高维输入的学习时间缩放

Carlos Stein Brito

NightCity Labs，里斯本，葡萄牙

arXiv:2603.01184v1 [cs.LG] 1 Mar 2026

摘要

从复杂数据中进行表征学习通常涉及具有大量参数的模型，而这些模型又需要大量数据样本。在神经网络模型中，模型复杂度随每个神经元的输入数量而增长，模型表达能力与学习时间之间存在权衡。对这一权衡的精确刻画将有助于解释在人工网络和生物网络中观察到的连接性与学习时间。我们针对执行独立成分分析的赫布学习模型，对学习时间如何依赖于输入维数进行了理论分析。基于高维空间的几何，我们表明学习动力学可简化为一维问题，学习时间仅依赖于初始条件。对于更高的输入维度，初始参数具有更小的学习梯度和更长的学习时间。我们发现学习时间呈超线性缩放，在高输入维度下会迅速变得难以承受。这些结果揭示了高维学习的一个根本限制，并有助于阐明神经网络的最优设计如何依赖于数据复杂度。我们的方法勾勒出一个用于分析神经网络模型中学习动力学与模型复杂度的新框架。

1 引言

训练神经网络需要大量数据和学习时间 [9]。机器学习近期的突破，很大程度上得益于计算能力和大规模数据集日益丰富 [11]。即使拥有如此丰富的资源，学习时间仍然是提升神经网络复杂度及其应用规模的主要障碍之一 [10]。

神经元的数量以及每个神经元的输入数量（扇入）都会影响学习时间。此外，生物神经网络拥有数十亿个神经元，且每个神经元都与数千个突触相连 [8], ，也可能受到类似权衡关系的限制。本文分析神经网络中的学习动力学，并探讨网络复杂度如何影响学习时间。

我们考虑在 N 维数据中寻找稀疏隐藏特征的无监督学习任务。该任务可由采用非线性赫布学习规则 [14]的单个神经元求解。我们表明，该问题的优化函数具有数量随维度呈指数增长的鞍点和极大值点。我们表明，当输入连接数量很大 N 时，随机初始突触权重极有可能几乎与隐藏特征正交，并处于鞍点区域，从而导致很长的学习时间。

基于高维空间的几何与统计性质，我们表明学习动力学可由单个动力系统很好地描述，其中唯一相关的变量是初始条件，由到最近隐藏特征的距离决定。我们表明，该初始重叠随输入数量的增加而减小，从而导致在更大维度下学习时间迅速增加。

我们的结果带来了一个令人惊讶的新发现：最优学习时间与单个神经元上的突触权重数量呈超线性依赖。这种依赖关系对突触连接性造成了根本性限制，因为当输入数量很大时，学习会慢得难以接受。我们认为，这种方法有助于理解其他网络和任务，例如计算机视觉模型中的卷积网络模型。更广泛地说，我们的方法为分析神经网络模型中的学习动力学和模型复杂度勾勒出一个新的框架。

2 结果

2.1 稀疏特征的非线性赫布学习

我们考虑投影寻踪问题：在神经元的 N输入中寻找隐藏的稀疏特征，这与独立成分分析中的情形类似。我们通过优化突触权重 $\mathbf{w}$ 来寻找隐藏模式，使其符合如下目标函数，

$$\max_{\mathbf{w},||\mathbf{w}||=1} \left\langle F(\mathbf{w}^T\mathbf{x}) \right\rangle \tag{1}$$

我们将 N 维输入建模为 K 个稀疏隐藏变量的线性组合。我们将权重的范数约束为 $|\mathbf{w}|_2 = 1$，只优化$\mathbf{w}$的方向，并使用白化输入，$\langle \mathbf{x}\mathbf{x}^T \rangle = I$。一般而言，对于此任务，稀疏分布的具体形状并不重要 [3, 14], ，我们考虑两种情形：对称（拉普拉斯）分布或非对称分布（χ^2）。本文在此针对对称情形展开分析。

许多可能的优化函数 F(。) 都适用于学习稀疏成分[3, 14]。通过随机梯度下降学习这一目标，会得到非线性赫布学习规则： $\Delta\mathbf{w}_t \propto \mathbf{x}_t\ f(\mathbf{w}^T\mathbf{x}_t)$，其中 $f(u) = \frac{\partial F}{\partial u}$。在模拟中，我们使用线性整流器，$f(u) = (u-2.)_+$。

2.2 输入权重优化曲面的几何

为了理解模型的学习动力学，我们首先对优化曲面进行解析描述。我们确定极小值、极大值点和鞍点的位置与数量；这些都是梯度下降方法中的临界点。鞍点的导数为零，但在某些参数方向上曲率为正，在另一些方向上则为负。图 1a-c 展示了二维优化曲面上这些特殊点的典型几何形态。

给定具有 $K = N$ 的输入分布，且每个基准方向上各有一个隐藏特征时，优化函数在每个 N 隐藏特征处都有一个极小值$\mathbf{w}^*$ ，即$\mathbf{w}^* = (0, \ldots, 0, \pm 1, 0, \ldots, 0)$，因此总共有 $2N$个极小值。同时，在每个 2^N对称方向处都有指数级多个极大值，$\mathbf{w}^{max} = (\pm 1, \pm 1, \ldots, \pm 1, \pm 1)/\sqrt{N}$。

此外，鞍点的数量还要多得多。虽然难以直观呈现，但任何具有 k对称分量（$1 < k < N$）的方向，若其他分量为零，都是鞍点。对于 k 非零分量而言，它是极大值点；对于其他 $N - k$ 零分量而言，它是极小值。我们在方法部分证明，这意味着鞍点总数的量级为 3^N 。

图 1d 展示了三维情况下的这些性质， $N = 3$。6 个主方向表示极小值点，隐藏特征位于这些方向上。每个方向都被一个吸引域所包围。8 个对称方向是极大值点，而 12 个部分对称点是鞍点。

连接极大值点和鞍点的参数区域梯度较小，且远离极小值。从这些区域开始会导致学习时间较长。请注意，极大值点和鞍点的存在是由问题中的对称性导致的。此前在其他网络模型中也观察到了由对称性导致的鞍点 [5, 6, 17]。鞍点在数值上的占优也与这些其他研究结果一致 [5, 6, 17]。与这些研究不同的是，那些研究中的对称性源于网络层内神经元置换的不变性，而我们的对称性则对应单个神经元的权重。

2.3 准正交 高维空间中的正交随机方向

上文展示的三维示例有助于直观理解投影寻踪特征学习的优化曲面的一些一般性质。然而，高维空间具有一些在低维情形中并不明显、违背直觉的几何性质。我们在分析中将利用的关键性质是随机向量之间的角距离，因为权重初值通常呈随机分布。

在讨论随机向量之前，我们先考虑一种“最坏情况”。在三维情形中，任意向量与某个坐标轴方向之间的角距离至多为 $55°$ ，最大距离出现在极大值点处（最坏情况）。然而，在更高维度中，这一距离渐近趋近于 $90°$；在N 维空间中，极小值与极大值之间的重叠为 $d_{\max} = \mathbf{w}_{\max}^T \mathbf{w}^* = 1/\sqrt{N}$ （见方法部分）。

随机方向$\mathbf{w}^R$ 的平均距离也呈现类似的衰减趋势 d ，即到最近极小值的距离为 $d \approx \frac{\sqrt{2}}{\log(N)\sqrt{N}}$ （见方法部分）。这一结果意味着，初始参数与隐藏特征的重叠程度很小，二者起始时几乎呈正交角。尽管到目前为止，我们假设隐藏特征的数量 K 与输入维度 $K = N$相同，但我们的结果也可以推广到 $K < N$ 和 $K > N$ （见方法部分），此时重叠为

$$d^K \approx \frac{\sqrt{2\ \log(K)}}{\sqrt{N}} \tag{2}$$

因此，当 N 较大时，即使隐藏特征的数量相对于维度数量而言很大（$K > N$），期望重叠仍然很小。这是一个副产品，源于在高维空间中可以生成指数数量的随机向量，并且它们以很高的概率彼此近似正交 [4]。图 2 通过随机生成的方向的模拟展示了这些依赖关系。图 2b 展示了给定输入维数 N 时，初始重叠 d^K 随隐藏特征数量的增加仅呈对数增长。

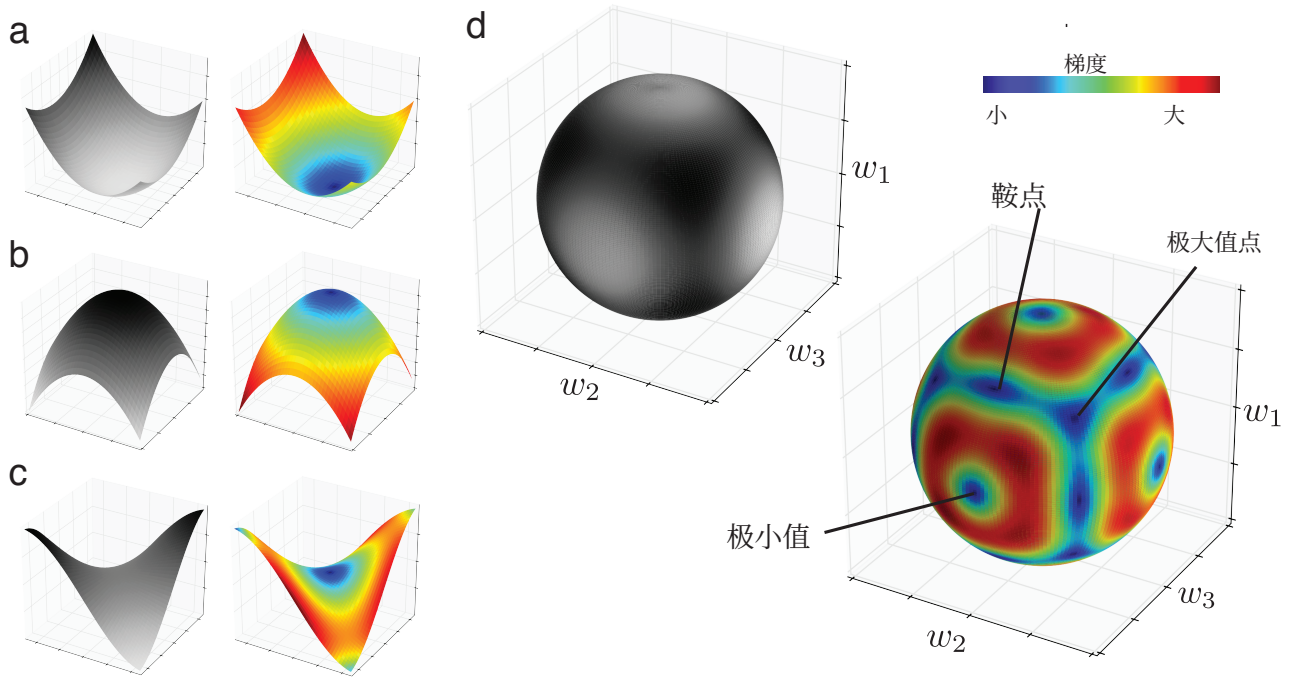

图1. 突触权重优化曲面的几何。 **(a)** 凸曲面上具有吸引域的典型极小值。（左）灰度表示优化函数值，极小值以白色显示。（右）彩色热图表示梯度幅值，零梯度以深蓝色显示。**(b)** 凹曲面上的典型极大值点，并具有一个不稳定平衡点。**(c)** 鞍点在某些参数方向上呈凹，而在其他方向上呈凸。**(d)** 三个权重的梯度幅值曲面。每个维度表示一个突触权重的值。当我们强制满足 $|\mathbf{w}|_2 = 1$ 时，参数空间被约束到单位球面。主方向是极小值，即隐藏模式的方向。对称方向（所有权重具有相同幅值）是极大值点。部分对称方向（其中两个权重具有相同幅值，第三个为零）是鞍点。蓝色区域表示较低的梯度幅值，因而学习动力学较慢。

我们可以将这些事实与从三维情形中获得的直觉联系起来。当维度数量增加时，充满鞍点和极大值点的蓝色区域面积（见图 1d）会呈指数级增长，远远超过由极小值周围吸引域构成的区域面积。形式上，维度越高，给定方向周围的极冠面积相对于球面的总面积会呈指数级减小 [12]。

这些结果表明，当输入权重较多时，初始随机权重与隐藏特征近似正交，并且位于梯度较小的参数区域中，该区域充满了鞍点和极大值点。

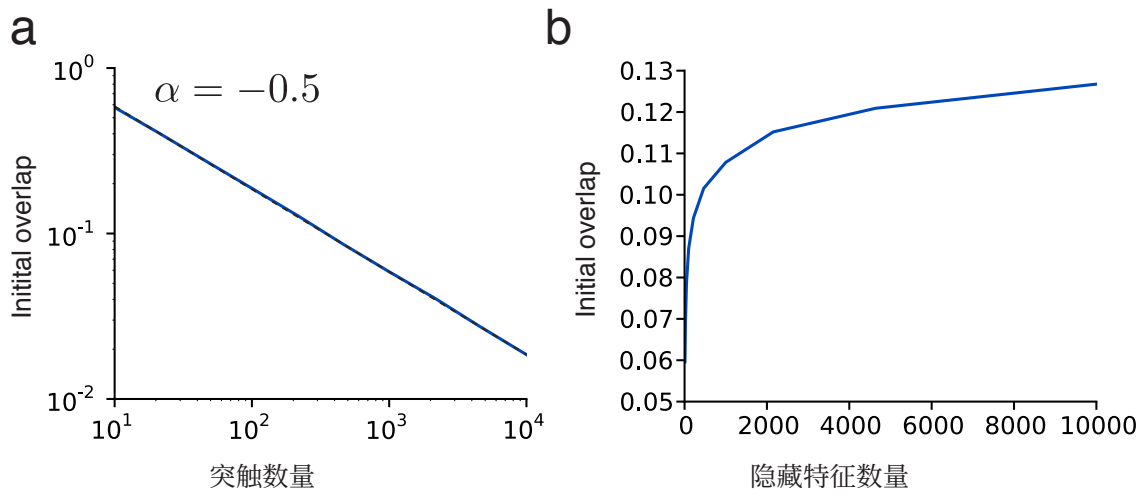

图2. 在高维情形下，初始参数与隐藏特征的重叠很小。（a）初始参数与某一隐藏特征的期望重叠随 $\sqrt{N}$衰减（ $K = 10$固定）。虚线为指数为 $\alpha = -0.5$的幂律。（b）增加隐藏特征的数量对期望初始重叠仅有对数效应（ $N = 1000$固定）。

2.4 约化为一维动力学

掌握了优化问题的一般景观后，我们现在来分析临界点和初始条件附近的梯度。这里假设维度较大， $N \gg 1$ and$K = N$，且初始参数与隐藏特征的重叠较小， $d \to 0$。

在投影寻踪问题中，神经元的总输入是 N 个隐藏模式的混合，并由突触强度加权，即 $u = \mathbf{w}^T\mathbf{x} = w_1x_1 + \cdots + w_Nx_N$。由于当$N$ 较大时，初始的各个权重 w_i^0 很小，我们可以援引中心极限定理，并得出结论：随着 N 增大，总输入 u（$\mathbf{x}$）收敛于正态分布。

我们假设，在学习过程中，权重将收敛到最近的隐藏模式$\mathbf{w}^* = \mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$，并沿最短路径从初始权重$\mathbf{w}^0 = (w_1^0, \ldots, w_{j-1}^0, w_j^0 = d_0, w_{j+1}^0, \ldots, w_N^0)$出发，其中 j 是具有最大重叠的特征， d_0是初始重叠。

我们将输入改写为 $u = w_jx_j + \sum_{j\neq k} w_k\ x_k$，并再次运用中心极限定理，得到近似式 $u = w_jx_j + \sum_{j\neq k} w_k\ g = w_j\ l + \sqrt{1 - w_j^2}\ g$，其中 l 是拉普拉斯变量， g 是正态分布变量；此外，我们利用了权重已归一化这一事实： $\sum w_j^2 = 1$。

因此，沿着从$\mathbf{w}^0$ 到$\mathbf{w}^*$ 的路径，我们可以将 u 的变化简化为初始（高斯）分布与最终（稀疏）分布之间的线性过渡，

$$u_t = d_t\ l + \sqrt{1 - {d_t}^2}\ g \tag{3}$$

由重叠 d 参数化。

式3中的表述表明，输入 u_t 的分布仅取决于重叠d_t。这意味着 $\langle\ F(u)\ \rangle$ 的梯度仅取决于 d_t，从而使我们能够通过单一的一维系统来研究学习˙动力学。取 $F(d_t) = \langle\ F(u_t)\ \rangle$，其中 u_t 由

式3，梯度下降给出

$$\Delta d_t \propto \frac{\partial \hat{F}(d_t)}{\partial d} \tag{4}$$

我们在图3中说明，对于不同的维度 N，完整的 N维学习动力学如何可以由有效一维动力学（式4）表示。图3a展示了最大重叠 d_t 在$\mathbf{w}_t$ 与隐藏特征之间的时间演化。结果表明，初始重叠取决于维度：维度越高，平均初始重叠越低，因而学习时间越长。

这也说明，正如约化到一维系统后所预期的那样，所有运行都具有相似的轨迹形态。在图3b中，我们通过将轨迹的时间参考平移到重叠越过$d_t = 0.75$的时刻，来突出典型学习动力学。

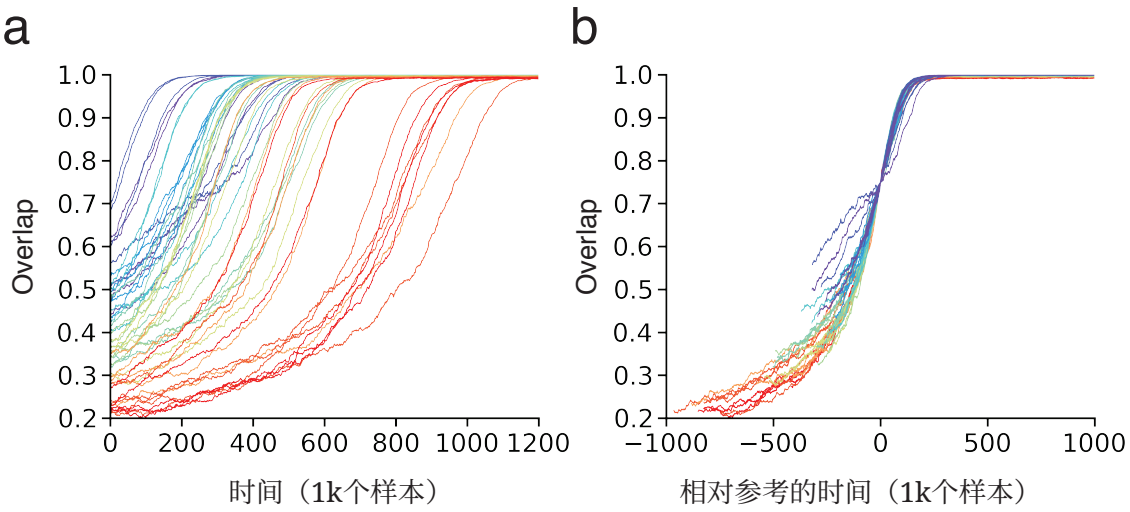

图3. 高维优化中的典型学习动力学。 **(a)** 权重与隐藏特征的重叠在不同维度 N下的演化。颜色热度表示 N，从 $N = 10$ （紫色）变化到 $N = 160$ （红色）。每个N一次运行，采用随机初始权重。**(b)** 将相同轨迹平移到重叠达到 $d = 0.75$ 的参考时间，以突出学习动力学的相似性。

对于较小的重叠，我们可以分析梯度 $\mu(d) = \partial \hat{F}(d_t)/\partial d$的性质。我们对目标函数 $F(d)$ 在 $d = 0$ 附近进行泰勒展开；对于对称特征分布，前三阶矩均为零，因此得到

$$\hat{F}(d) \propto a + d^4 \implies \mu(d) \propto d^3 \tag{5}$$

在图4d中，我们通过在模拟中计算小重叠时的梯度，验证了这一标度律。非对称稀疏分布具有非零三阶矩，因此会产生 $a\mu(d) \propto d^2$ 标度。

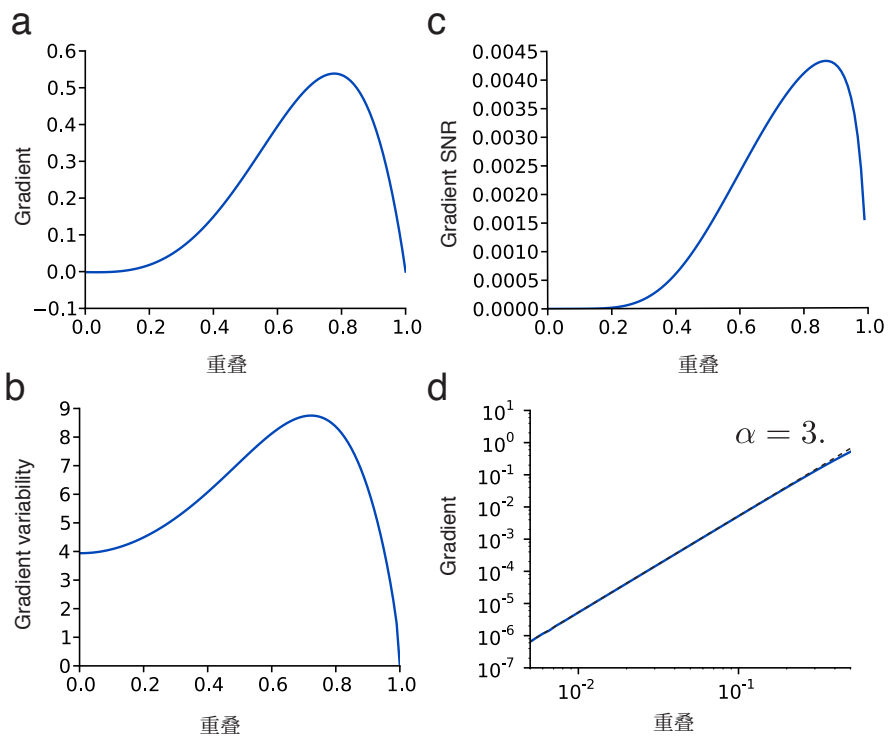

图4. 梯度对重叠的依赖关系 d **（针对对称稀疏分布。）(a)** 当重叠趋于零时，梯度幅值 $\mu(d)$ 消失。**(b)** 对于较小的重叠，梯度变异性 $\sigma(d)$不会发生显著变化。**(c)** 梯度信噪比 $\mu(d)^2/\sigma(d)^2$ 遵循梯度幅值的变化趋势。**(d)** 对于较小的重叠 $d \to 0$，梯度与重叠呈幂律依赖关系（虚线表示指数为$\alpha = 3.$的幂律）。

2.5 学习时间对输入维数的依赖关系

为了估算学习时间，我们考虑给定重叠 d 时的最优学习率 η。经过一次学习和归一化步骤后，我们得到（见方法）：

$$\Delta d \approx \eta\mu - \frac{1}{2}\eta^2 N\sigma^2 d \tag{6}$$

其中正项由梯度 μ 给出，负项则因梯度 σ 的噪声而随 N 缩放，该噪声存在于所有权重维度中。我们推导出最优学习率，以及 d 的最优变化，

$$\eta^* = \frac{2\mu}{N\sigma^2 d} \implies \Delta d \approx \frac{2}{Nd}\frac{\mu^2}{\sigma^2} \tag{7}$$

其中我们可以看到对梯度信噪比的依赖。在图4中，我们通过模拟公式4的简化的一维模型，展示了依赖于重叠 d的梯度统计量。我们看到，对于较小的重叠，梯度信噪比迅速趋近于零（图4c）。此外，信噪比曲线遵循梯度幅值 μ的曲线（图4a），因为在较小重叠时梯度变异性 σ 近似恒定（图4b）。

我们现在可以根据最优学习率下的变化 Δd ，从起点 d_0 到目标 d^*来估计总学习时间，

$$T(d_0) \approx \int_{d_0}^{d^*} \frac{Nd}{2} \frac{\sigma^2}{\mu^2} \partial \tilde{d} \tag{8}$$

我们已成功推导出式 (2)、(8) 和 (5) 中的一系列依赖关系，这些关系共同表明，学习时间对输入维数呈超线性依赖。对于对称情形，我们有

$$\left.\begin{array}{l} d_0 \propto \sqrt{log(K)/N} \\ \mu \propto d^3 \\ T \propto \int_{d_0} \frac{Nd}{2} \frac{\sigma^2}{\mu^2} \, \partial d \end{array}\right\} \Longrightarrow T \propto \frac{N^3}{log(K)^2} \tag{9}$$

而对于非对称情形，当 $\mu \propto d^2$ 时，我们有

$$T \propto \frac{N^2}{log(K)} \tag{10}$$

在图 5 中，我们展示了对于非对称情形，我们的梯度下降模拟的学习时间按 $log^{N^2}_{K()}$ 进行缩放。

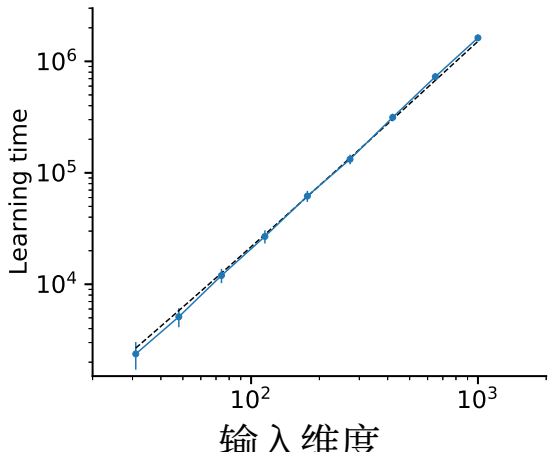

图 5. 学习时间对输入数量的依赖关系。 输入维度为 N、潜在特征为 $K = N$（蓝色）的学习时间与理论预测的标度关系 $T \propto log^{N^2}_{(N)}$一致。学习时间定义为重叠达到 $d = 0.7$ 所需的时间，并对 100 次模拟取平均。

3 讨论

3.1 向替代性神经网络模型的扩展 ls

我们研究了一种特定的无监督学习模型。不过，我们的结果可能也适用于其他学习模型。对于任何总输入是线性投影 $u = \mathbf{w}^T\mathbf{x}$ 的网络模型，突触强度的梯度都可以像我们的研究中那样形式化表示为 $g = \partial_w \left\langle F(\mathbf{w}^T\mathbf{x}) \right\rangle_x$。强化学习或监督学习网络就是重要的例子。

对于我们的无监督学习场景，所作的关键假设是：优化函数仅取决于输入的高阶统计量。未来的研究应确定我们能否将该方法扩展到其他学习范式，在这些范式中，目标值或奖励也发挥作用。

3.2 卷积神经网络理论

大型层次化神经网络模型最近在各种人工智能任务中取得了令人瞩目的成果 [1, 11]。其中很大一部分应用依赖于卷积神经网络的实现，在这类网络中，每个神经元的感受野大小都受到限制 [9, 13]。当输入具有高维度时，例如图像，这种架构约束至关重要。

尽管这种方法广受欢迎且已被证实高效，但对于有限输入维度所带来的功能增益，一直缺乏明确的理论解释。我们的结果为局部感受野和连接约束带来的性能提升提供了一种潜在解释，并通过分析表明，更大的感受野大小可能使学习变得不可行。在后续工作中，我们正在研究所提出的标度律是否适用于大规模卷积网络。

3.3 对神经网络学习动力学的启示

我们所采用的无监督学习范式已被证明是理解神经网络中学习如何展开的一种有用原型。它使我们能够精确计算极值点（如鞍点）的数量、它们的分布方式，以及它们如何决定学习动力学。

我们的曲率分析与先前的研究相一致；这些研究收集了证据，表明鞍点数量众多，并且可能是神经网络学习面临的一项重要障碍 [5, 6,17]。在 Dauphin 等人 [6], 的研究中，作者使用数值模拟来探究大型网络中是否存在鞍点。我们的结果或许有助于为他们的发现提供分析层面的解释。

在 Saxe、McClelland 和 Ganguli [17], 的研究中，作者研究了一种多层线性网络模型，并利用优化梯度对学习动力学进行定性刻画。然而，线性模型的选择导致了不同的结论；他们的结果取决于多层架构的性质，而不取决于每个神经元的学习特性或突触数量。在他们的模型中，单层网络不存在鞍点，优化中的对称性源于多个层，这为大型神经网络中的对称性提供了一个互补来源

。我们的分析能够研究非线性多层网络；在正在进行的工作中，我们探究如何将我们的结果推广到多个层。

3.4 局部感受野与大脑皮层突触数量的理论

皮层神经元拥有大约 1000 至 10000 个输入突触，而由于空间是大脑中的一项强约束，人们认为，连接性特征可能由表征能力与体积之间的权衡决定 [16]。

我们的结果为突触密度提供了另一种权衡。锥体神经元的突触可塑性具有赫布机制，这些机制可以映射为本文研究的非线性赫布学习规则 [3, 15]。对于这类模型，我们已经证明，学习实际上受到每个神经元输入突触数量的限制；即使不存在物理约束，学习时间之间的权衡仍会施加一个数量级为数千个突触的限制。

我们的理论还对感觉发育中的时间尺度和学习动力学作出了预测。给定神经活动的统计特性以及突触可塑性模型，我们可以利用我们的理论估计学习所需的感觉数据量。这将为感觉发育中关键期的时间尺度提供一种具有理论依据的解释[2]。

4 方法

4.1 极大值点与鞍点的数量

我们在此简要说明如何统计对称潜在特征的突触优化曲面上的极大值点与极小值。极大值点位于超立方体的边上，其中所有权重具有相同的幅值 $w_i = \pm 1/\sqrt{N}$。每个权重有两种可能，即负或正，共计 2^N 种组合。

鞍点具有一些为零的权重，而其余权重具有相同的幅值 $|w_{i_1}| = \cdots = |w_{i_k}| \neq 0$。因此，每个权重有三种可能，即正、负或零，共计 3^N种组合。例外是极小值、极大值点以及零向量 $\mathbf{w} = 0$，共计 $3^N - 2^N - 2N - 1$ 个鞍点。

4.2 数据生成

一般而言，我们的无监督学习问题的 N维输入$\mathbf{x}$ 由 K 个非高斯变量的白化线性混合生成，$\mathbf{x} = \sum_{i=1}^{K} \mathbf{w}_i l_i$，其中在对称情形下为归一化拉普拉斯变量（$l_i \sim \text{Laplace}(0, 1/\sqrt{2})$），或在非对称情形下为归一化 χ^2 变量（$l_i \sim \chi_q^2(-q, 1/\sqrt{2q})$，且 $q = 10$），混合向量$\mathbf{w}_i$ 为N维。

对于 $K = N$，我们使用正交混合向量，$\mathbf{w}_i = (0, \ldots, 0, w_i = 1, 0, \ldots, 0)$。对于$K < N$，我们添加高斯变量以补足 N 维，并通过 $\mathbf{x} = \sum_{i=1}^{K} \mathbf{w}_i l_i + \sum_{i=K+1}^{N} \mathbf{w}_i g_i$ 生成输入，其中 $g_i \sim$服从正态分布 (0,1)，混合向量与前面相同

情况。对于 $K > N$，我们通过 $\tilde{\mathbf{x}} = \sum_{i=1}^{K} \mathbf{w}_i l_i$ 生成输入，其中使用满足 $|\mathbf{w}_i| = 1$ 的随机向量 $\mathbf{w}_i$。随后对白化后的数据进行处理，$\mathbf{x} = \mathbf{M}\ \tilde{\mathbf{x}}$。

4.3 高维空间中随机向量之间的期望重叠

两个随机方向之间的平均重叠可以根据随机方向$\mathbf{w}^R$ 在一个 N维球面中的生成方式推导得出：每个分量都取自一个正态分布变量，然后将所得向量归一化为单位范数，

$$\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_N),\ w_i \sim N(0,1) \tag{11}$$

$$\mathbf{w}^R = \frac{\tilde{\mathbf{w}}}{|\tilde{\mathbf{w}}|} \tag{12}$$

由于当 N较大时，范数 $|\tilde{\mathbf{w}}|$ 可以用 $\sqrt{N}$很好地近似，我们得出随机向量的每个分量都服从 $w_i^R \sim N(0,\ \frac{1}{\sqrt{N}})$。不失一般性地，若将参考方向设为$\mathbf{w}^* = (1,\ 0,\ ...,\ 0)$，则重叠服从分布 $d \sim N(0,\ \frac{1}{\sqrt{N}})$。

改为考虑 K 个随机方向$\mathbf{w}^k$，它们与参考方向$\mathbf{w}^* = (1,\ 0,\ ...,\ 0)$之间的最大期望重叠 d^K，由 K个首分量中的最大值$d^K = max_k\ w_1^k \approx \sqrt{\frac{2\log(K)}{N}}$给出；这里对于较大的 K，我们使用了 K 个正态分布变量的期望极值 $max_k\ g_k \approx \sqrt{2logK}$ [7]。

由此可知，若在一个 N 维空间中有 K 个随机隐藏特征，则初始随机参数与其中一个隐藏特征的最大预期重叠约为 $\sqrt{\frac{2\log(K)}{N}}$。

4.4 小重叠的梯度缩放定律

我们分析权重与隐藏特征 $d \to 0$ 之间重叠较小时的梯度幅值。我们考虑投影输入 u 的统计累积量，其中 $u = d\ l + \sqrt{1-d^2}\ g$。由于方差为常数，第二累积量为常数，$\kappa_2^u = 1$。对于 $m > 2$，高阶累积量仅取决于 l、$\kappa_m^u = \kappa_m^l d^m$，因为正态变量 g 的高阶累积量为零，$\kappa_m^g = 0,\ m > 2$。

我们将分析分为两种情况，分别对应隐藏特征 l 的对称分布和非对称分布。对于对称分布，奇数阶累积量 κ_{2m+1}^u 为零。我们计算优化函数在 $u = 0$ 处的泰勒展开，

$$\langle F(u) \rangle = \langle F(0) + uF^{(1)}(0) + u^2F^{(2)}(0)/2 + u^3F^{(3)}(0)/3! + u^4F^{(4)}(0)/4! + ... \rangle \tag{13}$$

$$= a + \frac{F^{(4)}(0)}{4!}\langle u^4 \rangle + ... = a + \frac{F^{(4)}(0)}{4!}\ \kappa_4^l\ d^4 + O(d^6) \tag{14}$$

$$\approx a + b\ d^4 \tag{15}$$

对于某些常数 a 和 b，这表明优化函数对 d存在四阶依赖。对于非对称分布，我们有

$$\langle F(u) \rangle \approx a + b\ d^3 \tag{16}$$

重要的是，这些方程使我们能够将梯度下降形式化地简化为一维问题。令 $\mathbf{w}_j = e_j$ 为最接近初始权重的隐藏特征，使得重叠由权重 $d = w_j$的第 j 个分量给出，并且不失一般性地假设 $w_j > 0$ 。我们发现，在重叠较小时，重叠对应的梯度呈超线性依赖：

$$\langle \hat{F}(d) \rangle \propto a + d^4 \implies \frac{\partial \langle \hat{F}(d) \rangle}{\partial d} \propto d^3 \tag{17}$$

对于对称分布，以及

$$\langle \hat{F}(d) \rangle \propto a + d^3 \implies \frac{\partial \langle \hat{F}(d) \rangle}{\partial d} \propto d^2 \tag{18}$$

对于非对称分布。

4.5 随机梯度的学习时间

我们计算重叠较小时的最优学习率。经过一次更新并进行归一化后，我们得到

$$\mathbf{w}_{t+1} = \frac{\mathbf{w}_t + \eta \mathbf{g_t}}{||\mathbf{w}_t + \eta \mathbf{g_t}||^2} \tag{19}$$

考 虑沿隐藏特征$\mathbf{w}_d$方向的分量 d，我们得到

$$d_{t+1} \approx d_t + \eta\mu - \frac{1}{2}\eta^2 N\sigma^2 d_t \tag{20}$$

$$\implies \Delta d \approx \eta\mu - \frac{1}{2}\eta^2 N\sigma^2 d \tag{21}$$

其中我们利用了 $||\mathbf{w}_t||^2 = 1$，以及 $\mu = \langle \mathbf{w}_d^T \mathbf{g_t} \rangle$, $\sigma^2 = \langle (\mathbf{w}_d^T \mathbf{g_t})^2 \rangle$。对学习一步后的重叠进行优化，我们得到

$$\eta^* = argmax_\eta \langle \Delta d \rangle = \frac{2\mu}{N\sigma^2 d} \tag{22}$$

为了估算学习时间，我们考虑从 d_0 开始，在小重叠 $d \to 0$ 下，分量 d 上的平均动力学，

$$T \approx \int_{d_0} \frac{1}{\Delta d} \partial d \ \propto \ N \int_{d_0} \frac{d}{\mu^2} \partial d \tag{23}$$

其中我们只保留了随 N 缩放的变量。对于具有 K 个隐藏特征的对称情形，我们得到学习时间的标度律，

$$T \propto N \int_{d_0} \frac{d}{d^6} \partial d \ \propto \ N[d^{-4}]_{d=\sqrt{log(K)/N}} \tag{24}$$

$$\implies T \propto \frac{N^3}{log(K)^2} \tag{25}$$

以及非对称情形下，

$$T \propto \frac{N^2}{log(K)} \tag{26}$$

参考

[1] 约书亚·本吉奥、Aaron Courville 和 Pierre Vincent。《表征学习：综述与新视角》。载于：*IEEE* 模式分析与机器智能汇刊35.8（2013），第 1798–1828 页。[2] Nicoletta Berardi、Tommaso Pizzorusso 和 Lamberto Maffei。《感觉发育过程中的关键期》。载于：神经生物学当前观点10.1（2000），第 138–145 页。[3] Carlos S. N. Brito 和 Wulfram Gerstner。《非线性赫布学习：感受野形成的统一原则》。英文。载于：*PLOS*计算生物学12.9（2016 年 9 月），e1005070。[4] Tony Cai、范剑青和Tiefeng Jiang。《球面随机堆积中的角度分布》。载于*arXiv:1306.0256 [*数学、统计学*]*（2013 年 6 月）。[5] Anna Choromanska 等人。《多层网络的损失曲面》。载于：*arXiv* 预印本 *arXiv:1412.0233*（2014）。[6] Yann N Dauphin 等人。《识别并攻克高维非凸优化中的鞍点问题》。载于：神经信息处理系统进展第 *27* 卷。Z. Ghahramani 等人编。Curran Associates, Inc.，2014，第 2933–2941 页。[7] Herbert Aron David 和 Haikady Navada Nagaraja。《次序统计量》。Wiley Online Library，1970。[8] 埃里克·坎德尔、J. H. Schwartz 和 T. M. Jessell。《神经科学原理》。Appleton & Lange，2000。[9] Alex Krizhevsky、Ilya Sutskever 和杰弗里·辛顿。《使用深度卷积神经网络进行 ImageNet 分类》。载于：神经信息处理系统进展。2012，第 1097–1105 页。[10] Quoc V. Le。《通过大规模无监督学习构建高层特征》。载于*2013* 年 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2013，第 8595–8598 页。[11] 杨立昆、约书亚·本吉奥和杰弗里·辛顿。《深度学习》。英文。载于：《自然》521.7553（2015 年 5 月），第 436–444 页。[12] Shengqiao Li。《超球冠面积和体积的简明公式》。载于亚洲数学与统计期刊4.1（2011），第 66–70 页。[13] Volodymyr Mnih 等人。《通过深度强化学习玩雅达利游戏》。载于：*arXiv* 预印本 *arXiv:1312.5602*（2013）。[14] E. Oja、H. Ogawa 和 J. Wangviwattana。《非线性约束赫布网络中的学习》。载于人工神经网络（1991），第 385–390 页。[15] J. P Pfister 和 W. Gerstner。《脉冲时间依赖可塑性模型中的脉冲三联体》。载于神经科学期刊26.38（2006），第 9673–9682 页。[16] Marta Rivera-Alba 等人。《布线经济性和体积排斥决定果蝇大脑中的神经元排布》。载于：当代生物学21.23（2011），第 2000–2005 页。

[17] Andrew M. Saxe、James L. McClelland与Surya Ganguli。“深度线性神经网络中学习非线性动力学的精确解”。载于：*arXiv*预印本 *arXiv:1312.6120*（2013）。