐近等于一个预解式

$$\widehat{M}(z) := \left(-\Sigma\check{m}(z)z - z\right)^{-1} \tag{47}$$

$\check{\ }$总体协方差 Σ 的，而 G 渐近上是单位矩阵的一个标量倍数。
最近，文献 [46] 中针对 G 得到了函数型迹局部定律（尽管其对谱参数的依赖远非最优）： $\widehat{\ }$

定理 A.3 ([46], 命题 2.4). 若 $\|\Sigma\| \le C$ 且 $\mathcal{X}$ 满足：存在正常数 c、 C、 σ

$$P(|f(\mathcal{X}) - \mathbf{E}\, f(\mathcal{X})| \ge t) \le Ce^{-c(t/\sigma)^2} \quad \forall \text{ 1-Lipschitz } f : (\mathbb{R}^{d\times n}, \|\cdot\|_F) \to (\mathbb{R}, |\cdot|), \tag{48}$$

则对于任意确定性矩阵 A 和 $|z| \lesssim 1$， 以高概率有[3]。

$$\left|\langle A(\widehat{\Sigma} - z)^{-1} - A(-\check{m}(z)z\Sigma - z)^{-1}\rangle\right| \le \frac{\sqrt{\langle AA^*\rangle \log n}}{n \operatorname{dist}(z, \mathbb{R}_+)^9}, \tag{49}$$

其中， $\check{m} = \check{m}(z)$是标量方程的唯一解

$$1 - \frac{d}{n} + z\check{m} = -\frac{d}{n}\langle(\Sigma\check{m} + 1)^{-1}\rangle. \tag{50}$$

请注意，收敛速率为 $1/n$、误差以A 的弗罗贝尼乌斯范数表示的函数型迹意义表述，会自动包含各向同性局部定律这一特例。事实上，对于 $A = xy^\top$， 可得

$$y^\top\left((\widehat{\Sigma} - z)^{-1} - (-\check{m}z\Sigma - z)^{-1}\right)x \prec \frac{\|x\|\|y\|}{\sqrt{n}\delta^9}, \tag{51}$$

此处及下文中，我们记 $\delta \equiv \delta(z) := \operatorname{dist}(z,\ \mathbb{R}_+)$。本文将函数型迹局部定律$\check{\ }$从 [46] 推广到 G的情形，并得到以下结果：

命题 A.4（格拉姆矩阵的函数型局部定律）。 在定理 A.3 的假设下，我们有 t

$$\left|\langle A(\check{\Sigma} - z)^{-1}\rangle - \check{m}(z)\langle A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}. \tag{52}$$

[1]具有所有阶的有限矩[2]参见 [42][3]中的定义 2.7。 [46] 中的陈述字面上给出的是 $\operatorname{Im}\ z$， 而非 $\operatorname{dist}(z, \mathbb{R}_+)$；但证明逐字来看给出了更强的界，因为 $\operatorname{Im} z$ 仅被用作某类矩阵 $AA^* - z$的最小奇异值的下界。

请注意，命题A.4 中的界弱于定理 A.3中的界，而且这两个结果在对谱参数的依赖方面都远弱于定理 A.2 。鉴于相关结果，自然可以提出以下猜想：

猜想 A.5. 假设 x 的二次型集中于

$$\left|\frac{x^\top Ax}{d} - \langle \Sigma A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\sqrt{d}} \tag{53}$$

对任意确定性矩阵 A，且 $\|\Sigma\| \lesssim 1$。那么我们有如下函数迹估计

$$\begin{aligned}\left|\langle zA(\widehat{\Sigma}-z)^{-1}\rangle - \langle A(-\breve{m}(z)\Sigma - I)^{-1}\right| &\prec \frac{\langle AA^*\rangle^{1/2}}{n\delta}\\ \left|\langle A(\breve{\Sigma}-z)^{-1}\rangle - \breve{m}(z)\langle A\rangle\right| &\prec \frac{\langle AA^*\rangle^{1/2}}{n\delta}.\end{aligned} \tag{54}$$

请注意，定理 A.3 所要求的利普希茨集中远强于猜想 A.5 中的二次型集中，因为它意味着列向量 x满足 $\mathcal{X}$

$$P(|f(x) - \mathbf{E}\, f(x)| \geq t) \leq C\exp\Big(-\frac{t^2}{C\lambda_f^2}\Big) \tag{55}$$

对所有 λ_f-利普希茨 f： $\mathbb{R}^d \to \mathbb{R}$。因此，根据 Hanson-Wright（[54], 定理 2.4）

$$P\Big(\left|\frac{x^\top Ax}{d} - \langle \Sigma A\rangle\right| \geq \frac{t\langle AA^*\rangle^{1/2}}{\sqrt{d}} + \frac{t\|A\|}{d}\Big) \leq Ce^{-\min\{t^2,t\}/C} \tag{56}$$

并且，由于还有 $\|A\| \leq \sqrt{d}\langle AA^*\rangle^{1/2}$，我们以高概率有

$$\left|\frac{x^\top Ax}{d} - \langle \Sigma A\rangle\right| \leq \log d\Big(\frac{\langle AA^*\rangle^{1/2}}{\sqrt{d}} + \frac{\|A\|}{d}\Big) \lesssim \log d\frac{\langle AA^*\rangle^{1/2}}{\sqrt{d}}. \tag{57}$$

现在我们转向命题A.4的证明。我们需要以下关于预解函数利普希茨性的结果，参见例如 [46]

引理 A. 6. 映射 $\breve{G}\colon \mathcal{X} \to \big(\mathcal{X}^\top\mathcal{X}/n - z\big)^{-1}$ 关于 *Frobenius* 范数是 $(3\delta^{-2}|z|^{1/2}n^{-1/2})$-*Lipschitz* 的，且 *m.*

命题 *A.4*的证明。 记 $\breve{m} \equiv \breve{m}(z)$。由舒尔补公式可得

$$\breve{G}_{ii} = -\Big(z + z\frac{x_i^\top \hat{G}^{(i)}x_i}{n}\Big)^{-1} = -\Big(z + zc\langle \Sigma\hat{G}^{(i)}\rangle\Big)^{-1} + O\Big(\frac{1}{\sqrt{n}\delta^9}\Big) = \breve{m} + O\Big(\frac{1}{\sqrt{n}\delta^9}\Big), \tag{58}$$

利用

$$\langle \Sigma\hat{G}^{(i)}\rangle = \langle \Sigma\breve{G}\rangle + \frac{1}{n}\left\langle \Sigma\frac{\hat{G}^{(i)}x_ix_i^\top\hat{G}^{(i)}}{1 + x_i\hat{G}^{(i)}x_i/n}\right\rangle = -\frac{1}{z}\langle \Sigma(\breve{m}\Sigma + I)^{-1}\rangle + O\Big(\frac{1}{n\delta^9}\Big) \tag{59}$$

and

$$z - c\langle \Sigma(\breve{m}\Sigma + I)^{-1}\rangle = z - \frac{c}{\breve{m}} + \frac{c}{\breve{m}^2}\langle(\breve{m}\Sigma + I)^{-1}\rangle = z - \frac{c}{\breve{m}} - \frac{1}{\breve{m}}\Big(1 - c + z\breve{m}\Big) = -\frac{1}{\breve{m}} \tag{60}$$

在最后一步中。接下来，对于非对角元素，同样由舒尔补可得

$$\breve{G}_{ij} = z\breve{G}_{ii}\breve{G}^{(i)}_{jj}\frac{x_i^\top\hat{G}^{(ij)}x_j}{n} = z\breve{G}_{ii}\Big(\breve{G}_{jj} - \frac{\breve{G}_{ij}\breve{G}_{ji}}{\breve{G}_{jj}}\Big)\frac{x_i^\top\hat{G}^{(ij)}x_j}{n}. \tag{61}$$

这里由第一　　t 等式已经给出了一个界的大小 $n^{-1/2}\delta^{-4}$ 由此可得。因此，结合公式（58），可得

$$\breve{G}_{ij} = \breve{m}^2 z\frac{x_i^\top\hat{G}^{(ij)}x_j}{n} + O\Big(\frac{1}{n\delta^8}\Big), \tag{62}$$

因此　　由零均值假设可知 $\mathbf{E}\ \breve{G}_{ij} = O(1/n\delta^8)$。结合方程（58），可得　　t

$$\Big\|\mathbf{E}\,\breve{G} - \breve{m}(z)I\Big\|_{\mathrm{F}} = O\Big(\frac{1}{\delta^9}\Big). \tag{63}$$

记作

$$\left|\langle A\check{G}\rangle - \check{m}(z)\langle A\rangle\right| \le \left|\langle A\check{G}\rangle - \mathbf{E}\langle A\check{G}\rangle\right| + \left|\mathbf{E}\langle A\check{G}\rangle - \check{m}(z)\langle A\rangle\right|. \tag{64}$$

注意，由引理 A.6 和柯西-施瓦茨不等式可知，

$$\text{the map } \mathcal{X} \to \left\langle A\left(\frac{\mathcal{X}^\top\mathcal{X}}{n} - z\right)\right\rangle \quad \text{is} \quad \frac{3|z|^{1/2}\langle AA^*\rangle^{1/2}}{n\delta^2}\text{-Lipschitz}, \tag{65}$$

因此，

$$\left|\langle A\check{G}\rangle - \mathbf{E}\langle A\check{G}\rangle\right| \prec \frac{|z|^{1/2}\langle AA^*\rangle^{1/2}}{n\delta^2} \tag{66}$$

此外，由（63），可得

$$\left|\mathbf{E}\langle A\check{G}\rangle - \check{m}(z)\langle A\rangle\right| \le \frac{1}{\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|\mathbf{E}\,\check{G} - \check{m}(z)I\right\|_F \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}. \tag{67}$$

命题的结论由（64）、（66）和（67）得出。 □

A.3 随机特征模型

我们考虑一个单层随机特征模型，并对每个元素应用标量函数 $\sigma_1(x)$。

$$\sigma_1\Big(\frac{W_1X_0}{\sqrt{d}}\Big), \quad X_0 \in \mathbb{R}^{d\times n}, \quad W_1 \in \mathbb{R}^{k_1\times d}. \tag{68}$$

我们需要以下假设。

假设 A.7 (高斯权重)。 W_1 的各项独立同分布。 $\mathcal{N}(0,1)$ 元素。

假设 A.8 (正交有界数据)。 对于正常数 r_1， X_0 满足

$$\left\|\frac{X_0^\top X_0}{d} - r_1 I\right\|_{\max} \prec \frac{1}{\sqrt{n}}, \quad \left\|\frac{X_0}{\sqrt{d}}\right\|_{op} \prec 1. \tag{69}$$

Assu设 A.9 (非线性函数)。 标量函数 σ_1 是 λ_σ-*Lipschitz* 的，并且满足 $\langle\sigma_1\rangle_{\mathcal{N}(r_1)} = 0$，其中 *re*

$$\langle f\rangle_{\mathcal{N}(\sigma^2)} := \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} f(x)\exp\Big(-\frac{x^2}{2\sigma^2}\Big)\,\mathrm{d}x. \tag{70}$$

假设 A.10 (比例区间)。 对于某些常数 c_1， c_2，

$$c_1 n \le \min\{d, k_1\} \le \max\{d, k_1\} \le c_2 n, \quad 0 < c_1 < c_2 < \infty. \tag{71}$$

为简单起见，我们将权重矩阵的方差设为 1，尽管通过对函数 σ_1进行缩放，结果可以很容易地推广到任意方差 Δ。令 $\tilde{w}_i$ 表示 W_1的第 i行。我们定义

$$X_1 := \sigma_1\Big(\frac{W_1X_0}{\sqrt{d}}\Big) = \left(\sigma_1\Big(\frac{X^\top\tilde{w}_1}{\sqrt{d}}\Big)\cdots\sigma_1\Big(\frac{X^\top\tilde{w}_{k_1}}{\sqrt{d}}\Big)\right)^\top \in \mathbb{R}^{k_1\times n} \tag{72}$$

as a 具有独立同分布的行及其对应样本协方差矩阵的矩阵

$$\widehat{\Sigma} := \frac{X_1^\top X_1}{k_1} = \frac{1}{k_1}\sigma_1\Big(\frac{X_0^\top W_1^\top}{\sqrt{d}}\Big)\sigma_1\Big(\frac{W_1X_0}{\sqrt{d}}\Big). \tag{73}$$

对于 $\xi_{ij} := \tilde{w}_i^\top x_j \sim \mathcal{N}\left(0, \|x_j\|^2/d\right)$，我们有 $(X_1)_{ij} = \sigma_1(\xi_{ij})$，其中 x_j是 X_0的第 j列。为了分析高斯变量的函数，我们使用以下分解。

引理 A.11 (厄米特分解)。 对于任意利普希茨连续的 f 和任意 $\sigma > 0$，我们有 σ-厄米特展开[4]，

$$f(x) = \sum_{k\geq 0} \frac{\sigma^k}{k!} \mathrm{He}_k\left(\frac{x}{\sigma}\right) \langle f^{(k)} \rangle_{\mathcal{N}(\sigma^2)} \tag{74}$$

其中

$$\mathrm{He}_k(x) := (-1)^k \exp\left(\frac{x^2}{2}\right) \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp\left(-\frac{x^2}{2}\right) \tag{75}$$

w第 i 个 $\mathrm{He}_k(x)$ 是标准厄米特多项式 $\mathrm{He}_0(x) = 1$、$\mathrm{He}_1(x) = x$、$\mathrm{He}_2(x) = x^2 - 1$ 等 .

注意，厄米特多项式关于高斯密度两两正交。更准确地说，

$$\mathbf{E}\, \mathrm{He}_k(N_1)\, \mathrm{He}_j(N_2) = \delta_{jk} k!\, \mathrm{Cov}(N_1, N_2)^k \tag{76}$$

对于联合高斯 N_1, N_2 其中 $\mathbf{E}\, N_1 = \mathbf{E}\, N_2 = 0$ 且 $\mathbf{E}\, N_1^2 = \mathbf{E}\, N_2^2 = 1$。通过两次应用 (74)，并利用 (76)，我们得到帕塞瓦尔恒等式

$$\langle f^2 \rangle_{\mathcal{N}(\sigma)} = \sum_{k\geq 0} \frac{\sigma^{2k}}{k!} \langle f^{(k)} \rangle^2_{\mathcal{N}(\sigma)}. \tag{77}$$

在深度随机特征模型的确定性等价物证明中，我们依赖于 [45, 46] 中提出的技术，这些技术利用测度集中理论分析随机矩阵。这种方法特别适用于常见的神经网络架构，因为可以将从一层到另一层的变换视为利普希茨映射。下面的引理证明了所需函数的利普希茨性。

引理 A.12. 设 $f(x)$ 是一个 λ-利普希茨函数。设 $x, y, w \in \mathbb{R}^d$, $W \in \mathbb{R}^{k\times d}$ 和 $X \in \mathbb{R}^{d\times n}$。假设逐元素应用 $f(x)$，则以下映射均为利普希茨映射：

$$w \to f\left(\frac{x^\top w}{\sqrt{d}}\right) \qquad \textit{and} \qquad W \to f\left(\frac{WX}{\sqrt{d}}\right), \tag{78}$$

其利普希茨常数分别为 $\lambda\left\|x/\sqrt{d}\right\|$ 和 $\lambda\left\|X/\sqrt{d}\right\|$。此外，在事件 $Q := \{|f(x^\top w/\sqrt{d})| \lesssim 1 \wedge |f(y^\top v/\sqrt{d})| \lesssim 1|\}$ 下，映射

$$w \to f\left(\frac{x^\top w}{\sqrt{d}}\right) f\left(\frac{y^\top w}{\sqrt{d}}\right) \tag{79}$$

也满足利普希茨条件，对应的常数为 $\alpha \lesssim \lambda\left(\left\|x/\sqrt{d}\right\| + \left\|y/\sqrt{d}\right\|\right)$。

证明。 第一个和第二个映射的利普希茨性质直接由柯西-施瓦茨不等式得出。对于第三个映射，由于利普希茨函数的乘积不一定是利普希茨函数，因此需要限定在“良好”事件 Q 上。为简便起见，记 $f(a,b) := f\left(a^\top b/\sqrt{d}\right)$。在 Q 下，对于某些向量 $u, v \in \mathbb{R}^d$，我们可以写成，

$$\begin{aligned}
&|f(x,w)f(y,w) - f(x,v)f(y,v)| \\
&\quad \leq |f(x,w)f(y,w) - f(x,w)f(y,v)| + |f(x,w)f(y,v) - f(x,v)f(y,v)| \\
&\quad = |f(x,w)||f(y,w) - f(y,v)| + |f(y,v)||f(x,w) - f(x,v)| \\
&\quad \lesssim \lambda\left(\left\|x/\sqrt{d}\right\| + \left\|y/\sqrt{d}\right\|\right)\|w - v\|.
\end{aligned} \tag{80}$$

□

回顾一下记号

$$\begin{aligned}
r_2 &:= \langle \sigma_1^2 \rangle_{\mathcal{N}(r_1)} \\
\kappa_1^1 &:= \langle \sigma_1' \rangle_{\mathcal{N}(r_1)} \\
\kappa_*^1 &:= \sqrt{\langle \sigma_1^2 \rangle_{\mathcal{N}(r_1)} - r_1(\kappa_1^1)^2}
\end{aligned} \tag{81}$$

用于命题 3.1的证明。我们陈述一些技术性引理。

[4]注意，尽管其 表面上涉及导数，但光滑性并非必要，因为通过分部积分，导数可以转移到光滑的高斯权重 t.

引理 A.13。 对于 $w \sim \mathcal{N}(0,\ I)$，$\lambda_\sigma$-利普希茨函数 $\sigma(x)$ 以及 $\left\|x/\sqrt{d}\right\| \lesssim 1$，以高概率

$$\left|\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\right| \lesssim 1. \tag{82}$$

证明。由于映射 $w \to \sigma\left(\frac{x^\top w}{\sqrt{d}}\right)$ 是 $\left\|x/\sqrt{d}\right\|\lambda_\sigma$-Lipschitz 的，根据高斯集中定理（例如参见 [55] 中的定理 5.2.2），有

$$P\left(\left|\sigma\left(\frac{x^\top w}{\sqrt{d}}\right) - \mathbf{E}_w\,\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\right| \geq t\right) \leq e^{-\frac{t^2}{2\|x/\sqrt{d}\|^2\lambda_\sigma^2}}. \tag{83}$$

接下来，根据方程（90），对每个 $i \in [n]$,

$$\mathbf{E}_w\,\sigma\left(\frac{x^\top w}{\sqrt{d}}\right) = \langle\sigma\rangle_{\mathcal{N}(\|x\|^2/d)} = O(1/\sqrt{n}), \tag{84}$$

这意味着，以高概率,

$$\left|\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\right| \lesssim 1. \tag{85}$$

□

引理 A.14。对于 $w \sim \mathcal{N}(0, I)$，随机变量 $\sigma\left(\frac{x^\top w}{\sqrt{d}}\right)\sigma\left(\frac{w^\top y}{\sqrt{d}}\right)$ 以高概率为次高斯。其次高斯范数为 $O(\lambda_\sigma(\left\|x/\sqrt{d}\right\| + \left\|y/\sqrt{d}\right\|))$

证明。由引理A.13、引理A.12 和高斯集中定理可得。 □

引理 A.15。对于矩阵 A、$B \in \mathbb{R}^{n\times n}$，有

1. $\|AB\|_F \leq \|A\|\|B\|_F$，2. $\mathrm{Tr}(AB) \leq \|A\|_F\|B\|_F$，3. $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$，若 A 和 B 可逆。

引理 A.16。对于任意半正定矩阵 Y 和任意 $z \in \mathbb{C} \setminus \mathbb{R}_+$，有

$$\left\|(Y - z)^{-1}\right\| \leq \mathrm{dist}(z, \mathbb{R}_+)^{-1}. \tag{86}$$

命题 *3.1* 的证明。定义总体协方差矩阵

$$\Sigma_X := \mathbf{E}_w\,\sigma\Big(\frac{X_0^\top w}{\sqrt{d}}\Big)\sigma\Big(\frac{w^\top X_0}{\sqrt{d}}\Big) \in \mathbb{R}^{n\times n}, \quad w \sim \mathcal{N}(0, I). \tag{87}$$

利用厄米特级数展开（74）和（76），对于固定的 X_0，我们可以写出显式形式

$$\Sigma_X = \sum_{a\geq 0}\frac{1}{a!}D_X^{(a)}\Big(\frac{X_0^\top X_0}{d}\Big)^{\odot a}D_X^{(a)}, \tag{88}$$

其中我们定义了对角矩阵

$$D_X^{(a)} := \mathrm{diag}\Big(\langle\sigma^{(a)}\rangle_{\mathcal{N}(\|x_1\|^2/d)}, \ldots, \langle\sigma^{(a)}\rangle_{\mathcal{N}(\|x_n\|^2/d)}\Big). \tag{89}$$

由假设A.8 和标准扰动分析可得

$$\langle\sigma\rangle_{\mathcal{N}(\|x_i\|^2/d)} = \langle\sigma\rangle_{\mathcal{N}(r_1)} + O\Big(\frac{1}{\sqrt{n}}\Big) = O\Big(\frac{1}{\sqrt{n}}\Big) \tag{90}$$

and

$$\langle\sigma'\rangle_{\mathcal{N}(\|x_i\|^2/d)} = \langle\sigma'\rangle_{\mathcal{N}(r_1)} + O\Big(\frac{1}{\sqrt{n}}\Big). \tag{91}$$

因此，我们可以得出，对于非对角线 $i \neq j$,

$$\begin{aligned}(\Sigma_X)_{ij} &= \sum_{a\geq 0} \frac{\langle \sigma^{(a)}\rangle_{\mathcal{N}(\|x_i\|^2/d)}\langle \sigma^{(a)}\rangle_{\mathcal{N}(\|x_j\|^2/d)}}{a!}\left(\frac{x_i^\top x_j}{d}\right)^a \\ &= \langle \sigma'\rangle^2_{\mathcal{N}(r_1)}\frac{x_i^\top x_j}{d} + O\left(\frac{1}{n}\right) = (\Sigma_{\rm lin})_{ij} + O\left(\frac{1}{n}\right)\end{aligned} \tag{92}$$

对于对角线元素，我们可以直接由（87）写出，

$$\begin{aligned}(\Sigma_X)_{ii} &= \langle \sigma^2\rangle_{\mathcal{N}(\|x_i\|^2/d)} = \langle \sigma'\rangle^2_{\mathcal{N}(r_1)}\frac{\|x_i\|^2}{d} + \left(\langle \sigma^2\rangle_{\mathcal{N}(r_1)} - r_1\langle \sigma'\rangle^2_{\mathcal{N}(r_1)}\right) + O\left(\frac{1}{\sqrt{n}}\right) \\ &= (\Sigma_{\rm lin})_{ii} + O\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{93}$$

对所有指标求和 i, j 可得

$$\|\Sigma_X - \Sigma_{\rm lin}\|_{\rm F} = O(1). \tag{94}$$

令 $\breve{m}(\Sigma,\ z)$ 为下列方程的解:

$$\breve{m} = \frac{d-n}{nz} - \frac{d}{zn}\langle(\Sigma\breve{m}+1)^{-1}\rangle, \tag{95}$$

以及 $\breve{m}_X := \breve{m}(\ \Sigma_X,\ z)$ $\breve{m}_{\rm lin} := \breve{m}(\Sigma_{\rm lin},\ z)$。考虑以下近似序列（在函数迹意义下 e):

$$\left(\frac{X_1^\top X_1}{k_1} - z\right)^{-1} \approx \left(-\breve{m}_X z\Sigma_X - z\right)^{-1} \approx \left(-\breve{m}_{\rm lin} z\Sigma_X - z\right)^{-1} \approx \left(-\breve{m}_{\rm lin} z\Sigma_{\rm lin} - z\right)^{-1}. \tag{96}$$

第一个近似由定理 A.3 应用于矩阵 $\mathcal{X} = X_1^\top$ 得出。矩阵 $\mathcal{X}$ 根据引理 A.12 和高斯集中定理而集中。

第二种　近似需要证明函数 $\breve{m}(\Sigma, z)$ 的稳定性。具体而言，我们写作 te

$$\begin{aligned}&\left|\left\langle A\left[\left(-\breve{m}_X z\Sigma_X - z\right)^{-1} - \left(-\breve{m}_{\rm lin} z\Sigma_X - z\right)^{-1}\right]\right\rangle\right| \\ &= \left|\left\langle A\left[\left(-\breve{m}_X z\Sigma_X - z\right)^{-1}(z(\breve{m}_X - \breve{m}_{\rm lin})\Sigma_X)\left(-\breve{m}_{\rm lin} z\Sigma_X - z\right)^{-1}\right]\right\rangle\right| \\ &\leq \frac{|\breve{m}_X - \breve{m}_{\rm lin}|}{|z|^2\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|\left(\breve{m}_X\Sigma_X + I\right)^{-1}\right\|\left\|\left(\breve{m}_{\rm lin}\Sigma_X + I\right)^{-1}\right\|\|\Sigma_X\|_F \\ &\leq |z|^{-2}|\breve{m}_X - \breve{m}_{\rm lin}|\langle AA^*\rangle^{1/2}.\end{aligned} \tag{97}$$

现在，我们分析 $\breve{m}_X$ 与 $\breve{m}_{\rm lin}$之间的差异。根据（95），我们可以写出

$$\begin{aligned}\Delta := |\breve{m}_X - \breve{m}_{\rm lin}| &= \frac{d}{|z|n^2}\operatorname{Tr}\left[(\breve{m}_X\Sigma_X + I)^{-1} - (\breve{m}_{\rm lin}\Sigma_{\rm lin} + I)^{-1}\right] \\ &\lesssim \frac{1}{|z|n}\operatorname{Tr}\left[(\breve{m}_X\Sigma_X + I)^{-1}(\breve{m}_{\rm lin}\Sigma_{\rm lin} - \breve{m}_X\Sigma_X)(\breve{m}_X\Sigma_{\rm lin} + I)^{-1}\right] \\ &\leq \frac{1}{|z|n}\left\|(\breve{m}_X\Sigma_X + I)^{-1}\right\|_F\|\breve{m}_{\rm lin}\Sigma_{\rm lin} - \breve{m}_X\Sigma_X\|_F\left\|(\breve{m}_{\rm lin}\Sigma_{\rm lin} + I)^{-1}\right\| \\ &\leq \frac{1}{|z|\sqrt{n}}\|\breve{m}_{\rm lin}\Sigma_{\rm lin} - \breve{m}_X\Sigma_X\|_F \\ &\leq \frac{1}{|z|\sqrt{n}}\|\breve{m}_{\rm lin}\Sigma_{\rm lin} - \breve{m}_{\rm lin}\Sigma_X\|_F + \frac{d}{|z|n^{3/2}}\|\breve{m}_{\rm lin}\Sigma_X - \breve{m}_X\Sigma_X\|_F \\ &= \frac{|\breve{m}_{\rm lin}|}{|z|\sqrt{n}}\|\Sigma_{\rm lin} - \Sigma_X\|_F + \frac{\|\Sigma_X\|}{|z|\sqrt{n}}\Delta.\end{aligned} \tag{98}$$

由于 $\|\Sigma_X\||z|^{-1}n^{-1/2} \ll 1$，利用（94）可得 $|\breve{m}_X - \breve{m}_{\rm lin}| \lesssim |z|^{-1}n^{-1/2}$，因此，对于第二个近似 n,

$$\left|\left\langle A\left[\left(-\breve{m}_X z\Sigma_X - z\right)^{-1} - \left(-\breve{m}_{\rm lin} z\Sigma_X - z\right)^{-1}\right]\right\rangle\right| \lesssim \frac{1}{\delta^3\sqrt{n}}\langle AA^*\rangle^{1/2}. \tag{99}$$

对于第三个近似，我们可以写成

$$\begin{aligned}
&\left|\left\langle A\Big[\Big(-\breve{m}_{\mathrm{lin}} z\Sigma_X - z\Big)^{-1} - \Big(-\breve{m}_{\mathrm{lin}} z\Sigma_{\mathrm{lin}} - z\Big)^{-1}\Big]\right\rangle\right| = \frac{1}{|z||\breve{m}_{\mathrm{lin}}|} B, \\
\text{where}\quad B &:= \left|\langle A(\Sigma_X + 1/\breve{m})^{-1} - A(\Sigma_{\mathrm{lin}} + 1/\breve{m})^{-1}\rangle\right| \\
&\le \frac{1}{\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|(\Sigma_X + 1/\breve{m})^{-1} - (\Sigma_{\mathrm{lin}} + 1/\breve{m})^{-1}\right\|_{\mathrm{F}} \\
&= \frac{1}{\sqrt{n}}\langle AA^*\rangle^{1/2}\left\|(\Sigma_X + 1/\breve{m})^{-1}(\Sigma_{\mathrm{lin}} - \Sigma_X)(\Sigma_{\mathrm{lin}} + 1/\breve{m})^{-1}\right\|_{\mathrm{F}} \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^2\sqrt{n}},
\end{aligned} \tag{100}$$

其中在最后一个不等式中，我们使用了（94）。

结合所有近似结果，我们证明了

$$\left|\left\langle A\Big[\Big(\frac{X_1^\top X_1}{k_1} - z\Big)^{-1} - \Big(-\breve{m}_{\mathrm{lin}} z\Sigma_{lin} - z\Big)^{-1}\Big]\right\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}}. \tag{101}$$

接下来，我们将验证：当我们将矩阵 X_0 替换为 X_1，并将 r_1 替换为 r_2时，假设 A.8 仍然成立。具体而言，我们希望证明，以高概率有

$$\left\|\frac{X_1^\top X_1}{k_1} - r_2 I\right\|_{\max} = O\left(\frac{1}{\sqrt{n}}\right). \tag{102}$$

注意，方程（92、93）表明

$$\|\Sigma_X - r_2 I\|_{\max} = O\left(\frac{1}{\sqrt{n}}\right). \tag{103}$$

我们有

$$\left(\frac{X_1^\top X_1}{k_1}\right)_{ij} = \frac{1}{k_1}\sum_{l=1}^{k_1}\sigma\left(\frac{x_i^\top \tilde{w}_l}{\sqrt{d}}\right)\sigma\left(\frac{\tilde{w}_l^\top x_j}{\sqrt{d}}\right) = \frac{1}{k_1}\sum_{l=1}^{k_1} Y_l, \quad \text{where } Y_l := \sigma\left(\frac{x_i^\top \tilde{w}_l}{\sqrt{d}}\right)\sigma\left(\frac{\tilde{w}_l^\top x_j}{\sqrt{d}}\right). \tag{104}$$

注意到 Y_l 是独立随机变量，根据引理 A.14 可知， Y_l 的次高斯范数为$O\left(\lambda_\sigma\left\|X/\sqrt{d}\right\|\right)$ 。因此，根据 Hoeffding 不等式，我们有

$$P\left(\left|\left(\frac{X_1^\top X_1}{k_1}\right)_{ij} - (\Sigma_X)_{ij}\right| \ge t\right) \le 2e^{-\frac{ct^2 k_1}{\lambda_\sigma\|X/\sqrt{d}\|}}, \tag{105}$$

由此应用并集界，我们可以推导出

$$\left\|\frac{X_1^\top X_1}{k_1} - \Sigma_X\right\|_{\max} = O\left(\frac{1}{\sqrt{n}}\right). \tag{106}$$

结合方程（103）和（106），我们得到所需的最大范数界。接下来，利用标准的 ε-网论证（例如参见 [46], 命题 3.4），我们可以证明

$$\frac{1}{\sqrt{d}}\|X_1 - \mathbf{E}\,X_1\| \prec 1. \tag{107}$$

由于 $\sqrt{n}\|\mathbf{E}\,X_1\|_{\max} \lesssim 1$，因此

$$\left\|\frac{\mathbf{E}\,X_1}{\sqrt{d}}\right\| \le \sqrt{\frac{nk_1}{d}}\|\mathbf{E}\,X_1\|_{\max} \lesssim \sqrt{\frac{k_1}{d}} \lesssim 1. \tag{108}$$

最后，dist$(-1/\breve{m}(z)\ \mathbb{R}_+) \ge$dist$(z, \mathbb{R}_+)$这一断言可直接由不动点方程推出，例如参见 [46]中的命题 6.2。

□

定理的证明 3.3。 这可通过反复应用命题直接得出3.1 直到我们得到 h

$$\Big(\frac{X_\ell^\top X_\ell}{k_\ell} - z_\ell\Big)^{-1} \approx c_1 \cdots c_\ell\Big(\frac{X_0^\top X_0}{d} - z_0\Big)^{-1} \tag{109}$$

在最后一层，其中“≈”应按照命题 3.1的含义理解。现在，利用 $X_0X_0^\top/n$ 是一个总体协方差矩阵为 Ω_0,的样本协方差矩阵，可得

$$\Big(\frac{X_0^\top X_0}{d}-z_0\Big)^{-1}=\frac{d}{n}\Big(\frac{X_0^\top X_0}{n}-\frac{d}{n}z_0\Big)^{-1}\approx\frac{d}{n}\Big(\frac{d}{n}m_{\mu(\Omega_0)\boxtimes\mu_{\mathrm{MP}}^{d/n}}\Big(\frac{d}{n}z_0\Big)+\frac{d-n}{dz_0}\Big), \tag{110}$$

其中我们在最后一步再次使用了命题 A.4。 □

B 总体协方差的闭式公式

B.1 多层线性化

在本附录中附录中，我们给出了总体协方差闭式表达式的（启发式）推导 s:

$$\Omega_L \coloneqq \mathbf{E}\left[\varphi(\mathrm{x})\varphi(\mathrm{x})^\top\right], \qquad \Phi_{L^\star L} \coloneqq \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi(\mathrm{x})^\top\right], \qquad \Psi_{L^\star} \coloneqq \mathbf{E}\left[\varphi^\star(\mathrm{x})\varphi^\star(\mathrm{x})^\top\right]. \tag{111}$$

这一推导已出现在 [12], 中，为确保内容完整，我们在此将其列出。

结果回顾 考虑 dRF（2）和目标函数（25），以及数据 $x \sim \mathcal{N}(0,\Omega_0)$。$\Omega_0$ 假设具有广延的弗罗贝尼乌斯范数和迹，即存在常数 c, c'，使得渐近地（注意 $k_0 = d$）

$$c < \frac{1}{d}\operatorname{tr}\Omega_0^2 = \frac{1}{d}||\Omega_0||_F^2 < c' < \infty, \qquad c < \frac{1}{d}\operatorname{tr}\Omega_0 < c' < \infty. \tag{112}$$

就极限谱密度 μ 而言，这些假设意味着一阶矩和二阶矩有限且非零。考虑由以下递推关系定义的方差序列

$$r_{\ell+1}^{(\star)} = \Delta_{\ell+1}^{(\star)} \mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[\sigma_\ell^{(\star)}(z)^2\right] \tag{113}$$

满足初始条件

$$r_1^{(\star)} = \Delta_1^{(\star)} \frac{1}{d}\operatorname{tr}\Omega_0 \tag{114}$$

以及 GET [5, 36, 52] 系数

$$\kappa_1^{\ell(\star)} = \frac{1}{r_\ell^{(\star)}} \mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[z\sigma_\ell^{(\star)}(z)\right] \qquad \kappa_*^{\ell(\star)} = \sqrt{\mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[\sigma_\ell^{(\star)}(z)^2\right] - r_\ell^{(\star)}\left(\kappa_1^{\ell(\star)}\right)^2}. \tag{115}$$

定义矩阵序列

$$\Omega_{\ell+1}^{\mathrm{lin}} = \kappa_1^{\ell 2}\frac{W_{\ell+1}\Omega_\ell^{\mathrm{lin}}W_{\ell+1}^\top}{k_\ell} + \kappa_*^{\ell 2} I_{k_{\ell+1}} \tag{116}$$

$$\Psi_{\ell+1}^{\mathrm{lin}} = \kappa_1^{\star\ell 2}\frac{W_{\ell+1}^\star\Psi_\ell^{\mathrm{lin}}W_{\ell+1}^{\star\top}}{k_\ell^\star} + \kappa_*^{\star\ell 2} I_{k_{\ell+1}^\star} \tag{117}$$

采用初始化

$$\Omega_0^{\mathrm{lin}} \coloneqq \Psi_0^{\mathrm{lin}} = \Omega_0, \tag{118}$$

以及矩阵

$$\Phi_{\ell^\star\ell}^{\mathrm{lin}} = \prod_{r=1}^{\ell}\prod_{s=1}^{\ell^\star}\kappa_1^r\kappa_1^{s\star} \times \frac{W_{\ell^\star}\cdot\ldots\cdot W_1^\star\cdot\Sigma\cdot W_1^\top\cdot\ldots\cdot W_\ell^\top}{\prod_{r=0}^{\ell-1}\prod_{s=0}^{\ell^\star-1}\sqrt{k_r k_s^\star}}. \tag{119}$$

于是 $\Omega_L \approx \Omega_L^{\mathrm{lin}}$，$\Psi_{L^\star} \approx \Psi_{L^\star}$ 和 $\Phi_{\ell^\star\ell} \approx \Phi_{\ell^\star\ell}^{\mathrm{lin}}$。$A \approx B$ 可理解为 $||A-B||_F^2/d = \mathcal{O}(1/d)$。

示例： $L = 2$ 具体而言，我们给出一个关于 $L^\star = 1$，$L = 2$（随机特征教师、两层深度随机网络学生）的示例。学生的递推关系（116）（119）为 $L = 2$

$$\Omega_2 = (\kappa_1^1)^2(\kappa_1^2)^2\frac{W_2W_1\Sigma W_1^\top W_2^\top}{k_1 d} + (\kappa_1^2)^2(\kappa_*^1)^2\frac{W_2W_2^\top}{k_1} + (\kappa_*^2)^2 I_{k_1} \tag{120}$$

$$\Psi_1 = (\kappa_1^{1\star})^2\frac{W_1^\star\Sigma W_1^{\star\top}}{d} + (\kappa_*^{1\star})^2 I_{k_1^\star} \tag{121}$$

$$\Phi_{1,2} = \kappa_1^1\kappa_1^2\kappa_1^{1\star}\frac{W_1^\star\Sigma W_2^\top W_1^\top}{d\sqrt{k_1}} \tag{122}$$

等效线性网络 请注意，线性化意味着可以将第 ℓ层视为一个带噪声的线性层，

$$\varphi_\ell(x)^{\text{lin}} \approx \kappa_1^\ell \frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot x + \kappa_*^\ell \xi_\ell \tag{123}$$

其中， $\xi_\ell \in \mathbb{R}^{k_\ell}$ 是与各层独立的 i.i.d. 高斯噪声，并且在教师和学生之间也相互独立（假设教师和学生的权重是独立抽取的）。教师的情况类似：

$$\varphi_\ell^\star(x) \approx \kappa_1^{\star\ell} \frac{1}{\sqrt{k_{\ell-1}^\star}} W_\ell^\star \cdot x + \kappa_*^{\star\ell} \xi_\ell^\star \tag{124}$$

这为重新推导关系式（116）和（119）提供了一种简单的方法。

B.2 Ω_L 的推导概要

我们首先推导两个连续层后激活的协方差之间的关系，然后进行迭代。注意，由于针对 $\Psi_{L^\star}$ 的计算经适当调整后完全相同，我们这里只讨论 Ω_L。

单层传播 考虑辅助单层问题

$$h(x) = \sigma\left(\frac{1}{\sqrt{d}} W \cdot x\right) \tag{125}$$

其中 $x \sim \mathcal{N}(0,\ \Sigma)$。递归地假设 Σ 满足性质（112）。后激活的总体协方差 h 为

$$\Omega_{ij} = \langle h_i(x) h_j(x) \rangle_x = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \frac{w_i^\top \Sigma w_i}{d} & \frac{w_i^\top \Sigma w_j}{d} \\ \frac{w_i^\top \Sigma w_j}{d} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \frac{w_i^\top \Sigma w_i}{d} & \frac{w_i^\top \Sigma w_j}{d} \\ \frac{w_i^\top \Sigma w_j}{d} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}}} \sigma(u)\sigma(v). \tag{126}$$

注意到有

$$\mathbb{E}_w \frac{w^\top \Sigma w}{d} = \frac{\Delta}{d} \operatorname{tr} \Sigma \equiv r, \tag{127}$$

根据假设，其阶数为 1。将 $\Sigma = U \Lambda U^\top$ 对角化，并注意到 $U^\top w$ 仍为具有独立元素的高斯分布，

$$\mathbb{V}_w \left[\frac{w^\top \Sigma w}{d}\right] = \frac{1}{d^2} \sum_{i=1}^d \lambda_i^2 \mathbb{V}_w \left[(U^\top w)_i^2\right] = \frac{2\Delta}{d^2} \operatorname{tr} \Sigma^2 = \frac{2\Delta}{d} \frac{||\Sigma||_F^2}{d} = \mathcal{O}\left(\frac{1}{d}\right) \tag{128}$$

只要 $||\Sigma||_F^2/d$ 有限。我们使用了以下事实： 1– 自由度变量的方差为 2。将 r 的定义代入上式，可得对于 $i \neq j$:

$$\begin{aligned} \Omega_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{r^2 - \mathcal{O}\left(\frac{1}{d}\right)}(ru^2 + rv^2)} e^{\frac{1}{r^2 - \mathcal{O}\left(\frac{1}{d}\right)} \frac{w_i^\top \Sigma w_j}{d} uv}}{2\pi \sqrt{r^2 - \mathcal{O}\left(\frac{1}{d}\right)}} \sigma(u)\sigma(v) \\ &= \left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)\right)^2 + \frac{1}{r} \frac{w_i^\top \Sigma w_j}{d} \left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} z\sigma(z)\right)^2 + \mathcal{O}\left(\frac{1}{d}\right) \\ &= \kappa_1^2 \times \frac{w_i^\top \Sigma w_j}{d}. \end{aligned} \tag{129}$$

在对角线上（$i = j$），这变为

$$\Omega_{ii} = \int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)^2 = \kappa_*^2 + r\kappa_1^2 \tag{130}$$

从而得到

$$\Omega = \kappa_1^2 \frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k \tag{131}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right] \qquad \kappa_*^2 = \mathbb{E}_z^{\mathcal{N}(0,r)}\left[\sigma(z)^2\right] - r \times \kappa_1^2 \tag{132}$$

这将 GET [5] 推广 [56] 扩展到任意输入协方差。

逐层迭代， (115) 和 (116可由单层结果 (132) 和 (131) 通过直接递推得到。只需将 (113) 与单层方差 r (127) 联系起来。

$$\begin{aligned} r_{\ell+1} &= \Delta_{\ell+1}\frac{1}{k_\ell}\operatorname{tr}\Omega_\ell \\ &= \Delta_{\ell+1}\left(\frac{1}{k_\ell}\left(\kappa_1^\ell\right)^2\operatorname{tr}\left[\frac{W_\ell\Omega_{\ell-1}W_\ell^\top}{k_{\ell-1}}\right] + \left(\kappa_*^\ell\right)^2\right) \\ &= \Delta_{\ell+1}\left(\left(\kappa_1^\ell\right)^2 r_\ell + \left(\kappa_*^\ell\right)^2\right) \\ &= \Delta_{\ell+1}\mathbb{E}_z^{\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(z)^2\right] \end{aligned} \tag{133}$$

我们使用了

$$\begin{aligned} \frac{1}{k_\ell}\operatorname{tr}\left[\frac{W_\ell\Omega_{\ell-1}W_\ell^\top}{k_{\ell-1}}\right] &= \frac{1}{k_{\ell-1}}\sum_{i=1}^{k_{\ell-1}}\lambda_i^{\ell-1}\frac{1}{k_\ell}\left(U^\top W_\ell^\top W_\ell U\right)_{ii} \\ &= \frac{1}{k_{\ell-1}}\sum_{i=1}^{k_{\ell-1}}\lambda_i^{\ell-1}\Delta_\ell \\ &= \Delta_\ell\frac{1}{k_{\ell-1}}\operatorname{tr}\Omega_{\ell-1} = r_\ell \end{aligned} \tag{134}$$

我们利用了 $W_\ell U$ 也是一个独立同分布的高斯矩阵这一点。最后，必须检查关于 Σ 的假设，即 $||\Sigma||_F^2/d$，$\operatorname{tr}\Sigma/d = \mathcal{O}(1)$，是否也适用于 Ω。由于 $W\Sigma W^\top$ 是半正定的，显然有

$$\frac{1}{k}||\kappa_1^2\frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k||_F^2 \geq \kappa_*^2 > 0. \tag{135}$$

利用三角不等式和弗罗贝尼乌斯范数的次乘性，可以得到上界：

$$\begin{aligned} \frac{1}{k}||\kappa_1^2\frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k||_F^2 &\leq \frac{1}{k}||\kappa_1^2\frac{W\Sigma W^\top}{d}||_F^2 + \kappa_*^2 \\ &\leq \kappa_*^2 + \frac{||W||_F^4}{d^2}\frac{||\Sigma||_F^2}{k} \\ &\leq \kappa_*^2 + c' < \infty. \end{aligned} \tag{136}$$

我们使用了 $||W||_F^2/dk = 1$ 几乎必然渐近成立这一点。接下来考察迹，

$$\frac{1}{k}\operatorname{Tr}\left[\kappa_1^2\frac{W\Sigma W^\top}{d} + \kappa_*^2 I_k\right] = \kappa_*^2 + \frac{\kappa_1^2}{kd}\operatorname{Tr}\left[\Sigma W^\top W\right]. \tag{137}$$

界定

$$0 \leq \frac{\kappa_1^2}{kd}\operatorname{Tr}\left[\Sigma W^\top W\right] = \frac{\kappa_1^2}{kd}\sum_{i=1}^{k} w_i^\top\Sigma w_i = \kappa_1^2\frac{1}{d}\operatorname{Tr}\{\Sigma\} \leq \kappa_1^2 c', \tag{138}$$

其中，最后一个界渐近地几乎必然成立。

B.3 $\Phi_{L^\star L}$的推导概要

现在我们转而讨论两个权重相互独立的随机网络的后激活之间的交叉协方差 $\Phi_{L^\star L}$。同样，我们先给出一个预备结果，研究两个相关高斯变量在权重独立抽取的非线性层中传播时的统计特性。

两个高斯变量在两层中传播 考虑两个联合高斯变量 $u \in \mathbb{R}^d, \quad v \in \mathbb{R}^k$

$$(u, v) \sim \mathcal{N}\begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix} \tag{139}$$

各自独立地通过一个非线性层传播

$$h^\star(u) = \sigma_\star\left(\frac{1}{\sqrt{d_\star}} W_\star \cdot u\right), \qquad h(v) = \sigma\left(\frac{1}{\sqrt{d}} W \cdot v\right). \tag{140}$$

权重 $W_\star \in \mathbb{R}^{k_\star \times d_\star}$ 和 $W \in \mathbb{R}^{k \times d}$ 的元素分别独立采样自高斯分布，方差分别为 $\Delta_\star$和 $\Delta_\circ$ 交叉协方差 Φ^h 的第 i, j–个元素可表示为

$$\Phi^h_{ij} = \langle h^\star_i(u) h_j(v) \rangle_{u,v} = \int \frac{e^{-\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} \frac{w_i^{\star\top}\Sigma w_i^\star}{d_\star} & \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} \\ \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix} \frac{w_i^{\star\top}\Sigma w_i^\star}{d_\star} & \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} \\ \frac{w_i^{\star\top}\Sigma w_j}{\sqrt{d_\star d}} & \frac{w_j^\top \Sigma w_j}{d} \end{pmatrix}}} \sigma_\star(x)\sigma(y) \tag{141}$$

与之前一样，随机变量 ${w_i^\star}^\top \Sigma w_i^\star / d_\star$ 和 $w_j^\top \Sigma w_j / d$ 集中在其均值附近

$$r_\star \equiv \frac{\Delta_\star}{d_\star} \operatorname{tr} \Psi \qquad\qquad r \equiv \frac{\Delta}{d} \operatorname{tr} \Omega \tag{142}$$

将这些定义代入上式:

$$\begin{aligned}
\Phi^h_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}(r x^2 + r_\star y^2)} e^{\frac{1}{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}\frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}} xy}}{2\pi\sqrt{r_\star r - \mathcal{O}\left(\frac{1}{d}\right)}} \sigma_\star(x)\sigma(y) \\
&= \left(\int \frac{e^{-\frac{1}{2r_\star}z^2}}{\sqrt{2\pi r_\star}} \sigma_\star(z)\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} \sigma(z)\right) + \frac{1}{r_\star r}\frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}}\left(\int \frac{e^{-\frac{1}{2r_\star}z^2}}{\sqrt{2\pi r_\star}} z\sigma_\star(z)\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}} z\sigma(z)\right) + \mathcal{O}\left(\frac{1}{d}\right) \\
&:= \kappa_1 \kappa_1^\star \times \frac{w_i^{\star\top}\Phi w_j}{\sqrt{d_\star d}}
\end{aligned} \tag{143}$$

从而得到

$$\Phi^h = \kappa_1 \kappa_1^\star \frac{W_\star \Phi W^\top}{\sqrt{d_\star d}} \tag{144}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right] \qquad\qquad \kappa_1 = \frac{1}{r_\star}\mathbb{E}_z^{\mathcal{N}(0,r_\star)}\left[z\sigma_\star(z)\right] \tag{145}$$

一个高斯变量通过一个层传播我们还需要另一个结果，再次讨论两个相关的高斯变量，其中只有一个通过非线性层传播。考虑两个联合高斯变量 $u \in \mathbb{R}^{d_\star}$, $v \in \mathbb{R}^d$

$$(u, v) \sim \mathcal{N}\begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix} \tag{146}$$

只有 v 通过非线性层传播

$$h(v) = \sigma\left(\frac{1}{\sqrt{k}} W \cdot v\right). \tag{147}$$

条目 $W \in \mathbb{R}^{k\times d}$ 是从方差为 Δ 的高斯分布中独立采样得到的。 $h(v)$ 与 u 之间交叉协方差 Φ 的第 $i,j-$ 个元素可表示为

$$\Phi^h_{ij} = \langle u_i h_j(v)\rangle_{u,v} = \int \frac{e^{-\frac{1}{2}\begin{pmatrix}x & y\end{pmatrix}\begin{pmatrix}\Psi_{ii} & \frac{\Phi_i w_j}{\sqrt{k}} \\ \frac{\Phi_i w_j}{\sqrt{k}} & \frac{w_j^\top \Sigma w_j}{k}\end{pmatrix}^{-1}\begin{pmatrix}x\\ y\end{pmatrix}}}{\sqrt{\det 2\pi \begin{pmatrix}\Psi_{ii} & \frac{\Phi_i w_j}{\sqrt{k}} \\ \frac{\Phi_i w_j}{\sqrt{k}} & \frac{w_j^\top \Sigma w_j}{k}\end{pmatrix}}} x\sigma(y) \tag{148}$$

与之前一样，随机变量 $w_j^\top \Sigma w_j/k$ 集中在其均值附近

$$r \equiv \frac{\Delta}{k}\operatorname{tr}\Omega \tag{149}$$

将此定义代入上式:

$$\begin{aligned}\Phi^h_{ij} &= \int \frac{e^{-\frac{1}{2}\frac{1}{\Psi_{ii}r-\mathcal{O}\left(\frac{1}{d}\right)}(rx^2+\Psi_{ii}y^2)} e^{\frac{1}{\Psi_{ii}r-\mathcal{O}\left(\frac{1}{d}\right)}\frac{\Phi_i w_j}{\sqrt{k}}xy}}{2\pi\sqrt{\Psi_{ii}r-\mathcal{O}\left(\frac{1}{d}\right)}} x\sigma(y) \\ &= \left(\int \frac{e^{-\frac{1}{2\Psi_{ii}}z^2}}{\sqrt{2\pi\Psi_{ii}}}z\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}}\sigma(z)\right) + \frac{1}{\Psi_{ii}r}\frac{\Phi_i w_j}{\sqrt{k}}\left(\int \frac{e^{-\frac{1}{2\Psi_{ii}}z^2}}{\sqrt{2\pi\Psi_{ii}}}z^2\right)\left(\int \frac{e^{-\frac{1}{2r}z^2}}{\sqrt{2\pi r}}z\sigma(z)\right) + \mathcal{O}\left(\frac{1}{d}\right) \\ &:= \kappa_1 \times \frac{\Phi_i w_j}{\sqrt{k}}\end{aligned} \tag{150}$$

得到

$$\Phi^h = \kappa_1 \frac{\Phi W^\top}{\sqrt{k}} \tag{151}$$

with

$$\kappa_1 = \frac{1}{r}\mathbb{E}_z^{\mathcal{N}(0,r)}\left[z\sigma(z)\right]. \tag{152}$$

迭代为了证明（32），我们对（144）进行$\min(L, L_\star)$ 次迭代，随后进行$\max(L, L_\star) - \min(L, L_\star)$ 次单层关系（151）的迭代，从而完成数据在更深层（教师（25）或学生（2））神经网络中的传播。

B.4 协方差的谱

本节推导线性化协方差（21）的谱，这一结果本身也具有独立的研究价值。

有用的恒等式 首先回顾一些有用的事实。对于具有独立同分布高斯条目的 $W \in \mathbb{R}^{k_\ell \times k_{\ell-1}}$，以及当 $k_{\ell-1} \to \infty$ 时具有极限谱密度 $\mu_{\ell-1}$ 的确定性矩阵 $\Sigma \in \mathbb{R}^{k_{\ell-1}\times k_{\ell-1}}$，根据 $XX^\top$ 和 $X^\top X$ 除零特征值外具有相同谱这一事实，我们有:

$$\mu_{\frac{1}{k_{\ell-1}}W\Sigma W^\top} = \frac{k_{\ell-1}}{k_\ell} \times \left[\frac{k_\ell}{k_{\ell-1}} \otimes \mu_{\frac{1}{k_\ell}\Sigma^{\frac{1}{2}}W^\top W\Sigma^{\frac{1}{2}}}\right] + \frac{k_\ell - k_{\ell-1}}{k_\ell}\delta \tag{153}$$

$\frac{1}{k_\ell}\Sigma^{\frac{1}{2}}W^\top W\Sigma^{\frac{1}{2}}$ 的谱为

$$\mu_{\mathrm{MP}}^{\frac{k_{\ell-1}}{k_\ell}} \boxtimes \mu_{\ell-1} \tag{154}$$

其中 $\mu^{\frac{k_{\ell-}}{1}}{}^{k_\ell}{}_{\mathrm{MP}}$ 是长宽比为 $k_{\ell-1}/k_\ell$的马尔琴科-帕斯图尔分布。用斯蒂尔切斯变换表示:

$$m_{\frac{1}{k_{\ell-1}}W\Sigma W^\top}(z) = \left(\frac{k_{\ell-1}}{k_\ell}\right)^2 \times m_{\frac{1}{k_\ell}\Sigma^{\frac{1}{2}}W^\top W\Sigma^{\frac{1}{2}}}\left(\frac{k_{\ell-1}}{k_\ell}z\right) + \left(\frac{k_{\ell-1}}{k_\ell} - 1\right)\frac{1}{z} \tag{155}$$

利用 Marčenko–Pastur 映射，并采用简写 $\gamma_\ell = \frac{k_{\ell-1}}{k_\ell}$，可得 $\frac{1}{k_{\ell-1}} W\Sigma W^\top$的斯蒂尔切斯变换是以下方程的解

$$m(z) = \int \frac{(\gamma_\ell - 1)xm(z) - \gamma_\ell}{zxm(z) + \gamma_\ell z} d\mu_{\ell-1}(x) = \frac{\gamma_\ell - 1}{z} - \frac{\gamma_\ell^2}{z} \int \frac{1}{xm(z) + \gamma_\ell} d\mu_{\ell-1}(x) \tag{156}$$

谱 $\Omega_\ell^{\text{lin}}\Omega_\ell^{\text{lin}}$的谱分布 μ_ℓ 由以下递推关系给出

$$\mu_\ell = \left(\kappa_1^\ell\right)^2 \otimes \left[\frac{k_{\ell-1}}{k_\ell} \times \left[\frac{k_\ell}{k_{\ell-1}} \otimes \mu_{\text{MP}}^{\frac{k_{\ell-1}}{k_\ell}} \boxtimes \mu_{\ell-1}\right] + \frac{k_\ell - k_{\ell-1}}{k_\ell}\delta\right] \oplus \left(\kappa_*^\ell\right)^2 \tag{157}$$

初始条件为 $\mu_0 = \mu_{\Omega_0 \circ}$ 由此可得

$$m_\ell(z) = \frac{\gamma_\ell - 1}{z - \left(\kappa_*^\ell\right)^2} - \frac{\gamma_\ell^2 \left(\kappa_1^\ell\right)^2}{(z - \left(\kappa_*^\ell\right)^2) m_\ell(z)} m_{\ell-1}\left(-\frac{\gamma_\ell}{\left(\kappa_1^\ell\right)^2 m_\ell(z)}\right). \tag{158}$$

数值方案 现在我们讨论一种求解 (158) 的数值方案。注意，每个 $m_{\ell-1}$ 仅在

$$z_{\ell-1} \equiv -\frac{\gamma_\ell}{\left(\kappa_1^\ell\right)^2 m_\ell(z)} \tag{159}$$

为了进行数值求解，我们保留两个数组（m_0, ..., m_L）和（z_0, ...z_L），其中 $m_\ell \equiv m_\ell(z_\ell)$。为简单起见，考虑输入协方差为单位矩阵的情况，即

$$m_0(z_0) = \frac{1}{1 - z_0} \tag{160}$$

然后迭代，直至收敛

$$\forall 0 \le i \le \ell - 1, \qquad z_i \leftarrow -\frac{\gamma_{i+1}}{\left(\kappa_1^i\right)^2 m_{i+1}} \tag{161}$$

并保留

$$z_L = \lambda + i\eta \tag{162}$$

其中 $\eta = 0^+$ and λ 是我们希望计算密度 $\mu_L(\lambda)$ 的取值点。接下来我们更新

$$\forall 1 \le i \le L, \qquad m_i \leftarrow \frac{\gamma_i - 1 - \sqrt{(\gamma_i - 1)^2 - 4\frac{m_{i-1}\gamma_i^2}{\left(\kappa_1^i\right)^2}\left(z_i - \left(\kappa_*^i\right)^2\right)}}{2\left(z_i - \left(\kappa_*^i\right)^2\right)} \tag{163}$$

其中，我们直接求解了更新方程（158），从经验上看，这比直接迭代（158）收敛得更好。图 4 展示了 157 中的理论渐近分布，该分布对应 3–层随机特征网络，其激活函数为 Sigmoid 和符号函数。可以发现，该理论分布与利用 10^5 个独立样本估计总体协方差得到的数值估计高度吻合。图 5 展示了矩形 $L = 5$ 层双曲正切神经网络的渐近分布。与 [8] 在类似模型中对共轭核的观察结果一致，分布的支撑集随深度增加而扩大，同时小特征值的密度也随之增加。值得注意的是，在多种情形下，小特征值的存在已被发现 [15, 57–59] 会在使用（ℓ_2）进行回归时带来额外的隐式正则化。这一观点将在第 2 节中进一步讨论 5。

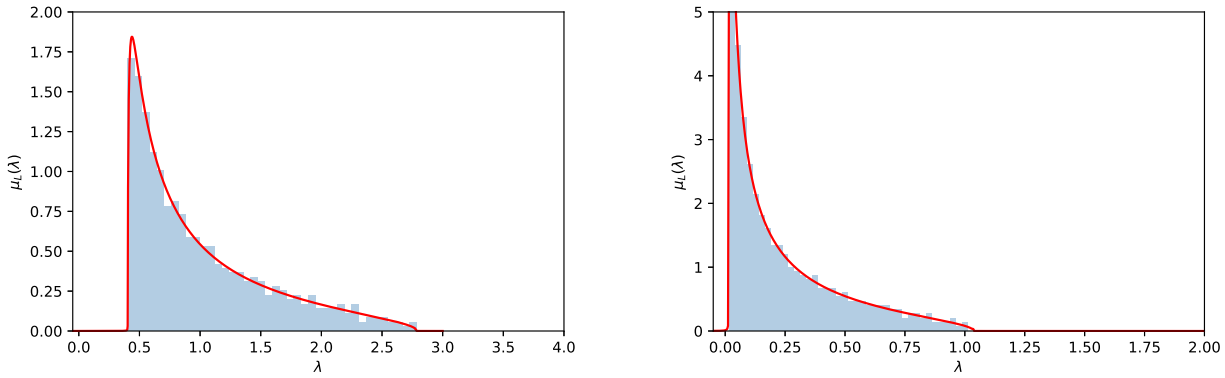

图 4：具有 Ω_2 (21) 的 $2-$ 隐藏层神经网络（2），其架构为 $\gamma_1 = 6/5$，$\gamma_2 = 3/5$，激活函数为 $\sigma_1 = \sigma 2 =$ 双曲正切(2·)（上图），以及 $\gamma_1 = 7/10$，$\gamma_2 = 6/5$，激活函数为$\sigma_1 = \sigma 2 =$ 符号函数 （下图）。（红色）通过求解递推关系（157）得到的理论渐近谱分布（数值方案的更多细节见附录 B）；（蓝色）根据 10^5 个样本在维度 $d = 1000$下的样本协方差估计得到的经验分布。

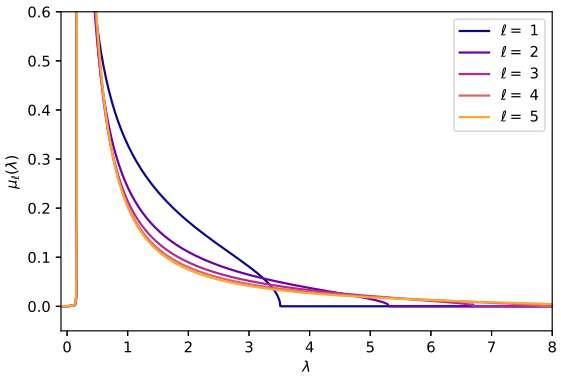

图 5：网络中后激活 $h_\ell(\mathrm{x})$ 的渐近谱分布 μ_ℓ 随 $1 \le \ell \le L = 5$ 的演变。该网络的架构为 $\gamma_1 = \ldots = \gamma_5 = 1$，采用 $\sigma_1 = \ldots \sigma_5 = \tanh$ 激活函数，输入为各向同性数据 $\Omega_0 = I_d$。经过非线性层传播后，分布的支撑集趋于扩展，小特征值的密度也会增加。

C 岭回归的误差普适性

在本附录中，我们将详细推导（27）和推论 4.1。首先，我们先回顾该推论的设定。这里，我们关注的是渐近均方测试误差的刻画：

$$\mathcal{E}_{\text{gen.}}(\hat{\theta}) = \mathbb{E}\left(y - \frac{\hat{\theta}^\top \varphi(\mathrm{x})}{\sqrt{k}}\right)^2 \tag{164}$$

其中 $\varphi:\ \mathbb{R}^d \to \mathbb{R}^k$ 是 L 层随机特征，定义见（2），而 $\hat{}\theta \in \mathbb{R}^k$ 是岭估计量：

$$\begin{aligned}\hat{\theta} &= \operatorname{argmin}\left[\sum_{\mu=1}^{n}\left(y^\mu - \frac{\hat{\theta}^\top \varphi(\mathrm{x}^\mu)}{\sqrt{k}}\right)^2 + \frac{\lambda}{2}||\theta||_2^2\right] \\ &= \frac{1}{\sqrt{k}}\left(\lambda I_k + \frac{1}{k} X_L X_L^\top\right)^{-1} X_L y\end{aligned} \tag{165}$$

其中，按照正文中的记号，我们将特征矩阵定义为 $X_L \in \mathbb{R}^{k\times n}$ 将 $\varphi(\mathrm{x}^\mu)$按列堆叠而成，并将标签向量记为 $y \in \mathbb{R}^n$。特别地，在推论4.1 中，我们关注标签在加性高斯噪声影响下由具有相同架构的 L层随机特征目标函数生成的情形。具体而言，

这可以写为：

$$y^{\mu} = \frac{\theta_{\star}^{\top}\varphi(\mathrm{x}^{\mu})}{\sqrt{k}} + z^{\mu} \tag{166}$$

其中 $\theta_{\star} \sim \mathcal{N}(0_k,\ I_k)$ 和 $z^{\mu} \sim \mathcal{N}(0,\ \Delta)$ 相互独立。请注意，就此处的讨论而言，我们不需要假设输入 $\mathrm{x}^{\mu} \in \mathbb{R}^d$ 服从高斯分布，只需假设数据矩阵 $X_0 \in \mathbb{R}^{d\times n}$ 满足集中条件（11）。特别地，这意味着本附录中的结果适用于在给定训练输入 X_0的条件下的测试误差（164）。

从这里开始，计算过程是标准的，并且与其他在不同假设下推导岭回归闭式渐近结果的研究密切相关，例如 [50, 60–62]。首先，注意我们可以改写为：

$$\begin{aligned}
\mathcal{E}_{\text{gen.}}(\hat{\theta}) &= \mathbb{E}_{\mathbf{z},\theta_{\star},z,\mathrm{x}}\left(\frac{\theta_{\star}^{\top}\varphi(\mathrm{x})}{\sqrt{k}} + z - \frac{\hat{\theta}^{\top}\varphi(\mathrm{x})}{\sqrt{k}}\right)^2 \\
&\overset{(a)}{=} \frac{1}{k}\mathbb{E}_{,\mathbf{z},\theta_{\star}}(\hat{\theta}-\theta_{\star})^{\top}\mathbb{E}_{\mathrm{x}}\left[\varphi(\mathrm{x})\varphi(\mathrm{x})^{\top}\right](\hat{\theta}-\theta_{\star}) + \Delta \\
&\overset{(b)}{=} \frac{1}{k}\mathbb{E}_{\mathbf{z},\theta_{\star}}\left[(\hat{\theta}-\theta_{\star})^{\top}\Omega_L(\hat{\theta}-\theta_{\star})\right]
\end{aligned} \tag{167}$$

其中在 (a) 中，我们利用了测试样本的独立性，显式地取了 z 平均值；在 (b) 中，我们使用了定义（87）。着重考察：

$$\begin{aligned}
\hat{\theta} - \theta_{\star} &= {}^{1}\!/\!\sqrt{k}\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1} X_L^{\top}({}^{1}\!/\!\sqrt{k}\, X_L^{\top}\theta_{\star} + \mathbf{z}) - \theta_{\star} \\
&= \left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\left({}^{1}\!/\!_k X_L X_L^{\top} - I_k\right)\theta_{\star} + {}^{1}\!/\!\sqrt{k}\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1} X\,\mathbf{z} \qquad (168) \\
&\overset{(c)}{=} -\lambda\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\theta_{\star} + {}^{1}\!/\!\sqrt{k}\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1} X\,\mathbf{z} \qquad (169)
\end{aligned}$$

其中在 (c) 中，我们使用了以下形式的伍德伯里恒等式：

$$\lambda\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1} = I_k - {}^{1}\!/\!_k\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1} X_L X_L^{\top} \tag{170}$$

将上述结果代入（167）：

$$\begin{aligned}
\mathcal{E}_{\text{gen.}}(\hat{\theta}) &= {}^{\lambda^2}\!/\!_k\,\mathbb{E}_{\theta_{\star}}\left[\theta_{\star}^{\top}\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\theta_{\star}\right] + \\
&\quad + {}^{1}\!/\!_{k^2}\,\mathbb{E}_{\mathbf{z}}\left[\mathbf{z}^{\top} X_L^{\top}\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1} X_L\,\mathbf{z}\right] + \Delta \\
&\overset{(d)}{=} \lambda^2\left\langle\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\right\rangle + \\
&\quad + \Delta\left\langle {}^{1}\!/\!_k X_L X_L^{\top}\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L^{\top}\right)^{-1}\right\rangle + \Delta
\end{aligned} \tag{171}$$

其中，在（d）中，我们对噪声和目标权重取了期望，并结合迹的循环性质，使用了定义 $\langle\cdot\rangle \equiv {}^{1}\!/\!_k\mathrm{tr}(\cdot)$。我们可以通过在第二个迹项中加上再减去 λI_k，将上述表达式改写为适用于定理 3.3 的形式。这就得到正文中引用的表达式（27）：

$$\begin{aligned}
\mathcal{E}_{\text{gen.}}(\hat{\theta}) &= \Delta\left(\left\langle\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L\right)^{-1}\right\rangle + 1\right) + \lambda(\lambda-\Delta)\left\langle\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L\right)^{-2}\right\rangle \\
&= \Delta\left(\left\langle\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L\right)^{-1}\right\rangle + 1\right) - \lambda(\lambda-\Delta)\partial_{\lambda}\left\langle\Omega_L\left(\lambda I_k + {}^{1}\!/\!_k X_L X_L\right)^{-1}\right\rangle
\end{aligned} \tag{172}$$

请注意，最后一个表达式需要对预解式的导数应用定理 3.3 。一般而言，可以通过将某个非负矩阵 $G(z)^2 = (H-z)^{-2}$ 的平方预解式写成柯西积分的形式来证明这一点 $H \geq 0$ 。

$$G(z)^2 = G'(z) = \frac{1}{2\pi\mathrm{i}}\oint_{\gamma}\frac{1}{(w-z)^2}G(w)\,\mathrm{d}w, \tag{173}$$

其中，γ是任意一个围绕 z 且不穿过 $\mathbb{R}_+$的围道。这样，某种形如 $|\langle A(G-M)\rangle| \prec \epsilon\,\mathrm{dist}(z,\mathbb{R}_+)^{-k}$的局部定律就可以传递到导数上，得到

$$\langle A(G'(z)-M'(z)\rangle = \frac{1}{2\pi\mathrm{i}}\oint_{\gamma}\frac{1}{(w-z)^2}\langle A(G(w)-M(w))\rangle\,\mathrm{d}w = O_{\prec}\Big(\frac{\epsilon}{\mathrm{dist}(z,\mathbb{R}_+)^{k+1}}\Big), \tag{174}$$

通过令 γ 为以 z为圆心、半径为$\mathrm{dist}(z,\mathbb{R}_+)/2$ 的小圆，并利用确定性等价物在 $\mathbb{R}_+$之外也全纯这一点。

因此，在高维极限下， n， k_ℓ， $d\to\infty$在固定比率 $\alpha = {}^{n}/_{d}$和 $\gamma_\ell = {}^{k_\ell}/_{d}$下，若输入数据 $X_0\in\mathbb{R}^{d\times n}$ (11及深度随机特征的架构满足定理 3.3 的假设，且 $\lambda>0$[5]，则可应用定理 3.3 写出测试误差的渐近极限：

$$\lim_{k\to\infty}\mathcal{E}_{\text{gen.}}(\hat{\theta}) = \mathcal{E}^{\star}_{\text{gen.}}(\lambda,\Delta,\alpha,\gamma_\ell,\kappa_1^\ell,\kappa_\star^\ell) \equiv \Delta\left(\langle\Omega_L\rangle\breve{m}_L(-\lambda)+1\right) - \lambda(\lambda-\Delta)\langle\Omega_L\rangle\partial_\lambda\breve{m}_L(-\lambda) \tag{175}$$

其中 $\breve{m}_L(z)$ 可针对给定的正则化强度 $\lambda>0$，噪声水平 $\Delta>0$、样本复杂度 $\alpha>0$ 和特征架构 $(\gamma_\ell,\ \sigma_\ell)^{\ell\in[L]}$，根据式 (18) 递推计算。另一方面，由递推关系 (21) 可知，最后一层协方差 $\langle\Omega_L\rangle$的迹可写成紧凑形式

$$\langle\Omega_L\rangle = \sum_{\ell=1}^{L-1}\left(\kappa_*^\ell\right)^2\prod_{\ell'=\ell+1}^{L}\left(\kappa_1^{\ell'}\right)^2\Delta_{\ell'} + \left(\kappa_*^L\right)^2 + \frac{1}{d}\langle\Omega_0\rangle\prod_{\ell=1}^{L}\left(\kappa_1^\ell\right)^2\Delta_\ell \tag{176}$$

仅用系数 (10) 表示。

请注意，(175) 与在等价高斯数据集 $\mathcal{D}=\{(v^\mu,y^\mu)\}_{\mu\in[n]}$上进行岭回归的渐近测试误差公式完全一致：

$$y^\mu = \frac{\theta_\star^\top v^\mu}{\sqrt{k}} + z^\mu, \qquad\qquad v^\mu\sim\mathcal{N}(0_k,\Omega_L). \tag{177}$$

据我们所知，该结果最早由 [50] 推导得出。这确立了该模型渐近测试误差的高斯普适性。

C.1 可能的扩展

现在，我们讨论上述普适性结果的一些可能扩展。不过，这些扩展需要更复杂的分析，我们将其留待未来工作。这里的目标仅仅是指出定理 3.3 中确定性等价物的其他可能应用。

确定性末层权重：第一个扩展是将上述结果推广到确定性末层权重 $\theta_\star$。事实上， [63] 表明，对于确定性目标函数 $y^\mu = {}^{1}/_{\sqrt{k}}\,\theta_\star^\top\varphi(\mathrm{x}^\mu)$[6], 上的岭回归，可以通过广义交叉验证 （GCV）估计量渐近估计测试误差，其定义如下：

$$\mathrm{GCV}_\lambda = \lambda\left\langle(\lambda I_k+\hat{\Omega}_L)^{-1}\right\rangle\mathcal{E}_{\text{train.}}(\hat{\theta}) \tag{178}$$

其中，$\hat{}\Omega_L = {}^{1}/_{n}X_LX_L^\top$是特征的样本协方差矩阵，而 $\mathcal{E}_{\text{train.}}(\hat{}\theta)$ 是与岭估计量相关的训练误差：

$$\mathcal{E}_{\text{train.}}(\hat{\theta}) = \frac{1}{n}\sum_{\mu=1}^{n}\left(y^\mu - \frac{\hat{\theta}^\top\varphi(\mathrm{x}^\mu)}{\sqrt{k}}\right)^2 \tag{179}$$

特别地，证明了：

定理 C.1（ 的定理 8）。假设

$$\left|\langle\left(\frac{X_L^\top X_L}{n}-z\right)^{-1}\rangle - \breve{m}(z)\right| + \left|v^\top\left[\left(\frac{X_LX_L^\top}{zn}-I\right)^{-1} + \left(\Omega_L\breve{m}(z)+1\right)^{-1}\right]v\right| \prec \frac{\sqrt{\operatorname{Im}\breve{m}(z)}}{\sqrt{n\operatorname{Im}z}}, \tag{180}$$

对所有满足 $v^\top\Omega_L v\le 1$ 的确定性向量 v， 其中

$$\breve{m}(z) = \frac{k_L-n}{nz} + \frac{k_L}{n}m_{\mu(\Omega_L)\boxtimes\mu_{\mathrm{MP}}^{k_L/n}}(z). \tag{181}$$

那么对所有 $\lambda>0$， 均有

$$\left|GCV_\lambda - \mathcal{E}_{\text{gen.}}(\hat{\theta})\right| \lesssim n^{-1/2+o(1)}\theta_\star^\top\Omega_L\theta_\star\left[\frac{||\Omega_L||_{\text{op.}}}{\lambda} + \left(\frac{\operatorname{tr}\Omega_L}{\lambda n}\right)^{3/2}\right] \tag{182}$$

[5]严格来说，我们无需假设正则化参数在此处有正下界。只需令其衰减速度慢于 $n^{-1/18}$，定理 3.3 即可适用。 [6]为简便起见，此处讨论无噪声 $\Delta=0$ 情形。关于含噪目标函数的讨论，见 [63] 附录 B

应用固定权重 A.3 定理 $W_1, \dots, W_L$ 可得[7] 假设（180）在比例区间 $k_L \sim n$内成立，但误差的 z依赖性会更差，即为 $\mathsf{dist}(z, \mathbb{R}_+)^{-9}$，而非 $(\operatorname{Im} \check{m}(z)/\operatorname{Im} z)^{1/2}$；且仅对有界向量成立 $\|v\| \lesssim 1$。

相反，我们的结果定理 3.3 证明了（180）的一个初步版本，其中的显式确定性等价物仅取决于输入总体协方差 Ω_0 而非输出总体协方差 Ω_L，代价是误差增大了 $\sqrt{n}$倍。我们的误差率能否得到改善，从而推出等式（180），这是一个值得探讨的问题，留待未来研究。

一般情形： 如引言中所述，在一般情形下，我们关注的目标是 t:

$$f_\star(\mathrm{x}^\mu) = \frac{1}{\sqrt{k_\star}} \theta_\star^\top \varphi_\star(\mathrm{x}^\mu), \qquad \theta_\star \sim \mathcal{N}(0_{k_\star}, I_{k_\star}), \tag{183}$$

其中 $L_\star$ 多层随机特征 $\varphi_\star$： $\mathbb{R}^d \to \mathbb{R}^{k_\star}$ 不一定与 L 多层随机特征 φ： $\mathbb{R}^d \to \mathbb{R}^k$相同。如引言中所述，这将隐流形模型 （HMM）作为一个特例，该模型由 [52] 提出，用于描述结构化高维数据，其中标签仅取决于低维“潜在空间”的坐标。第 4 节中，我们给出了在此情形下计算误差的精确但启发式的公式（适用于任意凸损失）。然而，使用随机矩阵理论方法证明该公式在岭回归情形下成立颇具挑战，因为这是一个失配模型。事实上，若直接写出此情形下的测试误差表达式：

$$\begin{aligned} \mathcal{E}_{\text{gen.}}(\hat{\theta}) &= \mathbb{E}_{\theta_\star, \mathrm{x}} \left(\frac{\theta_\star^\top \varphi_\star(\mathrm{x})}{\sqrt{k_\star}} + z - \frac{\hat{\theta}^\top \varphi(\mathrm{x})}{\sqrt{k}} \right)^2 \\ &= \langle \Psi_{L_\star} \rangle + \frac{2}{\sqrt{k_\star k}} \mathbb{E}_{\theta_\star} \left[\theta_\star^\top \Phi_{L_\star L} \hat{\theta} \right] + \frac{1}{k} \mathbb{E}_{\theta_\star} \left[\hat{\theta}^\top \Omega_L \hat{\theta} \right] \end{aligned} \tag{184}$$

这里我们提醒读者回顾以下定义：

$$\Psi_\ell = \mathbb{E}\left[h_\ell(\mathrm{x}) h_\ell(\mathrm{x})^\top \right], \qquad \Phi_{\ell\ell'} = \mathbb{E}\left[h_\ell(\mathrm{x}) h_{\ell'}(\mathrm{x})^\top \right]. \tag{185}$$

事实上，将定理 3.3 应用于上述表达式并不像前面那样直接。为说明这一点，我们关注第二项：

$$\mathbb{E}_{\theta_\star} \left[\theta_\star^\top \Phi_{L_\star L} \hat{\theta} \right] = \operatorname{tr}\left[\Phi_{L_\star L} (\lambda I_k + {}^1\!/_k X_L X_L^\top)^{-1} X_L^\top X_{L_\star} \right] \tag{186}$$

其中我们定义了目标特征矩阵 $X_{L_\star} \in \mathbb{R}^{k_\star \times n}$，其列由以下各项给出 $\varphi(\mathrm{x}^\mu) \in \mathbb{R}^{k_\star}$。若直接处理，这将需要比定理 3.3 所提供的更精细的确定性等价物。可能的替代方法包括：将模型设定错误改写为有效加性噪声（例如参见 [20]的附录 B），并推导类似于假设 180 的局部定律，同时控制噪声（讨论见 [63] 的附录 B）；或者采用 [4] 中的线性铅笔方法。这为未来的研究提供了一个有趣的方向。

[7]严格来说，这需要论证 以证明，深度随机特征模型在固定权重下以高概率满足关于 {v1} 的 Lipschitz 集中性关于 X_0

D 一般情形下的精确渐近结果

在本附录中，我们详细给出 dRF（2）在深度随机网络目标函数（25）上的测试误差的精确渐近刻画，分别针对回归（图1）和分类（图2）。

推导的核心是 [11], 的定理，该定理通过协方差矩阵 $\Psi_{L_\star}$、Ω_L 和 $\Phi_{L_\star L}$，完整刻画了 GCM 的测试误差（29）。在 [11], 的原始工作中，dRF 模型的这些矩阵必须通过蒙特卡洛算法进行数值估计。然而，在本文中，（21）、（31）和（32）给出的闭式表达式（我们将在下一小节回顾）使我们能够得到完全解析的公式。我们将依次详细介绍岭回归和逻辑回归读出层的这些刻画。

D.1 网络激活二阶统计量回顾

在详细给出岭回归和逻辑回归测试误差的渐近刻画之前，我们先回顾线性化矩阵 $\Psi^{\text{lin}}_{L_\star}$、$\Omega^{\text{lin}}_L$ 和 $\Phi^{\text{lin}}_{L_\star L}$（21、31、30）的表达式。利用猜想 4.3，随后便可将这些矩阵代入 [11] 中的公式，从而得到完全解析的测试误差公式，其表达仅涉及目标网络权重（25）和系数（10）。以下表达式由显式求解递推关系（21、31、30）得到

$$\Omega^{\text{lin}}_L = \left(\prod_{\ell'=1}^{L}\frac{\kappa_1^{\ell'}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=1}^{L}\frac{\kappa_1^{\ell'}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + \sum_{\ell=1}^{L-1}(\kappa_*^{\ell})^2 \left(\prod_{\ell'=\ell+1}^{L}\frac{\kappa_1^{\ell'}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=\ell+1}^{L}\frac{\kappa_1^{\ell'}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + (\kappa_*^{L})^2 I_{k_L} \tag{187}$$

$$\Psi^{\text{lin}}_{L_\star} = \left(\prod_{\ell'=1}^{L_\star}\frac{\kappa_1^{\ell'\star}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=1}^{L_\star}\frac{\kappa_1^{\ell'\star}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + \sum_{\ell=1}^{L_\star-1}(\kappa_*^{\ell\star})^2 \left(\prod_{\ell'=\ell+1}^{L_\star}\frac{\kappa_1^{\ell'\star}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right)^\top \Omega_0 \left(\prod_{\ell'=\ell+1}^{L_\star}\frac{\kappa_1^{\ell'\star}W_{\ell'}^\top}{\sqrt{k_{\ell'-1}}}\right) + (\kappa_*^{L_\star\star})^2 I_{k_{L_\star}} \tag{188}$$

$$\Phi^{\text{lin}}_{LL_\star} = \left(\prod_{\ell=L^\star}^{1}\frac{\kappa_1^{\ell\star}W_{\ell}^\star}{\sqrt{k_{\ell}^\star}}\right)\cdot\Omega_0\cdot\left(\prod_{\ell=1}^{L}\frac{\kappa_1^{\ell}W_{\ell}^\top}{\sqrt{k_{\ell}}}\right) \tag{189}$$

在教师深度为 $L_\star = 0$（即其架构中没有隐藏层）的特殊情况下，上述表达式简化为

$$\Psi^{\text{lin}}_0 = \Omega_0 \tag{190}$$

$$\Phi^{\text{lin}}_{LL_\star} = \Omega_0\cdot\left(\prod_{\ell=1}^{L}\frac{\kappa_1^{\ell}W_{\ell}^\top}{\sqrt{k_{\ell}}}\right). \tag{191}$$

$L_\star = 0$, $L = 1$ 这种情况在文献中被称为 [5, 52] 隐藏流形模型。本文分析了其推广到具有 $L > 1$ 个隐藏层的深度学习器的情形。

D.2 岭回归

我们考虑在数据集 θ 上训练 dRF（2）的读出权重 $\mathcal{D} = \{x^\mu, y^\mu\}_{\mu=1}^n$，其中 $x^\mu \sim \mathcal{N}(0_d, \Omega_0)$ 相互独立。标签由一个深度随机网络给出

$$y^\mu = \frac{\theta_\star^\top \varphi_\star(x^\mu)}{\sqrt{k_{L_\star}}} + z^\mu, \tag{192}$$

其中 $z^\mu \sim \mathcal{N}(0,\ \Delta)$ 是高斯加性噪声，教师特征映射为

$$\varphi_\star(x) = \varphi^\star_{L_\star} \circ \ldots \circ \varphi^\star_1(x). \tag{193}$$

注意，c 与（25）相比，为了清晰起见，我们采用了记号 $\theta_\star := W^\star_{L_\star+1} \in \mathbb{R}^{k_{L_\star}}$ 定义 ng

$$\rho := \frac{\theta_\star^\top \Psi_{L_\star}\theta_\star}{k_{L_\star}}, \tag{194}$$

我们考虑通过最小化风险，使用岭回归训练学习器 dRF（θ 的最后一层2）这一问题

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{\mu=1}^{n} \left(y^{\mu} - \frac{\theta^{\top} \varphi(x^{\mu})}{\sqrt{k_L}} \right)^2 + \frac{\lambda}{2} ||\theta||^2 \right\}. \tag{195}$$

基于定理 [11] 和猜想4.3，该经验风险最小化算法达到的均方误差为

$$\epsilon_g := \mathbb{E}_{\mathcal{D}} \mathbb{E}_{x \sim \mathcal{N}(0_d, \Omega_0)} \left(f_{\star}(x) - \frac{\hat{\theta}^{\top} \varphi(x)}{\sqrt{k_L}} \right)^2 = \rho + q - 2m, \tag{196}$$

其中 q， m 为方程组的解

$$\begin{cases} \hat{V} = \frac{1}{\gamma_L} \frac{\alpha}{1+V} \\ \hat{q} = \frac{1}{\gamma_L} \alpha \frac{\rho+q-2m}{(1+V)^2} \\ \hat{m} = \frac{1}{\sqrt{\gamma_L \gamma^{\star}_{L_{\star}}}} \frac{\alpha}{1+V} \end{cases}, \qquad \begin{cases} V = \frac{1}{k_L} \operatorname{tr} \left(\lambda I_d + \hat{V} \Omega_L^{\mathrm{lin}} \right)^{-1} \Omega_L^{\mathrm{lin}} \\ q = \frac{1}{k_L} \operatorname{tr} \left[\left(\hat{q} \Omega_L^{\mathrm{lin}} + \hat{m}^2 \Phi_{LL_{\star}}^{\mathrm{lin}\top} \theta_{\star} \theta_{\star}^{\top} \Phi_{LL_{\star}}^{\mathrm{lin}} \right) \Omega_L^{\mathrm{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\mathrm{lin}} \right)^{-2} \right] \\ m = \sqrt{\frac{\gamma_L}{\gamma^{\star}_{L_{\star}}}} \frac{\hat{m}}{k_L} \operatorname{tr} \Phi_{LL_{\star}}^{\mathrm{lin}\top} \theta_{\star} \theta_{\star}^{\top} \Phi_{LL_{\star}}^{\mathrm{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\mathrm{lin}} \right)^{-1} \end{cases}. \tag{197}$$

D.3 逻辑回归

现在我们转向分类场景，此时标签由带有符号读出的深度随机网络给出

$$y^{\mu} = \operatorname{sign} \left(\frac{\theta_{\star}^{\top} \varphi_{\star}(x^{\mu})}{\sqrt{k_{L_{\star}}}} \right). \tag{198}$$

注意，这对应于 $\sigma_{L_{\star}+1} =$ 符号函数。dRF 的读出权重 θ 通过逻辑回归进行训练，采用经验风险最小化

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{\mu=1}^{n} \ln \left(1 + e^{-y^{\mu} \frac{\theta^{\top} \varphi(x^{\mu})}{\sqrt{k_L}}} \right) + \frac{\lambda}{2} ||\theta||^2 \right\}. \tag{199}$$

同样地，引入以下 [11] 辅助函数

$$Z(y, \omega, V) := \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{y\omega}{\sqrt{2V}} \right) \right)$$

以及 $f(y,\ \omega,\ V)$，其定义为以下方程的解

$$f(y, \omega, V) = \frac{y}{1 + e^{y(V f(y,\omega,V) + \omega)}}.$$

由 [11] 和猜想 4.3 可得，相关的测试误差为

$$\epsilon_g := \mathbb{E}_{\mathcal{D}} \mathbb{P}_{x \sim \mathcal{N}(0_d, \Omega_0)} \left(f_{\star}(x) \neq \operatorname{sign} \left(\frac{\hat{\theta}^{\top} \varphi(x)}{\sqrt{k_L}} \right) \right) = \frac{1}{\pi} \arccos \frac{m}{\sqrt{\rho q}}, \tag{200}$$

其中 m， q 是方程组的解

$$\begin{cases} \hat{V} = -\frac{\alpha}{\gamma_L} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \left[\sum_{y=\pm 1} Z \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) \partial_{\omega} f(y, \sqrt{q}\xi, V) \right] \\ \hat{q} = \frac{\alpha}{\gamma_L} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \left[\sum_{y=\pm 1} Z \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) f(y, \sqrt{q}\xi, V)^2 \right] \\ \hat{m} = \frac{\alpha}{\sqrt{\gamma_L \gamma^{\star}_{L_{\star}}}} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \left[\sum_{y=\pm 1} \partial_{\omega} Z \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) f(y, \sqrt{q}\xi, V) \right] \\ V = \frac{1}{k_L} \operatorname{tr} \left(\lambda I_d + \hat{V} \Omega_L^{\mathrm{lin}} \right)^{-1} \Omega_L^{\mathrm{lin}} \\ q = \frac{1}{k_L} \operatorname{tr} \left[\left(\hat{q} \Omega_L^{\mathrm{lin}} + \hat{m}^2 \Phi_{LL_{\star}}^{\mathrm{lin}\top} \theta_{\star} \theta_{\star}^{\top} \Phi_{LL_{\star}}^{\mathrm{lin}} \right) \Omega_L^{\mathrm{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\mathrm{lin}} \right)^{-2} \right] \\ m = \sqrt{\frac{\gamma_L}{\gamma^{\star}_{L_{\star}}}} \frac{\hat{m}}{k_L} \operatorname{tr} \Phi_{LL_{\star}}^{\mathrm{lin}\top} \theta_{\star} \theta_{\star}^{\top} \Phi_{LL_{\star}}^{\mathrm{lin}} \left(\lambda I_d + \hat{V} \Omega_L^{\mathrm{lin}} \right)^{-1} \end{cases} \tag{201}$$

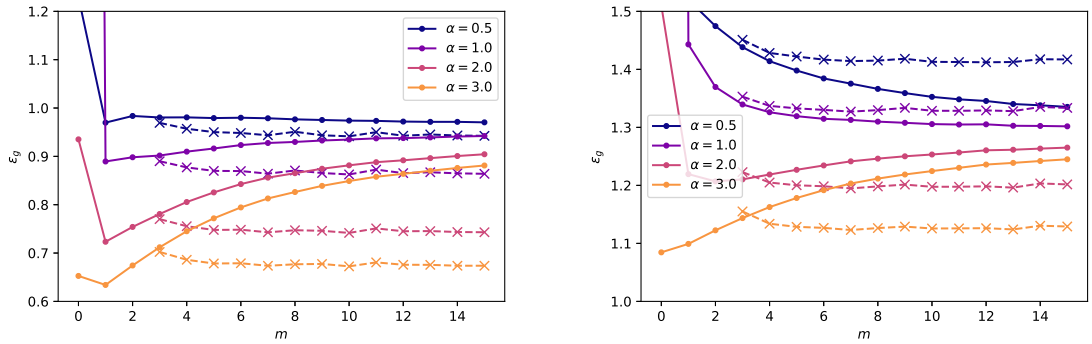

图6：在具有 $L_\star = 1$ 符号激活和宽度 $\gamma_1^\star = 4$的双层目标函数上执行回归任务时的测试误差。实线表示深度为 m 且宽度为 $\gamma_1 = ... = \gamma_m = \gamma$的深度随机特征的测试误差，而虚线表示架构为 $\gamma_1 = \gamma_3 = \gamma$ 和 $\gamma_2 = (m-2) \times k$的宽而浅的深度随机特征网络的测试误差。所有数值均采用猜想4.3的精确渐近刻画进行评估，另见附录D。参数 m（用于参数化这两个神经网络的参数数量）取值范围为0到15。对于 $\gamma = 4$ （左图），深层架构始终不如宽型架构。对于更接近插值峰值的 $\gamma = 1$ （右图），隐式深度诱导正则化意味着更深的架构表现优于更宽的架构。

E 架构诱导的隐式正则化

机器学习研究中的一个重要追求，是从理论上理解网络架构与其学习能力之间的相互作用。虽然这是一个颇具挑战、尚未解决的问题，但对 dRF（2）的研究，也就是研究在初始化时冻结中间层的神经网络，可以让我们取得一些进展，并对这些问题获得初步认识。它构成了一个高度风格化但用途广泛的试验场，可供探索一些问题，并有望为理解端到端训练的神经网络迈出最初的探索步伐。

第5 节正文讨论了 dRF 架构中由深度引起的正则化。本节将猜想4.3 作为获取渐近测试误差的灵活工具，进一步探讨其他架构特征对 dRF 性能的作用。我们的目的主要是补充第E节的讨论，并强调一些值得关注的观察结果。更完整的研究超出了本文的范围，将留待未来工作。在本节中，我们简要讨论两个问题：

- 在参数数量固定的情况下，采用较深的架构还是较宽的架构更好?
- 狭窄的（瓶颈）隐藏层对测试误差有何影响?

E.1 更深还是更宽

F对于给定数量的参数 $m\gamma d$，对于 $m \in [14]$ 我们探讨

- 一个深度为 $L = m$ 、宽度序列为 $\gamma_1 = ... = \gamma_m = \gamma$的矩形深度网络。
- 一个宽网络，其宽度为 $\gamma_1 = \gamma_3 = \gamma$ 和 $\gamma_2 = (m-2)\gamma$，深度为 $L = 3$。

对于 $m \geq 3$，从一个采用符号激活的双层目标函数中进行学习。两个神经网络的所有层均采用双曲正切作为激活函数。请注意，在这两种架构中，可训练参数的数量始终相同，因为读出层的宽度无论如何都是 γd。

图6 比较了深层架构（虚线）与宽架构（实线）。总体而言，宽架构的测试误差更小，这与我们的直觉一致：额外的层会引入更多有效噪声，因此通常会损害 dRF 的学习能力，见图6右图。不过，正如正文第5 节所讨论的，这种噪声所诱导的隐式正则化在某些情况下有助于缓解过拟合。对于含噪目标，在插值峰值附近尤其如此。图6 左图展示了这样一种情况：在小样本场景下，深层架构优于宽架构 $\alpha = 0.5$，1。如果对显式正则化 λ 进行优化，这种效应便会消失。

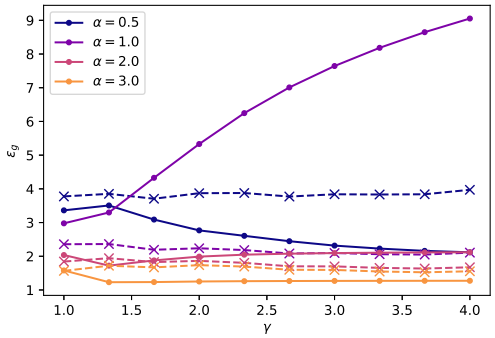

图 7：对 $L_\star = 1$ 目标函数进行回归，采用符号激活，宽度为 $\gamma_1^\star$。虚线表示测试误差（根据猜想4.3的尖锐渐近结果计算，另见附录D），对应于 $L = 4$ 深度随机特征网络，其中 $\gamma_1 = \gamma_2 = \gamma_4 = \gamma$且设有瓶颈第三层 $\gamma_3 = 1/2$。实线对应于没有瓶颈的矩形神经网络 $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma$。在接近插值峰值处（$\alpha = 1,\ 2$），瓶颈所引入的正则化能够减轻过拟合，从而降低测试误差。

E.2 瓶颈隐藏层

另一个值得关注的问题是非常窄的隐藏层所产生的影响。图7 考察了以下目标函数上的性能：$L_\star = 1$ 采用符号激活且宽度为 $\gamma_1^\star$ 的 $L = 4$ 个 dRF，$\gamma_1 = \gamma_2 = \gamma_4 = \gamma$ 并设有瓶颈第三层 $\gamma_3 = 1/2$。参数 γ 在 1 到 4 之间变化。正如直觉所预期的，瓶颈通过强制形成中间低维表示，产生了正则化效应。尽管瓶颈通常会造成信息损失，但在正则化有益的情况下（例如接近插值峰或处于噪声环境中），它会带来好处。图7展示了这样一个例子。同样，如果对显式正则化 λ 进行调节，这种效应便会消失。