arXiv:2602.02224v1 [cs.LG] 2 Feb 2026

谱叠加：特征几何理论

Georgi Ivanov [1 2 †] **Narmeen Oozeer** [3] **Shivam Raval** [2] **Tasana Pejovic** [3] **Shriyash Upadhyay** [3] **Amir Abdullah** [3 4 ‡]

摘要

神经网络通过叠加表示的特征数量超过其维度，迫使特征共享表征空间。现有方法将激活值分解为稀疏线性特征，却舍弃了几何结构。我们通过分析由权重导出的矩阵的谱（特征值、特征子空间等），建立了一套研究特征几何结构的理论。具体而言，我们引入框架算子 $F = WW^\top$，由此得到一种谱测度，用以描述每个特征如何在各特征子空间间分配范数。以往的工具能够描述特征之间的两两相互作用，而谱方法则能捕捉整体几何结构（“所有特征如何相互作用？”）。在叠加玩具模型中，我们运用这套理论证明，容量饱和会迫使特征局部化：特征坍缩到单个特征子空间上，组织成紧框架，并可通过关联方案进行离散分类，从而涵盖先前研究中的所有几何结构（单形、多边形、反棱柱）。谱测度形式体系适用于任意权重矩阵，因此能够在玩具模型之外诊断特征局部化。这些结果指向一个更广泛的研究方向：将算子理论应用于可解释性。

1. 引言

神经网络在激活空间中的表示特征远多于其维度，这种现象称为 叠加(Elhage 等人，2022年；Chan，2024年）。当模型需要用 f 个特征 d 个维度$f > d$时，它无法为每个特征分配各自正交的方向。相反，特征必须共享表征空间，压缩到彼此重叠的方向中，从而利用可用的高维几何结构。

叠加的直接后果是特征干扰：多个特征争用相同的表征方向，因此修改一个特征不可避免地会扰动其他特征。在实践中，我们发现激活引导方法（Turner 等人，2024年；Rimsky 等人，2024年；Chalnev 等人，2024年）往往会产生意料之外的副作用：推动一个概念时，会无意中改变其他概念（Siu 等人，2025年；Raedler 等人，2025年；Nguyen 等人，2025a）。当前的可解释性方法通过将激活值分解为稀疏、独立的线性特征来处理干扰。稀疏自编码器（SAE）及相关方法试图恢复一个特征字典，使其中的特征在不同输入中稀疏地激活，从而试图消除叠加造成的压缩。我们希望找到一个基底，使每个特征都对应一个单一且可解释的方向。

然而，这种方法舍弃了一些重要的东西。模型可能导致特征共享维度，至少有两个截然不同的原因：

- **偶然干扰**。有些特征在激活时呈负相关，它们很少或从不在同一个输入中共同出现（Lecomte等人，2023年）。如果“海洋生物学”和“抽象代数”都是几乎不会同时出现的稀疏特征，模型就能高效地让它们共用相同的维度，并通过上下文加以区分。这种共享是一种我们希望消除的压缩技巧：这些特征在概念上彼此独立，只是为了提高效率而被打包在一起。
- **结构性干扰**。 其他特征共享维度，是因为数据中存在真实结构，使这种安排顺理成章。星期几可能被编码在同一个子空间中，恰恰是因为模型需要表示它们之间的关系、它们的周期结构和顺序，以及星期一位于星期日和星期二“之间”这一事实（Engels 等人，2024年）。在这里，几何并非需要消除的伪影，而是概念之间关系的一种有意义的表示。

越来越多的研究表明，模型确实编码了特征之间有意义的几何关系，

†通讯作者。‡资深作者。 [1]Theopha [2]哈佛大学 [3]火星人 [4] Thoughtworks。联系邮箱：<georgi@theopha.com>。

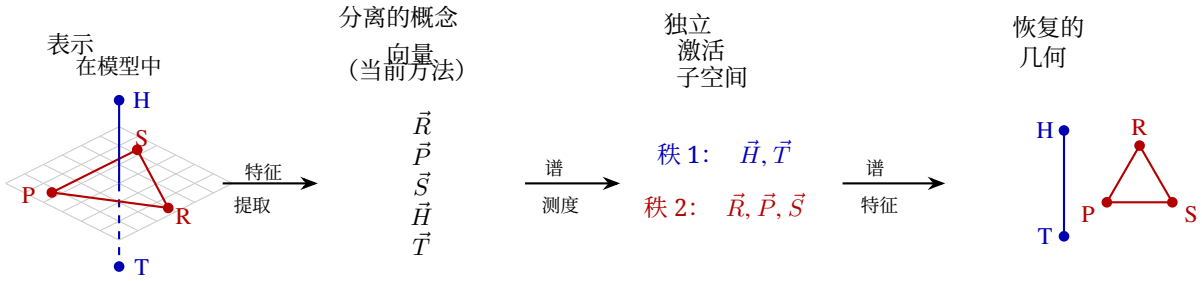


图*1*。 该示意图展示了“石头(R)/布(P)/剪刀(S)”和“正面(H)/反面(T)”嵌入三维空间的概念。这是结构性干扰的一个例子：模型会使彼此相关的概念相互干扰（手势游戏特征会相互干扰，硬币特征也会基于各自集合中对象之间的关系而相互干扰，但两个集合之间不会相互干扰）。当前仅提取特征的方法会生成一个包含5个向量的基。我们使用谱测度，按照特征所在的特征子空间对激活空间中的特征分组，从而捕捉特征之间的关系。当每个特征都谱局域化（占据单个特征子空间）时，我们可以通过补充谱特征来覆盖并恢复完整的几何结构。算子理论中的这些工具让我们得以探索干扰的整体特性：不仅了解成对特征如何相互作用，还能了解所有 特征之间的关系。

这说明模型存在稀疏线性分解所忽略的结构共享。求解模运算的模型会在表征空间中描绘螺旋线（Nanda 和 Chan，2025年）。空间推理任务会产生几何嵌入（Gurnee 和 Tegmark，2024年）。Evo 2 DNA基础模型会根据生物物种之间的系统发育关系，在曲流形上组织这些物种（Nguyen 等人，2025b）。在不同领域和架构中也观察到了类似的几何组织现象（Park 等人，2025年；Anthropic，2025年；Bhaskara 等人，2025年；Google，2025年；Conmy 等人，2025年；Pan 等人，2025年；Nguyen 和 Mo，2025年）。这些发现表明，几何并非偶然产物，而是特征共存与相互作用的核心机制。

当特征嵌入具有非平凡几何结构时，将它们视为独立方向会丢失信息。考虑一个简化示例（图 1）：模型可能如何表示游戏玩法——五个特征分别代表“石头”“布”“剪刀”“正面”和“反面”，并嵌入三维空间。石头剪刀布特征在同一平面内构成一个三角形；正反面特征则位于另一条独立的轴上。这就是结构性干扰：手势游戏特征共享一个子空间，因为它们之间的关系很重要；硬币特征也同样如此，但这两组特征彼此正交。当前的特征提取方法会返回五个独立向量。几何结构——哪些特征共享子空间，哪些特征彼此正交，以及它们如何排列——都丢失了。

我们希望还原这种全局结构：不仅要研究孤立的特征，还要了解它们如何共同组织成可能承载功能意义的子几何结构。Prakash 等人(2025年) 表明，要在 Gemma2-2B-Instruct 中诱发拒绝行为，需要消融 2,538 个 SAE特征；当其他特征受到抑制时，“备用”特征会被激活。这表明，模型的功能存在于集体几何结构之中，而非单个方向。

我们的贡献如下：

- 我们通过深入理解我们的机会游戏示例，阐明了使用谱理论（特征子空间等）研究干扰的动机：只有当特征共享激活空间中的子空间（即特征子空间）时，才会发生干扰。
- 我们建立了叠加的谱理论，提供了可推广的工具，可应用于玩具模型和现实场景，以研究特征几何。
- 我们将这些工具应用于经典的叠加玩具模型（Elhage 等人，2022年），完整刻画了在该场景中形成的特征几何。

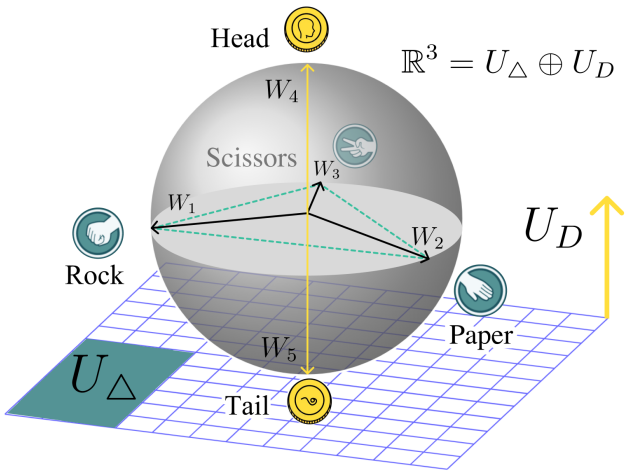
2. 谱方法的动机

本节将通过更加仔细地考察我们的机会游戏示例（图 2），阐明使用谱理论研究特征几何的动机。该示例有三个值得注意的特征：

1. **几何结构具有置换不变性**。在簇内置换特征向量，不会改变全局结构和相互作用。2. **正反面和石头剪刀布的子空间彼此独立**。 任何抛硬币结果都无法提高玩石头剪刀布的胜算，反之亦然，因此这些概念编码在正交子空间中。

3. **特征的几何排列编码了精确的对称性**。在石头剪刀布中：循环规则规定石头胜剪刀，剪刀胜布，布胜石头。在正反面中：正面或反面构成二元对立。

研究特征之间干扰的经典方法是格拉姆矩阵 $M = W^\top W$，其中每个元素$M_{ij} = \langle W_i,\ W_j\rangle = W_i^\top W_j$都是列向量之间的欧几里得范数。我们可以尝试用这种方法来确定哪些向量会相互作用，哪些不会。



$$W = \begin{bmatrix} \vdots & & \vdots \\ W_1 & \cdots & W_5 \\ \vdots & & \vdots \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -1 & 0 & 0 \\ \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

图 *2.* 更详细地从几何角度和矩阵形式来看机会游戏的例子。这里，$W_\triangle = [W_1, W_2, W_3]$ 分别表示“石头”“布”和“剪刀”这几个特征概念，它们在 xy 平面上排列成一个等边三角形。同样，$W_D = [W_4, W_5]$ 对应“正面”和“反面”，并沿着 z 轴构成一对对跖点。

鉴于我们将 W（图 2）写成这样，以下操作可行：直接进行矩阵乘法可得一个分块矩阵 $M = M_\triangle \oplus M_D =$（$W_\triangle^\top W_\triangle$）$\oplus$（$W_D^\top W_D$）。不过，这种简洁的块对角结构是索引选择造成的假象。只有在局部特征簇 A.1（通过矩阵 P_γ）内进行的置换 $\Omega_\triangle = \{1,2,3\}$ 和 $\Omega_D = \{4,5\}$ 才会保持分块分解不变。

设想我们重新排列列向量或旋转矩阵；在 W 或 M中，这种结构将不再直观可见，但底层几何保持不变。

本节的目标是找到一种方法，以基不变的方式表示 W 中特征的全局对称性（从而捕捉几何结构）。谱分析恰好能让我们做到这一点。

2.1. 几何不变性

概括而言，我们的论证如下：几何结构会产生对称性约束；这些约束可以表示为矩阵之间的代数关系；而这些关系完全由谱数据刻画。在本小节中，我们假设几何已知，并说明谱结构如何编码几何。在第 2.2 小节中，我们反向推导：仅给定权重矩阵时，如何从激活空间的谱中恢复几何。

我们已经确定，需要以置换不变的方式获取有关干涉的信息。因此，考虑任意的列标号 $\Omega_\triangle = \{i_1, i_2, i_3\}$ 和 $\Omega_D = \{i_4, i_5\}$，以及 $W_\triangle = [W_{i_1}, W_{i_2}, W_{i_3}]$ 和 $W_D = [W_{i_4}, W_{i_5}]$。

保持等边三角形不变的群是二面体群 $D_3 \cong S_3$。我们可以用置换记号表示这些元素 $\sigma = (i\ j\ k)$，其中 $i \to j \to k \to i$表示下标被交换的顺序（图 2）。

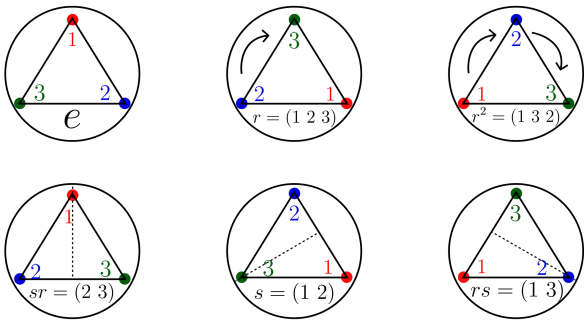


图 *3.* D_3 中各元素的置换记号。

我们可以用 Ω 上的邻接矩阵来表示群作用 $\times\Omega$；每个矩阵代表稳定轨道（对单个指标施加所有可能的置换后形成的轨迹）$\sigma \in D_3$。在我们的语境中，每个顶点都可在 1 步之内到达。因此，除表示静止轨道$\mathcal{O}_0 = \text{diag}(\Omega)$ 的平凡邻接矩阵 $A_0 = I_3$外，我们只有 $A_1 = J_3 - I_3$ (见 A.3.1)，其中 $J_3 \in \mathbb{R}^3$ 为全1矩阵。我们将关注 $\mathcal{A}_\triangle = \text{span}\{A_0, A_1\}$，称为 Bose-Mesner 代数。为什么要研究这个代数，而不直接对 $M_\triangle$ 进行对角化？答案是，$\mathcal{A}_\triangle$ 恰好包含与所有对称运算都可交换的矩阵——它是 D_3 的中心化子（见 A.1）。任何 D_3-不变算子，包括 $M_\triangle$，都必须属于这个代数，因此可在其谱投影算子 S_0、S_1上分解。这比一般的对角化更有力：它不仅告诉我们特征值是什么，还揭示了这些特征值为何源自对称性。Bose-Mesner 代数是关联方案理论中的核心对象（Bailey，2004），并且在层 $\mathbb{R}^{\Omega_\triangle} \cong \mathbb{R}^3 \cong V_0 \oplus V_1$上具有自身的谱理论（引理 A.5）。相应的层投影算子 S_0和 S_1 构成了 $\mathcal{A}$更方便的基。它们之间存在正交关系（引理 A.6）：

$$A_i = \sum_{e=0}^{1} C(i,e) S_e, \quad S_e = \sum_{i=0}^{1} D(e,i) A_i \tag{2.1}$$

$C(i,e)$ 的各项是 A_i 在 W_e上的特征值，构成特征标表 $C \in \mathbb{R}^{2\times 2}$。 $D(e,i)$ 的各项构成一个矩阵，该矩阵是 C 的逆矩阵。同一引理还指出 $S_0 = |\Omega_\triangle|^{-1} J = J/3$ ，并给出了部分系数的明确取值，例如$C(0,0) = C(0,1) = 1$ 以及 $C(1,0) = a_i = 2$。这里， a_i是非平凡轨道的度数，表示在群作用下一个指标可以到达的邻居数量。利用后者的信息以及（2.1），我们可以明确计算出 $S_1 = A_0 - S_0 = I_3 - \frac{1}{3}J_3$。借助这两个投影矩阵，我们可以还原 $V_0 = \mathrm{span}(\mathbf{1}_3)$ 和 $V_1 = \mathbf{1}_3^{\perp}$。最后，通过求解 $A_1 = 2S_0 + C(1,1)S_1$，结合引理 A.6 给出的 $C(1,1) = -1$ 及其对应的 D，即可补全特征标表：

$$C = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}, \quad D = \frac{1}{3}\begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}$$

利用特征标以及 $M_\triangle$ 具有 D_3不变性（证明见 A.3.1），我们可以将其表示在基 A_i 下：

$$M_\triangle = I - \frac{1}{2}(J - I) = \frac{3}{2}I - \frac{1}{2}J = \frac{3}{2}S_1$$

现在，我们来看二边形，其对应的群为 $\Gamma \cong S_3 \cong C_2$。这是一个含有两个元素$\{e, s\}$的循环群，即单位元 e 以及 s ，后者的阶为 2，这意味着 $s^2 = 1$。从图形上看，这表示交换二边形的两个节点（也就是从正面-反面变为反面-正面）。根据这是一个单纯形这一事实（即可以从一个节点到达其他任何节点），我们有相同数量的轨道/关联类$\mathcal{C}_0 = \mathrm{diag}(\Omega_D)$ 和 $\mathcal{C}_1 = \{(4,5);(4,5)\}$，其邻接矩阵分别为 $A'_0 = I_2$和 $A'_1 = J_2 - I_2$。类似地，将引理 A.5 应用于 $\mathcal{A}_D = \mathrm{span}(A'_0, A'_1)$，得到 $\mathbb{R}^{\Omega_D} \cong \mathbb{R}^2 = U_0 \oplus U_1$；根据引理 A.6 $S'_0 = \frac{1}{2}I_2$， $S'_1 = I - \frac{1}{2}J$。如下图所示，空间 $\mathbb{R}^{\Omega_\triangle}$（左）和 $\mathbb{R}^{\Omega_D}$（右）分解为质心子空间（红色箭头）和差子空间（阴影平面/黄色对角线，对应于 S_1 和S'_1）。恒等式 $M_\triangle = \frac{3}{2}S_1$ 和秩$(M_\triangle) = 2$ 表明，在每个对称簇中，格拉姆算子都起到投影算子的作用：它以特征值 0 消去质心，并分别将差平面缩放 $\lambda_\triangle = \frac{3}{2}$ 或 $\lambda_D = 2$。由此立即可知，内在秩降低：秩$(M_\triangle) = 2$，且秩$(M_D) = 1$。

然而，一旦特征索引被打乱，这一描述在Gram矩阵条目层面便不再可见。经索引置换后仍然留存的，只有谱内容（特征值/特征子空间，直至重新标记），因此我们接下来转向一个本身在重新标记下保持不变的激活空间算子。

2.2. 谱桥梁

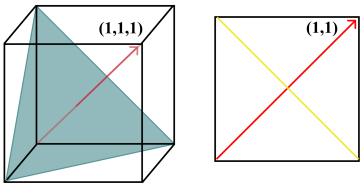


图 4. 关联方案分层

第 2.1 节表明，已知 几何被编码于谱数据之中。但在实践中，我们只能观测到权重矩阵 W，而看不到底层的特征结构。如何在事先并不知晓几何的情况下将其恢复？关键在于在激活空间而非索引空间中开展工作。格拉姆矩阵 $M = W^\top W$ 在索引置换下按 $M \mapsto P^\top M P$变换，从而打乱其条目。相比之下，框架算子 $F = WW^\top$是不变的：$(WP)(WP)^\top = WW^\top = F$。这使得 F成为基不变几何分析的自然对象。

到目前为止，我们将底层对称性描述为 Gram 的谱模态 M （依赖于索引），位于抽象的概念/索引空间 $\mathbb{R}^{\Omega_\triangle} \oplus \mathbb{R}^{\Omega_D}$。然而，实际决定干扰与容量的几何存在于激活空间 $\mathbb{R}^3$。打乱特征索引对应于右乘 W 一个置换矩阵P_γ，即 $W \to WP_\gamma$ （证明见 A.1）。这会共轭 Gram$M \to P^\top M P$，从而掩盖各项中的局部块结构。因此，若我们想要一种真正与标签无关的几何引导的干涉描述，就必须转向一个激活空间对象，它在此类重新标记下保持不变，并保留我们此前发现的谱信息。典范的此类对象是框架算子$F := WW^\top \in \mathbb{R}^{3\times 3}$，它在索引置换下保持不变 $(WP_\gamma)(WP_\gamma)^\top = WP_\gamma P_\gamma^\top W^\top = WW^\top = F$。尽管将 $M_\triangle$与 M_D 在概念空间中重新拼接起来并非平凡（由于索引具有任意性），但对框架算子 $F = \sum_{i=1}^{n} W_i W_i^\top$而言，这一过程则变得直接：

$$F = \sum_{i\in\Omega_\triangle} W_i W_i^\top + \sum_{i\in\Omega_D} W_i W_i^\top = F_\triangle + F_D \tag{2.2}$$

我们使用框架算子，是因为 M 与 F之间存在精确的谱对应。设 $M = \sum_e \lambda_e E_e$ 为格拉姆矩阵的谱分解B.1。那么，根据谱对应引理B.3， M 与 F 具有相同的非零特征值 $\lambda_e > 0$ ，且E_e 的本原幂等元 M 可提升为

$$P_e = \lambda_e^{-1} W E_e W^\top. \tag{2.3}$$

我们可以应用此引理，推导出一个推论，从而确立局部簇投影算子与框架分解之间的精确关系（推论6）： $F =$

$2P_D + \frac{3}{2}P_\triangle$，其中 P_D 和 $P_\triangle$ 分别是投影到二边形和三角形所在子空间上的投影算子。此外，对于 $C \in \{\triangle, D\}$，我们有以下推论：$\frac{1}{\lambda_C} = \frac{\text{维数}(U_C)\sum_{I\in\Omega_C}\|W_i\|^2}{\text{秩}(P_C)\,|\Omega_C|}$，也就是说，分数维数 $D_i = d/p$（表示共享维数的特征所占的比例）可以从谱特征中恢复。在下一节中，我们将详细分析这一结果的含义，以及此处采用的思路如何推广到任意权重矩阵的情形。

3. 可泛化工具

在上一节中，我们展示了如何从潜在空间中的框架算子中恢复几何，而该算子能够编码结构化表示。本节将重点介绍上一节中一组可推广到任意权重矩阵的工具。具体而言，我们将展示，特征沿特征子空间排列这一事实如何决定它们的相互作用模式（由谱桥梁给出）、单个特征所消耗的容量（分数维数），以及该特征的局域程度。

谱桥梁： 第 2 节引入的谱桥梁可推广到任意权重矩阵 W，并对输入空间（通过格拉姆矩阵 M）与激活空间（通过框架算子F）之间的交织关系进行分类。我们能够根据 F的谱行为恢复几何，并借助关联方案（附录A.1）根据其不变性对其分类。更具体地说，W 使 M 与 F的分解相互交织。因此，M 与 F 的非零特征子空间一一对应。将框架算子分解为正交子空间还有一个同样重要的结果：特征交互仅限于共享的谱子空间之内。支撑集位于同一诱导特征子空间中的分量可以相互作用，支撑集位于不同特征子空间中的分量则彼此不相互作用；而与 $\mathrm{Im}(W)$ 正交的方向不起作用，也就是说，它们会被 $W^\top$ 消去，因此不参与特征交互或容量消耗。

谱测度： 全局经验谱密度（Martin 和 Mahoney，2021a年）描述的是容量存在于何处，而谱测度的集合描述的是各个特征如何使用该容量。因此，两个模型可能呈现出相似的重尾全局谱，同时在特征局部化、干扰和有效维度方面具有截然不同的模式，而这些由谱测度揭示。

本文主要研究每个特征的维数；该概念在《玩具模型》中定义（Elhage 等人，2022年）用于研究编码特征之间的相互作用 s:

$$D_i = \frac{\|W_i\|^4}{\sum_{j=1}^{n}\left(W_i^\top W_j\right)^2} \tag{3.1}$$

它表示与其他特征的叠加程度，其中 $D_i \approx 1$ 表示特征向量 W_i 拥有其专属维度（即它近似正交于W_j，其中 $i \neq j$），而 $0 \leq D_i \ll 1$ 表示它在叠加中与许多其他特征共同相互作用。当前形式隐藏了一些更有趣的关系。我们可以将其改写为 Gram M 与框架算子 F 矩阵的表达式，如下所示（引理B.9和B.10）：

$$D_i = \frac{(M_{ii}^2)}{(M^2)_{ii}} = \frac{1}{\kappa_i}\|W_i\|^2, \tag{3.2}$$

其中 $\kappa_i = \frac{W_i^\top F W_i}{\|W_i\|^2}$ 是 F在 W_i处的瑞利商。注意，$|||W_i||| > 0$ 在初始化时以概率1 成立，并且在我们观察到的容量饱和下，它对秩坍塌是稳定的（在注意到 ϵu 与 $\epsilon \to 0$保持界B.13之后）。

我们可以将 Gram 第一种形式视为由成对相关性导出的信噪比，而第二种形式则将其重新表述为一个谱问题。现在我们引入谱测度，它的作用是衡量特征如何将其质量分配到各个特征模态上。

考虑权重矩阵及其对应的 Gram 算子和框架算子 M 和 F。如前所示，激活空间中所有非平凡的交互结构都由谱分解（公式2.3）所刻画，其中 $P_e(t)$ 是相互正交的投影算子，对应于独立的谱模态。为描述单个特征如何参与这一结构，对于每个特征向量 W_i，我们定义$p_{i,\,e} := \frac{\|P_e W_i\|^2}{\|W_i\|^2}$ 它衡量该特征的平方范数中有多少比例支撑在谱模态 e上。我们注意到，这些系数均为非负数，且引理B.11表明它们的和为一，因此构成一个概率分布。这促使我们定义谱测度

$$\mu_i(t) := \sum_{e\in\mathcal{E}_+} p_{i,e}(t)\,\delta_{\lambda_e(t)}. \tag{3.3}$$

谱测度简洁地描述了特征如何将其质量分配到不同的交互模态中。此外，现在可以从概率角度解读算子层面的量。引理 B.2 表明，μ_i 的所有矩都可以通过框架算子的幂求得。特别是，Rayleigh 商可表示为

p y g
$\kappa_i = \frac{W_i^\top F W_i}{\|W_i\|^2} = \mathbb{E}_{\mu_i}[\lambda].$

最后，分数维数可以写为

$$D_i(t) = \frac{\|W_i(t)\|^2}{\mathbb{E}_{\mu_i(t)}[\lambda]}, \tag{3.4}$$

这明确表明，有效维度取决于一个特征如何在整个谱中分配其质量。虽然经验谱分布提供了一个全局概述，说明容量如何分布在不同谱尺度上，但它并未描述各个特征如何参与这一结构。为此，我们的框架引入了一种针对每个特征的谱度量。对于每个特征向量 W_i，系数$p_{i,\ e}$ 在公式3.3 中构成了一个关于谱模态的概率分布，并量化了该特征的平方范数在各个特征子空间中的支撑比例。

这一构造使我们能够在单个特征的层面区分两种状态。如果某个特征的谱测度 μ_i 将质量分散到多个特征值上，表明它参与了多种交互模态，那么就称该特征为离域。相反，如果 μ_i 集中于少数几个特征值，或收缩为单个狄拉克质量，则称该特征为局域。在极端情况下$\mu_i = \delta_\lambda$，该特征完全位于框架算子的单个特征子空间内。

4. 求解叠加玩具模型

本节将展示，如何利用上述谱框架中的一些工具，彻底解释玩具模型中的叠加现象。我们首先通过实证观察发现，训练后的模型运行在接近容量饱和的状态，这促使我们考察这样一种状态：分数维数耗尽权重矩阵的秩。在这一假设下，我们证明所有特征都必须在谱上局域化，收缩至框架算子的各个特征子空间（定理 1）。局域化后，特征不再在不同特征子空间之间分配容量：分数维数由特征范数和谱尺度线性决定（推论 2），而每个特征子空间内的特征则组织成紧框架（定理 3）。这种紧框架结构使我们能够通过关联方案识别特征几何（定理 C），从而实现离散几何分类。最后，我们将这些静态结果与训练动态联系起来，说明梯度流如何引发谱质量传输和特征值漂移，以及为何容量饱和的紧框架构型会成为稳定的不动点。综上，这些结果完整地揭示了在谱约束下，叠加如何演化为结构化且稳定的几何。

Anthropic 的《叠加玩具模型》（TMS）论文（Elhage 等人，2022年）是理解多义性的奠基性技术贡献之一。研究者通过研究合成输入向量 x 来探索这一现象，这些向量模拟了底层特征的属性。每个向量都有对应的稀疏度 S_i （即$\mathbb{P}(x_i = 0) = S_i$ 以及 $x_i \sim$ 均匀分布 (0,1）（否则）和重要性 I_i，对应每个维度 x_i。将投影定义为线性映射 $h = Wx$，其中每一列 W_i 对应于低维空间中表示某个特征 x_i 的方向。为了恢复原始向量，我们使用转置 W^T，得到重构 x'。加入 ReLU 非线性后，模型变为$x' = \text{ReLU}(W^T h + b) = \text{ReLU}(W^T W x + b)$。

实验设置

为研究玩具模型中叠加现象的出现，我们正在进行一项包含 3,200 次实验的高通量参数扫描，采用按特征重要性加权的均方误差（MSE） $L = \sum_x \sum_i I_i (x_i - x'_i)^2$，与原始论文（Elhage 等人，2022年）一致。我们将输入维度固定为 $n = 1024$ 个特征，并为 $M \in [16, 512]$ 线性抽样 32 个离散值，为稀疏度 $(S \in [0.0, 0.99])$ 抽样 50 个离散值(每次运行使用 2 个随机种子）。我们设定均匀重要性 $I_i = I = 1$ 和均匀稀疏性 $S_i = S$，但在附录 G 中讨论非均匀稀疏性的结果。

4.1. 容量饱和即谱测度局域化

我们的第一个重要经验观察$(E1)$ 是，玩具模型使用了其隐藏维度容量中几乎全部的容量 $\sum_i D_i \approx m$。下图展示了所有 3200 次训练运行中各个特征分数维数之和（定义见3.2）：

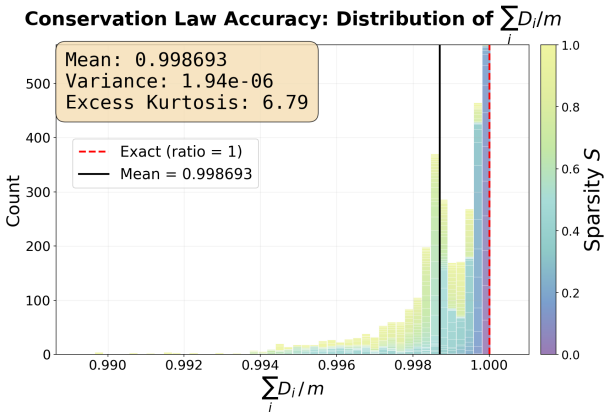


图 5. 3,200 次运行中的容量饱和

这是我们用来论证理论分析及后续预测的最重要事实。它大幅降低了训练动态的复杂度，具体而言，它迫使所有单个特征都局限于框架算子的一个且仅一个**特征子空间**，该子空间由相应的投影算子 P_k刻画：

定理 1（谱局域化）。 假设模型达到了分数维数容量上界，即 $\sum_{i=1}^n D_i = \text{秩}(W) = m$。那么，对于每个特征 i，谱测度都会收缩为单个狄拉克质量 $\mu_i = \delta_{\lambda_k}$，其中心位于某个特征值——

值 $\lambda_k > 0$。这等价于 $FW_i = \lambda_k W_i$。

证明。 定义杠杆分数 $\ell_i := W_i^\top F^+ W_i$，于是 $\sum_i \ell_i = \mathrm{tr}(W^\top F^+ W) = \mathrm{tr}(F^+ F) = \mathrm{rank}(W)$。我们可以对每个特征$W_i \in \mathrm{Im}(F)$ 应用引理 B.13，利用 $a := F^{1/2}x$和 $b := F^{+/2}x$:

$$\|W_i\|^4 \le (W_i^\top F W_i)(W_i^\top F^+ W_i),$$

重新排列后，我们得到每个特征的分数维数上界 $D_i = \|W_i\|^4/(W_i^\top F W_i) \le \ell_i$。对所有特征求和，得到界 $\sum_i D_i \le \sum_i \ell_i = \mathrm{rank}(W)$。这意味着，对 $\sum_i D_i = \mathrm{rank}(W) = m$ 的解读是：所有分数维数界都达到饱和，即 $D_i = \ell_i$。根据引理 B.16 以及μ_i的有限原子性（在引理 B.14 中形式化），$\kappa_i = \frac{W_i^\top F W_i}{\|W_i\|^2} = \mathbb{E}_{\mu_i}[\lambda]$和 $\frac{\ell_i}{\|W_i\|^2} = \frac{W_i^\top F^+ W_i}{\|W_i\|^2} = \mathbb{E}_{\mu_i}[\lambda^+]$.因此 $D_i = \|W_i\|^2/\mathbb{E}_{\mu_i}[\lambda^+]$。尽管我们现在已经通过谱测度重新诠释了 D_i 和 ℓ_i，但前者的柯西–施瓦茨（CW）界是在 $(\mathbb{R}^m, \langle\cdot,\cdot\rangle)$ 中，而后者则位于 $L^2(\mu_i)$ 中（引理 B.14），并且依赖于特征。根本的不匹配在于，特征向量 W_i 属于 $\mathbb{R}^m$，而控制容量和干扰的量则是对 F特征值的谱平均。这进一步证明了有必要将 W视为交织算子。弥合两者的方法是定义循环子空间 $\mathcal{K}_x := \mathrm{span}\{x, Fx, F^2x, \dots\}$，它是 $\mathbb{R}^m$ 空间中最小的 F-不变切片，能够涵盖所有二次型 $W_i^\top h(F) W_i$ 以及向量 $F^{1/2}W_i, F^{+/2}W_i$。根据循环空间等距映射 $U_{W_i}: L^2(\mu_i) \to \mathcal{K}_{W_i}$（引理 B.16），函数 $u(\lambda) = \sqrt{\lambda}$和 $v(\lambda) = \lambda^{+/2}$ 分别与向量$F^{1/2}W_i/\|W_i\|$ 和 $F^{+/2}W_i/\|W_i\|$完全对应。因此，$L^2(\mu_i)$ 中的柯西–施瓦茨（CW）界实际上就是 $\mathcal{K}_{W_i} \subseteq \mathbb{R}^m$上的欧几里得柯西–施瓦茨不等式，这使得等式 $D_i = \ell_i$ 等价于$\mathbb{E}_{\mu_i}[\lambda]\mathbb{E}_{\mu_i}[\lambda^+] = 1$，因为 $\langle u(\lambda)\, v(\lambda)\rangle_{L^2(\mu_i)} = 1$（引理 B.15）。当且仅当$\sqrt{\lambda} = c\lambda^{+/2}$ μ_i-几乎处处成立时，该等式成立，也就是说，若 λ 几乎处处为 μ_i常数。因此，μ_i 收缩为狄拉克质量 $\delta_{\lambda(i)}$，这意味着 W_i 完全属于 F的某个正特征值特征空间，等价地，$FW_i = \lambda(i) W_i$。证毕。

□

这一结果实际上相当有力。它让我们能够立即把握特征的全局几何。第一个推论（证明见 C）是：如果谱局域化成立，则分数维数与范数呈线性缩放关系，其中斜率对应于特征值的倒数 λ_k:

推论2 （射影线性）。 假设谱局域化成立。那么，任意特征的分数维数 D_i 均由其特征范数 $\propto \|W_i\|^2$以线性方式确定，斜率 k 等于该特征所处子空间的特征值 λ_e 的倒数 $D_i = \lambda_e^+ \|W_i\|^2$

由于模型的容量已饱和，因此其秩也已饱和（如上所证），这意味着我们可以利用这些信息，恢复该特征子空间中所含特征的精确密度（进而恢复其分数维数）。为考察射影线性推论在整个训练扫描中的有效性，我们分析了来自 3,200 次实验运行（共计 3,276,800 个特征）的特征簇的几何稳定性。对于每个特征 w_i，我们计算了其谱测度权重 $p_{i,\ e} = \|P_e w_i\|^2/\|w_i\|^2$。我们通过簇中各特征的最大谱质量均值，量化每个簇 C 的谱局域化程度，定义为 $\mathcal{L}_C = \frac{1}{|C|}\sum_{i\in C} \max_e(p_{i,\ e})$。我们将其绘制为线性决定系数 R^2的函数。误差棒表示每个簇的绝对误差 $|k\lambda - 1|$：

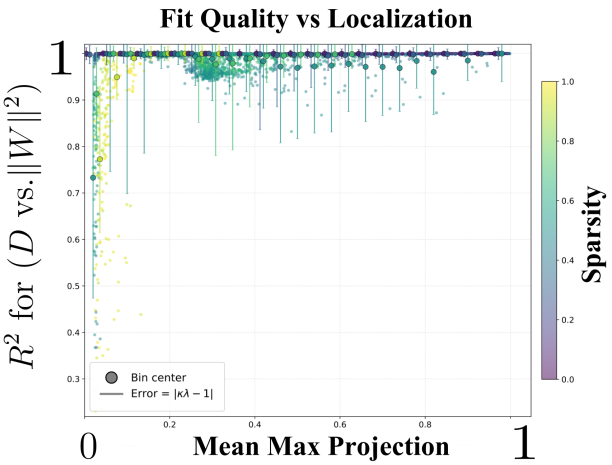


图6。射影线性测试：纵轴表示线性拟合的决定系数 R^2 λ_k^{-1}；误差线表示绝对误差 $|k\lambda - 1|$；x轴表示谱测度局域化的平均最大值指标

误差棒上新增的梯度着色揭示了特征值斜率拟合与特征稀疏度之间的紧密关系——较低的稀疏度会因更强的干扰而迫使特征形成更紧凑的配置（因此也形成更紧凑的特征子空间），而稀疏度较高的特征则在整个谱中分布得更分散。这一观察与我们针对容量界微小偏差的扰动分析相符（附录 $\sigma_i = 1 - D_i/\ell_i$ D）。这也解释了守恒定律中的两个峰值，它们按照稀疏度明显分层。虽然所有实验中的松弛量都很小，但高稀疏度区域的轻微偏差会导致

少数特征广泛离域化。谱局域化也会直接影响模型所学习的几何类型（命题 C）：

定理 3 (紧框架分解)。 假设所有特征 $i \in \{1, \cdots, n\}$均满足谱局域化。令 $\Lambda = \{\lambda_1, \cdots, \lambda_k\}$ 为$F = WW^\top$的不同特征值集合。那么，特征集可划分为互不相交的簇 $C_k = \{i|\lambda(i) = \lambda_k\}$ 和 $F = \bigoplus_{k=1}^m \lambda_k P_{V_k}$，其中 $V_k =$ 张成$\{W_i|i \in C_k\}$。此外，在每个簇 C_k内，权重子矩阵 W_{C_k} 构成一个紧框架：$\sum_{i \in C_k} W_i W_i^\top = \lambda_k I_k$ 在 V_k 上，常数为 λ_k。

这一结果完整描述了 TMS 中学习到的几何（Elhage 等人，2022年），包括所有单形、多边形和正方反棱柱。需要注意的是，对于单纯形以外的任何形状，紧框架刻画都不是唯一的。尽管如此，我们可以利用谱对应（引理 B.3），根据关联方案的谱特征对精确几何进行分类（证明见C）：

定理 4（几何的谱识别）。 假设谱局域化成立，则根据定理 *3*，特征簇 C_k构成一个紧框架。设 $M_k = W_{C_k}^\top W_{C_k}$ 为局部格拉姆矩阵。那么，簇 C_k的几何是一个关联方案，当且仅当格拉姆矩阵 $\mathcal{A}$ 的特征子空间 M_k 与代数 $\{W_0, W_1, \cdots, W_s\}$ 的典范分层 $\mathcal{A}$重合。

此定理直接推论出单纯形代数分类 A.3.1：也就是说，如果紧框架由 $d+1$ 个特征组成，位于 d 维空间 $\binom{d+1}{d} I_d$)中，那么它必然是单纯形几何（证明见C）：

推论 5 （单纯形识别）。 特征簇 C_k 的几何是单纯形，当且仅当对应的格拉姆矩阵 M_k 恰有两个特征子空间 $W_0 = \ker(M_k) = \text{span}(\mathbf{1})$， $W_1 = \text{Im}(M_k) = \mathbf{1}^\perp$ 其中$\Lambda(W_0) = 0$且 $\Lambda(W_1) = \lambda_k = \frac{|C_k|}{\dim(V_k)}$

我们分析了这些特征为何能在训练过程中持续存在，详见 F.1；我们的假设是，它们在梯度下降下保持稳定。

5. 局限性与结论

可解释性旨在从神经网络的内部结构入手理解它们。当前方法聚焦于识别特征——即激活空间中稀疏且可解释的方向。本文认为，特征之间的几何关系同样重要，而且可以借助经典数学工具恢复这种几何结构。通过研究权重诱导算子的特征子空间，而非单个特征向量，我们能够保留特征之间的几何关系，包括共享子空间、对称性和结构化干扰——而这些都会被稀疏或线性分析所舍弃。该框架将关于关系（成对）的问题转化为关于结构（整体）的问题。与其问“特征 i 和特征 j 有何关系 ”，不如问 “ 特征空间的几何结构是什么，i 和 j 在其中处于什么位置？”这就好比，知道城市之间的成对距离，与拥有一张地图之间的区别。

本文提出的图景仍有局限。要使谱测度能够完全刻画干扰，模型需要表现出谱局域化（HT-SR 现象学对此有所提示，但尚未直接测量）。若要完整恢复几何结构，我们需要容量饱和，而我们目前仅在玩具模型中通过实证验证了这一点。此外，我们针对玩具模型得到的结果格外简洁，主要是因为该玩具模型是自编码器，梯度仅通过格拉姆矩阵传递。

尽管存在这些局限，与单个特征向量相比，权重诱导算子似乎更适合作为分析叠加的基本单位。我们认为，这也说明可以运用算子理论来研究模型为何会呈现出特定行为。

算子理论在研究复杂物理系统方面已取得广泛成功。谱定理在量子力学中的作用——将状态分解为可观测量的特征基——与我们的谱局域化结果类似，后者将特征分解到框架算子的特征子空间中。表示论对对称作用的分类，与我们运用关联方案对特征几何进行离散分类相呼应。这些并非仅仅是类比：其数学结构完全相同，这表明，可解释性研究或许也能受益于那些曾使复杂物理系统变得可解析的工具。

然而，更根本的是，通过考察不同空间如何相互交织，算子理论旨在提供映射，而非仅仅进行两两比较。我们希望通过用算子理论的术语分析模型的组成部分，能够从单纯研究特征转向研究模型行为。

6. 影响声明

本文提出了一个新颖的理论框架，利用谱工具分析神经网络中的叠加现象。其主要贡献在于方法论：引入基不变诊断，用于分类学习到的表征中的特征几何和干扰。本研究具有基础性，不会带来新的模型能力、数据集或应用。研究结果旨在支持可解释性和表征分析领域的研究。尽管对内部表征的深入理解可能间接为下游方法提供参考，但本研究本身风险极低，也不会产生直接的社会影响。

参考文献

Micah Adler 和 Nir Shavit。《叠加中的神经计算复杂性》。arXiv:2409.15318，2024年。

Michael Aizenman 和 Stanislav Molchanov。大无序和极端能量下的局域化：一种初等推导。《数学物理通讯》*157:245-278*，1993。

Anthropic。当模型操纵流形：计数任务的几何。内部演示/研究札记，2025年。

Michael Atiyah 和 Paul Sutcliffe。物理学、化学与几何中的多面体。https://arxiv.org/pdf/math-ph/0303071，2003。

R. A. Bailey。关联方案：试验设计、代数与组合数学。剑桥大学出版社，2004。

Anirudh R. Bhaskara、Pradeep Kr. Yadama、Thomas Pell 和 Danica J. Sutherland。投射假设：稀疏自编码器与概念几何之间的对偶性。arXiv 预印本，2025年。

Sviatoslav Chalnev、Matthew Siu 和 Arthur Conmy。通过针对稀疏自编码器特征改进引导向量。*arXiv* 预印本 *arXiv:2411.02193*，2024年。

Lawrence Chan。叠加并不“只是”神经元的多义性。在 *AIAlignment Forum*，2024年。

Arthur Conmy、Thomas Pell、Pradeep Kr. Yadama 和 Danica J. Sutherland。从平坦到层次化：使用匹配追踪提取稀疏表示。arXiv 预印本，2025年。

Maximilian Dreyer、Erblina Purelku、Johanna Vielhaben、Wojciech Samek 和 Sebastian Lapuschkin。PURE：通过识别相关回路，将多义神经元转化为纯粹特征。CVPR 2024 Workshops；arXiv:2404.06453，2024年。

J T Edwards 和 D J Thouless。无序系统中的局域化数值研究。*Journal of Physics C:SolidStatePhysics*，1972年。

Nelson Elhage、Tristan Hume、Catherine Olsson、Nicholas Schiefer、Tom Henighan、Shauna Kravec、Zac Hatfield-Dodds、Robert Lasenby、Dawn Drain、Carol Chen、Roger Grosse、Sam McCandlish、Jared Kaplan、Dario Amodei、Martin Wattenberg 和 Christopher Olah。《叠加玩具模型》。arXiv 预印本 arXiv:2209.10652，2022年。

Joshua Engels、Isaac Liao、Eric J. Michaud 和 Max Tegmark。并非所有语言模型特征都是一维线性的。arXiv:2405.14860，2024年。

Chris Godsil 和 Gordon Royle。《代数图论》。Springer，2001年。Google。《深度序列模型往往会记忆几何结构》。Google Research Blog／内部报告，2025年。Wes Gurnee 和 Max Tegmark。《语言模型表征空间和时间》。arXiv:2310.02207，2024年。Robin Hesse、Jonas Fischer、Simone Schaub-Meyer 和 Stefan Roth。《分离卷积神经网络中的多语义通道》。CVPR 2025 Workshops；arXiv:2504.12939，2025年。

Lilian Hollard、Lucas Mohimont、Nathalie Gaveau 和 Luiz-Angelo Steffenel。《探索最先进的低参数视觉模型中的叠加与干涉》。arXiv:2507.15798，2025年。

Dirk Hundertmark。《增长过程与界面模型的分析与随机性理论》，第 9 章：安德森局域化简介。Oxford Academic，2008年。

TosioKat¯o. 线性算子摄动理论。Springer-Verlag，1980。

David Klindt、Charles O’Neill、Patrik Reizinger、Harald Maurer 和 Nina Miolane。从叠加到稀疏编码：神经网络中的可解释表示。arXiv:2503.01824，2025年。

Victor Lecomte、Kushal Thaman、Rylan Schaeffer、Naomi Bashkansky、Trevor Chow 和 Sanmi Koyejo。是什么导致了多语义性？混合选择性的另一种起源解释：源于偶然因素。arXiv:2312.03096，2023年。

Charles Martin 和 Michael Mahoney。重尾普适性能够预测超大型预训练深度神经网络的准确率趋势。https://arxiv.org/pdf/1901.08278，2020。

Charles Martin 和 Michael Mahoney。深度神经网络中的隐式自正则化：来自随机矩阵理论的证据及其对学习的启示。*Nature Communications 4122*，2021a年。

Charles Martin 和 Michael Mahoney。《在无法访问训练或测试数据的情况下预测最先进神经网络的质量趋势》。*Nature Communications4122*，2021b年。

Alexander D. Mirlin、Yan V. Fyodorov、Frank-Michael Dittes、Javier Quezada 和 Thomas H. Seligman。《幂律随机带状矩阵系综中从局域本征态到扩展本征态的转变》*PhysicalReviewE 54, 3221*，1996年。

Neel Nanda 和 Lawrence Chan。语言模型利用三角学进行加法。Mechanistic Interpretability Research，2025年。

DuyNguyen、Archiki Prasad、Elias Stengel-Eskin 和 Mohit Bansal。通过定向干预对语言模型进行多属性引导。*arXiv* 预印本 *arXiv:2502.12446*，2025a。

Eric Nguyen、Michael Poli、Armin Thomas、Brian Hie、Stephen Quake 和 Christopher R´e。在 Evo 2 中寻找生命之树。Arc Institute/斯坦福大学，2025b。

Tuan M. Nguyen 和 Ronald Mo。角度引导：通过激活空间中的旋转实现行为控制。新加坡国立大学，2025年。

Alexander Pan、Yi Chen、Qiyang Chen、Xingang Fang、Weidong Fu 和 Xiaozhe Jin。大语言模型对齐的隐藏维度：多维安全性分析。麻省理工学院/Safety Research，2025年。

Kiho Park、Jung-Hoon Kim、Alice Jiang 和 Victor Veitch。大语言模型中类别概念与层级概念的几何。芝加哥大学研究，2025年。

Lukas Pertl、Han Xuanyuan 和 Pietro Li‘o。《图神经网络中的叠加》。arXiv:2509.00928，2025年。

Nirmalendu Prakash、Yeo Wei Jie、Amir Abdullah、Ranjan Satapathy、Erik Cambria 和 Roy Ka Wei Lee。《不止于“对不起，我做不到”：剖析大语言模型拒绝》。*arXiv* 预印本 *arXiv:2509.09708*，2025年。

Jonas B Raedler、Weiyue Li、Alyssa Mia Taliotis、Manasvi Goyal、Siddharth Swaroop 和 Weiwei Pan。《干预保真度的必要性：引导大语言模型时产生的意外副作用》。发表于*ICML 2025* 可靠与负责任基础模型研讨会，2025年。

Nina Rimsky、Nick Gabrieli、Julian Schulz、Meg Tong、Evan Hubinger 和 Alexander T urner。《通过对比式激活添加引导 Llama 2》。发表于计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），第 15504–15522 页，2024年。

Vincent Siu、Nicholas Crispino、David Park、Nathan W Henry、Zhun Wang、Yang Liu、Dawn Song 和 Chen-guang Wang。Steeringsafety：一种针对大语言模型表征引导的系统性安全评估框架。*arXiv* 预印本 *arXiv:2509.13450*，2025年。

Steve Smale。下个世纪的数学问题。Mathematical Intelligencer，第20卷，1998。

Terence Tao。随机矩阵理论专题。美国数学学会，2012。

Alexander Matt Turner、Lisa Thiergart、Gavin Leech、David Udell、Ulisse Mini 和 Monte MacDiarmid。《激活添加：无需优化即可引导语言模型》。arXiv 预印本 arXiv:2308.10248，2024年。

附录

A 特征几何.12 A.1 轨道代数推导.12 A.2 关联方案.14 A.3 代数分类.15 A.3.1 单纯形代数.15 A.4 紧框架.17 B 谱理论.17 B.1 谱测度.22 B.2 希尔伯特空间.24 B.6 微扰理论.21 B.7 格拉姆特征投影的一阶漂移.21 B.8 谱测度矩.23 B.9 循环空间与伪逆柯西–施瓦茨不等式.24 B.10 $L^2(\mu)$（有限原子测度）.25 B.11 $L^2(\mu)$中的柯西–施瓦茨不等式.25 B.12 $L^2(\mu_x)$循环等距映射 $\simeq \mathcal{K}_x$ .26 B.13 定理证明与推论.27 B.14 扰动情形.29 C 局域化与离域化.33 C.1 HT-SR.33 C.2 安德森局域化.33 D 梯度流.34 D.1 显式形式.35 D.2 关联方案约化.36 E 非均匀稀疏性.41 F 相关工作.43

A. 特征几何

A.1. 轨道代数推导

设定一个大小为 $|C| = p$ 的特征簇 C。我们可以为顶点集合 $V = \{1, \cdots, p\}$ 编号。根据多胞体几何，我们可以假设存在一个抽象对称群 $\Gamma \subseteq S_p$ （凯莱定理），它在 V 上传递地作用，即$\forall i,\ j \in V,\ \exists \gamma \in \Gamma$，使得 $\gamma \cdot i = j$。简单来说，这意味着每个顶点都可以从任意其他顶点出发到达。借助给定顶点 i 的“轨道”这一概念，可以非常简洁地表述这一点。轨道定义为$\Gamma \cdot i = \{\gamma \cdot i \,|\, \gamma \in \Gamma\}$，也就是说，由元素 γ 从 i 出发可以到达的所有顶点所构成的集合。于是，传递性等价于该群只有一个轨道。回到线性代数领域，我们可以将 $P_\gamma \in R^{p\times p}$ 定义为 γ 置换矩阵表示:

$$(P_\gamma)_{ij} = \delta_{i,\gamma(j)} = \begin{cases} 1 & \text{if } i = \gamma(j) \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

关于置换矩阵，我们有以下一个简洁的结论:

引理 A.1 （置换矩阵的正交性）。 对于任意 $\gamma \in \Gamma$， $P_\gamma^{-1} = P_\gamma^\top$

证明。 根据定义（5），我们有 $(P_\gamma^\top)_{ij} = (P_\gamma)_{ji} = \delta_{j,\ \gamma(i)}$，因此:

$$(P_\gamma P_\gamma^\top)_{ij} = \sum_{k=1}^{p} (P_\gamma)_{ik}(P_\gamma^\top)_{kj} =$$

$$\sum_{k=1}^{p} \delta_{i,\gamma(k)}\delta_{j,\gamma(k)} = \delta_{ij} = (I_p)_{ij},$$

其中 $I_p \in \mathbb{R}^{p\times p}$ 是单位矩阵。不失一般性， $P_\gamma^\top P_\gamma$ 也会得到相同的结果。 □

事实上，这直接推得

$$P_\Gamma = \{P_\gamma \in \mathbb{R}^{p\times p} : \gamma \in \Gamma\} \subseteq O(p),$$

其中 $O(p)$ 是 p 维正交群，定义为由 $p\times p$ 正交矩阵构成的群。这就是包含所有多胞体几何对称性的总括性“母”对称性:

$$P_{D_n} \subset O(2) \subset O(3) \subset \cdots \subset O(n-1) \subset O(n)$$

这种矩阵形式化还使我们能够利用中心化子（交换子）严格定义簇构型的“稳定性”:

$$\mathcal{A}_\Gamma := \{A \in \mathbb{R}^{p\times p} : P_\gamma A P_\gamma^\top = A, \forall \gamma \in \Gamma\} \tag{A.2}$$

这是所有在任意置换 P_γ 的共轭作用下保持不变的 $\mathbb{R}^{p\times p}$ 矩阵所组成的集合。事实上，它不仅仅是一个集合，还是一个代数。要理解其中的原因，我们需要将注意力转向有序对 $V \times V$ 的集合。该集合具有一个由 Γ 作用于其上的典范群作用，其定义如下:

$$\pi : \Gamma \times (V \times V) \to V \times V : \quad \gamma \cdot (i,j) = (\gamma(i), \gamma(j)),$$

该作用自然地将所有 $|V \times V| = p^2$ 对划分为互不相交的轨道 $\mathcal{O}_1,\ \cdots,\ \mathcal{O}_R$，其中 $R + 1 = r(\Gamma)$ 是置换群的秩（拉格朗日定理）。我们可以为每个轨道 $\mathcal{O}_r$ 定义一个轨道矩阵 A_r，如下所示:

$$(A_r)_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{O}_r \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

由于我们的群 Γ 在 V上具有传递性，对角线 $\mathcal{O}_0 = \{(i,\ i)\}$ 自然构成一个轨道，其表示为单位矩阵 I_p。

引理 A.2 (中心化子成员关系)。 矩阵 X 属于 $\mathcal{A}_\Gamma$，当且仅当其元素在 Γ于 $V \times V$上的各个轨道上为常数。

证明。 ($\Rightarrow$)令 $\gamma \in \Gamma$和 $X \in \mathcal{A}_\Gamma$ 为任意元素。根据定义 (6):

$$P_\gamma X P_\gamma^\top = X \quad \Leftrightarrow \quad P_\gamma X = X P_\gamma \quad \text{(Lemma 4.3)}$$

根据定义 (5)，对于所有 i， j，左侧为:

$$\sum_{k=1}^{p} (P_\gamma)_{ik} X_{kj} = \sum_{k=1}^{p} \delta_{i,\gamma(k)} X_{kj} = X_{\gamma^{-1}(i),j}$$

右侧同理:

$$\sum_{k=1}^{p} X_{ik} (P_\gamma)_{kj} = \sum_{k=1}^{p} X_{ik} \delta_{k,\gamma(j)} = X_{i,\gamma(j)}$$

因此，$X_{i,\ \gamma(j)} = X_{\gamma^{-1}(i)\ j}$。由于这对所有 i， j均成立，我们可以令 $i = \gamma^{-1}(k)$，即

$$X_{\gamma^{-1}(\gamma(k)),j} = X_{k,j} = X_{\gamma(k),\gamma(j)}$$

由于 γ 是任意的，正向证明至此完成。反向证明（$\Leftarrow$）由 WLOG 得出：取上文左右两侧中任意的 i， j 项即可。 □

引理 A.3 （中心化子的基）。 $\{A_0, \cdots, A_R\}$ 是 $\mathcal{A}_\Gamma$ 的基

证明。 前述 引理直接意味着 $\forall_{r\in\{0,\cdots,R\}} A_r \subseteq \mathcal{A}_\Gamma$，因为如果 $(i,\ j) \in \mathcal{O}_r$，根据 t
轨道的定义:

$$\forall \gamma \in \Gamma, \quad \gamma \cdot (i,j) = (\gamma(i), \gamma(j)) \in \mathcal{O}_r,$$

这意味着 $(A_r)_{i,\ j} = 1 = (A_r)_{\gamma(i),\ \gamma(j)}$，类似地 $\forall (i,\ j) \notin \mathcal{O}_r$，$(A_r)_{i,\ j} = 0 = (A_r)_{\gamma(i),\ \gamma(j)}$。事实上，这可以扩展到 $\text{span}\{A_0, \cdots, A_R\} \subseteq \mathcal{A}_\Gamma$，因为这些轨道互不相交:

$$X = c_0 A_0 + \cdots c A_R \Leftrightarrow X_{ij} = c_r = X_{\gamma(i),\gamma(j)}, \tag{A.4}$$

其中 $(i,\ j) \in \mathcal{O}_r$。在（8）中用任意 A_r 替换 X，意味着 $\{A_0, \cdots, A_R\}$ 线性无关。此外，由于任意不变矩阵 $X \in \mathcal{A}_\Gamma$ 都必须在这些轨道上保持常数，根据（8）和前述引理， X 可以唯一写成:

$$X = \sum_{r=0}^{R} c_r A_r,$$

借此我们完成了等式另一方向的证明，得出 $\mathcal{A}_\Gamma \subseteq \text{span}(A_0,\ \cdots,\ A_R\}$ 实际上是一个基。

□

我们还需要的最后一个性质是中心化子在矩阵乘法下的代数闭性:

引理 A.4 （代数闭性）。若 X， $Y \in \mathcal{A}_\Gamma$，则 $XY \in \mathcal{A}_\Gamma$

证明。 令 $X, Y \in \mathcal{A}_\Gamma$。根据成员资格的定义，我们有 $P_\Gamma X = X P_\Gamma$ 和 $P_\gamma Y = Y P_\gamma$。将XY乘以 P_γ，得到:

$$P_\gamma (XY) = (P_\gamma X) Y = (X P_\gamma) Y$$
$$= X (P_\gamma Y) = X (Y P_\gamma) = (XY) P_\gamma$$

因此 $XY \in \mathcal{A}_\Gamma$。 □

我们可以将后两个引理应用于任意两个轨道矩阵的乘积

$$(A_r A_s)_{ij} = \sum_k (A_r)_{ik}(A_s)_{kj},$$

其中各个元素可解释为从 i 到 j 的长度为 2 的“路径”数量，第一步的轨道类型为 r，第二步的轨道类型为 s:

$$A_r A_s \in \text{span}\{A_0, \cdots A_R\} \Rightarrow A_r A_s = \sum_{u=0}^{R} c_{rs}^u A_u \tag{A.5}$$

$$c_{rs}^u = |\{k \in V | (i,k) \in \mathcal{O}_r \text{ and } (k,j) \in \mathcal{O}_s\}| \tag{A.6}$$

对于任意 $(i,\ j) \in \mathcal{O}_u$。事实上，请注意，若我们选择 $(i,\ j) \in \mathcal{O}_u$，根据引理 A.4 的不变性，这一计数必定与 $(i,\ j)$ 在该轨道中的具体选择无关。这些结构常数实际上称为关联方案 $\{A_0, \cdots, A_R\}$ (Bailey，2004) 的交数；在代数图论中（Godsil 和 Royle，2001年）也以此名称为人所知。为了说明叠加玩具模型（Elhage 等人，2022年）所描述的散点图中的“粘滞点”，我们需要证明，一旦矩阵进入这种状态，它就会相对于给定多边形的群对称性变得“稳定”。

A.2. 关联方案

每个关联 方案 $\{A_0, \cdots A_s\}$ 在具有 s 个关联类的集合 Ω 上构成一个 Bose-Mesner 代数 :

$$\mathcal{A} = \left\{ \sum_{i=0}^{s} \mu_i A_i : \mu_0, \cdots, \mu_s \in \mathbb{R} \right\}$$

我们有如下简洁的谱理论，陈述为定理 2.6（Bailey，2004）：

引理 A.5. 设 $\{A_0, A_1, \cdots A_s\}$ 为Ω 上一个关联方案的邻接矩阵，并设 $\mathcal{A}$ 为其 *Bose-Mesner* 代数。则 $\mathbb{R}^\Omega$ 具有 $s+1$ 个带正交投影算子的正交子空间 W_0、W_1、$\ldots W_s$（层），使得:

(i) $\mathbb{R}^\Omega = W_0 \oplus W_1 \oplus \cdots \oplus W_s$

(ii) W_0、$W_1 \ldots$、W_s 中的每一个都是 $\mathcal{A}$中每个矩阵的子特征空间;

(iii) 对于 $i = 0,\ 1, \ldots\ s$，邻接矩阵 A_i 是 S_0、S_1、$\cdots$、S_s 的线性组合（注意$S_0 = {}_{|}{}^{1}{}_{\Omega|} J$） —

(iv) 对于 $e = 0,\ 1, \ldots s$，层投影算子 S_e 是 A_0； A_1、$\cdots$、A_s的线性组合

对于 $i \in \mathcal{K}$（关联方案的指标集）和 e（层投影算子的指标集），我们定义 $C_{(i,e)}$ 为 W_e 上 A_i 的特征值，并定义 $D_{(i,e)}$ 为以下展开中的系数:

$$A_i = \sum_{e \in \mathcal{E}} C_{(i,e)} S_e, \quad S_e = \sum_{i \in \mathcal{K}} D_{(e,i)} A_i$$

以下 事实成立（引理 2.9、定理 2.12、推论 2.14 和推论 2.15，出自（Bailey，2004）):

引理 A.6。 矩阵 $C \in \mathbb{R}^{\mathcal{K} \times \mathcal{E}}$ 和:

(i) $D \in \mathbb{R}^{\mathcal{E}\times\mathcal{K}}$ 互为逆矩阵，并且：

$$C_{(0,e)} = 1, \quad C_{(i,0)} = a_i, \quad D_{(0,i)} = a_i, \quad D_{(e,0)} = \frac{d_e}{|\Omega|},$$

其中 $a_i = p_{ii}^0$ 是第 i个关联类的价，且 $d_e = \mathit{tr}(S_e)$；

(ii)

$$\sum_e C_{(i,e)} C_{(j,e)} d_e = \delta_{ij} a_i |\Omega|,$$

其中 δ_{ij} 是狄拉克函数，即当 $i = j$ 时取 $\delta_{ij} = 1$，否则取0 。

(iii)

$$\sum_{i\in\mathcal{K}} \frac{1}{a_i} C_{(i,e)} C_{(i,f)} = \delta_{ij} \frac{|\Omega|}{d_e}$$

(iv)

$$\sum_{i\in\mathcal{K}} D_{(e,i)} D_{(f,i)} a_i = \delta_{ef} \frac{d_e}{|\Omega|}$$

(v)

$$\sum_e \frac{1}{d_e} D_{(e,i)} D_{(e,j)} = \frac{\delta_{ij}}{|\Omega| a_i}$$

A.3. 代数分类

A.3.1. 单纯形代数

任意单纯形的一个性质是其双内积，即对于单位向量 $\{x_i\}_{i\in V} \subset \mathbb{R}^{p-1}$，我们有 $X_{ii} = \langle x_i, x_i\rangle = 1$ 对所有 i 和 $X_{ij} = \langle x_i, x_j\rangle = \frac{1}{1-p}$均成立。 我们选择 $\Gamma = S_p$ 作为群。根据 X的定义，其元素值仅取决于两个指标是否相等。又由于 γ 是双射，我们有 $\gamma^{-1}(i) = \gamma^{-1}(j) \Leftrightarrow i = j$，因此根据中心化子的定义可得：

$$(P_\gamma M P_\gamma^\top)_{ij} = M_{\gamma^{-1}(i),\gamma^{-1}(j)} = \begin{cases} 1 & i = j \\ \frac{1}{1-p} & i \neq j \end{cases} = M_{ij}$$

这将 $V \times V$ 上的群作用分成两个轨道，因为在典范对角线 $\mathcal{O}_0$ 之外。将另一个轨道定义为：

$$\mathcal{O}_{off} = \{(i,j) \in V \times V : i \neq j\}$$

根据 γ 的单射性， $i \neq j$ 推出 $\gamma(i) \neq \gamma(j)$，因此 $\gamma(i,j) \in \mathcal{O}_{off}$。现在，任取 $\mathcal{O}_{\text{off}}$ 中的 (i,j) 和 (k,ℓ)；于是 $i \neq j$ 和 $k \neq \ell$。我们可以显式构造 $\gamma \in S_p$，使得 $\gamma(i) = k$ 且 $\gamma(j) = \ell$。在 $\{i,j\}$ 上定义 γ 为 $\gamma(i) = k$， $\gamma(j) = \ell$。由于 $k \neq \ell$，它在 $\{i,j\}$ 上是单射。将 γ 扩展为整个 V 上的一个双射，方法是将 $V \setminus \{i,j\}$ 中剩余的 $p-2$ 个元素双射地映射到 $V \setminus \{k,\ell\}$（这是可行的，因为这两个集合的有限基数相同）。这个扩展是一个置换 $\gamma \in S_p$。于是 $\gamma \cdot (i,j) = (k,\ell)$。因此，恰有两个轨道，也就是说 $r(\Gamma) = 2$ 和 $R = 1$。根据 A.3，Γ- 不变矩阵具有如下形式

$$A = aI + c(J - I),$$

其中 J 是 m 全 1 矩阵。利用 $A_0 = I, A_1 = J - I$，我们还可以计算交数：

$$A_0 A_0 = A_0, \quad A_0 A_1 = A_1, \quad A_1 A_0 = A_1$$

$$A_1^2 = (J-I)^2 = J^2 - 2J + I = pJ - 2J + I$$

$$= (p-2)J + I = (p-2)A_1 + (p-1)A_0$$

因此，唯一非零的交数为：

$$c_{00}^0 = 1, \quad c_{01}^1 = c_{10}^1 = 1, \quad c_{11}^0 = p-1, \quad c_{11}^1 = p-2$$

利用 A.6(i) 来表示

$$C_{(0,0)} = 1, \quad C_{(0,1)} = 1, \quad C_{(1,0)} = a_1 = c_{11}^0 = p-1$$

我们可以显式地计算投影算子：

$$A_0 = S_0 + S_1, \quad A_1 = (p-1)S_0 + C_{(1,1)}S_1,$$

其中，由于 $S_0 = \frac{1}{p}J$ [A.5(iii)],我们有 $S_1 = I - \frac{1}{p}J$。因此：

$$J - I = \frac{p-1}{p}J + C_{(1,1)}\left[I - \frac{1}{p}J\right]$$

$$\frac{C_{(1,1)}+1}{p}J = (1 + C_{(1,1)})I,$$

这意味着 $C_{(1,1)} = -1$。 利用 B.3 我们可以将找到的本原幂等元转换到特征潜在空间 $\mathbb{R}^m$

$$WS_0W^\top = \frac{1}{p}WJW^\top = \frac{1}{p}W\mathbf{1}\mathbf{1}^\top W^\top = \frac{1}{p}(W\mathbf{1})(W\mathbf{1})^\top = \frac{1}{p}\left(\sum_{i=1}^p W_i\right)\left(\sum_{i=1}^p W_i\right)^\top$$

根据单纯形格拉姆矩阵的定义，我们有

$$\langle W_i, W_i\rangle = 1, \quad \langle W_i, W_j\rangle = \frac{1}{1-p} = -\frac{1}{p-1} = \frac{1}{p-1}$$

那么，对于任意 k：

$$\langle W_k, \sum_{i=1}^p W_i\rangle = \sum_{i=1}^p \langle W_k, W_i\rangle = \langle W_k, W_k\rangle + \sum_{i\neq k}\langle W_k, W_i\rangle = 1 + (p-1)\frac{-1}{p-1} = 0$$

这实际上意味着

$$\langle \sum_{k=1}^p W_k, \sum_{i=1}^p W_i\rangle = \sum_{k=1}^p \langle W_k, \sum_{i=1}^p W_i\rangle = 0 \Leftrightarrow WS_0W^\top = 0$$

根据同一定义

$$M = I - \frac{1}{p-1}(J-I) = \frac{p}{p-1}I - \frac{1}{p-1}J = \frac{p}{p-1}\left(I - \frac{1}{p}J\right) = \frac{p}{p-1}S_1$$

引理 A.7 (二边形)。 考虑 *2* 中的设定。我们可以用与推导三角形（实际上也包括上文的单纯形）相同的方法，推导出二边形的精确关联方案：

证明。这里， $\Gamma \cong S_3 \cong C_2$ 是由两个元素组成的循环群 $\{e, s\}$，即单位元 e，而 s 是一个二阶元，也就是说 $s^2 = 1$。这在视觉上表示二边形的两个节点互换（即从正面-反面变为反面-正面）。根据这是一个单纯形这一事实（即从一个节点可以到达其他每个节点），我们有相同数量的轨道/关联类 $\mathcal{C}_0 = \mathsf{diag}(\Omega_D)$ 和 $\mathcal{C}_1 = \{(4,5);\ (4,5)\}$，其邻接矩阵为$A_0' = I_2$ 和 $A_1' = J_2 - I_2$。类似地，对 $\mathcal{A}_D = \mathsf{span}(A_0', A_1')$ 应用引理 A.5，得到 $\mathbb{R}^{\Omega_D} \cong \mathbb{R}^2 = U_0 \oplus U_1$，并由引理 A.6 $S_0' = \frac{1}{2}I$, $S_1' = I - \frac{1}{2}J$ 以及：

$$C' = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad D' = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$
$$M_D = I - (J - I) = 2I - J = 2S_1'$$

□

A.4. 紧框架

引理 A.8。 设 w_1、 $\cdots$、 $w_p \in \mathbb{R}^d$ 是构成一个紧框架的单位向量：

$$\sum_{j=1}^{p} w_j w_j^\top = \frac{p}{d} I_d$$

设 $P_{ij} = \langle w_i, w_j \rangle$ 为它们的 $p \times p$ 格拉姆矩阵。那么，对每个 i：

$$(P^2)_{ii} = \frac{p}{d} \Rightarrow D_i = \frac{1}{(P^2)_{ii}} = \frac{d}{p}$$

Proof. $(P^2)_{ii} = \sum_j (w_i^\top w_j)^2 = w_i^\top \left(\sum_j w_j w_j^\top\right) w_i = w_i^\top \frac{p}{d} I w_i = \frac{p}{d}$. By (L4.1) $P_{ii} = 1$. □

后一项结果使我们能够用分数维数的倒数来量化观测到的簇值。具体来说，请注意，根据引理 4.2，我们可以对叠加玩具模型中观测到的几何形状进行分类（Elhage 等人，2022年），并使用紧框架参数$(p,\ d)$表示其对应的分数维数：

- (2，1) - 含 $D_i = 1/2$ 的二边形
- (3，2) - 含 $D_i = 2/3$ 的三角形
- (4，3) - 含 $D_i = 3/4$ 的四面体
- (5，2) - 带有 $D_i = 2/5$ 的五边形
- (8，3) - 带有 $D_i = 3/8$ 的正方反棱柱

然而，需要注意的是，仅凭这一点并不足以区分不同类型的对称性。例如，正方反棱柱和立方体都具有 $(p,\ d) = (8,\ 3)$，但前者是保持稳定的那个，这一点在原始论文以及更一般的汤姆孙问题中均有观察到（Atiyah 和 Sutcliffe，2003）。

B. 谱理论

引理 B.1 (格拉姆算子的谱分解（Kat¯o，1980），定理 2.10）。 设 $W \in \mathbb{R}^{m \times n}$ 并设M：$= W^\top W \in \mathbb{R}^{n \times n}$ 为格拉姆矩阵。那么 M 是对称半正定的，并具有正交谱分解

$$M = \sum_{e \in \mathcal{E}} \lambda_e E_e,$$

其中， $\lambda_e \in \mathbb{R}_{\geq 0}$ 是互不相同的特征值， $E_e \in \mathbb{R}^{n \times n}$ 是对应的正交谱投影算子。具体而言，

$$E_e^\top = E_e, \qquad E_e^2 = E_e, \qquad E_e E_f = 0 \ (e \neq f), \qquad \sum_{e \in \mathcal{E}} E_e = I_n,$$

and

$$ME_e = E_eM = \lambda_e E_e \qquad (\forall e \in \mathcal{E}).$$

如果0 是一个特征值，我们将其记为 $\lambda_0 = 0$，而 E_0 是投影到$\ker(M)$上的正交投影算子。

引理 B.2 (框架与格拉姆核)。 以下结论可由像空间与核的正交分解直接推出：

(i) $\ker(WW^\top) = \ker(W^\top)$ *(ii)* $\ker(W^\top W) = \ker(W)$

证明。 (⇒) $\ker(W^\top) \subseteq \ker(WW^\top)$。任取 $y \in \ker(W^\top)$。则，

$$W^\top y = 0 \Rightarrow WW^\top y = W0 = 0 \Rightarrow y \in \ker(WW^\top)$$

(⇐) 任取 $y \in \ker(WW^\top)$。则 $WW^\top y = 0$。由于右侧为 0，我们可以在左侧乘以 $y^\top$，从而得到：

$$y^\top WW^\top y = y^\top 0 = 0 \Rightarrow (W^\top y)^\top W^\top x = ||W^\top y|| = 0,$$

根据范数的定义，这意味着 $W^\top y = 0$，即 $y \in \ker(W^\top)$。由此，我们从两个方向证明了等式 $\ker(W^\top) = \ker(WW^\top)$。交换 $W^\top$ 和 W 后，(*ii*) 的证明完全相同。 □

引理 B.3 （谱对应）。 设 $M = W^TW = \sum_e \lambda_e E_e$ 为格拉姆矩阵的谱分解。定义 $P_e := \lambda_e{}^+ WE_eW^\top \in \mathbb{R}^{m\times m}$，对每个 $e = \{e : \ \lambda_e > 0\}$。则：

(i) P_e 对称 $P_e^\top = P_e$ 且幂等 $P_e^2 = P_e$ （即正交投影算子）

(ii) $P_eP_f = 0$ *for* $e \neq f$

(iii) 框架算子分解为 $F = WW^\top = \sum_e \lambda_e P_e$。

(iv) $Im(P_e) = W(Im(E_e))$

证明。 (i) 对称性可由 E_e 的对称性直接得出：

$$P_e^\top = \lambda_e{}^+(WE_eW^\top)^\top = \lambda_e{}^+WE_e^\top W^\top = \lambda_e{}^+WE_eW^\top = P_e$$

利用 E_e、$W^\top W = M$ 与 $ME_e = \lambda_e E_e$ 的幂等性，可得：

$$P_e^2 = \left(\lambda_e{}^+WE_eW^\top\right)\left(\lambda_e{}^+WE_eW^\top\right) =$$

$$= \left(\lambda_e{}^+\right)^2 WE_eME_eW^\top = \lambda_e{}^+WE_eW^\top = P_e$$

因此，由于 P_e 是对称且幂等的，它是到 $\mathrm{Im}(P_e)$ 上的正交投影算子（Kat¯o, 1980）。

（ii）计算：

$$P_eP_f = \left(\lambda_e{}^+WE_eW^\top\right)\left(\lambda_f^+WE_fW\top\right) =$$

$$\lambda_e^+\lambda_f^+WE_eME_fW^\top = \lambda_e^+\lambda_f^+WE_e\left(\sum_g \lambda_g E_g\right)E_f = \lambda_e{}^+WE_eE_f = 0$$

由于 $E_eE_f =$ 除非 $g = f$，否则为 0。由于每个 P_e 也是对称的，因此这意味着对于 $e \neq f$，$\mathrm{Im}(P_e) \perp \mathrm{Im}(P_f)$

(iii) 对 $\lambda_e > 0$， $\lambda_e P_e = W E_e W^\top$ 进行改写，并注意到根据 B.2(ii) $WE_0 = 0$:

$$\sum_e \lambda_e P_e = \sum_e W E_e W^\top = W \left(\sum_e E_e \right) W^\top = W \left(I_n - E_0 \right) W^\top = W W^\top - W E_0 W^\top = W W^\top$$

(iv) (⇒) 令 $y \in \mathbb{R}^m$ 为任意元素。则

$$P_e y = {\lambda_e}^+ W E_e W^\top y$$

令 $u := E_e W^\top y$。则 $u \in \text{Im}(E_e)$，并且

$$P_e y = {\lambda_e}^+ W u \in W(\text{Im}(E_e)),$$

因此 $\text{Im}(P_e) \subseteq W(\text{Im}(E_e))$，因为 y 是任意的
(⇐) 令 $u \in \text{Im}(E_e)$ 为任意元素。则 $E_e u = u$ 且 $Mu = \sum_e E_e u = \lambda_e u$。在 Wu上计算 P_e，可得:

$$P_e(Wu) = {\lambda_e}^+ W E_e W^\top W u = {\lambda_e}^+ W E_e M u =$$

$$= {\lambda_e}^+ W E_e (\lambda_e u) = W(E_u u) = W u$$

由此我们完成了 $W(\text{Im}(E_e)) \subseteq \text{Im}(P_e)$，因此等式 $\text{im}(P_e) = W(\text{Im}(E_e)$)

□

由 (*A*.3)(*iv*) 立即得到一个推论，即 dim $\text{Im}(P_e)$ = dim $\text{Im}(E_e)$。证明很简单，只需注意到 W是单射，因为如果 $u \in \text{Im}(E_e)$ 且 $Wu = 0$，那么 $Mu = W^\top W u = 0 = Mu = \lambda_e u$且 $\lambda_e > 0$。

引理 B.4 (谱交织)。 定义 $P_e := {\lambda_e}^+ W E_e W^\top \in \mathbb{R}^{m \times m}$ 对每个 $e = \{e \colon \lambda_e > 0\}$，其中 $M = W^T W = \sum_e \lambda_e E_e$ 是格拉姆算子的谱分解。则:

(i) $P_e W = W E_e$和 $W^\top P_e = E_e W^\top$。

(ii) $F P_e = P_e F = \lambda_e P_e$。

(iii) $\mathbb{R}^m = (\bigoplus_e \text{像}(P_e)) \oplus \text{核}(W^\top)$

证明。 (i)

$$P_e W = {\lambda_e}^+ W E_e W^\top W = {\lambda_e}^+ W E_e M = {\lambda_e}^+ W (\lambda_e E_e) = W E_e$$

$$W^\top P_e = {\lambda_e}^+ W^\top W E_e W^\top = {\lambda_e}^+ M E_e W^\top = E_e W^\top$$

(ii)

$$F P_e = W W^\top {\lambda_e}^+ W E_e W^\top = {\lambda_e}^+ W (W^\top W) E_e W^\top =$$

$$= {\lambda_e}^+ W M E_e W^\top = {\lambda_e}^+ W (\lambda_e E_e) W^\top = \lambda_e P_e$$

$$P_e F = {\lambda_e}^+ W E_e W^\top W W^\top = {\lambda_e}^+ W E_e M W^\top = \lambda_e P_e$$

(iii) 令 $M^+ = \sum_e {\lambda_e}^+ E_e$为 M的 Moore-Penrose 伪逆。则:

$$\sum_e P_e = \sum_e {\lambda_e}^+ W E_e W^\top = W M^+ W^\top$$

且 $W M^+ W^\top$ 是到 $\text{Im}(W) = \ker(W^\top)^\perp = \ker(W W^\top)^\perp$ 由 B。 2.

□

推论 6 (三角形-二边形情形)。 设定与第 2 节相同。于是， $F = 2P_D + \frac{3}{2}P_\triangle$，其中 P_D 和 $P_\triangle$分别是投影到二边形和三角形所在子空间上的投影算子。

证明。 现在，让我们定义 $U_\triangle$: 张成$\{W_i : \Omega_\triangle\} \subset \mathbb{R}^3$ 以及 U_D : 张成$\{W_i : \ i \in \Omega_D\} \subset \mathbb{R}^3$ 为了将其与谱描述联系起来 $M_\triangle$ 以及 M_D 我们发现的结果（它们彼此等价，因此可以推广到任意 C），让我们定义从一个簇 $C \in \{\triangle, D\}$ 到整个概念空间 E_C 的注入算子： $\mathbb{R}^{\Omega_C} \to \mathbb{R}^\Omega$，其中$(E_C x)^i = \sum_{j\in\Omega_C} x_i \delta_{ij}$ （图7）。

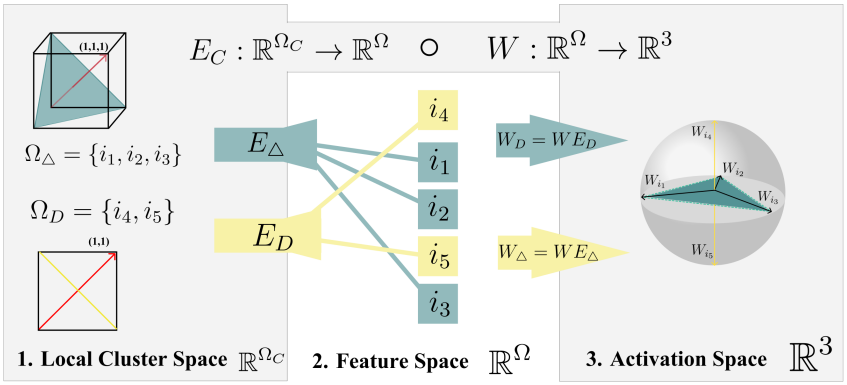


图 7. W 是谱模态的交织算子

那么，若 $W_C := WE_C \in \mathbb{R}^{3\times|\Omega_C|}$，则有 $M_C = W_C^\top W_C$和 $F_C = W_C W_C^\top = \sum_{i\in\Omega_C} W_i W_i^\top$，因为对集合求和具有置换不变性。此外，我们的谱结果可直接转移： $M_C = \lambda_C S_1^{(C)}$,$S_0^{(C)} = \frac{1}{|\Omega_C|}\mathbf{1}\mathbf{1}^\top$, $S_1^{(C)} = I - S_0^{(C)}$。由于质心被映射到核，这意味着 $S_0^{(C)}\, M_C\mathbf{1} = \lambda_C S_1^{(C)}\mathbf{1} = 0$。对于注入后的矩阵还有一个直接推论：左乘 $\mathbf{1}^\top$ 得到:

$$0 = \mathbf{1}^\top M_C \mathbf{1} = \mathbf{1}^\top W_C^\top W_C \mathbf{1} = \|W_c\mathbf{1}\|^2 = \left\| \sum_{i\in\Omega_C} W_C \right\|^2$$

若 $M_C x = \lambda_C x$ 且 $\lambda_C > 0$，则 $F_C W_C x = W_C M_C x = \lambda_C (W_C x)$. 因此 W_C 将 M_C 的 λ_C-特征子空间映射到 F_C 的 λ_C-特征子空间。由于 $M_C = \lambda_C S_1^{(C)}$，概念空间中的 λ_C-特征子空间为 $\mathrm{Im}(S_1^{(C)})$（“差分”子空间），而 (0)-特征子空间为 $\mathrm{Im}(S_0^{(C)})$ = span(1)，我们刚刚证明它被 W_C 映为零。因此，该簇产生的唯一非平凡激活几何为 $U_C := W_C\big(\mathrm{Im}(S_1^{(C)})\big) = \mathrm{Im}(W_C) \subseteq \mathbb{R}^3$, 并且 F_C 在 U_C 上的作用为 λ_C，在 $U_C^\perp$ 上的作用为 (0)。等价地，$F_C = \lambda_C P_C$, 其中 P_C 是到 U_C 上的正交投影算子。最后，定义 $P_C := \lambda_C^{-1} W_C S_1^{(C)} W_C^\top$. 则利用 $M_C S_1^{(C)} = \lambda_C S_1^{(C)}$ 直接计算可知， P_C 是一个投影算子（幂等且对称）：

$$\begin{aligned} P_C^2 &= \lambda_C^{-2} W_C S_1 W_C^\top W_C S_1 W_C^\top = \lambda_C^{-2} W_C S_1 M_C S_1 W_C^\top \\ &= \lambda_C^{-1} W_C S_1 W_C^\top = P_C. \end{aligned}$$

那么显然， $F_C = W_C W_C^\top = W_C(S_0 + S_1)W_C^\top = W_C S_1 W_C^\top = \lambda_C P_C$，因为 $W_C S_0 = 0$，这源于质心抵消。由此，我们恢复了式（2.2）。现在，我们可以利用全部谱信息，给出全局和局部几何分类；若格拉姆矩阵被任意打乱，这些分类原本是无法获得的。由于 $F_C = \lambda_C = P_C$，我们有:

$$\mathrm{tr}(F_C) = \lambda_C \,\mathrm{tr}(P_C) = \lambda_C \,\mathrm{rank}(P_C) = \lambda_C \,\dim(U_C)$$

但同时，$\mathrm{tr}(F_C) = \sum_{i\in\Omega_C} \|W_i\|^2$，因此对于我们的单位范数特征，有 $|\Omega_C|$，即 $\frac{1}{\lambda_C} = \dim(U_C)$ $\frac{\sum_{I\in\Omega_C}\|W_i\|^2}{\mathrm{rank}(P_C)\,|\Omega_C|}$，由此我们得到主示意图2。

□

B.1. 微扰理论

引理 B.5 (格拉姆算子的解析展开)。 设 $x \mapsto M(x) \in \mathbb{R}^{n\times n}$ 在$x=0$的某个邻域内解析，且 $M(0)=M$。那么存在矩阵 $M_{(k)} \in \mathbb{R}^{n\times n}$，使得

$$M(x) \;=\; M + xM^{(1)} + x^2 M^{(2)} + \cdots, \qquad M^{(k)} \;=\; \frac{1}{k!}\,\frac{d^k}{dx^k} M(x)\Big|_{x=0}.$$

在我们的设定中，可以取 $M(x) = W(x)^\top W(x)$ 以及 $F(x) = W(x)W(x)^\top$。当 $M(\cdot)$ 沿 *Gram* 流 $\dot M = -\{M, \Phi\}$ 演化，且在快照处导数存在时，可以将 $M^{(1)}$ 认定为 $\dot M(0) = -\{M, \Phi\}$。

参考文献：(Kat¯o，1980)，第 II 章，第 1 节，式 (1.2)。

引理 B.6 (的特征值的约化预解式 M)。 固定 M 的一个特征值 λ，并令 E 为相应的谱投影算子。 M 在 λ 处的约化预解式是算子 S_λ，它在不变补空间 $(I_n - E)\mathbb{R}^n$上对 $M - \lambda I_n$ 求逆。等价地， S_λ 由以下条件刻画：

$$(M-\lambda I_n)S_\lambda = S_\lambda(M-\lambda I_n) = I_n - E, \qquad S_\lambda E = E S_\lambda = 0.$$

（对于对称的 $M = \sum_f \lambda_f E_f$，可以显式写为 $S_\lambda = \sum_{f:\ \lambda_f \neq \lambda} (\lambda_f - \lambda)^{-1} E_f$。）

参考文献：(Kat¯o, 1980), 第 I 章，第 5.3 节，公式 (5.26)–(5.29)（通过预解式的全纯部分及在补空间上的逆算子恒等式进行定义）。

引理 B.7 (格拉姆特征投影的一阶漂移)。 设 $M(x) = M + xM_{(1)} + O(x^2)$ 为解析微扰，设 $E(x)$ 为对应于 λ 的孤立特征值（等价地，孤立的 λ 群）的谱投影算子。令 $E := E(0)$，并设 S_λ 为 M 在 λ 处的约化预解式。则首个泰勒系数$E^{(1)} := \frac{d}{dx}E(x)\big|_{x=0}$ 满足

$$E^{(1)} \;=\; -\,E\,M^{(1)}\,S_\lambda \;-\; S_\lambda\,M^{(1)}\,E.$$

特别地，沿着 *Gram* 流 $M = -\{M, \Phi\}$，将 $M_{(1)} = \dot M$ 代入即可得到瞬时特征投影漂移。

参考文献：(Kat¯o，1980)，第 II 章，第 2 节，式（2.14）（半单情形；对于对称 M，该条件自动满足）。

引理 *B.8*（*Gram* 特征值的一阶漂移）。设 $M(x) = M + xM^{(1)} + O(x^2)$ 为解析函数，且 λ是¯ M 的一个 E 特征值，其 $d =$ 谱投影算子E的秩为 λ。令 $x(M)$ 表示从 λ分岔出的 $x(\lambda)$ 特征值的加权平均（等价地，$\bar\lambda(x) = \frac{1}{d}\mathrm{tr}(M(x)E(x))$，其中 $E(x)$ 是相应的谱投影算子）。则

$$\bar\lambda^{(1)} := \frac{d}{dx}\bar\lambda(x)\Big|_{x=0} \;=\; \frac{1}{d}\,\mathrm{tr}\big(M^{(1)}E\big).$$

沿着 *Gram* 流，$\dot{\bar\lambda}(t) = \frac{1}{d}\mathrm{tr}(\dot M(t)\ E(t)\ \dot M(t) = -\{M(t)\ \Phi(t)\}$。若 λ是单重的，且对应单位特征向量为 u，则可化简为 $\lambda^{(1)} = u^\top M^{(1)} u$。

参考文献：（Kat¯o，1980），第 II 章，§2，公式 (2.32)（按本 PDF 的编号）。

B.2. 谱测度

概述：引理 B.9 将特征 i 的维度表示为由格拉姆矩阵函数中的权重提取的 D_i。 B.10 则给出了 D_i 的对应表达式，即以特征范数表示，并按 Rayleigh 的逆值进行缩放

的系数　　框架算子（它决定了框架算子可能取值的特征值范围）

引理 B.9。令 $W_i \in \mathbb{R}^m \backslash \{0\}$ 为权重矩阵 W 的第 i 列，且 $M = W^\top W$。则：

$$D_i = \frac{M_{ii}^2}{(M^2)_{ii}} \tag{A1}$$

证明。根据格拉姆矩阵的定义，$M_{ij} = \langle W_i, W_j \rangle$，即 $M_{ii} = \|W_i\|^2$ 和 $(M^2)_{ii} = \sum_j M_{ij} M_{ji} = \sum_j (W_i^\top W_j)^2$。因此：

$$D_i = \frac{\|W_i\|^2}{\sum_j (\hat{W}_i^\top W_j)^2} = \frac{\|W_i\|^2}{\frac{1}{\|W_i\|^2} \sum_j (W_i^\top W_j)^2} =$$

$$\frac{\|W_i\|^4}{\sum_j (W_i^\top W_j)^2} = \frac{M_{ii}^2}{(M^2)_{ii}}$$

□

Lem引理 B.10。令 $W_i \in \mathbb{R}^m \backslash \{0\}$ 为 $W \in \mathbb{R}^{m \times n}$ 的第 i 列，且 D_i 如 $(A1)$ 中所定义。则　：

$$D_i = \frac{1}{\kappa_i} \|W_i\|^2, \quad \kappa_i = \frac{W_i^\top F W_i}{\|W_i\|^2}$$

证明。重写为：

$$\forall_i \, (M^2)_{ii} = e_i^\top M^2 e_i = e_i^\top (W^\top W)(W^\top W) e_i =$$

$$= (We_i)^\top F (We_i) = W_i^\top F W_i$$

因此，利用引理 A1

$$D_i = \frac{M_{ii}^2}{(M^2)_{ii}} = \frac{\|W_i\|^4}{W_i^\top F W_i} = \frac{\|W_i\|^2}{W_i^\top F W_i} \|W_i\|^2 = \frac{1}{\kappa_i} \|W_i\|^2,$$

□

通过将分数维数表示为 $M = W^\top W$和 $F = WW^\top$ 的泛函（B.9、B.10），我们可以使用 B.3 将分数维数改写为纯粹的谱形式。由于后续所有分析都将是动态的，因此不失一般性地，我们可以对权重矩阵的任意特定时间快照 $W(t)$ 及其相应的 Gram 算子和框架算子 $M(t) = W(t)^\top W(t)$ 以及 $F(t) = W(t) W(t)^\top$ 使用之前的结果。利用 B.3(iii) 提供给我们的谱分解 $F(t) = \sum_{e \in \mathcal{E}_+} \lambda_e(t) P_e(t)$，我们可以使用谱权重定义逐特征谱测度 $\mu_i(t)$：

引理 B.11（谱测度）。固定权重矩阵的一个快照 $W \in \mathbb{R}^{m \times n}$，并令 $M := W^\top W$ 和$F := WW^\top$。令 $M = \sum_{e \in E} \lambda_e E_e$ 为引理 *B.1* 中的正交谱分解，并且对于每个 $e \in E_+ := \{e : \lambda_e > 0\}$，令 $P_e := \lambda_e^+ W E_e W^\top$ 为引理 *B.3* 中诱导的激活空间投影算子。对于任意特征（列）$W_i \neq 0$，定义

$$p_{i,e} := \frac{\|P_e W_i\|^2}{\|W_i\|^2} \qquad (e \in E_+), \qquad \mu_i := \sum_{e \in E_+} p_{i,e} \, \delta_{\lambda_e}.$$

则：

1. $p_{i,\,e} \geq 0$对所有 l $e \in E_+$ 和 $\sum_{e \in E_+} p_{i,\,e} = 1$。特别地，$\mu_i$ 是一个定义良好的概率测度（有限　　*e*

原子型）且支撑集为 $\{\lambda_e:\ e\in E_+\}\subset(0,\infty)$。等价地，对于任意博雷尔集 $B\subset\mathbb{R}$，

$$\mu_i(B)\ =\ \sum_{e\in E_+:\,\lambda_e\in B} p_{i,e}.$$

2. 权重具有等价的闭式表达

$$p_{i,e}\ =\ \frac{\lambda_e\,(E_e)_{ii}}{\|W_i\|^2}\qquad (e\in E_+).$$

证明。 我们只使用引理 B.1–B.4。
(1) 由于每个 P_e 都是正交投影算子（引理 B.3(i)），$\|P_eW_i\|^2\ge 0$，因此$p_{i,\,e}\ge 0$。为证明权重之和为 1，先注意到 $W_i=We_i\in\mathrm{Im}(W)$。根据引理 B.4(iii)，$\mathbb{R}^m=\big(\bigoplus_{e\in E_+}\mathrm{Im}(P_e)\big)\oplus\ker(W^\top)$，且 $\sum_{e\in E_+}P_e$是投影到 $\mathrm{Im}(W)\ =\ker(W^\top)^\perp$上的正交投影算子。因此 $\sum_{e\in E_+}P_eW_i=W_i$。由于这些投影算子两两正交（引理 B.3(ii)），向量 $\{P_eW_i\}_{e\in E_+}$ 两两正交，因此

$$\|W_i\|^2\ =\ \Big\|\sum_{e\in E_+}P_eW_i\Big\|^2\ =\ \sum_{e\in E_+}\|P_eW_i\|^2.$$

除以 $\|W_i\|^2>0$ 得到$\sum_{e\in E_+}p_{i,\,e}=1$。因此$\mu_i=\sum_{e\in E_+}p_{i,\,e}\delta_{\lambda_e}$的总质量为1，所以它是一个概率测度；它对博雷尔集的作用是有限原子测度的标准作用。
(2) 由于 P_e是正交投影算子（引理 B.3(i)），$\|P_eW_i\|^2\ =\ W_i^\top P_eW_i$。利用 $P_e\ =\ \lambda_e^+WE_eW^\top$（引理 B.3）和 $W_i=We_i$，我们计算

$$W_i^\top P_eW_i=e_i^\top W^\top(\lambda_e^+WE_eW^\top)We_i=\lambda_e^+\,e_i^\top(W^\top W)\,E_e\,(W^\top W)\,e_i=\lambda_e^+\,e_i^\top ME_eMe_i.$$

根据引理 B.1，$ME_e\ =\ \lambda_eE_e$与 $E_eM\ =\ \lambda_eE_e$，故 $ME_eM\ =\ \lambda_e^2E_e$。因此 $W_i^\top P_eW_i\ =\ \lambda_e^+\lambda_e^2\,e_i^\top E_eE_ei\ =\lambda_e(E_e)_{ii}$。除以 $\|W_i\|^2$ 即得到$p_{i,\,e}$的所述公式。 □

引理 B.12（谱测度矩）。 固定 i，其中 $W_i\ \neq\ 0$，并令 $\mu_i\ =\ \sum_{e\in E_+}p_{i,e}\delta_{\lambda_e}$ 如引理 ***B.11*** 中所述。则对每个整数 $r\ge 1$,

$$\mathbb{E}_{\mu_i}[\lambda^r]\ :=\ \int\lambda^r\,d\mu_i(\lambda)\ =\ \frac{W_i^\top F^rW_i}{\|W_i\|^2}.$$

此外，将 *Moore–Penrose* 伪逆记作 $F^+=\sum_{e\in E_+}\lambda_e^+P_e$,

$$\frac{W_i^\top F^+W_i}{\|W_i\|^2}\ =\ \int\lambda^+\,d\mu_i(\lambda)\ =\ \mathbb{E}_{\mu_i}[\lambda^+],\qquad \textit{where } \lambda^+:=\frac{1}{\lambda}\ \textit{on}\ (0,\infty).$$

证明。 由引理 B.3(iii)，$F\ =\ \sum_{e\in E_+}\lambda_eP_e$；由引理 B.3(ii)，投影算子满足 $P_eP_f\ =\ \delta_{ef}P_e$（正交幂等元）。因此，对任意整数 $r\ge 1$,

$$F^r=\Big(\sum_{e\in E_+}\lambda_eP_e\Big)^r=\sum_{e\in E_+}\lambda_e^rP_e,$$

因为除非 $e_1=\cdots=e_r$，否则所有混合积 $P_{e_1}\cdots P_{e_r}$ 都为零。因此

$$W_i^\top F^rW_i=\sum_{e\in E_+}\lambda_e^r\,W_i^\top P_eW_i=\sum_{e\in E_+}\lambda_e^r\,\|P_eW_i\|^2,$$

这里我们利用了 P_e 是正交投影算子（引理 B.3(i)），因此 $W_i^\top P_e W_i = \|P_e W_i\|^2$。两边除以 $\|W_i\|^2$，并回想起$p_{i,\ e} = \|P_e W_i\|^2/\|W_i\|^2$，得到

$$\frac{W_i^\top F^r W_i}{\|W_i\|^2} = \sum_{e\in E_+} p_{i,e}\lambda_e^r = \int \lambda^r \, d\mu_i(\lambda) = \mathbb{E}_{\mu_i}[\lambda^r].$$

伪逆恒等式完全相同：根据定义 $F^+ = \sum_{e\in E_+} \lambda_e^+ P_e$，因此

$$\frac{W_i^\top F^+ W_i}{\|W_i\|^2} = \sum_{e\in E_+} p_{i,e}\lambda_e^+ = \int \lambda^+ \, d\mu_i(\lambda) = \mathbb{E}_{\mu_i}[\lambda^+].$$

□

这使我们能够立即将瑞利商重新表述为谱测度的一阶矩 μ_i（t）：**推论 7**。 $\kappa_i = \mathbb{E}_{\mu_i(t)}[\lambda]$

这意味着，在整个训练过程中，瑞利商遵循如下动态规律：

$$\dot\kappa_i(t) = \sum_k \dot\lambda_k(t) p_{ik}(t) + \sum_k \lambda_k(t)\dot p_{ik}(t),$$

其中，第一项表示特征值漂移，第二项表示特征在特征子空间之间的质量传输（例如特征值跳跃）。

引理 B.13（循环空间、分数幂与伪逆柯西–施瓦茨不等式）。 设 $F \in \mathbb{R}^{m\times m}$ 为对称半正定矩阵，其谱分解为 $F = \sum_{e=0} \lambda_e P_e$、　　$\lambda_k \geq 0$，其中 P_k 为正交投影算子，$P_k P_\ell = \delta_{k\ell} P_k$，且 $\sum_k P_k = I$。令

$$P := P_{\mathrm{Im}(F)} = \sum_{\lambda_k>0} P_k = I - P_0, \qquad F^{1/2} := \sum_k \lambda_k^{1/2} P_k, \qquad F^{+/2} := \sum_k \lambda_k^{+1/2} P_k.$$

对于 $x \in \mathbb{R}^m$，定义循环子空间

$$\mathcal{K}_x := \mathrm{span}\{x, Fx, F^2x, \dots\}.$$

那么：

1. $\mathcal{K}_x$ = 张成$\{P_k x :\ P_k x \neq 0\}$。特别地，维数($\mathcal{K}_x$) $\leq m$，并且存在 $r \leq m$，使得 $\mathcal{K}_x$ =张成 $\{x, Fx, \dots, F^{r-1}x\}$。
2. $F^{1/2}x \in \mathcal{K}_x$和 $F^{+/2}x \in \mathcal{K}_x$。
3. 对所有 $x \in \mathbb{R}^m$,

$$(x^\top Px)^2 \ \leq\ (x^\top Fx)\,(x^\top F^+ x),$$

并且如果 $x \in$像(F)（等价地 $Px = x$），那么

$$(x^\top x)^2 \ \leq\ (x^\top Fx)\,(x^\top F^+ x).$$

此外，等式成立当且仅当 $F^{1/2}x$与 $F^{+/2}x$线性相关，等价地当且仅当对某个$c \in \mathbb{R}$有 $Fx = c\,Px$（并且当 $x \in \mathrm{Im}(F)$时当且仅当 $Fx = cx$）。

证明。**(1)** 对于每个整数 $t \geq 0$,

$$F^t x = \sum_{k=0}^{K} \lambda_k^t \, P_k x,$$

因此 $\mathcal{K}_x \subseteq \mathrm{span}\{P_k x : \ P_k x \neq 0\}$。
对于反向包含关系，设 $S := \{k : \ P_k x \neq 0\}$，并设 $\{\lambda_k\}_{k\in S}$ 是 x 中出现的不同特征值。对于每个固定的 $j \in S$，选择一个次数为 $\leq |S|-1$ 的多项式 q_j，使得对所有 $k \in S \setminus \{j\}$ 都有 $q_j(\lambda_j) = 1$ 且 $q_j(\lambda_k) = 0$（在有限集合 $\{\lambda_k\}_{k\in S}$ 上进行拉格朗日插值）。于是

$$q_j(F)x = \sum_{k\in S} q_j(\lambda_k)\, P_k x = P_j x,$$

由于 $q_j(F)x$ 是 $\{F^t x\}_{t=0}^{|S|-1}$ 的线性组合，我们有 $P_j x \in \mathcal{K}_x$。因此 $\mathrm{span}\{P_k x : \ P_k x \neq 0\} \subseteq \mathcal{K}_x$，从而证明了相等性。稳定化 $\mathcal{K}_x = \mathrm{span}\{x, Fx, \ldots, F^{r-1}x\}$ 可由 $r := \dim(\mathcal{K}_x) \leq |S| \leq m$ 得出。(2) 利用谱展开，

$$F^{1/2}x = \sum_k \lambda_k^{1/2} P_k x, \qquad F^{+/2}x = \sum_{\lambda_k>0} \lambda_k^{-1/2} P_k x,$$

且根据 (1)，每个 $P_k x \in \mathcal{K}_x$，因此两个向量都属于 $\mathcal{K}_x$。
(3) 注意 $F^{1/2}F^{+/2} = F^{+/2}F^{1/2} = P$。令 $a := F^{1/2}x$，且 $b := F^{+/2}x$。那么

$$\langle a, b\rangle = x^\top F^{1/2}F^{+/2}x = x^\top P x, \qquad \|a\|^2 = x^\top F x, \qquad \|b\|^2 = x^\top F^{+} x.$$

在 $\mathbb{R}^m$ 中应用柯西–施瓦茨不等式，得到 $(x^\top P x)^2 = \langle a,\ b\rangle^2 \leq \|a\|^2\|b\|^2 = (x^\top F x)(x^\top F^{+} x)$。若 $x \in \mathrm{Im}(F)$，则 $Px = x$，由此可得所述的简化形式。
柯西–施瓦茨不等式取等号当且仅当 a 和 b 线性相关，即 $F^{1/2}x = c\, F^{+/2}x$。左乘 $F^{1/2}$ 得到 $Fx = c\, F^{1/2}F^{+/2}x = c\, Px$。若 $Px = x$，则可化简为 $Fx = cx$。 □

引理 B.14 ($L^2(\mu)$ 针对有限原子测度。) 设 $\mu = \sum_{k=1}^{s} p_k\, \delta_{\lambda_k}$ 是 $\mathbb{R}$ 上的一个概率测度，满足 $p_k > 0$ 和 $\sum_k p_k = 1$。那么，$L^2(\mu)$ 可通过 $f \mapsto (f(\lambda_1)\ldots, f(\lambda_s))$ 与 $\mathbb{R}^s$ 等同，其内积为

$$\langle f, g\rangle_{L^2(\mu)} = \int f(\lambda) g(\lambda)\, d\mu(\lambda) = \sum_{k=1}^{s} p_k\, f(\lambda_k) g(\lambda_k),$$

其范数为 $\|f\|^2_{L^2(\mu)} = \sum_{k=1}^{s} p_k\, |f_{(\lambda_k)}|^2$。特别地，$\{\lambda_1, \ldots, \lambda_s\}$ 上的每个函数都属于 $L^2(\mu)$，而 $L^2(\mu)$ 是一个有限维希尔伯特空间。

证明。由于 μ 支撑于有限集 $\{\lambda_1, \ldots, \lambda_s\}$，且其质量 p_k 均严格为正，因此两个函数 f, g 几乎处处相等，当且仅当对所有 k，$f(\lambda_k) = g(\lambda_k)$。因此，一个 $L^2(\mu)$ 等价类由其取值向量 $(f(\lambda_1)\ldots, f(\lambda_s)) \in \mathbb{R}^s$ 唯一确定。

此外，

$$\int |f(\lambda)|^2\, d\mu(\lambda) = \sum_{k=1}^{s} p_k\, |f(\lambda_k)|^2 < \infty$$

对于每个这样的向量（有限和），因此支撑集上的每个函数都属于 $L^2(\mu)$。内积公式直接由原子测度下积分的定义得到。完备性则源于 $L^2(\mu)$ 是有限维的。

□

引理 B.15 （实数域中的柯西–施瓦茨不等式 L^2（μ））。设 $(X, \mathcal{B}, \mu)$ 是一个概率空间，且 $u, v \in L^2$（μ）。则

$$|\langle u, v\rangle_{L^2(\mu)}|^2 \leq \|u\|^2_{L^2(\mu)}\, \|v\|^2_{L^2(\mu)}.$$

如果 $\|v\|_{L^2\ (\mu)} > 0$，则等号成立当且仅当存在 $c \in \mathbb{R}$ 使得 $u = c\, v$ μ-几乎处处成立。（如果 $\|v\|_{L^2\ (\mu)} = 0$，则 $v = 0$ μ-几乎处处成立，不等式显然成立。）

证明。 如果 $\|v\|_2 = 0$，则结论显然成立。否则定义

$$c := \frac{\langle u, v\rangle}{\|v\|_2^2}.$$

Then

$$0 \le \|u - cv\|_2^2 = \|u\|_2^2 - 2c\langle u, v\rangle + c^2\|v\|_2^2 = \|u\|_2^2 - \frac{\langle u, v\rangle^2}{\|v\|_2^2},$$

重新排列后得到 $\langle u,\ v\rangle^2 \le \|u\|_2^2\|v\|_2^2$。当且仅当 $\|u - cv\|_2^2 = 0$ 时等号成立，即 $u = cv$ μ-几乎处处成立。 □

引理 B.16 (循环等距映射 $L^2(\mu_x) \simeq \mathcal{K}_x$)。 设 $F \in \mathbb{R}^{m\times m}$ 为对称半正定矩阵，其谱分解为 $F = \sum_{k=0}^{K} \lambda_k P_k$。取定 $x \in \mathrm{Im}(F) \setminus \{0\}$，并定义（正谱）谱权重和测度

$$p_k := \frac{\|P_k x\|^2}{\|x\|^2} \quad (\lambda_k > 0), \qquad \mu_x := \sum_{\lambda_k>0} p_k\, \delta_{\lambda_k}.$$

对于定义在 $\mathrm{supp}(\mu_x)$上的 f，在 $\mathrm{Im}(F)$ 上定义函数演算如下

$$f(F) := \sum_{\lambda_k>0} f(\lambda_k)\, P_k.$$

定义 U_x： $L^2(\mu_x) \to \mathcal{K}_x$ 如下

$$U_x(f) := \frac{1}{\|x\|}\, f(F)x.$$

因此 U_x 是一个定义良好的酉同构（线性、双射且保内积）。特别地，对于所有f、 $g \in L^2$（μ_x），

$$\langle U_x(f), U_x(g)\rangle_{\mathbb{R}^m} = \int f(\lambda)g(\lambda)\, d\mu_x(\lambda),$$

并且对于定义在$\mathrm{supp}(\mu_x)$上的任意实值 h，

$$\frac{x^\top h(F)x}{\|x\|^2} = \int h(\lambda)\, d\mu_x(\lambda).$$

证明。定义良好性。 若 $f = g$ μ_x-几乎处处成立，则对于满足 $p_k > 0$的每个原子，都有 $f(\lambda_k) = g(\lambda_k)$。若 $p_k = 0$，则 $P_k x = 0$，因此改变 $f(\lambda_k)$不会改变 $f(F)x$。所以 $f(F)x = g(F)x$，且 U_x在 $L^2(\mu_x)$等价类上定义良好。

等距。 利用正交性 $P_k P_\ell = \delta_{k\ell} P_k$，

$$\begin{aligned}
\langle U_x(f), U_x(g)\rangle &= \frac{1}{\|x\|^2}\,\langle f(F)x,\, g(F)x\rangle \\
&= \frac{1}{\|x\|^2}\left\langle \sum_{\lambda_k>0} f(\lambda_k)P_k x,\ \sum_{\lambda_\ell>0} g(\lambda_\ell)P_\ell x \right\rangle \\
&= \frac{1}{\|x\|^2}\sum_{\lambda_k>0} f(\lambda_k)g(\lambda_k)\,\|P_k x\|^2 = \sum_{\lambda_k>0} p_k f(\lambda_k)g(\lambda_k) = \int fg\, d\mu_x.
\end{aligned}$$

满射性。 根据引理 B.13(1)， $\mathcal{K}_x$ = 张成$\{P_k x :\ P_k x \neq 0\}$ = 张成$\{P_k x :\ p_k > 0\}$。给定 $y = \sum_{p_k>0} a_k P_k x \in \mathcal{K}_x$，定义 f在 支撑集(μ_x) 上为 $f(\lambda_k) := a_k\|x\|$。于是 $f \in L^2(\mu_x)$（有限支撑集），并且

$$U_x(f) = \frac{1}{\|x\|}\sum_{p_k>0} f(\lambda_k)P_k x = \sum_{p_k>0} a_k P_k x = y.$$

单射性由等距性推出：若 $U_x(f) = 0$则$\|f\|^2_{L^2(\mu_x)} = \|U_x(f)\|^2 = 0$，因此 $f = 0\ \mu_x$-几乎处处成立。故 U_x是酉的。

二次型恒等式。对于任意 h 在 $\mathrm{supp}(\mu_x)$上的，

$$x^\top h(F)x = \sum_{\lambda_k>0} h(\lambda_k)\, x^\top P_k x = \sum_{\lambda_k>0} h(\lambda_k)\, \|P_k x\|^2,$$

并除以 $\|x\|^2$，得到 $\int h\, d\mu_x$。

□

C. 定理证明与推论

推论2（射影线性）。假设谱局域化成立。那么，任意特征的分数维数 D_i 由其特征范数 $\propto \|W_i\|^2$ 线性决定，斜率 k 等于其所占据子空间的特征值λ_e 的倒数 $D_i = \lambda_e^+ \|W_i\|^2$。

推论*2*（射影线性）的证明。根据引理 B.10，

$$D_i \;=\; \frac{1}{\kappa_i}\,|W_i|^2, \qquad \kappa_i \;=\; \frac{W_i^\top F W_i}{|W_i|^2}.$$

U在谱局域化（定理 1）下，$FW_i = \lambda_e\, W_i$ 对于某个 F特征值 $\lambda_e > 0$。因此

$$W_i^\top F W_i \;=\; \lambda_e\, W_i^\top W_i \;=\; \lambda_e\, |W_i|^2,$$

因此 $\kappa_i = \lambda_e$ 所以

$$D_i \;=\; \frac{|W_i|^2}{\lambda_e} \;=\; \lambda_e^+\, |W_i|^2,$$

其中 $\lambda_e^+ = 1/\lambda_e$ 因为 $\lambda_e > 0$在定理1中。

□

定理 3（紧框架分解）假设谱局域化（定理*1*）对所有特征$i \in \{1, \dots, n\}$均成立。令 $\Lambda = \{\lambda_1, \dots, \lambda_r\}$ 为 $F = WW^\top$的互不相同的正特征值，并定义指标簇及其张成空间

$$C_k := \{\, i : \lambda(i) = \lambda_k \,\}, \qquad V_k := \mathrm{span}\{\, W_i : i \in C_k \,\}.$$

于是 $\{C_k\}_{k=1}^r$ 将 $\{1, \dots, n\}$划分为若干部分，子空间 V_k 两两正交，并且

$$F \;=\; \sum_{k=1}^{r} \lambda_k\, P_{V_k} \quad \textit{(on } \mathrm{Im}(W)\textit{; plus } 0 \textit{ on } \ker(F)\textit{)}.$$

此外，在每个簇 C_k中，

$$\sum_{i \in C_k} W_i W_i^\top \;=\; \lambda_k\, I \quad \textit{on } V_k,$$

即 (W_{C_k}) 为 V_k 的紧框架，框架常数为 λ_k。

证明。根据定理 1，每个 W_i 都是 F 的特征向量，对应特征值为 $\lambda(i) > 0$，即 $FW_i = \lambda(i)\, W_i$。固定不同的特征值 $\lambda_k \neq \lambda_\ell$。由于 F 是对称半正定的，不同特征值对应的特征子空间彼此正交，因此当 $\lambda(i) \neq \lambda\, j$ 时，$W_i \perp W_j$。所以集合 C_k 彼此不相交，并构成 $\{1, \dots, n\}$ 的划分；其张成空间 V_k 两两正交。

对于任意 $v \in V_k$，记 $v = \sum_{i \in C_k} \alpha_i W_i$。于是

$$Fv = \sum_{i \in C_k} \alpha_i F W_i = \sum_{i \in C_k} \alpha_i \lambda_k W_i = \lambda_k v,$$

因此 $V_k \subseteq \mathrm{Im}\ P_k$。反过来，由于 $F = WW^\top$，我们有 $\mathrm{Im}(F) = \mathrm{Im}(W) = \mathrm{span}\{W_i\}_i$。$F$的任意特征值为 $\lambda_k > 0$ 的特征向量 v 都属于 $\mathrm{Im}(F)$，因此可唯一分解为 $v = \sum_{\ell=1}^{r} v_\ell$，其中 $v_\ell \in V_\ell$（正交直和）。应用 F 可得

$$Fv = \sum_{\ell=1}^{r} \lambda_\ell v_\ell, \qquad \text{but also} \qquad Fv = \lambda_k v = \sum_{\ell=1}^{r} \lambda_k v_\ell,$$

因此，对所有 ℓ，$(\lambda_\ell - \lambda_k)v_\ell = 0$，这迫使 $\ell \neq k$ 满足 $v_\ell = 0$。于是

$$\mathrm{Im}(P_k) \cap \mathrm{Im}(W) = V_k,$$

并且 F 在 V_k上作用为 $\lambda_k I$。由此得到

$$F = \sum_{k=1}^{r} \lambda_k P_{V_k} \quad \text{on } \mathrm{Im}(W),$$

并且 $F = 0$ 在 $\ker(F) = \mathrm{Im}(W)^\perp$。
对于紧框架断言，定义

$$F_k := \sum_{i \in C_k} W_i W_i^\top.$$

取 $v \in V_k$。对于任意 $j \notin C_k$，不同特征子空间的正交性给出 $W_j^\top v = 0$，因此

$$Fv = \sum_{i=1}^{n} W_i W_i^\top v = \sum_{i \in C_k} W_i W_i^\top v = F_k v.$$

但对于 $v \in V_k$，$Fv = \lambda_k v$，因此对于所有 $v \in V_k$，即在 V_k 上的 $F_k = \lambda_k I$。这就是框架常数为 λ_k的紧框架条件。

□

定理 4（几何的谱识别）假设谱局域化（定理 *1*）成立，使得每个簇 C_k都如定理 *3* 所述构成一个紧框架。令 $M_k := W_{C_k}^\top W_{C_k}$为局部格拉姆矩阵。则簇 C_k 的几何是一个关联方案代数 $\mathcal{A}$ 的实例，当且仅当 M_k 的特征子空间分解与 $(W_0, \ldots\ldots, W_s)$的典范层 $\mathcal{A}$相一致。

证明。我们使用标准的关联方案事实（附录 A.2）。如果 $(A_0, \ldots\ldots, A_s)$ 是 Ω 上某个关联方案的邻接矩阵，则其 Bose–Mesner 代数 $\mathcal{A}$ 具有典范正交分解

$$\mathbb{R}^\Omega = W_0 \oplus \cdots \oplus W_s$$

具有正交投影算子（$S_0, \ldots\ldots, S_s$），且每个 $X \in \mathcal{A}$都保持每个分层 W_e。此外，(S_e)构成 $\mathcal{A}$ 的一组两两正交幂等元基，因此每个 $X \in \mathcal{A}$ 都有唯一展开式

$$X = \sum_{e=0}^{s} \theta_e S_e,$$

并在 $W_e = \mathrm{Im}(S_e)$ 上作为标量 θ_e 起作用。
(⇒) 如果簇几何是 $\mathcal{A}$ 的一个实例，那么定义该几何的不变矩阵（此处为 M_k）属于 $\mathcal{A}$。因此$M_k = \sum_e \theta_e S_e$，所以每个层 W_e 都是一个不变子空间，在该子空间上，M_k 的作用为标量乘法。等价地，M_k 的谱分解由层分解所决定。

($\Leftarrow$) 反过来，假设 M_k 的特征子空间分解与各层相一致，即 M_k 在$\mathbb{R}^\Omega = \bigoplus_e W_e$上为块标量矩阵。那么存在标量 θ_e，使得对所有 e 都有 $M_k|_{W_e} = \theta_e I$。因此 $M_k = \sum_e \theta_e S_e \in \mathcal{A}$，因为$S_e \in \mathcal{A}$ 且 $\mathcal{A}$ 是一个线性空间。由此，簇几何是关联方案代数的一个实例。 □

推论 5 (单纯形识别)簇 C_k 是单纯形几何，当且仅当其格拉姆矩阵 M_k 恰好有两个不变子空间

$$W_0 = \ker(M_k) = \operatorname{span}(\mathbf{1}), \qquad W_1 = \operatorname{Im}(M_k) = \mathbf{1}^\perp,$$

其特征值为 $\Lambda(W_0) = 0$ 和 $\Lambda(W_1) = \lambda_k = \frac{|C_k|}{\dim(V_k)}$。

证明。 在单纯形关联方案（附录 A.3.1）中，Bose–Mesner 代数恰好有两个层：常向量空间 $W_0 =$ span(1)及其正交补 $W_1 = \mathbf{1}^\perp$，对应的投影算子为 $S_0 = \frac{1}{p}J$ 和$S_1 = I - \frac{1}{p}J$，其中 $p = |C_k|$。根据定理 4，属于单纯形方案等价于谱分解/不变子空间分解恰好沿着这两个层进行。

关于特征值：在单纯形模型中，向量是单位向量且中心化（和为零），而附录 A.3.1 给出

$$M_k = \frac{p}{p-1} S_1.$$

因此， $M_k\mathbf{1} = 0$和 M_k 在 $\mathbf{1}^\perp$上充当标量 $\frac{p}{p-1}$。

在由 p 个向量张成 d 维子空间的单位紧框架的紧框架参数化下，框架常数为 p/d（迹恒等式；参见引理 A.7）。对于单纯形， $p = d + 1$，因此

$$\frac{p}{p-1} = \frac{d+1}{d} = \frac{p}{d} = \frac{|C_k|}{\dim(V_k)},$$

与所述 λ_k 相符。 □

D. 扰动情形

定理 8 (缺陷分解：近饱和 $\Leftrightarrow$ 总松弛量较小)。 设 $W \in \mathbb{R}^{m\times n}$ 的列为w_1、.....、 w_n，且框架算子为 F：$= WW^\top$。令 r：$=$秩(W) $=$秩(F)。定义杠杆分数ℓ_i：$= w_i^\top F^+ w_i$以及分数维数 D_i：$= \|w_i\|^4/(w_i^\top F w_i)$（对于 $w_i \neq 0$）。定义相对松弛量σ_i：$= 1 - D_i/\ell_i \in [0,1]$。则：

*1. (*容量预算*)* $\sum_{i=1}^n \ell_i = r$。

*2. (*逐点界*)* $D_i \le \ell_i$ 对所有 i成立，且等号成立当且仅当 μ_i 是狄拉克质量（等价地： w_i 是F的特征向量）。

*3. (*精确缺陷恒等式*)*

$$r - \sum_{i=1}^n D_i \;=\; \sum_{i=1}^n (\ell_i - D_i) \;=\; \sum_{i=1}^n \ell_i\,\sigma_i.$$

特别地，若 $r = m$，则 $m - \sum_i D_i = \sum_i \ell_i\sigma_i$。

证明。 (1) 利用 $\ell_i = w_i^\top F^+ w_i$ 和迹的循环性，

$$\sum_{i=1}^n \ell_i = \sum_{i=1}^n w_i^\top F^+ w_i = \operatorname{tr}\left(\sum_{i=1}^n F^+ w_i w_i^\top\right) = \operatorname{tr}\left(F^+ \sum_{i=1}^n w_i w_i^\top\right) = \operatorname{tr}(F^+F).$$

对于对称半正定矩阵 F，矩阵 F^+F是到 Im(F) 上的正交投影算子，因此 tr(F^+F) $=$rank(F) $= r$。

(2) 固定 i并令 $x := w_i$。由于 x是 W的一列，我们有 $x \in \mathrm{Im}(W) = \mathrm{Im}(F)$。令 $a := F^{1/2}x$和$b := F^{+1/2}x$。根据柯西–施瓦茨不等式，

$$\langle a, b\rangle^2 \le \|a\|^2\|b\|^2.$$

计算每一项：

$$\langle a, b\rangle = x^\top F^{1/2}F^{+1/2}x = x^\top (F^+F)^{1/2}x = x^\top \Pi_{\mathrm{Im}(F)}x = \|x\|^2,$$

因为 $x \in \mathrm{Im}(F)$。此外 $\|a\|^2 = x^\top Fx$且 $\|b\|^2 = x^\top F^+x = \ell_i$。因此

$$\|x\|^4 \le (x^\top Fx)(x^\top F^+x) \quad\Longrightarrow\quad \frac{\|x\|^4}{x^\top Fx} \le x^\top F^+x \quad\Longrightarrow\quad D_i \le \ell_i.$$

柯西–施瓦茨不等式中等号成立当且仅当 a与 b线性相关，即 $F^{1/2}x = cF^{+1/2}x$（对某个 $c \in \mathbb{R}$）。乘以 $F^{1/2}$得到 $Fx = c\,\Pi_{\mathrm{Im}(F)}x = cx$，因此 x是 F的特征向量（特征值为 $c > 0$，因为$x \in \mathrm{Im}(F)$ 且 $x \neq 0$）。等价地，x完全位于单个正特征值特征空间中，这即表明特征谱测度 μ_i 是狄拉克质量。

(3) 根据定义 $\ell_i\sigma_i = \ell_i - D_i$，因此

$$\sum_{i=1}^n \ell_i\sigma_i = \sum_{i=1}^n (\ell_i - D_i) = \Big(\sum_{i=1}^n \ell_i\Big) - \sum_{i=1}^n D_i.$$

利用 (1)，$\sum_i \ell_i = r$，得到 $r - \sum_i D_i = \sum_i \ell_i\sigma_i$。满秩特例是带有$r = m$的同一单位元。

□

定理 9 （ε 饱和下的离域残差界）。假设 $\sum_{i=1}^n D_i \ge (1-\varepsilon)\, r$。那么对于任意 $\tau \in (0,1)$，

$$\frac{1}{r}\sum_{i:\,\sigma_i \ge \tau} \ell_i \;\le\; \frac{\varepsilon}{\tau}.$$

证明。由定理 8(3)，

$$\sum_{i=1}^n \ell_i\sigma_i = r - \sum_{i=1}^n D_i \le \varepsilon r.$$

在指标集 $S_\tau := \{i:\ \sigma_i \ge \tau\}$上，我们有 $\ell_i\sigma_i \ge \tau\ell_i$，因此

$$\varepsilon r \;\ge\; \sum_{i=1}^n \ell_i\sigma_i \;\ge\; \sum_{i\in S_\tau} \ell_i\sigma_i \;\ge\; \tau \sum_{i\in S_\tau} \ell_i.$$

两边除以 τr，得到 $(1/r)\sum_{i:\ \sigma_i\ge\tau} \ell_i \le \varepsilon/\tau$。

□

定理 10 （特征向量残差等于特征值方差）。设 $F = \sum_{k:\ \lambda_k>0} \lambda_k P_k$ 是 F 在 $\mathrm{Im}(F)$ 上的谱分解。对于 $w_i \neq 0$，定义权重 $p_{ik} := \|P_k w_i\|^2/\|w_i\|^2$ 以及相应的谱测度$\mu_i := \sum_{k:\ \lambda_k>0} p_{ik}\,\delta_{\lambda_k}$。设 $\kappa_i := \mathbb{E}_{\mu_i}[\lambda]$。则

$$\frac{\|Fw_i - \kappa_i w_i\|^2}{\|w_i\|^2} \;=\; \mathrm{Var}_{\mu_i}(\lambda) \;=\; \mathbb{E}_{\mu_i}\big[(\lambda - \kappa_i)^2\big].$$

特别是，

$$\mathrm{CV}_i := \frac{\sqrt{\mathrm{Var}_{\mu_i}(\lambda)}}{\mathbb{E}_{\mu_i}[\lambda]} = \frac{\|Fw_i - \kappa_i w_i\|}{\kappa_i \|w_i\|}$$

当且仅当 w_i 是 F 的特征向量（等价地，μ_i 是狄拉克分布）时，取值为零。

证明。 设 $w_i = \sum_k P_k w_i$。 则

$$Fw_i = \sum_k \lambda_k P_k w_i, \qquad \kappa_i w_i = \sum_k \kappa_i P_k w_i,$$

so

$$Fw_i - \kappa_i w_i = \sum_k (\lambda_k - \kappa_i)\, P_k w_i.$$

利用不同谱子空间之间的正交性，

$$\|Fw_i - \kappa_i w_i\|^2 = \sum_k (\lambda_k - \kappa_i)^2 \|P_k w_i\|^2.$$

两边除以 $\|w_i\|^2$，得到

$$\frac{\|Fw_i - \kappa_i w_i\|^2}{\|w_i\|^2} = \sum_k p_{ik}(\lambda_k - \kappa_i)^2 = \mathbb{E}_{\mu_i}\big[(\lambda - \kappa_i)^2\big] = \mathrm{Var}_{\mu_i}(\lambda).$$

当且仅当 λ 在 μ_i-几乎处处为常数时，方差为零；也就是说，μ_i 是某个 $\lambda(i) > 0$上的狄拉克分布，等价地，w_i 完全位于单个特征子空间中，即 $Fw_i = \lambda(i) w_i$。将公式除以 κ_i^2 并开平方，即可立即得到变异系数 CV_i。 □

定理 11 （带局部化 $\Rightarrow$ 拟紧性（康托罗维奇界））。 固定 i 并假设 μ_i 支撑集为$[\lambda_i^-, \lambda_i^+] \subset (0, \infty)$。令 $\kappa_i^\star := \lambda_i^+ / \lambda_i^- \geq 1$。则

$$\kappa_i\, h_i = \mathbb{E}_{\mu_i}[\lambda]\, \mathbb{E}_{\mu_i}[\lambda^{-1}] \;\leq\; \frac{(\kappa_i^\star + 1)^2}{4\kappa_i^\star},$$

因此

$$\sigma_i = 1 - \frac{D_i}{\ell_i} \leq 1 - \frac{4\kappa_i^\star}{(\kappa_i^\star + 1)^2}.$$

等价地，$\omega_i := (\lambda_i^+ - \lambda_i^-)/(\lambda_i^+ + \lambda_i^-) \in [0, 1)$，

$$\sigma_i \leq \omega_i^2.$$

证明。 步骤 1（此处使用的标量形式的康托罗维奇不等式）。设 X 是一个支撑集为$[a, b] \subset (0, \infty)$的任意随机变量。在 $[a, b]$上逐点有，

$$(x - a)(b - x) \geq 0 \iff x^2 - (a + b)x + ab \leq 0 \iff x + \frac{ab}{x} \leq a + b.$$

取期望可得

$$\mathbb{E}[X] + ab\, \mathbb{E}[X^{-1}] \leq a + b.$$

根据算术—几何平均不等式，$\mathbb{E}[X] + ab\, \mathbb{E}[X^{-1}] \geq 2\sqrt{ab\, \mathbb{E}[X]\mathbb{E}[X^{-1}]}$，因此

$$2\sqrt{ab\, \mathbb{E}[X]\mathbb{E}[X^{-1}]} \leq a + b \implies \mathbb{E}[X]\mathbb{E}[X^{-1}] \leq \frac{(a + b)^2}{4ab}.$$

步骤 2（应用于 $\lambda \sim \mu_i$）。取 $X = \lambda, a = \lambda_i^-, b = \lambda_i^+$。于是

$$\mathbb{E}_{\mu_i}[\lambda]\mathbb{E}_{\mu_i}[\lambda^{-1}] \le \frac{(\lambda_i^- + \lambda_i^+)^2}{4\lambda_i^-\lambda_i^+} = \frac{(\kappa_i^\star + 1)^2}{4\kappa_i^\star}.$$

这是第一个显示的不等式。
步骤 3（转化为松弛界）。利用谱权重 p_{ik}，可得

$$w_i^\top F w_i = \sum_k \lambda_k \|P_k w_i\|^2 = \|w_i\|^2 \mathbb{E}_{\mu_i}[\lambda], \qquad w_i^\top F^+ w_i = \sum_k \lambda_k^{-1} \|P_k w_i\|^2 = \|w_i\|^2 \mathbb{E}_{\mu_i}[\lambda^{-1}].$$

因此，$\kappa_i = \mathbb{E}_{\mu_i}[\lambda]$ 和 $\ell_i = \|w_i\|^2 h_i$ 满足 $h_i := \mathbb{E}_{\mu_i}[\lambda^{-1}]$。此外，$D_i = \|w_i\|^4/(w_i^\top F w_i) = \|w_i\|^2/\kappa_i$。因此

$$\frac{D_i}{\ell_i} = \frac{\|w_i\|^2/\kappa_i}{\|w_i\|^2 h_i} = \frac{1}{\kappa_i h_i} = \frac{1}{\mathbb{E}_{\mu_i}[\lambda]\mathbb{E}_{\mu_i}[\lambda^{-1}]}.$$

结合关于 $\kappa_i h_i$ 的上界可得

$$\frac{D_i}{\ell_i} \ge \frac{4\kappa_i^\star}{(\kappa_i^\star + 1)^2} \quad \Longrightarrow \quad \sigma_i = 1 - \frac{D_i}{\ell_i} \le 1 - \frac{4\kappa_i^\star}{(\kappa_i^\star + 1)^2}.$$

最后，利用 $\omega_i = (\lambda_i^+ - \lambda_i^-)/(\lambda_i^+ + \lambda_i^-)$ 和 $\kappa_i^\star = \lambda_i^+/\lambda_i^-$，

$$1 - \frac{4\kappa_i^\star}{(\kappa_i^\star + 1)^2} = \frac{(\kappa_i^\star - 1)^2}{(\kappa_i^\star + 1)^2} = \left(\frac{\lambda_i^+ - \lambda_i^-}{\lambda_i^+ + \lambda_i^-}\right)^2 = \omega_i^2.$$

□

定理12(　　与近饱和相一致的三类分类法。 定理 *11* 和推论 *9* 给出以下依据：

1. 局　域化（狄拉克）特征：$\sigma_i = 0$（等价地，变异系数$_i = 0$），即 μ_i 是狄拉克，且 w_i 是特征向 *r.*

2. 带局部化特征：$\omega_i \ll 1$，因此 $\sigma_i \le \omega_i^2 \ll 1$，尽管 μ_i 不一定是狄拉克。

3. 宽带（非局部化残差）特征：$\omega_i = \Theta(1)$ 且 变异系数$_i$ 非平凡；这些会导致 $\sigma_i > 0$，但在ε近饱和条件下，其总杠杆质量很小。

证明。 (1) 根据定理 8(2)，$\sigma_i = 0$ 当且仅当 $D_i = \ell_i$ 当且仅当 w_i 是特征向量，等价地 μ_i 是狄拉克。引理 10 表明这也等价于变异系数$_i = 0$。
(2) 若 μ_i 支撑在窄带上，且 $\omega_i \ll 1$，定理 11给出 $\sigma_i \le \omega_i^2 \ll 1$，刻画了完美局域化的“扰动”。

(3) 若 $\omega_i = \Theta(1)$（宽支撑集），则通常 方差$_{\mu_i}(\lambda)$ 不可忽略，因此根据引理 变异系数$_i$ 非平凡；引理 10还表明 $\sigma_i > 0$，除非 μ_i 退化为狄拉克。在 ε-饱和条件下，推论9 对任意满足 $\sigma_i \ge \tau$的集合的杠杆加权质量给出上界，这迫使宽带/高松弛量群体仅占杠杆预算的一小部分。

□

微扰状态下的经验解释。W 在不同稀疏度下始终保持满秩，而 $\sum_i D_i/m$ 则以依赖于稀疏度的方式略低于 1。在满秩状态下，定理 $r = m$，定理 8 迫使单位元 $m - \sum_i D_i = \sum_i \ell_i \sigma_i$，因此，观察到的未达到饱和的偏差不能归因于秩缺陷；它是一个总松弛项。

与此同时，本文指出，稀疏度越高，特征在“谱上的分布越分散”（即谱局域化程度越低）。引理 10 将“谱上的分布越分散”形式化为更大的特征值离散度 方差$_{\mu_i}(\lambda)$(等价于更大的变异系数$_i$)，这正是无法成为特征向量/狄拉克态的障碍。这解释了为何稀疏度增加可能与更强的离域化诊断相关，即使全局容量使用率仍接近饱和。

“松弛量最小”与“存在一定程度的离域化”看似并存，这一现象可由推论 9解释：如果 $\sum_i D_i \geq (1-\varepsilon)r$较小 ε (近饱和)，那么对于任意固定的离域阈值 τ，

$$\frac{1}{r}\sum_{i:\sigma_i \geq \tau} \ell_i \leq \frac{\varepsilon}{\tau}.$$

因此，近饱和并不排斥离域／宽带特征；它排斥的是这些特征承载大量总杠杆质量。实证结果显示，高松弛特征的杠杆加权尾部质量在不同稀疏度下始终很小，符合这一预测：离域是允许的，但仅限于杠杆质量中一个很小的“残余”子集（附录 I）。

最后，定理 11给出了一种清晰的微扰机制：如果大多数特征仍保持带局域，且相对半宽较小 $\omega_i \ll 1$，那么它们会自动具有 $\sigma_i \leq \omega_i^2 \ll 1$，因此对总松弛量的贡献可以忽略不计。由稀疏度驱动的离域化，进而可理解为少数特征的 逐渐展宽 μ_i（宽带残余），这会提高它们的 变异系数$_i$ 和 σ_i，但不会实质性地改变全局预算，因为它们的总杠杆权重很小。

E. 局域化与离域化

E.1. HT-SR

Charles Martin 和 Michael Mahoney 的实证研究表明，随着模型的泛化能力提高，模型在训练过程中会“自正则化”。所谓的重尾自正则化（HT-SR）现象学（Martin 和 Mahoney，2021a年；b；2020年）受到随机矩阵理论（RMT）的启发，其研究对象是层权重矩阵的经验谱密度（ESD）。更具体地说，考虑实值权重矩阵 W$_l \in \mathbb{R}^{N\times M}$，其奇异值分解（SVD）为 W$_l$ = U$\mathbf{\Sigma}$V*。这里，$\nu_i = \mathbf{\Sigma}_{ii}$ 是第 i 个奇异值，$p_i = \nu_i^2 / \sum_i \nu_i^2$是奇异值上的谱概率分布，而 X$_l$ = $\frac{1}{N}$W$_l^T$W$_l$是样本协方差矩阵。通过计算其特征值 Xv$_i$ = λ_iv$_i$，其中 $\forall_{i=1}$、$\cdots$、$_M\lambda_i = \nu_i^2$，作者发现，经验谱分布（ESD）从 Marchenko-Pastur（MP）／Tracy-Widom（TW）谱统计逐渐转变为 $P_{W_{ij}}(X) \sim \frac{1}{x^{1+\mu}}$，且具有 $\mu > 0$，即幂律（PL）／重尾（HT）统计。尽管 HT-SR 只关注特征值的行为，我们认为，这一观察只是数学物理中安德森局域化研究计划的表层特征（Hundertmark，2008年）。

E.2. 安德森局域化

关于随机算子和无序系统谱行为的文献将 Tracy-Widom 谱分布归类为具有离域特征向量的高斯普适类的特征值统计，例如高斯酉系综（Tao，2012），而幂律/重尾谱统计则代表具有局域特征向量的随机矩阵（Aizenmanand Molchanov，1993年；Mirlinetal.，1996年；Edwards 和 Thouless，1972）。

F. 梯度流

F.1. 梯度流动力学

到目前为止，我们的所有推导都基于这样一个假设：模型已经达到容量饱和，因此所得结果是静态描述。如果要给出其中一种动力学描述，就需要谨慎处理。对数学物理学中类似系统进行显式求解（Atiyah 和 Sutcliffe，2003；Smale，1998）仍然是一个悬而未决的问题。尽管如此，我们在第3 节引入的工具，已经足以让我们掌握这些动力学，从而表明我们最初关于容量饱和的假设是有充分依据的。

定理13 (梯度流**)**。 设 $M = W^\top W$为自编码器，且 $\Phi = \mathrm{Sym}(\nabla_M L)$ 为梯度核，其中*F.2*:

$$\Phi = -\mathbb{E}_x[\boldsymbol{\delta}_{(x)}\mathbf{x}^\top + x\boldsymbol{\delta}_{(x)}^\top]$$

$$\boldsymbol{\delta}_{(x)} = I \odot (\mathbf{x} - \mathbf{x}') \odot \mathbb{I}(M\mathbf{x} + b > 0).$$

于是:

$$\dot{M} = -\{M, \Phi\} = -(M\Phi + \Phi M) \tag{F.1}$$

推论 14(投影算子动力学**)**。

$$\dot{E}_e = \sum_{e \neq f} \frac{\lambda_e + \lambda_f}{\lambda_e - \lambda_f} \left(E_e \Phi E_f + E_e \Phi E_f \right) \tag{F.2}$$

推论 15（特征值漂移）。

$$\dot{\lambda}_e = -2\lambda_e \frac{\mathrm{tr}(E_e \Phi E_e)}{\dim(E_e)} \tag{F.3}$$

推论 16 （谱质量传输）。 定义 $q_{i,\ e}\ (t)\ =\ (E_e\ (t)\ )_{\ ii} = \lambda_e^+ p_{i,\ e} M_{ii}$，表示支撑在特征子空间 e上的特征 i的质量。因此，其时间演化为:

$$\dot{q}_{i,e} = \sum_{e \neq f} \mathcal{T}^{(i)}_{e \to f} = \sum_{e \neq f} 2 \left(\frac{\lambda_e + \lambda_f}{\lambda_e - \lambda_f} \right) (E_e \Phi E_f)_{ii} \tag{F.4}$$

推论 17 （容量饱和的稳定性）。 如果梯度核 Φ 与格拉姆算子的谱投影算子 $[E_e, \Phi] = 0 \quad \forall \lambda$可交换，则该配置是一个谱不动点。均匀稀疏性下的紧框架满足此条件，这意味着

令 $W \in \mathbb{R}^{m \times n}$ 为权重矩阵， $M = W^\top W \in \mathbb{R}^{n \times n}$ 为其格拉姆矩阵。权重上的标准梯度流为 $\dot{W} = -\nabla_W L$。我们假设梯度是根据参数坐标上的普通点积定义的，即将参数视为处于平坦的 $\mathbb{R}^{mn}$ 中，不存在曲率或加权。使用弗罗贝尼乌斯内积 $\langle A,\ B \rangle_F = \mathrm{tr}(A^\top B)$，我们有:

$$dL = \sum_{a,b} \frac{\partial L}{\partial W_{ab}} dW_{ab} = \mathrm{tr}\left((\nabla_W L)^\top dW \right) = \langle \nabla_W L, dW \rangle_F$$

让我们推导 M 上的诱导流。首先，计算微分:

$$dM = d(W^\top W) = dW^\top W + W^\top dW$$

.

$$\begin{aligned} \langle \nabla_M L, dM \rangle_F = \mathrm{tr}\left((\nabla_M L)^\top dM \right) = \\ = \mathrm{tr}\left((\nabla_M L)^\top (dW^\top W + W^\top dW) \right) = \\ = \mathrm{tr}\left((\nabla_M L)^\top dW^\top W \right) + \mathrm{tr}\left((\nabla_M L)^\top W^\top dW \right) \end{aligned}$$

为了将其转换为 ∇W 位于右侧的弗罗贝尼乌斯内积，我们可以利用转置的循环性质，以及 $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$:

$$\begin{aligned} \mathrm{tr}\left(W^\top dW \nabla_M L \right) = \mathrm{tr}((\nabla_M L) W^\top dW) = \\ = \mathrm{tr}((W(\nabla_M L)^\top)^\top dW) = \langle W(\nabla_M L)^\top, dW) \end{aligned}$$

同样，对于右侧，我们有:

$$\begin{aligned} \mathrm{tr}((\nabla_M L)^\top W^\top dW) = \mathrm{tr}((W \nabla_M L)^\top dW) = \langle W \nabla_M L, dW \rangle \\ \Rightarrow dL = \langle W(\nabla_M L)^\top + W \nabla_M L, dW \rangle = \langle \nabla_W L, dW \rangle \\ \Rightarrow \nabla_W L = W \left(\nabla_M L + (\nabla_M L)^\top \right) \end{aligned}$$

因此， M 上的诱导流为:

$$\dot{M} = \dot{W}^\top W + W^\top \dot{W} = -\left(\nabla_W L \right)^\top W - W^\top (\nabla_W L) =$$

$$\begin{aligned} = -(\nabla_M L + (\nabla_M L)^\top)^\top W^\top W \\ -W^\top W (\nabla_M L + (\nabla_M L)^\top) \\ = -(\Psi + \Psi^\top) M - M(\Psi + \Psi^\top) \end{aligned}$$

这实际上表明，只有对称化梯度会驱动 Gram 流（根据弗罗贝尼乌斯内积，这一点是意料之中的），并且说明我们需要将核梯度定义为：

$$\Phi := 2\text{Sym}(\Psi) = (\Psi + \Psi^\top),$$

因此，我们诱导出的梯度流变为：

$$\dot{M} = -\{\Phi, M\} = -(\Phi M + M\Phi) = -\{M, \Phi\} \tag{F.5}$$

由梯度核 $\Phi = \text{Sym}(\nabla_M L)$ 驱动。

F.2. 显式形式

回顾均方误差损失函数：

$$L = \frac{1}{2}\sum_x p(x)\sum_i I_i\left(x_i - \text{ReLU}\left(\sum_k M_{ik}x_k + b_i\right)\right)^2$$

根据链式法则，对于 $u_i = \sum_k M_{ik}x_k + b_i$ 和 $x_i' = \text{ReLU}(u_i)$，我们有：

$$\frac{\partial L}{\partial M_{ij}} = \sum_l \frac{\partial L}{\partial x_l'}\cdot\frac{\partial x_l'}{\partial u_l}\cdot\frac{\partial u_l}{\partial M_{ij}},$$

其中

$$\frac{\partial L}{\partial x_l'} = -I_l(x_l - x_l'), \quad \frac{\partial x_l'}{\partial u_l} = \mathbb{I}(u_l > 0), \quad \frac{\partial u_l}{\partial M_{ij}} = \frac{\partial}{\partial M_{ij}}\left(\sum_k M_{lk}x_k + b_k\right),$$

其中后者仅当 $l = i$时非零。因此，在期望形式下

$$\Rightarrow \frac{\partial L}{\partial M_{ij}} = -\mathbb{E}_x\left[I_i(x_i - x_i')\cdot\mathbb{I}\left(\sum_k M_{ik}x_k + b_i > 0\right)\cdot x_j\right]$$

为恢复矩阵形式，我们定义有效误差向量：

$$\boldsymbol{\delta}_{(x)} = I \odot (\mathbf{x} - \mathbf{x}') \odot \mathbb{I}(M\mathbf{x} + b > 0),$$

其中 $\odot$ 为Hadamard积。可以看到，元素$(\nabla_M L)_{ij}$ 对应于第i个误差项 δ_i 与第 j个输入 x_j之积的期望。因此：

$$\begin{aligned}\nabla_M L &= -\mathbb{E}_x[\boldsymbol{\delta}_{(x)}\mathbf{x}^\top] \\ \Phi = -\mathbb{E}_x[\boldsymbol{\delta}_{(x)}\mathbf{x}^\top + \mathbf{x}\boldsymbol{\delta}_{(x)}^\top] &= -\sum_x p(x)\left[\boldsymbol{\delta}_{(x)}\mathbf{x}^\top + \mathbf{x}\boldsymbol{\delta}_{(x)}^\top\right]\end{aligned} \tag{2}$$

稀疏性机制源于将 Φ 分解为对角部分（特征收益） Φ_{ii} 和非对角部分（干扰）Φ_{ij} （$i \neq j$）。更具体地说，注意：

$$\begin{aligned}\Phi_{ij} &= -\sum_x p(x)(\delta_{(x),i}x_j + x_i\delta_{(x),j}) \\ \Rightarrow \Phi_{ii} &= -2\sum_x p(x)\delta_{(x),i}x_i = -2\sum_x p(x)\left[I_i(x_i - x_i')\cdot\mathbb{I}(u_i > 0)\cdot x_i\right]\end{aligned}$$

F.3. 关联方案约化

引理 F.1（Γ 对 M、G、Φ 的不变性）。

根据引理 4.5 和 4.7，我们得到关联方案上的如下展开：

$$M(t)=\sum_{r=0}^{R}\theta_r(t)A_t \quad , \quad \Phi(t)=\sum_{r=0}^{R}\varphi_r(t)A_r$$

L让我们使用诱导梯度流方程 (1) 对系数作等同，并直接代入 y:

$$M\Phi=\left(\sum_r\theta_rA_r\right)\left(\sum_s\varphi_sA_s\right)=\sum_{r,s}\theta_r\varphi_s(A_rA_s)=$$

$$=\sum_{r,s}\theta_r\varphi_s\sum_u c_{rs}^uA_u=\sum_u\left(\sum_{r,s}\theta_r\varphi_sc_{rs}^u\right)A_u$$

$$\Phi M=\sum_{r,s}\varphi_r\theta_s(A_rA_s)=\sum_u\left(\sum_{r,s}\varphi_r\theta_sc_{rs}^u\right)A_u$$

$$\dot{M}=-(M\Phi+\Phi M)=$$

$$=-\sum_u\left(\sum_{r,s}\theta_t\varphi_sc_{rs}^u+\sum_{r,s}\varphi_r\theta_sc_{rs}^u\right)A_u$$

$$\Rightarrow\dot{\theta}_u=-\sum_{r,s}\theta_r\varphi_sc_{rs}^u-\sum_{r,s}\varphi_r\theta_sc_{rs}^u=-\sum_{r,s}\theta_r\varphi_s(c_{rs}^u+c_{sr}^u) \tag{F.6}$$

定理 18 (梯度流（定理 6）)。设 $M=W^\top W$，并定义（对称）梯度核 $\Phi=\mathrm{Sym}(\nabla_ML)$，其显式形式为

$$\Phi=-\mathbb{E}_x[\delta(x)x^\top+x\delta(x)^\top],\qquad \delta(x)=I\odot(x-x')\odot\mathbf{1}(Mx+b>0),$$

如附录 *E.1 /* 方程 *(E.1)* 所示。则诱导 *Gram* 流为

$$\dot{M}=-\{M,\Phi\}=-(M\Phi+\Phi M).$$

证明。设 $W(t)$ 遵循权重空间梯度流 $\dot{W}=-\nabla_WL$。对 $M=W^\top W$ 求导：

$$\dot{M}=\dot{W}^\top W+W^\top\dot{W}=-(\nabla_WL)^\top W-W^\top(\nabla_WL).$$

To e用 ∇_ML 表示 ∇_WL，使用 $dM=dW^\top W+W^\top dW$ 和 Frobenius 内积：

$$dL=\langle\nabla_ML,dM\rangle=\langle\nabla_ML,dW^\top W+W^\top dW\rangle=\langle W(\nabla_ML+(\nabla_ML)^\top),dW\rangle.$$

因此

$$\nabla_WL=W(\nabla_ML+(\nabla_ML)^\top)=W\Phi,$$

其中 $\Phi=(\nabla_ML+(\nabla_ML)^\top)=2\,\mathrm{Sym}(\nabla_ML)$。代入 $\dot{M}$ 得

$$\dot{M}=-(W\Phi)^\top W-W^\top(W\Phi)=-(\Phi W^\top W+W^\top W\Phi)=-(\Phi M+M\Phi)=-\{M,\Phi\}.$$

Φ 的显式形式在附录 E.1 中计算得出，从而得到所述表达式。 □

定理 19（投影算子动力学（推论 7））。设 $M=\sum_e \lambda_e E_e$ 是 M 的谱分解，其中 E_e 为正交谱投影算子。在 *Gram* 流 $M=-\{M,\Phi\}$ 下，

$$\dot{E}_e=\sum_{f\neq e}\frac{\lambda_e+\lambda_f}{\lambda_f-\lambda_e}\left(E_e\Phi E_f+E_f\Phi E_e\right).$$

证明。对 $M^{(1)}=\dot{M}$ 应用 Kato 一阶特征投影漂移公式（引理 B.7）。对于带有投影算子 E_e 的孤立特征值 λ_e，引理 B.7 给出

$$\dot{E}_e=-E_e\dot{M}\,S_{\lambda_e}-S_{\lambda_e}\dot{M}\,E_e,$$

其中，对于对称的 M，约化预解式为（引理 B.6）

$$S_{\lambda_e}=\sum_{f\neq e}\frac{1}{\lambda_f-\lambda_e}E_f.$$

代入 $\dot{M}=-\{M,\Phi\}=-(M\Phi+\Phi M)$，得到

$$\dot{E}_e=E_e(M\Phi+\Phi M)S_{\lambda_e}+S_{\lambda_e}(M\Phi+\Phi M)E_e.$$

利用 $ME_f=\lambda_f E_f$ 与 $E_eM=\lambda_e E_e$，对于 $f\neq e$，我们有

$$E_e(M\Phi+\Phi M)E_f=E_eM\Phi E_f+E_e\Phi ME_f=(\lambda_e+\lambda_f)\,E_e\Phi E_f.$$

因此

$$E_e(M\Phi+\Phi M)S_{\lambda_e}=\sum_{f\neq e}\frac{\lambda_e+\lambda_f}{\lambda_f-\lambda_e}E_e\Phi E_f,$$

并且类似地

$$S_{\lambda_e}(M\Phi+\Phi M)E_e=\sum_{f\neq e}\frac{\lambda_e+\lambda_f}{\lambda_f-\lambda_e}E_f\Phi E_e.$$

相加即得所述公式。 □

定理 20（特征值漂移（推论 8））**.** 沿着 *Gram* 流 $\dot{M}=-\{M,\Phi\}$，特征值满足

$$\dot{\lambda}_e=-2\lambda_e\frac{\operatorname{tr}(E_e\Phi E_e)}{\dim(E_e)}.$$

证明。引理 B.8 给出了特征值 λ_e 的一阶特征值漂移，其谱投影算子为 E_e，秩为 $d_e=\operatorname{tr}(E_e)$：

$$\dot{\lambda}_e=\frac{1}{d_e}\operatorname{tr}(\dot{M}\,E_e).$$

现在 $\dot{M}=-(M\Phi+\Phi M)$，因此

$$\operatorname{tr}(\dot{M}E_e)=-\operatorname{tr}(M\Phi E_e)-\operatorname{tr}(\Phi ME_e).$$

利用循环性和 $ME_e=\lambda_e E_e$，

$$\operatorname{tr}(M\Phi E_e)=\operatorname{tr}(\Phi E_eM)=\lambda_e\operatorname{tr}(\Phi E_e),\qquad \operatorname{tr}(\Phi ME_e)=\operatorname{tr}(\Phi\lambda_eE_e)=\lambda_e\operatorname{tr}(\Phi E_e).$$

此外，由于交叉项在投影算子分解下消失，$\operatorname{tr}(\Phi E_e)=\operatorname{tr}(E_e\Phi E_e)$。因此

$$\operatorname{tr}(\dot{M}E_e)=-2\lambda_e\operatorname{tr}(E_e\Phi E_e),$$

再除以 $d_e=\dim(E_e)$，即可得到所述结论。 □

定理 21（谱质量传输（推论 9））。 定义

$$q_{i,e}(t) := (E_e(t))_{ii} = \lambda_e^+ \, p_{i,e}(t) \, M_{ii}(t),$$

可解释为支撑在特征子空间 e 中的坐标/特征 i 的“质量”。于是

$$\dot{q}_{i,e} = \sum_{f \neq e} T^{(i)}_{e \to f}, \qquad T^{(i)}_{e \to f} := 2\Big(\frac{\lambda_e + \lambda_f}{\lambda_e - \lambda_f}\Big) \, (E_e \Phi E_f)_{ii}.$$

证明。 恒等式 $q_{i,e} = (E_e)_{ii} = \lambda_e^+ \, p_{i,e} \, M_{ii}$ 由引理 B.11(2) 及 $M_{ii} = \|W_i\|^2$ 给出。对 $q_{i,e} = (E_e)_{ii}$ 求导:

$$\dot{q}_{i,e} = (\dot{E}_e)_{ii}.$$

代入推论 7。利用 Φ 和 E_e 的对称性，注意对于 $f \neq e$,

$$(E_f \Phi E_e)_{ii} = e_i^\top E_f \Phi E_e e_i = (E_f e_i)^\top \Phi (E_e e_i) = (E_e e_i)^\top \Phi (E_f e_i) = (E_e \Phi E_f)_{ii}.$$

因此，配对项相加，得到

$$\dot{q}_{i,e} = \sum_{f \neq e} 2\Big(\frac{\lambda_e + \lambda_f}{\lambda_e - \lambda_f}\Big) \, (E_e \Phi E_f)_{ii}.$$

按陈述中的定义令 $T^{(i)}_{e \to f}$，即可得到传输形式。 □

令 $\Gamma \leq S_{n_i}$ 作用于块 Ω_i 的指标，并令 $\{P_\gamma\}_{\gamma \in \Gamma}$ 为相应的置换矩阵。定义交换子（中心化子）

$$\mathcal{A}_\Gamma := \{A \in \mathbb{R}^{n_i \times n_i} : \; P_\gamma A P_\gamma^\top = A \; \forall \gamma \in \Gamma\}.$$

等价地， $\mathcal{A}_\Gamma = \{A : \quad P_\gamma A = A P_\gamma \; \forall \gamma\}$。

[梯度在 Γ-共轭作用下的等变性]

引理 F.2. 假设 $L : \quad \mathbb{R}^{n_i \times n_i} \to \mathbb{R}$ 是 *Fr´echet* 可微的（关于弗罗贝尼乌斯内积），并且满足

$$L(P_\gamma M P_\gamma^\top) = L(M) \qquad \forall \gamma \in \Gamma, \; \forall M.$$

那么梯度是 Γ-等变的:

$$\nabla L(P_\gamma M P_\gamma^\top) = P_\gamma \big(\nabla L(M)\big) P_\gamma^\top \qquad \forall \gamma \in \Gamma, \; \forall M.$$

证明。 固定 $\gamma \in \Gamma$ 并定义 $\Psi(M) := P_\gamma M P_\gamma^\top$。不变性假设为 $L(\Psi(M)) = L(M)$，对所有 M 均成立。在 M 处沿任意方向 H 求导:

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} L(\Psi(M + \varepsilon H)) = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} L(M + \varepsilon H).$$

由于 $\Psi(M + \varepsilon H) = \Psi(M) + \varepsilon(P_\gamma H P_\gamma^\top)$，根据梯度的定义可得

$$\big\langle \nabla L(\Psi(M)), \, P_\gamma H P_\gamma^\top \big\rangle_F = \big\langle \nabla L(M), \, H \big\rangle_F.$$

利用正交共轭下的弗罗贝尼乌斯不变性， $\langle A, \quad P_\gamma H P_\gamma^\top \rangle_F = \langle P_\gamma^\top A P_\gamma, \quad H \rangle_F$，我们得到

$$\big\langle P_\gamma^\top \nabla L(\Psi(M)) P_\gamma, \, H \big\rangle_F = \big\langle \nabla L(M), \, H \big\rangle_F \qquad \forall H.$$

因此 $P_\gamma^\top \nabla L(P_\gamma M P_\gamma^\top) P_\gamma = \nabla L(M)$，这等价于该主张。 □

[$\mathcal{A}_\Gamma$-对（对称化）梯度的成员资格]

Lemma F.3. 假设引理 *F.3*的假设成立。若 $M \in \mathcal{A}_\Gamma$，则 $\nabla L(M) \in \mathcal{A}_\Gamma$。因此 y,

$$\Phi(M) := \mathrm{Sym}(\nabla L(M)) := \tfrac{1}{2}\big(\nabla L(M) + \nabla L(M)^\top\big) \in \mathcal{A}_\Gamma.$$

证明。若 $M \in \mathcal{A}_\Gamma$，则 $P_\gamma M P_\gamma^\top = M$对所有 γ。引理 F.3给出

$$\nabla L(M) = \nabla L(P_\gamma M P_\gamma^\top) = P_\gamma(\nabla L(M))P_\gamma^\top,$$

因此 $\nabla L(M) \in \mathcal{A}_\Gamma$。由于 $P_\gamma A P_\gamma^\top = A$蕴含 $P_\gamma A^\top P_\gamma^\top = A^\top$，我们也有 $\nabla L(M)^\top \in \mathcal{A}_\Gamma$。由于 $\mathcal{A}_\Gamma$是一个线性子空间，它们的平均值 Sym($\nabla L(M)$) 属于 $\mathcal{A}_\Gamma$。 □

$\mathcal{A}_\Gamma$ $\mathcal{A}_\Gamma$ 的代数闭性是一个矩阵代数：若 X，$Y \in \mathcal{A}_\Gamma$ 则对于所有 $\alpha \in \mathbb{R}$，有 $X + Y \in \mathcal{A}_\Gamma, \alpha X \in \mathcal{A}_\Gamma$，并且 $XY \in \mathcal{A}_\Gamma$。此外，若 $X \in \mathcal{A}_\Gamma$，则 $X^\top \in \mathcal{A}_\Gamma$，因此 Sym($X$) $\in \mathcal{A}_\Gamma$。

证明。根据定义关系 $P_\gamma(\cdot)P_\gamma^\top = (\cdot)$，线性性显然成立。对于乘积，若 X，$Y \in \mathcal{A}_\Gamma$，则对每个$\gamma \in \Gamma$,

$$P_\gamma(XY)P_\gamma^\top = (P_\gamma X P_\gamma^\top)(P_\gamma Y P_\gamma^\top) = XY,$$

因此 $XY \in \mathcal{A}_\Gamma$。对于转置，若 $P_\gamma X P_\gamma^\top = X$，则取转置得 $P_\gamma X^\top P_\gamma^\top = X^\top$（因为 $P_\gamma^\top = P_\gamma^{-1}$），所以 $X^\top \in \mathcal{A}_\Gamma$。最后，Sym($X$) $= \frac{1}{2}(X + X^\top)$ 由线性性属于 $\mathcal{A}_\Gamma$。 □

引理 F.4（Gram 流的不变块闭包）。假设对所有 t 有 $\Phi(t) \in \mathcal{A}_\Gamma$，并考虑 *Gram* 流

$$\dot{M}(t) = -(M(t)\Phi(t) + \Phi(t)M(t)).$$

若在某个时刻 t_0 有 $M(t_0) \in \mathcal{A}_\Gamma$，则对所有 $t \geq t_0$ 均有 $M(t) \in \mathcal{A}_\Gamma$。

证明。由于 $\dot{M}\ t_0\ \Phi\ t_0 \in \mathcal{A}_\Gamma$ $\mathcal{A}_\Gamma$ 3 $\dot{M}\ t_0\ \Phi\ t_0 \in \mathcal{A}_\Gamma$ $\Phi\ t_0\ M\ t_0 \in (\ )\ (\ )$ 且其为代数（引理 F.），我们有 ( ) ( ) 和 ( ) ( ) $\mathcal{A}_\Gamma$，因此 $M(t_0) \in \mathcal{A}_\Gamma$。更一般地，只要 $M(t) \in \mathcal{A}_\Gamma$，根据同样的闭包性，我们就得到 $M(t) \in \mathcal{A}_\Gamma$，所以该向量场处处与线性子空间 $\mathcal{A}_\Gamma$ 相切于 $\mathcal{A}_\Gamma$上。因此，从 $M(t_0) \in \mathcal{A}_\Gamma$出发的解对所有 $t \geq t_0$均保持在 $\mathcal{A}_\Gamma$ 中。 □

引理 F.5（ Gram 流的不变块闭包）。假设 $\Phi(t) \in \mathcal{A}_\Gamma$ 对所有 t 均成立，并考虑 *Gram* 流

$$\dot{M}(t) = -(M(t)\Phi(t) + \Phi(t)M(t)).$$

若在某个时刻 t_0 有 $M(t_0) \in \mathcal{A}_\Gamma$，则对所有 $t \geq t_0$ 均有 $M(t) \in \mathcal{A}_\Gamma$。

证明。任取 $\gamma \in \Gamma$，并定义 $\widetilde{M}(t) := P_\gamma M(t) P_\gamma^\top$。于是

$$\dot{\widetilde{M}}(t) = P_\gamma \dot{M}(t) P_\gamma^\top = -\big(\widetilde{M}(t)\,\widetilde{\Phi}(t) + \widetilde{\Phi}(t)\,\widetilde{M}(t)\big),$$

其中 $\widetilde{\Phi}(t) := P_\gamma \Phi(t) P_\gamma^\top$。由于 $\Phi(t) \in \mathcal{A}_\Gamma$，对所有 t，我们有 $\widetilde{\Phi}(t) = \Phi(t)$，因此

$$\dot{\widetilde{M}}(t) = -(\widetilde{M}(t)\Phi(t) + \Phi(t)\widetilde{M}(t)).$$

此外，由于 $M(t_0) \in \mathcal{A}_\Gamma$，我们有 $\widetilde{M}(t_0) = P_\gamma M(t_0) P_\gamma^\top = M(t_0)$。因此，$M$和 $\widetilde{M}$满足相同的常微分方程$\widetilde{\ }$，且在 t_0处具有相同的初始条件。根据线性矩阵常微分方程解的唯一性，对于所有 $t \geq t_0$，均有 $M(t) = M(t)$。因此，对于所有 $\gamma \in \Gamma$和所有 $t \geq t_0$，均有 $P_\gamma M(t) P_\gamma^\top = M(t)$，即对于所有 $t \geq t_0$，均有 $M(t) \in \mathcal{A}_\Gamma$。 □

定理 22 (容量饱和的稳定性（推论 10）**)**。 若一个配置的梯度核 Φ 与每个格拉姆谱投影算子都可交换，则该配置是谱不动点：

$$[E_e, \Phi] = 0 \quad \forall e.$$

特别地，具有均匀稀疏性的紧框架几何满足此条件（因此是谱不动点）。

证明。 若对所有 e 都有 $[E_e, \Phi] = 0$，则对于 $f \neq e$,

$$E_e \Phi E_f = E_e E_f \Phi = 0,$$

，因为对易性给出 $E_e \Phi = \Phi E_e$，而正交性给出 $E_e E_f = 0$。代入推论 7，得到对所有 e 都有 $E_e = 0$。因此，推论 9 给出 $\dot{q}_{i,\ e} = 0$ 对所有 $i,\ e$均成立：特征子空间之间不存在谱质量传输，因此谱分解在动力学上稳定。

关于紧框架／均匀稀疏性的论断：附录 E.2（引理 E.1）在同一个关联方案（轨道）基 $\{A_r\}$中展开了 $M(t)$ 和 $\Phi(t)$:

$$M(t) = \sum_{r=0}^{R} \theta_r(t) A_r, \qquad \Phi(t) = \sum_{r=0}^{R} \phi_r(t) A_r.$$

当中心化子是 Bose–Mesner 代数（交换关联方案）时，所有 A_r 都彼此对易，因此 $M(t)$ 和$\Phi(t)$ 也对易。在一个对称的交换族中， Φ与 M 的谱投影算子 $\{E_e\}$ 对易（同时对角化），所以对所有 e 都有 $[E_e, \Phi] = 0$，即为谱不动点。 □

G. 非均匀稀疏性

我们观察到均匀稀疏性具有相同类型的射影线性:

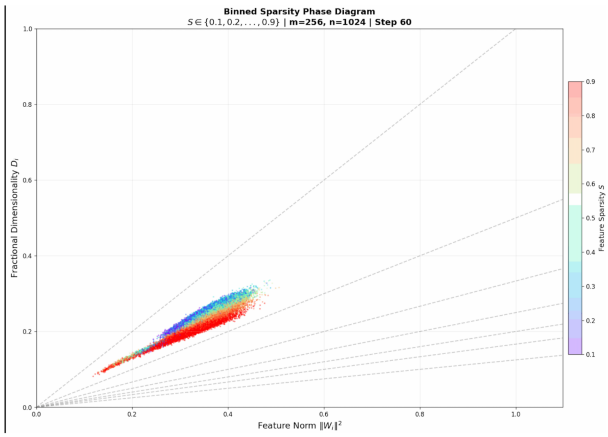


图*8.* 图像1的描述

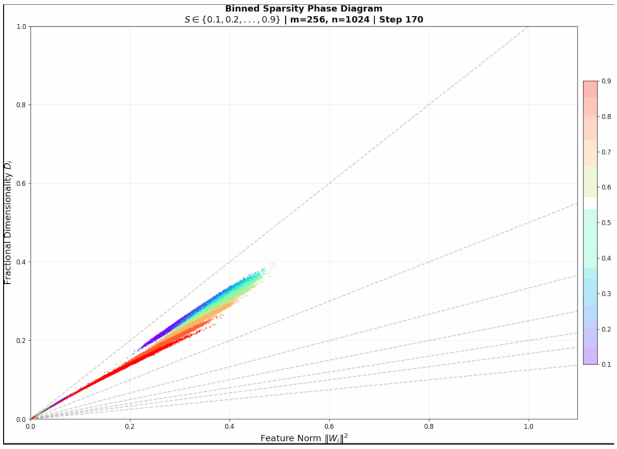


图*9.* 图像2的描述

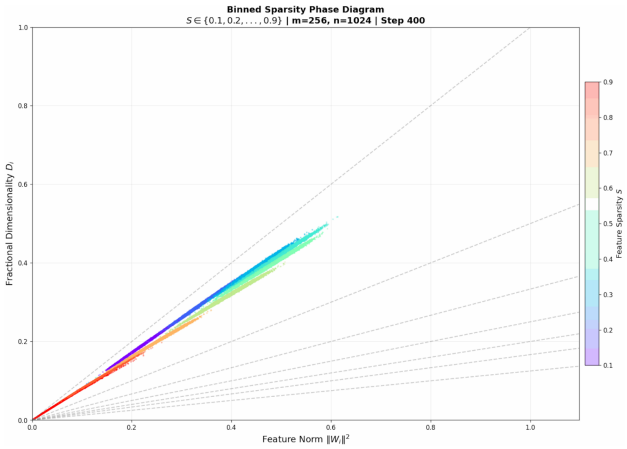


图*10.* 图像3的描述

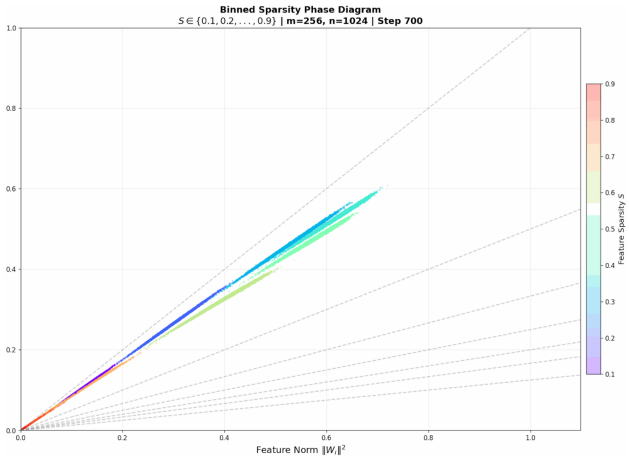


图*11.* 图像4的描述

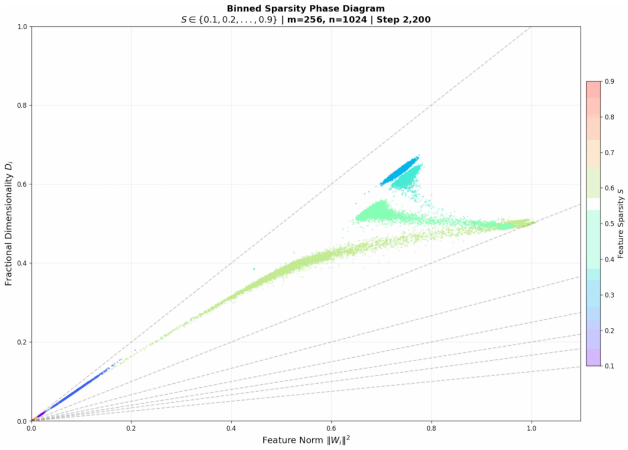


图*12*。 图像5的描述

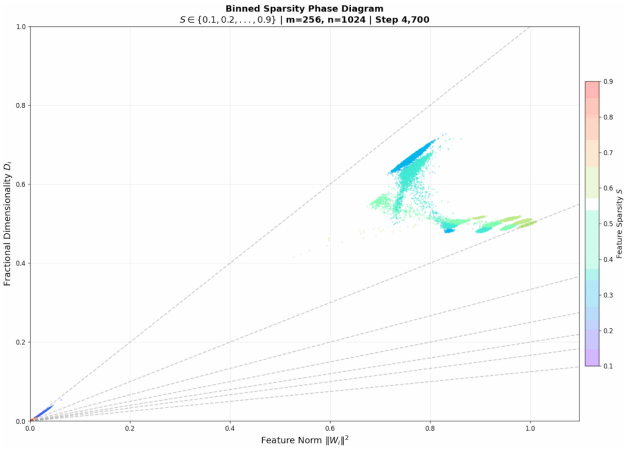


图*13*。 图像6的描述

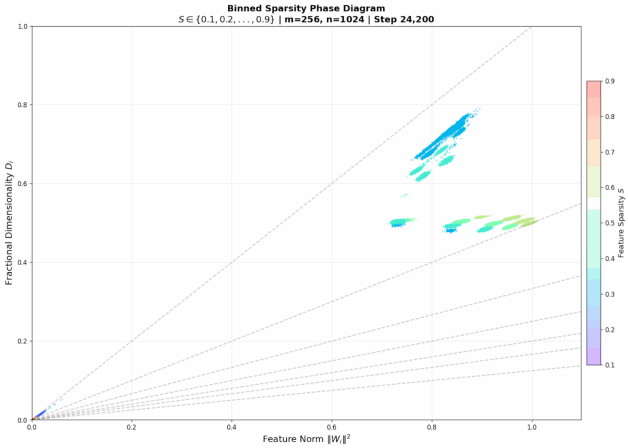


图*14*。 图像7的描述

H. 相关工作

在严谨考察后一种模型之前，我们先介绍多语义性相关研究的背景：

• **Lecomte等人（2023/2024）（Lecomte等人，2023年）偶然多语义性** Lecomte等人认为，即使容量足以支持单语义编码，多语义性也可能偶然地产生。通过理论分析和实验，他们表明，随机初始化、正则化和神经噪声可能在训练早期使多个特征偶然地对齐到同一个神经元上，之后梯度动力学会进一步强化这种重叠。这一发现补充了容量驱动的叠加理论：多语义性可以成为一种稳定的吸引子，但并非实现良好性能的绝对必要条件。

• **Adler与Shavit（2024/2025）（Adler与Shavit，2024年）——叠加计算的复杂性** Adler与Shavit为使用叠加特征进行计算建立了复杂性理论界，证明一大类任务（例如置换、成对逻辑）至少需要$\Omega(\sqrt{m'}\log m')$ 个神经元和$\Omega(m'\log m')$ 个参数，才能在叠加状态下计算m' 个输出。他们还给出了几乎匹配的构造性上界（例如，使用$O(\sqrt{m'}\log m')$个神经元实现成对与运算），揭示了效率提升幅度虽大，却并非没有上限。这些结果区分了在叠加中“表示”特征（成本低得多）和利用它们进行“计算”（可证明成本更高），为任何可解释性或压缩方法都必须遵守的原则性界限奠定了基础。

• **Klindt等人（2025年）（Klindt等人，2025年）——从叠加到稀疏编码**Klindt等人提出了一条有原则的方法路径，通过利用三个组成部分，将叠加激活转化为可解释因子：可识别性（分类表示可以通过可逆线性变换恢复潜在特征）、稀疏编码/字典学习（用于寻找概念对齐基），以及定量可解释性指标。从这一视角来看，深度网络通常会以线性方式叠加概念；事后进行稀疏编码可以逆转这种混合，在无需重新训练模型的情况下得到单语义方向。

• **Hollard等人（2025年）（Hollard等人，2025年）低参数视觉模型中的叠加**Hollard等人研究了现代 <! 1。5M参数卷积神经网络，并表明瓶颈设计和超线性激活会加剧特征图中的干扰（特征重叠），从而限制准确率缩放。通过系统地改变瓶颈结构，他们找出了能够减少干扰的设计选择，并提出了“NoDepth瓶颈”，使模型在严格的参数预算内提升了在ImageNet上的缩放表现。

• **Pertl等人（2025年）（Pertl等人，2025年）——图神经网络中的叠加** Pertl等人通过提取节点级和图级的特征方向，并研究其基不变重叠，分析了图神经网络中的叠加现象。他们观察到重叠呈现宽度驱动的相模式、拓扑引发的节点级混合，以及通过更尖锐的池化来缓解这种现象并增强轴对齐；浅层模型可能会陷入亚稳态低秩嵌入。

• **Hesse 等人（2025年）（Hesse 等人，2025年）解缠 CNN 中的多义性通道** Hesse 等人提出了一种算法手术，通过利用输入混合单元的不同上游激活模式，将一个多义性 CNN 通道拆分为多个通道，使每个通道仅响应单一概念。该方法无需重新训练即可重连预训练网络，从而生成明确的单义性通道，提升特征可视化的清晰度，并支持使用标准机制工具。不同于“虚拟”分解，这种方法能够生成与概念对齐的具体网络组件，展示了事后减少叠加的一条实用途径。

• **Dreyer 等人（2024 年）（Dreyer 等人，2024年）PURE：基于电路的分解** Dreyer 等人提出了 PURE，这是一种事后电路方法，通过识别负责各个概念的不同上游子图（电路），将多义性神经元分解为多个虚拟的单义性单元。该方法改善了概念级可视化效果（例如通过基于 CLIP 的评估），且不会修改权重，而是将行为重新归因到彼此不相交的计算路径上。PURE 表明，多语义性在很大程度上源于电路叠加；分离电路可以在不改变架构的情况下，得到更纯粹的概念单元。

I. 图表

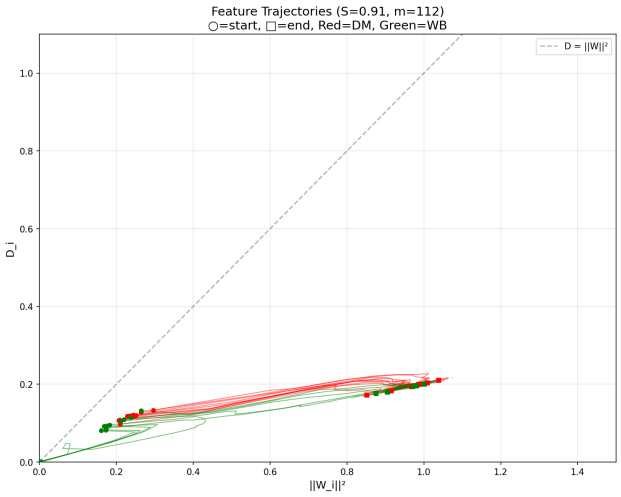


图*15.* 图像5的描述

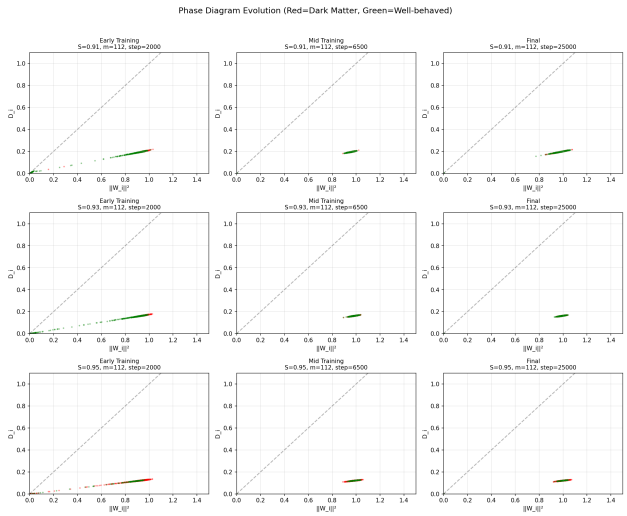


图*16.* 图像6的描述

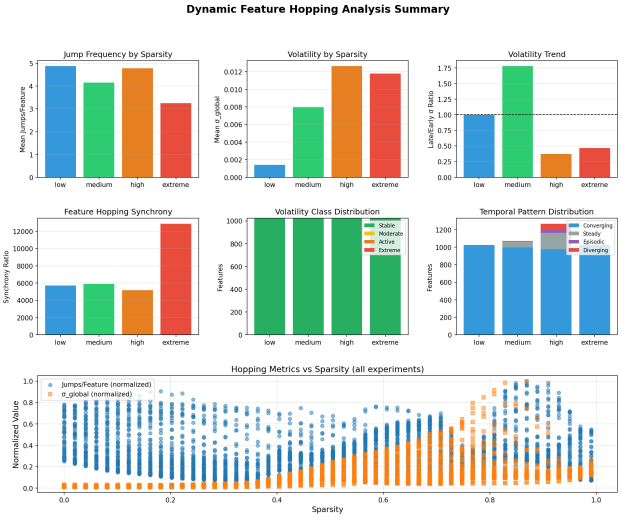


图*17.* 图像5的描述

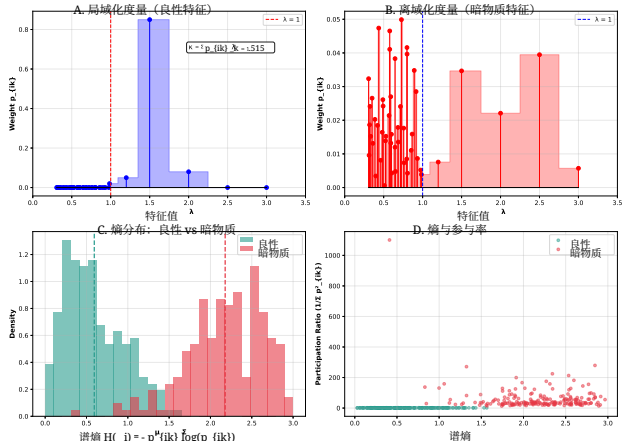


图*18.* 图像6的描述

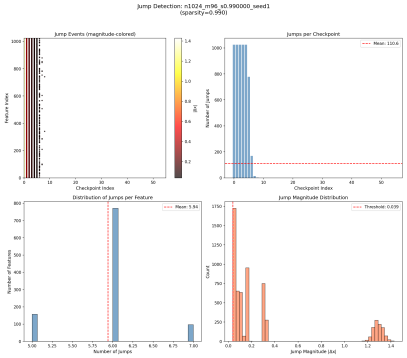


图*19.* 跳跃 m96

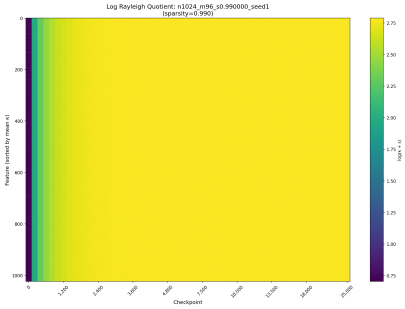


图*20.* Rayleigh m96

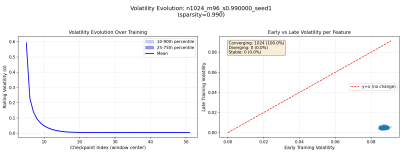


图*21.* 波动率 m96

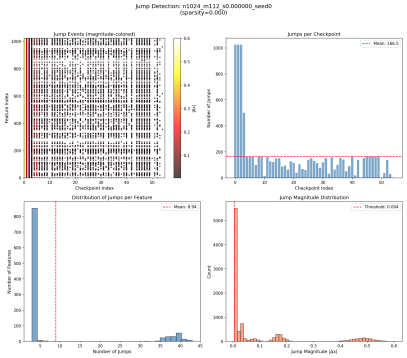


图*22.* 跳跃 m112-0

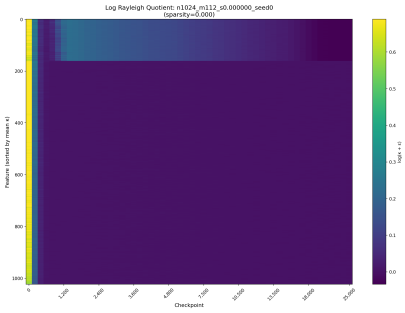


图*23.* Rayleigh m112-0

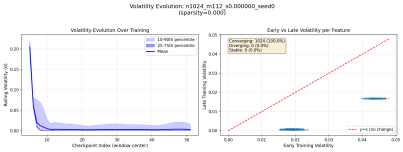


图*24.* 波动率 m112-0

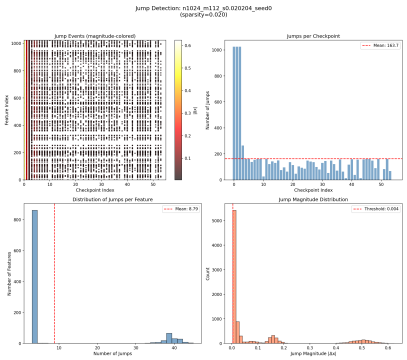


图 *25.* 跳跃 m112-2

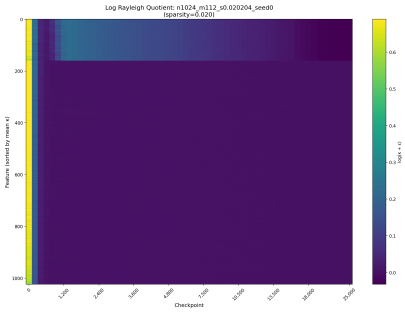


图 *26.* Rayleigh m112-2

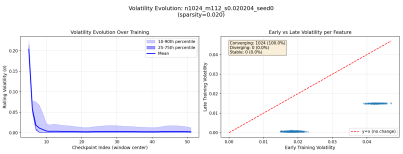


图 *27.* 波动率 m112-2

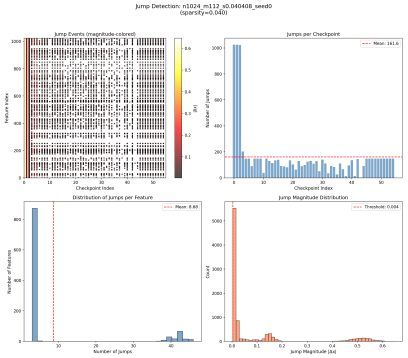


图 *28.* 跳跃 m112-4

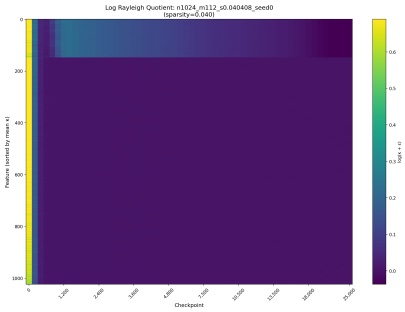


图 *29.* Rayleigh m112-4

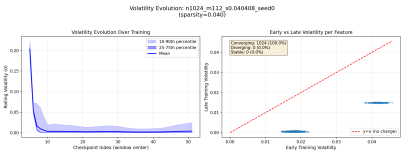


图 *30.* 波动率 m112-4

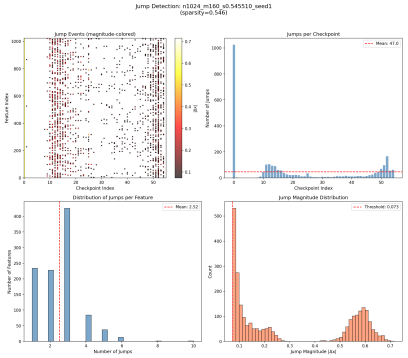


图 *31.* 跳跃 m160

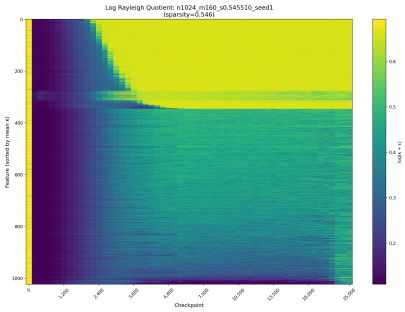


图 *32.* Rayleigh m160

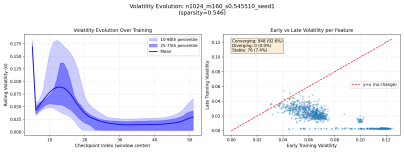


图 *33.* 波动率 m160

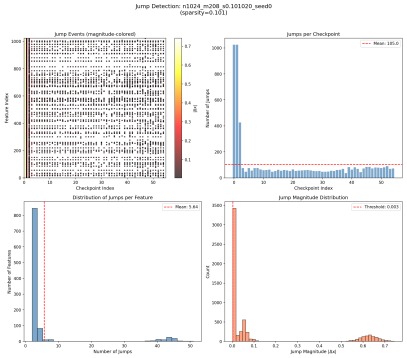


图 *34.* 跳跃 m208

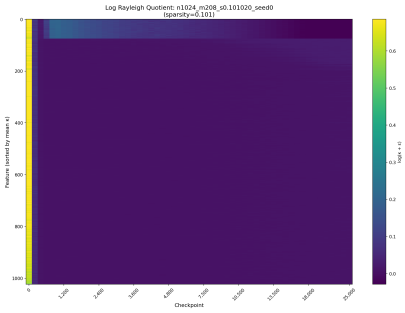


图 *35.* Rayleigh m208

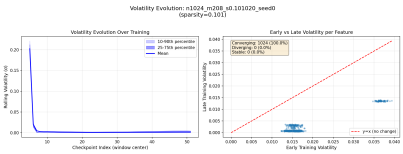


图 *36.* 波动率 m208

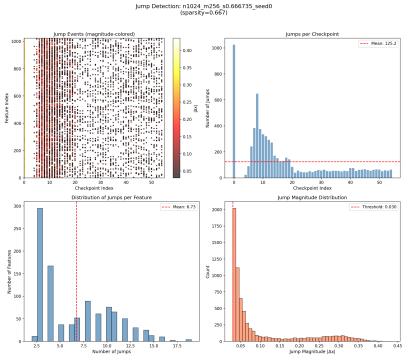


图 *37.* 跳跃 m256

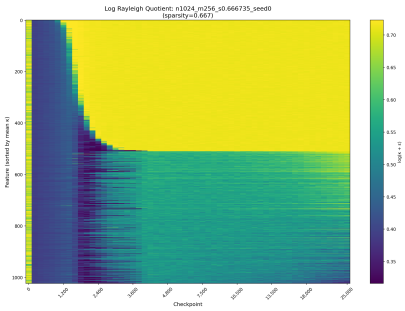


图 *38.* Rayleigh m256

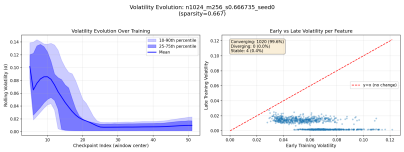


图 *39.* 波动率 m256