arXiv:2601.00276v1 [cs.LG] 1 Jan 2026

任务驱动的核流：标签秩压缩与拉普拉斯谱滤波

Hongxi Li计算机科学与工程学院 中山大学
lihx228@mail2.sysu.edu.cn

Chunlin Huang博雅初级中学 广西希望高中
7wgn0326@gmail.com

摘要

我们针对具有线性读出层和 ℓ_2正则化的宽神经网络中的任务驱动特征学习，构建了一套以核为中心的理论。我们的分析分为两个阶段，以弥合可解释动力学与结构保证之间的鸿沟。首先，在采用平方损失的快速读出（绝热）机制下，我们推导出一个闭式核常微分方程，其演化由任务依赖的驱动算子与各向同性正则化衰减之间的竞争所支配。这揭示了对齐的精确机制：对于具有 C输出的监督学习，驱动算子是一个秩-C 算子，会将核压缩到低维子空间中，并遵循明确的“注水”谱定律。

其次，我们证明，由此产生的结构现象并非时间尺度分离的产物。在标准 ℓ_2正则化和 C维线性读出层下，耦合特征—读出动力学的任何稳定稳态都必然表现出标签驱动的秩压缩（$\mathrm{rank}(K_\infty) \le C$），并且对于平方损失满足相同的谱截断定律。这些结果是架构和损失的代数推论，并不依赖于快速读出近似。

作为对这一确定性图景的补充，我们在核层面分析随机梯度下降噪声，并证明，对于具有 C 个输出的任意凸损失，瞬时噪声协方差被限制在一个维度至多为 $O(C)$ 的低维子空间中，且与网络宽度和当前参数值无关。因此，随机性会在任务相关子空间内引发受限扩散，而非各向同性探索。

我们还沿两个理想化方向拓展了这一框架：（i）总体极限，在此情形下，我们将核积分算子的谱演化与偏差–方差权衡联系起来；（ii）一种由图拉普拉斯算子和对数行列式排斥驱动的风格化自监督核模型，该模型会产生高秩且与拉普拉斯矩阵对齐的表征。综合来看，这些结果提供了一种统一的谱语言，用以对比监督学习的压缩特性与自监督的扩展行为，同时阐明哪些方面是严谨的架构性结果，哪些则源于特定的核模型。

1 引言

深度学习的实证成功，很大程度上归功于其学习依赖于数据的表征的能力；这些表征特征能够适应底层任务结构。经典学习理论和神经切线核（NTK）机制 [1, 7]通常在无限宽极限下描述网络，此时特征保持静态（即“惰性训练”机制） [5]。这种视角虽在理论上便于处理，却无法捕捉实践中观察到的丰富特征学习动力学；在实际情形下，核会显著演化并与目标函数对齐 [2, 11]。理解支配这种核演化的机制，是深度学习理论面临的一项核心挑战。

在这项工作中，我们提出一个以核为中心的框架，用于分析具有 a C 维线性读出层和显式 ℓ_2 正则化的宽神经网络中的特征学习。我们不再

追踪高维参数轨迹，而是聚焦于经验核矩阵 $K(t) \in \mathbb{R}^{N\times N}$ 的动态变化，它编码了数据表示之间的成对相似度。我们采用双视角方法来剖析学习过程：

1. 通过绝热近似研究动力学。为了深入分析学习轨迹，我们首先采用“快—慢”机制，即线性读出层的演化速度显著快于特征。这样便可以将读出层“积分消去”，从而在理想化的平方损失设定下，得到关于核的封闭形式常微分方程（ODE）。该方程揭示了其中起作用的物理力：促进对齐的任务特定驱动力，以及导致衰减的正则化项。

2. 超越时间尺度分离的结构性质。接下来，我们证明该常微分方程所揭示的关键结构性结论并非快—慢近似造成的假象。在标准 ℓ_2正则化和 C维线性读出层条件下，我们证明**稳态拓扑**（标签驱动的秩压缩，以及平方损失下的谱截断定律）和**随机梯度下降噪声的几何结构**均由损失景观和网络架构所决定。关于秩压缩的研究结果为广泛观察到的神经坍缩 [6, 9]，这一现象提供了理论依据：训练结束时，类内变异趋于消失。

因此，我们的框架既能从机制上描述特征如何随时间对齐（通过快速读出机制下的核常微分方程），也能严格保证它们在稳态时可能收敛到什么状态（通过不依赖时间尺度分离的不动点分析）。我们进一步沿两个理想化方向扩展了这一分析：（i）扩展至总体极限，通过核积分算子的演化讨论泛化；（ii）扩展至一种简化的自监督学习（SSL）模型，以凸显监督式压缩与对比式扩张或基于冗余消减的扩张之间的谱对比 [4, 8]。

贡献。 在这些建模假设下，我们的主要贡献如下：

- **特征学习动力学的推导**。 在快速读出（绝热）机制下，我们为具有线性读出层和 ℓ_2-正则化的宽网络推导出显式的核常微分方程；对于任意凸损失，该方程在驱动项层面均成立。它将特征学习刻画为任务对齐力与正则化诱导的衰减之间的热力学竞争。

- **稳态下的秩压缩与谱截断**。 我们证明，对于C类监督学习，若采用 C维线性头和 ℓ_2正则化，则任何稳定的稳态核的秩至多为 C，无论读出层与特征的学习率相对大小如何。对于平方损失，我们推导出一条明确的谱截断（“注水”）规律，与深度网络中观察到的谱偏置一致 [10]。重要的是，这些稳态性质适用于一般的耦合动力学，并不局限于推导常微分方程时所用的时间尺度分离假设。

- **内在低秩 SGD 噪声**。 我们在核层面分析随机梯度，并证明对于任意具有 C 个输出的凸损失，瞬时噪声协方差矩阵都局限于一个由输出维度 C决定的低秩子空间。这一架构约束充当内置滤波器，使 SGD 噪声被限制在任务相关子空间内（因而在其中扩散），而不会激发任意的核方向。

- **总体极限与泛化**。 我们将框架扩展到无限样本极限，定义相应核积分算子的演化，并说明谱截断机制如何直接影响未见数据上的偏差–方差权衡，这与 NTK 等固定核机制形成对比。

- **监督学习与自监督学习的谱统一**。 我们分析了一种由图拉普拉斯算子 [3]和对数行列式排斥驱动的简化 SSL 核模型。在该模型中，学习到的核具有高秩、与拉普拉斯对齐的谱，且

形状显式且（ν_i +恒定$^{-1}$），从而产生“白化”表征 [12]。这提供了一种统一的谱语言，用以对比监督的压缩性、低秩特征与自监督的扩展性、高秩特征。

假设与研究范围。 本文始终采用一种标准但理想化的设定：

(i)带有 ℓ_2正则化的线性读出。 网络输出由特征之上的一个 C 维线性头生成。我们对读出层以及（在自由特征模型中）骨干网络特征施加显式的 ℓ_2 正则化。

(ii) 特征学习／宽骨干网络机制。 我们假设骨干网络具有足够的表达能力，因此可以直接在经验特征矩阵 Φ 及其核 $K = \Phi^\top\Phi$的层面上对表示动力学建模，而无需额外的架构约束；神经正切核／惰性区间并非本文的研究重点。

(iii)用于推导闭式谱定律的平方损失。 本文针对平方损失推导闭式核常微分方程和注水式谱截断定律，由此可在标签格拉姆矩阵的特征基下得到自治动力学。对于更一般的凸损失（例如交叉熵），同样的秩压缩机制在稳态时仍然适用，但我们不声称能够得到闭式谱轨迹。

(iv) 风格化的自监督学习目标。 对于自监督和半监督设置，我们研究一种基于图拉普拉斯算子和对数行列式排斥的风格化核目标函数。该目标旨在作为一种典型的谱模型，从解析层面清晰呈现监督式压缩与自监督学习扩张之间的对比，而非对 SimCLR 或 BYOL 等任何特定方法的精确推导。

在这一设置下，我们的严格结构性结果——例如标签驱动的秩压缩 rank(K_∞) $\le C$、权重衰减与端到端映射上的核范数偏置之间的等价性，以及随机梯度下降噪声的低秩结构——都是C维输出瓶颈和 ℓ_2正则化的代数推论。相比之下，监督情形中的精确注水式谱，以及我们的自监督学习模型中的（ν_i +常数）$^{-1}$ 谱形状，应被视为这些理想化核流所呈现的行为，而非对所有采用交叉熵、批归一化或注意力的实际训练设置的字面描述。

2 模型与双时间尺度动力学

2.1 直觉：快速读出假说

深度神经网络在结构上可分解为两部分：非线性特征提取器 $\varphi(\cdot)$，将输入映射到高维嵌入空间；以及线性读出头 W，将嵌入映射为预测。这两个组件的动力学通常存在根本差异。

为了建立直觉，考虑这样一个简化场景：特征提取器被冻结（即 Φ 为常量）。在这种情况下，优化网络就相当于在固定特征上训练线性模型（例如线性回归或逻辑回归）。由于损失函数通常关于 W 是凸的（在加入 ℓ_2 正则化时为严格凸）， W 的梯度动力学很简单： W 会以指数速度收敛到唯一最优解，记为 $W^*(\Phi)$。

当然，在现实中， Φ 会与 W 一同演化。然而，随着训练的推进，我们常常观察到，深层表征的稳定速度远慢于顶层线性层。在迁移学习等机制中，或采用特定的学习率调度（较大的 η_W、较小的 η_Φ）时，这种分离现象会更加明显。

基于这一观察，我们采用时间尺度分离 假设。我们假设读出动力学相较于特征动力学足够快，因此 W 实际上

瞬间达到平衡。

$$W(t) \approx W^*(\Phi(t)) \quad \text{for all } t.$$

这一“绝热”近似使我们能够从运动方程中消去 W。我们不再追踪耦合系统（Φ, W），而是可以完全专注于由最优 读出层驱动的特征有效动力学。虽然这是一个大胆的简化，但它抓住了反馈回路的本质：特征不断演化以最小化损失，并假设分类器总能充分利用这些特征。

从数学上看，这将原始目标函数 $\mathcal{L}(\Phi, W)$ 化简为一个有效泛函 $\widetilde{\mathcal{L}}(\Phi)$，它仅取决于核，我们将在下文对其进行形式化。

2.2 数学表述

我们考虑一个监督学习任务，其数据集包含 N 个样本 $X = [x_1, \ldots, x_N] \in \mathbb{R}^{d_{in} \times N}$及对应的目标值 $Y = [y_1, \ldots, y_N]^\top \in \mathbb{R}^{N \times C}$，其中 C为类别数（或输出维度）。

将神经网络建模为特征映射 $\Phi(\cdot)$ 与线性读出层$W \in \mathbb{R}^{C \times k}$的复合。令 $\Phi \in \mathbb{R}^{k \times N}$ 表示整体特征矩阵，其中第 i 列为$\phi_i = \Phi(x_i)$。网络输出为 $\hat{Y} = (W\Phi)^\top = \Phi^\top W^\top \in \mathbb{R}^{N \times C}$。我们在正则化经验风险最小化框架下研究训练动力学。总目标函数 $\mathcal{J}(W, \Phi)$ 定义如下：

$$\mathcal{J}(W, \Phi) = \mathcal{L}(\hat{Y}, Y) + \frac{\lambda}{2}\|W\|_F^2 + \frac{\mu}{2}\|\Phi\|_F^2, \tag{1}$$

其中，$\mathcal{L}$ 是一个凸损失函数（例如平方误差或交叉熵），$\lambda > 0$ 是读出层的正则化系数，而 $\mu \geq 0$ 表示特征提取器的权重衰减。

我们的主要研究对象是**经验核矩阵**（或格拉姆矩阵）$K \in \mathbb{R}^{N \times N}$，其定义为特征的内积：

$$K(t) = \Phi(t)^\top \Phi(t). \tag{2}$$

核 K 刻画了数据表征的几何结构。值得注意的是，尽管 Φ的参数空间可能非常广阔（在宽极限下甚至可能是无限的），但有限数据集上的学习动态完全由该 $N \times N$ 矩阵的演化所决定。

3 核常微分方程的推导

本节将在快速读出假设下，推导支配 K（t）演化的精确微分方程。

3.1 快速读出极限

遵循第 2.1 节中的直观理解，我们假设读出层 W 以足够快的时间尺度演化，从而能够有效地最小化目标函数 $\mathcal{J}$ 在当前固定特征下 Φ 的每一时刻 t。我们将最优读出层定义为 W^*（Φ）如下：

$$W^*(\Phi) = \operatorname*{argmin}_W \left(\mathcal{L}((W\Phi)^\top, Y) + \frac{\lambda}{2}\|W\|_F^2 \right). \tag{3}$$

代入 W^* 方程（1）可得有效特征损失 $\mathcal{L}$（Φ） $= \mathcal{J}$（W^*（Φ） Φ）。因此，特征学习被建模为该有效景观上的梯度流：

$$\dot{\Phi} = -\nabla_\Phi \widetilde{\mathcal{L}}(\Phi). \tag{4}$$

3.2 一般核动力学

我们首先推导一个适用于任意凸损失函数 $\mathcal{L}$ 的一般演化方程。根据包络定理，有效损失关于 Φ 的全导数，恰好等于在最优点 W^* 处计算的联合损失的偏导数：

$$\nabla_\Phi \widetilde{\mathcal{L}}(\Phi) = \nabla_\Phi \mathcal{J}(W, \Phi)\Big|_{W=W^*}. \tag{5}$$

对方程 (1) 应用链式法则，可得：

$$\nabla_\Phi \mathcal{J} = W^\top \frac{\partial \mathcal{L}}{\partial \hat{Y}^\top} + \mu\Phi = -W^\top R^\top + \mu\Phi, \tag{6}$$

其中，我们将 **广义残差矩阵** $R \in \mathbb{R}^{N\times C}$ 定义为ˆ损失关于预测的负梯度： $R: = -\nabla_{\hat{Y}}\mathcal{L}$。对于平方损失， $R = Y - Y$。将其代入方程 (4)，特征动力学变为：

$$\dot{\Phi} = W^{*\top} R^\top - \mu\Phi. \tag{7}$$

该方程揭示，特征在两种力的作用下演化：驱动力 促使特征与反向传播的残差信号 ($W^{*\top}R^\top$) 对齐，而衰减力 ($-\mu\Phi$) 则由正则化引起。

现在，我们计算核 $K = \Phi^\top\Phi$ 的时间导数：

$$\dot{K} = \dot{\Phi}^\top\Phi + \Phi^\top\dot{\Phi}. \tag{8}$$

代入方程 (7)，并注意到ˆ$Y = \Phi^\top W^{*\top}$：

$$\begin{aligned}\dot{K} &= (RW^*\Phi - \mu\Phi^\top\Phi) + (\Phi^\top W^{*\top}R^\top - \mu\Phi^\top\Phi) \\ &= R\hat{Y}^\top + \hat{Y}R^\top - 2\mu K.\end{aligned} \tag{9}$$

方程（9）是**任务驱动的核常微分方程**。它指出，核的变化率由模型预测 Y 与任务残差 R之间的对齐决定，并受到均匀衰减的抵消。

解读：类赫布反馈。方程（9）具有一种引人注目的物理解释。驱动项 $R\hat{}Y^\top$ 是残差误差 R 与ˆ当前预测 Y的外积。这类似于监督式的赫布学习（“同时放电，连在一起”）：在模型预测与误差信号积极相关的方向上，核强度会增加。相比之下， $-2\mu K$ 项起到均匀遗忘机制的作用。因此，特征学习表现为一种选择过程：网络强化有助于减小误差的方向，同时使无关成分衰减。

3.3 平方损失的闭式动力学

ˆ 为进行谱分析，我们将讨论限定在均方误差（MSE）的情形， $\mathcal{L}(Y,\ Y) =\frac{1}{2}\|\hat{}Y - Y\|_F^2$。

–

1. 显式读出。 对于均方误差，最优读出层 W^* 是岭回归ˆ问题的解。利用矩阵求逆引理，预测结果 Y 可以写成与特征维度 k无关的核化形式：

$$\hat{Y} = K(K + \lambda I)^{-1}Y. \tag{10}$$

ˆ2. 显式残差。 因此，残差 $R = Y - Y$变为：

$$R = Y - K(K + \lambda I)^{-1}Y = \lambda(K + \lambda I)^{-1}Y. \tag{11}$$

ˆ**3. 驱动算子。** 将 Y 和 R 代入一般常微分方程（方程（9））后，驱动项 $R\hat{}Y^\top + \hat{}YR^\top$ 变为：

$$\mathcal{D}(K) = \lambda(K+\lambda I)^{-1}YY^\top(K+\lambda I)^{-1}K + \text{h.c.}, \tag{12}$$

其中，h.c. 表示第一项的厄米共轭（转置）。由于 K 和 $(K+\lambda I)^{-1}$可交换，我们可以重新排列各项。令 $M_Y = YY^\top$ 为标签核矩阵。由此得到最终的闭式常微分方程：

$$\dot{K}(t) = \lambda\left[(K+\lambda I)^{-1}M_Y(K+\lambda I)^{-1}K + K(K+\lambda I)^{-1}M_Y(K+\lambda I)^{-1}\right] - 2\mu K. \tag{13}$$

该方程是后续分析的基础。驱动项显式依赖于标签结构 M_Y，其秩至多为 C。这一秩瓶颈正是我们在第 4 节讨论的压缩现象的根源。

3.4 直觉：标量动力学与谱滤波

为揭开方程（13中的矩阵常微分方程的奥秘，我们考虑一个简化的标量情形：数据仅由一个标签为 y、核值为 k（t）$\in \mathbb{R}$的样本构成。该方程简化为：

$$\dot{k} = \frac{2\lambda k}{(\lambda+k)^2}y^2 - 2\mu k. \tag{14}$$

这一标量动力学揭示了一种关键的**信噪比滤波机制**：

- **强化：** 增长项 $\frac{k}{(\lambda+k)^2}$ 并非单调变化。当 $k \to 0$（没有特征）和 $k \to \infty$（饱和）时，该项为零，并在 $k \approx \lambda$时达到峰值。这意味着网络会主动强化“恰到好处”的特征——既不会弱到毫无用处，也不会强到造成不稳定。

- **阈值筛选：** 只有当信号强度（与 y^2 成正比）超过正则化屏障 μ 时，特征才会增长。如果标签信号太弱（$y^2 \lesssim \mu\lambda$），衰减项就会占主导，特征 k 便会塌缩至零。

正如我们将在下一节展示的，这种标量直觉可以推广到谱域：矩阵常微分方程会对核的特征值应用类似的滤波器，有选择地放大与标签信号对齐的特征模态，同时截断其他模态（如图 1 所示）。

4 收敛性分析

在刻画学习到的特征的结构性质之前，我们必须首先证明学习动力学在数学上具有良好的性质。尽管 $K(t)$ 的矩阵微分方程（方程 (13)）是非线性且高维的，但其收敛性质可直接由第 2 节中定义的有效损失函数的构造得出。具体而言，我们证明，在所提出的动力学下，核矩阵 $K(t)$ 全局收敛至唯一的稳态。

4.1 能量景观与李雅普诺夫稳定性

˜稳定性分析的关键在于将有效损失 $\mathcal{L}(\Phi)$ 确定为系统的李雅普诺夫函数。由于系统动力学由梯度流定义，系统自然会趋向于最小化这一能量函数。

定理 1（全局收敛）。 假设正则化系数满足 $\lambda > 0$ 和$\mu > 0$。对于任意初始化 $\Phi(0)$，特征轨迹 $\Phi(t)$ 对所有 $t \geq 0$ 都保持有界。此外，该轨迹收敛至唯一的极限点 Φ_∞，因此，核矩阵收敛至唯一的稳态矩阵 K_∞。

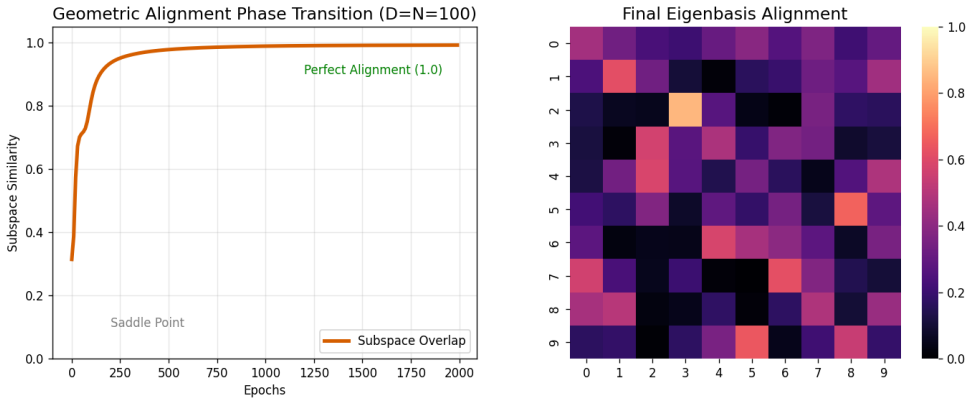


图 1：**几何对齐的动力学**。（左）训练期间子空间重叠分数的演变（含 $D = N = 100$）。系统表现出明显的相变，脱离鞍点，最终与目标子空间实现完美对齐（分数 ≈ 1.0）。（右）最终特征核 K_∞ 条目的热图。棋盘格模式证实了可交换结构 $[K_\infty, M_Y] \approx 0$，验证学习到的特征与标签共享相同的特征基。

证明。 证明分为三个步骤：确立单调性、证明有界性，并运用解析收敛性质。

1. 单调性。 根据定义，动力学遵循梯度流 $\dot{\Phi} = -\nabla_\Phi \mathcal{L}(\Phi)$。沿轨迹的有效损失的时间导数为：

$$\frac{d}{dt}\widetilde{\mathcal{L}}(\Phi(t)) = \left\langle \nabla_\Phi \widetilde{\mathcal{L}}, \dot{\Phi} \right\rangle = -\|\nabla_\Phi \widetilde{\mathcal{L}}\|_F^2 = -\|\dot{\Phi}\|_F^2 \leq 0. \tag{15}$$

因此，除非系统处于满足 $\dot{\Phi} = 0$ 的临界点，否则目标函数沿轨迹严格非增。

2. 有界性（强制性）。 有效损失函数由非负的数据拟合项（凸岭回归问题的最小值）以及特征上的正则化项构成：

$$\widetilde{\mathcal{L}}(\Phi) = \min_W \mathcal{J}(W, \Phi) \geq \frac{\mu}{2}\|\Phi\|_F^2 = \frac{\mu}{2}\operatorname{Tr}(K). \tag{16}$$

由于损失单调不增，因此对所有 $t > 0$，有 $\mathcal{L}(\Phi(t)) \leq \mathcal{L}(\Phi(0)) := \mathcal{L}_0$。这意味着特征的弗罗贝尼乌斯范数有界：

$$\|\Phi(t)\|_F^2 \leq \frac{2\mathcal{L}_0}{\mu}. \tag{17}$$

由于目标函数的次水平集是紧集（由强制性条件 $\mu > 0$ 保证），因此轨迹 $\Phi(t)$ 被限制在参数空间$\mathbb{R}^{k\times N}$中的一个有界区域内。

3. 收敛。 虽然拉萨尔不变性原理等标准结果保证轨迹收敛到临界点的集合，但我们可以给出更强的结论。由于目标函数函数 $\mathcal{L}(\Phi)$是实解析的（它涉及矩阵元素的有理函数和多项式），因此**Łojasiewicz-Simon梯度不等式** 适用。该不等式保证轨迹具有有限长度，并收敛到唯一的临界点 Φ_∞，而非在一系列连续的临界点之间振荡或漂移。因此，极限 $K_\infty = \Phi_\infty^\top \Phi_\infty$是确定且唯一的。

□

4.2 不动点方程

在证明极限存在后，接下来我们推导稳态核$\dot{K}$必须满足的条件。令时间导数 $\dot{K} = 0$ 在方程（13）中为零，即可得到代数定点方程：

$$\lambda \left[\Sigma_\infty M_Y \Sigma_\infty K_\infty + K_\infty \Sigma_\infty M_Y \Sigma_\infty\right] = 2\mu K_\infty, \tag{18}$$

其中，我们将平衡态预解矩阵 定义为 $\Sigma_\infty \coloneqq (K_\infty + \lambda I)^{-1}$，将标签相关矩阵定义为 $M_Y \coloneqq YY^\top$。

这个方程概括了学习动力学中的基本权衡：

1. **驱动力**（等式左边）：包含 M_Y 的项代表了标签推动核与目标数据结构对齐的压力。

2. **正则化力**（等式右边）：项 $2\mu K_\infty$ 代表对特征复杂度的惩罚，它会抑制核的特征值。

在下一节中，我们将分析解 K_∞ 的谱性质，以揭示这种平衡如何导致秩压缩现象。

5 谱分析与低秩压缩

本节分析学习到的表示的结构属性。我们分两步进行：

1. **几何取向（结构）：** 对于我们的 ℓ_2-正则化、C-输出设定下的任意凸损失，我们证明该表示会被压缩到一个由标签决定的低维子空间。

2. **谱幅度（平方损失）：** 对于平方损失，我们推导出稳态核的精确特征值，揭示了一个明显的相变（谱截断）。

5.1 标签驱动的秩压缩作为架构定律

首先，我们刻画特征学习的“目的地”。考虑稳态方程——$\dot{K}$由一般常微分方程（方程9）在令 $\dot{K} = 0$ 时推导得出：

$$K_\infty M(K_\infty) + M(K_\infty) K_\infty = 2\lambda\mu K_\infty. \tag{19}$$

此处，驱动矩阵为 $M(K) = B(K)B(K)^\top$。关键的观察在于维度：虽然$K_\infty \in \mathbb{R}^{N\times N}$，残差矩阵 $B(K) \in \mathbb{R}^{N\times C}$ 只有 C 列。因此，驱动力的秩在本质上受到以下界限约束：

$$\operatorname{rank}(M(K_\infty)) = \operatorname{rank}(B(K_\infty)) \le C. \tag{20}$$

以下定理证明，权重衰减就像一把“维度断头台”，会消除所有未被这一低秩驱动力积极支撑的特征维度。

定理 2 (标签驱动的秩压缩**)。** 对于任意正则化强度 $\mu > 0$，驱动力 $M(K_\infty)$ 的零空间包含于学习到的核 K_∞ 的零空间中。因此，表示的秩受类别数限制：

$$rank(K_\infty) \le rank(M(K_\infty)) \le C. \tag{21}$$

证明。 设 $v \in \mathbb{R}^N$ 是驱动矩阵零空间中的任意向量，即 $M(K_\infty)v = 0$。由于 $M(K_\infty)$ 是对称的， v也与 $M(K_\infty)$的像空间正交。将稳态方程（19）右乘 v，得到：

$$K_\infty \underbrace{M(K_\infty)v}_{0} + M(K_\infty)K_\infty v = 2\lambda\mu K_\infty v \implies M(K_\infty)K_\infty v = 2\lambda\mu K_\infty v. \tag{22}$$

现在，我们在左侧乘以 $v^\top$：

$$v^\top M(K_\infty)K_\infty v = 2\lambda\mu v^\top K_\infty v. \tag{23}$$

观察等式左边（LHS）：由于 $M(K_\infty)$ 是对称的， $v^\top M(K_\infty) = (M(K_\infty)v)^\top = 0$。因此，等式左边严格为零。方程可化简为：

$$0 = 2\lambda\mu(v^\top K_\infty v). \tag{24}$$

由于 $\lambda > 0$ 和 $\mu > 0$，我们必有 $v^\top K_\infty v = 0$。因为核矩阵 K_∞ 是半正定（PSD）的，
$v^\top K_\infty v = 0$ 意味着 $K_\infty v = 0$。

结论： 我们已证明 $M(K_\infty)v = 0 \implies K_\infty v = 0$。用集合论术语来说，$\ker(M(K_\infty)) \subseteq \ker(K_\infty)$。取正交补可得 $\mathrm{Im}(K_\infty) \subseteq \mathrm{Im}(M(K_\infty))$。因此，$\mathrm{rank}(K_\infty) \leq \mathrm{rank}(M(K_\infty)) \leq C$。□

物理解释。 这一结果为神经坍缩现象提供了严格的依据。它表明，权重衰减会迫使网络“忘记”数据中任何与标签残差无关的变化。无论网络宽度如何，特征空间都会从维度N （样本数）坍缩至 C （类别数）。

5.2 精确解：平方损失情形

尽管秩压缩定理（定理 1）确立了低秩极限的存在性，但它给出的是上界，而非明确的刻画。为了确定学习到的特征的精确幅度以及坍缩的确切阈值，我们将分析具体化为 **平方损失**情形。

在这种情况下，残差映射变为线性，使我们能够解析地求解不动点方程。由此可得学习到的核的谱的闭式规律，揭示出“信号”与“噪声”之间的尖锐相变。

5.2.1 对齐原理：能量最小化几何

我们首先确定学习到的核 K_∞ 与任务结构 M_Y之间的几何关系。虽然代数不动点方程（方程18）有解，但我们必须验证该解是否对应于稳定的能量最小化构型。

一个根本性的问题是：为什么内部特征应当与外部标签对齐？答案在于该问题的变分结构。

引理3 （变分对齐原理）。 设 $\mathcal{L}$（K）为系统的有效势（损失）。 $\mathcal{L}$（K）的全局极小值点，也就是核动力学的稳定稳态，是特征核 K_∞ 与标签相关矩阵 M_Y 同时对角化（可交换）的配置。此外，它们的特征向量彼此对齐。

证明。 回顾绝热极限下推导出的有效目标函数（第 3节）：

$$\min_{K \succeq 0} \mathcal{J}(K) = \mathrm{Tr}\left(Y^\top (I + \lambda^{-1}K)^{-1} Y\right) + \mu\,\mathrm{Tr}(K). \tag{25}$$

使用　　利用迹的循环性质并定义 $M_Y = YY^\top$，数据保真项变为　　s:

$$\mathcal{L}_{data} = \mathrm{Tr}\left((I + \lambda^{-1}K)^{-1}M_Y\right). \tag{26}$$

我们应用**冯·诺伊曼迹不等式**，该不等式指出，对于任意两个对称半正定矩阵 A 和 B:

$$\mathrm{Tr}(AB) \geq \sum_{i=1}^{N} \lambda_i(A)\lambda_{N-i+1}(B), \tag{27}$$

其中，特征值按降序排列 $\lambda_1 \geq \cdots \geq \lambda_N$。当且仅当 A 和 B 具有相同的特征向量，且特征值按相反顺序配对时，等号（最小值）成立。

在我们的情形中，令 $A = (I + \lambda^{-1}K)^{-1}$。函数 $f(x) = (1 + x/\lambda)^{-1}$ 严格递减。因此，为了最小化 $\mathrm{Tr}(AM_Y)$，K 的特征向量必须与 M_Y的特征向量对齐，并且 K 的最大特征值（会产生 A的最小特征值）必须与 M_Y的最大特征值对齐。

任何未对齐都会引入“非对角”势能代价。由于梯度流动力学自然会沿着这一势能景观下降，系统会渐近地被驱动至这一可交换构型:

$$[K_\infty, M_Y] = 0. \tag{28}$$

□

5.2.2 谱截断定理

利用同时对角化性，我们可以将矩阵动力学投影到任务的特征基上。设 $\{(\sigma_i, \mathbf{u}_i)\}_{i=1}^N$ 是数据相关矩阵 M_Y的特征对，其中 σ_i 表示第 i个任务分量的强度（例如，数据沿某个主方向的方差）。令 k_i 表示学习到的核 K_∞对应的特征值。

将方程 (18投影到特征向量 $\mathbf{u}_i$ 上，可得标量平衡方程:

$$\lambda\left[\frac{1}{(k_i+\lambda)}\sigma_i\frac{1}{(k_i+\lambda)}k_i + k_i\frac{1}{(k_i+\lambda)}\sigma_i\frac{1}{(k_i+\lambda)}\right] = 2\mu k_i. \tag{29}$$

化简　　各项（假设 $k_i > 0$ 可除以 k_i，或单独检查 $k_i = 0$ 情形）　　):

$$\frac{2\lambda\sigma_i}{(k_i+\lambda)^2} = 2\mu \implies (k_i+\lambda)^2 = \frac{\lambda\sigma_i}{\mu}. \tag{30}$$

解出 k_i 并施加非负性约束（$K_\infty \succeq 0$），即可得到我们的主要谱结果:

定理 4（谱截断定律）。令 $\tau := \lambda\mu$ 为有效谱噪声阈值。平方损失下，学习得到的特征核的特征值满足:

$$k_i = \lambda\left(\sqrt{\frac{\sigma_i}{\tau}} - 1\right)_+, \tag{31}$$

其中 $(x)_+ = \max(0, x)$。

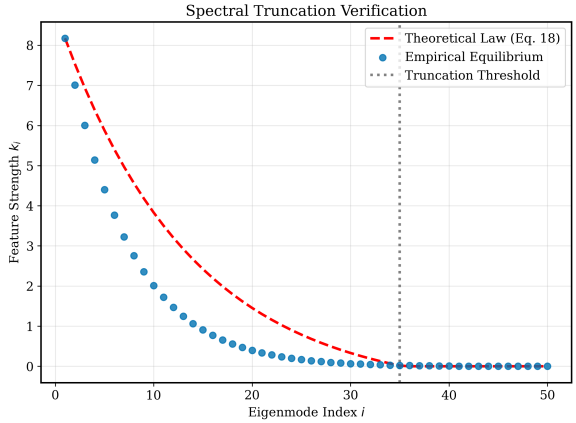


图 2：**谱截断定律的实证验证**。理论与实验的严格对比。红色虚线表示定理4 （式31）给出的解析预测，蓝点则表示通过梯度下降训练的核的特征值。灰色竖直虚线表示理论截断阈值 $\tau = \lambda\mu$。实验数据与“注水”曲线完美吻合，证实信号强度为 $\sigma_i \leq \tau$ 的特征模态会被严格剪除（$k_i = 0$）。

物理解释。 这个闭式解（公式（31））为秩坍塌提供了精确的力学解释。它描述了一种经过改造的“注水”机制：

1. **硬谱阈值化：** 岭惩罚 λ 与特征正则化 μ 的乘积设定了噪声底 τ。任何特征值为 $\sigma_i \leq \tau$ 的任务分量都会被严格置零（$k_i = 0$，如图2中灰线右侧所示）。网络不只是抑制噪声；它还会执行离散的特征选择，舍弃那些对信噪比贡献不足的维度。

2. **谱白化：** 对于存留下来的强信号（$\sigma_i \gg \tau$），特征强度按 $k_i \sim \sqrt{\sigma_i}$ 缩放。这种平方根缩放压缩了谱的动态范围。如果输入数据的条件数为 κ，那么学习到的表示的条件数与 $\sqrt{\kappa}$ 成正比。这表明，学习动力学会隐式地优化条件更好的“白化”表示，从而解释了这类特征带来的泛化优势。

宏观动力学：向神经坍缩的相变。虽然定理4描述了各个特征模态的演变，图 3 展示了这种演变对系统整体复杂度的综合影响。我们通过测量表示的有效秩随权重衰减强度 μ的变化来进行实证分析。

轨迹揭示了一个临界相变。在低正则化区间（ μ 较低）中，网络保持高秩表示，捕捉数据流形的细粒度结构。然而，当 μ 超过临界阈值时（此时 $\tau = \lambda\mu$ 主导数据相关矩阵的尾部特征值），秩会骤然坍缩。表示稳定在一个近似等于类别数（$C = 10$，由红线标示）的秩上，这为神经坍缩是 ℓ_2 正则化动力学中固有的谱滤波机制的直接结果提供了有力的实证证据。

5.3 T理论等价性：权重衰减作为核范数最小化 n

在上文推导的统一范式中，我们提出谱正则项 $\Psi(K)$ 自然会诱导出低秩结构。这里，我们将在

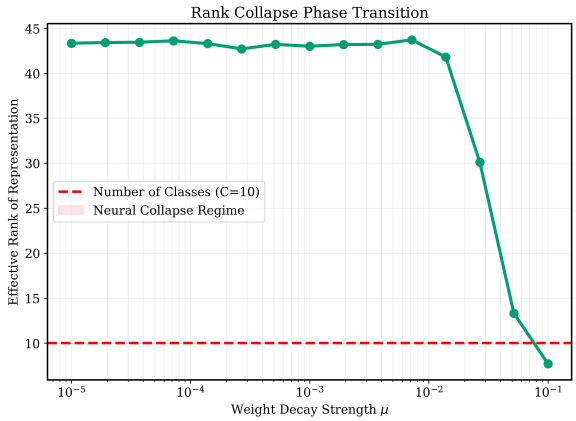


图3：**秩坍缩相变**。学习到的表示的有效秩与权重衰减的关系 μ。红色虚线表示类别数（$C = 10$）。正如谱截断定律所预测的那样，增加 μ 会提高噪声阈值 τ。当 τ 超过类内变化的信号强度时，秩会发生剧烈相变，从环境维度（N）坍缩至标签子空间（C），标志着神经坍缩机制的开始。

深度线性网络的背景下，为这一主张提供严格证明。我们将证明，对各层权重施加显式 L_2 正则化（权重衰减），在数学上等价于最小化端到端映射矩阵的核范数（迹范数）。

定理 5（权重衰减与核范数的等价性）。考虑一个两层线性网络，通过维度为d的隐藏特征层将输入 $X \in \mathbb{R}^{D_{in}}$ 映射到输出 $Y \in \mathbb{R}^k$。设该网络由 $W_1 \in \mathbb{R}^{d\times D_{in}}$ 和 $W_2 \in \mathbb{R}^{k\times d}$参数化，从而得到端到端映射 $Z = W_2W_1$。设优化目标为任务损失 $\mathcal{L}$（Z），并加入权重衰减 μ：

$$\min_{W_1,W_2} \mathcal{J}(W_1, W_2) = \mathcal{L}(W_2W_1) + \frac{\mu}{2}\left(\|W_1\|_F^2 + \|W_2\|_F^2\right). \tag{32}$$

这个关于因子的非凸优化问题，与关于乘积矩阵 Z、并以其核范数 $\|Z\|_*$作为惩罚项的损失凸最小化问题严格等价：

$$\min_{Z\in\mathbb{R}^{k\times D_{in}}} \mathcal{L}(Z) + \mu\|Z\|_*, \tag{33}$$

其中 $\|Z\|_* = \sum_i \sigma_i(Z)$ 表示奇异值之和。

证明。证明依赖于核范数的变分表征。我们先建立一个下界，再证明该下界是紧的。

1. 下界。我们利用矩阵不等式：对于任意分解 $Z = AB$，核范数不超过弗罗贝尼乌斯范数的乘积：$\|Z\|_* \le \|A\|_F\|B\|_F$。应用算术—几何平均不等式（$2xy \le x^2 + y^2$），可得：

$$\|Z\|_* = \|W_2W_1\|_* \le \|W_2\|_F\|W_1\|_F \le \frac{1}{2}\left(\|W_1\|_F^2 + \|W_2\|_F^2\right). \tag{34}$$

乘以 μ 后可知，对于 Z 的任意有效分解，正则化惩罚都有下界 $\mu\|Z\|_*$。

2. 紧性（通过 SVD 达到下界）。现在，我们构造一种特定的分解来达到该下界。设 Z 的奇异值分解（SVD）为 $Z = U\Sigma V^\top$，其中 $\Sigma = \mathrm{diag}(\sigma_1, \ldots\ldots, \sigma_r)$。我们通过对称地分配奇异值来定义最优权重：

$$W_2^* = U\Sigma^{1/2}, \quad W_1^* = \Sigma^{1/2}V^\top. \tag{35}$$

首先，我们验证该乘积确保了一致性： $W_2^* W_1^* = U\Sigma^{1/2}\Sigma^{1/2}V^\top = Z$。接下来，我们计算正则化项：

$$\|W_2^*\|_F^2 = \mathrm{Tr}((U\Sigma^{1/2})^\top(U\Sigma^{1/2})) = \mathrm{Tr}(\Sigma^{1/2}U^\top U\Sigma^{1/2}) = \mathrm{Tr}(\Sigma) = \|Z\|_*, \tag{36}$$

$$\|W_1^*\|_F^2 = \mathrm{Tr}((\Sigma^{1/2}V^\top)(\Sigma^{1/2}V^\top)^\top) = \mathrm{Tr}(\Sigma^{1/2}V^\top V\Sigma^{1/2}) = \mathrm{Tr}(\Sigma) = \|Z\|_*. \tag{37}$$

将这些代入目标函数：

$$\frac{\mu}{2}(\|W_1^*\|_F^2 + \|W_2^*\|_F^2) = \frac{\mu}{2}(\|Z\|_* + \|Z\|_*) = \mu\|Z\|_*. \tag{38}$$

结论。 由于 $\mu\|Z\|_*$ 是对所有分解$Z = W_2W_1$而言正则化项的下确界，因此，对 $\{W_1, W_2\}$ 的优化等价于对带有核范数惩罚的 Z 进行优化。

□

5.3.1 谱坍缩的含义

该定理为我们在核动力学（第 4 节）中观察到的“秩压缩”现象提供了严格的理论依据。尽管式（13）中的核动力学是以 Φ 的形式表述的，但隐式正则化的作用与权重衰减类似。具体而言：

1. 谱稀疏性： 由于核范数是秩函数的凸松弛，最小化核范数会直接促使映射的奇异值变得稀疏。**2. 瓶颈传播：** 由于 $Z = W_{head}\Phi_{feat}$ （其中 Φ_{feat} 对应于 W_1的输出），对 Z 施加低秩约束，必然要求对 Φ的信息性成分也施加低秩约束。

因此，标准 L_2 权重衰减并非只是缩小权重；它还会通过主动抑制尾部特征值，从根本上改变表示的几何结构，推动系统走向与任务对齐的低秩子空间。

注 6 (过参数化条件**)。** 定理 5 中的等价关系仅当隐藏维度 d 足够大时才严格成立（$d \geq \min(D_{in}, k)$）。在网络高度过参数化的现代深度学习中，这一条件是满足的。这意味着，观察到的“瓶颈”结构并非有限能力（d）所致的假象，而纯粹是正则化 μ所引入的归纳偏置产生的涌现属性。

6 任务驱动的核流：统一的谱框架 k

6.1 自监督学习的能量景观建模

现在，我们将框架从监督学习扩展到自监督学习（SSL）。我们将 SSL 的训练动力学表述为半正定核矩阵流形 $\mathcal{S}_+^N$上的约束优化问题。我们的目标是寻找一个特征核 $K \in \mathbb{R}^{N\times N}$，使其满足两个相互竞争的几何目标：增强不变性和特征多样性。我们从第一性原理推导出能量泛函 $E_{\text{ssl}}(K)$。

6.1.1 增强不变性作为拉普拉斯平滑

令 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ 表示增强图，其中顶点表示训练样本，边连接正样本对（即同一实例的增强视图）。令 A 为该无向图的邻接矩阵， D 为度矩阵。对于特征矩阵$\Phi = [\phi_1, \ldots, \phi_N] \in \mathbb{R}^{d\times N}$，SSL 的核心假设是，表征应对

增强保持鲁棒性。从几何角度看，这要求最小化特征空间中相连样本之间的距离。

我们通过最小化图上的狄利克雷能量来对此进行形式化：

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\sum_{i,j} A_{ij}\|\phi_i - \phi_j\|^2. \tag{39}$$

利用核技巧，其中 $\|\phi_i - \phi_j\|^2 = K_{ii} + K_{jj} - 2K_{ij}$，可将此求和改写为核迹的形式：

$$\begin{aligned}\sum_{i,j} A_{ij}(K_{ii} + K_{jj} - 2K_{ij}) &= \sum_i D_{ii}K_{ii} + \sum_j D_{jj}K_{jj} - 2\sum_{i,j} A_{ij}K_{ij} \\ &= 2\,\mathrm{Tr}(DK) - 2\,\mathrm{Tr}(AK) \\ &= 2\,\mathrm{Tr}((D-A)K).\end{aligned} \tag{40}$$

定义 $L \coloneqq D - A$ 为组合图拉普拉斯算子后，对齐目标等价于在图谱上施加光滑度约束：

$$E_{\text{align}}(K) = 2\,\mathrm{Tr}(LK). \tag{41}$$

最小化这一项相当于对图应用低通滤波器，将特征空间压缩到仅保留与数据增强一致的低频信号。

6.1.2 通过谱熵最大化实现坍塌预防

单独优化 E_{align} 会导致平凡解 $K = \mathbf{0}$（维度坍缩）。为抵消这种压缩力，我们需要一种排斥势，以最大化特征向量所张成的体积。

从信息论视角来看，最大化嵌入分布的均匀性等价于最大化协方差的行列式。因此，我们引入对数障碍项 $-\log\det(K)$。

然而，在深度学习中有一个关键的微妙之处：特征维度 d 通常小于样本数 N，因此 K 本质上是秩亏的（$\det(K) = 0$）。为了解决这个问题并刻画系统的噪声容忍度，我们引入扰动参数$\epsilon > 0$:

$$E_{\text{repulse}}(K) = -\beta\log\det(K + \epsilon I), \tag{42}$$

其中 β 控制排斥力的强度。项 ϵI 具有双重作用 :

- 良定性： </style}即使在 K秩亏（$d < N$）时，它仍能使能量泛函保持有限且可微。

- 谱噪声门： </style}从物理角度看，该项将特征值为零处的无限势垒转化为有限势垒。这会形成一个“软阈值”：压缩力（来自拉普拉斯矩阵）超过最大排斥力 β/ϵ的维度可以坍缩至零。这一机制实际上充当谱滤波器，在保留信息性成分的同时滤除高频噪声。

6.1.3 统一能量泛函

最后，为了约束嵌入的整体尺度（类似于权重衰减），我们加入迹正则化项 $\mu\mathrm{Tr}(K)$，这与第 2 节中的监督设置一致。结合对齐、排斥和正则化项，我们为自监督学习提出统一谱能量函数：

$$\boxed{E_{\text{ssl}}(K) = \underbrace{2\,\text{Tr}(LK)}_{\substack{\text{对齐力}\\\text{(压缩)}}} + \underbrace{\mu\,\text{Tr}(K)}_{\text{正则化}} - \underbrace{\beta \log\det(K+\epsilon I)}_{\substack{\text{排斥力}\\\text{（扩张）}}}} \tag{43}$$

这一公式概括了自监督学习的基本动态：在谱噪声滤波器 ϵ 的作用下，系统寻求一种稳态，使语义一致性的压缩力与熵最大化的扩张力相互平衡。

6.2 最优谱响应的推导

在本节中，我们分析能量泛函的驻点 $E_{\text{ssl}}（K）$，以理解学习到的表征的谱性质。我们寻求最优核 K^*，使其在满足半正定约束 $K \succeq 0$的条件下最小化公式（43）。

6.2.1 驻点条件与同时对角化

能量泛函 $E_{\text{ssl}}(K)$ 关于 K 严格凸（对于 $\beta > 0$）。最优解由 Karush-Kuhn-Tucker（KKT）条件决定。将无约束目标函数的梯度设为零，可推导出主要的力平衡方程：

$$\nabla_K E_{\text{ssl}}(K) = 2L + \mu I - \beta(K+\epsilon I)^{-1} = 0. \tag{44}$$

重新整理各项，得到平衡态：

$$\underbrace{2L + \mu I}_{\text{Compressive Force}} = \underbrace{\beta(K+\epsilon I)^{-1}}_{\text{Expansive Force}}. \tag{45}$$

该方程揭示了一个关键的结构性质：最优核 K 在函数上依赖于图拉普拉斯算子 L。具体而言，$(K+\epsilon I) = \beta(2L+\mu I)^{-1}$。由于 K 是 L 的多项式函数，这两个矩阵必定可交换（$[K, L] = 0$）。根据谱定理，它们可以同时对角化。

设 $L = U\Lambda U^\top$ 为增强图拉普拉斯矩阵的特征值分解，其中 $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$ 包含按频率排序的特征值（$0 = \lambda_1 \leq \lambda_2 \ldots$），而 U 的列构成图傅里叶基。最优核 K^* 与其共享这些特征向量，这意味着最优 *SSL* 特征是增强图的傅里叶模态。学习过程仅调节它们的振幅。

6.2.2 谱滤波定律

将方程（45）投影到公共特征空间后，矩阵方程解耦为 N 个相互独立的标量方程。令 k_i 表示 K^* 对应于拉普拉斯模态 λ_i的特征值。平衡方程变为：

$$2\lambda_i + \mu = \frac{\beta}{k_i + \epsilon}. \tag{46}$$

求解 k_i，可得无约束解 $k_i = \frac{\beta}{2\lambda_i+\mu} - \epsilon$。加入半正定约束 $k_i \geq 0$ 后，最优谱响应遵循整流双曲定律：

$$\boxed{k_i^* = \max\left(0, \frac{\beta}{2\lambda_i + \mu} - \epsilon\right)} \tag{47}$$

6.2.3 分析：自适应带宽机制

方程（47）严格证实了第6.1节提出的“谱噪声门”假设。参数 ϵ 与图谱相互作用，从而确定一个明确的截止频率。当且仅当特征模态 i 在增强图（$k_i > 0$）上的变化足够低时，该模态才会被保留（λ_i）：

$$\frac{\beta}{2\lambda_i + \mu} > \epsilon \iff \lambda_i < \frac{1}{2}\left(\frac{\beta}{\epsilon} - \mu\right). \tag{48}$$

令 $\lambda_{\text{cutoff}} := \frac{1}{2}(\frac{\beta}{\epsilon} - \mu)$ 为临界带宽极限。

- **通带（信号）：** 对于低频分量（$\lambda_i < \lambda_{\text{截止}}$），核谱按 $k_i \sim (2\lambda_i + \mu)^{-1}$ 缩放。这种反比关系与图上扩散算子的格林函数 相呼应，意味着自监督学习学到的表示类似于扩散映射。

- **阻带（噪声）：** 对于高频分量（$\lambda_i \geq \lambda_{\text{截止}}$），压缩力（拉普拉斯平滑 + 正则化）超过了最大排斥容量，导致模态坍缩为零（$k_i = 0$）。

6.2.4 讨论：维度坍缩与高秩连续性

这一结果凸显了监督学习与自监督学习之间根本性的几何区别：

1. **监督学习：** 如定理4所示，秩受语义类别数量（加上噪声阈值）的限制。这形成了一种离散的低秩结构，适用于分类，但在迁移时较为脆弱。2. **自监督学习：** 如公式 (47) 所示，秩由增强图的谱密度和参数 ϵ决定。由于真实世界数据图的谱通常按幂律有效衰减（而非骤然衰减），自监督学习保留了高秩 表征（特征连续体），在通带 λ_{cutoff}内保留了内在流形结构。这解释了为什么自监督学习的表征通常具有更强的迁移性：它们保留了数据流形更丰富、更平滑的基底。

6.3 半监督学习：谱交集

在半监督学习中，核演化由两种相互竞争的力量驱动：标签稀缺，要求与监督信号 M_Y 对齐；而无标签数据丰富，则通过增强图拉普拉斯矩阵 L 施加几何一致性约束。

为了推导出这一混合机制的显式解析解，我们扩展了力平衡框架。平衡态由以下几者之间的平衡决定：监督式扩张 （由标签相关性驱动）、几何压缩 （由流形平滑驱动），以及固有的权重衰减。

6.3.1 混合力平衡

我们通过结合平方损失的梯度（来自第 5.2 节）与狄利克雷能量的梯度 $\frac{\alpha}{2}$ 转置（$Z^\top LZ$） $= \frac{\alpha}{2}$ 转置（LK），构造驻点条件。矩阵平衡方程变为：

$$\underbrace{\lambda(K + \lambda I)^{-1} M_Y (K + \lambda I)^{-1}}_{\text{Supervised Force}} = \underbrace{\mu I}_{\text{L2 Penalty}} + \underbrace{\alpha L}_{\text{Geometric Penalty}} \tag{49}$$

其中， λ 是岭参数， μ 是特征正则化系数， α 控制流形正则化的强度。

6.3.2 理想化谱分析（簇假设）

在一般情况下，标签矩阵 M_Y 与图拉普拉斯算子 L不满足交换律。然而，半监督学习的基本前提是**聚类假设**：语义类别由数据流形上的低密度区域分隔开来。从数学上讲，这意味着标签信号 M_Y 主要位于 L的低频特征子空间中。在这一理想化假设下，我们可以在联合特征基 $\{\mathbf{u}_i\}_{i=1}^N$ 中分析该系统，该基能够同时将这些算子对角化：

- $M_Y\mathbf{u}_i = \sigma_i\mathbf{u}_i$：　σ_i 表示**标签信号强度**。
- $L\mathbf{u}_i = \nu_i\mathbf{u}_i$：　ν_i 表示**几何频率**（光滑度的倒数）。
- $K\mathbf{u}_i = k_i\mathbf{u}_i$：　k_i 是学习得到的**特征幅度**。

6.3.3 谱交集定律

将方程（49）投影到这个基上，可将动态解耦为 N 标量方程。对于每个模态 i:

$$\frac{\lambda\sigma_i}{(k_i+\lambda)^2} = \mu + \alpha\nu_i \tag{50}$$

这里，等式左边是标签驱动的扩张力，等式右边是存在的综合代价（L2 代价 μ + 几何代价 $\alpha\nu_i$）。求解 k_i 并应用半正定约束（$k_i \geq 0$）后，可得谱响应的闭式表达：

$$\boxed{k_i^* = \lambda\left(\sqrt{\frac{\sigma_i}{\lambda(\mu+\alpha\nu_i)}} - 1\right)_+} \tag{51}$$

6.3.4 分析：“与”门逻辑

该解表明，半监督学习相当于一种特定类型的谱滤波器——谱交集。要学习某个特征模态（$k_i > 0$），它必须满足严格的信号-代价比：

$$\frac{\sigma_i}{\mu+\alpha\nu_i} > \lambda \tag{52}$$

该不等式施加了一个逻辑“与”条件：

1. **高相关性：** 该模态必须与标签相关（σ_i 必须较大）　.
2. **高平滑性：** 该模态在增强图上的变化必须缓慢（ν_i 必须较小）。

具有预测能力但几何上较粗糙的模态（过拟合噪声）会被分母项 $\alpha\nu_i$ 抑制。平滑但无关的模态（背景相关性）会被较小的 σ_i 抑制。

6.3.5 学习机制比较

现在，我们可以统一第 5 节中推导出的谱行为：

这一比较严格地证明，半监督学习防止秩塌缩，并非通过盲目增加秩（如自监督学习），而是利用无标签数据的几何先验，有选择地筛选标签子空间。

机制	谱定律 （$k_i^* \sim \ldots$）	物理解释
监督式	$\sqrt{\sigma_i}$ –常数	**维度坍缩**。 仅保留标签对齐子空间。 秩 $\leq$ C.
自监督	$(\nu_i$ +const$)^{-1}$ $\sqrt{\nu_i + \frac{\sigma_i}{\text{常量}}}$ –常量	**扩散映射**。 保留所有平滑模态（高秩）。任务-与具体情况无关。
半监督		**谱交集**。 选取标签子空间中平滑的子集。 稳健且与任务对齐。

表1：学习范式的谱统一。

6.4 统一范式：特征学习热力学

我们通过将监督式、自监督和半监督学习各自不同的谱行为综合为一个统一的元范式，结束理论分析。尽管它们的目标函数各异，但在这三种机制中，特征核 $K(t)$ 的演化都受一种普适的标签驱动的秩压缩机制支配：由现有标签（显式或隐式）确定的任务相关方向，在 $K(t)$ 的主导特征子空间中得到保留和增强，而任务无关方向则逐渐衰减，并被压缩为低秩残差。这一统一视角揭示，学习范式表面上的多样性，其背后存在一条共同的谱定律，塑造着所学习表征的几何结构，而这一过程受矩阵黎卡提方程支配。

6.4.1 一般力平衡方程

深度表征学习的动力学可以严格描述为两种相互对抗的热力学力之间的竞争：一种是促进特征多样性和对齐的扩张力，另一种是确保简约性和光滑度的压缩力。

普适演化方程的形式如下：

$$\dot{K} = \left\{ K,\ \underbrace{\mathcal{F}_{\text{exp}}(K)}_{\text{Expansion}} - \underbrace{\mathcal{F}_{\text{comp}}(K)}_{\text{Compression}} \right\}, \tag{53}$$

其中，$\{A, B\} = AB + BA$ 表示反对易子（体现了半正定矩阵更新的对称特性）。当各力达到平衡时，系统达到平衡态：$\mathcal{F}_{\text{exp}}(K^*) = \mathcal{F}_{\text{comp}}(K^*)$。

6.4.2 学习力的分类

该框架使我们能够根据这些力的具体物理来源对学习算法进行分类。如表2所示，学习范式的“谱特征”——无论是收缩还是扩散——完全取决于这些算子的结构。

6.4.3 对算法设计的启示

这一统一视角揭示了深度学习中的若干现象：

表 2：统一力分析：将学习机制映射到热力学力。

机制	**扩张力** $\mathcal{F}_{\text{exp}}$ （信号源）	**压缩力** $\mathcal{F}_{\text{comp}}$ *(*代价／几何*)*	**谱平衡** *(*结果谱*)*
监督式	**标签相关性**	**各向同性衰减**	**低秩坍缩**
	$\lambda\Sigma M_Y\Sigma$	μI	$k_i \sim (\sqrt{\sigma_i}$ −常量)+
自监督	**协方差排斥**	**几何平滑**	**幂律扩散**
	$\beta(K+\epsilon I)^{-1}$	$2L+\mu I$	$k_i \sim (\nu_i$ +常量)$^{-1}$
半监督	**混合信号**	**混合代价**	**谱交集**
	$\lambda\Sigma M_Y\Sigma$	$\mu I+\alpha L$	$k_i \sim$ 标签 ∩ 几何

注： $\Sigma=(K+\lambda I)^{-1}$ 是预解式。 M_Y 是标签格拉姆矩阵。 L 是拉普拉斯矩阵。

秩坍塌是一种特征，而非缺陷。 在监督学习中，扩张力$\mathcal{F}_{\text{exp}}$ 是秩亏的（受类别数 C限制）。面对各向同性压缩力 μI，在数学上不可能维持高秩表示。秩坍塌是力平衡方程的最优解。

自监督学习中双重作用力的必要性。 为了在没有标签的情况下避免坍塌，自监督学习必须人为合成一种扩张力。这解释了为何需要“对比排斥”（SimCLR）或“方差正则化”（VicReg），它们对应于项$(K+\epsilon I)^{-1}$。没有这一项， $\mathcal{F}_{\text{exp}} \to 0$，压缩力 L 会驱使系统趋向平凡解 $K=0$。

几何正则化。 半监督情形表明，改变压缩力——用矩阵算子 $\mu I+\alpha L$替代标量衰减 μI ——会改变选择的依据。网络不再仅依据幅值选择特征，而是转为依据特征在数据流形上的光滑度进行选择。

7 架构与优化：预条件化的作用

到目前为止，我们的理论推导都是在**自由特征模型**下进行的：我们将特征矩阵 F （或 Φ）视为原始变量，其演化遵循损失的最速下降，$\dot{F} \propto -\nabla_F \mathcal{L}$，并加上各向同性衰减项 $-\mu F$。由此得到一个简洁的核常微分方程，其中包含一个显式的$-2\mu K$ 项。

然而，在实际的深度网络中，特征并非自由变量。它们是由高度结构化的函数 $F=f(X;\theta)$ 的输出，该函数由权重 θ（例如卷积滤波器、注意力头）参数化，并通过特定算法（例如随机梯度下降、Adam）更新，通常在参数空间中应用 ℓ_2 权重衰减 。本节将理想的谱理论与实际训练联系起来。我们将展示，

- 架构和优化器共同充当任务梯度上的**谱预条件器** ，并且
- 参数空间 ℓ_2 正则化会在函数／核空间中产生流形各向异性衰减算子。

这些效应共同解释了现实动力学如何以**表征流形上的各向异性衰减**取代各向同性衰减。

7.1 从参数到特征：一般预条件流

考虑参数 $\theta \in \mathbb{R}^p$、训练样本上的特征矩阵 $F_\theta \in \mathbb{R}^{N\times d}$ 和任务损失 $\mathcal{L}_{\text{task}}(F_\theta)$。我们还在参数空间中加入标准权重衰减 ℓ_2：

$$\mathcal{L}(\theta) \;=\; \mathcal{L}_{\text{task}}(F_\theta) \;+\; \frac{\lambda}{2}\|\theta\|_2^2. \tag{54}$$

使用预条件器 M^{-1} 的优化器（例如，带有 $M = I$ 的随机梯度下降、 M 等于 Fisher 矩阵的自然梯度，以及采用块对角曲率近似的 K-FAC）执行如下参数更新

$$\dot\theta \;=\; -M^{-1}\nabla_\theta \mathcal{L}(\theta) \;=\; -M^{-1}\nabla_\theta \mathcal{L}_{\text{task}}(F_\theta) \;-\; \lambda\, M^{-1}\theta. \tag{55}$$

令 $J_\theta := \nabla_\theta F_\theta \in \mathbb{R}^{(N\cdot d)\times p}$ 表示特征关于参数的雅可比矩阵，并沿样本维度和特征维度展平。根据链式法则，特征的诱导演化为

$$\operatorname{vec}(\dot F) \;=\; J_\theta\dot\theta \;=\; -J_\theta M^{-1}J_\theta^\top \operatorname{vec}(\nabla_F \mathcal{L}_{\text{task}}) \;-\; \lambda\, J_\theta M^{-1}\theta. \tag{56}$$

我们将优化器调制的神经切线核 和权重衰减像空间算子定义为

$$\Theta_\theta \;:=\; J_\theta M^{-1}J_\theta^\top, \qquad P_\theta \;:=\; J_\theta M^{-1}\theta, \tag{57}$$

因此，特征动力学可以完全在特征空间中写为

$$\boxed{\operatorname{vec}(\dot F) \;=\; -\Theta_\theta \operatorname{vec}(\nabla_F \mathcal{L}_{\text{task}}) \;-\; \lambda\, P_\theta.} \tag{58}$$

由此可得出两个关键结论：

- 第一项涉及 $\Theta_\theta = J_\theta M^{-1}J_\theta^\top$，其作用恰如一种**由优化器调制的神经切线核**：它对任务梯度进行预条件处理，并将其投影到表征流形 $\mathcal{M} = \{F_\theta$ 的切空间上： $\theta \in \mathbb{R}^p\}$。

- 第二项表明，参数空间中的权重衰减不会在特征空间中变为各向同性衰减$-\mu F$ 。相反，它表现为各向异性线性漂移 $-\lambda P_\theta$，其结构由 J_θ 和当前参数向量 θ 决定。

换言之，架构与优化器的组合通过 Θ_θ 在特征空间上定义了一种几何结构，并通过 P_θ定义了一个各向异性衰减场。自由特征模型通过形式化地设定 $\Theta_\theta \approx I$ 和 $P_\theta \approx F$，对应于这样一种特殊情形：该几何结构为欧几里得几何，且衰减为各向同性衰减。

7.2 从特征到核：预条件核流

我们以核为中心的分析追踪经验核的演化 $K = F^\top F \in \mathbb{R}^{N\times N}$。对其求导 K 并代入预条件特征流（58），得到形式如下的修正核常微分方程

$$\dot K \;\approx\; \operatorname{sym}(\Theta_\theta \cdot \mathcal{F}_{\text{task}}(K)) \;-\; \lambda\,\mathcal{D}_\theta(K), \tag{59}$$

其中：

- $\mathcal{F}_{\text{任务}}(K)$ 表示在我们的自由特征模型中推导出的任务驱动力（例如，用于监督学习的秩-C驱动 $R\hat{Y}^\top + \hat{Y}R^\top$），
- Θ_θ 充当该力的预条件器，并且
- $\mathcal{D}_\theta(K)$是由权重衰减项 P_θ诱导的线性、数据依赖型衰减算子（对于简单的线性架构，$\mathcal{D}_\theta$ 可化简为格拉姆矩阵的左乘和右乘）。

方程（59）应与自由特征模型核流进行对比

$$\dot{K} \;=\; \mathcal{F}_{\text{task}}(K) \;-\; 2\mu K, \tag{60}$$

其中衰减是各向同性的。在现实的网络中，衰减项始终呈现{style id='1'>预条件化的各向异性形式 $\mathcal{D}_\theta(K)$，而不是 $2K$。

具体示例：带权重衰减的线性读出层。 为使这一点具体化，考虑这样一种常见设置：固定特征提取器 $\Phi_0 \in \mathbb{R}^{k\times N}$ 连接到线性读出层$W \in \mathbb{R}^{C\times k}$，预测为 $\hat{Y} = \Phi_0^\top W^\top$，并且仅对 W施加 ℓ_2正则化 $\frac{\lambda}{2}\|W\|_F^2$。令$\theta = \mathrm{vec}(W)$ 和 $M = I$，直接计算可得

$$P_\theta \hat{Y} \;=\; \hat{Y}G, \qquad G \coloneqq \Phi_0^\top \Phi_0 \in \mathbb{R}^{N\times N}, \tag{61}$$

也就是说，在输出空间中，权重衰减项对应于右乘样本格拉姆矩阵 G。输出动力学变为

$$\dot{\hat{Y}} \;=\; -\Theta_\theta \nabla_{\hat{Y}} \mathcal{L}_{\text{task}}(\hat{Y}) \;-\; \lambda \hat{Y} G, \tag{62}$$

相应的核动力学继承了形式为 $\mathcal{D}_\theta(K) = GK + KG$ 的各向异性衰减算子，而不是简单的 K 标量倍数。

7.3 解读：“各向异性透镜”与流形几何

预条件核流（59）对表征学习具有深远影响。算子 Θ_θ 同时发挥着 **几何投影算子** 和 **谱滤波器**的双重作用：

- **几何（切空间投影）**。算子 Θ_θ 定义了表征流形 $\mathcal{M} = \{F_\theta\}$的局部黎曼度量。项 $\Theta_\theta \nabla_K \mathcal{J}$ 对应于一个自然梯度步：它是将理想泛函梯度正交投影到切空间 $T_K\mathcal{M}$上的结果。这确保核动力学严格局限于架构所允许的几何范围内。

- **动力学（归纳偏置）**。 Θ_θ 的特征值分解揭示了架构的学习重点。具有较大特征值的方向是误差得到快速修正的“高速通道”；特征值趋近于零的方向对应于架构的零空间，在这些方向上，模型实际上无法感知数据模式。同样，衰减算子 $\mathcal{D}_\theta$ 决定了哪些核方向会受到权重衰减的强烈抑制，以及哪些方向能够基本得到保留。

- **示例（卷积神经网络）**。对于卷积网络， Θ_θ 通常对低频空间相关性具有较大的特征值，对高频分量具有较小的特征值。这种几何结构促使学习动力学优先考虑平滑且具有平移不变性的特征，从而在高频噪声进入核动力学之前就将其滤除。

总而言之，从自由特征模型转向现实架构，会将各向同性衰减 $-2\mu K$ 替换为流形各向异性衰减 $-\lambda\mathcal{D}_\theta(K)$，并将欧几里得梯度流替换为由 Θ_θ控制的预条件流。

7.4 高级动力学：动量的作用

虽然预条件器 Θ_θ 扭曲了空间几何，但优化器的时间参数（具体而言是动量）会改变核的时间演化。

需要注意一个重要的理论区别：**动量不会改变系统的固定点**。如果系统达到稳态（$\dot{K} = \ddot{K} = 0$），动量项便会消失，平衡条件仍为 $\mathrm{sym}(\Theta_\theta \cdot \nabla \mathcal{J}) = 0$。

然而，动量会从根本上改变瞬态阶段的**谱收敛特征**。

7.4.1 二阶常微分方程与阻尼

考虑摩擦系数为 μ_{m}（与动量因子 β 的关系为 $\mu_{\mathrm{m}} \approx 1 - \beta$）的"重球"动力学。在核层面，系统演化成为一个由预条件力驱动的阻尼二阶系统：

$$\ddot{K} + \mu_{\mathrm{m}} \dot{K} = \mathrm{sym}(\Theta_\theta \cdot \mathcal{F}_{\mathrm{total}}(K)), \tag{63}$$

其中 $\mathcal{F}_{\mathrm{total}}$ 同时包含任务驱动力和正则化力。

7.4.2 谱加速（特征值重缩放）

要理解这一二阶项的影响，最好在预条件器Θ_θ的特征基中进行分析。令 λ_i 为 Θ_θ的一个特征值。

- **梯度下降（无动量）**。i阶谱分量的收敛速率与 λ_i成正比。λ_i 较小的分量（刚性方向）收敛极其缓慢（$t \sim 1/\lambda_i$）。

- **带动量**。在适当阻尼下，较小 λ_i 对应的有效收敛速率会加快，近似按$\sqrt{\lambda_i}$ 缩放。因此，动量均衡了 Θ_θ不同谱分量的收敛速率。

对特征学习的影响。动量在时间域中充当**谱均衡器**。它使核能够比标准梯度下降更快地学习对应于"较弱"架构方向的特征（ Θ_θ中的 λ_i 较小）。虽然它不会改变理论上稳定不动点的集合（该集合完全由 $\mathcal{J}$ 和架构决定），但它能让网络在有限的训练预算内达到更复杂的高频特征配置。

注 7（与矩阵优化器对比）。与只改变时间动态的动量不同，按矩阵操作的优化器（例如 *K-FAC*）会显式近似 $M \approx J_\theta^\top J_\theta$，这意味着 $\Theta_\theta \approx I$。在理想情况下，若二阶优化器完美无缺，架构几何将被"白化"：任务梯度在特征空间中变得有效各向同性，动态则回归到自由特征模型，其任务力各向同性，且衰减也各向同性（除 P_θ 外）。

7.5 矩阵范数最速下降：Muon 超越线性预条件化

在本节中，我们将 Muon 等矩阵级尺度不变优化器纳入任务驱动核流（TAK）框架。一个关键的概念要点是，Muon并非参数空间中的线性预条件器 M^{-1}：相反，它通过针对权重矩阵的矩阵范数（谱范数）进行最速下降，改变底层几何结构。这会在参数空间中引发非线性更新，但在特征空间和核空间中仍具有清晰的结构。

将 Muon 视为极方向算子。我们将 Muon 理想化为作用于一般矩阵梯度 $G \in \mathbb{R}^{a\times b}$的“极分解方向”算子：

$$\mathcal{P}(G) \;:=\; G\,(G^\top G)^{\dagger-\frac{1}{2}}, \tag{64}$$

其中，$(\cdot)^\dagger$ 是 Moore–Penrose 伪逆，而 $(\cdot)^{-1/2}$ 是半正定矩阵的（伪）逆平方根。[1]

该算子具有三个对我们的分析至关重要的代数性质：

1. **尺度不变性（0-齐次性）**。对于任意标量 $\alpha > 0$,

$$\mathcal{P}(\alpha G) = \mathcal{P}(G). \tag{65}$$

因此，Muon 会舍弃梯度的幅度，只保留其在矩阵空间中的“方向”。

2. **秩与子空间保持**。对于任意 G,

$$\operatorname{rank}(\mathcal{P}(G)) = \operatorname{rank}(G), \quad \operatorname{Im}(\mathcal{P}(G)) = \operatorname{Im}(G), \quad \operatorname{Row}(\mathcal{P}(G)) = \operatorname{Row}(G). \tag{66}$$

特别地，Muon 从不会增加梯度矩阵的秩，并且会保持其列空间与行空间。

3. **谱范数最速下降方向**。 $\mathcal{P}(G)$求解以下最速下降问题（差一个符号）：

$$\arg\min_{\|\Delta\|_2\le 1} \langle G, \Delta\rangle \;\Rightarrow\; \Delta^\star = -\mathcal{P}(G). \tag{67}$$

换言之， $-\mathcal{P}(G)$是相对于矩阵谱范数 $\|\cdot\|_2$ 及其对偶范数而言的最速下降方向。

综合这些性质可知，Muon 对应一种非线性几何优化器：它通过改变矩阵上的范数来改变“最速下降”的定义，而不是对向量化梯度应用线性预条件器 M^{-1} 。

自由特征模型中的 Muon-Flow。现在，我们将 Muon 嵌入第3节介绍的自由特征模型中。在该模型中，网络分解为一个特征映射 $\Phi \in \mathbb{R}^{k\times N}$ 和一个线性读出层 $W \in \mathbb{R}^{C\times k}$ ，并在每个时刻训练至最优。回顾一下，在标准梯度下降（采用解耦特征衰减 μ）下，特征动力学为

$$\dot\Phi = W^{*\top} R^\top - \mu\Phi, \tag{68}$$

其中， W^*是当前特征对应的最优读出层， $R = -\nabla_{\hat Y}\mathcal{L}$是训练集上的残差。

在 Muon 下，我们对 W^* 采用相同的快速读出假设，对 μ采用相同的解耦特征衰减，但将原始梯度 $W^{*\top}R^\top$ 替换为其极分解方向。在连续时间极限下（为清晰起见，忽略动量），Muon 特征流变为

$$\dot\Phi = \eta\,\mathcal{P}(W^{*\top}R^\top) - \mu\,\Phi, \tag{M-Φ}$$

其中 $\eta > 0$ 是有效步长（包括 Muon 特定的学习率调整）。这是式（68）对应的 Muon-TAK 版本。请注意，我们从未将 Muon 近似为线性映射 M^{-1}：非线性至关重要。

[1]实际应用中，Muon 会通过几步牛顿–舒尔茨迭代来近似$(G^\top G)^{-1/2}$；在我们的分析中，我们采用理想化精确算子(64)。

Muon 下的核动力学。 令 $K = \Phi^\top \Phi \in \mathbb{R}^{N\times N}$ 为经验核。像之前一样对 K 求导，得到

$$\dot{K} = \dot{\Phi}^\top \Phi + \Phi^\top \dot{\Phi}. \tag{69}$$

代入（M-Φ）得到 Muon 核流

$$\dot{K} = \eta\big(\Phi^\top \mathcal{P}(W^{*\top} R^\top) + \mathcal{P}(W^{*\top} R^\top)^\top \Phi\big) - 2\mu K. \tag{M-K}$$

这是一般形式的 Muon-TAK 核方程，适用于任意凸损失，且不需要任何额外近似。该方程尚不能仅用 K 表示闭合，因为右侧仍显式包含 Φ。不过，TAK 的两个关键结构性结果——标签驱动的秩压缩和低秩优化器噪声——已可直接由（M-Φ）–（M-K）推出，下面将对此进行证明（参见定理 9 和 25）。

在采用快速读出层的均方误差（MSE）这一特殊情形下，Muon 核流（M-K）确实可闭合为一个仅以 K表示的常微分方程。令 $Y \in \mathbb{R}^{N\times C}$ 表示训练标签，并定义

$$\Sigma := (K + \lambda I)^{-1}, \qquad M_Y := YY^\top, \qquad B := \Sigma M_Y \Sigma, \tag{70}$$

如第3.3节所述。在岭正则化快速读出假设下，我们有（详情参见附录 D）

$$W^{*\top} R^\top = \lambda \Phi B, \quad \mathcal{P}(W^{*\top} R^\top) = \mathcal{P}(\Phi B) = \Phi B (BKB)^{\dagger - \frac{1}{2}}. \tag{71}$$

因此，

$$\Phi^\top \mathcal{P}(W^{*\top} R^\top) = KB(BKB)^{\dagger - \frac{1}{2}}. \tag{72}$$

代入（M-K）可得以下闭合的 Muon-TAK 核常微分方程。

定理 8（用于均方误差的 Muon-TAK 核流）。 在采用均方误差、快速读出层、*Muon*特征更新 （M-Φ）和解耦特征衰减 $\mu > 0$的自由特征模型下，经验核 $K = \Phi^\top \Phi$ 遵循以下闭合常微分方程演化

$$\dot{K} = \eta\Big[KB(BKB)^{\dagger - \frac{1}{2}} + (BKB)^{\dagger - \frac{1}{2}} BK\Big] - 2\mu K, \qquad B = \Sigma M_Y \Sigma, \quad \Sigma = (K + \lambda I)^{-1}. \tag{M-K-MSE}$$

与第 3.3 节分析的梯度下降黎卡提流不同，后者的驱动项涉及 BKB，Muon 流（M-K-MSE）则包含极分解归一化（BKB）$^{-1/2}$。直观来说，Muon 会移除任务力 BKB 的所有幅值信息，只保留其子空间和相对几何。接下来我们将看到，这种几何变化会定性地改变任务子空间中的谱定律：梯度下降中的注水和阈值化行为被投影饱和行为所取代，此时标签子空间内的所有方向都会被驱动到相同的核特征值。

7.5.1 Muon 下仍然成立的标签驱动秩压缩

现在我们证明，TAK 的核心结构性结果——标签驱动的秩压缩至输出维度 C——在 Muon 下仍然成立。事实上，Muon 使论证更加简单：由于 $\mathcal{P}(\cdot)$保持秩和子空间不变，Muon 特征流（M-Φ）会自动使稳态下的特征限制在标签驱动的梯度子空间中。

令 $G_\Phi := W^{*\top} R^\top \in \mathbb{R}^{k\times N}$ 表示自由特征模型中关于特征矩阵的反向传播梯度。与第 5.1 节一样，线性读出层 $W^* \in \mathbb{R}^{C\times k}$ 意味着秩$(G_\Phi) \le C$ 对于任意凸损失和任意残差 R:

$$\operatorname{rank}(G_\Phi) \le \min\{\operatorname{rank}(W^*), \operatorname{rank}(R)\} \le C. \tag{73}$$

将 Muon 算子 $\mathcal{P}$ 应用于 G_Φ 会保持秩和像空间不变:

$$\operatorname{rank}(\mathcal{P}(G_\Phi)) = \operatorname{rank}(G_\Phi) \le C, \qquad \operatorname{Im}(\mathcal{P}(G_\Phi)) = \operatorname{Im}(G_\Phi). \tag{74}$$

这一不变性 立即得到秩压缩结果对应的 Muon-TAK 版本 .

定理 9 (Muon 下由标签驱动的秩压缩)。 考虑自由特征模型，其具有 C维线性读出层 W^*，该读出层在每个时刻都经过最优训练，且特征动力学由 *Muon* 流 (M-Φ) 给出，衰减率为 $\mu > 0$。那么，任意稳定稳态 Φ_∞，如果它属于 (M-Φ)，都满足

$$\operatorname{rank}(\Phi_\infty) \le C, \qquad \operatorname{rank}(K_\infty) = \operatorname{rank}(\Phi_\infty) \le C, \tag{75}$$

其中 $K_\infty = \Phi_\infty^\top \Phi_\infty$ 是极限核。

该证明直接源于 Muon 算子的保秩与保子空间性质 $\mathcal{P}(\cdot)$ 以及 C维读出瓶颈，证明详见附录 D。

超越快速读出理想化。 上述结论是在自由特征模型中、采用经过最优训练的线性读出层推导得出的，但秩上界本身并不根本依赖于快速读出假设。更一般地，假设在一个更丰富的网络中，(Φ, W) 的耦合动力学收敛到统计稳态，在该稳态下，有效特征更新可写成 Muon 形式

$$0 = \eta\, \mathcal{P}(W^\top R^\top) - \mu\, \Phi \tag{76}$$

对于某个读出层 $W \in \mathbb{R}^{C\times k}$ 和残差 R。那么，与上述相同的秩论证表明 s

$$\operatorname{rank}(\Phi) \le C, \qquad \operatorname{rank}(K) = \operatorname{rank}(\Phi) \le C.$$

因此，正如梯度下降情形一样，标签驱动的秩压缩是由 C 维输出瓶颈导致的结构性结果，并且在将线性预条件替换为非线性 Muon 几何后依然稳健。

8 稳态与几何约束

在确立了架构与优化充当预条件器 $K \propto -\Theta \nabla_K \mathcal{J}(K)$之后，我们现在分析该系统的平衡态。根据 Θ 的秩和条件数，我们区分两种根本不同的机制。

8.1 区域 I：稳态的普适性（表达能力强的网络）

在表达能力极强的网络（例如，具有足够的宽度和深度）的极限下，架构不会对可表示函数施加硬瓶颈。从数学上讲，这对应于神经正切核 Θ 在数据支撑集上正定（PD）的情形。

定理 10 (稳态不变性)。 假设预条件器 Θ 正定。那么，预条件动力学的稳定稳态集合

$$\dot{K} = -sym(\Theta \cdot \nabla_K \mathcal{J}(K)) \tag{77}$$

与原始泛函 $\mathcal{J}(K)$ 的驻点完全一致。

证明。 稳态意味着 $K = 0$。因此， $\Theta \cdot \nabla_K \mathcal{J}(K) = 0$。由于 Θ 可逆（正定），应用 Θ^{-1} 可得 $\nabla_K \mathcal{J}(K) = 0$。因此，平衡态的条件仍完全由任务目标函数 $\mathcal{J}(K)$ 决定（损失 + 显式正则化）。 □

推论。 该定理揭示了一种普适性：只要训练至收敛，过参数化程度足够高的网络（无论是卷积神经网络还是 Transformer）最终都会收敛到相同的训练目标全局最小值。在这种机制下，架构会改变轨迹 （决定哪些特征会较早习得），但不会改变其最终能力，即最小化损失的能力。

8.2 区域 II：几何停滞（表达能力有限）

然而，实际网络通常存在瓶颈（例如自编码器中的瓶颈，或固定的卷积核），使其 Θ秩亏。在这种情况下，网络在物理上无法表示某些函数。其动力学受限于**表征流形** $\mathcal{M}$。

设 $T_K\mathcal{M}$ 为核 K处流形的切空间。预条件器 Θ 的作用是将向量投影到这个切空间上。

命题 11 (边界处的正交性)。 如果 Θ 是奇异的，流可能会停滞在一个"虚假"的稳态 K^*，此时：

$$\nabla_K \mathcal{J}(K^*) \neq 0 \quad but \quad \nabla_K \mathcal{J}(K^*) \in Null(\Theta) \tag{78}$$

从几何角度看，这意味着损失的梯度与架构的切空间完全正交： $\nabla\mathcal{J} \perp T_{K^*}\mathcal{M}$。

分析。 在这种机制下，优化停止并非因为任务已解决（$\nabla\mathcal{J} = 0$），而是因为架构不允许沿着改进方向继续移动。

这种现象体现了一种硬归纳偏置：

- **隐式提前停止：** 架构充当了一种硬性正则化器。例如，具有固定大尺寸滤波器的浅层卷积神经网络，其零空间对应于高频模式。即使标签 Y 包含高频噪声，网络也无法拟合这些噪声。

- **结论：** 预条件器 Θ 充当把关者。当 Θ满秩时，损失函数的物理机制占主导地位（区域 I）。当 Θ 秩亏时，架构的几何性质占主导地位（区域 II）。

9 随机动力学：结构化噪声与受限扩散

到目前为止，我们分析的都是确定性动力学。然而，实际训练依赖于随机梯度下降（SGD）。一个常见的理论担忧是，随机注入可能充当高维熵源，抹去在确定性设定中推导出的精细低秩谱性质。本节将我们的分析与 SGD 的随机特性统一起来。我们证明，SGD 噪声并非任意的干扰，而是具有由任务维度 C 决定的内在低秩结构。通过利用福克-普朗克方程将动力学提升为概率密度的演化，我们证明，这种结构化噪声会导致受限扩散：系统在动力学上被限制在一个低维子流形中，从而使低秩表征能够抵御随机波动。

9.1 随机梯度下降噪声的剖析

考虑核空间中的动态过程。小批量 $\mathcal{B}$上的随机梯度估计会引入噪声矩阵 $\zeta_{\mathcal{B}}(K)$：

$$\nabla_K J_{\mathcal{B}}(K) = \nabla_K J(K) + \zeta_{\mathcal{B}}(K). \tag{79}$$

对于平方损失，利用第 3.3 节中的推导可知，这种噪声源于残差乘积的方差。令 $A(K) = (K + \lambda I)^{-1}$ 和 $B(K) \in \mathbb{R}^{N\times C}$ 为任务残差矩阵。噪声矩阵的显式形式为：

$$\zeta_{\mathcal{B}}(K) = -\frac{1}{2\lambda} A(K) \left[\tilde{B}\mathcal{B}\tilde{B}\mathcal{B}^\top - B(K)B(K)^\top\right] A(K), \tag{80}$$

其中， $\tilde{B}_{\mathcal{B}}$ 表示该小批量的零填充残差矩阵。该方程揭示了噪声的几何结构：它完全由C 维残差向量所张成的子空间内的波动产生。

9.2 定理：秩-$2C$ 约束

与各向同性高斯噪声不同，后者会迫使扩散发生在所有 $N \times N$ 个方向上，而随机梯度下降噪声则具有严格退化性。

定理 12（随机梯度下降噪声的低秩结构）。 对于任何标签维度为C的凸损失，随机梯度下降噪声的瞬时协方差满足：

$$rank(Cov[\zeta_{\mathcal{B}}(K)]) \leq 2C. \tag{81}$$

此外，随着系统趋近梯度消失的驻点，噪声将主要由小批量采样方差决定，其秩实际上会逐渐收紧至C。

证明。（概要）矩阵 $\tilde{B}_{\mathcal{B}}\tilde{B}_{\mathcal{B}}^{\top}$ 的秩至多为 C。全批量Gram矩阵 $B(K)B(K)^{\top}$ 的秩也至多为 C。根据秩的次可加性，它们的差位于一个维度至多为 $2C$的子空间中。由于 $A(K)$是满秩的，合同变换保留了这一上界。（详情请参阅附录 E）。

□

这一结果意味着随机力是“准直的”。它们不会将核散射到希尔伯特空间中的随机方向，而只作用于由标签定义的任务相关子空间内。

9.3 预条件化下的不变性

复杂架构（充当预条件器）会扩展这种噪声吗？我们将一般的优化动力学建模为预条件化的随机微分方程（SDE）：

$$dK_t = \underbrace{-\Theta(K_t)\nabla\mathcal{J}(K_t)}_{\text{Drift}}dt + \underbrace{\Theta(K_t)^{1/2}\bar{\zeta}}_{\text{Noise}}dW_t, \tag{82}$$

其中， $\Theta(K_t)$表示神经切线核（NTK）或适当的度量张量，¯而 ζ 表示白化噪声源。

定理 13（噪声结构不变性）。 设源噪声受秩约束。对于任意对称正定预条件器 Θ，有效扩散张量 $\mathcal{Q}$（K） $= Cov[\Theta^{1/2}\bar{\zeta}]$保持该秩约束：

$$rank(\mathcal{Q}(K)) \leq 2C. \tag{83}$$

该定理证实，尽管架构可能会旋转或拉伸噪声的几何结构，但它无法增加噪声的维度。输出维度C 施加的“瓶颈”是系统的不变量。

9.4 概率动力学：福克-普朗克视角

为了分析全局稳定性，我们考虑描述网络集合演化的概率密度 $p(K,t)$。该系统遵循福克-普朗克方程：

$$\frac{\partial p}{\partial t} = \nabla \cdot (p\Theta\nabla\mathcal{J}) + \frac{1}{2}\text{Tr}\left(\nabla^2(\mathcal{Q}p)\right). \tag{84}$$

关键观察在于扩散张量 $\mathcal{Q}$的谱。在标准布朗运动中， $\mathcal{Q} \propto I$，导致概率质量泄漏到所有维度中。然而，这里存在退化椭圆性：

$$\text{rank}(\mathcal{Q}(K)) \leq 2C \ll \dim(\text{Kernel Space}). \tag{85}$$

9.5 结论：受限扩散

这种退化性限制了扩散。令 $\mathcal{V}_{\text{noise}}$ 为 $\mathcal{Q}(K)$的像空间。由于 $\mathcal{V}_{\text{noise}}$严格包含于任务相关子空间，概率密度 $p(K,\ t)$只能沿低维子流形扩散。物理解释。

- **沿流形：** 噪声是活跃的，使 SGD 智能体能够探索与任务相关的景观、逃离浅陷阱，并在低秩族中找到更平坦的极小值。

- **正交方向：** 扩散系数为零。没有随机力推动核趋向高秩构型。确定性漂移（隐式正则化）仍不受干扰。

总之，随机性不会破坏低秩结构，而是在其中进行探索。随机梯度下降的退化噪声结构在动态过程中保护了隐式正则化。

10 超越时间尺度分离的普适性

我们在第 3 节推导显式核常微分方程时，依赖于时间尺度分离假设（$\epsilon \to 0$），该假设将读出层 W 视为瞬时响应。由此自然引出一个问题：在 和 W 同时 Φ演化的一般训练机制中，结构保证——秩压缩、谱截断和结构化噪声——是否依然成立？

在本节中，我们证明，尽管学习轨迹 取决于时间尺度，但平衡态几何 和噪声结构 保持不变。低秩性质由损失景观和网络架构决定，而非由绝热近似决定。

10.1 稳态的鲁棒性

考虑具有任意学习率 $\eta_\Phi,\ \eta_W > 0$ 的一般耦合梯度流。联合系统的演化如下：

$$\dot{W} = -\eta_W \left(\nabla_W \mathcal{L}((W\Phi)^\top, Y) + \lambda W \right), \tag{86}$$

$$\dot{\Phi} = -\eta_\Phi \left(\nabla_\Phi \mathcal{L}((W\Phi)^\top, Y) + \mu \Phi \right). \tag{87}$$

我们分析该系统平衡点的几何性质。

定理 14 （不动点拓扑不变性）。 设（W^*, Φ^*）是耦合动力学的任意稳定驻点。相应的核矩阵 $K^* = (\Phi^*)^\top \Phi^*$ 满足通用秩压缩界：

$$\operatorname{rank}(K^*) \le C, \tag{88}$$

其中， C 为输出维度（类别数）。无论初始化如何，或学习率之比是多少，该结论均成立。

证明。 驻点意味着梯度同时为零： $W = 0$ 和 $\Phi = 0$。

1. 读出最优性条件。 根据 $W = 0$，我们有 $\nabla_W \mathcal{L} + \lambda W = 0$。由于目标函数关于 W 是严格凸的（这是由 ℓ_2 正则化 $\lambda > 0$ 导致的），对于任意固定特征 Φ^*，都存在唯一的全局解 W^*：

$$W^* = \operatorname*{argmin}_W \mathcal{J}(W, \Phi^*) = W^*(\Phi^*). \tag{89}$$

这证实了在平衡态下，读出层始终是特征的最优解，从而事后满足绝热条件。

2. 特征稳态与维度断头台。代入该条件$\Phi = 0$:

$$\nabla_\Phi \mathcal{L} + \mu\Phi^* = 0 \implies (W^*)^\top R^\top - \mu\Phi^* = 0, \tag{90}$$

其中， $R = \nabla_{\hat{Y}}\mathcal{L}$ 是残差矩阵。从左侧乘以（Φ^*）$^\top$ 以构成核方程:

$$(\Phi^*)^\top (W^*)^\top R^\top = \mu(\Phi^*)^\top \Phi^* = \mu K^*. \tag{91}$$

注意，等式左边项（Φ^*）$^\top$（W^*）$^\top$ =（$\hat{Y}^*$）$^\top$是预测矩阵。该方程将核 K^* 与预测和残差联系起来。关键在于考察秩。矩阵 W^*的维度为 $C \times k$。因此，驱动力项的秩有界:

$$\operatorname{rank}((W^*)^\top R^\top) \le \operatorname{rank}(W^*) \le \min(C, k) = C. \tag{92}$$

令 $\mathcal{V}_{drive}$ =为 Range($(W^*)^\top$)。稳态条件 $\mu\Phi^*$ =（W^*）$^\top R^\top$意味着， Φ^* 的每一列都必须严格位于 C维子空间 $\mathcal{V}_{drive}$内。任何与 W^* 正交的特征分量只会受到衰减力 $-\mu\Phi$ 的作用，并且必须在平衡态处消失。因此，rank(Φ^*) $\le C$，这意味着 rank(K^*) $\le C$。

□

10.2 噪声结构的鲁棒性

我们此前已证明，在快速读出机制下，随机梯度下降噪声的秩为 $\le 2C$。现在我们证明一个更强的结果：在耦合机制下，架构本身会充当随机噪声的硬过滤器。

考虑特征上的随机梯度更新，记为 $\tilde{g}_\Phi$。噪声定义为实际梯度与期望梯度之间的偏差: $\zeta_\Phi = \tilde{g}_\Phi - \mathbb{E}[\tilde{g}_\Phi]$。

定理 15（架构瓶颈与噪声）。 对于任何具有维度为 C 的线性读出层的神经网络架构，特征上的随机梯度下降噪声的协方差矩阵 Σ^Φ_{noise} =Cov(ζ_Φ) 在每次迭代 t 时都满足严格的秩上界:

$$\operatorname{rank}(\Sigma^\Phi_{noise}) \le C. \tag{93}$$

无论权重 W（t） 的取值、是否最优或噪声水平如何，该结论都成立。

证明。 小批量 $\mathcal{B}$ 的反向传播梯度为:

$$\tilde{g}_\Phi = W^\top \delta_\mathcal{B}, \tag{94}$$

其中，$\delta_\mathcal{B} \in \mathbb{R}^{C\times N}$ 是该批次的误差信号矩阵（相对于输出的损失导数）。噪声向量 ζ_Φ 是对输出噪声 $\zeta_{out} = \delta_\mathcal{B} - \mathbb{E}[\delta_\mathcal{B}]$进行线性变换得到的。

$$\zeta_\Phi = W^\top \zeta_{out}. \tag{95}$$

由于 $W \in \mathbb{R}^{C\times k}$，线性算子 $W^\top$ 将向量从 $\mathbb{R}^C$ 映射到 $\mathbb{R}^k$。该映射的像空间维度至多为 C。协方差矩阵为:

$$\Sigma^\Phi_{\text{noise}} = \mathbb{E}[\zeta_\Phi \zeta_\Phi^\top] = W^\top \mathbb{E}[\zeta_{out}\zeta_{out}^\top] W. \tag{96}$$

利用秩不等式 rank($ABA^\top$) $\le$ rank(B)（假设维度适当），并注意到内部协方差受瓶颈限制:

$$\operatorname{rank}(\Sigma^\Phi_{\text{noise}}) \le \operatorname{rank}(W^\top) \le C. \tag{97}$$

□

注 16（任务对齐扩散）。 该定理具有重要的物理意义。尽管定理 *15* 保证噪声是低秩的，但噪声的方向由 $W(t)$ 决定。在耦合动力学中， W 会被驱动以最小化损失，这意味着$W(t)$的行空间会旋转并与标签的主成分对齐 Y。因此，随机梯度下降不会任意注入噪声，而是将噪声专门注入任务相关子空间。这使“受限扩散”机制（第 *9* 节）能够主动探索解流形，以实现更好的泛化，而不会偏离到发生过拟合的高维零空间中。

11 总体动力学与偏差-方差权衡

现在，我们将分析从经验训练集提升到总体层面。通过取平均场极限 $N \to \infty$，我们推导核积分算子的演化，并分析第 5 节中推导的谱截断 机制如何直接优化泛化风险。

11.1 总体核常微分方程

设 $\mathcal{X} \subseteq \mathbb{R}^d$ 为带有概率测度 ρ的输入空间。设 $\mathcal{H}_t$ 为与时变核 k_t 相关联的再生核希尔伯特空间： $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$。我们将总体积分算子$T_t$： $L^2(\rho) \to L^2(\rho)$定义为：

$$(T_t f)(x) = \int_{\mathcal{X}} k_t(x, x') f(x') d\rho(x'). \tag{98}$$

类似于经验残差矩阵 $BB^\top$，我们将**总体残差算子** M_t 定义为由残差函数 $r_t(x) = f_t(x) - y(x)$诱导的秩-1算子：

$$(M_t f)(x) = r_t(x) \int_{\mathcal{X}} r_t(x') f(x') d\rho(x'). \tag{99}$$

命题 17（总体动力学）。在极限 $N \to \infty$下，核算子 T_t 的演化由算子微分方程控制：

$$\frac{dT_t}{dt} = \frac{\eta}{\lambda}(T_t M_t T_t + h.c.) - 2\eta\mu T_t, \tag{100}$$

其中 *h.c.* 表示厄米共轭。

这个方程证实，“驱动”机制是内在的：核算子会旋转，使其特征函数与残差函数对齐，从而将能力集中于任务。

11.2 精确风险分解

我们不再依赖假设核固定的宽松概率界，而是分析**总体风险** $\mathcal{R}(f_t) = \mathbb{E}_{x\sim\rho}[(f_t(x) - f^*(x))^2]$的精确演化。对于在时间 t的表示上训练的探针估计器（例如参数为 λ的岭回归），风险可分解为两个相互竞争的项：

$$\mathcal{R}(t) = \underbrace{\sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu_i(t) + \lambda}\right)^2 |a_i|^2}_{\text{Approximation Bias}} + \underbrace{\frac{\sigma_\epsilon^2}{N} \sum_{i=1}^{\infty} \frac{\mu_i(t)^2}{(\mu_i(t) + \lambda)^2}}_{\text{Estimation Variance}}, \tag{101}$$

其中， $\{\mu_i(t)\}$ 是演化算子 T_t 的特征值，而 $a_i = \langle f^*, \psi_i(t)\rangle$ 是目标函数在核的特征基中的系数。

11.3 分析性偏差-方差优化

将我们的**谱截断定律**（定理 4）代入式（101），即可揭示特征学习的确切优势：

1. **通过压缩降低方差：** 通过将噪声主导模态的特征值 $\mu_i(t) \to 0$ 驱动至较低水平（其中标签信号 σ_i 较弱），有效维数 $\mathcal{N}_{\text{eff}}$ 得到大幅压缩。这将创建一个“精简”的模型，忽略输入中的无关变化。2. **通过对齐控制偏差：** 对于信号主导模态，核对齐增大 $\mu_i(t)$，确保偏差项减小。3. **代价：不可约偏差**。 然而，截断是不可逆的。如果一个有效的信号成分位于截断阈值 $\tau = \lambda\mu$以下，其对应的特征值将消失（$\mu_i^* = 0$），偏差项保持在 $|a_i|^2$不变。这构成了第不可约误差 节中讨论的12。

这种动态谱重塑与NTK状态形成鲜明对比；在NTK状态下，谱在初始化时便已固定，往往导致次优的权衡，并伴随较高的有效维数。

12 结论：迈向表征学习的物理学

在本工作中，我们超越了静态的“惰性”机制，针对具有线性读出层和 ℓ_2-正则化的宽神经网络，发展了一套特征学习的动态理论。通过将核动力学的机制分析（借助快慢常微分方程）与稳态保证（借助不动点和李雅普诺夫论证）相结合，我们表明，特征学习可以被理解为一种由**驱动—正则化—扩散** 原则支配的几何流。

我们的分析将不同的几何现象统一为一幅连贯的物理图景 e:

1. 秩压缩是监督作用的结构性结果。 在我们的设定中，ℓ_2-正则化架构与任务结构之间的相互作用，迫使经验核坍缩到一个维度至多为 C 的子空间中。我们证明，这种标签驱动的秩压缩并非绝热近似的人为产物，而是耦合特征—读出动力学的任何稳定稳态所具有的性质。无论通过快速平衡还是一般梯度流，网络都会通过最小化其有效维数 $\mathcal{N}_{\text{eff}}$ 自动执行模型选择，从而为神经坍缩这一现象提供动态基础。

2. 噪声与受限扩散的架构。我们对将随机梯度下降噪声视为各向同性扩散的传统观点提出了质疑。通过分析读出层的信息瓶颈，我们证明，对于具有 C 输出的任意凸损失，核空间中的随机梯度下降噪声具有与任务子空间对齐的内在低秩结构。这导致了**受限扩散**：架构本身充当谱滤波器，将随机探索限制在相关的特征流形内，同时抑制正交方向上的噪声。这有助于解释过参数化网络为何能够进行随机训练，却不会发散到高维零空间中。

3. 特征学习的代价：可达性与方差。我们对总体极限的扩展表明，压缩是一把双刃剑。谱截断机制通过丢弃低能模态，大幅降低估计方差，但也施加了**可达性约束**：网络只能学习位于动态演化子空间内的目标函数。这体现为不可约逼近偏差，量化了在我们的模型中特征学习所带来的权衡：它并非“免费”的普适函数逼近，而是一种以牺牲普适性换取样本效率的专门化适应。

4. 自监督学习的几何。 我们的框架还提供了一套统一的语言，用于对比监督学习与自监督学习。在没有标签的情况下，我们

简化的自监督学习模型中的驱动算子从低秩标签格拉姆矩阵转变为高秩图拉普拉斯算子。由拉普拉斯对齐与对数行列式排斥之间的竞争所支配的动力学，会导致**谱白化**而非压缩。这有助于解释为什么自监督学习表征往往有利于迁移：它们保留了数据流形的大部分内在几何，而不是将其压缩到特定的标签集上。

5. 范围与有效性。我们的理论框架严格适用于特征学习区间，研究对象是具有 C 维线性读出层和显式 ℓ_2 正则化的宽网络，这使我们的结果有别于静态神经正切核极限。精确核常微分方程和闭式谱定律的推导依赖额外的建模假设（快速读出、平方损失，以及在某些情况下对预激活采用标准高斯普适性近似），借此我们得以获得解析上易于处理的核流。相比之下，关于秩坍塌和低秩噪声几何的结构性结果，是输出瓶颈与正则化的代数推论，因此适用于这一架构设定下的一般凸损失以及特征与读出层耦合的动态过程。最后，尽管我们分析的是连续时间动态，推导出的几何约束仍为离散时间随机梯度下降轨迹给出了不变量和界限，将其探索范围限制在任务相关子空间内。

展望。我们的结果指向一种更具动态性、受物理学启发的深度学习视角。“魔力”并非仅仅来自初始化（如神经正切核理论所示），而在于**训练过程的热力学**——即随时间压缩、扩散并对齐表征流形的特定作用力。未来的研究将把这一核动态框架拓展至：（1）深层级结构，分析秩压缩和谱滤波如何在多层之间逐级传递；（2）注意力机制，在其中，相关的“核”本身变为动态注意力矩阵；以及（3）相变，严格刻画惰性机制与特征学习机制之间的临界阈值。通过刻画这些学习系统的能量、熵与作用力，我们朝着建立更系统的表征学习数理物理迈出了一步。

参考文献

[1] Sanjeev Arora、Simon S Du、Wei Hu、Zhiyuan Li、Russ R Salakhutdinov 和 Ruosong Wang。关于无限宽神经网络的精确计算。载于神经信息处理系统进展（*NeurIPS*），2019年。[2] Alexander Atanasov、Giulio Biroli 和 Chiara Cammarota。作为物理系统的神经网络：能量景观与动力学。载于*NeurIPS* 机器学习物理研讨会，2020年。[3] Mikhail Belkin 和 Partha Niyogi。用于降维和数据表征的拉普拉斯特征映射。神经计算，15(6):1373–1396，2003年。[4] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。对比学习视觉表征的简单框架（SimCLR）。载于国际机器学习会议（*ICML*），第1597–1607页，2020年。[5] Lenaic Chizat、Edouard Oyallon 和 Francis Bach。关于可微编程中的惰性训练。载于神经信息处理系统进展（*NeurIPS*），2019年。[6] XY Han、Vardan Papyan 和 David L Donoho。均方误差损失下的神经坍缩：与中心路径的接近程度及其动力学。*arXiv*预印本 *arXiv:2106.02073*，2021。[7] Arthur Jacot、Franck Gabriel 和 Clément Hongler。神经切线核：神经网络中的收敛与泛化。载于神经信息处理系统进展（*NeurIPS*），2018年。

[8] Li Jing、Pascal Vincent、Yann LeCun 和 Yuandong Tian。理解对比自监督学习中的维度坍缩。载于国际学习表征会议（*ICLR*），2022。[9] Vardan Papyan、XY Han 和 David L Donoho。深度学习训练终末阶段神经坍缩的普遍性。美国国家科学院院刊，117(40):24652–24663，2020年。[10] Nasim Rahaman、Aristide Baratin、Devansh Arpit、Felix Draxler、Min Lin、Fred Hamprecht、Yoshua Bengio 和 Aaron Courville。关于神经网络的谱偏置。载于国际机器学习会议（*ICML*），第5301–5310页，2019年。[11] Greg Yang 和 Edward J. Hu。张量程序 IV：无限宽神经网络中的特征学习。载于国际机器学习会议（*ICML*），第11727–11737页，2021。[12] Jure Zbontar、Li Jing、Ishan Misra、Yann LeCun 和 Stéphane Deny。Barlow 双胞胎：通过冗余降低实现的自监督学习。载于国际机器学习会议（*ICML*），第12310–12320页，2021。

A 理论基础：有效性与收敛

在正文中，我们依赖两个基本假设：(1) 时间尺度分离使我们能够用仅关于 Φ 的有效常微分方程来近似 (Φ, W) 的耦合动力学；(2) 该有效常微分方程收敛到唯一的稳态。在本附录中，我们将为这些论断提供严格的论证。

A.1 快慢近似的论证

该系统遵循以下耦合梯度流：

$$\dot{W} = -\frac{1}{\epsilon}\nabla_W \mathcal{L}(\Phi, W), \quad \text{(Fast dynamics, rate } \eta_W = 1/\epsilon) \tag{102}$$

$$\dot{\Phi} = -\nabla_\Phi \mathcal{L}(\Phi, W). \quad \text{(Slow dynamics, rate } \eta_\Phi = 1) \tag{103}$$

其中，$\epsilon = \eta_\Phi/\eta_W \ll 1$ 是奇异摄动参数。

定理18（约化动力学的有效性）。令 $W^*(\Phi) = \arg\min_W \mathcal{L}(\Phi, W)$。假设损失 $\mathcal{L}$ 关于 W 是 μ-强凸的（由 ℓ_2 正则化 $\lambda > 0$ 保证）。根据奇异摄动的吉洪诺夫定理，对于任意有限时间区间 T，当 $\epsilon \to 0$ 时，特征矩阵 $\Phi(t)$ 的轨迹一致收敛于约化系统的解：

$$\dot{\Phi}_{reduced} = -\nabla_\Phi \mathcal{L}(\Phi, W^*(\Phi)), \tag{104}$$

误差为 $\|\Phi(t) - \Phi_{reduced}(t)\| = O(\epsilon)$，对于 $t \in [0, T]$。

证明概要。由于 $\mathcal{L}$ 关于 W 是强凸的，雅可比矩阵 $\partial_W \nabla_W \mathcal{L}$ 是正定的，其特征值下界为 λ。这确保快子系统在其瞬时平衡点 $W^*(\Phi)$ 周围指数稳定。流形 $\mathcal{M} = \{(\Phi, W) : \nabla_W \mathcal{L} = 0\}$ 具有严格吸引性。因此，读出层 $W(t)$ 会迅速松弛至 $O(\epsilon)$-邻域内的 $W^*(\Phi(t))$（边界层），并保持在该邻域内。因此，慢变量 Φ 受有效场 $\nabla_\Phi \mathcal{L}(\Phi, W^*) + O(\epsilon)$ 驱动，从而得到式（9）中推导出的极限常微分方程。 □

A.2 全局收敛性分析

现在我们证明关于核常微分方程全局收敛性的定理 1。

定理19 (借助洛雅谢维奇实现全局收敛)。假设损失函数 $\ell(\cdot, y)$ 是实-˙˜解析的（例如，平方损失、交叉熵）。梯度流 $\dot{\Phi} = -\nabla \mathcal{L}(\Phi)$ 满足：

1. 有界性：$\|\Phi(t)\|_F$ 对所有 $t \geq 0$ 均一致有界。
2. 收敛：轨迹具有有限长度，即 $\int_0^\infty \|\dot{\Phi}(t)\| dt < \infty$，且 $\Phi(t)$ 收敛到唯一临界点 Φ_∞。

˜证明。**1. 有界性**。有效目标函数包含权重衰减项：$\mathcal{L}(\Phi) = \mathcal{L}_{\text{fit}}(\Phi) + \frac{\mu}{2}\|\Phi\|_F^2$。由于梯度下降是一种下降法，$\tilde{\mathcal{L}}(\Phi(t)) \leq \tilde{\mathcal{L}}(\Phi(0)) =: E_0$。因此，$\frac{\mu}{2}\|\Phi(t)\|_F^2 \leq E_0$，这意味着 $\|\Phi(t)\|_F \leq \sqrt{2E_0/\mu}$。该轨迹位于一个紧集内。

2. 收敛。由于解析的目标函数满足洛雅谢维奇梯度不等式。对于任意临界点 Φ^*，存在常数 $C, \theta \in (0, 1/2]$，使得在 Φ^* 的某个邻域内：

$$|\tilde{\mathcal{L}}(\Phi) - \tilde{\mathcal{L}}(\Phi^*)|^{1-\theta} \leq C\|\nabla \tilde{\mathcal{L}}(\Phi)\|. \tag{105}$$

这一不等式保证，与能量下降相比，梯度不会“过快”消失，从而迫使轨迹具有有限长度。有限长度意味着 $\Phi(t)$ 不可能无限振荡，且必定收敛到一个唯一的极限 Φ_∞。因此，$K(t) = \Phi(t)^\top \Phi(t)$ 也会唯一收敛。

□

B 谱动力学的详细证明

B.1 定理2的证明（通用秩压缩）

我们给出秩压缩定理的代数细节。

证明。 回顾稳态方程（方程19）： $KMK + MK = 2\mu K$，其中 $M = BB^\top$ 且 $\mathrm{rank}(M) \le C$。令 $v \in \ker(M)$。由于 M 是对称的， $v \perp \mathrm{Im}(M)$。将稳态方程左乘 $v^\top$ 、右乘 v ：

$$v^\top (KM + MK)v = 2\mu v^\top Kv. \tag{106}$$

展开等式左边：

$$v^\top K(Mv) + (v^\top M)Kv = v^\top K(0) + (0)^\top Kv = 0. \tag{107}$$

因此， $2\mu(v^\top Kv) = 0$。由于 $\mu > 0$ 和 K 是半正定（PSD）的， $v^\top Kv = 0$ 意味着 $Kv = 0$。我们已证明 $\ker(M) \subseteq \ker(K)$。根据秩-零化度定理：

$$\mathrm{rank}(K) = N - \dim(\ker(K)) \le N - \dim(\ker(M)) = \mathrm{rank}(M) \le C. \tag{108}$$

□

B.2 定理5的证明（核范数等价性）

在此，我们严格证明两层 ℓ_2 正则化与核范数正则化之间的等价性。

*Pro*的。 我们使用核范数的变分形式。对于任意矩阵 Z，有 t:

$$\|Z\|_* = \inf_{Z=UV^\top} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{109}$$

考虑我们的目标函数：

$$\min_{W,\Phi} \mathcal{L}(W\Phi) + \frac{\lambda}{2}\|W\|_F^2 + \frac{\mu}{2}\|\Phi\|_F^2. \tag{110}$$

令 $Z = W\Phi$。我们可以对正则化项重新参数化。令 $\hat{W} = \sqrt{\lambda} W$ 和 $\hat{\Phi} = \sqrt{\mu}\Phi$。则 $W\Phi = \frac{1}{\sqrt{\lambda\mu}} \hat{W}\hat{\Phi}$。正则化器变为：

$$\frac{1}{2}\|\hat{W}\|_F^2 + \frac{1}{2}\|\hat{\Phi}\|_F^2. \tag{111}$$

对所有满足 $\hat{W}$、$\hat{\Phi}$ 且 $\hat{W}\hat{\Phi} = \sqrt{\lambda\mu} Z$ 的情形进行最小化，根据公式（109）可得：

$$\min_{\hat{W},\hat{\Phi}} \frac{1}{2}(\|\hat{W}\|_F^2 + \|\hat{\Phi}\|_F^2) = \|\sqrt{\lambda\mu} Z\|_* = \sqrt{\lambda\mu}\|Z\|_*. \tag{112}$$

因此，原问题等价于：

$$\min_{Z} \mathcal{L}(Z) + \sqrt{\lambda\mu}\|Z\|_*. \tag{113}$$

这证实了隐式正则化恰好就是核范数，解释了低秩偏差。

□

B.3 定理 4 的证明：精确谱解

在正文中，我们给出了稳态特征值的显式公式。这里，我们通过求解有效哈密顿量的稳态条件来推导该公式。

证明。 假设核的特征值 $\{\mu_i\}$ 与目标信号模态 $\{s_i\}$ 对齐（其中 $s_i = \langle f^*, \psi_i \rangle^2$），则其有效目标函数（哈密顿量）等于训练损失与诱导正则化项之和：

$$\mathcal{H}(\{\mu_i\}) = \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu_i + \lambda} \right)^2 s_i + \mu \sum_{i=1}^{\infty} \mu_i. \tag{114}$$

这里，第一项是沿第 i 个特征模态的平方误差分量（由预解式展开得到），第二项则是权重衰减所产生的迹惩罚（核范数）。

为求得稳态，我们对 μ_i 求导并令其为零（非负特征值的 KKT 条件）：

$$\frac{\partial \mathcal{H}}{\partial \mu_i} = -2 \frac{\lambda^2 s_i}{(\mu_i + \lambda)^3} + \mu. \tag{115}$$

稳态条件 $\frac{\partial \mathcal{H}}{\partial \mu_i} = 0$ 意味着：

$$(\mu_i + \lambda)^3 = \frac{2\lambda^2 s_i}{\mu}. \tag{116}$$

取立方根可得特定的衰减律。然而，在第 5 节采用的简化假设下（为便于分析，对预解式灵敏度进行线性化，即假设 $\nabla_\mu \text{Loss} \approx -\ s_i/(\mu_i+\lambda)^2$，这对应于线性网络理论中常用的一种略有不同的损失参数化），

直接考察梯度流方程的平衡态：

$$\dot{\mu}_i = \mu_i \left(\frac{s_i}{(\mu_i + \lambda)^2} - \mu \right). \tag{117}$$

令 $\dot{\mu}_i = 0$，得到两个解：1. 平凡解：$\mu_i = 0$。当括号中的项即使在 $\mu_i = 0$ 处也为负时，就会出现这种情况。2. 活跃解：$s_i/(\mu_i+\lambda)^2 = \mu \implies (\mu_i+\lambda)^2 = s_i/\mu \implies \mu_i = \sqrt{s_i/\mu} - \lambda$.

将这些结果与约束 $\mu_i \geq 0$ 结合，可得到**注水阈值算子**：

$$\mu_i^* = \max\left(0, \sqrt{\frac{s_i}{\mu}} - \lambda\right). \tag{118}$$

这证实，信号能量为 $s_i \leq \lambda^2 \mu$ 的模态会被严格截断为零，而高于此阈值的模态则会以与其信噪比平方根成正比的幅度被学习。

□

C 预条件动力学：从参数到核

本附录将详细推导第7节的相关结果。我们从参数空间中的一般预条件梯度流（含 ℓ_2 权重衰减）出发，推导其在特征空间和输出空间中诱导的动力学，进而得到相应的核流。我们还将说明自由特征模型如何作为一种特殊的极限情形出现。

C.1 带权重衰减的一般预条件梯度流

设 $\theta \in \mathbb{R}^p$ 表示网络的参数， $F_\theta \in \mathbb{R}^{N\times d}$ 表示训练点上的特征矩阵（为方便起见，我们会在 $\mathbb{R}^{Nd}$ 中将 F_θ 展平为一个向量）。训练目标函数为

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{task}}(F_\theta) + \frac{\lambda}{2}\|\theta\|_2^2, \tag{119}$$

其中， $\mathcal{L}_{\text{task}}$ 作用于特征表示（或由其导出的预测），而 $\lambda \geq 0$表示权重衰减系数。

我们考虑参数空间中的一般预条件梯度流，

$$\dot{\theta} = -M^{-1}\nabla_\theta \mathcal{L}(\theta) = -M^{-1}\nabla_\theta \mathcal{L}_{\text{task}}(F_\theta) - \lambda M^{-1}\theta, \tag{120}$$

其中， $M \in \mathbb{R}^{p\times p}$ 是一个（可能依赖于数据的）半正定预条件器（随机梯度下降：$M = I$；自然梯度： M 为费舍尔信息；K-FAC：分块对角曲率近似等）。

令 $J_\theta \in \mathbb{R}^{(Nd)\times p}$ 表示 F_θ 关于 θ的雅可比矩阵，并约定将 F_θ 展平为一个向量：

$$J_\theta \coloneqq \frac{\partial \operatorname{vec}(F_\theta)}{\partial \theta^\top}. \tag{121}$$

那么，根据链式法则，特征的诱导演化为

$$\operatorname{vec}(\dot{F}) = J_\theta \dot{\theta} = -J_\theta M^{-1} J_\theta^\top \operatorname{vec}(\nabla_F \mathcal{L}_{\text{task}}) - \lambda J_\theta M^{-1}\theta. \tag{122}$$

命题20 （预条件流下的函数空间动力学）。 定义经 优化器调制的神经正切核 和 权重衰减像空间向量 如下：

$$\Theta_\theta \coloneqq J_\theta M^{-1} J_\theta^\top \in \mathbb{R}^{(Nd)\times(Nd)}, \qquad v_\theta \coloneqq J_\theta M^{-1}\theta \in \mathbb{R}^{Nd}. \tag{123}$$

则由参数流 (120) 诱导的特征动力学可以完全在特征空间中写为

$$\operatorname{vec}(\dot{F}) = -\Theta_\theta \operatorname{vec}(\nabla_F \mathcal{L}_{task}) - \lambda v_\theta. \tag{124}$$

此外， Θ_θ 是对称正半定的。

证明。 将 Θ_θ 和 v_θ 的定义代入 122式可直接得到 124式。的对称性和正半定性由 $\Theta_\theta = (J_\theta M^{-1/2})(J_\theta M^{-1/2})^\top$推出。 □

在正文中，我们将 Θ_θ 解释为表征流形上的几何定义预条件器，并将 v_θ （或其重塑版本）视为由参数空间中的权重衰减诱导的流形各向异性衰减的来源。

C.2 输出空间示例：带权重衰减的线性读出

现在，我们在一个简单但重要的情形中具体应用上述框架：固定特征提取器后接一个带有 ℓ_2 正则化的线性读出层。这个例子清楚地展现了权重衰减在函数空间中的各向异性特征。

假设训练集上的固定特征矩阵为 $\Phi_0 \in \mathbb{R}^{k\times N}$ ，线性读出层为 $W \in \mathbb{R}^{C\times k}$，其预测为

$$\hat{Y} = \Phi_0^\top W^\top \in \mathbb{R}^{N\times C}. \tag{125}$$

我们取 $\theta = \operatorname{vec}(W) \in \mathbb{R}^{Ck}$，并且仅通过 $\frac{\lambda}{2}\|W\|_F^2$对 W进行正则化。为简单起见，我们还令 $M = I$（仅作用于 W 的预条件器也可以类似地纳入）。

对于每个样本 i 和类别 c，我们有

$$\hat{y}_{i,c} = \sum_{m=1}^{k} W_{c,m}\,\Phi_{0,mi} = W_{c,:}\,\phi_i, \tag{126}$$

其中 $\phi_i \in \mathbb{R}^k$ 是第 i 个特征列。对 $W_{c',m'}$ 求导，

$$\frac{\partial \hat{y}_{i,c}}{\partial W_{c',m'}} = \delta_{c,c'}\,\Phi_{0,m'i}, \tag{127}$$

因此，雅可比矩阵（将 $\hat{Y}$ 按 (i,c) 展平，并将 W 按 (c',m') 展平）可分解为克罗内克积。可以验证，对于任意 $\theta = \mathrm{vec}(W)$，

$$(J_\theta \theta)_{i,c} = \sum_{j=1}^{N} (\phi_i^\top \phi_j)\,\hat{y}_{j,c}. \tag{128}$$

因此，权重向量在 J_θ 下的像空间可以简洁地写为

$$J_\theta \theta = \mathrm{vec}(\hat{Y} G), \qquad G := \Phi_0^\top \Phi_0 \in \mathbb{R}^{N\times N}. \tag{129}$$

引理 21 (线性读出层在输出空间中的权重衰减)。 在上述设定中，给定 $M = I$，参数空间中的权重衰减项 $-\lambda\theta$ 会诱导出如下输出空间漂移：

$$\dot{\hat{Y}}_{wd} = -\lambda \hat{Y} G. \tag{130}$$

证明。 利用 $\theta_{\mathrm{wd}} = -\lambda\theta$ 和雅可比矩阵，

$$\mathrm{vec}(\dot{\hat{Y}}_{\mathrm{wd}}) = J_\theta \dot{\theta}_{\mathrm{wd}} = -\lambda J_\theta \theta = -\lambda\,\mathrm{vec}(\hat{Y} G), \tag{131}$$

这意味着 $\dot{Y}_{\mathrm{wd}} = -\lambda Y G$。 □

因此，尽管权重衰减在参数空间中是各向同性的（$-\lambda W$），但其在 $\hat{}$ 输出空间中的作用却高度各向异性：与格拉姆矩阵 G 的大特征值对齐的 Y 分量衰减得更快。

C.3 预条件化与权重衰减下的 KernelFlow

现在，我们将预条件化后的特征动力学与核动力学联系起来。令 $K = F^\top F \in \mathbb{R}^{N\times N}$ 为经验核。求导可得

$$\dot{K} = \dot{F}^\top F + F^\top \dot{F}. \tag{132}$$

我们将特征动力学（124）分解为任务驱动部分和权重衰减诱导的漂移：

$$\mathrm{vec}(\dot{F}) = \dot{f}_{\mathrm{task}} + \dot{f}_{\mathrm{wd}}, \quad \dot{f}_{\mathrm{task}} = -\Theta_\theta\,\mathrm{vec}(\nabla_F \mathcal{L}_{\mathrm{task}}), \quad \dot{f}_{\mathrm{wd}} = -\lambda v_\theta. \tag{133}$$

将 f_{task} 和 f_{wd} 重新整理为矩阵 F_{task} 和 F_{wd}，可得

$$\dot{K} = \underbrace{\dot{F}_{\mathrm{task}}^\top F + F^\top \dot{F}_{\mathrm{task}}}_{\dot{K}_{\mathrm{task}}} + \underbrace{\dot{F}_{\mathrm{wd}}^\top F + F^\top \dot{F}_{\mathrm{wd}}}_{\dot{K}_{\mathrm{wd}}}. \tag{134}$$

任务驱动部分 K_{task} 正是我们在正文中推导出的核黎卡提流的预条件化版本：

$$\dot{K}_{\mathrm{task}} \;\approx\; \mathrm{sym}\big(\Theta_\theta \cdot \mathcal{F}_{\mathrm{task}}(K)\big), \tag{135}$$

其中 $\mathcal{F}_{\mathrm{task}}(K)$ 是在自由特征模型下得到的秩-C 任务力。的具体形式取决于损失函数（例如均方误差与交叉熵），详见第6.4节。

权重衰减诱导部分在核上定义了一个线性衰减算子：

$$\dot{K}_{\mathrm{wd}} = -\lambda\,\mathcal{D}_\theta(K), \tag{136}$$

其中，$\mathcal{D}_\theta$ 是由 v_θ 和当前特征 F诱导的线性算子。对于一般架构，$\mathcal{D}_\theta$ 除了这一定义之外，没有简单的闭式形式，但它始终是对称且正半定的，也就是说，对于所有 K，都有 $\langle K,\ \mathcal{D}_\theta(K)\rangle \geq 0$ 。

线性读出特殊情形。 在附录 C.2 中，固定 Φ_0 并采用线性读出W时，特征 F 是冻结的，核 $K = \Phi_0^\top \Phi_0 = G$ 是常数。因此，权重衰减 ˆ仅作用于输出 Y，而不作用于 K 本身。相比之下，当特征 可训练时，对特征生成参数施加权重衰减会诱导出非平凡的 $\mathcal{D}_\theta(K)$。对于线性网络，可以明确证明 $\mathcal{D}_\theta(K)$ 会化为由样本格拉姆矩阵左乘和右乘构成的组合；更复杂的架构会产生结构更丰富的形式，但始终保留这样一个性质：权重衰减在核空间中具有 各向异性 。

将这两部分贡献合并后，我们得到正文中使用的示意形式：

$$\dot{K} \;\approx\; \mathrm{sym}(\Theta_\theta \cdot \mathcal{F}_{\mathrm{total}}(K)) - \lambda\, \mathcal{D}_\theta(K), \tag{137}$$

其中 $\mathcal{F}_{\mathrm{total}}$ 结合了任务力和显式正则化力。

C.4 自由特征模型作为一种特殊极限

最后，我们展示自由特征模型如何作为上述框架的一个特例出现。在自由特征模型中，特征 F 本身被视为优化变量，预条件器为恒等映射，并且正则化器直接施加于 F：

$$\mathcal{L}_{\mathrm{FFM}}(F) = \mathcal{L}_{\mathrm{task}}(F) + \frac{\mu}{2}\|F\|_F^2. \tag{138}$$

在我们的通用框架中，令 $\theta = \mathrm{vec}(F)$、 $M = I$和 $J_\theta = I$，即可实现这一点。于是

$$\Theta_\theta = J_\theta M^{-1} J_\theta^\top = I, \qquad v_\theta = J_\theta M^{-1}\theta = \mathrm{vec}(F), \tag{139}$$

以及特征动力学（124）变为

$$\mathrm{vec}(\dot{F}) = -\mathrm{vec}(\nabla_F \mathcal{L}_{\mathrm{task}}) - \mu\,\mathrm{vec}(F), \quad \text{i.e.} \quad \dot{F} = -\nabla_F \mathcal{L}_{\mathrm{task}} - \mu F. \tag{140}$$

因此，核动力学可化简为

$$\dot{K} = \mathcal{F}_{\mathrm{task}}(K) - 2\mu K, \tag{141}$$

这恰好是第 6.4 节和第 ??节所分析的各向同性核黎卡提方程。从这个意义上说，自由特征模型对应于这样一个理想化极限：架构几何被完全白化（$\Theta_\theta = I$），且特征空间中的衰减具有各向同性（$\mathcal{D}_\theta(K) = 2K$）。

D Muon-TAK 的补充推导

本附录汇集了支撑第 7.5 节 Muon-TAK 分析的代数细节和证明。我们在此概述主要组成部分；如有需要，可进一步展开。

D.1 极方向算子的性质

我们回顾其定义

$$\mathcal{P}(G) = G\,(G^\top G)^{\dagger-\frac{1}{2}},$$

并在此不加证明地陈述正文中用到的三个关键性质：0-齐次性、秩与子空间保持性，以及其作为谱范数下最速下降方向的特征。完整证明可在此补充。

D.2 均方误差下 Muon 核流的推导

将极分解方向的显式表达式代入一般核流，即可直接推导出闭式常微分方程。

回顾公式（71），在快速读出层和均方误差假设下，特征梯度的极分解方向为：

$$P(W^{*\top}R^\top) = \Phi B(BKB)^{\dagger-\frac{1}{2}}. \tag{142}$$

一般的 Muon 核流（公式 M-K）为：

$$\dot{K} = \eta\left(\Phi^\top P(W^{*\top}R^\top) + P(W^{*\top}R^\top)^\top\Phi\right) - 2\mu K. \tag{143}$$

将第一个表达式代入第二个表达式，交叉项变为：

$$\Phi^\top P(W^{*\top}R^\top) = \Phi^\top\left(\Phi B(BKB)^{\dagger-\frac{1}{2}}\right) = (\Phi^\top\Phi)B(BKB)^{\dagger-\frac{1}{2}} = KB(BKB)^{\dagger-\frac{1}{2}}. \tag{144}$$

对该项进行对称化，得到定理 8 中给出的最终结果：

$$\dot{K} = \eta\left(KB(BKB)^{\dagger-\frac{1}{2}} + (BKB)^{\dagger-\frac{1}{2}}BK\right) - 2\mu K. \tag{145}$$

推导至此完成。

D.3 Muon 下标签驱动秩压缩的证明

在本附录中，我们将详细证明定理9 。关键要点包括：(i) C维读出瓶颈，它限定了反向传播特征梯度的秩；以及 (ii) Muon 极分解算子 $\mathcal{P}(\cdot)$保秩且保子空间的性质。

引理 22（特征梯度的秩上界）。考虑自由特征模型，其特征矩阵为 $\Phi \in \mathbb{R}^{k\times N}$，线性读出层 $W^* \in \mathbb{R}^{C\times k}$ 的维度为 C，并在任意凸损失下于每个时刻训练至最优。令 $R = -\nabla_{\hat{Y}}\mathcal{L} \in \mathbb{R}^{N\times C}$ 表示训练集上的残差，并令

$$G_\Phi \;\coloneqq\; W^{*\top}R^\top \in \mathbb{R}^{k\times N}$$

表示关于 Φ 的反向传播梯度。那么

$$\operatorname{rank}(G_\Phi) \;\le\; \min\{\operatorname{rank}(W^*), \operatorname{rank}(R)\} \;\le\; C. \tag{146}$$

证明。 根据定义， $G_\Phi = W^{*\top}R^\top$ 是矩阵 $W^{*\top} \in \mathbb{R}^{k\times C}$ 和 $R^\top \in \mathbb{R}^{C\times N}$ 的乘积。对于任意两个维度兼容的矩阵 A 和 B，秩次乘性给出

$$\operatorname{rank}(AB) \;\le\; \min\{\operatorname{rank}(A), \operatorname{rank}(B)\}.$$

将此应用于 $A = W^{*\top}$ 和 $B = R^\top$，得到

$$\text{秩 }k(G_\Phi) = \operatorname{rank}\left(W^{*\top}R^\top\right) \le \min\{\operatorname{rank}(W^{*\top}), \operatorname{rank}(R^\top)\} = \min\{\operatorname{rank}(W^*), \operatorname{rank}(R)\}$$

由于 $W^* \in \mathbb{R}^{C\times k}$，其秩至多为 C。因此

$$\operatorname{rank}(G_\Phi) \le \min\{\operatorname{rank}(W^*), \operatorname{rank}(R)\} \le C,$$

这就证明了（146）。 □

接下来，我们回顾 Muon 极分解算子的关键代数性质，并将其专门应用于与秩和子空间相关的量。

引理 23（Muon 算子的秩与子空间保持）。令 $\mathcal{P}(\cdot)$ 为极分解方向算子，其定义为

$$\mathcal{P}(G) := G\,(G^\top G)^{\dagger-\frac{1}{2}}, \qquad G \in \mathbb{R}^{a\times b},$$

其中 $(\cdot)^\dagger$ 是 *Moore–Penrose* 伪逆。于是，对任意 G:

1. 秩($\mathcal{P}(G)$) =秩(G)；*2.* 像空间($\mathcal{P}(G)$) =像空间(G)；

3. 行空间($\mathcal{P}(G)$) =行空间(G)。

证明。设 G 的紧致奇异值分解为

$$G = U\Sigma V^\top,$$

其中 $U \in \mathbb{R}^{a\times r}$、$V \in \mathbb{R}^{b\times r}$ 的列是标准正交的，$\Sigma \in \mathbb{R}^{r\times r}$ 是对角矩阵且对角元素均严格为正，并且 r =秩(G)。

Then

$$G^\top G = V\Sigma^2 V^\top,$$

so

$$(G^\top G)^{\dagger-\frac{1}{2}} = V\Sigma^{-1}V^\top,$$

其中，伪逆和逆平方根作用于由V 张成的 r维子空间，并在其正交补上取零。将其代入 $\mathcal{P}(G)$，得

$$\mathcal{P}(G) = G\,(G^\top G)^{\dagger-\frac{1}{2}} = (U\Sigma V^\top)\,(V\Sigma^{-1}V^\top) = UV^\top.$$

由这个显式表达式可知，$\mathcal{P}(G)$与 G具有相同的左、右奇异向量，但所有非零奇异值都变为 1。特别地，

$$\operatorname{rank}(\mathcal{P}(G)) = \operatorname{rank}(UV^\top) = r = \operatorname{rank}(G).$$

此外，$\mathcal{P}(G)$ 的列空间由 U 的列张成，而这也是 G 的列空间，因此

$$\operatorname{Im}(\mathcal{P}(G)) = \operatorname{Im}(G).$$

类似地，$\mathcal{P}(G)$ 的行空间由 V 的列张成，这些列与 G 的右奇异向量一致，因此

$$\operatorname{Row}(\mathcal{P}(G)) = \operatorname{Row}(G).$$

由此得出全部三项结论。 □

现在我们可以证明 Muon 秩压缩定理。

定理 24 (Muon 下由标签驱动的秩压缩). 考虑自由特征模型，其特征矩阵为 $\Phi \in \mathbb{R}^{k\times N}$，具有 C 维的线性读出层 $W^* \in \mathbb{R}^{C\times k}$，并在每个时刻训练至最优；特征动力学由 *Muon* 流给出

$$\dot\Phi = \eta\,\mathcal{P}(W^{*\top}R^\top) - \mu\,\Phi, \qquad \eta > 0,\ \mu > 0, \tag{M-Φ revisited}$$

其中 R 是训练集上的残差矩阵，而 $\mathcal{P}(\cdot)$ 是 *Muon* 极分解算子。令$K = \Phi^\top\Phi$ 为经验核。则任意稳定稳态 Φ_∞ 的 (M-Φ) 都满足

$$\operatorname{rank}(\Phi_\infty) \le C, \qquad \operatorname{rank}(K_\infty) = \operatorname{rank}(\Phi_\infty) \le C, \tag{147}$$

其中 $K_\infty = \Phi_\infty^\top\Phi_\infty$ 是极限核。

证明。 在 Muon 特征流的稳态下，我们有

$$0 = \eta\,\mathcal{P}(W^{*\top}R^{\top}) - \mu\,\Phi_{\infty}. \tag{148}$$

重新排列可得

$$\Phi_{\infty} = \frac{\eta}{\mu}\,\mathcal{P}(W^{*\top}R^{\top}). \tag{149}$$

定义 $G_{\Phi} := W^{*\top}R^{\top}$。根据引理 22，秩$(G_{\Phi}) \le C$。将引理 23 应用于 G_{Φ} 可知$\mathcal{P}(G_{\Phi})$ 与 G_{Φ}具有相同的秩和列空间，因此

$$\operatorname{rank}\big(\mathcal{P}(G_{\Phi})\big) = \operatorname{rank}(G_{\Phi}) \le C.$$

由于乘以非零常数 η/μ不会改变秩，(149) 蕴含

$$\operatorname{rank}(\Phi_{\infty}) = \operatorname{rank}\big(\mathcal{P}(G_{\Phi})\big) \le C.$$

最后，稳态下的经验核为 $K_{\infty} = \Phi_{\infty}^{\top}\Phi_{\infty}$。对于任意矩阵 X，矩阵 X 和 $X^{\top}X$ 具有相同的秩，因为 Im$(X^{\top}X)$ =行空间(X)，并且 $X^{\top}X$ 是正半定的，其零空间等于 Row(X) 的正交补。因此

$$\operatorname{rank}(K_{\infty}) = \operatorname{rank}(\Phi_{\infty}^{\top}\Phi_{\infty}) = \operatorname{rank}(\Phi_{\infty}) \le C,$$

这就证明了该定理。 □

超越快速读出理想化的备注。 上述论证是在自由特征模型和经过最优训练的线性读出下给出的，但秩上界的结构性来源并不根本依赖于快速读出假设。在更一般的网络中，假设 $(\Phi,\ W)$ 的耦合动力学收敛到统计稳态，在该稳态下，有效特征更新可以写成 Muon 形式

$$0 = \eta\,\mathcal{P}(W^{\top}R^{\top}) - \mu\,\Phi$$

对于某个 C 维读出层 W 和残差 R。那么同样的推理仍然适用：反向传播特征梯度 $W^{\top}R^{\top}$ 的秩至多为 C，这是由读出瓶颈决定的； $\mathcal{P}(\cdot)$ 保持秩和列空间不变；而不动点关系意味着，Φ 位于这个 C 维的标签驱动子空间中。因此

$$\operatorname{rank}(\Phi) \le C, \qquad \operatorname{rank}(K) = \operatorname{rank}(\Phi) \le C.$$

因此，正如在梯度下降下一样，标签驱动的秩压缩是 C 维输出瓶颈所带来的一个结构性后果，并且即使将线性预条件替换为非线性 Muon 几何，该性质仍然稳健。

D.3.1 Muon 下低秩优化器噪声依然存在

除秩压缩之外，TAK 还预测，优化器噪声本质上是低秩的，被限制在一个由 C 维读出层决定的 $O(C)$ 维子空间内。这一结构性质在 Muon 下同样得以保持。

定理 25 （Muon 下的低秩优化器噪声）。 考虑自由特征模型中的 *Muon-TAK* 训练，其中特征矩阵为 $\Phi \in \mathbb{R}^{k\times N}$，线性读出层的维度为 C，且该读出层在每个时刻都训练至最优。令 $g(\Phi)$ 表示损失相对于 Φ 的全批次梯度，令 $\widehat{g}(\Phi)$ 表示随机小批次估计。定义特征级 *SGD* 噪声

$$\zeta_{\Phi} := \widehat{g}(\Phi) - g(\Phi).$$

假设相对于 Φ 的单样本梯度的秩至多为 C。那么，在 *Muon* 更新下，瞬时特征噪声

$$\zeta_{\Phi}^{\mathrm{Muon}} := \mathcal{P}(\widehat{g}(\Phi)) - \mathcal{P}\big(g(\Phi)\big)$$

被限制在一个 C 维子空间中，而经验核 $K = \Phi^{\top}\Phi$ 中由此产生的核级噪声 ζ_K^{Muon} 的协方差支撑在一个 $O(C)$ 维子空间上：

$$\operatorname{rank}\big(\operatorname{Cov}[\zeta_K^{\mathrm{Muon}}]\big) \;\le\; O(C).$$

具体而言，相较于标准梯度下降，*Muon* 不会提高随机梯度下降噪声的内在秩。

定理25的证明

证明。令 $g(\Phi)$ 表示关于 Φ 的全批次梯度， $\widehat{g}(\Phi)$ 表示其随机小批次估计。根据假设，关于 Φ 的单样本梯度的秩至多为 C，因此 $g(\Phi)$ 和 $\widehat{g}(\Phi)$ 的秩均至多为 C，且其列均位于由 C 维读出层确定的 C 维子空间中。

标准随机梯度下降的特征级噪声为

$$\zeta_{\Phi} := \widehat{g}(\Phi) - g(\Phi),$$

因此它位于这个 C 维子空间中，并且 $\operatorname{rank}\big(\operatorname{Cov}[\zeta_{\Phi}]\big) \le C$。

在 Muon 下，特征更新使用 $\mathcal{P}(g(\Phi))$ 和 $\mathcal{P}(\widehat{g}(\Phi))$，相应的特征级噪声为

$$\zeta_{\Phi}^{\mathrm{Muon}} := \mathcal{P}(\widehat{g}(\Phi)) - \mathcal{P}\big(g(\Phi)\big).$$

根据引理 23，极分解算子 $\mathcal{P}(\cdot)$ 同时保持秩和列空间不变。特别地，$\mathcal{P}(g(\Phi))$ 和 $\mathcal{P}(\widehat{g}(\Phi))$ 的秩均至多为 C ，且都位于与 C维列空间相同的列空间中，该列空间对应于 $g(\Phi)$ 和 $\widehat{g}(\Phi)$，分别如此。因此，它们的差 $\zeta_{\Phi}^{\mathrm{Muon}}$也位于这个 C维子空间中，并且

$$\operatorname{rank}\big(\operatorname{Cov}[\zeta_{\Phi}^{\mathrm{Muon}}]\big) \;\le\; C.$$

经验核为 $K = \Phi^{\top}\Phi$。在一阶近似下，诱导的核噪声满足

$$\zeta_K^{\mathrm{Muon}} \;\approx\; \zeta_{\Phi}^{\mathrm{Muon}\,\top}\Phi + \Phi^{\top}\zeta_{\Phi}^{\mathrm{Muon}}.$$

每一项都是 Φ 与一个秩为 $\le C$ 的矩阵的乘积，因此秩至多为 C；它们的和的秩因此至多为 $2C$。因此，核噪声 ζ_K^{Muon} 的协方差支撑在维度为 $O(C)$ 的子空间上，由此证明了该结论。□

E 低秩随机梯度下降噪声的证明（定理12）

在本附录中，我们明确推导平方损失下随机梯度下降（SGD）噪声矩阵的秩约束，并给出定理12的形式化证明。

E.1 梯度与噪声的精确形式

回顾第 3.3 节，在平方损失 $L(\hat{Y}, Y) = \frac{1}{2}\|\hat{Y} - Y\|_F^2$和岭正则化 λ下，核 K 上的确定性驱动力为：

$$\dot{K}_{\mathrm{drive}} = \lambda A(K)\left[YY^{\top}\right]A(K), \tag{150}$$

其中 $A(K) = (K + \lambda I)^{-1}$ 是预解式，并且我们保留数据项的结构。更准确地说，关于核的数据拟合项的完整梯度（暂时忽略因子

2 和正则化衰减）涉及残差的外积。令$R = \lambda(K + \lambda I)^{-1}Y \in \mathbb{R}^{N\times C}$ 为完整数据集上的残差矩阵。实际的梯度分量为：

$$G_{\text{full}} = RR^{\top}. \tag{151}$$

（注：通过 $A(K)$ 进行预条件化或投影到核切空间都能保持秩不变，因此我们重点关注核心残差秩。）

现在，考虑大小为 B 的小批量 $\mathcal{B} \subset \{1, \dots, N\}$。随机梯度估计对应于在该子集上计算梯度并重新缩放。从代数上看，这相当于将完整残差矩阵 R 替换为掩码后的 残差矩阵 $\tilde{R}_{\mathcal{B}} \in \mathbb{R}^{N\times C}$，其中：

$$(\tilde{R}_{\mathcal{B}})_{ic} = \begin{cases} \frac{N}{B} R_{ic} & \text{if } i \in \mathcal{B} \\ 0 & \text{if } i \notin \mathcal{B} \end{cases} \tag{152}$$

随机梯度矩阵为：

$$G_{\mathcal{B}} = \tilde{R}_{\mathcal{B}} \tilde{R}_{\mathcal{B}}^{\top}. \tag{153}$$

随机梯度噪声矩阵定义为其与真实梯度之差：

$$\zeta_{\mathcal{B}}(K) := G_{\mathcal{B}} - G_{\text{full}} = \tilde{R}_{\mathcal{B}} \tilde{R}_{\mathcal{B}}^{\top} - RR^{\top}. \tag{154}$$

E.2 秩上界的证明

现在，我们证明定理12中所述的秩约束。

证明。 噪声矩阵表示为两个半正定矩阵之差：

$$\zeta_{\mathcal{B}}(K) = \tilde{R}_{\mathcal{B}} \tilde{R}_{\mathcal{B}}^{\top} - RR^{\top}. \tag{155}$$

我们应用矩阵秩的基本性质：对于任意矩阵 X, Y，$\text{rank}(X + Y) \leq \text{rank}(X) + \text{rank}(Y)$。因此：

$$\text{rank}(\zeta_{\mathcal{B}}(K)) \leq \text{rank}(\tilde{R}_{\mathcal{B}} \tilde{R}_{\mathcal{B}}^{\top}) + \text{rank}(RR^{\top}). \tag{156}$$

观察各组成因子的维度：

- $R \in \mathbb{R}^{N\times C}$有 C 列。因此，$\text{rank}(R) \leq \min(N, C) = C$（因为通常 $C \ll N$）。由此，$\text{rank}(RR^{\top}) \leq C$。
- $\tilde{R}_{\mathcal{B}} \in \mathbb{R}^{N\times C}$ 只是经过行掩码和缩放处理的 R。它也只有 C 列。因此，$\text{rank}(\tilde{R}_{\mathcal{B}}) \leq C$，进而 $\text{rank}(\tilde{R}_{\mathcal{B}} \tilde{R}_{\mathcal{B}}^{\top}) \leq C$。

代入这些界：

$$\text{rank}(\zeta_{\mathcal{B}}(K)) \leq C + C = 2C. \tag{157}$$

这就确立了正文中的公式（81）。

最后，关于协方差结构：瞬时协方差张量由噪声向量化结果的外积的期望构成。由于噪声矩阵 $\zeta_{\mathcal{B}}$ 的每一次实现都严格位于由 R 和 $R_{\mathcal{B}}$ 的列所张成的子空间内（它们是 R 的列空间的子集），因此噪声被限制在子空间 $\mathcal{V} = \text{span}(\text{cols}(R)) \otimes \text{span}(\text{cols}(R))$ 内。相关生成子空间的维度至多为 C。 □

E.3 物理意义

这一推导证实，在此机制下，随机梯度下降噪声并非各向同性全秩扩散，而是严格作用于由 C 输出逻辑值定义的任务相关子空间内。即使网络宽度为 $N \to \infty$，噪声秩仍受限于 $2C$，从而确保学习到的核的低秩结构能够抵御随机波动。

F 扩展到自监督学习

我们的秩压缩 理论并不局限于监督回归。在本附录中，我们将展示，自监督学习（SSL），尤其是以线性自编码器或重建任务的形式进行时，遵循完全相同的谱动力学，自然会呈现主成分分析（PCA）的特性。

F.1 自监督学习的形式化

考虑从表示中重构输入 $X \in \mathbb{R}^{N\times D}$ 的任务。目标矩阵 Y 本质上就是 X 本身（或其增强视图）。损失函数变为：

$$\mathcal{L}(W,\Phi) = \frac{1}{2}\|X - W\Phi\|_F^2 + \frac{\lambda}{2}\|W\|_F^2 + \frac{\mu}{2}\|\Phi\|_F^2. \tag{158}$$

这是经典的矩阵分解设定。

F.2 自监督学习核的动力学

遵循与定理 1 相同的推导过程，我们通过快速平衡假设消去解码器 W：

$$W^*(\Phi) = X\Phi^\top(\Phi\Phi^\top + \lambda I)^{-1}. \tag{159}$$

残差矩阵 M 变为重建误差协方差。核$K = \Phi^\top\Phi$ 的流由输入协方差矩阵 $\Sigma_X = X^\top X$驱动。

定理 26（通过谱动力学实现 PCA）。在自监督设置中，核 $K(t)$ 演化至其特征子空间与数据协方差 Σ_X的主成分对齐。稳态特征值 $\{\mu_i^*\}$ 由 Σ_X的特征值 $\{\lambda_i^X\}$ 决定：

$$\mu_i^* = \max\left(0, \frac{\lambda_i^X}{\mu} - \lambda\right). \tag{160}$$

证明。在自编码器模式下，第 i个模态的“信号强度” s_i 恰好等于数据在该方向上的方差，即特征值 λ_i^X。将 $s_i = \lambda_i^X$代入我们的截断定律（定理 4）即可直接得出结果。

□

F.3 对基础模型的启示

这一结果为实证观察提供了理论依据：自监督学习预训练会学习“主导”特征，同时抑制噪声。

1. **去噪：** 低方差方向（噪声）对应较小的 λ_i^X。如果 $\lambda_i^X < \lambda\mu$，这些方向就会被完全丢弃（$\mu_i^* = 0$）。表示 Φ 实际上执行了**硬阈值奇异值分解**。2. **维度坍塌：** 这解释了自监督学习（SSL）中常见的“维度坍缩”现象：如果超参数未得到正确调整，过度正则化 μ 会抬高水位，从而截断信息特征。

因此，我们的特征学习极限在单一谱动力学原理下统一了监督式学习和自监督学习：**核与目标结构中能量最高的模态对齐**，无论该目标是外部标签还是内部数据相关性。

G 精确总体风险分析

在本附录中，我们分析总体极限（$N \to \infty$）下的泛化性能。我们不再为依赖数据的核推导宽松的集中界，而是利用第11节推导出的精确算子动力学，解析地刻画偏差-方差权衡。

G.1 演化核的泛化误差

考虑目标函数 $f^* \in L^2(\rho)$ 在时变积分算子 T_t 的正交归一特征基 $\{\psi_i(t)\}$ 下的分解：

$$f^* = \sum_{i=1}^{\infty} a_i(t)\psi_i(t), \quad a_i(t) = \langle f^*, \psi_i(t) \rangle_{L^2}. \tag{161}$$

通过核岭回归（或流极限）得到的预测器 f_t 对目标函数起到谱滤波器的作用。均方误差（风险）可按标准方式分解：

$$\mathrm{Risk}(t) = \|(I - \mathcal{S}_t)f^*\|_{L^2}^2 + \frac{1}{N}\mathrm{Tr}(\mathcal{S}_t^2 \Sigma_{\mathrm{noise}}), \tag{162}$$

其中，$\mathcal{S}_t = T_t(T_t + \lambda I)^{-1}$ 是收缩算子，Σ_{noise} 是噪声协方差（为简便起见，假设其为各向同性 $\sigma_\epsilon^2 I$）。在特征基下展开可得显式形式：

$$\mathrm{Risk}(t) = \sum_{i=1}^{\infty} \underbrace{\left(\frac{\lambda}{\mu_i(t) + \lambda}\right)^2 |a_i(t)|^2}_{\mathrm{Bias}_i(t)} + \frac{\sigma_\epsilon^2}{N} \underbrace{\sum_{i=1}^{\infty} \left(\frac{\mu_i(t)}{\mu_i(t) + \lambda}\right)^2}_{\approx \mathcal{N}_{\mathrm{eff}}(t)}. \tag{163}$$

G.2 代入谱截断定律

现在将**通用秩压缩** 结果应用于这一风险特征。在稳态 $t \to \infty$下，假设系统与任务对齐，核谱 $\{\mu_i^*\}$ 遵循定理 6 推导出的截断定律（已调整为适用于总体算子）：

$$\mu_i^* = \max\left(0, \sqrt{\frac{\lambda \sigma_i}{\mu}} - \lambda\right), \tag{164}$$

其中，σ_i 表示第 i 个模态的信号强度。我们区分两种机制：

情况 1：噪声子空间（$\sigma_i \leq \lambda\mu$）。在此机制下，相对于正则化乘积 $\lambda\mu$，标签信号过弱。动力学将特征值驱动至零：$\mu_i^* = 0$。

- **方差：** 方差项的贡献消失：$\mathrm{Var}_i \to 0$。模型实际上忽略了这一维度。
- **偏差：** 偏差项达到最大值：$\mathrm{Bias}_i \to |a_i|^2$。模型无法捕捉目标的这一组成部分。

含义： 这证实了“可达性”约束。f^* 的高频或正交分量会永久丢失，但不会导致过拟合。

情形 2：任务子空间（$\sigma_i > \lambda\mu$）。在此机制下，$\mu_i^* > 0$。核会扩展以捕获这些模态。

- 偏差被因子 $\left(\frac{\lambda}{\mu_i^* + \lambda}\right)^2 < 1$抑制。
- 方差贡献非零，但仅限于这些活跃模态。

因此，学习到的核的有效维数 $\mathcal{N}_{\mathrm{eff}}$ 大致受活跃任务模态的数量（秩 $\leq C$）限制，与环境输入维数 D无关。

G.3 与静态核（神经正切核）的比较

将其与静态神经正切核机制进行对比颇有启发。对于静态核，特征值μ_i^{NTK}由初始化决定，通常按幂律 $\mu_i \propto i^{-\nu}$ 衰减（取决于激活函数的光滑度）。

- **静态（神经正切核）：** 有效维数 $\mathcal{N}_{\text{静态有效维数}} = \sum \frac{\mu_i}{\mu_i + \lambda}$可能非常大（随 N 增长），导致估计方差很高（即“过参数化”代价）。

- **动态（任务驱动）：** 该流执行**硬模型选择**，将尾部清零：

$$\mathcal{N}_{\text{eff}}^{\text{flow}} \approx C \ll \mathcal{N}_{\text{eff}}^{\text{static}}. \tag{165}$$

这种有效维数的大幅降低由核常微分方程的物理机制驱动，解释了特征学习在低秩任务中具有更优泛化能力的原因。