arXiv:2611.28824v1 [cond-mat.dis-nn] 21 Nov 2026

层次树数据上线性分类的统计力学

J´érôme Garnier-Brun，* Federica Kulka 和 Carlo Lucibello
Bocconi Institute for Data Science and Analytics (BIDSA)
Department of Computing Sciences, Bocconi University, Milan, Italy（日期：2026年11月24日）

深度网络的成功，很大程度上归功于其利用数据结构的能力，从语言语法到图像的多尺度组织皆是如此。然而，许多针对结构化数据的统计力学方法依赖于高斯等价性；当少数潜变量影响所有输入坐标时，其假设可能不再成立。在这项工作中，我们通过一个简单的设定迈出超越这一框架的一步：训练一个线性分类器，根据经典树上广播过程的 N 个叶节点推断根节点。尽管设定简单，数据在承载标签的方向上已呈现非高斯统计，同时也表现出显著的有限尺寸效应。通过保留该方向上的精确分布，并将横向涨落近似为高斯分布，我们得到一种有限深度的副本理论，能够准确描述测得的学习曲线。在深度较大且正则化足够弱时，我们发现了双下降现象，其两个阶段在尺度上彼此分离。在广播噪声低于重构阈值的某个范围内，第一次下降发生时的样本数按 N^{ω} 缩放，其中 $0<\omega<1$ 由数据统计量决定；而过拟合峰则出现在 N 个样本的量级上。除了线性分类器之外，这一设定还可作为受控的起点，用于研究层次化数据上的更深层架构和生成模型。

I. 引言

统计力学长期以来一直为研究人工神经网络提供工具，从感知机的存储容量 [1, 2] 到教师–学生场景下的学习曲线 [3, 4] （参见 [5, 6] 及其中的参考文献），如今仍是描述高维学习的首选工具之一，包括深度网络 [7] 和基于注意力的模型 [8, 9]。然而，人们普遍认为，从真实数据中学习依赖于利用数据的结构，而这正是当前统计力学理论落后的地方 [10]。事实上，统计力学方法的核心建立在测度集中之上；这对于具有许多相互独立自由度的输入而言很自然，但与真正的低维结构先验上存在冲突。在后一种情况下，输入的统计特性跨越多个尺度组织起来，而这些尺度的数量远小于数据的维度 N，例如只有 $O(\log N)$ 个尺度，隐藏层次结构顶端的少量潜变量会影响所有输入。一方面，因此，结构化数据的可处理模型仍相对简单；例如，在 Goldt等人 [11], 提出的隐流形模型中，潜在维度仍按比例增长至N。另一方面，具有真正层级结构的模型，例如 Cagnetta等人 [12] 提出的随机层次模型（RHM）及相关的基于树的模型 [13], , 迄今仍未能接受学习的渐近分析。对它们的研究反而依赖于信噪比论证以及与受控预言机的比较，或依赖于证明适当的架构能够高效地学习它们 [14–16]。其他关于层次化学习的理论工作则将层级结构置于目标函数中，而非数据中，此时输入仍服从高斯分布 [17–23]。

在本文中，我们通过研究线性分类器的经验风险最小化迈出了建立真正层次化数据学习理论的第一步：该分类器的任务是根据基于树的广播过程的叶节点推断树的根节点。在我们的设定中，所有变量都是二元自旋，该过程只有一个参数，即子节点与其父节点不同的概率（见图 1）。与 RHM 的词汇表和生成规则相比，这是一个极大的简化。尽管如此，它已经展现出直接源于数据层级结构的关键技术挑战：**(i)** 沿着承载标签的方向呈现非高斯统计特性，使高斯等价性失效；**(ii)** 强烈的有限尺寸效应；**(iii)** 由数据结构决定的样本复杂度。通过分离出非高斯方向，并在复制计算中跟踪有限深度，我们兼顾了这三点，并将学习曲线描述到样本数为次线性的范围。在深度趋于无穷的极限下，这使我们得以揭示双下降现象，其各阶段在尺度上彼此分离：第一次下降仅由数据的统计特性决定，所需样本数为 $\sim N^{\omega}$，$\omega<1$ 随后，在正则化足够弱时，出现过拟合峰以及第二次下降，位于

比例区间。

与 RHM 相比，浅层网络会遭遇维数灾难，而广播模型则提供了一个有趣的研究机会：仅凭叶节点的简单线性统计量，就能达到接近贝叶斯错误率的分类误差，因此，我们在此研究的线性分类器层面上，这个问题就已经具有研究意义。置信传播这一贝叶斯最优估计器，仍可通过对叶节点进行分层的非线性处理，取得优于普查的表现。因此，同一模型也为研究超越本文范围的表达能力更强的架构开辟了道路，而本文提出的方法可为此提供一个起点。

本文的组织结构如下。在第II节中，我们介绍

* 通讯作者： jerome.garnier@unibocconi.it

广播模型，并推导与线性学习相关的统计量，即叶节点协方差的层次结构、普查的非高斯分布律及其有限深度修正。在第III节中，我们建立经验风险最小化的副本理论：首先研究无限深度下比例区间中的情形（第IIIA节），然后研究有限深度情形（第IIIB节），最后研究样本数低于比例关系的情形（第IIIC节），其中会出现尺度分离的双下降现象。我们将在第IV节讨论研究结果及可能的扩展。

II. 基于树的层级数据

A. 模型介绍

我们考虑深度为 ℓ 的二叉树上的广播过程，如图1所示。根自旋 $x_0 = \pm 1$均匀抽取， $\mathbb{P}(x_0 = 1) = 1/2$，每个节点将其取值传递给两个子节点，子节点彼此独立地以概率 $1-\varepsilon$ 复制该值，否则将其翻转。因此，每条边都相当于一个翻转概率为 ε 的二元对称信道，这一概率起到噪声的作用。我们将范围限制在“铁磁”情形 $0 \le \varepsilon \le \frac{1}{2}$；由对称性可知，“反铁磁”情形 $\frac{1}{2} \le \varepsilon \le 1$ 。重复这一马尔可夫步骤 ℓ 个世代后，会生成 $N = 2^\ell$ 个叶自旋 $\{x_i\}_{i=1,\ \ldots,\ N}$，它们构成输入数据。这一过程作为树上信息流的典型模型，在概率论、信息论以及统计物理中都由来已久；在统计物理中，它可以理解为树上的伊辛模型 [24–29]。

对于后续计算，使用转移矩阵描述广播过程十分方便，

$$\boldsymbol{\pi}(\varepsilon) = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}, \tag{1}$$

其特征值为 **1** 和 θ，其中我们引入了互补量

$$\theta = 1 - 2\varepsilon, \qquad 0 \le \theta \le 1, \tag{2}$$

由于 $\mathbb{E}[x_{\text{child}} \mid x_{\text{parent}}] = \theta x_{\text{parent}}$，该量可以理解为父子相关性。

在本文中，我们研究根据叶节点 $\boldsymbol{x} =$（x_1、……、x_N）预测根节点的线性分类器，其形式为 $\hat{x}_0 =$sgn($\boldsymbol{w} \cdot \boldsymbol{x}$)，其中 $\boldsymbol{w}$ 通过有限数量 P 的示例进行学习。自然的参考点是普查，其定义简单地为

$$C^{(\ell)} = \sum_{i=1}^{N} x_i, \qquad \hat{x}_0^{\text{census}} = \operatorname{sgn} C^{(\ell)}, \tag{3}$$

即对叶节点采用多数规则， $\boldsymbol{w} = \mathbf{1}$。在全文中，零分类间隔 $\boldsymbol{w} \cdot \boldsymbol{x}$ 均通过公平硬币来裁决。

对于 ± 1 对称信道，已知 [26, 27] 当生成 ε过程的参数足够小时，计数估计量能够实现非平凡的重构准确率

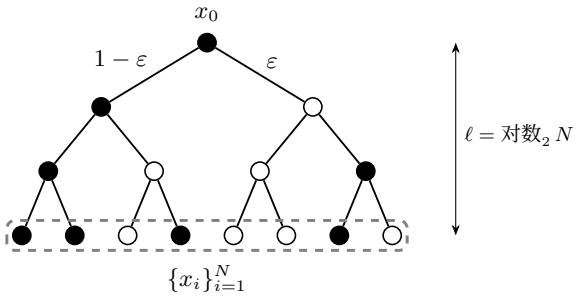

图 1. 广播树模型。根自旋 $x_0 \in \{-1, +1\}$均匀分布；每条边都会独立地以概率 ε翻转其父节点自旋。深度为 ℓ 的树的 $N = 2^\ell$ 叶节点 $\{x_i\}_{i=1}^N$ 构成输入。图中展示了一种实现，其中 $\ell = 3$ （黑色：$+1$，白色： -1）。

尽管沿树向下传播会引发涨落数量呈指数级增长，该过程仍携带一些可检测的信号。渐近而言，这一现象通过在临界值 ε_c 处发生的相变实现；该临界值由著名的 Kesten–Stigum（KS）判据 [30], 定义，对于我们的二叉树，该判据可写为

$$r = 2\theta^2 > 1, \tag{4}$$

即 $\varepsilon < \varepsilon_c = \frac{1}{2}(1 - 1/\sqrt{2}) \approx 0.146$。我们将 r 称为 KS 参数。KS 判据可推广至分支比为 d 的正则树，其中 $r = d\theta^2$ 和 $N = d^\ell$。分析方法可调整后用于这种情形，但谱的重数及其他数值系数也取决于d；此处我们仅讨论二叉树。若改为改变每个变量的状态数，例如采用 q态 Potts 广播模型，问题则会复杂得多；当 q [28] 足够大时，即使低于 KS 阈值，也可能实现重构。

如引言中所述，普查通常并非贝叶斯最优。由于生成模型是一棵树，最优估计量可通过置信传播（BP）精确计算 [29], ，在这里即对叶自旋进行简单的向上递归（附录A）。在我们考虑的二元对称信道中，与这一基准相比，普查的表现仍然相差不大，因为两种估计量都会在同一个值 $\varepsilon = \varepsilon_c$ [24, 25]处变得无法提供信息；而当 $\varepsilon < \varepsilon_c$ 时，普查的准确度仍接近贝叶斯最优的准确度，见附录A中的图7。

B. 精确统计量

我们首先描述广播数据的统计量。由于本文考虑的任务是给定完整的叶节点序列来恢复根自旋，因此我们聚焦于根节点条件统计。

1. Leaf correlation structure

如上所述，沿着一条根到叶路径，转移矩阵 $\boldsymbol{\pi}$ 的特征值为 $\mathbf{1}$ 和 θ。因此，记作 $x_0 = y$,

$$\mathbb{P}(x_i = \sigma \mid y) = \tfrac{1}{2}(1 + \sigma y \theta^{\ell}), \qquad \mathbb{E}[x_i \mid y] = y\theta^{\ell}. \quad (5)$$

除了这些单独的叶节点与根节点相关性之外，该数据模型的一个关键要素是，即使相距很远，叶节点之间仍然相关，因为它们继承了层级结构中共同祖先的涨落。对于两个叶节点（i, j），若其最近共同祖先位于距根节点深度 k_{ij} 处，则该祖先以下的条件独立性给出

$$\mathbb{E}[x_i x_j \mid y] = \theta^{2(\ell - k_{ij})}, \quad (6)$$

因此

$$\Sigma_{ij} := \mathrm{Cov}(x_i, x_j \mid y) = \theta^{2(\ell - k_{ij})} - \theta^{2\ell}. \quad (7)$$

令 $k_{ii} = \ell$，即可得到对角项 $\Sigma_{ii} = 1 - \theta^{2\ell}$，即给定根节点时叶节点的方差。随着共同祖先深度增加，协方差自然增大，也就是说，两个叶节点在树中的距离越近，协方差越大；而树的两半在给定根节点的条件下相互独立。仅凭协方差矩阵的各项，无法看出哪些集体涨落幅度较大，尤其是哪些涨落会影响普查。相反，我们利用树的自相似性，沿深度递归地对 $\boldsymbol{\Sigma}$ 进行对角化。

由于树的两半在给定根节点的条件下相互独立，按叶节点在树中的位置排序后，深度为 ℓ 的树的协方差矩阵呈分块对角形式,

$$\boldsymbol{\Sigma}^{(\ell)} = \begin{bmatrix} \mathbf{A}_{\ell-1} & 0 \\ 0 & \mathbf{A}_{\ell-1} \end{bmatrix}, \quad (8)$$

其中 $\mathbf{A}_{\ell-1}$ 是以全局根为条件时，一个半树叶节点的协方差。对于这两个叶节点，它们的共同祖先位于各自半树顶部的更近一层，相比完整树而言，式（7）给出

$$\mathbf{A}_{\ell-1} = \boldsymbol{\Sigma}^{(\ell-1)} + (1 - \theta^2)\, \theta^{2(\ell-1)}\, \mathbf{1}\mathbf{1}^{\top}. \quad (9)$$

秩一项是在给定全局根的情况下，该半树自身根的方差；这一信息以幅度 $\theta^{\ell-1}$传递给它的所有叶节点。

的所有行都包含相同的元素（排列顺序可能不同），因此它们的和相等，均匀向量$\mathbf{1}$ 是一个特征向量。由于 $\mathbf{1}\mathbf{1}^{\top}$ 仅沿着$\mathbf{1}$方向作用，特征值为 $\mathbf{1}^{\top}\mathbf{1} = 2^{\ell-1}$，更新（9）不会改变所有与$\mathbf{1}$ 正交的特征向量及其特征值，而会将均匀向量对应的特征值移动

$$2^{\ell-1}(1 - \theta^2)\, \theta^{2(\ell-1)} = (1 - \theta^2)\, r^{\ell-1}, \quad (10)$$

其中 $r = 2\theta^2$ 为上述 KS 参数。

将 $\mathbf{A}_{\ell-1}$ 的每个特征向量放在任一分块上，便得到 $\boldsymbol{\Sigma}^{(\ell)}$ 的一个特征向量，因此所有重数都翻倍。特别是，两个平移后的均匀向量合并为整棵树的均匀向量和根对比向量 $\mathbf{1}_L - \mathbf{1}_R$ ($\mathbf{1}_{L,\ R}$ 分别是两半部分的指示向量)，它们具有相同的特征值。从单层开始，$\boldsymbol{\Sigma}^{(1)} = (1 - \theta^2)\mathbf{I}_2$, 深度为 k 的树的均匀特征值满足 $\mu_k = \mu_{k-1} + (1 - \theta^2) r^{k-1}$，其中 $\mu_1 = 1 - \theta^2$。因此，对于 $0 < \theta < 1$，$\boldsymbol{\Sigma}^{(\ell)}$ 有 ℓ 个不同的特征值

$$\mu_k = (1 - \theta^2) \sum_{j=0}^{k-1} r^j = (1 - \theta^2)\, \frac{r^k - 1}{r - 1}, \quad (r \neq 1), \quad (11)$$

对于 $1 \le k \le \ell$，而恰好在 KS 转变处 $r = 1$我们有 $\mu_k = k/2$。它们的重数分别为 $2^{\ell-k}$（对应 $k < \ell$），最大特征值 μ_ℓ的重数为 2。注意，μ_k 与 ℓ无关；增加深度只会增加一个新的、更大的特征值，并改变重数。

重要的是，普查方向是一个特征向量，

$$\boldsymbol{\Sigma}\hat{\mathbf{e}} = \mu_{\mathrm{sig},\ell}\hat{\mathbf{e}}, \qquad \mu_{\mathrm{sig},\ell} = \mu_\ell, \qquad \hat{\mathbf{e}} = \frac{\mathbf{1}}{\sqrt{N}}, \quad (12)$$

因此普查涨落与横向涨落彼此不相关。除普查方向外，学习理论中只会涉及 $\boldsymbol{\Sigma}$ 的谱，而不会涉及其特征向量。正如我们将看到的，将信号方向与正交（横向）方向分开处理很方便。我们将完整谱测度和横向谱测度定义为

$$\begin{aligned} \rho_\ell &= \sum_{k<\ell} 2^{-k} \delta_{\mu_k} + 2^{1-\ell} \delta_{\mu_\ell}, \\ \rho_{\perp,\ell} &= \rho_\ell - N^{-1} \delta_{\mu_{\mathrm{sig},\ell}}. \end{aligned} \quad (13)$$

横向测度去除了一个最高模态，其总质量为 $1 - 1/N$；其平均值仍按 N 归一化。极限谱密度为

$$\rho(\mu) = \sum_{k=1}^{\infty} 2^{-k} \delta(\mu - \mu_k). \quad (14)$$

当 $r < 1$时，特征值保持有界；而当 $r > 1$时，$\mu_k \asymp r^k$ 无界增长，不过大部分质量由较小的 k 承载，仍集中在数量级为 $\mathbf{1}$ 的特征值上。因此，对于固定的 $1 < r < 2$，当 $r^p \ge 2$时，极限概率密度的谱矩 $\langle \mu^p \rangle_\rho = \sum_k 2^{-k} \mu_k^p$ 发散。特别地，当 $\sqrt{2} \le r < 2$时，二阶矩发散；而一阶矩（即给定根节点时叶节点的方差）则趋于 $\mathbf{1}$。不过，有界预解式的极限平均值仍然有限（见下文），因为发散的特征值所承载的质量趋于零。

2. Non-Gaussianity of the census

现在我们聚焦于普查量 $C^{(\ell)} = \sum_i x_i$，即上文引入的参考线性统计量。普查均值为 $y(2\theta)^\ell$，其条件方差为 $N\mu_{\mathrm{sig},\ \ell}$。等价地，单位普查方向 $\hat{\mathbf{e}} \cdot \mathbf{x}$ 上的投影 $\hat{\mathbf{e}} = \mathbf{1}/\sqrt{N}$ 的均值为 $y s_\ell$，信号振幅为

$s_\ell = \sqrt{N\theta^\ell}$，因此 $s_\ell^2 = r^\ell$。对于 $\theta > 0$，我们引入归一化有符号普查量，其无限深度极限对于 $1 < r < 2$ 是非退化的

$$c_\ell := \frac{yC^{(\ell)}}{(2\theta)^\ell}. \tag{15}$$

叶节点不含有关于根节点信息的端点 $\theta = 0$ 不予考虑。根据定义，$\mathbb{E}[c_\ell] = 1$，而归一化普查方差为

$$\gamma_\ell := \mathrm{Var}(c_\ell) = \frac{\mu_{\mathrm{sig},\ell}}{s_\ell^2} = (1-\theta^2)\sum_{j=1}^{\ell} r^{-j}. \tag{16}$$

对于 $r > 1$，我们可以写成

$$\gamma_\ell = \gamma_\infty(1 - r^{-\ell}), \qquad \gamma_\infty = \frac{1-\theta^2}{2\theta^2 - 1}. \tag{17}$$

在 KS 阈值处，式（16）则给出 $\gamma_\ell = \ell/2$，它会随深度增大而发散。归一化普查的信噪比可直接写为 $\mathrm{SNR}_\ell = 1/\sqrt{\gamma_\ell}$。

在有限深度下，c_ℓ 的完整分布可以通过对子树进行卷积递归精确计算（附录B）。然而，这种精确计算在深度较大时会变得难以实现；更重要的是，它本身并不能提供关于 $\ell \to \infty$ 渐近行为的信息。因此，我们通过关于深度的分布递归来刻画归一化普查。令 η_1, η_2 为相互独立的边符号，且满足 $\mathbb{P}(\eta_i = -1) = \varepsilon$，于是 $\mathbb{E}[\eta_i] = \theta$。根节点的两个子节点携带着乘以这些边符号后的根自旋，并且各自作为一棵深度为 $\ell - 1$的独立子树的根。由于普查是两棵子树普查之和，归一化普查满足如下分布递归

$$c_\ell \overset{d}{=} \frac{\eta_1 c_{\ell-1}^{(1)} + \eta_2 c_{\ell-1}^{(2)}}{2\theta}, \qquad c_0 = 1, \tag{18}$$

其中 $c_{\ell-1}^{(1)}$，$c_{\ell-1}^{(2)}$ 是 $c_{\ell-1}$ 的独立副本，且与边符号相互独立。

对于 $r > 1$ $(\varepsilon < \varepsilon_c)$，$c_\ell$ 当 $\ell \to \infty$ 趋于c时收敛到极限 $\mathbb{E}[c] = 1$，其中 c的方差为 $\mathrm{Var}(= \gamma_\infty$ [30, 31]），其分布是方程 (18) 的分布不动点。实际中，我们通过群体动力学对其进行采样。一个大型样本池通过以下方式更新：将每个成员替换为方程 (18) 的右端；计算时，从当前样本池中随机抽取两个成员，并使用两个全新的边符号。从 $c_0 = 1$开始，经过 ℓ轮同步更新后，该样本池便可近似表示 c_ℓ的分布，但存在有限总体误差。条件 $r^{-\ell} \ll 1$ 控制相对方差修正；抽样误差必须另行检查。图 ε 展示了不同 2值下得到的极限分布。

的矩可由该不动点逐阶推出（见附录C）。具体而言，计算三阶累积量 κ_3 即可得到偏度

$$\frac{\kappa_3}{\gamma_\infty^{3/2}} = -\frac{2(2\theta^2-1)^{3/2}}{(4\theta^2-1)(1-\theta^2)^{1/2}} < 0. \tag{19}$$

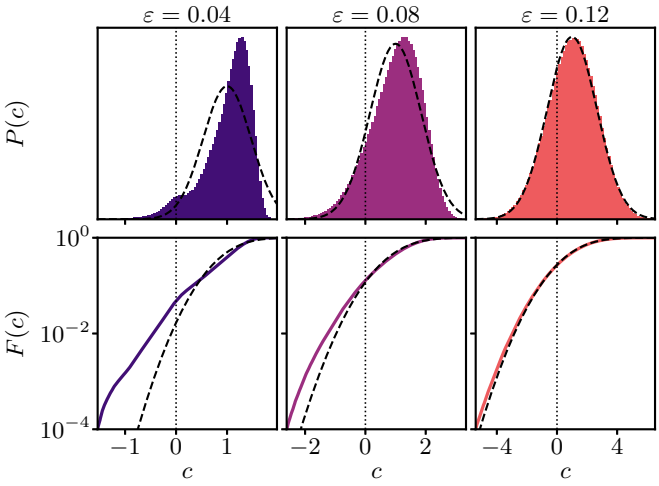

图 2. 无限深度归一化普查量的分布 c，通过对递归关系（18）进行群体动力学模拟得到（2×10^6 个样本，迭代至有效 $r^{-\ell} < 10^{-4}$，并在每次扫描后重新缩放，使经验均值为单位值）。上图：概率密度 P；下图：累积分布 F，两者均与匹配的高斯分布 $\mathcal{N}$（1，γ_∞）进行比较（虚线）。点线标示决策边界，因此 F（0）即为普查误差。左：小噪声下，高斯分布低估了误差和左尾；中：噪声较大时，分布虽不匹配，但由于累积分布在 0 处相近，普查误差仍接近其高斯对应值；右：接近 $\varepsilon_c \approx 0.146$ 时，两种分布趋于重合。

对于每个 $1 < r < 2$，偏度都严格为负，且仅在 $r \to 1$时才趋于零，即趋近 Kesten–Stigum 阈值时。因此，即使在渐近情况下，极限普查也呈现非高斯分布，其左侧权重较大（图2）。[1] 这种权重偏高源于树顶端发生的事件，例如两条根边之一翻转，详见下文。

在根分类的背景下，偏离高斯性的影响在决策边界处最为显著。图2的第二行对此作了说明：虽然分布仍然是非高斯的，但决策边界处的累积分布可能已经接近其高斯对应分布。在小 ε时，非高斯性对计数预测的影响可以精确刻画。若计数服从具有相同均值和方差的高斯分布，那么计数估计量的根预测误差为 $\Phi(-1/\sqrt{\gamma_\infty}) \leq \mathrm{e}^{-1/(2\gamma_\infty)}$，其中 Φ是标准正态累积分布函数。由于 $\gamma_\infty = 4\varepsilon + O(\varepsilon^2)$，其在小噪声下为$\mathrm{e}^{-1/(8\varepsilon)}$量级。实际的计数误差 $\mathcal{E}_{\mathrm{cens}}^{(\ell)} = \mathbb{P}(c_\ell < 0) + \frac{1}{2}\mathbb{P}(c_\ell = 0)$则为 ε量级。事实上，若两条根边中恰有一条发生翻转（其概率为 $2\varepsilon(1-\varepsilon)$），那么树的两半将携带相反的信号，此时计数就相当于抛一枚公平硬币。以

[1] 有限深度普查 c_ℓ 同样呈现非高斯特征；其精确分布见附录B。

两条根边，在此条件下可得（附录B）

$$\mathcal{E}^{(\ell)}_{\rm cens} = \varepsilon + (1-2\varepsilon)\, a_{\ell-1} \geq \varepsilon, \tag{20}$$

其中，$a_{\ell-1}$ 是两个相互独立、根节点对齐的深度为 $(\ell-1)$ 的普查之和的误差，其中包含平局概率的一半。因此，在小噪声下，普查的任何高斯描述都无法重现其误差，反而会对误差作指数级低估。正因如此，我们将为这一数据模型建立的学习理论必然需要考虑这种非高斯性，以免在小 ε下严重低估误差。

3. *Finite-depth effects*

当前数据模型（以及一般而言的基于树的数据）有一个独特之处：序列长度 N很快变得很大，而树的深度 $\ell \sim \log N$ 则相对小得多。关于 ℓ的代数修正会变成 $1/\log N$的幂，而关于ℓ 的指数修正则会变成 $1/N$的幂。普查方差呈现出后一种行为，并且在重构阈值附近收敛得越来越慢。另一个不同寻常之处是，这些修正的重要程度可能因 ε的取值而有很大差异。

最明显的例证来自归一化普查的方差，见公式（17）。其中，相对有限深度修正为

$$\frac{\gamma_\infty - \gamma_\ell}{\gamma_\infty} = r^{-\ell} = N^{-\log_2 r}, \tag{21}$$

N中的幂律，其指数为 $\log_2 r \in (0,\ 1)$，并在 Kesten–Stigum 阈值处趋于零。恰好在 Kesten–Stigum 阈值处，γ_ℓ= $\ell/2 = \frac{1}{2}\log_2 N$ 永不收敛，而普查信噪比按 $\sqrt{2/\log_2 N}$衰减。在 Kesten–Stigum 阈值附近展开，$r-1 \simeq 4\sqrt{2}(\varepsilon_c - \varepsilon)$，因此渐近区域 $\ell(r-1) \gg 1$ 只有在临界窗口之外才会达到

$$|\varepsilon_c - \varepsilon| \lesssim \frac{1}{4\sqrt{2}\,\log_2 N}. \tag{22}$$

请注意，该窗口仅在主导阶上是对称的。在此之外，普查在转变两侧表现不同：γ_ℓ 在 $r > 1$时收敛，而在 $r < 1$时则按 $r^{-\ell}$ 增长。这解释了为什么我们随后观察到转变两侧具有不同大小的有限尺寸效应。

由这一观察，我们得出结论：为了充分描述即使在较大 N 下进行的实验，我们的学习理论也需要控制某些有限尺寸效应。

III. 非高斯学习理论

现在我们转向一个经过训练、旨在预测根节点的线性分类器——即感知机。其权重 $\boldsymbol{w} \in \mathbb{R}^N$ 通过在由 P 个独立叶节点-根节点对 $\{(\boldsymbol{x}^\nu,\ y^\nu)\}_{\nu=1}^P$组成的训练集上进行经验风险最小化（ERM）得到，

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{\nu=1}^{P} V\!\left(\frac{y^\nu\, \boldsymbol{w}\cdot\boldsymbol{x}^\nu}{\sqrt{N}}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \right\}, \tag{23}$$

采用凸损失 V 和强度为 $\lambda > 0$ 的岭正则化，我们的主要示例是逻辑损失 $V(z) = \log(1+\mathrm{e}^{-z})$。随后，新树的根节点被预测为 $\hat{}x_0 = \mathrm{sgn}(\hat{}\boldsymbol{w}\cdot\boldsymbol{x})$，我们则测量测试误差 $\mathcal{E} = \mathbb{P}(\hat{}x_0 \neq y)$。

在固定维度下的无限数据极限中，式（23）中的正则化变得可以忽略，因为损失是在样本上求和的，而 $\hat{}\boldsymbol{w}$ 因此趋近于未正则化总体损失的极小值点。事实上，这个极小值点就是普查，因而它是自然的参考点。的确，交换任意节点下方的两棵子树，不会改变（$\boldsymbol{x}$，y）的联合分布，而连续进行此类交换可以将任意叶节点映射到其他任意叶节点。由于我们采用凸 V，将总体损失的一个极小值点在这些交换下取平均，可得到一个在所有叶节点上具有等权重的极小值点。因此，普查就是总体逻辑预测器；更一般地说，对于任何在零点处递减的凸损失，它都是总体损失的极小值点。[2] 现在，我们希望理解在考虑有限数量的样本时会发生什么；在这种情况下，我们自然会预期 $\hat{}\boldsymbol{w}$ 会在普查方向 $\hat{}\mathbf{e}$附近发生涨落。

A. 比例区间内的渐近结果

按照统计力学分析经验风险最小化的惯例，我们从考察渐近比例区间开始，其中 N，$P \to \infty$ 在固定 $\alpha = P/N$的条件下。通过研究配分函数，可以得到极小值点的典型性质（23）

$$Z_\beta = \int \mathrm{d}\boldsymbol{w}\, \exp\left(-\beta\Big[\sum_{\nu=1}^{P} V(z^\nu) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\Big]\right), \tag{24}$$

其中 $z^\nu = y^\nu \boldsymbol{w}\cdot\boldsymbol{x}^\nu/\sqrt{N}$为训练间隔，相应的自由能

$$f_N = -\lim_{\beta\to\infty} \frac{1}{\beta N}\langle \log Z_\beta\rangle, \tag{25}$$

其中，零温极限 $\beta\to\infty$ 选出极小化器，尖括号表示对训练样本取平均。我们使用复制方法 [29, 32], 对训练集取平均，得到 f_N

[2]对于 $0 < \varepsilon < 1/2$，叶节点与根节点正相关，因此，当 V 在零点递减时，总体损失也会沿着普查方向从 $\boldsymbol{w} = 0$ 开始下降，从而确保不会选中零向量。

在主导阶 N， $P \to \infty$ 在固定 α下，对于凸损失，采用适用的复制对称假设，并将推导留待附录D。

在标准方法中，对数据取平均很直接，因为假设输入服从高斯分布，且可能具有结构化协方差。然而，我们在此强调，数据的相关投影通常并非高斯分布。为考虑这种非高斯性，关键步骤是将权重分解为沿普查方向和垂直于普查方向的分量， $\boldsymbol{w} = \sqrt{N} m \hat{\mathbf{e}} + \boldsymbol{w}_\perp$ with $\hat{\mathbf{e}} \cdot \boldsymbol{w}_\perp = 0$，其中 $m = \mathbf{1} \cdot \boldsymbol{w}/N$是权重沿普查方向的投影。随着信号沿 $\hat{\mathbf{e}}$ 随深度增长， $y \hat{\mathbf{e}} \cdot \boldsymbol{x} = s_\ell c_\ell$ with $s_\ell^2 = r^\ell$，我们引入重新缩放后的信号重叠 $\bar{m} = s_\ell m$，当 $\ell \to \infty$时它仍为一阶量。于是，样本的间隔可精确分解为

$$z = \bar{m}\, c_\ell + \xi, \qquad \xi = \frac{y \boldsymbol{w}_\perp \cdot \boldsymbol{x}}{\sqrt{N}}. \tag{26}$$

这将携带标签的方向的非高斯性集中体现在标量 c_ℓ中，其精确分布见第II B2节。横向场 ξ的均值为零，方差为 $q_\perp = \boldsymbol{w}_\perp^\top \boldsymbol{\Sigma}_\perp \boldsymbol{w}_\perp / N$，其中 $\boldsymbol{\Sigma}_\perp = (\mathbf{I} - \hat{\mathbf{e}}\hat{\mathbf{e}}^\top)\boldsymbol{\Sigma}(\mathbf{I} - \hat{\mathbf{e}}\hat{\mathbf{e}}^\top)$ 是横向协方差，其谱为 $\rho_{\perp,\ell}$；由于 $\hat{\mathbf{e}}$ 是 $\boldsymbol{\Sigma}$的特征向量，它与 c_ℓ 不相关。我们将其近似为一个独立于 c_ℓ的高斯变量。这是该理论的核心近似。

复制对称鞍点将自由能的计算化简为一个只涉及少数序参量的优化问题。除重标度后的重叠$\bar{m}$外，这些序参量还包括重叠 $q_\perp$ 和横向权重的磁化率 χ ——它们由两个副本的重叠 $\boldsymbol{w}_\perp^{a\top} \boldsymbol{\Sigma}_\perp \boldsymbol{w}_\perp^b / N$ $Q_{a\neq b}^\perp = q_\perp$ $\chi =$ 定义

为以及 $\lim_{\beta\to\infty} \beta(Q_{aa}^\perp - q_\perp)$——以及它们的共轭变量 $\hat{q}$和 $\hat{\chi}$。由此得到的鞍点值 $f^{(\ell)}$ 具有标准形式

$$f^{(\ell)} = \operatorname*{extr}_{\bar{m}, q_\perp, \chi, \hat{q}, \hat{\chi}} \left[g_I + g_S + \alpha\, g_E \right], \tag{27}$$

其中包含相互作用项、熵项和能量项

$$\begin{aligned} g_I &= \frac{\hat{q}\chi - \hat{\chi} q_\perp}{2}, \quad g_S = \frac{\widetilde{\lambda}_\ell \bar{m}^2}{2} - \frac{\hat{q}}{2} \left\langle \frac{\mu}{\lambda + \mu\hat{\chi}} \right\rangle_{\rho_{\perp,\ell}}, \\ g_E &= \left\langle \min_z \left[\frac{(z-h)^2}{2\chi} + V(z) \right] \right\rangle_{c,\zeta}. \end{aligned} \tag{28}$$

这里，$\langle \cdot \rangle_{\rho_{\perp,\ell}}$ 表示对特征值 μ的平均，$\boldsymbol{\Sigma}_\perp$的概率密度为 $\rho_{\perp,\ell}$；而$\langle \cdot \rangle_{c,\zeta}$ 表示对普查分布和独立标准高斯变量 ζ的平均，其中$h = \bar{m} c_\ell + \sqrt{q_\perp}\,\zeta$ 是一个样本的场。由于沿$\hat{\mathbf{e}}$ 方向的信号随 s_ℓ增长，岭参数以重缩放强度 $\widetilde{\lambda}_\ell = \lambda / s_\ell^2 = \lambda r^{-\ell}$作用于普查坐标，而在横向方向上保持其原始值 λ 。对于 $r > 1$，大多数横向特征值为常数量级，但最大的特征值会随树深度增长。我们保留了对树深度 ℓ的显式依赖

~，通过重缩放岭参数 λ_ℓ、 c_ℓ 的分布以及谱$\rho_{\perp,\ell}$，然后再取它们在下文中的无限深度极限。记 $z_\star(h) = \arg\min_z [(z-h)^2/(2\chi) + V(z)]$ 为 g_E中的极小值点；对于凸的 V，该极小值点是唯一的。由驻点条件可得

$$\begin{aligned} \chi &= \left\langle \frac{\mu}{\lambda + \mu\hat{\chi}} \right\rangle_{\rho_{\perp,\ell}}, \\ \hat{q} &= \alpha \langle V'(z_\star)^2 \rangle_{c,\zeta}, \\ q_\perp &= \hat{q} \left\langle \frac{\mu^2}{(\lambda + \mu\hat{\chi})^2} \right\rangle_{\rho_{\perp,\ell}}, \\ \hat{\chi} &= \alpha \left\langle \frac{V''(z_\star)}{1 + \chi V''(z_\star)} \right\rangle_{c,\zeta}, \\ \widetilde{\lambda}_\ell\, \bar{m} &= -\alpha \langle c_\ell V'(z_\star) \rangle_{c,\zeta}. \end{aligned} \tag{29}$$

假设新样本的间隔也服从分布 h ，并具有横向高斯涨落，则对于$q_\perp > 0$ ，预测测试误差可由鞍点方程得出：

$$\mathcal{E} = \left\langle \Phi\left(-\frac{\bar{m} c_\ell}{\sqrt{q_\perp}} \right) \right\rangle_c. \tag{30}$$

矩匹配高斯普查分布相当于取 $c_\ell \sim \mathcal{N}(1, \gamma_\ell)$, 此时，方程（29）化为具有相同协方差的高斯混合分布的方程 [33]（附录 D）。下文中，我们将方程（29）的鞍点称为混合理论。该鞍点将普查分布的精确分布律与高斯横向平均结合起来。

对于 $1 < r < 2$ 且固定 α 时，三个依赖深度的~输入收敛到 $c_\ell \to c$、 $\rho_\perp$、$\ell \to \rho$ 和 $\lambda_\ell \to 0$，由此定义无限深度自由能 $f = \lim_{\ell\to\infty} f^{(\ell)}$，其中公式 (29) 的最后一式变为 $\langle c\, V'(z_\star) \rangle_c$，$\zeta = 0$。实际上，我们通过总体动力学对 c 的不动点分布律进行采样，如图 2 所示，并在得到的总体上计算关于 c 的平均值。谱平均只涉及有界预解式，因此即使 ρ 的矩可能发散，它们仍会收敛，正如第II.B.1节所预期的那样。我们将它们截断在水平 K，其误差关于 K 呈指数级小（附录 E）。由于完整谱和横向谱仅在其最大特征值 μ_ℓ的权重上有所不同，因此这一截断谱为二者共有，也与其极限 ρ 相同。

不同取值下的渐近测试误差 α， ε， λ 如图所示。3此外，我们还展示了等价的高斯鞍点，它是通过用 $\mathcal{N}(1,\ \gamma_\infty)$ 替换普查分布得到的。正如上文关于普查分布非高斯性的讨论所预期的那样，在 Kesten–Stigum 阈值附近，高斯近似与非高斯结果一致；但在小噪声下，它会将误差低估几个数量级。相比之下，非高斯鞍点在两端都与普查误差接近，$\alpha \to 0$ 并且 $\alpha \to \infty$。在较小的岭参数下，误差在两者之间呈非单调变化，并在 α 处达到最大值，其量级为一到十；随着 α 减小，该最大值向更大的 ε 移动。这个最大值与训练权重拟合横向噪声方向有关

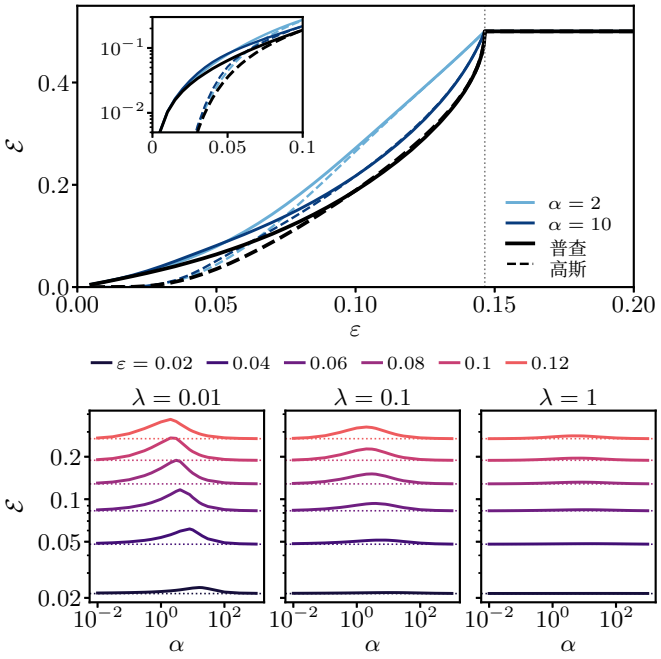

图3. 副本理论无限深度极限预测的测试误差。上图：当 , 时，误差随 ε $\lambda =$ 0.01 $\alpha = 2$ 10 的变化；图中比较了非高斯鞍点（实线）和高斯鞍点（虚线）的结果，以及普查误差（黑色实线）和其高斯对应分布 $\Phi(-1/\sqrt{\gamma_\infty})$（黑色虚线）。点线标出 ε_c，超过该值后，在无限深度下叶节点不再携带关于根节点的信息。插图以对数尺度展示了 $\varepsilon \le 0.1$ 时相同的曲线。下图：误差随 α 的变化，展示了三个岭参数 λ 取值及若干个 ε 取值时的结果，并以点线标出相应的普查误差。

，类似于样本级双下降 [34]，并且可以通过增大正则化 [35]来平滑。对于有限的 λ，我们没有发现明确的插值阈值。

起初，误差在 $\alpha \to 0$ 时趋近于普查基准，这令人惊讶，也表明存在一个次比例区间。事实上，由于首先取极限 $\ell \to \infty$， $\alpha \to 0$ 仍对应于发散的样本数 $P = \alpha N$，这意味着远少于 N 个样本就已经足以使误差接近普查误差。在探讨该区间之前，我们先考察在鞍点中保持有限 ℓ 的影响。

B. 有限深度树

到目前为止，我们研究了无限深度渐近行为。然而，正如第IIB 3节所述，我们预计这些结果无法很好地描述有限深度N 实验，因为层级结构会带来显著的有限深度效应。

必须区分三种效应：混合模型对深度的依赖、其大-N复制鞍点的修正，以及混合模型与基于横向涨落的高斯假设所得到的广播数据之间的差异。我们对第一种效应进行了量化。下面的计算并未确定复制近似中精确的 $O(1/N)$ 余项，也未给出横向场高斯近似的收敛速率。

我们从协方差谱入手。首先注意到，式（29）中谱平均的被积函数随 μ增加，但会趋于饱和，

$$0 \le \frac{\mu}{\lambda + \mu\hat{\chi}} \le \frac{1}{\hat{\chi}}, \qquad 0 \le \frac{\mu^2}{(\lambda + \mu\hat{\chi})^2} \le \frac{1}{\hat{\chi}^2}, \tag{31}$$

因此，大特征值的贡献由其在谱中的权重决定，而非其大小。此外，$\rho_{\perp,\ell}$ 恰好是截断于高度 ℓ的极限概率密度ρ，因此二者自然共享特征值 μ_k 及其权重 2^{-k} （适用于 $k \le \ell$），区别仅在于质量 $2^{-\ell}$，它将质量分配给较大的特征值μ_k，$k > \ell$。结合饱和效应，有限深度谱平均与其极限值的差异至多为 $2^{-\ell}\hat{\chi}^{-j} = O(1/N)$， $j = 1$，2（附录 E）。因此，在固定的正值$\hat{\chi}$下，用极限谱替代有限深度谱会在这些谱平均中引入 $O(1/N)$阶误差，与预期的在鞍点处计算自由能所产生的修正处于同一阶。请注意，当 $\alpha \to 0$时，这一结论不再成立，因为$\hat{\chi} \propto \alpha$ [式（29）]，，正如我们将在第III C 节看到的那样。

现在我们转向普查平均值。由于无限树是深度为 ℓ 的树，其每个叶节点都以一棵独立的无限子树为根，因此我们可以在同一棵树上写出 $c = c_\ell + \delta_\ell$，其中 δ_ℓ 汇集了这些子树的涨落，且 $\mathbb{E}[\delta_\ell \mid c_\ell] = 0$ 和 $\mathbb{E}[\delta_\ell^2] = \gamma_\infty - \gamma_\ell = \gamma_\infty r^{-\ell}$，因此在任何光滑平均中，一阶项都会抵消，展开到二阶可得

$$\left|\langle g(c)\rangle_c - \langle g(c_\ell)\rangle_c\right| \le \tfrac{1}{2}\|g''\|_\infty \gamma_\infty r^{-\ell}, \tag{32}$$

对于任意二次可微且二阶导数有界的 g，上述结论均成立。对于逻辑损失，这适用于式（29）中的普查平均值以及式（30）中的测试误差。因此，普查会使这些平均值产生 $r^{-\ell} = N^{-\text{对数}^2\, r}$ 阶的修正，重缩放岭参数 $\tilde{\lambda}_\ell = \lambda r^{-\ell}$ 也是如此。这些估计表明，对于 $1 < r < 2$ 且 $\alpha > 0$ 固定时，

$$f^{(\ell)} = f + O(r^{-\ell}), \tag{33}$$

预测测试误差也具有相同的阶数。这里的深度修正至关重要，因为当 $r < 2$ 时，它们比 $O(1/N)$ 谱修正和鞍点修正衰减得更慢。通过使用有限深度输入 c_ℓ, $\rho_{\perp,\ell}$ 和 λ_ℓ 求解~方程（29），即可得到下文所用的有限深度鞍点。

我们在图 4 中将有限深度鞍点与广播树上的经验风险最小化进行了比较；数值实现的详细信息见附录 F。在所有深度下，两者都具有非常好的一致性，包括超过

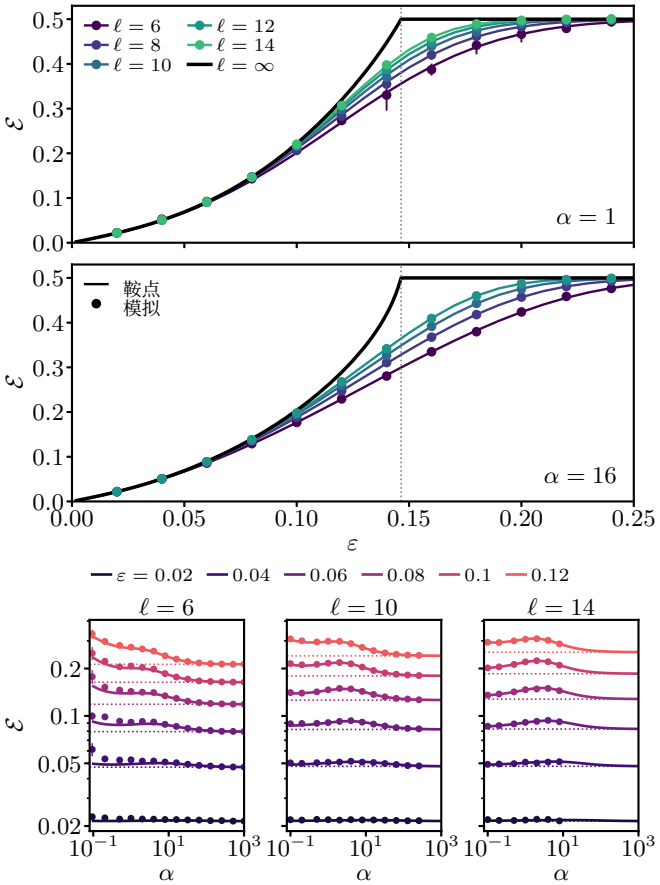

图4. 有限深度树上逻辑回归的测试误差，$\lambda = 0.1$。上方两个面板：误差随 ε 变化的曲线，分别对应 $\alpha = 1$ 和$\alpha = 16$，深度为 $\ell = 6$–14，结果来自有限深度混合鞍点（曲线）和广播树上的经验风险最小化（点，近似95%区间，见附录F），并与无限深度鞍点（黑色）比较。虚线标记 $\varepsilon_{c\infty}$。下方：误差随 α 变化的曲线，深度分别为 $\ell = 6$、10和14，对应若干 ε 取值（颜色同图3），结果来自有限深度鞍点（曲线）和模拟（点），并以虚线标出相应的有限深度普查误差。

阈值的情况，此时误差仅随 $1/2$ 缓慢趋近于 ℓ。请注意，这种一致性在训练后的分类器的序参量层面同样成立，参见附录G。有限深度混合鞍点在阈值两侧都保持良定义。限制条件$r > 1$ 是归一化普查具有前述收敛分析中所使用的有限方差的无限深度极限所必需的。相比之下，正如我们在第 ε_c 节 II B3 中预期的那样，对于接近 ε_c 的可达深度树，无限深度鞍点无法进行描述。图 4 的底行展示了相应的学习曲线，其为α的函数。在较小深度下，误差现在会随 α 单调下降，这与标准的学习图景一致。在我们考察的中等深度下，对于 ε 接近 ε_c 的情况，最小误差在 $\alpha \to \infty$ 时达到，而不是在 $\alpha \to 0$ 时同样达到；这同样符合传统学习场景。

C. 次比例区间

到目前为止，我们观察到，在无限深度下，当 $\alpha \to 0$ 时，经验风险最小化的误差已经与普查误差相匹配。这一观察表明，只需使用比例趋于零的样本，就能学得普查方向，$P \ll N$。即使在有限 ℓ 的情况下，普查准确率似乎也能在较小的 α 时迅速达到。现在，我们将精确确定这一现象发生的尺度，以及有限树深度在此情境下的影响。

为了确定学习的尺度，我们首先考虑训练集的带符号均值，

$$\hat{\boldsymbol{v}} = \frac{1}{P}\sum_{\nu} y^{\nu}\boldsymbol{x}^{\nu}. \tag{34}$$

每个训练样本都通过其带符号的叶节点 $y^{\nu}\boldsymbol{x}^{\nu}$对 $\hat{\boldsymbol{v}}$ 作出贡献。沿信号方向，这一贡献是第 II B 2 节中的带符号普查量 $y^{\nu}\hat{\mathbf{e}} \cdot \boldsymbol{x}^{\nu} = s_\ell c_\ell^{\nu}$，其均值为 s_ℓ。这些贡献在训练集上相干叠加；将普查量的经验平均值$\bar{c}_P = P^{-1}\sum_\nu c_\ell^{\nu}$（均值为 1，方差为 γ_ℓ/P）记为——$\hat{\boldsymbol{v}}$的重叠度恰好为 $\bar{m} = s_\ell^2\bar{c}_P/\sqrt{N}$。当$P \gg \gamma_\ell$ 时，该平均值集中于 $\bar{c}_P \simeq 1$，因此重叠是确定的 $\bar{m} \simeq s_\ell^2/\sqrt{N}$。另一方面，在横向方向上，这些贡献的均值为零，协方差为 $\boldsymbol{\Sigma}_\perp$，因此它们的平均值$\hat{\boldsymbol{v}}_\perp$ 的协方差为 $\boldsymbol{\Sigma}_\perp/P$。因此 $\mathbb{E}[q_\perp] = \mathrm{Tr}\ \boldsymbol{\Sigma}_\perp^2/(NP)$ $= B_\ell/P$，其中期望是对训练集取的，并且

$$B_\ell = \frac{1}{N}\mathrm{Tr}\,\boldsymbol{\Sigma}_\perp^2 = \sum_{k=1}^{\ell-1} 2^{-k}\mu_k^2 + 2^{-\ell}\mu_\ell^2 \tag{35}$$

是横向谱的二阶矩。如果 $\bar{c}_P$和 $q_\perp$ 集中在这些值附近，并且如第 III A 节所述，横向场服从高斯分布且与普查相互独立，那么式（30）的测试误差取决于比值 $\bar{m}^2/q_\perp \simeq Ps_\ell^4/(NB_\ell)$。对于 $1 < r < 2$，这表明当该比值为数量级为一时，开始出现相对于随机猜测的数量级为一的优势（$1/2 - \mathcal{E}$），这对应于数量级为

$$P_\star = \frac{NB_\ell}{s_\ell^4}. \tag{36}$$

.....的行为 B_ℓ 源于第 II B1 节的谱分析。公式（35）中的各项重要地按（$r^2/2$）^{k}进行缩放，因此当 B_ℓ 时仍保持有限，其极限为 $r < \sqrt{2}$，其中极限 $B_\infty = (4 - r^2)/[2(2 - r^2)]$。恰好在 $r = \sqrt{2}$ 时，$B_\ell \asymp \ell$这意味着每个足够大的谱层级都贡献一个数量级为一的量，因此发散会沿层级结构累积。最后，当 $\sqrt{2} < r < 2$ 时，$B_\ell \asymp r^{2\ell}/N$ 由最大特征值主导。将这些结果与 $s_\ell^2 = r^\ell$ 和 $N = 2^\ell$的表达式结合起来，样本尺度因此表现为

$$P_\star \asymp \begin{cases} N^{\omega(\varepsilon)}, & 1 < r < \sqrt{2}, \\ \log N, & r = \sqrt{2}, \\ 1, & \sqrt{2} < r < 2, \end{cases} \qquad \omega(\varepsilon) = 1 - 2\log_2 r, \tag{37}$$

其中 $r=\sqrt{2}$对应于 $\varepsilon_\star = \left(1-2^{-1/4}\right)/2 \approx 0.080$。换句话说，当从下方趋近 KS 阈值时，该指数趋于 $\omega = 1$；当趋近于 $\varepsilon_\star$时，则趋于 $\omega = 0$，此时边际尺度为 $\log N$。在 $\varepsilon_\star$ 以下，$P_\star$ 保持在一的数量级。在 KS 阈值处，普查误差趋于 $1/2$，因此形式上的尺度 $P_\star \asymp N$ 并不意味着非平凡的重构。样本尺度会随着层级相关性的强度而连续调节。请注意，这一尺度并非一个基本极限，因为普查本身不需要训练。更确切地说，它表示使用本文所考虑的估计量找到普查方向所需的样本数量，而这些估计量对该方向没有先验偏好。如下文所示，它不取决于所采用的具体损失或正则化。

有趣的是，$\varepsilon_\star$ 本身可以被视为一个 KS 阈值。这是因为，将两棵相互独立的广播树逐节点相乘，并使每棵树都与各自的根节点对齐，所得结果本身也是一棵广播树。它在根节点处等于 $+1$，并且当且仅当两棵树中恰有一棵沿某条边翻转时才会沿该边翻转，即概率为$\varepsilon' = 2\varepsilon(1-\varepsilon)$，因此 $\theta' = \theta^2$ 和 $r' = r^2/2$。带符号均值正是依靠这类乘积来对测试树进行分类的。其间隔（式 (26) 中令 $\boldsymbol{w} = \hat{\boldsymbol{v}}$）确实与 $\sum_\nu \sum_i u_i^\nu$ 成正比，其中 $u_i^\nu = (yx_i)(y^\nu x_i^\nu)$ 是由测试树和训练树 ν 形成的重叠树的叶节点，因此它等于各个 P 重叠树的普查之和。重叠树的成对相关性是原树对应相关性的平方，即 $\mathbb{E}[u_i u_j] = \mathbb{E}[x_i x_j \mid y]^2$，因此，按原树的定义方式定义的归一化普查的方差为$\gamma'_\ell = P_\star + \gamma_\ell(2+\gamma_\ell)$，也就是说，$P_\star$，相差一个数量级为一的项。因此，样本尺度 $P_\star$ 就是为分辨其普查的单位均值而必须取平均的重叠树数量；式 (37) 中的三个区间对应于重叠树自身的 KS 阈值 $r' = 1$两侧的情形，即 $r = \sqrt{2}$。对于 $\varepsilon > \varepsilon_\star$，因此，单个训练样本所携带的信息会随着 $\ell \to \infty$ 而趋于零；而对于$\varepsilon < \varepsilon_\star$，无论深度如何，其分类效果都优于随机猜测，误差等于重叠树的普查误差。

对于 $\varepsilon_\star < \varepsilon < \varepsilon_c$，在这一尺度下，式（34）中带符号均值的学习曲线可以精确求得。令$P/P_\star \to \tau \in (0,\ \infty)$ 为 $\ell \to \infty$，将上文在式（$\bar{m}$ 和 $q_\perp$）中推导出的数值代入式（30），可得

$$\mathcal{E}_{\text{on}}(\tau) = \left\langle \Phi(-\sqrt{\tau}\, c) \right\rangle_c . \tag{38}$$

请注意，在此区域内，高斯普查结果$\Phi(-\sqrt{\tau/(1+\gamma_\infty \tau)})$在数值上与这一预测非常接近，因为只有在$\Phi$较小时，非高斯效应才会变得显著。

此时 ε 较小。

对于 $\varepsilon_c < \varepsilon < 1/2$，式（36）中的尺度则给出 $\tau = P/P_\star \propto \alpha r^{2\ell} \to 0$，其中 α 固定。在带符号均值近似下，有限深度表达式 $\langle \Phi(-\sqrt{\tau} c_\ell) \rangle$ 随后预测出量级为 $\sqrt{\varepsilon} \propto \sqrt{\alpha}\, r^\ell$ 的优势。接近 Kesten–Stigum 阈值时，这一优势缓慢消失，与图4中可见的有限深度优势一致。式（38）采用具有有限方差的极限单位均值普查，因此不适用于 $r \le 1$。

值得注意的是，第IIIB 节的有限深度鞍点

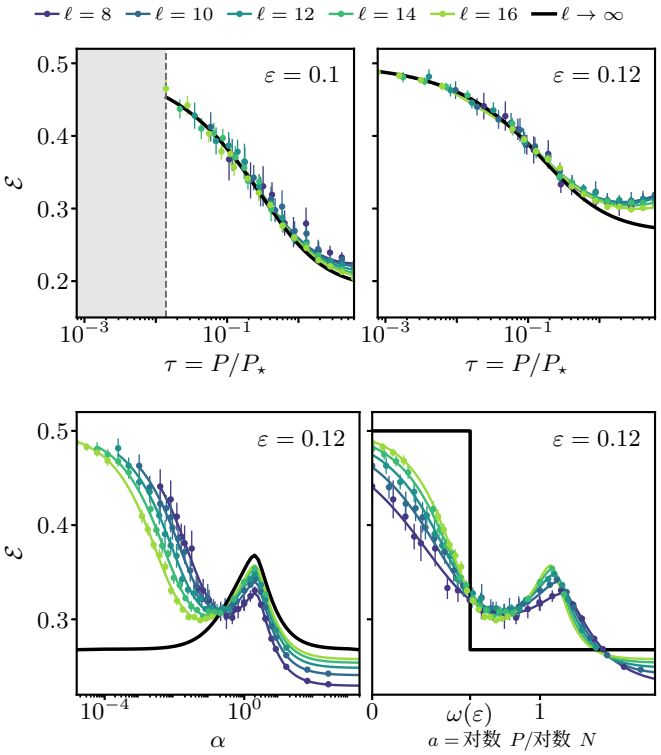

图5. 在 $\lambda = 0.01$、深度 ℓ =8–16 时的次比例区间：基于广播树的模拟结果（圆点，近似 95% 区间，见附录F）与有限深度鞍点（曲线）。上图：在 $\tau = P/P_\star$ 处，测试误差随 $\varepsilon = 0.1$（左）和 $\varepsilon = 0.12$（右）的变化，并与起始极限$\mathcal{E}_{\text{on}}(\tau)$（黑色）比较。左图中，阴影区域低于在 $\ell = 16$处、使用单个训练样本时 τ 可达到的最小值。下图：在 $\varepsilon = 0.12$ 处，测试误差随 $\alpha = P/N$（左）的变化，黑色曲线为无限深度鞍点；以及随 $a = \log P/\log N$（右）的变化，黑色曲线为$a = \omega(\varepsilon)$处的无限深度阶跃，从 $1/2$ 到极限普查误差。比例区间的峰值在此极限下收缩为单点$a = 1$，因此未绘出。

在此次比例区间内，对经验风险最小化有相关预测。对于固定的 $1 < r < \sqrt{2}$、$\lambda > 0$ 和 $P = \tau P_\star$，且 $\tau > 0$ 固定时，样本比率 α 趋于零。为了研究这一范围，我们因此在 α 的主导阶求解鞍点方程（29），并将每个序参量保留在其首个非零阶。共轭变量 $\hat{q}$ 和 $\hat{\chi}$ 的阶为 α。由于 ρ 的前两个矩有限，利用控制收敛定理，可在谱平均中取极限，从而得到 $\chi \simeq \langle \mu \rangle_{\rho_\perp,\ell}/\lambda = 1/\lambda$ 为一阶量，而 $q_\perp \simeq \hat{q} B_\ell/\lambda^2$ 的阶为 α。只要 $\bar{m}$ 也足够小，腔场 $h = \bar{m} c_\ell + \sqrt{q_\perp}\,\zeta$ 就很小。在主导阶下，g_E 中满足 $z_\star = h - \chi V'(z_\star)$ 的极小值点 $z_\star$，因此对所有样本都具有相同的间隔 z_0，它是 $\lambda z_0 = -V'(z_0)$ 的解。于是，对 c_ℓ 和 ζ 的平均可以在 $\hat{q}$ 和 $\bar{m}$ 的方程中进行，

$z_\star = z_0$，得到

$$\hat{q} \simeq \alpha V'(z_0)^2, \quad \widetilde{\lambda}_\ell \, \bar{m} \simeq -\alpha V'(z_0)\langle c_\ell \rangle = \alpha |V'(z_0)|, \tag{39}$$

利用 $\langle c_\ell \rangle = 1$ 以及 V 的单调递减性。 $\bar{m}$ 的方程中，保留有限深度至关重要，因为方程两边都很小。右边来自 αg_E 中的损失，其阶为 α；左边来自 g_S 中的正则化，带有趋于零的重缩放岭参数 $\widetilde{\lambda}_\ell = \lambda r^{-\ell}$。两者平衡后得到 $\bar{m} \simeq \alpha r^\ell |V'(z_0)|/\lambda$，再结合 $q_\perp \simeq \alpha V'(z_0)^2 B_\ell/\lambda^2$，可得

$$\frac{\bar{m}^2}{q_\perp} \simeq \frac{\alpha r^{2\ell}}{B_\ell} = \frac{P}{P_\star} = \tau, \tag{40}$$

其中，损失和岭参数均消去了，因为它们只通过公共因子 $|V'(z_0)|/\lambda$ 进入。这也验证了起始尺度上的 $\bar{m} = O(r^{-\ell}) \to 0$ 。因此，鞍点近似预测出与带符号均值相同的领先阶测试误差， $\mathcal{E}_{\rm on}(\tau)$。再次强调，只有保留有限深度输入$\widetilde{\ }$，这一数据区间才是可达的：在固定 $\lambda_\ell = 0$ 下取 α，如图 3 的无限深度极限所示，在令 $\alpha \to 0$ 之前，会丢弃信号与重缩放正则化之间的平衡，而正是这种平衡导致了起始现象。

图 5 在数据稀缺区间检验了这些预测，并与数值实验结果进行了对比。当以 τ （顶行）为横轴绘图时，模拟在较小 τ下趋近于起始曲线。在较大的 τ下，随着经验风险最小化器开始拟合训练集，其结果最终会稳定在起始曲线之上，且早于比例区间中的过拟合峰。此外，实验表明，直到样本比 α 达到 10^{-4} 量级时，经验风险最小化器都能由有限深度鞍点很好地描述。作为指数 $a = \log P/\log N$ 的函数，有限深度 ℓ 误差缓慢趋近于 $a = \omega(\varepsilon)$ 所预测的尖锐阶跃，即 $\ell \to \infty$，此处达到普查准确率；随后一直到 $a = 1$ 的比例区间之前，保持在普查平台上。

对于 $0 < \varepsilon < \varepsilon_\star$，不同的是， $P_\star$ 会随 ℓ 增长而保持有限。这表明存在一个用于分辨信号的量级为一的样本尺度，而非一个能够稳定恢复普查方向或其准确度的尺度。在 P固定时，训练普查和占主导地位的横向模态仍会发生涨落，因此，前述自平均起始公式并不适用。对这一情形的描述必须保留训练集涨落，不过仍有可能研究其分布和平均误差。边缘情形 $r = \sqrt{2}$ 同样需要对数样本尺度的单独分析。

总的来说，对于 $\varepsilon_\star < \varepsilon < \varepsilon_c$ 且岭参数足够小并保持固定时，理论预测了以下学习
大深度下的曲线，如图6所示。在 $P \sim P_\star$附近，误差从 $1/2$ 下降至普查误差，遵循 $\mathcal{E}_{\rm on}$。在比例区间 $P = \alpha N$ 中，对横向涨落的拟合随后产生图3中的峰值，其位置在 α 处，量级为一；随着 λ 增大，该峰值会变得平滑，

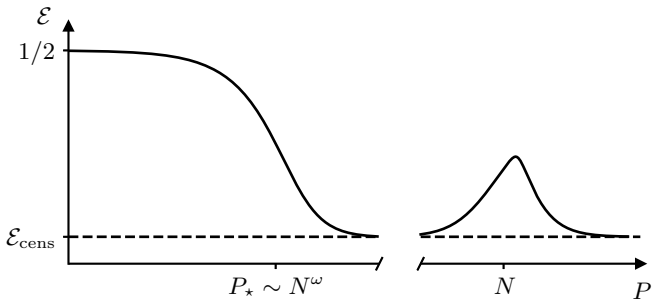

图6. 大深度下的学习曲线，针对 $\varepsilon_\star < \varepsilon < \varepsilon_c$，显示为样本数 P 的对数尺度函数。在 $P_\star \sim N^\omega$附近，误差从 $1/2$ 下降至普查误差，遵循方程（38）的起始极限；随后，误差在普查平台上保持不变，直至进入比例区间 $P \sim N$。在此区间内，对横向涨落的拟合会产生一个依赖于正则化的峰值，之后普查误差再次恢复。这两部分分别是其自身尺度上的 $\ell \to \infty$ 极限，分别对应于 $\tau = P/P_\star$ 和 $\alpha = P/N$，此处取值为 $\varepsilon = 0.12$ 和 $\lambda = 0.01$。坐标轴的断裂隐藏了平台，其范围随 $N^{1-\omega}$增长。

之后，普查误差在 $\alpha \to \infty$处恢复。因此，学习曲线呈现双下降，但两个下降发生在彼此分离的尺度上：第一次下降仅由数据统计量决定，发生在 $P \sim N^\omega$处，远低于 $P \sim N$处的峰值；后者则取决于正则化，尤其取决于在 λ 还是 λ/α 保持不变，而 P 不断增大。两个尺度之间，误差处于普查平台，即图 $\alpha \to 0$ 3所示的极限。然而，这两个尺度之间的间隔仅随 $N^{1-\omega}$增长。在可达到的深度下，图5中可以清楚看到第一次下降、峰值和第二次下降，有限深度鞍点对它们作出了准确描述；但在两次下降之间，误差仅降至高于普查误差的一个局部最小值，而非形成平台。接近 Kesten–Stigum 阈值时， $\omega \to 1$ ，且起始点与比例区间合并，此时第IIIB 节中的有限深度效应开始起主导作用。

IV. 讨论

在这项工作中，我们运用统计力学工具研究了一种可解析求解的层次化数据模型。基于树的广播数据模型十分简单，使我们得以针对线性分类器的学习曲线，构建并检验一种混合理论，适用于样本数量从低于比例增长到与维度成比例增长的情形；该理论考虑了非高斯性、有限尺寸效应，以及由层级结构直接导致的样本复杂度降低。我们认为，这是朝着以统计力学描述如何从内在具有非高斯性和结构化特征的数据中学习迈出的第一步，体现了 M´ezard [10] 所提出的研究纲领的精神。

在较小的 ε下，要获得准确描述，就必须保留普查的非高斯分布律

鞍点方程中的这一事实可能令人惊讶，因为已有大量研究结果证明了高维中高斯等价性的稳健性 [36–40]。此类高斯等价性结果都要求满足模型特定假设，包括控制沿相关方向的投影以及适当的集中性质。它们并不意味着任意相关输入都具有普适性。在可重构广播模型中，yˆ $\mathbf{e} \cdot \boldsymbol{x}/s_\ell = c_\ell$ 即使在无限深度下仍然是非高斯的，仅匹配输入的均值和协方差也无法重现其分类误差。这与参考文献中讨论的普适性失效机制一致。 [41, 42]我们的混合理论保留沿ˆ**e**方向的精确分布，并将横向场视为独立高斯变量，这类似于参考文献中的渐近描述 [42, 43]以及条件高斯等价性 [44, 45]。我们的设定并未被这些结果直接涵盖，而且我们尚未证明横向场在此处会变成高斯变量。更一般地说，这揭示了真正层次化数据所面临的一个核心矛盾：高斯描述保留了层次化协方差，却抹去了层级结构在与任务相关的少数方向上留下的非高斯统计量。我们提出的混合理论通过分离出这些方向并将其余部分视为高斯变量，在不牺牲可解性的前提下，为解决这一矛盾提供了一种途径。

从这个意义上说，广播模型尽管形式简单，而且具有随维度增长的异常强信号，却提供了一个可检验这一方法并将其推向更复杂架构的初步场景。在线性经验风险最小化于大样本比下收敛到的普查中，其误差可能大于置信传播的贝叶斯最优预测，从而为研究更深层架构如何逼近贝叶斯最优性留下空间。逐阶学习目标的随机特征模型 [46], 将是自然的第一步。一种可能性是，对普查的非线性修正需要更高次幂的N 样本，从而产生一个进一步且彼此分离的、朝向贝叶斯错误率的下降过程。要确定这种情况是否会发生，需要考察所选特征映射的谱权重和目标对齐性；这并不能仅由线性分析推出。真正的特征学习，例如通过逐层训练 [15], 或与树相匹配的卷积架构，反而可能使这第三次下降在更少的样本下发生。不过需要注意的是，在该模型中，置信传播相对于普查的增益很小，这意味着第三次下降可能难以在数值实验中观测到，反而需要一种渐近理论。

另一个可能从这种简单的层级数据模型中受益的方向，是自监督学习和生成建模。的确，值得探究根节点所携带的强信号是否会使新序列的生成相对容易，抑或树内微弱的非高斯相关性反而难以复现。生成式扩散已在层级数据上得到研究，但研究大多聚焦于生成动力学 [47] ，并通过数值实验探讨泛化与训练集记忆化之间的相互作用 [48, 49]。在我们的语境中，广播模型上的离散扩散（例如采用文献 [50], 的方法）似乎更有希望为这些问题提供解析结果，而不仅限于高斯数据的连续扩散 [51, 52]。特别是，它可能为文献 [49], 中有偏泛化现象提供一个可解的实例。该文提出的假设是：由于数据集规模有限，高阶统计量未能被正确解析，从而造成模型偏向训练数据；按其构造，这一机制要求数据为非高斯数据。

致谢

J.G.-B. 的工作获得了欧盟地平线欧洲计划 Marie Sk-lodowska-Curie 资助，资助协议编号为 101210798。我们感谢 H. Tabanelli 和 L. Gianmanco 提供的宝贵讨论。我们承认在整个研究过程中使用了大型语言模型，尤其用它们编写数值代码和图表脚本、协助推导，并为手稿修改提出建议。所有结果均经作者核查；作者对内容承担全部责任，并撰写了最终文本。

[1] E. Gardner，《神经网络模型中的相互作用空间》，《物理学杂志A：数学与一般》**21**，257（1988）。[2] E. Gardner 和 B. Derrida，《神经网络模型的最优存储性质》，《物理学杂志A：数学与一般》**21**，271（1988）。[3] E. Gardner 和 B. Derrida，《关于网络最优存储容量的三篇未完成作品》，《物理学杂志A：数学与一般》**22**，1983（1989）。[4] H. S. Seung、H. Sompolinsky 和 N. Tishby，《从样例中学习的统计力学》，《物理评论A》**45**，6056（1992）。[5] A. Engel 和 C. Van den Broeck，*Statistical Mechanics of Learning* 2001（剑桥大学出版社，）。[6] M. Gabri´e、S. Ganguli、C. Lucibello 和 R. Zecchina，《神经网络：从感知机到深度网络》，载于 *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking After 40 Years* 2023 477（世界科学出版社，）第 – 497 页。[7] J. Barbier、F. Camilli、M.-T. Nguyen、M. Pastore 和 R. Skerk，《深度学习的统计物理：插值附近多层感知机的最优学习》，

《物理评论X》**16**, 031014 (2026). [8] V. Erba、E. Troiani、L. Biggio、A. Maillard 和 L. Zde- borov´a, 《双线性序列回归：一种从高维词元长序列中学习的模型》, 《物理评论X》**15**, 021092 (2025). [9] F. Boncoraglio、E. Troiani、V. Erba 和 L. Zdeborov´a, 《注意力索引模型的贝叶斯最优学习》, 《神经信息处理系统进展》**38**, 105029 2025( ). [10] M. M´ezard, 《自旋玻璃理论及其新挑战：结构化无序》, 《印度物理学杂志》**98**, 3757 (2024). [11] S. Goldt、M. M´ezard、F. Krzakala 和 L. Zdeborov´a, 《建模数据结构对神经网络学习的影响：隐流形模型》, 《物理评论X》**10**, 041044 (2020). [12] F. Cagnetta、L. Petrini、U. M. Tomasini、A. Favero 和 M. Wyart, 《深度神经网络如何学习组合数据：随机层次模型》, 《物理评论X》**14**, 031001 (2024). [13] J. Garnier-Brun、M. M´ezard、E. Moscato 和 L. Sagli- etti, 《Transformer 如何学习结构化数据：层级滤波的启示》, 载于 *International Conference on Machine Learning* 2025 18831 18847 (PMLR,) 第 – 页。[14] S. Mei, 《U-Net 即置信传播：生成式层级模型中的高效分类、去噪与扩散》, arXiv 预印本 arXiv:2404.18444 (2024). [15] Y. Ren、Y. Dandi、F. Krzakala 和 J. D. Lee, 《随机层级模型的可证明学习及层级浅层到深层的串联》, 载于 *Conference on Learning Theory (COLT)* 2026 2601 19756 ( ) arXiv: . . [16] J. Gaitonde、F. Koehler、E. Mossel、J. Shin 和 A. Sly, 《具有可预测缩放定律和可证明推理优势的层级语言模型》, arXiv 预印本 arXiv:2605.13687 (2026). [17] Z. Wang、E. Nichani 和 J. D. Lee, 《用三层神经网络学习层级多项式》, arXiv 预印本 arXiv:2311.13774 (2023). [18] H. Fu、Z. Wang、E. Nichani 和 J. D. Lee, 《用三层网络学习多个非线性特征的层级多项式》, arXiv 预印本 arXiv:2411.17201 (2024). [19] Y. Dandi、L. Pesce、L. Zdeborov´a 和 F. Krzakala, 《深度的计算优势：用梯度下降学习高维层级函数》, 载于 *Advances in Neural Information Processing Systems* (2025) arXiv:2502.13961。[20] L. Defilippis、F. Krzakala、B. Loureiro 和 A. Maillard, 《层级多指标模型学习中的最优缩放定律》, arXiv 预印本 arXiv:2602.05846 (2026). [21] H. Tabanelli、Y. Dandi、L. Pesce 和 F. Krzakala, 《通过层级谱方法深度学习组合目标》, arXiv 预印本 arXiv:2602.10867 (2026). [22] A. Wortsman-Zurich、H. Tabanelli、Y. Dandi、F. Krza- kala 和 B. Loureiro, 《序贯特征恢复中的缩放定律：一个可解层级模型》, arXiv 预印本 arXiv:2605.14567 (2026). [23] Y. Dandi、M. Vilucchio、L. Arnaboldi、H. Tabanelli 和 F. Krzakala, 《深度学习即神经低阶滤波：层级特征学习的谱理论》, arXiv 预印本 arXiv:2605.13612 (2026). [24] P. M. Bleher、J. Ruiz 和 V. A. Zagrebnov, 《关于贝特格点伊辛模型极限吉布斯态的纯度

，统计物理学杂志 **79** ，473

(1995). [25]W. Evans、C. Kenyon、Y. Peres 和 L. J. Schulman, 《树上的广播与伊辛模型》, 应用概率年刊 10, 410 (2000). [26] E. Mossel 和 Y. Peres, 《树上的信息流》, 应用概率年刊 13，817 (2003). [27] E. Mossel, 《综述：树上的信息流》, 载于 *Graphs,Morphisms and Statistical Physics*, 离散数学与理论计算机科学 DIMACS 系列，第 63 卷（美国数学会，2004），第 155–170 页。[28] M. M´ezard 和 A. Montanari, 《树上的重构与自旋玻璃转变》, 统计物理学杂志 124，1317 (2006). [29] M. M´ezard 和 A. Montanari, , ,（牛津大学出版社，）。[30] H. Kesten 和 B. P. Stigum, 《多维 Galton–Watson 过程的极限定理》, 数学统计年刊 37，1211 (1966). [31] J. D. Biggins, 《分枝随机游走中的鞅收敛》, 应用概率杂志 14，25 (1977). [32] M. M´ezard、G. Parisi 和 M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications* 9 ，第 卷（世界科技出版公司，1987）。[33] B. Loureiro、G. Sicuro、C. Gerbelot、A. Pacco、F. Krza- kala 和 L. Zdeborov´a, 《利用广义线性模型学习高斯混合模型：高维情形下的精确渐近》, 神经信息处理系统进展（NeurIPS）34, 10144 (2021). [34] P. Nakkiran, 《更多数据可能损害线性回归：样本级双下降》, arXiv 预印本 arXiv:1912.07242 (2019). [35] P. Nakkiran、P. Venkat、S. Kakade 和 T. Ma, 《最优正则化可以缓解双下降》, 载于 *International Conference on Learning Representations (ICLR)* (2021). [36]S. Goldt、B. Loureiro、G. Reeves、F. Krzakala、M. M´ezard 和 L. Zdeborov´a, 《浅层神经网络学习中生成模型的高斯等价性》, 机器学习研究论文集（MSML）145, 426 (2022). [37] H. Hu 和 Y. M. Lu, 《随机特征高维学习的普适性定律》, IEEE 信息论汇刊 69，1932 (2023). [38] A. Montanari 和 B. N. Saeed, 《经验风险最小化的普适性》, 载于 *Conference on Learning Theory (COLT)* 2022 4310 4312 (PMLR,) 第 – 页。[39] B. Loureiro、C. Gerbelot、H. Cui、S. Goldt、F. Krzakala、M. M´ezard 和 L. Zdeborov´a, 《基于教师—学生模型、面向真实数据集的通用特征映射学习曲线》, 神经信息处理系统进展（NeurIPS）34 (2021). [40] Y. Dandi、L. Stephan、F. Krzakala、B. Loureiro 和 L. Zdeborov´a, 《广义线性模型中高斯混合模型的普适性定律》, 神经信息处理系统进展（NeurIPS）36 (2023). [41] L. Pesce、F. Krzakala、B. Loureiro 和 L. Stephan, 《高斯数据是否足够？高维广义线性估计中普适性的范围与局限》, 载于 *International Conference on Machine Learning (ICML)* 2023 27680 27708 (PMLR,) 第 – 页。[42] X. Mai 和 Z. Liao

，高斯普适性失

效性在高维线性因子混合分类中的表现，发表于 *International Conference on Learning Representations (ICLR)* 2025 2410 05609 ( )

arXiv: . . [43] C. Yaakoubi、C. Louart、M. Tiomoko 和 Z. Liao，《高维经验风险最小化中高斯普适性失效的表征》，发表于 *Interna-tional Conference on Machine Learning (ICML)* (2026) arXiv:2604.03146。[44] H. Cui、L. Pesce、Y. Dandi、F. Krzakala、Y. M. Lu、L. Zdeborov´a 和 B. Loureiro，《双层网络中一步梯度步之后的特征学习渐近行为》，发表于 *International Conference on Machine Learning* (PMLR, 2024) 第 9662–9695 页。[45] G. G. Wen 、 H. Hu 、 Y. M. Lu 、 Z. Fan 和 T. Misi- akiewicz，《高斯等价性何时失效，以及如何修复：具有二次标度的随机特征的非普适行为》，arXiv 预印本 arXiv:2512.03325 (2025)。[46] F. Aguirre-L´opez、S. Franz 和 M. Pastore，《随机特征与多项式规则》，SciPost Physics **18**, 039 (2025)。[47] A. Sclocchi、A. Favero 和 M. Wyart，《扩散模型中的相变揭示了数据的层次结构》，Proceedings of the National Academy of Sciences **122** (2025), arXiv:2402.16991。[48] A. Favero 、 A. Sclocchi 和 M. Wyart，《规模更大不总是意味着记忆：过参数化扩散模型中的提前停止》，arXiv 预印本 arXiv:2505.16959 (2025)。[49] J. Garnier-Brun、L. Biggio、D. Beltrame、M. M´ezard 和 L. Saglietti，《生成式扩散中的有偏泛化》，发表于 *Proceedings of the 43rd International Conference on Machine Learning (ICML)* 2026 ( )。

[50] F. Elgorni 和 C. Lucibello，(2026)，准备中。[51] G. Biroli、T. Bonnaire、V. de Bortoli 和 M. M´ezard，《扩散模型的动力学机制》，Nature Commu- nications**15**, 9957 (2024)。[52] T. Bonnaire、R. Urfin、G. Biroli 和 M. M´ezard，《扩散模型为何不会记忆：隐式动力学正则化在训练中的作用》，发表于 *Advances in Neural Information Processing Systems* (2025) arXiv:2505.17638。[53] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cour- napeau、M. Brucher、M. Perrot 和 ´E. Duchesnay，《Scikit-learn：Python 机器学习》，Journal of Ma- chine Learning Research **12**, 2825 (2011)。

附录 A：置信传播

由于生成模型是一棵树，因此可以通过置信传播（BP）精确计算贝叶斯最优根节点预测器 [29]，，沿树向上传递消息，从叶节点传到根节点。首先，利用自旋的二态性质，将转移矩阵的各个元素 $\boldsymbol{\pi}$（ε）写成

$$\pi_{x_i x_j} = \frac{1+\theta x_i x_j}{2}, \qquad \theta = 1-2\varepsilon, \tag{A1}$$

其中， x_i 是父节点自旋， x_j 是其子节点之一。按照参考文献 [29]，的标准记法，每条边（ij）是一个因子节点 b，而由 b发送给父节点的消息更新为

$$\hat{\nu}_{b\to i}(x_i) = \sum_{x_j=\pm 1} \pi_{x_i x_j}\, \nu_{j\to b}(x_j). \tag{A2}$$

我们可以将概率向量 ν 化为一个标量磁化强度，

$$m_j = \nu_{j\to b}(+1) - \nu_{j\to b}(-1) = 2\nu_{j\to b}(+1) - 1, \tag{A3}$$

从而可以改写为

$$\nu_{j\to b}(x_j) = \frac{1+m_j x_j}{2}. \tag{A4}$$

符号 m 仅在本附录中使用，不应与学习问题中的序参量混淆。在置信传播更新方程中代入，

$$\begin{aligned}\hat{\nu}_{b\to i}(x_i) &= \sum_{x_j=\pm 1} \frac{1+\theta x_i x_j}{2}\,\frac{1+m_j x_j}{2} \\ &= \frac{1+\theta m_j x_i}{2}.\end{aligned} \tag{A5}$$

记 j、 k 为 i的两个子节点，它们通过因子 b、 c与其相连，并记 a 为通向其父节点的边，则变量节点 i 处的更新会将传入的消息 $\hat{}\nu_{b\to i}$和 $\hat{}\nu_{c\to i}$相乘，

$$\nu_{i\to a}(x_i) \propto (1+\theta m_j x_i)(1+\theta m_k x_i). \tag{A6}$$

利用 $x_i^2 = 1$ 并强制满足归一化条件，可得

$$\nu_{i\to a}(x_i) = \frac{1}{2} + \frac{1}{2}\,\frac{\theta(m_j+m_k)}{1+\theta^2 m_j m_k}\, x_i. \tag{A7}$$

与方程（A4）对照可知，贝叶斯最优归结为沿树向上传递磁化强度：从一对子节点（j， k）传递到它们的父节点 i，如下所示：

$$m_i = \frac{\theta(m_j+m_k)}{1+\theta^2 m_j m_k}, \tag{A8}$$

或者等价地， $m = \tanh h$、 $h_i = \operatorname{artanh}(\theta m_j)$ $+\operatorname{artanh}(\theta m_k)$。递归从已观测到的叶节点开始初始化，其中 $m_i = x_i$。由于图是一棵树，贝叶斯最优是精确的 [29]。在根节点处，均匀先验给出相同的更新，因此 $m_0 = \mathbb{E}[x_0 \mid x_1, \ldots, x_N]$。因此，贝叶斯最优估计器为 $\hat{}x_0^{\mathrm{BP}} = \operatorname{sgn} m_0$；若出现平局，则抛掷一枚公平硬币，给定叶节点后，其正确概率为 $(1+|m_0|)/2$ 。

这一迭代在此阶段完全依赖具体实例，因为 m_j 和 m_k 取决于这些节点下方的子树。然而，它们的统计量遵循分布上的封闭递归。将子树的所有自旋翻转，会使其磁化强度 m 变为 $-m$；因此，给定其顶端节点的自旋x_i ， m_i 服从 $x_i\tilde{m}$，其中 $\tilde{m}$表示根对齐磁化，即以顶端自旋为正为条件的磁化。对于边符号 η_1， η_2 的

第 IIB 2节， $\mathbb{P}(\eta=-1)=\varepsilon$，深度为$\ell$ 的树的根对齐磁化满足

$$\tilde{m}_\ell \overset{d}{=} \frac{\theta\big(\eta_1\tilde{m}_{\ell-1}^{(1)}+\eta_2\tilde{m}_{\ell-1}^{(2)}\big)}{1+\theta^2\eta_1\eta_2\,\tilde{m}_{\ell-1}^{(1)}\tilde{m}_{\ell-1}^{(2)}}, \tag{A9}$$

其中 $\tilde{m}_0=1$，其中 $\tilde{m}_{\ell-1}^{(1)}$，$\tilde{m}_{\ell-1}^{(2)}$ 是 $\tilde{m}_{\ell-1}$的相互独立副本，并且与边符号相互独立。这是普查递归（式（18））的非线性类比。根据对称性，BP 错误率仅由根对齐分布决定，这与第 IIB 2 节中的普查误差相对应。

$$\begin{aligned}\mathcal{E}_{\mathrm{BP}}^{(\ell)} &= \mathbb{P}(\tilde{m}_\ell<0)+\frac{1}{2}\mathbb{P}(\tilde{m}_\ell=0)\\ &= \frac{1}{2}\big(1-\mathbb{E}[|\tilde{m}_\ell|]\big).\end{aligned} \tag{A10}$$

该递归精确地定义了有限深度分布，但枚举其支集很快就会变得不切实际。因此，我们像处理普查一样，用群体动力学对其进行近似：构造一个包含 M 条消息、初始于 $\tilde{m}=1$ 的池，并从当前池中随机抽取两条消息，再取两个新的边符号，将每个成员替换为在这些变量上计算的式（A9）右端。经过 ℓ 次更新后，该池中的样本 $\tilde{m}_\ell$ 了有限 M 涨落，而 $\ell\to\infty$ 极限则是式（A9）的分布固定点。由于式（18）和式（A9）使用相同的边符号，因此可以将两个递归联合应用于一个由成对变量（$\tilde{m}$，c）组成的池，从而在相同的树上对 BP 与普查输出的联合分布进行抽样。

最后，KS 判据可由平凡固定点的线性稳定性 $\tilde{m}=0$ （方程（A9））得到；该固定点不包含任何关于根节点的信息。在线性阶下，$\tilde{m}_\ell \simeq \theta(\eta_1\tilde{m}_{\ell-1}^{(1)}+\eta_2\tilde{m}_{\ell-1}^{(2)})$，并对 $\mathbb{E}[\eta]=\theta$取平均，

$$\mathbb{E}[\tilde{m}_\ell]\simeq 2\theta^2\,\mathbb{E}[\tilde{m}_{\ell-1}] = r\,\mathbb{E}[\tilde{m}_{\ell-1}]\,. \tag{A11}$$

因此，当 $r=2\theta^2>1$，即 $\varepsilon<\varepsilon_c$ 时，平凡固定点变得不稳定。仅凭线性稳定性，无法排除在 $r<1$ 时存在与平凡固定点共存的非平凡固定点；对于更大的字母表，这种情况确实会发生，此时在 KS 阈值以下仍可能实现重构 [27, 28]。然而，对于二元对称信道，有一个定理指出，当 $r\le 1$ [24, 25] 时无法实现重构，因此贝叶斯最优与普查具有相同的阈值 ε_c，二者仅在 $\varepsilon<\varepsilon_c$ 时的准确度不同。

由联合群体动力学得到的普查与 BP 错误率，在图7中进行了比较。它们的差异在所有噪声水平下都很小。值得注意的是，在较小的 ε处，两种估计量都近乎完美，因此差异消失；在无限深度下，当达到两者都失效的阈值时，差异也会消失；而在中等噪声下差异最大，并且随着树的深度增加而增大。

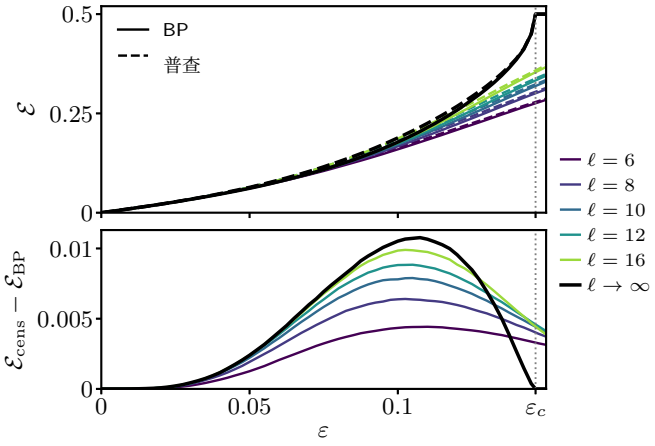

图 7。根节点的贝叶斯最优（BP）重构与普查重构。上图：由两种估计量的联合群体动力学得到的 BP（实线）和普查（虚线）误差随 ε 的变化，其中深度 ℓ 和 $\ell\to\infty$ （黑色）逐渐增加。下图：两者的差异，该差异在所有噪声水平下都很小；在无限深度下，于 KS 阈值 ε_c处消失，此时两种估计量均失效。

附录 B：精确普查分布

根据广播过程的自旋翻转对称性，不失一般性地令 $y=+1$。令 n_k 是根为+1的深度为k 的子树中 $+1$ 叶节点的数量，并且 $p_k(n)=\mathbb{P}(n_k=n)$是其在 $n=0$、……、2^k上的分布。深度为k 的子树是以其根节点的两个子节点为根的深度为$(k-1)$的子树的并集。每个子节点以概率 $1-\varepsilon$为+1，此时其计数服从分布p_{k-1}；以概率 ε为 -1，此时根据自旋翻转对称性，其 $+1$叶节点计数服从翻转后的分布 $p_{k-1}(2^{k-1}-n)$。因此，子树的计数服从

$$g_{k-1}(n)=(1-\varepsilon)\,p_{k-1}(n)+\varepsilon\,p_{k-1}(2^{k-1}-n), \tag{B1}$$

由于两个子树相互独立，p_k 是离散卷积 $p_k=g_{k-1}*g_{k-1}$，

$$p_k(n)=\sum_{m=0}^{2^{k-1}} g_{k-1}(m)\,g_{k-1}(n-m), \tag{B2}$$

其中 $g_{k-1}(m)=0$ 在 $0\le m\le 2^{k-1}$之外，起始于$p_0(n)=\delta_{n,\,1}$。这是分布递归（18）的精确有限深度对应形式：边符号 $\eta_{1,\,2}$是（B1）中带有翻转概率的混合，而两个子树普查之和就是卷积（B2）。等价地，可以逐层揭示树，并通过二项转移核传播每一层的 $+1$ 节点数；不过，对子树进行递归更简单，因为每个深度只涉及单个标量的分布。

由于 $C^{(\ell)}=2n_\ell-N$，归一化普查取值为

$$\mathbb{P}\left(c_\ell=\frac{2n-N}{(2\theta)^\ell}\right)=p_\ell(n),\qquad n=0,\dots,N, \tag{B3}$$

在间距为 $2/(2\theta)^\ell$ 的格点上，当 $2\theta>1$ 时，该间距趋于零，尤其是当 $r>1$ 时。

正文中的 方程（20）描述了普查准确率，它可由普查递归方程（18）在给定两条根边的条件下推得。事实上， c_ℓ 的符号与 $S=\eta_1 c_{\ell-1}^{(1)}+\eta_2 c_{\ell-1}^{(2)}$ 的符号相同；将平局计为二分之一时，普查误差为 $\mathbb{P}(S<0)+\frac{1}{2}\mathbb{P}(S=0)$。若两条边都保留（其概率为 $(1-\varepsilon)^2$,），则 $S=c_{\ell-1}^{(1)}+c_{\ell-1}^{(2)}$，我们将其误差记为 $a_{\ell-1}$。若两条边都翻转（其概率为 ε^2），则 S 变号，误差变为 $1-a_{\ell-1}$。若恰有一条边翻转（其概率为 $2\varepsilon(1-\varepsilon)$），则 $S=\pm(c_{\ell-1}^{(1)}-c_{\ell-1}^{(2)})$。由于两棵子树的普查结果独立且同分布，这一结果关于零对称，因此误差恰为 $\frac{1}{2}$。将这三种情形相加，

$$\begin{aligned}\mathcal{E}_{\mathrm{cens}}^{(\ell)} &= (1-\varepsilon)^2 a_{\ell-1}+\varepsilon^2(1-a_{\ell-1})+\varepsilon(1-\varepsilon)\\ &= \varepsilon+(1-2\varepsilon)\,a_{\ell-1},\end{aligned}\tag{B4}$$

即式（20），且至少为 ε，因为 $a_{\ell-1}\ge 0$ 且$\varepsilon<1/2$。

从实现角度来看，步骤 k （B2）对两个长度为 $2^{k-1}+1$的向量进行卷积，因此计算成本主要由最后一步决定。采用直接求和时，总体复杂度为 $O(N^2)$；采用快速傅里叶变换时，复杂度为 $O(N\log N)$。有限深度鞍点计算采用精确的离散递归。其支撑集随 ℓ呈指数增长，而群体动力学保持固定的样本量，为求得极限分布提供了一种实用途径。数值实现以及无限深度平均值所用的近似方法详见附录 F。

附录 C：极限普查的矩

对于 $r>1$归一化普查 c_ℓ 收敛到极限c（第 IIB 2节），其分布满足分布固定点方程

$$c \overset{d}{=} \frac{\eta_1 c^{(1)}+\eta_2 c^{(2)}}{2\theta},\tag{C1}$$

其中 $c^{(1)}$, $c^{(2)}$ 是 c 的独立副本，且与边符号 η_1, η_2 相互独立。我们逐阶计算其矩$M_p=\mathbb{E}[c^p]$，并在本附录末尾检验这些矩是 c_ℓ 的矩的极限。

将 (C1) 提升到 p次方，对二项式展开，并对独立的符号和子树取平均，得到

$$M_p=\sum_{j=0}^{p}\binom{p}{j}\psi_j\psi_{p-j}M_jM_{p-j},\tag{C2}$$

其中 $M_0=1$ 和

$$\psi_j=\frac{\mathbb{E}[\eta^j]}{(2\theta)^j}=\begin{cases}\theta/(2\theta)^j, & j\ \text{odd},\\ 1/(2\theta)^j, & j\ \text{even}.\end{cases}\tag{C3}$$

我们首先注意到， M_p 只通过项 $j=0$ 和 $j=p$ 出现在等式右侧，且两者的系数均为 ψ_p，因此该方程实际上变为

$$(1-2\psi_p)M_p=\sum_{j=1}^{p-1}\binom{p}{j}\psi_j\psi_{p-j}M_jM_{p-j},\tag{C4}$$

只要 $1-2\psi_p\neq 0$，各阶矩便可依次得到。对于 $p=1$，两边均为零，因为 $2\psi_1=1$。这反映了 (C1) 在 $c\to tc$ 下的不变性，该不变性使均值无法确定，因此我们令 $M_1=1$，此设定继承自 $\mathbb{E}[c_\ell]=1$。

依次求解 (C4)，右端分别为 $1/2$、$3M_2/(4\theta^2)$ 和 $M_3/(2\theta^2)+3M_2^2/(8\theta^4)$，对于 $p=2$、3、4，我们得到

$$\begin{aligned}M_2&=\frac{\theta^2}{2\theta^2-1}=\frac{r}{2(r-1)},\\ M_3&=\frac{3\theta^2}{(2\theta^2-1)(4\theta^2-1)}=\frac{3r}{2(r-1)(2r-1)},\\ M_4&=\frac{3\theta^4(12\theta^2-5)}{(2\theta^2-1)^2(4\theta^2-1)(8\theta^4-1)}\\ &=\frac{3r^2(6r-5)}{4(r-1)^2(2r-1)(2r^2-1)}.\end{aligned}\tag{C5}$$

第二累积量 $\kappa_2=M_2-1$ 恢复了 γ_∞ 的 (17。第三和第四累积量 $\kappa_3=M_3-3M_2+2$和$\kappa_4=M_4-4M_3-3M_2^2+12M_2-6$分别为

$$\begin{aligned}\kappa_3&=-\frac{2(1-\theta^2)}{4\theta^2-1}=-\frac{2-r}{2r-1},\\ \kappa_4&=\frac{6(1-\theta^2)(4\theta^4-\theta^2-1)}{(4\theta^2-1)(8\theta^4-1)}\\ &=\frac{3(2-r)(2r^2-r-2)}{2(2r-1)(2r^2-1)}.\end{aligned}\tag{C6}$$

将 $\gamma_\infty=(2-r)/[2(r-1)]$,归一化后，偏度和超额峰度为

$$\begin{aligned}\frac{\kappa_3}{\gamma_\infty^{3/2}}&=-\frac{2(2\theta^2-1)^{3/2}}{(4\theta^2-1)(1-\theta^2)^{1/2}}\\ &=-\frac{2\sqrt{2}\,(r-1)^{3/2}}{(2r-1)(2-r)^{1/2}},\\ \frac{\kappa_4}{\gamma_\infty^2}&=\frac{6(r-1)^2(2r^2-r-2)}{(2-r)(2r-1)(2r^2-1)}.\end{aligned}\tag{C7}$$

这些数值与通过种群动力学估计的累积量，以及文献 [16] 中独立得到的广播和的渐近矩相符。在整个 $1<r<2$ 范围内，偏度始终严格为负，即对每个 $0<\varepsilon<\varepsilon_c$都如此；只有在趋近 Kesten–Stigum 阈值时，偏度才趋于零，如 $-2\sqrt{2}(r-1)^{3/2}$ 所示。相较之下，超额峰度会在 $2r^2-r-2$ 的正根处改变符号，即 $r_\kappa=(1+\sqrt{17})/4\approx 1.28$，或 $\varepsilon_\kappa\approx 0.1$。当 $\varepsilon<\varepsilon_\kappa$ 时它为正，当 $\varepsilon_\kappa<\varepsilon<\overline{\varepsilon_c}$ 时为负，并在 Kesten–Stigum 阈值处按 $-6(r-1)^2$ 趋于零。因此，基于峰度的检验

单独使用会错误地认为，在$\varepsilon \approx 0.1$附近的高斯普查呈高斯分布，而此时偏度仍显著不为零。接近 Kesten–Stigum 阈值时，偏度的影响也超过峰度，因为 $(r-1)^{3/2} \gg (r-1)^2$。因此，我们使用偏度作为非高斯性的指标。

在小噪声下，$\theta = 1-2\varepsilon$ 给出

$$\kappa_2 \simeq 4\varepsilon, \qquad \kappa_3 \simeq -\frac{8\varepsilon}{3}, \qquad \kappa_4 \simeq \frac{16\varepsilon}{7}, \tag{C8}$$

因此，偏度 $\simeq -1/(3\sqrt{\varepsilon})$ 和超额峰度$\simeq 1/(7\varepsilon)$ 发散，这是根节点附近罕见翻转的特征，而非微弱中心极限定理修正的特征，这与误差下限一致 (20)。

同样的展开也适用于有限深度。将其应用于 (18)，可将矩 $M_p^{(\ell-1)} = \mathbb{E}\big[c_{\ell-1}^p\big]$ 映射为

$$M_p^{(\ell)} = \sum_{j=0}^{p} \binom{p}{j} \psi_j \psi_{p-j} M_j^{(\ell-1)} M_{p-j}^{(\ell-1)}, \tag{C9}$$

从所有 p 的 $M_p^{(0)} = 1$ 开始。这现在是显式迭代，而非不动点条件。当 p =2 时，它写作 $\gamma_\ell = (\gamma_{\ell-1} + 1 - \theta^2)/r$，其中 $\gamma_0 = 0$，其解为 $\gamma_\ell = \gamma_\infty(1 - r^{-\ell})$，如 (17) 所示。在每个阶数上$p \geq 2$，$M_p^{(\ell)}$都是 $M_{(\ell-1)p}$的仿射函数，斜率为$2\psi_p \leq 2(2\theta)^{-p} \leq 1/r$，因为 $2\theta \geq r > 1$，并由低阶矩驱动。根据对 p 的归纳，所有矩都收敛到不动点值 M_p，修正项为 $O(r^{-\ell})$。由于c_ℓ 是收敛到 c [30, 31]的鞅，且其所有矩都保持有界，这些极限就是 c 的矩。

附录 D：副本计算

我们使用著名的副本技巧推导 $f^{(\ell)}$ 方程（27）和（28）的鞍点自由能 $\langle \log Z_\beta \rangle = \lim_{n\to 0} n^{-1} \log \langle Z_\beta^n \rangle$，先对 $\langle Z_\beta^n \rangle$ 整数个副本进行计算 n，再将其延拓至 $n \to 0$。这些步骤对于感知机而言是标准的，包括具有一般协方差的高斯混合输入；详细推导请参见文献。[5, 33] 这里仅作简要介绍，并重点讨论混合模型所特有的部分，即对普查方向的单独处理。

根对齐输入可以分解为

$$y\boldsymbol{x} = s_\ell c_\ell \, \hat{\mathbf{e}} + \boldsymbol{x}_\perp, \tag{D1}$$

其中，$\boldsymbol{x}_\perp$ 与 $\hat{\mathbf{e}}$正交，均值为零，协方差为 $\boldsymbol{\Sigma}_\perp$，且与 c_ℓ不相关。对于副本 $a = 1, \ldots\ldots, n$，我们记 $\boldsymbol{w}^a = \sqrt{N} m_a \hat{\mathbf{e}} + \boldsymbol{w}_\perp^a$，并令 $\bar{m}_a = s_\ell m_a$。副本 a 在样本 ν 上的间隔为

$$\frac{y^\nu \boldsymbol{w}^a \cdot \boldsymbol{x}^\nu}{\sqrt{N}} = \bar{m}_a c_\ell^\nu + \xi_a^\nu, \qquad \xi_a^\nu = \frac{\boldsymbol{w}_\perp^a \cdot \boldsymbol{x}_\perp^\nu}{\sqrt{N}}, \tag{D2}$$

因此，在权重固定时，数据仅通过 c_ℓ^ν 与横向场 ξ_a^ν的联合分布进入。由于

如第IIIA节所述，我们取这些场为联合高斯分布、均值为零，且与 c_ℓ^ν独立，其协方差为$Q_{ab}^\perp = \boldsymbol{w}_\perp^{a\top} \boldsymbol{\Sigma}_\perp \boldsymbol{w}_\perp^b / N$。由于 $\hat{\mathbf{e}}$ 是 $\boldsymbol{\Sigma}$的一个特征向量，这与 $\boldsymbol{w}_\perp^{a\top} \boldsymbol{\Sigma} \boldsymbol{w}_\perp^b / N$一致。对所有权重都要求该高斯分布，相当于取 $\boldsymbol{x}_\perp = \boldsymbol{\Sigma}_\perp^{1/2} \boldsymbol{z}_\perp$，其中 $\boldsymbol{z}_\perp$ 是一个与 c_ℓ独立的标准高斯向量，由此得到我们称为混合模型的方案。正则化项可分解为 $\lambda \|\boldsymbol{w}^a\|^2 = N\lambda_\ell \bar{m}_a^2 + \lambda \|\boldsymbol{w}_\perp^a\|^2$。由于样本彼此独立，数据仅通过$\bar{m}_a$ 和 $Q^\perp$进入，并且

$$\begin{aligned} \langle Z_\beta^n \rangle &= \int \prod_a \mathrm{d}\bar{m}_a \, \mathrm{e}^{-\frac{\beta N \lambda_\ell}{2} \sum_a \bar{m}_a^2} \int \mathrm{d}Q^\perp \, \mathrm{e}^{N\mathcal{G}(Q^\perp)} \, \mathcal{Z}_n^P, \\ \mathcal{Z}_n &= \Big\langle \prod_a \mathrm{e}^{-\beta V(\bar{m}_a c_\ell + \xi_a)} \Big\rangle_{c,\, \boldsymbol{\xi} \sim \mathcal{N}(0, Q^\perp)}, \end{aligned} \tag{D3}$$

相差一个常数雅可比因子。这里，$\mathcal{G}$ 是具有固定重叠的横向副本的岭参数加权体积，

$$\begin{aligned} \mathrm{e}^{N\mathcal{G}(Q^\perp)} = \int \prod_a \mathrm{d}\boldsymbol{w}_\perp^a \, \mathrm{e}^{-\frac{\beta\lambda}{2} \sum_a \|\boldsymbol{w}_\perp^a\|^2} \\ \times \prod_{a \leq b} \delta\big(N Q_{ab}^\perp - \boldsymbol{w}_\perp^{a\top} \boldsymbol{\Sigma} \boldsymbol{w}_\perp^b\big) . \end{aligned} \tag{D4}$$

我们用指数因子表示这些约束

$$\exp\Big[-\frac{N}{2} \sum_{ab} \hat{Q}_{ab} Q_{ab}^\perp + \frac{1}{2} \sum_{ab} \hat{Q}_{ab} \, \boldsymbol{w}_\perp^{a\top} \boldsymbol{\Sigma} \boldsymbol{w}_\perp^b \Big], \tag{D5}$$

$\hat{}$，其中对称共轭矩阵为 Q。

我们采用复制对称假设，该假设由损失函数的凸性所支持。考虑到接下来要取的 $\beta \to \infty$ 极限，我们将其写为

$$Q_{ab}^\perp = q_\perp + \frac{\chi}{\beta} \delta_{ab}, \quad \hat{Q}_{ab} = \beta^2 \hat{q} - \beta \hat{\chi} \, \delta_{ab}, \quad \bar{m}_a = \bar{m}, \tag{D6}$$

因此 $q_\perp = Q_{a\neq b}^\perp$且 $\chi = \lim_{\beta\to\infty} \beta(Q_{aa}^\perp - q_\perp)$，并将对 $-(\beta N n)^{-1} \log \langle Z_\beta^n \rangle$ 的各项贡献汇总为 $n \to 0$，然后 $\beta \to \infty$。式（D5）的第一项贡献为 $(\hat{q}\chi - \hat{\chi} q_\perp)/2 = g_I$，而信号坐标上的岭参数贡献为 $\lambda_\ell \bar{m}^2/2$。

在权重积分中，式（D5）的第二项为$\beta^2 \hat{q} \|\boldsymbol{\Sigma}^{1/2} \sum_a \boldsymbol{w}_\perp^a\|^2 - \frac{\beta\hat{\chi}}{2} \sum_a \boldsymbol{w}_\perp^{a\top} \boldsymbol{\Sigma} \boldsymbol{w}_\perp^a$。通过对 $\boldsymbol{u} \sim \mathcal{N}(0,)$施加 Hubbard–Stratonovich 变换，可将第一部分线性化并使各副本解耦。随后，每个副本在横向子空间上都有一个高斯积分，其精度矩阵为 $\beta(\lambda + \hat{\chi}\boldsymbol{\Sigma}_\perp)$，源项为 $\beta\sqrt{\hat{q}}\, \boldsymbol{\Sigma}_\perp^{1/2} \boldsymbol{u}$。配方后，在 $\boldsymbol{u}$ 处对 $n \to 0$取平均，横向权重积分（式（D4））的贡献最终为

$$\frac{\beta\hat{q}}{2} \operatorname{Tr} \frac{\boldsymbol{\Sigma}_\perp}{\lambda + \hat{\chi}\boldsymbol{\Sigma}_\perp} = \frac{\beta N \hat{q}}{2} \Big\langle \frac{\mu}{\lambda + \mu\hat{\chi}} \Big\rangle_{\rho_\perp, \ell}, \tag{D7}$$

遍历横向子空间的迹，其中 $\boldsymbol{\Sigma}_\perp$ 的谱与 $\boldsymbol{\Sigma}$的谱相同，只多出一个$\mu_{\mathrm{sig},\ \ell}$

被移除，即 $\rho_{\perp,\,\ell}$ 在归一化 13之后的（ N。横向岭未进行重新缩放，因此裸 λ 出现了。加上信号岭后，得到 g_S。对数行列式为 $O(N\log\beta)$，除以 β后消失。

在数据项中，副本对称场为$\xi_a=\sqrt{q_\perp}\,\zeta+\sqrt{\chi/\beta}\,\zeta_a$，其中 ζ，ζ_a 是相互独立的标准高斯变量，并记作 $h=\bar{m}c_\ell+\sqrt{q_\perp}\,\zeta$，

$$\mathcal{Z}_n=\left\langle\left(\left\langle \mathrm{e}^{-\beta V(h+\sqrt{\chi/\beta}\,\zeta_a)}\right\rangle_{\zeta_a}\right)^n\right\rangle_{c,\zeta}. \tag{D8}$$

对 n 作一阶展开，$\log\mathcal{Z}_n=n\langle\log\langle\cdots\rangle_{\zeta_a}\rangle_{c,\,\zeta}+O(n^2)$。作变量变换 $z=h+\sqrt{\chi/\beta}\,\zeta_a$，

$$\left\langle \mathrm{e}^{-\beta V(h+\sqrt{\chi/\beta}\,\zeta_a)}\right\rangle_{\zeta_a}=\sqrt{\frac{\beta}{2\pi\chi}}\int \mathrm{d}z\,\mathrm{e}^{-\beta\left[\frac{(z-h)^2}{2\chi}+V(z)\right]}, \tag{D9}$$

并且由于凸V时括号内的函数有唯一极小值，拉普拉斯方法恰好给出 $-\beta^{-1}\log\langle\cdots\rangle_{\zeta_a}\to\min_z[(z-h)^2/(2\chi)+V(z)]$（当 $\beta\to\infty$ 时），前因子仅贡献 $O(\beta^{-1}\log\beta)$。因此，数据的贡献为正文中的αg_E 。普查仅通过其精确分布的一维平均进入。将各项贡献相加，并在大 N 下求极值，得到$f_{(\ell)}$。

关于 $\hat{q}$, χ、$\hat{\chi}$, $q_\perp$ 和 $\bar{m}$ 对 (27) 的驻定条件依次给出五个方程 (29)。 g_E 的导数由 $z_\star=h-\chi V'(z_\star)$ 得出；对于 h 中最小值的导数，它给出 $V'(z_\star)$，而 χ 中的 $-V'(z_\star)^2/2$ 与固定 c_ℓ 时对 ζ 作高斯分部积分一起给出 $q_\perp$ 中的导数，即 $\partial_h z_\star=1/(1+\chi V''(z_\star))$。新样本的间隔就是 h 本身，因此得到测试误差 (30)。

取 $c_\ell\sim\mathcal{N}(1,\gamma_\ell)$ 即可恢复高斯鞍点。由斯坦引理可得

$$\langle c_\ell V'(z_\star)\rangle_{c,\zeta}=\langle V'(z_\star)\rangle_{c,\zeta}+\gamma_\ell\bar{m}\left\langle\frac{V''(z_\star)}{1+\chi V''(z_\star)}\right\rangle_{c,\zeta}, \tag{D10}$$

结合 $\hat{\chi}$ 的方程，信号方程变为

$$(\tilde{\lambda}_\ell+\gamma_\ell\hat{\chi})\,\bar{m}=-\alpha\langle V'(z_\star)\rangle_{c,\zeta}. \tag{D11}$$

~两边乘以 s_ℓ，并利用 $\bar{m}=s_\ell m$、$\lambda_\ell s_\ell^2=\lambda$ 和 $\gamma_\ell s_\ell^2=\mu_{\mathrm{sig},\,\ell}$，可得 $(\lambda+\mu_{\mathrm{sig},\,\ell}\hat{\chi})m=-\alpha s_\ell\langle V'(z_\star)\rangle_{c,\,\zeta}$，这就是协方差为 $\mathbf{\Sigma}$ 的高斯协变量的信号方程。将重叠识别为 $q=q_\perp+\mu_{\mathrm{sig},\,\ell}m^2$ 后，其余方程与高斯协变量的方程一致，因为完整谱和横向谱除 $O(1/N)$（最大特征值权重）外相同。

附录 E：谱截断与有限深度修正

在 $\ell=\infty$ 式（29）的谱平均是在 ρ 式（14）的无限测度上进行的。我们将其截断并界定

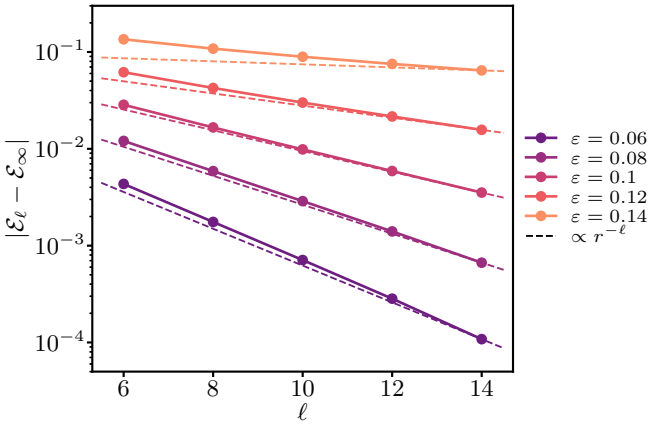

图 8. 在 $\lambda=0.1$ 和 $\alpha=8$ 处，有限深度鞍点误差与其无限深度极限之间的距离；图中展示了 ε 的几个取值，并用虚线标出了普查速率 $r^{-\ell}$。随着 ε 接近 ε_c，收敛速度变慢，此时深度尺度 $1/\log r$ 发散。

由此产生的误差，假设 $\hat{\chi}>0$。被积函数$\mu^j/(\lambda+\mu\hat{\chi})^j$，$j=1$，2 随 μ 增大并趋于饱和，

$$0\le\frac{\mu}{\lambda+\mu\hat{\chi}}<\frac{1}{\hat{\chi}},\qquad 0\le\frac{\mu^2}{(\lambda+\mu\hat{\chi})^2}<\frac{1}{\hat{\chi}^2}. \tag{E1}$$

令 $\rho^{(K)}=\sum_{k\le K}2^{-k}\delta_{\mu_k}$ 为在高度 K 处截断的 ρ 。由于舍弃质量为 $\sum_{k>K}2^{-k}=2^{-K}$，

$$0\le\left\langle\frac{\mu^j}{(\lambda+\mu\hat{\chi})^j}\right\rangle_\rho-\left\langle\frac{\mu^j}{(\lambda+\mu\hat{\chi})^j}\right\rangle_{\rho^{(K)}}\le\frac{2^{-K}}{\hat{\chi}^j}. \tag{E2}$$

根据（13），$\rho_{\perp,\,\ell}=\sum_{k<\ell}2^{-k}\delta_{\mu_k}+2^{-\ell}\delta_{\mu_\ell}=\rho^{(\ell)}$，这意味着有限深度横向测度恰好是将 ρ 在高度 ℓ 处截断，并且界（E2）对 $\rho_{\perp,\,\ell}$ 成立，其中 2^{-K} 被 $2^{-\ell}=1/N$ 替换。完整测度 ρ_ℓ 则将 $\{\mu_k\}_{k>\ell}$ 的质量 $2^{-\ell}$ 下移至 μ_ℓ，而被积函数的单调性给出相同的界。舍弃尾部也可以替换为其饱和值$2^{-K}\hat{\chi}^{-j}$。由于 $1-(1+x)^{-j}\le jx$，其中 $x=\lambda/(\mu\hat{\chi})$，因此在固定 λ 和 $\hat{\chi}>0$ 时，误差随 $1<r<2$ 降至 $2^{-K}\hat{\chi}^{-j}\,j\lambda/(\hat{\chi}\mu_{K+1})=O((2r)^{-K})$。

图 8 展示了有限深度下鞍点误差以 $r^{-\ell}$ 普查修正的速率收敛至其无限深度极限；该速率主导 $O(1/N)$ 谱修正。

附录 F：数值方案

正文图中经验风险最小化对应的点，是使用 Scikit-learn 的 L-BFGS 求解器进行逻辑回归得到的 [53]（不含截距项，输入除以 $\sqrt{N}$，并采用式（23）中的正则化约定，在此实现中对应于 $C=1/\lambda$ ）。每个点表示多次独立重复实验的测试误差平均值，每次实验都有各自的训练

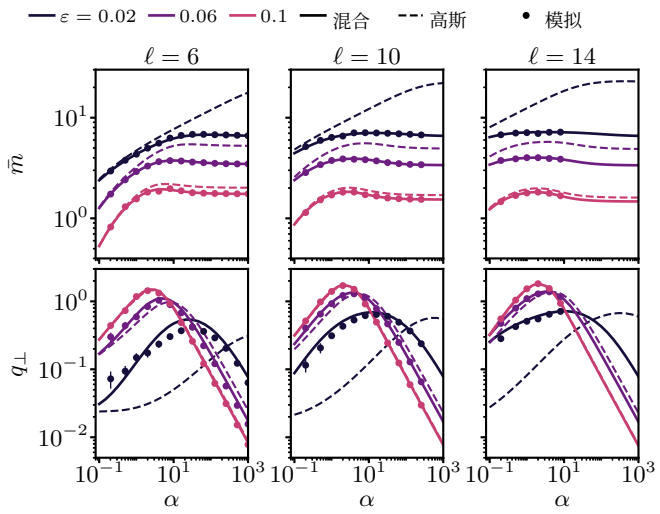

图 9. 广播树上逻辑回归的序参量，在 $\lambda = 0.1$ 条件下，相对于 α 深度 $\ell = 6$、10 和 14，以及 ε 的三个取值。信号重叠 $\bar{m}$（上）和横向重叠 $q_\perp$（下），分别来自有限深度混合鞍点（实线）、高斯鞍点（虚线）以及图 4 模拟中训练得到的权重（圆点，约 95% 区间）。

集，以及由 2×10^5 棵新树组成的独立测试集。图中绘制的区间是近似的 95% 置信区间，计算方法为重复实验均值的标准误差乘以 1.96，因此同时包含训练和测试的涨落。

当所有训练标签都相同时——对于平衡标签，这种情况以概率 2^{1-P} 发生——岭参数正则化目标仍然定义良好，并且具有唯一的极小值点。我们通过将每个训练样本分别以权重纳入 $(\boldsymbol{x}, y)$ 和 $(-\boldsymbol{x}, -y)$ 来拟合它，权重为 $1/2$，这使每个有符号间隔（进而目标函数）保持不变，同时为拟合接口提供所需的两个类别。

鞍点方程使用 Julia 求解器求解，该求解器可按需提供，并采用有限深度下精确的普查分布（附录B）、无限深度下的种群动力学，以及附录 E 中的谱截断。

附录 G：训练后的分类器的序参量

给定 $\boldsymbol{w}$ 根据经验风险最小化数值实验得到的训练权重，序参量 $\bar{m} = s_\ell \mathbf{1} \cdot \boldsymbol{w}/N$和 $q_\perp = \boldsymbol{w}_\perp^\top \boldsymbol{\Sigma}_\perp \boldsymbol{w}_\perp / N$ 可由协方差（7）精确得出。图9沿着图4底行的学习曲线，将它们与有限深度的混合鞍点和高斯鞍点进行了比较。混合鞍点能够准确描述 $\bar{m}$，而高斯鞍点则高估了它；在 ε 较小、 α较大时，这种高估愈发明显。因此，非高斯普查分布也会改变学得的权重。横向重叠也得到了较好的描述，只有在噪声水平最低时例外：此时它在深度较小时偏离鞍点，随着深度增加则逐渐趋近鞍点。使用精确普查分布和独立高斯横向场在合成数据上训练，似乎可以消除这种偏差，这表明偏差完全源于理论所忽略的普查与横向场之间的依赖关系。不过，这对测试误差影响不大。事实上，在信噪比较大时 $\bar{m}^2/q_\perp$，间隔 h与 c_ℓ 同号，只有$|c_\ell| \lesssim \sqrt{q_\perp}/\bar{m}$的样本较为罕见，因此 $\mathcal{E} \simeq \mathcal{E}_{\text{cens}}^{(\ell)}$ 不受 $q_\perp$具体取值的影响；这一点由图4中较小 ε 区域与数值模拟的完美吻合可见。